

Tim Hector

Smart Speaker im Dialog

Empirische Linguistik / Empirical Linguistics



Herausgegeben von
Wolfgang Imo und Constanze Spieß

Band 21

Tim Hector

Smart Speaker im Dialog



Sprachliche Praktiken mit Voice User Interfaces

DE GRUYTER

Der Peer Review wird in Zusammenarbeit mit themenspezifisch ausgewählten externen Gutachterinnen und Gutachtern durchgeführt. Unter <https://www.degruyterbrill.com/serial/el-b/html> finden Sie eine aktuelle Liste der Expertinnen und Experten, die für die Reihe begutachtet haben.

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 262513311 – SFB 1187

Finanziert aus dem DFG-geförderten Open-Access-Publikationsfonds der Universität Siegen für Monografien

ISBN 978-3-11-157383-0

e-ISBN (PDF) 978-3-11-157433-2

e-ISBN (EPUB) 978-3-11-157489-9

ISSN 2198-8676

DOI <https://doi.org/10.1515/9783111574332>



Dieses Werk ist lizenziert unter der Creative Commons Namensnennung 4.0 International Lizenz. Weitere Informationen finden Sie unter <http://creativecommons.org/licenses/by/4.0/>.

Die Creative Commons-Lizenzbedingungen für die Weiterverwendung gelten nicht für Inhalte (wie Grafiken, Abbildungen, Fotos, Auszüge usw.), die nicht im Original der Open-Access-Publikation enthalten sind. Es kann eine weitere Genehmigung des Rechteinhabers erforderlich sein. Die Verpflichtung zur Recherche und Genehmigung liegt allein bei der Partei, die das Material weiterverwendet.

Library of Congress Control Number: 2025937015

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

© 2025 bei den Autorinnen und Autoren, publiziert von Walter de Gruyter GmbH, Berlin/Boston, Genthiner Straße 13, 10785 Berlin

Dieses Buch ist als Open-Access-Publikation verfügbar über www.degruyterbrill.com.

Korrektur: Dr. Moritz Bensch

Satz: Integra Software Services Pvt. Ltd.

Druck und Bindung: CPI books GmbH, Leck

www.degruyterbrill.com

Fragen zur allgemeinen Produktsicherheit:

productsafety@degruyterbrill.com

Danksagung

Dieses Buch ist die leicht überarbeitete Fassung der Arbeit, die am 12. April 2024 von der Philosophischen Fakultät der Universität Siegen als meine Dissertationschrift im Fach Angewandte Sprachwissenschaft angenommen wurde. Bei aller Mühe hat es mir große Freude bereitet, diese Schrift zu verfassen – vor allem dank der vielen Menschen, die mich dabei begleitet und unterstützt haben. Ihnen möchte ich an dieser Stelle von Herzen danken und einige von ihnen namentlich nennen, etliche sogar mehrfach.

Mein erster und ganz besonderer Dank gilt meinem Betreuer Stephan Habscheid, der mich mit umfassender Fachkompetenz sowie großer Professionalität und Integrität durch die Promotion begleitet und mich durchgängig unterstützt hat. Ebenso herzlich danke ich meiner Zweitbetreuerin Karola Pitsch von der Universität Duisburg-Essen für ihre wertvollen Anregungen und die intensive Auseinandersetzung mit meiner Arbeit.

Wolfgang Imo und Constanze Spieß danke ich für die Aufnahme der Arbeit in die Reihe „Empirische Linguistik“. In diesem Zusammenhang danke ich auch den anonymen Gutachter*innen für die gründliche Durchsicht und die hilfreichen Kommentare zu einer früheren Version dieses Texts sowie Erika Linz für ihre Ratschläge. Bei den Mitarbeiter*innen bei De Gruyter Brill bedanke ich mich für die Unterstützung und Beratung bei der Vorbereitung meines Buchs für die Veröffentlichung. Moritz Bensch danke ich für das professionelle Korrektorat. Der von der DFG geförderte Sonderforschungsbereich (SFB) 1187 „Medien der Kooperation“ sowie der ebenfalls DFG-geförderte Open-Access-Publikationsfonds der Universität Siegen für Monografien haben meine Forschung bzw. deren Publikation finanziert.

Mein Promotionsstudium begann in etwa zeitgleich mit meiner Tätigkeit im Teilprojekt B06 „Un/erbetene Beobachtung in Interaktion: Intelligente Persönliche Assistenten“ in der zweiten Laufzeit des SFBs „Medien der Kooperation“ (2020–2023). Das Projekt leiten gemeinsam Stephan Habscheid und Dagmar Hoffmann und ich hatte das Glück, mit ihnen und einem großartigen Team an wissenschaftlichen Angestellten zusammenzuarbeiten: David Waldecker, Christine Gebhard (geb. Hrnca), Niklas Strüver und Kathrin Englert. Ihnen allen danke ich für die inspirierende, stets konstruktive und zwischenmenschlich außergewöhnlich angenehme Zusammenarbeit. Für die Unterstützung insbesondere bei der Arbeit an den Transkripten danke ich unseren studentischen Mitarbeiter*innen Franziska Niersberger-Guéye, Franziska Petri, Sarah Diehl, Marie Dienst, Johanna Klein, Katharina Meiers, Leonie Tittel, Chris Dangelmaier und Finnja Strunk. Außerdem danke ich allen anonymen Studienteilnehmer*innen.

Der SFB bot ein ideales wissenschaftliches Umfeld für meine Arbeit – anregend, bereichernd und kollegial. Im integrierten Graduiertenkolleg des SFBs

(MGK) konnte ich meine Arbeit regelmäßig diskutieren und wertvolle Rückmeldungen erhalten. Allen Mitgliedern des SFBs gilt mein ausdrücklicher Dank – erwähnen möchte ich an dieser Stelle Tanja Aal, Hendrik Bender, Marcus Burkhardt, Jason Chao, Susanne Förster, Carolin Gerlitz, Sam Hind, Max Kanderske, Julia Kurz, Danny Lämmerhirt, Gaia Mosconi, Kevin Onland, Sheree May Plinzsaßmannshausen, Inga Schuppener, Erhard Schüttpelz, Tatjana Seitz, Yarden Skop, Niklas Strüver, Fernando van der Vlist und Daniela van Geenen. Mein großer Dank geht auch an die Menschen in der Koordination des SFBs, die mich bei zahlreichen Anliegen hilfsbereit und professionell unterstützt haben. Dafür danke ich Anja Höse, Timo Kaerlein, Karina Kirsten, Susanne Kokel, Johannes Schick und Dominik Schrey.

Einige Menschen haben mich während der gesamten Promotionszeit fachlich intensiv begleitet, mir wesentliche Denkanstöße gegeben und viele meiner Textentwürfe kritisch kommentiert. Besonders danken möchte ich dafür Niklas Strüver, Benedikt Merkle, Yarden Skop, Susanne Förster und Katharina Hector. Ihnen sowie Laura und Brian Krämling und Lenia Zinßer-Kraatz danke ich außerdem für ihre Mühen mit dem ersten Lektorat meiner Arbeit. Ein großer Dank geht außerdem an Sina van Oostrum, nicht nur für ihre Hilfe bei der Bearbeitung von Abbildungen. Wertvolle Diskussionen hatte ich auch mit Anastasia-Patricia Och, Viviane Börner und Katharina Hofius sowie mit den Teilnehmer*innen an den Treffen des „Doktorand*innennetzwerks Gesprächsforschung“. Eine Promotion begonnen hätte ich nicht ohne den Rat und die Unterstützung von Katharina König aus Münster, der ich den Einstieg in die Arbeit mit gesprochener Sprache und auch darüber hinaus viel verdanke. Christine Gebhard stand mir über die Zusammenarbeit im Projektkontext hinaus auch als Mentorin in allen Stadien der Promotionszeit zur Seite, wofür ich ihr sehr dankbar bin.

Meinen Freund*innen danke ich für ihre Ermutigung und ihre Anteilnahme: Bereits seit vielen Jahren begleiten mich Leonard Conradi, Niclas Helten, Laura und Brian Krämling, Susan Krautmacher, Lisa Leuchtenberg, Johannes Stottmeister, Daisy Törner, Sina van Oostrum, Lenia Zinßer-Kraatz und Lukas Kraatz. Außerdem danke ich in diesem Sinne auch Benedikt Merkle und Zélie Jouenne, Susanne Saliger und Frederike Schmitz, Kurt Jansson, Johanna Niesyto, Nicole Ebber und Moritz Rahm sowie noch einmal Niklas Strüver für ihr Interesse und ihre Resonanz während der Entstehung dieser Arbeit.

Schließlich danke ich meiner gesamten Familie für ihre bedingungslose Unterstützung auf meinem beruflichen Weg und bei meinem Promotionsvorhaben. Meiner Mutter Ulrike und meinen Geschwistern Kim, Katharina, Moritz und Paul Hector möchte ich diese Arbeit daher widmen.

Inhaltsverzeichnis

Danksagung — V

1 Einleitung, Forschungsfragen, Aufbau der Arbeit — 1

2 Theorie und Begriffe — 8

- 2.1 Praxeologische und medientheoretische Grundlagen — 8
 - 2.1.1 Soziale Praxis und Ko-Operation — 10
 - 2.1.2 Soziale Praktiken: Kleinste Einheit des Sozialen — 13
 - 2.1.3 Sprachliche Praktiken: Ressourcen für Bausteine der Interaktion — 15
 - 2.1.4 Medien- und Interface-Praktiken — 18
 - 2.1.5 Zwischenfazit Praxeologie und Medientheorie — 25
- 2.2 Zum Verhältnis von Mensch und Maschine — 28
 - 2.2.1 Interaktion und Gespräch — 29
 - 2.2.2 Dialog zwischen Mensch und Maschine — 31
 - 2.2.3 Charakteristika von Mensch-Maschine-Dialogen — 34
 - 2.2.4 Beteiligung maschineller Entitäten an der Praxis — 40
 - 2.2.5 Zwischenfazit zum Verhältnis von Mensch und Maschine — 43
- 2.3 Aneignung und Domestizierung — 44
 - 2.3.1 Kommunikative Aneignung von Medien — 46
 - 2.3.2 Zum Domestizierungsansatz — 50
 - 2.3.3 Zwischenfazit Aneignung und Domestizierung — 66

3 Forschungsüberblick und Gegenstandsbestimmung — 68

- 3.1 Mensch und Maschine in der empirischen Linguistik — 68
 - 3.1.1 Chatbots — 69
 - 3.1.2 Embodied Conversational Agents — 72
 - 3.1.3 Humanoide Roboter — 74
 - 3.1.4 Telefonbasierte Sprachdialogsysteme — 76
- 3.2 Smart Speaker — 78
 - 3.2.1 Historische Entwicklung der Sprachtechnologie — 78
 - 3.2.2 Funktionsweise und technischer Hintergrund — 82
 - 3.2.3 Smart Speaker als Intermediäre und Plattformen — 89
 - 3.2.4 Nutzung und Aneignung von Smart Speakern — 93
 - 3.2.5 Smart Speaker und Gender — 101
 - 3.2.6 Datenschutz, Datenverwertung und Privatsphäre — 104
- 3.3 Untersuchte Smart Speaker-Modelle — 110
 - 3.3.1 Amazon Echo — 111

3.3.2	Google Home und Google Nest —	117
3.3.3	Apple HomePod —	120
4	Methodologische Einordnung —	122
4.1	Ethnomethodologie als praxeologische Methodologie —	122
4.2	Sprache-in-Interaktion —	126
4.2.1	Konversationsanalyse —	127
4.2.2	Gesprächsanalyse —	130
4.2.3	Interaktionale Linguistik —	131
4.2.4	Gesprochene-Sprache-Forschung —	134
4.3	Sprache und Ethnografie —	135
4.4	Multimodalität —	139
4.5	Zwischenfazit —	145
5	Methodik: Projektkontext, Untersuchungsdesign, Erhebung, Datentypen und -aufbereitung —	148
5.1	Projektvorstellung und Untersuchungsdesign —	148
5.2	Datentyp I: Videoaufzeichnungen der Smart-Speaker-Ersteinrichtung —	150
5.3	Datentyp II: Audioaufzeichnungen der Smart-Speaker-Nutzung —	152
5.4	Datenaufbereitung —	157
5.5	Untersuchte Haushalte und Auswertung —	159
5.5.1	Haushalt 1: Amazon: Echo Dot mit „Alexa“ (Faßbender) —	163
5.5.2	Haushalt 7: Apple: HomePod mit „Siri“ (Waldes) —	166
5.5.3	Haushalt 8: Google Nest mit „Google Home“ (Matthäi) —	168
5.6	Transkriptionskonventionen —	170
6	Analyse sprachlicher Praktiken —	174
6.1	Organisation dyadischer VUI-Dialoge —	174
6.1.1	Anredeformen im VUI-Dialog —	177
6.1.2	Sequenzialität und Sequenzorganisation —	204
6.1.3	Turns und Turn-Taking zwischen Mensch und VUI —	220
6.1.4	Reparaturen —	239
6.1.5	Zwischenfazit zu dyadischen VUI-Dialogen —	263
6.2	VUI-Dialoge in Mehrparteienkonstellationen —	266
6.2.1	Sequenzielle Einbindung von VUI-Dialogen —	270
6.2.2	VUIs im Zentrum sozialer Praxis —	285
6.2.3	Sprachliche Praktiken in der (Nicht-)Beteiligung von VUIs —	318

6.2.4	Zwischenfazit zu VUI-Dialogen in Mehrparteienkonstellationen — 371

7 Schlussbetrachtung — 374

7.1	Zusammenfassung der Ergebnisse — 374
7.1.1	Dyadische VUI-Dialoge — 374
7.1.2	VUI-Dialoge in Mehrparteienkonstellationen — 379
7.2	Diskussion der Ergebnisse — 381
7.2.1	Sprachliche Praktiken – Medienpraktiken im Interface — 382
7.2.2	Sprachliche Praktiken in VUI-Dialogen als Sprachregister? — 383
7.2.3	Zur ‚Beteiligung‘ von VUIs an Praxis und Gespräch — 388
7.2.4	Domestizierung im <i>display</i> sprachlicher Praktiken — 389
7.2.5	Smart Speaker im bewohnten Raum — 393
7.2.6	Methodologische Reflexion und Anschlussmöglichkeiten — 396
7.3	Ausblick — 400

8 Literatur — 407

Register — 453

1 Einleitung, Forschungsfragen, Aufbau der Arbeit

Als im *Business Insider* im November 2022 zu lesen war, dass der stimmbasierte digitale Assistent Alexa ein Milliardenverlustgeschäft sei und die erhofften Gewinne bei Weitem nicht einfahren konnte (vgl. Kim 2022), schien damit das Ende stationärer Sprachassistentensysteme eingeläutet zu sein. *Ars Technica* berichtete, frühere Angestellte von Amazon hätten Alexa als „colossal failiure of imagination“ bezeichnet (Amadeo 2022). Ein Zurückfahren von Investitionen würde, so die Analyse der Tech-Expert*innen, zum Ausbleiben von Innovationen führen und somit früher oder später zwangsläufig dazu, dass Updates und notwendige Infrastrukturen für die Geräte nicht mehr bereitgestellt würden (vgl. Ashworth 2022). Dafür spricht, dass sich die Nutzung stationärer Sprachassistentensysteme wie Amazons Echo Dot oder der von Apple vertriebene HomePod mit Siri nur auf wenige Anwendungsfälle beschränkt (vgl. Ammari et al. 2019) und die Verbreitung von Smart Speakern in Deutschland bei ca. 25 Prozent stagniert (vgl. Statista 2019; 2025b). Global ausgerichtete Prognosen sagen nur noch einen leichten Zuwachs beim Absatz der Geräte bis 2028 vorher – dieser soll allerdings die Absatzzahlen aus den Jahren 2019 bis 2022 nicht mehr übertreffen (vgl. Statista 2025c). Der Aufwind, den der Diskurs um KI-Anwendungen insbesondere durch die Verfügbarkeit des Sprachgenerators ChatGPT erfahren hat, wirkt sich zwar auch auf Smart Speaker aus (vgl. Malik 2023), allerdings ist eine Tendenz zur Marktsättigung erkennbar, die in Kombination mit der aufgedeckten Unwirtschaftlichkeit ein baldiges Ende der Geräte in der bisherigen Form in den Augen vieler wahrscheinlich werden lässt (vgl. Mansholt 2022). Warum also 2025 noch eine Studie über Smart Speaker vorlegen? Dafür sprechen im Wesentlichen drei Gründe.

Erstens sind Smart Speaker und die in ihnen verbauten stimmbasierten Benutzer*innen-Schnittstellen (Voice User Interfaces, VUIs) eine trotz dieser Ankündigungen weit verbreitete Benutzer*innen-Schnittstelle: Ein Viertel aller Befragten in Deutschland sind Nutzer*innen eines solchen Geräts und insofern sind Untersuchungen zum tatsächlichen sprachlichen Gebrauch der Interfaces schon aufgrund ihrer Verbreitung von Relevanz, auch dann, wenn sie als Phänomen wieder verschwinden sollten. Es ist zwar ohnehin anzunehmen, dass VUIs als Interface-Typ auch unabhängig von ihrer Existenz in Smart Speakern der Betreiberfirmen Amazon, Google und Apple in der einen oder anderen Form Bestand haben werden. Falls nicht, kann in deren Untersuchung nachträglich ein wichtiger medien- und sprachhistorischer Beitrag zur Rolle von Interfaces im Umgang mit Medientechnologien gesehen werden.

Zweitens ist die Untersuchung des Austauschs zwischen Menschen und Maschinen unter linguistischen und interaktionstheoretischen Gesichtspunkten interessant: Die Analyse von „Sprache-in-Interaktion“ (Imo 2013), konkret konversationeller, dialogischer und an der Oberfläche gesprächsähnlicher Sprachformen aus dem Bereich der Mündlichkeit, erlaubt Rückschlüsse auf Sprache im Gebrauch auch außerhalb von Mensch-Maschine Dialogen. So zeigt etwa Pitsch (2015) am Beispiel von Ko-Konstruktionen, dass erlernte und im sprachlichen Regelwissen der Sprecher*innen verankerte gesprächsorganisatorische Strategien in Mensch-Maschine-Dialogen aufgebrochen werden. So werden über Mensch-Maschine-Dialoge einzelne konversationelle Phänomene untersuchbar – als Muster, an denen sich Sprecher*innen im Mensch-Maschine-Austausch orientieren und sie damit für die Analyse freilegen (siehe auch Pitsch 2023).

Drittens sind Smart Speaker Bestandteil von „Smart Homes“ – Wohnumgebungen, die mit digitalen und vernetzten Technologien ausgestattet sind (vgl. Wilson/Hargreaves/Hauxwell-Baldwin 2015: 463–464). Bestandteile eines Smart Homes müssen sowohl miteinander als auch mit der Außenwelt über das häusliche Netzwerk verbunden sein, um zu funktionieren. Diese Vernetzung erfordert Interoperabilität der verschiedenen Dienste (z. B. Regulierung der Heizung, Öffnen und Schließen der Vorhänge, Energie- und Heizungsmanagement, Ein- und Ausschalten von Lichtern). Die häufigste Verwendungsweise von VUIs ist die Steuerung solcher Smart Home-Anwendungen (vgl. Bitkom 2022: 28). Das liegt nicht nur an der „hands free“-Steuerung von VUIs, was für bestimmte Zwecke vorteilhaft ist (z. B. für das Einstellen von Timern in der Küche beim Kochen oder das Schließen der Vorhänge vom Bett aus), sondern, so argumentiert Strüver (2023a), auch daran, dass sich Dienstanbieter wie Amazon als zentrale Schnittstelle zur Herstellung von Interoperabilität zwischen den einzelnen Geräten und Herstellern etabliert haben und, so ist anzunehmen, dies zukünftig weiter ausbauen wollen, um diesen Status weiter zu manifestieren. So wird für die Nutzer*innen die Handhabung und Verwaltung über nur ein einziges Gerät bzw. eine einzige Anwendung ermöglicht, zugleich entstehen Schnittstellen für die Hersteller. Insofern ist die linguistische Untersuchung von Smart Speakern auch vor dem Hintergrund der Integration sprachlicher Praktiken in die zunehmende Technologisierung und Vernetzung des Haushalts lohnenswert.

Im Zentrum dieser Arbeit stehen also, wie bereits deutlich geworden ist, stationäre Sprachassistenzsysteme, die hier als *Smart Speaker* bezeichnet werden. So sollen Verwechslungen mit Begriffen wie „Intelligenter Persönlicher Assistent“, „Voice Assistant“, „Sprachdialogsystem“ oder „Sprachassistent“ u. a. vermieden werden, die zwar auf diesen, aber auch auf anderen Geräten installiert und insofern eher als Software zu konzeptualisieren sind, da sie keine eigene Hardware haben, sondern die Audio-Systeme und teilweise auch Bildschirme von

Smartphones, Tablets, Laptops und anderen Geräten mitnutzen. Eine deutschsprachige Alternative wäre „Intelligenter Lautsprecher“, dieser Begriff hat sich allerdings nicht durchgesetzt. Smart Speaker bestehen aus verschiedenen Bestandteilen, darunter am wichtigsten: Lautsprecher, Mikrofone und Recheneinheiten, die u. a. die lokale Erkennung gesprochener Sprache durchführen und eine Verbindung mit dem Internet herstellen. Diese Elemente sind in einem kompakten Gerät zusammengefasst und kombinieren auf Grundlage der verfügbaren Hardware verschiedene technische Funktionen, v. a. Spracherkennung und Sprachsynthese, Natural-Language-Processing und Informationsgewinnung – d. h., sie stellen ein VUI bereit. Smart Speaker sind also alleinstehende Geräte mit verschiedenen Komponenten in einem Gehäuse, und zwar in mehrfacher Hinsicht: erstens als materielle „Deckelhaube“ (Kittler 2002: 24–25) die den Blick auf die Bestandteile im ‚Innenen‘ sowie die ablaufenden Prozesse versperrt (vgl. Kittler 1986: 5) und den Smart Speaker von außen als unveränderliches Ganzes erscheinen lässt. Zweitens lassen sie sich mit einem „strategischen weiten Sinne“ (Heilmann 2019: 40) als Gehäuse betrachten: Das Gehäuse erscheint dann, wie der Medienwissenschaftler Till Heilmann (2019) vorschlägt, heuristisch betrachtet als „Grenzfläche“, welche die Computerbestandteile einerseits physisch umhüllt, diese aber andererseits mit der Umgebung sowie mit globalen Datacentern vernetzt und konzeptionell zudem bei der materiellen Umhüllung und Verbindung nicht stehen bleibt, sondern auch andere Prozesse der Vermittlung zwischen Computer und Umwelt einschließt (vgl. Heilmann 2019: 41). Das Gehäuse eines Smart Speakers ist so betrachtet eine physische Grenzfläche, die bidirektional durchlässig ist für akustische, unidirektional auch für visuelle und taktile Signale, denn das Gehäuse reagiert nicht nur auf Befehle, sondern auch auf Berührungen und sendet selbst durch entsprechend modifizierte Lichtelemente Zeichen. Das Gehäuse eines Smart Speakers als Grenzfläche verbindet also die Umgebung, in der er platziert ist, über diverse Leitungen und Verbindungen sowohl mit anderen Elementen der ‚Innenwelt‘ als auch mit der ‚Außenwelt‘ des Haushalts. Diese Verbindungen sind notwendig für das vollständige Funktionieren des Smart Speakers und Bestandteil der Hard- und Software-Kombinationen, die sich im Gehäuse vollziehen.

Die Vermittlung durch das Gehäuse zwischen Umwelt und Computer erscheint im Falle der Smart Speaker besonders tiefgreifend: Die stimmliche Ein- und Ausgabe mit einem VUI ist nach Erstinstallation als Modus für die Bedienung des Geräts vorgesehen, obschon sie stellenweise durch die Nutzung des Smart-

phones ‚unterstützt‘ wird.¹ Das Gehäuse verbirgt einerseits die verschiedenen Komponenten und Prozesse und ist andererseits ‚kommunikativ‘, es vermittelt Gehörtes ins Internet und gibt Informationen gesprochensprachlich oder durch andere akustische und visuelle Signale wieder. Besonders durch die mündlichen Wiedergaben entstehen – in unterschiedlicher Ausprägung – ‚Persönlichkeiten‘ des Smart Speakers: Statt visueller werden hier akustische Charakteristika, Metaphern und Ästhetiken gebraucht, um durch symbolische Ausdrücke Software und Nutzer*innen miteinander zu verbinden. Diese Verbindung wird mit einer Kontinuität hinterlegt, die auf der Illusion einer persönlichen Beziehung, einer „Persona“ aufbaut (vgl. Natale/Cooke 2021: 1009; siehe auch Lotze 2016: 63). Diese werden von den Herstellern gezielt zum Aufbau von ganzheitlich erscheinenden ‚persönlichen Assistent*innen‘ genutzt und entsprechend vermarktet (vgl. Dickel/Schmidt-Jüngst 2021) und als solche finden sie auch Eingang in den öffentlichen Diskurs (vgl. Lind 2021).

Zu diesen Eigenschaften gehören, so wie die gängigen Hersteller die Geräte bisher applizieren, auch Merkmale der (gesprochenen) Sprache und des Sprechens – und damit jedenfalls die Bedingung der Möglichkeit für eine Beteiligung an der sozialen Praxis und mithin an Gesprächen. Doch wie vollzieht sich empirisch der Austausch mit Smart Speakern bzw. VUIs, die im Fokus der Arbeit stehen sollen? Die VUIs als der ‚Gespräch führende‘ Bestandteil des Smart Speakers vermitteln die Erfassung und Wiedergabe von Informationen in gesprochensprachlicher Form (Stimmein- und -ausgaben). Kann ein VUI damit als Gesprächspartner*in fungieren? Kann es in Alltagssituationen als dritte oder vierte Person ‚mit am Tisch sitzen‘? Wie wird es konversationell in laufende Interaktionssituationen und somit in die soziale Praxis eingebunden? Welche Dynamiken (sozial, zeitlich, räumlich) lassen sich dabei beobachten, in die VUIs eingebunden sind und die sie mit hervorbringen? Daraus ergeben sich folgende **Leitfragen**, die im Rahmen der folgenden Arbeit beleuchtet werden sollen:

1. Welche sprachlichen Praktiken zeigen sich im Prozess der Domestizierung von stationären Sprachassistenzsystemen mit VUIs? Wie gehen Nutzer*innen dabei mit Störungen um?
2. Inwieweit lassen sich diese sprachlichen Praktiken als Abwandlungen bereits untersuchter sprachlicher Praktiken in zwischenmenschlichen Interaktionen beschreiben, inwieweit sind sie als emergente sprachlich-infrastrukturelle Elemente und Fortsetzung kommunikativer Routinen zu verstehen?

¹ Es existieren auch Modelle mit integriertem Bildschirm und Kamera, die allerdings nicht Gegenstand der vorliegenden Arbeit sind.

3. Wie wird der Gebrauch von Smart Speakern sprachlich in die soziale Praxis und in Gespräche eingebunden und welchen Beteiligungsstatus schreiben die Anwender*innen der Geräte diesen zu?

Die Arbeit will damit neben einer empirischen Betrachtung dieser für das Deutsche nicht systematisch untersuchten Dialogform auch einen grundsätzlichen bzw. konzeptionellen Beitrag zur Erforschung von mündlichen Dialogen mit Maschinen leisten, der auch für weitere Untersuchungen zu moderneren und auf „Künstlicher Intelligenz“ basierenden Systemen herangezogen werden kann. Dies soll dadurch ermöglicht werden, dass die Analyse nicht auf die technischen Möglichkeiten der in der vorliegenden Arbeit untersuchten Modelle limitiert ist, sondern deren konversationelle Aneignung bzw. *Domestizierung* ins Zentrum der Betrachtung rückt. Damit wird das Verhältnis von Nutzer*innen, Sprache bzw. Sprachgebrauch und Interface praxeologisch konturiert. Eine methodologische Besonderheit der Arbeit liegt insofern darin, dass Konzepte aus der Medien- und Kommunikationswissenschaft in die Gesprächsforschung hinein vermittelt und durch empirische Untersuchungen unterfüttert werden, die auf Daten basiert, die aus Alltagssituationen der Nutzer*innen stammen. Dazu eignet sich der Methodenkanon der (multimodalen) Interaktionsanalyse bestens. Die Arbeit schreibt damit auch eine durchaus auch mit Bezügen zur Gesprächsforschung etablierte Forschungstradition zur Medienrezeption und -aneignung fort, die sich in den 1990er- und 2000er-Jahren mit Fernsehen befasst hat und die zentrale Rolle konversationeller Äußerungen dabei erkannte (vgl. Holly/Püschel/Bergmann 2001; Holly/Püschel 1993a). Die Arbeit greift ferner, wenn auch als Nebenaspekt, auch die Diskussion um ein spezifisches sprachliches Register, „Computer Talk“ (Zoeppritz 1985; Krause/Hitzenberger 1992; für eine aktuelle Übersicht siehe auch Lotze 2025), wieder auf, obschon diesbezüglich sowohl konzeptionelle wie auch methodologische Einschränkungen bestehen, wie zu reflektieren sein wird.

Die Arbeit ist – wie sich schon durch die bisher verwendeten Termini andeutet – praxistheoretisch fundiert. Im weiteren Verlauf gehe ich mit Hirschauer (2016) davon aus, dass neben Menschen auch Gegenstände (oder Körper, Körperhaltungen, Raumumgebungen, Architekturen usw.) Beteiligte an der Praxis sein können. Da Gespräche eine (mehr oder weniger gerichtete) Form des Praxisvollzugs sind (siehe auch Goffman 1979: 6–7), kann eine Beteiligung von Smart Speakern am Vollzug dieser nicht *a priori* ausgeschlossen werden. Vielmehr stellt sich die Frage, wie die Beteiligten damit umgehen, dass sie einerseits mit einem Computer bzw. dessen Grenzfläche (sensu Heilmann 2019) – und damit auch der Außenwelt – in Verbindung stehen, zugleich aber die sozialen Regeln und Erwartungen für das Führen von

Gesprächen bedienen. Dabei befinden sich Anwender*innen und Smart Speaker gleichermaßen in einem Prozess kommunikativ vollzogener Domestizierung.

Die begrifflichen Grundlagen und Konzepte für diese Aspekte werden in **Kapitel 2** diskutiert. Dabei wird auf die praxistheoretische Grundierung sowie ein sich daraus ergebendes praxeologisches Verständnis von Medien ebenso eingegangen wie auf einen darauf aufbauenden Begriffsapparat zur Beschreibung der jeweiligen dialogischen Konstellationen. Das Kapitel fasst auch Debatten zu Verhältnis und Austausch zwischen Mensch und Maschine zusammen. Darüber hinaus werden Grundlagen der Medienaneignung und -domestizierung für den Kontext der Arbeit vorgestellt. Auf diesen Teil folgt ein Überblick über den Stand der Forschung zu Mensch-Maschine-Dialogen einschließlich einer historischen Betrachtung und eines Überblicks über die Funktionsweise (**Kapitel 3**). Das Kapitel leistet in diesem Zuge auch eine weitere Eingrenzung des Gegenstands. Ein Schwerpunkt liegt in diesem Kapitel auf der Darstellung des aktuellen Forschungsstands zu stationären Sprachassistenzsystemen. Das Kapitel schließt eine Funktions- und Modellbeschreibung für die untersuchten Smart-Speaker-Typen ein.

Aus linguistischer Sicht interessiert an den zuvor genannten Fragen besonders, ob und wie sich sprachliche Praktiken durch die Nutzung im VUI-Dialog verändern. Damit verortet sich die Arbeit methodologisch gesehen bei gebrauchsbasierten Ansätzen in der Linguistik und wendet Verfahren der ethnomethodologischen Konversations- bzw. Gesprächsanalyse an – mit einigen Erweiterungen und Modellierungen z. B. für Multimodalität und die Relevanz von ethnografischen Kontextbezüge. Dies wird in **Kapitel 4** ausgeführt, während **Kapitel 5** die Methodik der Arbeit detailliert beschreiben wird (Kontext und Vorgehensweise bei der Datenerhebung und -auswertung, die Kombination verschiedener Datentypen und Umfang des Korpus sowie gebildeter Kollektionen). Dabei wird auch auf den Projektkontext eingegangen, in dem die Arbeit entstanden ist, sowie auf die untersuchten Haushalte.

In **Kapitel 6** werden dann die qualitativ-explorativen Analysen der Arbeit vorgestellt. Diese fokussieren zunächst Dialoge zwischen genau einem VUI und einem menschlichen Beteiligten (**Kapitel 6.1**). So entsteht eine Grundlage von Erkenntnissen zu Basisproblemen der Gesprächsorganisation (vgl. Schegloff 2006) mit VUIs unter den Aspekten Anreden, Sequenzorganisation, Rederechtsverteilung sowie Reparaturmechanismen. Die Analysekategorien ergaben sich aus den Daten heraus und wurden dann mit Bezug auf die kanonische konversationsanalytische bzw. interaktional-linguistische Literatur weiter ausdifferenziert. Auf dieser Grundlage baut der zweite Analyseteil auf, in dem die Beteiligung von VUIs an Mehrparteieninteraktionen beleuchtet wird (**Kapitel 6.2**). Die Analysen stellen dabei auf das Aufdecken spezifischer Situationen und sprachlicher Verfahren ab, in und mit denen ein VUI (nicht) zu einem Beteiligten an der sozialen Praxis und an Gesprächen wird: Wer bzw. was an einem Gespräch ‚beteiligt‘ ist, ergibt sich

aus einer praxeologischen Perspektive nicht aus der Konstellation der menschlichen Teilnehmenden, sondern ist Gegenstand des praktischen Vollzugs. Vor dem Hintergrund des Versprechens eines ‚natürlichen Erlebens‘ seitens der Hersteller von Smart Speakern ist also danach zu fragen, wie die Geräte partizipieren können und wie dies sprachlich verfertigt wird. Mögliche Folgeuntersuchungen werden (nebst einer Zusammenfassung der Analyseergebnisse und Rückbindung an die Ausgangsfragen) in **Kapitel 7** diskutiert, das außerdem einen Ausblick auf gegenwärtige und zukünftige Herausforderungen durch VUIs enthält.

Zu Smart Speakern liegt bisher nur eine sehr beschränkte Zahl von Arbeiten aus der angewandten Linguistik vor, die sich empirisch mit den sprachlichen Praktiken dieser Geräte auseinandersetzen. Die bisher erschienenen Arbeiten stammen zu nicht unerheblichen Teilen aus demselben Kontext, in dem auch die vorliegende Arbeit entstanden ist – als Bestandteil des Projekts „Un/erbetene Beobachtung in Interaktion: Intelligente Persönliche Assistenten“ im von der Deutschen Forschungsgemeinschaft (DFG) geförderten Sonderforschungsbereich 1187 „Medien der Kooperation“ an der Universität Siegen,² an dem der Verfasser selbst beteiligt war (u. a. Hector/Hrncal 2020; Habscheid et al. 2021; Hector 2022; Hector et al. 2022; Habscheid/Hector/Hrncal 2023; Waldecker/Hector/Hoffmann 2024; Hector et al. 2023; Waldecker/Hector 2023; Hector/Hrncal 2024; Hector 2025; Habscheid et al. 2025a; Habscheid/Hector/Hrncal 2025).³ Insofern ist es ein Anliegen dieser Arbeit, die Geräte und die kookkurrenten sprachlichen Praktiken aus möglichst verschiedenen Perspektiven und mit einem in die Breite gehenden Ansatz bei der Suche nach auftretenden sprachlichen Form-Funktionszusammenhängen zu beleuchten, sodass die vertiefte Betrachtung von Folgefragen oder Einzelphänomenen dann im Anschluss daran erfolgen kann. Die bereits aus dem Projektzusammenhang heraus entstandenen Untersuchungen zum Gegenstand informieren die Arbeit, werden aber im Hinblick auf Fragestellungen und untersuchtes Material distinkt behandelt.

2 Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 262513311 – SFB 1187 Medien der Kooperation.

3 Arbeiten aus anderen Kontexten, v. a. aus dem englischsprachigen Raum, etwa von Beneteau et al. (2019), Porcheron et al. (2018) und Reeves/Porcheron (2023), sollen hier keineswegs übersehen werden; sie werden an späterer Stelle in dieser Arbeit diskutiert.

2 Theorie und Begriffe

Das folgende Kapitel dient der Begriffsbildung, die durch die Einführung verschiedener theoretischer und konzeptioneller Debatten und der detaillierten Herleitung ihrer Anwendung für den Kontext der vorliegenden Arbeit geleistet wird. Das Kapitel besteht aus drei Teilen: Am Anfang stehen Grundlagen aus den Bereichen Sozialtheorie, Sprach- und Medienwissenschaft zu den Begriffen *Praxis* und *Praktik* (Kap. 2.1). Daran schließt sich die ebenfalls sozialtheoretische, teilweise auf den vorherigen Begriffen aufbauende Debatte zum Verhältnis von Mensch und Maschine an (Kap. 2.2). Schließlich werden Begriffe und Konzepte aus der Medienaneignungs- bzw. Domestizierungsforschung vorgestellt und für den Kontext der Arbeit diskutiert (Kap. 2.3).

2.1 Praxeologische und medientheoretische Grundlagen

Die vorliegende Arbeit untersucht gebrauchsbasiert und auf einer empirischen Grundlage sprachliche *Praktiken* mit VUIs. Dieser Begriff wurde nicht zufällig gewählt: Die Analysen sollen sich mit Bezug auf den Praktiken-Begriff auf gleich zwei verschiedene Traditionen praxeologischer Theoriebildung stützen. Diese sollen im Folgenden zunächst kurz übersichtlich eingeführt werden, ehe sich eine genauere Betrachtung in den folgenden Unterkapiteln daran anschließt.

- (1) Erstens wird mit einem Verständnis von *sprachlichen Praktiken* im Sinne konversations- bzw. gesprächsanalytischer Forschung gearbeitet (siehe dazu Selting 2016). Sprachliche Praktiken sind in diesem Sinne wiederkehrend genutzte (Bündel von) „verbalen, vokalen und visuellen Ressourcen“ (Selting 2013: 589) bzw. sprachlich-interaktionalen Verfahren zur Lösung spezifischer kommunikativer Aufgaben. Sie stehen in einem konstitutiven Verhältnis zu den Handlungen, die sie hervorbringen (können). Von Interesse ist hier also mit Bezug auf den Gegenstand, welche Ressourcen in VUI-Dialogen⁴ zum Einsatz kommen, d. h., mit welchen sprachlich-interaktionalen Formen und Verfahren die Aufgabe eines Austauschs zwischen Mensch und Maschine bewältigt werden kann. Dabei orientiert sich die Analyse – mit leichten Anpassungen – an den von Schegloff (2006) formulierten allgemeinen Problemen der Gesprächsorganisation, entlang derer auch die Auswahl der tiefer analysierten Phänomenbereiche für dyadische VUI-Dialoge erfolgt.⁵ Dieser

4 Für eine Einordnung dieses Begriffs siehe Kap. 2.2.2.

5 Siehe die Einführung zum ersten Analyseteil in Kap. 6.1.

Praktikenbegriff wird einschließlich seiner Forschungstraditionen in Kap. 2.1.3 genauer vorgestellt.

- (2) Zweitens führen die Untersuchungen in sprachlich vermittelte *soziale Praktiken* hinein. Von besonderem Interesse sind dabei Domestizierungs- und Interface-Praktiken sowie das Gespräch als soziale Praktik,⁶ aber auch die sozialen Praktiken, die mit Smart Speakern ermöglicht und vollzogen werden (z. B. Üben, Testen, Vorführen, gemeinsam Radio Hören oder Kochen).⁷ Dabei beruhen die Analysen auf einem anders gelagerten Begriff von Praktiken: Soziale Praktiken werden als sinnhafte und als solche erkennbare Kombinationen körperlich-materieller, ko-operativer Verhaltensweisen verstanden, als „wechselseitig gefertigte gemeinsame Abläufe“ (Schüttpelz/Meyer 2017: 158). Am Vollzug von sozialen Praktiken können mehrere Menschen oder Artefakte sowie andere Einheiten beteiligt sein (vgl. Hirschauer 2016). Diese Begriffsbestimmung impliziert auch, dass einzelne Handlungen Teil von Praktiken sein können. Der linguistische Einsatz ist mithin anders ausgerichtet: Sprachliche Praktiken sind damit Bestandteil sozialer Praktiken, sie können diese (mit-)hervorbringen. Das Konstitutionsverhältnis ist im Vergleich zu sprachlichen Praktiken also genau umgekehrt: Handlungen bringen soziale Praktiken hervor. Durch die sprachlich-interaktionale Analyse kann insofern ein Zugang zu sozialen Praktiken gefunden werden.

Die Arbeit nimmt beide Traditionen auf und untersucht sowohl *sprachliche* Praktiken als auch sprachlich vermittelte *soziale* Praktiken mit Smart Speakern. Mit dem Begriff der *Praktik* ist dabei – in Abgrenzung zum Begriff der *Praxis* – in beiden Fällen auf ein übersituatives, raumzeitlich zwar verteiltes, aber gleichwohl wiederkehrendes und wiedererkennbares Element verwiesen, das Praktiken, sprachlichen wie sozialen, innewohnt und sie erst zur Praktik macht. Praxis hingegen ist der ‚Fluss der Ereignisse‘ – das situative, lokale „in einer wechselseitigen Verfertigung befindliche Geschehen“ (Schüttpelz/Meyer 2017: 158), in dem als solche erkenn- und verstehbare Praktiken zur ‚Aufführung‘ kommen können. Dieser Begriff von Praxis ist für die Analyse ebenso relevant, denn es soll auch die Frage beleuchtet werden, wer oder was an der „wechselseitigen Verfertigung“ mit welcher Rolle beteiligt ist – die Video-Interaktionsanalysen im zweiten Analy-

⁶ Siehe dazu Kap. 2.2.1.

⁷ Siehe die Analysen in Kap. 6.2.

seteil setzen genau auf dieser Ebene an. Dazu werden praxistheoretische Setzungen des Soziologen Stefan Hirschauer (2004; 2016) auf den Gegenstand der VUIs angewendet.

Der Aufbau der folgenden Unterkapitel gliedert sich insofern wie folgt: Zunächst wird der Begriff der *sozialen Praxis* (Kap. 2.1.1) nach Schüttpelz/Meyer, Goodwin und Hirschauer eingeführt. In diesem Zusammenhang wird die Praxeologie auch theoriegeschichtlich kurz verortet. Anschließend werden die Forschungstraditionen für die Termini *soziale Praktik* (Kap. 2.1.2) und *sprachliche Praktik* (Kap. 2.1.3) aufgeschlüsselt (um den Argumentationsgang zu vereinfachen, wird mit dem umfassenderen Begriff der sozialen Praktik begonnen, ehe der spezifischere Begriff der sozialen Praktik eingeführt wird). Abschließend werden diese praxeologischen Grundlagen auf den Bereich der Medientheorie angewendet, um ein praxeologisches Verständnis von Medien bzw. Interfaces darzulegen, die jeweils sich über die zugehörigen Praktiken analysieren lassen (Kap. 2.1.4).

2.1.1 Soziale Praxis und Ko-Operation

Das Feld der Praxeologie, regelmäßig auch als „Familie der Praxistheorien“ (Schäfer 2020: 13) bezeichnet, ist noch vergleichsweise jung. Sie ist vom Beginn ihrer Entwicklung an interdisziplinär (mit einer Fundierung in Soziologie bzw. Sozialtheorie und -philosophie), hat viele offene Enden und ist an ihren Rändern regelmäßig von Unschärfe mit fluiden Grenzen gekennzeichnet. Schäfer (2016: 9) versteht sie als „heterogene, aber dennoch definierbare Theoriebewegung“, ganz ähnlich Schatzki (2001: 13), der von einem „loose, but nevertheless definable movement of thought“ spricht. Diese Bewegung hat sich spätestens mit dem sog. „practical turn“ in den Sozial- und Kulturwissenschaften (Schatzki/Knorr-Cetina/Savigny 2001) als Paradigma etabliert und auch wenn Uneinigkeit darüber besteht, welche Theorien und welche Vertreter*innen der Praxistheorie zuzurechnen sind, gibt es doch erstens Denk- und Analyseperspektiven sowie zweitens theoretische Ausgangspunkte, die die verschiedenen Ausformungen der Praxistheorie einen und sie als Strömung klar erkennbar machen.

Zentraler gemeinsamer Ausgangspunkt und damit verbundenes Forschungsanliegen praxeologischer Theorie- und Begriffsarchitekturen ist der Versuch, den vermeintlichen Gegensatz von Gesellschaft und Individuum, den Dualismus von ‚Ordnen‘ und ‚Ordnung‘ und Dichotomien von Mikro- und Makroebene oder Subjekt und Objekt zu überwinden (vgl. Schäfer 2020: 18). Praxistheorie versteht sich als Antwort auf das Mikro-Makro-Problem des Strukturalismus in der Soziologie. Schatzki (2016) bringt dies mit dem Begriff der „flachen Ontologie“ auf den Punkt: In praxeologischen Anschauungsweisen gibt es nicht unterschiedliche Realitäts-

ebenen für die Mikroebene (wie Handlungen, Interaktionen und einzelne Individuen) auf der einen Seite und die Makroebene (wie Institutionen, Strukturen, Systeme u. a.) auf der anderen Seite. Es bestehen auch keine bestimmenden oder begründenden Effekte von der vermeintlich höheren Makro- auf die niedrigere Mikro-Ebene. Stattdessen existiert bloß eine einzige ontologische Ebene, in der sich die beiden Bereiche in einem „Konstitutionsverhältnis“ befinden, d. h. auseinander bestehen (vgl. Schatzki 2016: 31): Strukturen oder Systeme konstituieren sich nur *durch* Individuen, deren Handlungen und Interaktionen, und diese bilden immer auch die Strukturen oder Systeme. Praktiken sind insofern der konstitutive Ort des Sozialen, alle sozialen Phänomene sind „Konstellationen oder Aspekte von Praktiken oder in Praktiken begründet“ (Schatzki 2016: 30) und öffnen den Blick für die „beständige[n] Hervorbringungen“ (Schäfer 2020: 20) von Gesellschaft und Individuum mit einer *Prozessperspektive*.

Den praxistheoretischen Rahmen der vorliegenden Arbeit bildet eine Theorie- und Begriffsarchitektur, die auf den Setzungen von Charles Goodwin (insb. Goodwin 2018) sowie Schüttpelz/Meyer (2017) aufbaut und bereits kurz angerissen wurde: ein Verständnis von Praxis als „das in einer wechselseitigen Verfertigung befindliche Geschehen“ (Schüttpelz/Meyer 2017: 158); dies hat eine konsequente *Vollzugsperspektive* zur Folge. In etwa äquivalent dazu verhält sich der von Goodwin (2018) geprägte Begriff der „Co-Operative Action“ – beide Ansätze betonen einen fortlaufenden, hochgradig situativen und lokal entstehenden ‚Strom‘ sozialer Ereignisse. Der Begriff der Kooperation stellt dabei nicht die konsensuale Zusammenarbeit von zwei oder mehr Individuen dar, die – spieltheoretisch bzw. aus Sicht der biologischen Anthropologie gedacht – für einen bestimmten Preis ihren gemeinsamen und so auch ihren individuellen Nutzen maximieren, sondern vielmehr wird durch Ko-Operation – jetzt mit Bindestrich – wechselseitig ein Ablauf hervorgebracht (vgl. Goodwin 2018: 5–7). Diese Wechselseitigkeit entspricht im Englischen dem Konzept der „mutuality“, d. h. eine unausgesprochene, vorausgesetzte Wechselseitigkeit, auf die wir permanent angewiesen sind, um uns überhaupt mit anderen Menschen koordinieren und den ‚Strom‘ der Abläufe erzeugen zu können: „Without our permanent mutual assistance, we would not be able to speak and act at all“ (Schüttpelz/Meyer 2018: 175). Die Ziele der Beteiligten sind also nicht notwendigerweise die gleichen. Es handelt sich nicht um eine Wechselseitigkeit im Sinne einer „common“ oder – noch stärker – „joint“ action, sondern gerade das Verweigern einer „joint action“ kann Gegenstand einer Ko-Operation sein, die ebenfalls eine „mutuality“ voraussetzt. Handeln mit gemeinsamen Zielen (Interagieren) ist also eine mögliche Ausprägung, aber weder eine notwendige noch hinreichende Bedingung für Ko-Operation. Betrachten wir die Einbindung von VUIs in die soziale Praxis, ist dies eine wichtige Setzung, weil dem technischen Artefakt keine *intentionale* Wechselseitigkeit oder Zielorientie-

rung unterstellt werden muss, um Bestandteil des wechselseitig gefertigten Ablaufs zu sein.⁸ Schüttpelz/Meyer (2017) rücken von den Arbeiten Goodwins ausgehend den Praxisbegriff ins Zentrum ihrer Überlegungen und bauen darauf eine in sich konsistente Begriffsbildung auf: Praxis ist „das in einer wechselseitigen Verfertigung befindliche Geschehen“ (Schüttpelz/Meyer 2017: 158). Darauf – also auf diesen fortlaufenden ‚Fluss‘ oder ‚Strom‘ sozialer Ereignisse – wird im Folgenden als (*soziale*) *Praxis* referiert werden. Schüttpelz/Meyer (2017: 158–159) leiten darauf aufbauend die Begriffe *Kooperation* – „die wechselseitige Verfertigung gemeinsamer Ziele, Mittel oder Abläufe“ (Schüttpelz/Meyer 2017: 158) – *Interaktion*, *Praktiken*, *Handlungen*, *Routinen* und *Techniken* ab. Praxis und Kooperation werden also konzeptuell allen anderen Größen vorgelagert.

Der Begriff der *Handlung* (jedenfalls in einem klassischen Verständnis) ist bei Schüttpelz und Meyer und anderen praxeologischen Ansätzen hingegen weniger zentral. Das individuelle „Tun“ einer Handlungstheorie etwa nach Max Weber (1966[1921]) steht zur Erklärung der sozialen Ordnung in dieser Auffassung weniger im Fokus. Praxis wird in der Terminologie von Schüttpelz/Meyer (2017) und anderen praxeologischen Ansätzen überhaupt erst zu Handeln, wenn es *mit einem Anderen* und in einem bestimmten materialen Kontext vollzogen wird – und ist auch dann nur der „*Ausschnitt* [Herv. i. O.] einer Praktik, der von nur einem Körper getragen wird“ (Hirschauer 2016: 51). Goodwin (2018) entfaltet seine Idee zwar terminologisch ausgehend von „actions“, doch er nimmt außerdem an, dass diese nur dadurch entstehen, dass verschiedene allgemeingültige oder neu kombinierte, semiotisch in der spezifischen Situation immer wieder neu funktionalisierte Ressourcen miteinander in Verbindung gebracht werden und so ein neues „Ganzes“ schaffen (Goodwin 2018: 221). Erst durch diese Verbindung entsteht eine Handlung. Der *Prozess* der Verbindung ist entsprechend „co-operative action“ – also die Praxis (vgl. Habscheid et al. 2021: 27).

Für eine praxeologische Perspektive genügt es also Hirschauer (2004, 2016) zufolge nicht, das auf die einzelnen Körper zugeschnittenen bzw. von ihnen ausgehende Handeln anzusehen. Man muss die Praxis, die *co-operative action*, vielmehr als Prozess mit verschiedenen Beteiligten in den Blick nehmen. Für Hirschauer bedeutet das, dass in diesen sozialen Prozess mehrere *Partizipanden*⁹ involviert sind,

⁸ Siehe dazu auch Kap. 2.2.

⁹ Hirschauer (2004: 74–75) begründet hier explizit die Verwendung des Partizip Präsens: Dies „gehört zu den sog. Infiniten Verbformen, die anders als finite Formen in Person und Numerus unbestimmt sind. Infiniten Verben ist der Akteur egal: Etwas bleibt ‚rauchend‘ – ob als Mensch, als Colt, als Schlot oder als Vulkan“. Damit wird – wie im Folgenden näher erläutert – eine partizipatorische anstatt einer kontributorischen Perspektive fokussiert. Entsprechend wird diese

die ihren ‚Beitrag‘, ihren wahrnehmbaren Anteil, zum Vollzug der Praktik leisten, die Handlungen also auf mehrere Urheber „verteilt“ sind (Hirschauer 2016: 51). In Abgrenzung zu handlungstheoretischen Ansätzen, aber auch Gegenentwürfen wie der von Bruno Latour eingebrachten *non-human agency*¹⁰ schlägt Hirschauer (2004: 74) vor, eine „*partizipatorische* [Herv. i. O.] Perspektive auf jene Entitäten ein[zu] nehmen, die an Praxis teilhaben und in ihre Dynamik verwickelt sind“. Diese materiellen Entitäten können jedoch (a) einen sehr unterschiedlichen Aktivitätsstatus aufweisen, der zwischen proaktiven und inhibtiven sowie aktivem und passivem Handeln zu suchen ist (vgl. Hirschauer 2016: 49). Mit anderen Worten: Partizipanden am Vollzug der Praxis können eher nützlich oder hinderlich sowie explizit oder implizit daran beteiligt sein. Die für den Gegenstandsbereich der VUIs sehr interessante Pointe ist, dass außerdem (b) die Partizipanden nicht menschlich sein müssen und auch nicht in jedem Fall sind. Involvierte können auch nicht-menschliche Wesen (etwa Tiere), deren Körper, Dinge (z. B. eine Zahnbürste, eine Pfanne oder ein Smart Speaker), Dokumente oder deren Gestaltung, gegebene Umstände (z. B. das Wetter) oder räumliche „Settings“ sein. Damit ist klar, dass in dieser Hinsicht auch ein Smart Speaker mit einer synthetischen Stimme, die verbalsprachliche Laute erzeugt, ‚involviert‘ sein und z. B. bestimmte Äußerungen der menschlichen Partizipanden evozieren und andere supprimieren kann. Diese *partizipatorische* Betrachtungsweise soll auch bei der empirischen Untersuchung von Mensch-Maschine-Dialogen herangezogen werden: Sie ermöglicht es, maschinelle Artefakte differenziert als Partizipanden an der Praxis zu betrachten, ohne ihnen Subjektstatus, Intentionalität oder Handlungsmacht zuschreiben zu müssen.¹¹

2.1.2 Soziale Praktiken: Kleinste Einheit des Sozialen

Soziale Praktiken sind das, was der Soziologe Andreas Reckwitz (2003: 290) treffend als die „kleinste Einheit‘ des Sozialen“ beschrieben hat. Reckwitz stützt sich dabei auf den Sozialtheoretiker Theodore Schatzki, dessen Arbeiten auch im Folgenden leitend sein sollen und v. a. deshalb geeignet erscheinen, weil sie Rekurs auf den Sprachphilosophen Ludwig Wittgenstein nehmen und die hohe Relevanz

Schreibweise (gegenüber einer Schreibweise mit *-ant*, die eine *aktive* Rolle der Beteiligten schon voraussetzen würde) übernommen.

¹⁰ Zur Akteur-Netzwerk-Theorie (ANT), auf die an dieser Stelle nicht weiter eingegangen werden soll, siehe Latour (1996; 2007).

¹¹ Siehe dazu ausführlicher Kap. 2.2.4.

des Sprachlichen innerhalb des Gefüges sozialer Praktiken erkennen.¹² Dies erscheint vor dem Hintergrund gegenstandsadäquat, dass die Funktionsweise von VUIs dialogisch und auf gesprochene Sprache orientiert ist, sodass soziale Praktiken mit VUIs jedenfalls zu großen Teilen sprachlich vollzogen werden. Schatzki (1996: 89) spricht von sozialen Praktiken als „a temporally unfolding and spatially dispersed nexus of doings and sayings“. Damit sind soziale Einheiten beschrieben, die eine Verkettung von verteilten, verbal oder körperlich ausgeführten Handlungsschritten erfordern und durch diese konstituiert werden, z. B. Kochen, Wählen, Pflegen oder Korrigieren (vgl. Schatzki 1996: 89). An diesen können sprachliche Praktiken im Sinne von genutzten Ressourcenbündeln bzw. sprachlich-interaktionalen Verfahren einen (auch den wesentlichen) Anteil haben, insofern sie Handlungsschritte hervorbringen. Schatzki (1996: 89) beschreibt soziale Praktiken¹³ insofern als als Zusammenhang von Handlungen und Äußerungen („nexus of doings and sayings“): Diese sind miteinander erstens dadurch verbunden, dass die ausführenden Personen sie als eine Einheit verstehen (nicht notwendigerweise, dass sie von außen auch als solche beobachtbar sind – dies trifft nur auf einen Teil der Praktiken zu, s. u.). Zweitens sind sie durch eine normative Ebene in Form von Regeln, Vorgaben und Anweisungen miteinander verbunden und drittens durch eine kulturell geprägte Zielrichtung, die Schatzki (1996: 89) „teleoaffektive structure“ nennt und die die Ziele und Zwecke bestimmter Praktiken mit in den Aufbau dieser einfließen lässt. Die einzelnen Bestandteile, d. h. die Handlungen und Äußerungen, können dabei auch kausal miteinander verbunden sein. Diese Auffassung ist primär gemeint, wenn im Folgenden von *sozialen* Praktiken gesprochen wird.

Praktiken sind also einerseits konstitutiver Kontext, in dem ein bestimmtes Verhalten – in Form von körperlichem Verhalten, sprachlichen Äußerungen und durch Übung – zur Handlung werden kann (vgl. Schatzki 1996: 36). Die besondere Rolle des Sprachlichen kommt im von Ludwig Wittgenstein geprägten Begriff der „Sprachspiele“ zum Ausdruck (vgl. Wittgenstein 1990[1953]: 250; § 23). Wittgen-

12 Zur Relevanz von Schatzki für die Praxeologie siehe auch Schulz-Schaeffer (2010: 335).

13 Schatzki (1996: 90–98) unterscheidet auch noch zwei weitere Bestimmungen des englischen Begriffs „practice“: erstens verstanden als „Übung“ oder das „Einstudieren“, z. B. beim Klavierspielen „üben“ oder beim sportlichen Training. Dieser Auffassung geht Schatzki nicht weiter nach, auch wenn er sie selbst nicht für irrelevant hält und sie auch bei Goodwin in Form der Ausbildung durch Fähigkeiten und Kompetenzen eine gewichtige Rolle spielt (vgl. Goodwin 2018: 456). Der Idee des ‚Einstudierens‘ und ‚Erprobens‘ wird in der folgenden Arbeit zwar stellenweise ebenfalls Aufmerksamkeit geschenkt, allerdings bleibt dabei der Praktiken-Begriff außen vor. Zweitens wird „practice“ auch bei Schatzki (1996: 90) als „stream of activity“ bezeichnet und somit als Vollzugsbegriff eingeführt, was im Deutschen in etwa dem Terminus *Praxis* entspricht, siehe Kap. 2.1.1.

stein betont, dass Sprache in ihren Gebrauchskontext eingewoben ist und diesen (und reflexiv auch dessen Regeln) mit hervorbringt. Letzterer kann dabei sowohl lokal der Kontext einer ganz bestimmten Situation oder auch übergreifend ein bestimmter Lebensbereich (z. B. das Militär) oder anderer Zusammenhang sein – Praktiken können also auf ganz unterschiedlichen Analyseebenen beleuchtet werden. Andererseits entstehen also Praktiken auch erst aus den immer wieder so vollzogenen und für die Praktik relevanten Handlungsabläufen. Diese Handlungsabläufe unterliegen den für die Praktik spezifischen Regeln, doch da – mit Verweis auf den „unendlichen Regelregress“¹⁴ wie Wittgenstein (1990[1953]: 345–346; PU §§ 201–206) ihn herleitet – diese nicht durch weitere Regeln normiert werden können, ist es eben gerade deren Anwendung und ihr spezifischer Vollzug in einer gegebenen sozialen Situation, die sie konstituieren: „Regeln und Praktiken setzen also einander wechselseitig voraus“ (Habscheid 2016: 134).¹⁵ Das „Regelfolgen“ Wittgensteins macht deutlich, dass sich die konkreten Regelinstantiierungen, d. h. die *vollzogenen* Praktiken, durch die Deutungen der Regeln durch die Beteiligten, den situativen Kontext und andere Bedingungen (Wittgenstein 1990[1953]: 254; PU § 30), d. h. *in der Praxis* ergeben. Dabei sind Praktiken auch hier dadurch gekennzeichnet, dass der Sinn einer Praktik nicht identisch mit dem subjektiv gemeinten Sinn oder der Intention des Individuums sein muss und der Sinn so erst *in situ* entsteht (vgl. Lanwer/Coussios 2015: 128–129). Sprache stellt also mit ihren Regeln bzw. Sprachspielen einen Zugang zur Anatomie sozialer Praktiken dar – und insofern auch zu kulturellen und medialen Aspekten dessen (vgl. Meiler 2023), was sich im Hinblick auf die zu untersuchenden VUI-Dialoge aufgrund der bereits erwähnten Sprachlichkeit dieser besonders anbietet.

2.1.3 Sprachliche Praktiken: Ressourcen für Bausteine der Interaktion

Der hier ebenfalls verwendete Begriff der *sprachlichen Praktiken* liegt insofern, wie bereits eingangs dargestellt analytisch, auf einer anderen ‚Ebene‘. Sprachliche

14 Der infinite Regelregress (auch „Regelfolgen“), der auf Wittgenstein zurückgeht, besagt in Kürze zusammengefasst, dass Regeln zur Anwendung der Regel wiederum einer Regel bedürfen, wann diese Regel zu verwenden ist, und zu einer Regel zweiter Ordnung eine Regel dritter Ordnung hinzutritt, bis Regeln n-ter Ordnung aufgestellt sind, aber noch immer nicht die Anwendung der Regeln geklärt werden konnte (vgl. Hillebrandt 2014: 37; Habscheid 2016: 134); erst das tatsächliche „Befolgen“ der Regel kann dies auflösen.

15 Angewendet auf ein sprachliches Regelsystem bedeutet dies im Sinne Wittgensteins, dass „die Regeln der Sprache [...] keine natürlich gegebenen Ausgangspunkte der Praxis [sind], sie sind vielmehr deren Produkt“ (Hillebrandt 2014: 40).

Praktiken sind bei Goodwin (2018: 25) ähnlich wie etwa bei Schegloff (1997a; 2000a) als „specific grammatical practices“ bestimmt. Einbezogen werden auch Praktiken nicht-sprachlicher Art (Goodwin 2018: 46), denn es muss betont werden, dass damit keineswegs nur sprachliche Ressourcen i.e.S. erfasst werden sollen, sondern vielmehr alle in der Situation zur Verfügung stehenden sprachlichen, materialen und sozialen Mittel als Ressource zur Übermittlung von Zeichen eingesetzt werden können – man könnte hier präziser auch von „kommunikativen Praktiken“ sprechen (vgl. Habscheid 2016: 137), von denen sprachliche Praktiken ein Bestandteil sind.¹⁶ Sie ergeben sich aus der Praxis und sind deren Produkt. Goodwin versteht dabei unter Praktiken nicht einfach nur die unfunktionalisierten Ressourcen, sondern erst durch deren wiederholte, konglomerierte Nutzbarmachung werden Praktiken konstituiert. Goodwin geht dabei von einer temporalen Akkumulation einzelner Operationen aus. Genau diese Auffassung von Praktiken herrscht auch in der konversationsanalytischen Betrachtungsweise vor: Selting (2016: 28–29) spricht von „routinemäßiger Verwendung“ „verbaler, vokaler und, so weit relevant, visueller Ressourcen“. Genau dies entspricht dem Praktikenverständnis konversationsanalytischer Arbeiten (siehe etwa Schegloff 1997a) – Praktiken sind die Ressourcen-Bausteine, aus denen sich soziale Interaktion speist und von Interagierenden zusammensetzen lässt, um kommunikative Aufgaben zu bewältigen (siehe auch Schegloff 2006). Mit welchen Bausteinen die Aufgabe eines VUI-Dialogs bewältigt werden kann ist also die zentrale Frage, für die dieses Verständnis von sprachlichen Praktiken nutzbar gemacht wird.

Praktiken auf dieser Ebene dienen also dem Aufbau „soziale[r] Handlungen und Aktivitäten“ (Selting 2016: 28) in der Interaktion¹⁷ und es lässt sich noch radikaler formulieren, dass in der Interaktion alle sprachlichen und anderweitig verkörperten Ausdrücke, auch im Rückgriff auf körperliche oder materiale Ressourcen, dazu dienen, die Einheit *für das Gegenüber* verständlich zu machen – also

¹⁶ Aufgrund der sprachzentrierten Analyse bleibe ich gleichwohl bei dem Terminus „sprachliche Praktiken“, beziehe dabei aber auch multimodale Verfahren in die Analyse ein (siehe Kap. 4.4).

¹⁷ Wie hier deutlich wird, ist in dieser Auffassung, wie sie v. a. bei Selting, Schegloff und in anderen konversationsanalytischen Arbeiten (teils implizit) entwickelt wird, nicht die ‚Praxis‘ oder die ‚Ko-Operation‘ die entscheidende Bezugsgröße, sondern die Interaktion und die Verfahren, mit denen sie organisiert werden kann (vgl. Schegloff 2006: 70). Über diese Abweichung in der Fokussierung der einzelnen Domänen kann im Folgenden aber hinweggesehen werden, weil das Konzept von Interaktion, auf das Schegloff sich bezieht, dem Konzept der Ko-Operation bzw. der Praxis nicht unähnlich ist – auch er geht nicht von Subjekten aus, sondern betont die Wechselseitigkeit und den „common ground“, von dem aus eine Interaktion überhaupt erst möglich ist (vgl. Schegloff 2006: 89).

das Konzept des Recipient-Design (vgl. Sacks/Schegloff/Jefferson 1974: 727) Anwendung findet. Diese Ressourcen werden jedoch nur dann eine Praktik, wenn sie für einen routinisierten und wiederholt auftretenden Zweck eingesetzt werden, so z. B. der Einsatz bestimmter Partikeln zur Initiierung von Reparaturen (Schegloff 1997a). Selting (2016: 29) formuliert dieses Verhältnis zwischen Ressourcen, Praktik und Handlung so: „Durch die Verwendung einer für eine bestimmte Handlung designierten Praktik wird diese bestimmte Handlung ausgeführt, deren Erkennen ermöglicht oder deren Interpretation nahegelegt“. Dafür kommen bestimmte „Ressourcen oder Ressourcenbündel, d. h. sprachliche und nicht-sprachliche Formate“ zum Einsatz. Werden diese Ressourcen nun aber nicht für den routinisierten Zweck eingesetzt, sondern anderweitig funktionalisiert, sind sie eben nicht mehr eine „für eine bestimmte Handlung designierte[.] Praktik“. ¹⁸ Dabei ist die Bedeutung einer Praktik dennoch nicht *a priori* gesetzt: Praktiken sind „kontextsensitiv“ und können „flexibel an die je spezifischen Gegebenheiten angepasst werden“ (Selting 2016: 29). Diese Flexibilität ist im Hinblick auf VUI-Dialoge zu prüfen, die mit vorgegebenen Dialogstrukturen arbeiten und insofern bestimmte Eingaben an bestimmten konversationellen Stellen erwarten – die Flexibilität könnte an dieser Stelle v. a. seitens der Nutzer*innen bestehen, sich an den präskribierten Verlauf anzupassen (vgl. Habscheid 2022: 191).

Selting (2013) unterscheidet zwischen „verbalen“, „vokalen“ und „visuellen“ Ressourcen, die beim Vollzug von Praktiken zum Einsatz kommen. Mit Blick auf die „interaktionale Infrastruktur“ (Schegloff 2006) ist darüber hinaus nach den interaktionalen Merkmalen zu fragen: Konversationelle Strategien für die sequenzielle Organisation, die Rederechtsverteilung oder für Reparaturen sowie der Aufbau, das Zustandekommen und das Beenden der Interaktion sollen in den Blick genommen werden (siehe auch Schegloff/Sacks 1973). Diese sprachlich-interaktionalen Merkmale sind nicht zufällig genau die Ressourcen bzw. Eigenschaften der gesprochenen Sprache, an denen auch die Forschung zu Spracherkennung und VUIs ansetzt, um ihre angebotenen Technologien zu verbessern (vgl. Kap. 3). Es ist die Kombination sehr kleiner Einheiten von weniger als einer Sekunde, die mit Blick auf Tonhöhe, Sprechtempo, Sprechpausen und Zögerungen das synthetische Erzeugen

¹⁸ Wie Selting (2016: 50) ausführt, ist dies analog zum Verhältnis von „Kontextualisierungshinweis“ und „Kontextualisierungsverfahren“ in der Kontextualisierungstheorie nach Gumperz (1982) zu betrachten: Während die *Hinweise* („Contextualization Clues“ bei Gumperz) lediglich die Ressourcen sind, sind es die *Verfahren*, die einen bestimmten Kontext konstituieren, indem sie Bezüge zum Hintergrundwissen der Gesprächsbeteiligten herstellen (vgl. Auer 1986: 24). Auffällig ist, dass sich zwar Gumperz (1982) noch überwiegend auf Sprachliches bezog, während in der deutschsprachigen Weiterentwicklung von Auer (1986) bereits Ressourcen wie Blick, Gestik, Pausen u. a. einbezogen werden; siehe auch Kap. 4.4.

einer dem Menschlichen ähnlichen Mündlichkeit erst ermöglicht. Will man also, mit einer Engführung des Erkenntnisinteresses auf das Sprachliche, herausfinden, welche sprachlichen Praktiken im Kontext des Gebrauchs von VUIs zum Einsatz kommen, ist eine sequenzielle Analyse¹⁹ der einzelnen Handlungsschritte im Sinne der Konversationsanalyse und der Interaktionalen Linguistik durchaus sehr hilfreich. Selting/Couper-Kuhlen (2001a: 280) nennen die „Sprachanalyse und -synthese in elektronischen Dialogsystemen“ explizit als möglichen Ertrag der Interaktionalen Linguistik.

2.1.4 Medien- und Interface-Praktiken

Ein Ansatz zur Untersuchung von Medien, der von den Praktiken ausgeht, mit denen sie verfertigt werden – mithin *Medienpraktiken* – hat in den letzten Jahren eine erhöhte Aufmerksamkeit erfahren (siehe etwa Schüttpelz/Gießmann 2015; Dang-Anh et al. 2017; Schüttpelz/Meyer 2017; Meiler 2019; Schüttpelz et al. 2021). Diese Auffassung ist nicht nur eine Folge davon, dass der sogenannte „practice turn“ auch die Medienwissenschaften erreicht hat (vgl. Schüttpelz 2021), sondern auch eine Reaktion darauf, dass Medien sich spätestens seit der bereits von Kittler (1996: 131) beobachteten Konvergenz von Medien in einem „Supermedium Computer“ sich nicht ohne ihre sozialen Voraussetzungen beschreiben lassen. Entsprechend müssten, so die Argumentation, die Praktiken, mit denen Medien verfertigt werden, am Ausgangspunkt der Analyse stehen (vgl. Schüttpelz/Gießmann 2015: 8–9). Medien bringen in diesem Sinne also Medienpraktiken hervor und sind andersherum, in Anlehnung an das ethnomethodologische Programm Harold Garfinkels und die dort entworfene Idee von Reflexivität (vgl. Garfinkel 1967: 7–9; siehe auch Bergmann/Meyer 2021b: 39–42) zugleich das Ergebnis von Medienpraktiken (vgl. Schüttpelz/Gießmann 2015: 35–36): Medien konstituieren

¹⁹ Dabei soll die von Goodwin herausgestellte und insbesondere von Schüttpelz/Meyer (2018: 174, 183) pointierte ‚Abkehr‘ von der (vermeintlichen) „Reduktion“ der sozialen Ereignisse bzw. der Interaktion auf „chained (strategic) moves of individuals“ bzw. „aggregations of individual acts that are sequentially performed“ nicht übersehen werden: Praktiken zeigen sich in dieser Auffassung „simultaneously in the midst of single actions, rather than sequentially“ (Goodwin 2018: 13). Eine zu starke Engführung auf die sequenzielle Organisation verstellt Goodwin (2018: 41) zufolge den Blick für die relevanten Bestandteile des Vollzugs der Praktiken. Doch trotz der ganzheitlichen Betrachtungsweise sind es für die Erkenntnisinteressen an der Schnittstelle von Medienlinguistik und Gesprächsforschung gerade die sequenziellen Vollzüge, in denen die sprachlichen Praktiken zum Einsatz kommen, analytisch beobachtbar werden und auch einen analytischen Zugriff auf soziale Praktiken ermöglichen. Weder Goodwin noch Schüttpelz/Meyer negieren diese Inkrementalität von sprachlicher Entfaltung.

sich in einem reflexiven Wechselspiel immer wieder neu als Kooperationsbedingungen, hervorgebracht als das, was kooperativ durch Medienpraktiken als Kooperationsbedingungen verfertigt wird. Diese Kooperationsbedingungen müssen freilich nicht notwendigerweise digitale sein, auch wenn durch immer weiter in alle Lebensbereiche eingewobene Digitalität und Vernetzung durchaus spezifische Bedingungen entstehen (vgl. Kaerlein 2020: 46). Besonders in Bezug auf digitale Medien wird also ein Spannungsverhältnis sichtbar zwischen sich konventionalisierenden Medienpraktiken und ihrer fortlaufenden Aktualisierung durch den Vollzug.

Smart Speaker können in verschiedener Hinsicht als Kooperationsbedingungen verstanden werden. Für die Anwender*innen sind sie gleich in mehrfacher Hinsicht Bedingung für die Kooperation: einerseits für die Kooperation mit den Systemanbietern, die Sprachsteuerung durch das Übersetzen von Schallwellen in diskrete Zeichen (und zurück) anbieten und dabei zum Teil opake Datenverarbeitungs- und -verwertungsprozesse vollziehen,²⁰ andererseits mit Drittanbietern, die u. a. gestreamte Medieninhalte (Musik, Podcasts, Nachrichten oder auch interaktive Formate wie Spiele) bereitstellen. Nutzer*innen kooperieren also (mehr oder weniger wissentlich) mit den Anbietern der Systeme sowie mit Plattformen zur Bereitstellung von Inhalten.²¹ Untersuchungen zeigen, dass die Anwender*innen die hintergründig ablaufenden Prozesse dieser Kooperation zu großen Teilen nicht kennen können oder wollen – diese Kooperationsform ist damit ein geradezu typischer Fall von „Kooperation ohne Konsens“ nach Susan Leigh Star (1993) (vgl. Waldecker/Hector/Hoffmann 2024). Stationäre Sprachassistentensysteme sind zudem – für die Nutzer*innen nur vermittelt einseh- und konfigurierbar – Kooperationsbedingung für die Kooperation zwischen verschiedenen Plattformen, u. a. zwischen den Diensteanbietern der Sprachsteuerung und denen zur Bereitstellung der Inhalte, ebenso wie Werbetreibenden.²² Sie sind somit Bestandteil ganz unterschiedlicher Medienpraktiken, durch die sie zugleich immer wieder neu hervorgebracht werden.

Medienpraktiken konstituieren und vollziehen sich allerdings nicht im luftleeren Raum. Sie werden durch *Interfaces* konturiert – ein Begriff, der hier als wegweisend für die theoretische Fundierung der Analysen eingeführt werden soll und zunehmend das Interesse der medienwissenschaftlichen Forschung weckt; als Belege dafür sei auf die Verweise bei Distelmeyer (2020: 63) sowie auf

²⁰ Siehe dazu Kap. 3.2.2.

²¹ Siehe Kap. 3.2.3.

²² Für eine genauere Betrachtung siehe Kap. 3.2.2 und 3.2.3 sowie Hector (2025).

das derzeit entstehende Handbuch zur Einführung in die Interface Studies (siehe Ernst et al. i. V.) verwiesen. Interfaces waren in medienwissenschaftlichen Untersuchungen die längste Zeit ein eher diffuser und unscharfer Begriff (vgl. Wirth 2016: 17), doch durch theoretische Arbeit am Interface-Begriff²³ scheint sich eine Perspektive zu etablieren, die Interfaces nicht als neutrale Objekte sieht. Die Interface-Forschung hat dazu verschiedene Thesen formuliert, die – unterschiedlich nuanciert – auf einen „prozessualen“ (Kaerlein 2020: 50), man kann, je nach Fokus, auch sagen: relationalen oder praxeologischen Begriff von Interfaces abstellen.²⁴ Interfaces entstehen demzufolge erst dort, wo zwei unterschiedliche Einheiten aufeinandertreffen. Sie sind mit Kaerlein (2020: 50) „dasjenige, was erst in der Begegnung heterogener Entitäten Kontur gewinnt, aber die Beziehung zwischen den beteiligten Größen konstitutiv reguliert“. Diese pointierte Zusammenfassung der gegenwärtig vorherrschenden Position zu Interfaces baut auf Theoriebildungen der Medienwissenschaften bzw. der Software Studies auf, so postuliert etwa Galloway (2012: 33) „[A]n interface is not a thing, an interface is always an effect“, Hookway (2014: ix) beschreibt Interfaces als „a relation with technology rather than as a technology in itself“. Distelmeyer (2020: 63) weist darauf hin, dass Interfaces auch in der Begriffsgeschichte mit einer Fokussierung auf „Prozesse und Effekte konkreter Konstellationen“ verstanden wurden; ein solches Verständnis ist auch im Hinblick auf den Wortursprung konsequent, der im 19. Jahrhundert bei den chemischen Untersuchungen James Thomsons zu finden ist. Bei Thomson entsteht – vereinfacht gesagt – ein Interface beim *Kontakt* zweier Flüssigkeiten mit einer geteilten Grenzfläche (vgl. Hookway 2014: 59–60): Das Interface ist also nicht einfach schon gegeben, sondern kommt erst als Grenzfläche zweier ungleicher Ein-

²³ Für eine Übersicht über die Entwicklungslinien des Interface-Begriffs siehe Wirth (2016: 18–21) und ausführlich Wirth (2019: 68–81; 2025); der Begriff der *Interaktion* bzw. der *Interaktivität* hat sich demzufolge zwar ebenfalls etablieren können – bis hin zur Teildisziplin der Human-Computer-Interaction –, bleibt jedoch im Hinblick auf sein heuristisches Potenzial unbestimmt (siehe auch Manovich 2001) und bringt zudem weitere begriffliche Schwierigkeiten mit sich; er soll dementsprechend nicht für den Austausch zwischen Mensch und Computer verwendet werden (siehe auch Kap. 2.2).

²⁴ Cramer/Fuller (2008: 149) unterscheiden in einer Interface-Typologisierung fünf verschiedene Arten von Interfaces: User-Hardware, Hardware-Hardware, Hardware-Software, Software-Software und User-Software. Auch wenn im Hinblick auf den zu untersuchenden Gegenstand der stationären Sprachassistenzsysteme eine eingehende Betrachtung aller diese Typen lohnen würde, konzentrieren sich die Ausführungen im Folgenden ausschließlich auf User-Software-Relationen (trotz der Kritik von Cramer/Fuller (2008: 149), dass es gerade diese Beziehung ist, die in medienwissenschaftlichen Untersuchungen fälschlicherweise häufig als die einzige Form von Interfaces betrachtet wird).

heiten zustande, wenn diese in den Austausch treten. Konsequenterweise gehen neuere Ansätze (etwa Lipp/Dickel 2022) von einem „Interfacing“ aus, d. h. von den „manifold practices and activities that are needed to establish, maintain, and repair the human/machine interface“ (Lipp/Dickel 2022: 429) – und beziehen somit neben den Nutzer*innen auch Perspektiven von Ingenieur*innen oder Entwickler*innen ein. Interfaces sind entsprechend „the contingent, temporarily stabilized *products* [Herv. i. O.] of ongoing practices of interfacing“ (Lipp/Dickel 2022: 432). Diese Praktiken zeichnen sich, so Lipp/Dickel (2022: 439), dadurch aus, dass sie *gleichzeitig* menschliche und nicht-menschliche Einheiten sowohl verbinden als auch trennen.²⁵ Sie bestimmen „contact points and corridors that allow to ‚bridge‘ those differences“ (Lipp/Dickel 2022: 429).

Damit ist auf die regulierenden, steuernden und leitenden Eigenschaften von Interfaces verwiesen, die bei aller Situationsbezogenheit der Analyse nicht aus dem Blick geraten dürfen. Interfaces sind entscheidend für die Beziehungen zwischen Menschen und Computern (vgl. Ernst/Bächle 2020: 418): durch Interfaces „stehen nur bestimmte Handlungs- und Kommunikationsweisen zur Auswahl, die dadurch die Interaktion von menschlichen Akteuren mit Computersystemen nachhaltig prägen“ (Ernst/Bächle 2020: 418). Dadurch werden bestimmte Praktiken inhiert, zu anderen wird geradezu eingeladen, was auch konkrete Auswirkungen auf den Sprachgebrauch hat. Interfaces tragen also aktiv zur Verortung bestimmter Medienpraktiken auf dem Proaktiv-Inhibitiv-Spektrum nach Hirschauer (2016: 49) bei: „interfaces both engender and condition human/machine relations“ (Lipp/Dickel 2022: 429; vgl. auch Kaerlein 2020: 51). Damit wird auch noch einmal deutlich, dass Interfaces keine Nullstelle sind, sondern sich gerade dadurch auszeichnen, dass sie (Macht)asymmetrien zwischen zwei ungleichen Einheiten miteinander vermitteln (vgl. Cramer/Fuller 2008: 150). Computer-Interfaces wirken einerseits ermächtigend auf Nutzer*innen, weil diese in die Lage versetzt werden, etwas zu bestimmen – zugleich wirken sie andererseits ebenfalls auf die Nutzer*innen sowie die ausgeführten Praktiken regulierend (vgl. Distelmeyer 2018: 29). Von Interfaces geht also – wie Distelmeyer (2020) detailliert am Begriff des „Leitens“ illustriert – ein Machtpotenzial aus; konkret wirkt dies auf die Nutzer*innen z. B. in vorstrukturierten Oberflächen, die bestimmte Praktiken wahrscheinlicher machen als andere – z. B., im Zusammenhang mit grafischen Interfaces, die Zustimmung zu den Datenschutz-Bestimmungen wahrscheinlicher als deren Ablehnung oder das Verweilen auf einer Internetseite wahrscheinlicher als deren schnelles Verlassen (vgl. Kaerlein 2020: 52).

25 Siehe auch Kap. 3.1.

Eine Konzentration auf die Betrachtung von Interfaces erlaubt, so argumentiert Kaerlein (2020: 49–52), eine Vermittlung des Forschungsinteresses an Medienpraktiken mit dem an Infrastrukturen. Während der konstitutive Zusammenhang zwischen beiden Größen, die sich wechselseitig hervorbringen, in den letzten Jahren immer wieder betont und deutlich herausgearbeitet werden konnte (vgl. insbesondere Gießmann 2018; aber auch Couldry/Hepp 2021; Burkhardt et al. 2022), entstand in der Forschung zu diesen beiden Themenfeldern zugleich eine Lage, die Kaerlein (2020: 47–48) als „Diskrepanzverhältnis“ zwischen den situierten Praktiken der Anwender*innen auf der einen und den raumzeitlich davon distanzierenden Praktiken der Datenaus- und verwertung auf der anderen beschreibt.²⁶ Kaerlein (2020: 48) wirbt für ein „post-situationales“, wenngleich nicht weniger praxeologisches Medienverständnis, das nicht nur das Geschehen *in* einer Situation, sondern auch vor, nach und zwischen Situationen einbeziehen kann. Dies sei, so Kaerlein weiter, über die Beforschung von Interfaces zu leisten:

Interfaces strukturieren oder formatieren einerseits Praktiken durch ihre eingeschriebenen Affordanzen (a), und andererseits haben Praktiken nur dann eine infrastrukturierende Wirkung, d. h. eine kontinuierliche Einschreibung zur Folge, wenn sie von Interfaces geleitet und vermittelt werden (b). (Kaerlein 2020: 51)

Interfaces bzw. Interface-Praktiken zu untersuchen hilft dabei, dieses Wechselverhältnis nicht nur zu benennen, sondern es beschreiben und kritisieren zu können – denn entlang der Interfaces lassen sich Praktiken Schritt für Schritt als Infrastrukturen erkennen (vgl. Kaerlein 2020: 54). Sprachliche Praktiken mit VUIs im Kontext ihrer Domestizierung zu beschreiben (und ihre Verwebung mit der sozialen Praxis der Nutzer*innen), ist insofern auch ein Forschungsvorhaben, das Kaerleins (2020: 54) Aufruf „Follow the interfaces“²⁷ ernst nimmt und daran aus Sicht der Nutzer*innen ansetzt (siehe auch Hector 2025).

Wie Christoph Ernst (2017) herausstellt und im Rahmen einer Analyse exemplifiziert, entsteht im Austausch mit Interfaces *implizites Wissen* über diese Interfaces, das sich im Alltag der Nutzer*innen verankert.²⁸ Mit diesem impliziten

²⁶ Zu Datenverwertung bei der Nutzung von Smart Speakern siehe Kap. 3.2.6.

²⁷ Dieser Ausdruck ist eine Referenz auf die ANT-Forschung, die als ‚Motto‘ ein „Follow the actors“ ausgerufen hatte (vgl. Latour 1987).

²⁸ Der Begriff des Impliziten Wissens (*tacit knowledge*) hat seinen Ursprung bei Polanyi (1985 [1966]) und wurde bereits dort auf sprachliche Praktiken zwischen explizierbarem Regelwissen und implizitem Regelfolgen (im Sinne Wittgensteins, s. o.) angewendet (siehe auch Sharrock 2012; Loenhoff 2015; 2019).

Wissen schreiben sich Computer, ihre Operationen und auch die großflächige Datenverwertung und die computationale Vernetzung in die sprachlichen und sozialen Praktiken der Nutzer*innen ein (vgl. Ernst 2017: 111). User-Interfaces können wesentlicher Alltagsbestandteil der Anwender*innen sein bzw. werden und dabei implizites Wissen konstituieren, argumentiert auch Wirth (2016: 32): „A user interface can be a space of inhabitation, a comfort zone, which we tacitly know to use and which offers ways of operating and ‚handling‘, that appear to be ‚ready-to-hand‘ to the user“ – damit verbunden ist der Vorschlag, den alltäglichen Gebrauch in die Konzeptionierung des Interface-Begriffs als Medienpraktiken einzubeziehen (vgl. Wirth 2016: 33). Die sprachlichen Praktiken zur Bedienung von VUIs, ihre Ausbildung und Verfestigung, sind Ausdruck dieser sozialen Praktiken und machen „Übergänge zwischen situierter Kognition und dem impliziten Regelwissen von Praktiken“ (Ernst 2017: 101) sichtbar, die entlang der Interfaces vollzogen werden. Wenn sich Nutzer*innen dabei an etablierten, z. B. gesprächsorganisatorischen, Praktiken orientieren, die ebenfalls als implizites Wissen vorliegen, entstehen im Interface-Gebrauch epistemologische Kontinuitäten, deren Erscheinungsformen einer Untersuchung zugänglich sind (vgl. Ernst 2017: 109).

Mit anderen Worten: Einerseits ist zu erwarten, dass Nutzer*innen im ‚Interfacing‘ Bezug auf als implizites Wissen vorliegende Prinzipien der Gesprächsorganisation nehmen und dies durch sprachliche Praktiken zum Ausdruck bringen. Andererseits haben Nutzer*innen auch Vorstellungen von den Funktionsweisen der Systeme („conceptual model“),²⁹ die sie beim Eintritt in den Dialog mit den Interfaces berücksichtigen – dies zeigen Merkle/Hector (2025) am Beispiel wiederholter Reparaturversuche einer VUI-Nutzerin. Zwar besteht mit der gewählten Ausrichtung dieser Arbeit kein direkter Zugriff auf die „situierter Kognition“ der Nutzer*innen und ihr implizites Wissen. Aber: Es lässt sich über eine detaillierte Untersuchung sprachlicher Praktiken als Medienpraktiken zur Bedienung von VUIs verfolgen, wie sich VUIs in die sprachlichen und sozialen Praktiken zu injizieren suchen. In diesem Ausloten von Möglichkeiten und Grenzen symbolischer Operationen mit den Interfaces entstehen in einem Prozess der Domestizierung Medien als Gewohnheiten, Erwartungen, Vorstellungen, Möglichkeitsbedingungen für zukünftigen Austausch, von denen die sprachlichen Praktiken zeugen.

Die primär in den Medienwissenschaften beheimatete Interface-Forschung war und ist größtenteils auf die Untersuchung visueller Effekte und ihrer Ge-

²⁹ Mit dem Begriff *conceptual model* werden zusammenhängende Annahmen bzw. Vorstellungen über Möglichkeiten des Systems und die daraus folgenden möglichen Umgangsweisen mit einem Interface beschrieben, die im Design entlang von Möglichkeiten, Limitierungen und Verknüpfungen konzipiert und von den Nutzer*innen im Umgang mit den Interfaces (implizit mit-) entwickelt werden (vgl. Norman 1988: 12; siehe auch Ernst 2017: 100).

schichte konzentriert, teilweise unter dem Einbezug von Haptik (vgl. Borbach 2019). Die Betrachtung auch akustischer Interfaces ist noch neu und bislang sind nur wenige Arbeiten zu ihrer Untersuchung erschienen (vgl. Schoon/Volmar 2012; Borbach 2018; 2019; Borbach et al. 2025). Dabei muss betont werden, dass die mündliche Steuerung maschineller Anwendungen nicht *prinzipiell* neu ist, so haben z. B. Tonwahl- sowie Sprachdialogsysteme (etwa für vorstrukturierende Abfragen im Callcenter) eine in die Mitte des 20. Jahrhunderts zurückreichende Kulturgeschichte, die als Wegbereiter für die natürlichsprachige Bedienung von Interfaces verstanden werden kann (vgl. Volmar 2019), wobei sich eine Entwicklung von stark prästrukturierten Tasten- und Stimmein- und ausgaben hin zu zunehmend offenen, gesprächsähnlichen und mit Natural Language Processing verarbeiteten Dialogsystemen nachzeichnen lässt (vgl. Volmar 2019: 71–72).

Die Geschichte visueller Interfaces weist Parallelen auf: So setzen sich für visuelle Interfaces in den frühen 80er-Jahren „direct manipulation“-Interfaces (Shneiderman 1983) durch. Sie ermöglichen die Bedienung von Computern über visuelle Repräsentationen statt über symbolische Eingaben (z. B. das Icon eines Druckers für die Druckfunktion statt der zuvor notwendigen Tastatur-Eingabe ‚Strg + P‘). Shneiderman (1983: 66) erklärt den Durchbruch damit, dass mit den visuellen Repräsentationen der computationalen Objekte semantisches Wissen verbunden ist, das die Nutzer*innen sich leichter aneignen und anwenden können, als das zuvor notwendige syntaktische Wissen über die Gestaltung der Eingaben: „Dealing with the re-presentations of objects may be more ‚natural‘ and closer to innate human capabilities“, so Shneidermans (1983: 66) These.

In der Historie akustischer Interfaces vollzieht sich mit dem Entstehen von Voice Assistants eine ähnliche Entwicklung, die weg von prästrukturierten, symbolischen Eingaben und hin zu ‚natürlichsprachigen‘, gesprächsähnlichen Ein- und Ausgaben für die durchzuführenden Operationen führt. Diese Entwicklung steht im Einklang mit der zunehmenden Tendenz in der HCI und im Interface-Design zu Natural User Interfaces (NUIs), die zunehmend den gesamten Körper einbezieht – Stefan Rieger (2019: 190) spricht von einer „Naturalisierung“ sowie von einer „Rückkehr des Sensorischen [...]“, die auf nicht weniger als auf die Auflösung vormals getrennter Bereiche abzielt“. In der HCI-Forschung wird dieser Wiedereinbezug menschlicher Sinne und die Orientierung an menschlichen Kapazitäten und Sinneskonfigurationen als ‚natürlich‘ konzeptionalisiert, zugleich steht dieser Begriff in der Kritik (vgl. Norman 2010). Indes ist noch einmal herauszustellen, dass Nutzer*innen sich in der Bedienung akustischer Interfaces nicht nur an Ähnlichkeiten des Dialogs zu zwischenmenschlichen Interaktionen orientieren, sondern auch an den (unterstellten) technologischen Möglichkeiten und Grenzen der Interfaces (vgl. Ernst 2017) und ihren Sprachgebrauch daran anpassen (vgl. Habscheid 2022: 191). Insofern verfertigen die Nutzer*innen zu Teilen

auch symbolische Eingaben, die von der Gesprächsähnlichkeit wenig übrig lassen (vgl. Merkle/Hector 2025) – eine Folge von Recipient-Design im Interfacing.

Ein letzter Aspekt: Smart Speaker sind als stationäre Lautsprecher konfiguriert, sodass der Zugriff nicht nur in dyadischen Dialogsituationen möglich ist, sondern die Interface-Operationen prinzipiell auch anderen Anwesenden in Mehrparteien-Konstellationen zugänglich sind. Das unterscheidet Smart Speaker etwa auch von mit Sprachassistentz-Software ausgestatteten Smartphones, deren Sehfläche Nutzer*innen ‚standardmäßig‘ als privat behandeln (vgl. Oloff 2019). Damit erzeugen Smart Speaker durch ihre Interfaces eine Spannung zwischen Gemeinschafts- und Individualmedium (für diese Unterscheidung siehe Röser/Peil 2014), auch vor dem Hintergrund personalisierter Benutzer*innen-Konten und der möglichen Privatheit von Kommunikationen. VUI-Dialoge sind damit hörbar für andere Anwesende – und somit auch für die Beteiligung oder Kommentierung dieser. In der akustischen Wahrnehmung der Anwesenden ‚konkurrieren‘ VUI-Dialoge mit laufenden Gesprächen und treten mit diesen unweigerlich in eine Beziehung, was sich auch aufgrund ihrer oberflächlichen (mehr oder minder großen) Ähnlichkeit mit Gesprächen ergibt. Insofern können neben dyadischen Dialogen auch die interaktiven Dynamiken in Gesprächen durch die Interface-Operationen beeinflusst und in die Medienpraktiken einbezogen werden – in den Analysen werden entsprechend beide Konstellationen beleuchtet.

2.1.5 Zwischenfazit Praxeologie und Medientheorie

Die hier aus Gründen der terminologischen Klarheit zur Begriffsbestimmung aufgefächerten Begriffe sind empirisch eng miteinander verwoben, wie auch in den Analysen deutlich wird. Wie nicht nur die Arbeiten Charles Goodwins oder die Studien von Michael Tomasello (2005; 2009) zeigen, besteht eine enge Verbindung zwischen der menschlichen Kooperativität und dem Ursprung und der Entwicklung von Sprache(n). Sprache wiederum ist ein Mittel, das im Prozess der praktischen Verfertigung sozialer Ordnung eingesetzt wird. Die gebrauchsbasierte Linguistik kann also von einer sozialtheoretischen Unterfütterung profitieren, weil sprachliche Formen im Kontext ihrer zentralen Rolle in der reflexiven Hervorbringung sozialer Gegebenheiten verständlich werden. Andersherum haben sprachliche Formen wesentlichen Anteil an sozialen Praktiken und ihrer Entstehung, sodass praxeologische Forschungsarbeiten von linguistischen Betrachtungen Erkenntnisse über die Anatomie sozialer Praktiken erwarten dürfen. Dies gilt ganz besonders für die Untersuchung von Medientechnologien, die mit gesprochener Sprache arbeiten und insofern gleichermaßen sowohl sprachliche als auch soziale Praktiken potenziell rekonfigurieren oder hervorbringen.

Verdeutlicht werden sollten zwei wesentliche Unterscheidungen:

- (1) Während die *Praxis*, „das in einer wechselseitigen Verfertigung befindliche Geschehen“ (Schüttpelz/Meyer 2017: 158) selbst immer situativ, spezifisch in ihrer Situation, in ihren gegebenen raumzeitlichen Umständen und unter Berücksichtigung der konkreten Partizipanden, die das gemeinsame Geschehen wechselseitig verfertigen, neu entsteht und damit einmalig ist, sind demgegenüber *Praktiken* niemals ausschließlich situiert, sondern entstehen erst durch ihre übersituative Verteiltheit. Praktiken sind mit Hirschauer (2016: 46) „situationsübergreifende, analytische Einheiten“, deren „kollektiver Sinn“ (Hirschauer 2016: 56) sich nicht ohne kulturelles Hintergrundwissen erschließen lässt. Die übersituative Verteiltheit von Praktiken in den Blick zu nehmen, ist gerade das, was verschiedene praxeologische Ansätze miteinander verbindet (vgl. Hillebrandt 2014: 58).
- (2) Während die Untersuchung sprachlicher Praktiken auf der Ebene des Sprachgebrauchs und der dabei genutzten „verbalen, vokalen und visuellen Ressourcen“ (Selting 2013: 589) ansetzt, ist die Analyse sozialer Praktiken auf, wenn man so will, ‚größere Einheiten‘ fokussiert. Für den Gegenstand der Smart Speaker sind dies z. B. Üben, Testen und Vorführen, außerdem soll die soziale Praktik des Gesprächs im Zentrum des Interesses stehen. Die Konstitutionsverhältnisse sind jeweils umgekehrt: Sprachliche Praktiken konstituieren Handlungen, miteinander verbundene und verteilte Handlungsschritte können soziale Praktiken konstituieren. Sprachliche Praktiken können daran aber beteiligt sein und wie unter Rückgriff auf Ludwig Wittgenstein und das Konzept der „Sprachspiele“ deutlich wurde, sind sie sogar wesentlich an deren Hervorbringung beteiligt.

Sprachwissenschaftlich können also beide Domänen im Rahmen von Video-Interaktionsanalysen untersucht werden: Einerseits werden die Ressourcenbündel aufgedeckt, die für sprachliche Praktiken mit VUIs relevant sind. Andererseits werden über die Analyse des Sprachlichen auch die sozialen Praktiken und deren Anatomie verständlich. Die Analyse von Gesprächen mit ihren Beteiligungsstrukturen und materiellen Partizipanden kann dabei auch Aufschluss darüber geben, welchen (sozialen) Stellenwert ein Smart Speaker haben kann. Nicht zuletzt soll die Analyse auch zeigen, in welcher Beziehung der Sprachgebrauch zu den technischen Präkonfigurationen der Interfaces steht.

Linguistisch knüpft die Arbeit also sowohl an eine Perspektive aus der Gesprächsforschung als auch an eine bisher eher medienlinguistisch fruchtbar gemachte Sichtweise an. Letztere manifestiert sich erst in neuerer Zeit: Als ‚neue Medien‘ zur Jahrtausendwende zunehmend in den Blick interaktionsorientierter Sprachwissenschaft kamen, etablierten sich in der Linguistik zunächst das Kon-

zept der Kommunikationsformen, mit dem Medien v. a. von ihren technologischen und institutionellen Eigenschaften her in den Blick genommen wurden (vgl. Dürscheid 2005; Brock/Schildhauer 2017), sowie kulturlinguistische Ansätze, die mit einem ‚weiteren‘ Medienverständnis arbeiteten (vgl. Schneider 2008; Linz 2016; Schneider 2017).³⁰ Eine sozialtheoretische Perspektive, die nicht bereits von den Konventionen des Mediengebrauchs ausgeht, sondern den Gebrauch – die *Praxis* – in den Mittelpunkt stellt (und damit eine materielle, lokale und temporale Dynamik³¹ in den Blick nimmt, die mit anderen Ansätzen nicht gleichermaßen eingefangen werden kann), etabliert sich in der Medienlinguistik zunehmend (vgl. Deppermann/Feilke/Linke 2016; Meiler 2019; 2021; Habscheid et al. 2021: 24–29; Waldecker/Hector 2023); einem solchen Ansatz soll hier gefolgt werden.³²

Mit dieser Ausrichtung kann also zusammengefasst erstens Gegenstand der Untersuchung werden, wie die Nutzer*innen durch eine ‚Anhäufung‘ von sprachlichen und mithin sozialen Praktiken selbst das Gerät in der Praxis aneignen und es domestizieren³³ und wie sie zu kompetenten Nutzer*innen werden, und wie die sozialen Praktiken insbesondere im Verhältnis zu ihren Interfaces (als Interface-Praktiken) gestaltet sind. So kann in den Blick kommen, wie das Interface dabei als materieller Partizipand an konversationellen Vollzügen ‚beteiligt‘ wird und welche Verfahren dazu eingesetzt werden. Mit einer Analyse sprachlicher Praktiken im Gebrauch können zweitens sprachliche Praktiken entschlüsselt werden, die in Dialogen mit VUIs zum Einsatz kommen können. Der Charakter der Analyse lässt nicht zu, einen abschließend zu verstehenden ‚Katalog‘ von sprachlichen Praktiken aufzustellen, was angesichts der hohen Flexibilität und Kreativi-

30 Eine vertiefte Diskussion zum Medienbegriff in unterschiedlichen Teilen der Sprachwissenschaft soll an dieser Stelle unterbleiben, siehe dazu neben den genannten Belegstellen etwa Habscheid (2020; 2024) sowie Habscheid et al. (2021: 24–29).

31 Siehe Meiler (2021).

32 Sprachtheoretiker*innen wie Ludwig Wittgenstein oder Karl Bühler sowie einzelne Autor*innen wie Charles Goodwin oder einzelne Verwendungsweisen des Praktiken-Begriffs etwa bei Fiehler et al. (2004) sollen dabei nicht übersehen werden; im wissenschaftsgeschichtlichen Überblick siehe Meiler (2019). Gloning (2016: 460) merkt zur zunehmenden Etablierung des Praktikenbegriffs als ‚Innovation‘ in der Sprachwissenschaft kritisch an, dass die Linguistik in ihrem Denken schon lange Ideen des Gebrauchs und der Praxis von Sprache verankert und „nicht nur aus de Saussure und Chomsky“ bestehe und bestand; er nennt zahlreiche weitere Arbeiten, in denen sich diese Tradition zeige. Er argumentiert allerdings in Abgrenzung zu Pennycook (2010: 6), der einen Verzicht auf derartige Perspektiven mit der Verfestigung der Sprachwissenschaft und dem „Kampf“ um die Anerkennung als eigene Disziplin begründet – und wendet ein, dass im Forschungsdiskurs Arbeiten, die solche Perspektiven bereits früher eingenommen hatten, nicht genügend Berücksichtigung fänden. Diesen wird durch die Ankunft des „practice turn“ in der Linguistik jedenfalls nun zunehmend Aufmerksamkeit zuteil.

33 Zur Bildung dieser beiden Begriffe siehe Kap. 2.3.

tät beim Vollzug sprachlicher Praktiken auch gar nicht sinnvoll wäre. Sprachliche Praktiken machen aber die Transformation bestehender bzw. die Emergenz sprachlicher Praktiken an der Oberfläche sichtbar und lassen insofern Aussagen über das Verhältnis von Sprachgebrauch und Medien bzw. Interfaces zu – ob schon solche Veränderungen nicht einfach „ablesbar“ sind, sondern einer methodisch vorgenommenen Analyse bedürfen (vgl. Deppermann 2001: 59–60). Die Analyse sprachlicher Praktiken ist insofern auch geeignet, um die Veränderung und Neubildung von Praktiken durch bzw. mit VUIs zu erfassen.

2.2 Zum Verhältnis von Mensch und Maschine

Auf diesem praxeologischen Fundament sollen Formen des zwischenmenschlichen Austauschs sowie des Austauschs zwischen Mensch und Maschine sozialtheoretisch konzeptualisiert und begrifflich gefasst werden. Dabei beginnt das Kapitel mit einer Klärung der Termini ‚Interaktion‘ und ‚Gespräch‘ (Kap. 2.2.1), die als zwischenmenschliche Austausch-Situationen begriffen werden. Daran anschließend kann der Austausch zwischen Mensch und Maschine als ‚Dialog‘ gefasst, davon abgegrenzt und unter Bezug auf bisherige Forschung aus diesem Bereich genauer charakterisiert werden (Kap. 2.2.2 und 2.2.3), um die daraus abgeleitete Terminologie in den hieran anschließenden Teilen der Arbeiten sauber verwenden zu können, auch mit Blick auf die methodologische Verortung im Bereich der ethnomethodologischen Konversationsanalyse bzw. der ethnografischen Gesprächsanalyse.³⁴ Das Kapitel begegnet darüber hinaus auch soziologischen Debatten zur Handlungsträgerschaft bzw. Agency von Maschinen und greift dabei die praxeologischen Grundlagen wieder auf, um ein Verständnis von Mensch-Maschine-Verhältnissen zu entwickeln, auf das in den Analysen rekurriert werden kann (Kap. 2.2.4). Die Ausführungen zum theoretischen Hintergrund dieser Begriffe sollen insofern auf das beschränkt bleiben, was für das Verständnis ihres Gebrauchs in der Arbeit notwendig ist; ein Beitrag zu theoretischen Grundsatzfragen, die sich an allen drei Begriffen (Interaktion, Gespräch, Dialog) aufhängen, liegt nicht im Interesse dieser Arbeit. Die Notwendigkeit einer Klärung und genaueren Charakterisierung ergibt sich gleichwohl aus dem Untersuchungsgegenstand der Voice User Interfaces (VUIs) und der Frage, wie der Verlauf und das Produkt eines Austauschs zwischen Anwender*innen und VUIs begrifflich und konzeptionell gefasst werden können.

³⁴ Siehe Kap. 4.

2.2.1 Interaktion und Gespräch

Soziale Interaktionen sollen im Anschluss an den Soziologen Erving Goffman (1983: 2) als „environments in which two or more individuals are physically in one another's response presence“ begriffen werden. In fokussierten Interaktionen gehört dazu bei Goffman weiterhin ein geteilter Aufmerksamkeitsfokus, eine geteilte Wahrnehmung und die wechselseitige Wahrnehmung dieser Wahrnehmung (vgl. Goffman 1983: 3) sowie eine koordinierte Ausrichtung von Handlungen am Gegenüber. Damit folgt die vorliegende Arbeit den Ausführungen von Antonia Krummheuer (2010: 13–30), in denen unter Rückgriff auf die Studien Goffmans eine vergleichsweise eng gefasste Auffassung von sozialer Interaktion entwickelt wird. Krummheuer schlägt dies in ihrer Analyse zur Abgrenzung von Austausch-situationen mit einem „Embodied Conversational Agent“³⁵ vor und entwickelt für Letzteres den Begriff des „hybriden Austauschs“ (Krummheuer 2010: 5), um zu vermeiden, den Austausch zwischen menschlichen und technischen Entitäten *a priori* als soziale Interaktion zu bezeichnen. Krummheuer (2010: 323) kommt in ihrer Untersuchung zu dem Schluss, dass im hybriden Austausch zwar durchaus vorübergehend Eigenschaften von Interaktion aufzufinden sind, diese aber nicht kontinuierlich aufrecht erhalten bleiben und der technisch präfigurierte Charakter des Austauschs im Vordergrund steht: „Er [der Austausch, T.H.] ist somit weder eine soziale Interaktion noch ein rein instrumenteller Umgang mit einem technischen Artefakt. Der Austausch changiert vielmehr zwischen den Polen“ (Krummheuer 2010: 323). Die Nutzer*innen schreiben also weder sich noch dem technischen Artefakt konsequent zu, Interaktionspartner*innen zu sein.

Die Domäne der Interaktion ist also menschlichen Akteur*innen vorbehalten – diese Perspektive wird auch in der vorliegenden Arbeit übernommen – und bezeichnet lediglich solche Situationen, in denen mindestens zwei menschliche Beteiligte ko-präsent sind, einander wahrnehmen, diese Wahrnehmung wahrnehmen, und miteinander kommunizieren. Dabei teilen sie einen Aufmerksamkeitsfokus und richten ihre Handlungen aneinander aus (vgl. Goffman 1983: 2–3). Absichtlich wird hier insofern mit Goffman ein stärker an Subjekten orientiertes Interaktionsverständnis gewählt als etwa in den Arbeiten Garfinkels, der von der Interaktion ausgehend vielmehr ein Interesse an deren praktischer Hervorbringung und kontinuierlicher Konstitution entwickelte (vgl. vom Lehn 2021: 300), die auch in den folgenden Analysen im Zentrum stehen.³⁶ Diese

³⁵ Siehe Kap. 3.1.2.

³⁶ Siehe auch Kap. 4.1.

beiden Aspekte – Interaktion auf der einen und ihre praktische Hervorbringung auf der anderen Seite – sollen allerdings in der vorliegenden Arbeit zunächst *analytisch* getrennt werden, um zwischenmenschliche, soziale, sprachliche Interaktion und den Austausch mit einem VUI voneinander abzugrenzen und sprachliche Praktiken aufdecken zu können, die spezifisch für Letzteres sind. Damit soll gerade *nicht* ausgeschlossen sein, dass auch der Austausch zwischen Mensch und VUI phasenweise interaktionsähnliche Eigenschaften hat – vielmehr geht diese Festlegung gerade davon aus, dass hier auf der Phänomenebene eine Unschärfe besteht. Ähnlich wie in Krummheuers Arbeit kann dies aber nicht vor der Analyse festgelegt werden, da ja u. a. die sprachlichen Praktiken Gegenstand der Untersuchung sein sollen, mit denen VUIs an der Praxis ‚beteiligt‘ werden. Insofern scheint es sinnvoll, wenigstens einen Begriff für den rein zwischenmenschlichen, an Subjekten und Handlungen orientierten Austausch vorzubehalten.

Mit dem Begriff *Gespräch* wird hingegen im Folgenden eine soziale Praktik bezeichnet – die des konzentrierten Austauschs von Redebeiträgen und deren Organisation. Mit Goffman (1979: 6)³⁷ kann unterschieden werden zwischen „talk“ und „a talk“ – dem „Reden“ und dem „Gespräch“. Ein Gespräch benötigt zu seiner Konstitution ein irgendwie gestaltetes „Reden“, wobei dies nicht notwendigerweise verbalsprachlich sein muss – man denke etwa an Gebärdensprache. Hingegen entsteht nicht durch jedes Reden schon ein Gespräch: Reden kann nämlich auch Teil einer anderen Praktik sein, „one in which nonlinguistic events may have the floor“ (Goffman 1979: 15), etwa bei Verkaufssituationen, Begegnungen zwischen Gästen und Servicepersonal in einem Restaurant oder ärztlichen Untersuchungen. Umgangssprachlich ist dann zwar häufig auch von Gesprächen die Rede, obschon die vollzogene Praktik nicht primär ein Gespräch darstellt. Sehr viele Alltagspraktiken haben gesprächsartige Anteile, sind aber darum noch nicht oder nicht in erster Linie ein Gespräch. Im Folgenden sollen als Gespräch v. a. solche Situationen bezeichnet sein, in denen die Durchführung dessen im Zentrum steht. Solche Situationen können je nach sozialem Kontext und möglicher Zweckbindung der Gespräche sehr unterschiedlich gestaltet sein, wozu die Forschung zu kommunikativen Gattungen (Günthner/Knoblach 1994) einen großen Beitrag leisten konnte, auch im interkulturellen Vergleich (vgl. Günthner/Linke 2006).

37 Die im Folgenden verwendeten deutschen Begriffe sind der Übersetzung von Knoblach/Leuenberger/Schnettler (2017) entnommen.

Zugleich wird methodologisch bei der Untersuchung von Gesprächen als soziale Praktik in der vorliegenden Arbeit an die Gesprächsforschung nach Deppermann (2008) angeknüpft. Gespräche setzen sich demzufolge zusammen aus gemeinsam zu verfertigenen Handlungen: Die „Interaktivität“ und der „Handlungscharakter“ sind mit Deppermann (2000: 98) zwei der konstituierenden Prinzipien von Gesprächen. Zusammen mit drei weiteren Eigenschaften – ihrer Methodizität, ihrer Sequenzialität und ihrer Reflexivität – stellen sie die ‚Anatomie‘ von Gesprächen dar (hierfür und für das Folgende vgl. Deppermann 2000: 98). Die *Methodizität* verweist auf den mehr oder weniger hohen Grad an methodischer Routinisierung bei der Durchführung von Gesprächen. Die *Sequenzialität* ist als grundlegende Eigenschaft durch die Zeitlichkeit bedingt; sie gilt insbesondere für die verbalsprachlichen Anteile des Gesprächs und steht vor der Herausforderung, diese Anteile nicht überzufokussieren und die Gleichzeitigkeit anderer Bestandteile des Gesprächs nicht zu übersehen (vgl. Deppermann/Schmitt 2007: 27–29).³⁸ Die *Reflexivität* ist die wesentliche übersituative Eigenschaft von Gesprächen: Dadurch, dass die Gesprächsteilnehmenden ihr Gespräch in einem Kontext auf eine bestimmte Weise führen, „bestätigen [sie] stets aufs Neue die Wirklichkeit und Gültigkeit der betreffenden Kontexte“ (Deppermann 2000: 98; siehe auch Garfinkel 1967: 1) – und machen sie so übertragbar auch auf andere Situationen.

2.2.2 Dialog zwischen Mensch und Maschine

VUIs ermöglichen den Austausch zwischen Mensch und Maschine über dialogisch gestaltete verbale und non-verbale Signale. Der Austausch zwischen menschlichen und maschinellen Einheiten kann jedoch – das wird aus den obigen Ausführungen deutlich – nicht mit zwischenmenschlicher Interaktion oder einem Gespräch gleichgesetzt werden – zu verschieden sind die in einen Austausch tretenden Entitäten (vgl. Pfänder/Wagner 2009: 92; Krummheuer 2010). Für den ‚Austausch‘ mit den VUIs soll daher der Begriff *Dialog* verwendet werden, auch wenn diesem Begriff aus Sicht der empirischen Linguistik durchaus skeptisch begegnet wurde (vgl. Imo 2016: 351; Auer 2016). Dialog steht dabei nicht in (etymologisch ohnehin fragwürdiger) Opposition zu Monolog oder Polylog (vgl. Linell 2009: 3), auch soll mit dem Begriff nicht die praxeologische Analyse auf Sprachliches oder gar ‚Vernünftiges‘ reduziert werden, wovor Auer (2016: 358) warnt.³⁹ Vielmehr wird versucht, mit dem Dialogbegriff einen weiter gefassten, komplementär zu

³⁸ Siehe auch Kap. 4.4.

³⁹ Für die Methodologie der Arbeit siehe Kap. 4.

Interaktion verstandenen Begriff für den Austausch zwischen Mensch und Maschine zu entwickeln. Michail Bachtin grenzt in seinen sprachphilosophischen Arbeiten ein ‚enges‘ von einem ‚weiten‘ Verständnis von Dialog ab: Das enge Verständnis, der „real-life-dialogue“ (Bachtin 1986: 75), sind Gattungen des alltäglichen Sprechens (vgl. Velmezova 2015: 277). Linell (2009: 4) bezeichnet dies in der Hinführung zu seiner auf Bachtin gründenden Dialogtheorie als „[t]he most down-to-earth meaning of ‚dialogue‘“, als „direct interactive encounter between two or more, mutually co-present individuals ...“. Genau mit dieser Bedeutung von Dialog soll hier *nicht* gearbeitet werden.

Stattdessen wird ‚Dialog‘ in einem weiten, abstrakten Sinn verstanden (vgl. Linell 2009: 5). Bei Bachtin ist dieser weite Sinn von Dialog ein sehr breit gefasstes, nahezu ausuferndes Konzept, das er auf unterschiedliche Lebensbereiche und Analysefelder angewendet hat – darunter bewusstseinspsychologische, religiöse, kulturorientierte, ontologische und ethische bzw. philosophische Fragestellungen (vgl. Velmezova 2015: 277). Dabei steht im Mittelpunkt eine grundlegende Orientierung auf etwas Anderes oder eine*n Andere*n – die „other-orientation“ (Linell 2009: 13). Alle Diskursbeiträge zeichnen sich durch eine grundlegende Responsivität (auf irgendwelche anderen, kontextuellen Bezugspunkte) und Antizipation (möglicher Folgebezüge) aus – ohne, dass dabei geklärt sein muss oder sichtbar ist, wer oder was *genau* die andere Einheit ist und ohne, dass dies Menschen vorbehalten wäre. So haben auch (technische) Artefakte durch in sie eingeschriebene Eigenschaften responsive und antizipative Bedeutungspotenziale, die wiederum in Relation zu den Nutzer*innen der Artefakte treten (vgl. Linell 2009: 346).⁴⁰ Linell (2009: 350) empfiehlt explizit für rezente Entwicklungen im Bereich Künstlicher Intelligenz, dialogorientierte Ansätze zu nutzen und Dialoge (im weiteren Sinne) zwischen Nutzer*innen, Hard- und Software zu betrachten.

So mag der Dialogbegriff bei Bachtin einerseits teils eklektisch, schwierig zu operationalisieren und uneindeutig sein. Bachtin-Exegese ist, wie auch Auer (2016: 357) andeutet, ein ebenso umfangreiches wie, auch aufgrund der sprachlichen Barrieren zum Russischen, schwieriges Unterfangen, das hier nicht in der Tiefe verfolgt werden soll. Andererseits lädt der weitreichende Dialogbegriff, wie ihn im Anschluss daran auch Linell beschreibt, geradezu dazu ein, für noch unbe-

⁴⁰ Linell (2009: 346) führt zur weiteren Auseinandersetzung das Konzept der Affordanzen ein, mit dem Beziehungen beschrieben werden können zwischen (a) den Nutzungsszenarien, für die Artefakte vorgesehen sind, (b) dem, was sie tatsächlich an Nutzungsweisen „anbieten“, und (c) den darauf aufbauenden Praktiken, die Nutzer*innen damit vollziehen; eine vertiefte Auseinandersetzung damit führt an dieser Stelle zu weit und ist für die späteren Analysen nicht von Relevanz. Zum Affordanzkonzept siehe Gibson (1977), Zillien (2008; 2019) sowie Pentzold/Bischof (2019).

stimmte Kategorien des Austauschs verwendet zu werden, denen – schon durch ihre Interfaces – unbestreitbar ein hoher Grad an Dialogizität innewohnt und über die durch eine dialogische Analysehaltung Erkenntnisse gewonnen werden können. Auch Fritz/Hundsnurscher (1994a: XIV) verwenden den Begriff in dieser Weise und wenden ihn explizit auf den „Dialog“ zwischen Mensch und Maschine“ an – als deviante Form zu zwischenmenschlichen Interaktionen.

Es soll in diesem Sinne also von einem ‚Dialog‘ im weiteren Sinne gesprochen werden, wenn die Nutzer*innen und die VUIs in einen Austausch treten (*VUI-Dialog*). Dieser Dialog sieht zwar durch die Spezifik von VUIs an der Oberfläche einer Interaktion nach Goffman (oder einem Dialog im engeren Sinne nach Bachtin) durchaus ähnlich, wie wir noch sehen werden – und gerade das macht ihn als linguistischen Untersuchungsgegenstand interessant – soll damit aber nicht verwechselt werden. VUI-Dialoge sind Bezugnahmen auf technische Artefakte, wie sie auch für die nicht-stimmliche, sondern z. B. auf Maus und Tastatur basierende Bedienung von PCs angewendet werden könnten. Interfaces konstituieren sich – wie im vorherigen Kapitel dargestellt – prozessual an und in diesen Dialogen, d. h. im Sprechen und Hören; Interfaces sind „der Ort eines ‚Dialogs‘ zwischen Mensch und Computertechnologie“ (Ernst/Bächle 2020: 419) – ein Ort, der nur dann entsteht, wenn dieser Dialog auch stattfindet.

Vor dem Hintergrund, dass die genaue Funktionsweise von Smart Speakern und damit auch VUIs den Nutzer*innen zu großen Teilen opak bleibt und nicht offensichtlich ist, mit wem genau gesprochen wird und wohin das Gesprochene übertragen wird, scheint eine Betrachtung als Interaktion, die ja gerade die wechselseitige Orientierung am Gegenüber und die Wahrnehmungswahrnehmung zum Ausgangspunkt hat, auch theoretisch unpassend. Eine Betrachtung hingegen, die die grundlegende Responsivität und Antizipation als Eigenschaften der Äußerungen beider Parteien setzt, nicht aber die zeitlichen oder sozialen Bezugspunkte *a priori* festlegt, die vielmehr auch unklar, vielschichtig und diffus sein können, scheint für die Konzeptualisierung dieser Art von Äußerungen⁴¹ bestens geeignet.

Wie an den bisherigen Ausführungen deutlich wird, sind die Konzepte Interaktion, Gespräch und (VUI-)Dialog nur analytisch zu trennen. Aus Perspektive der Teilnehmenden vollziehen sie sich in einer einzigen sozialen Praxis, in der nicht nur die einzelnen Konzepte sich gleichzeitig vollziehen, sondern auch übergänglich sind. War ein Austausch zwischen Ärztin und Patient zunächst eine ärztliche Untersuchung, in der das Reden keine gesprächshafte Funktion hatte, kann der

41 Wie Velmezova (2015: 287–288) ausführt, sind es für Bachtin gerade die Äußerungen („utterances“), die in einer genaueren dialogischen Untersuchung vielversprechend sind.

daran anschließende Austausch durchaus eine spezifische Form des Gesprächs annehmen, z. B. ein Beratungs- oder Aufklärungsgespräch, das dann auch in einer veränderten körperlichen Konstellation stattfindet: Die Ärztin wird sich möglicherweise bemühen, einen besseren Blickkontakt zum Patienten herzustellen, ggf. ist auch entsprechendes Mobiliar vorhanden, das den Austausch rahmt und als Interaktionsressource dienen kann. Je nach Gesprächstyp sind also die körperlichen und lokalen Rahmenbedingungen anders. Zu beiden Teilen dieser exemplarischen sozialen Situation gehört außerdem (unterschiedlich ausgeformte) Interaktion.

Gerade daran zeigt sich aber im Hinblick auf die Fragestellungen die Sinnhaftigkeit einer analytischen Trennung: Der Charakter des sprachlichen Austauschs mit stationären VUIs in den häuslichen Alltagsumgebungen der Nutzer*innen ist erst durch wenige Arbeiten empirisch beleuchtet worden, die noch diskutiert werden (vgl. Porcheron et al. 2018; Beneteau et al. 2019; Habscheid 2022; Hector 2022; 2025; Habscheid/Hector/Hrncal 2023; 2025; Reeves/Porcheron 2023). In diesen Arbeiten – und auch in anderen Studien zu ‚artverwandten‘ Geräten, u. a. bei Krummheuer (2010) oder Pitsch (2015) – wird erstens deutlich, dass der Austausch an der Oberfläche Ähnlichkeit zu menschlichen Gesprächen hat – er vollzieht sich primär verbalsprachlich, handlungsorientiert, sequenziell und methodisch und erfüllt damit schon einige der weiter oben referierten Eigenschaften nach Deppermann (2000: 98–99). Zweitens vollzieht er sich parallel und potenziell überlagernd zu zwischenmenschlichen Gesprächen. Da gerade dieser Austausch, sein Verhältnis zu menschlichen Gesprächen und die ihn verfertigenden sozialen Praktiken Gegenstand der Untersuchung sein sollen, bietet es sich zum einen an, die Unbestimmtheit des Austauschs auch begrifflich zu fassen und von VUI-Dialogen (im Bachtinschen, weiten Sinn) zu sprechen. Es ist insofern das Anliegen auch dieses Begriffsinstrumentariums, von einer zu starren konzeptionellen Festlegung über den Austausch mit VUIs Abstand zu nehmen. Zum anderen soll eine möglichst große konzeptionell-begriffliche Sicherheit für die parallel vollzogene soziale Praxis geschaffen werden.

2.2.3 Charakteristika von Mensch-Maschine-Dialogen

Auch wenn mit ‚Dialog‘ also ein Begriff für den Austausch zwischen Mensch und Maschine gefunden ist, muss dieser insbesondere für die in späteren Kapiteln folgenden Analysen zum Gesprächsbeteiligten-Status und zur praktischen Einbindung in interaktive Prozesse der sozialen Praxis genauer bestimmt und charakterisiert werden. Dazu sollen zunächst die Arbeiten der Anthropologin Lucy Suchman als Pionierin auf dem Feld der Mensch-Maschine-Interaktion und der

interdisziplinären Betrachtung zwischen Computer- und Sozialwissenschaften ausgewertet werden, die auch methodologisch bereichernd sind. Unter besonderer Berücksichtigung der Analysen der Soziologin Antonia Krummheuer (aus dem Feld der Wissenschafts- und Techniksoziologie) sowie des sozialtheoretischen Entwurfs nach Stefan Hirschauer werden anschließend Konzepte zur Kommunikation zwischen Mensch und Maschine und zur ‚Beteiligung‘ von Maschinen am praktischen Geschehen vorgestellt und auf die Fragestellungen der vorliegenden Arbeit angewendet.

Suchman unternahm in ihrer Arbeit „Plans and Situated Actions“ (1987) als eine der ersten Forscher*innen den Versuch, zwischen Informatik und Sozialwissenschaften zu vermitteln (vgl. Suchman 2007: 4–5).⁴² Sie ging dabei der Frage nach, wann Kategorien, Eigenschaften und Zuschreibungen von „Mensch“ und „Maschine“ relevant werden und wie diese Unterschiede in bestimmten Situationen aufgeführt (englisch „enacted“) werden – und welche Konsequenzen dies diskursiv und materiell hat (vgl. Suchman 2007: 2). Mehr noch als die erste Ausgabe ihrer Arbeit von 1987 prüft sie in der zweiten Auflage: Wie werden die Grenzen gezogen und *in situ* von den Beteiligten relevant gesetzt? Suchman betont, dass sie kategorial-theoretische Debatten über das Sein und das Wesen von Maschinen und Menschen ablehnt und dieser mit empirischer Evidenz begegnen will – dieser Position schließt sich die vorliegende Arbeit an. In ihren Arbeiten geht Suchman mit Mitteln vor, die denen der ethnomethodologischen Konversationsanalyse ähnlich sind (vgl. Suchman 2007: 10–11). Suchman (1990) stellt heraus, dass in verschiedenen kommunikativen Arrangements auch verschiedene kommunikative Ressourcen im Gebrauch sind. Der Austausch zwischen Mensch und Maschine lässt sich insofern als eine Extremform von ressourcenlimitierter Kommunikation fassen, auf den die gleichen Methoden angewendet werden können – d. h. bei Lucy Suchman ethnomethodologisch-konversationsanalytische Verfahren (vgl. Suchman 1990: 43). Über diesen methodologischen ‚Trick‘ gelingt es Suchman, die Unterschiede zwischen „human interaction and machine operation“ aufzuzeigen und sich von „a priori identification“ (Suchman 1990: 46) der Eigenschaften zu lösen.

Durch ihre detaillierten Analysen konnte sie zeigen, dass die Ko-Konstruktion des Austauschs zwischen Mensch und Maschine wesentlich davon geprägt war, dass Asymmetrien zwischen den beiden Entitäten herrschten: Während die

⁴² Eher theoretisch ausgerichtete Versuche einer Vermittlung begannen bereits in den frühen 80er-Jahren (siehe etwa Friedrich 1980), doch entbehrten sie überwiegend empirischer Untersuchungen und fokussierten auch nicht auf die interaktionistischen Aspekte des Mensch-Maschine-Austauschs. Die ebenfalls zentralen Arbeiten von Agre/Chapman (1987) werden im weiteren Verlauf der Ausführungen noch diskutiert.

Anwender*innen keinen Zugang zu den in die Maschine eingebauten Skripten und damit zu deren ‚Erwartungen‘ hatten, kann die Maschine den größten Teil menschlicher Handlungen nicht wahrnehmen: Für die Ko-Konstruktion sind lediglich Eingaben relevant, die den Status der Maschine verändern können. Menschliches Handeln ist aber immer – selbst, wenn es vorgeplant ist – situativ und selbst das Planen selbst ist situiert. Planen sowie auch ‚das Handeln an dem Plan orientiert zu begreifen‘, werden als von den Beteiligten gemeinsam hervorbrachte Leistungen verstanden, die Garfinkel/Sacks (1979: 135) folgend, sprachlich „beobachtbar“ und „berichtbar“ (mithin „accountable“) gemacht werden (vgl. Suchman 2007: 13).⁴³ Das Handeln als geplantes Handeln zu verstehen und zu kennzeichnen, „accountable“ zu machen, ist insofern eine reflexive Sinnzuschreibung nach Garfinkel (1967: 1).

Suchman (1987) zeigt in ihren Analysen von Erstnutzungen eines Hilfesystems zur Bedienung eines Fotokopierers (siehe Suchman 2007: 126–168), dass die fehlende Fähigkeit zur reflexiven Hervorbringung und gerade die menschliche Fähigkeit zur Reflexivität und deren Anwendung Blindstellen bei der Maschine verursacht:

The new user of a system, however, is engaged in ongoing, situated inquiries regarding an appropriate next action. Although the instructions of the expert help system are designed in anticipation of the user's inquiries, problems arise from the user's ability to move easily between a simple request for a next action [...] and embedded requests for clarification of the actions described within a procedure. (Suchman 2007: 167)

Das Design einer Maschine ‚erzwingt‘, um diese Blindstellen und fehlende Reflexivität auszugleichen, eine bestimmte Reihenfolge für den Ablauf von Handlungen und erzeugt dadurch den Eindruck einer maschinellen Antwort – und „enables the appearance of instructions occasioned by the user's actions“ (Suchman 1990: 31). Die menschlichen Handlungen sind also Aufführungen in einem vorstrukturierten Plan, der den Austausch überhaupt ermöglicht – die *tatsächliche* Aufführung allerdings bleibt der Maschine weiterhin verborgen, und somit z. B. alle anderen soziomateriellen Vorgänge, Einstellungen, unterschiedliche Auffassungen, die die Nutzer*innen von der Planstruktur haben, Unterbrechungen und Störungen.

⁴³ Suchman (2007: 71) verweist auch auf die Arbeiten G. H. Meads zum Verhältnis von situierter Aktivität und geplanter Aktivität, wobei sie betont, dass der geplanten Aktivität schon die Vorstellung von der Situierung innewohnt. Diese sozialtheoretische Diskussion soll hier nicht weiter verfolgt werden.

Es zeigt sich also eine Differenz, die die Informationswissenschaftler Philip Agre und David Chapman⁴⁴ herausgearbeitet haben: „plans-as-program“ und „plans-as-communication“. Während die erste Perspektive v. a. die Ausführung von determinierten Handlungen oder Schritten in einer miteinander verbundenen Abfolge beschreibt, ist die zweite nicht-deterministisch. Sie geht von einer fortlaufenden Improvisation bei der Befolgung von Plänen aus, die als eine Ressource verstanden werden können, die bei der Bewältigung sozialer Aufgaben zum Einsatz kommt (vgl. Agre/Chapman 1987: 17). Ein Hauptanliegen, das Suchman, Agre und Chapman teilten, war die Vermittlung dieser Unterscheidung in die Sozialwissenschaften, und mehr noch in die Informatik hinein, in Verbindung mit daraus abgeleiteten Konsequenzen für das User-Design, um die Grenzen der ‚Planbarkeit‘ sozialer Handlungen einerseits, aber auch die Geplantheit der maschinellen Prozesse andererseits in eine Verbesserung der Systeme einfließen zu lassen.

Suchman, Agre und Chapman begreifen auf diese Weise den Austausch zwischen Mensch und Maschine als asymmetrischen Dialog, in dem unterschiedliche Konzepte vom *praktischen* Ablauf von Handlungen und unterschiedliche Ausdrucksressourcen zu Missverständnissen, Störungen und Abbrüchen führen können und andererseits Verfahren entwickelt werden, um diese Grenzen zu identifizieren und zu überwinden. Die dabei genutzte Vorgehensweise der ethnomethodologischen Konversationsanalyse soll auch in der vorliegenden Arbeit zum Einsatz kommen.⁴⁵ Eine Folge von Äußerungen wird sequenziell analysiert und es wird dabei an der Oberfläche zunächst nicht zwischen menschlichen und maschinellen Äußerungen unterschieden – die Situation wird als Ganzes betrachtet und alle Ausdrucksressourcen werden einbezogen. Die Erkenntnisse von Suchman über Asymmetrien und fehlende Ressourcen bzw. Wahrnehmung sowie die Vorstellungen von „plan-as-program“ und „plan-as-communication“ werden herangezogen, um interaktive Abläufe aus verschiedenen Perspektiven zu analysieren, Beteiligungsstrukturen herauszuarbeiten und Störungsquellen zu identifizieren.

Die Analyse soll dabei auch die Ergebnisse von Krummheuer einbeziehen und ein darauf sowie auf die Arbeiten von Suchman aufbauendes Verständnis

⁴⁴ Wie Suchman (2007: 13–14) beschreibt, haben sich Suchman, Agre und Chapman insbesondere in den 1980er-Jahren wechselseitig in ihrer Forschung beeinflusst; sie waren zur gleichen Zeit an der Erforschung der Einbettung technologischer Systeme in soziale Zusammenhänge interessiert – Agre und Chapman aus informationswissenschaftlicher Perspektive, Suchman aus anthropologischer. Agre entwickelte aufbauend auf dem Austausch seine Konzeption von „Critical Technical Practice“ (Agre 1997; siehe auch Hind/Seitz 2021).

⁴⁵ Siehe Kap. 4.

von VUI-Dialogen entwickeln. Ähnlich wie Suchman sucht auch Krummheuer (2010: 6) „keine ontologische Debatte darüber, ob Technik handeln kann oder nicht“. Stattdessen werden strikt empirisch Situationen analysiert, in denen Nutzer*innen mit einem virtuellen Agenten in Kontakt treten. Stellt Suchman v. a. die Asymmetrien heraus, die sich in der Kommunikation mit Maschinen ergeben, betont Krummheuer (2010: 10; 323–325) den *hybriden* und *ambigen* Charakter des Mensch-Maschine-Dialogs. Während Hybridität im menschlichen Alltag Krummheuer (2010: 323) zufolge ein gängiges Phänomen ist, das bei allen menschlichen Anwendungen von technischen oder dinglichen Artefakten entsteht, wird es im Rahmen der Austauschsituation zwischen Menschen und dialogischen Maschinen (wie in Krummheuers Beispiel dem virtuellen Agenten Max) relevant gesetzt: So betone einerseits der Agent regelmäßig sowohl seine Eigenschaft als technisches Artefakt als auch seine Rolle als sozialer Interaktionspartner, andererseits orientierten sich auch die Nutzer*innen auf den ‚Zwischenstatus‘ des Geräts. Dieses weist sowohl menschliche Eigenschaften auf (z. B. die grundlegende Sprachfähigkeit, Mimik und Gestik oder visuelle Repräsentation durch einen dem Menschen nachempfundenen Körper auf einem Bildschirm), wie auch maschinelle Charakteristika (z. B. eine stark funktionalistische Syntax oder die physische Beschaffenheit nicht aus ‚Fleisch und Blut‘, sondern aus Plastik, Draht und anderen Materialien zum Aufbau von Computern): „Der hybride Austausch bezeichnet somit die wechselseitige technisch vermittelte Koordination von Aktivitäten zwischen zwei Welten und zwei Wesen“ (Krummheuer 2010: 105). Dieses ‚Zusammenlaufen‘ von Welten bzw. Eigenschaften, die zuvor unterschiedlichen Domänen angehörten (Menschen und Computern) und nun nicht mehr getrennt voneinander auftreten, wird sowohl durch die Nutzer*innen wie auch durch die Maschine selbst (bzw. durch die einprogrammierten Dialog-Bestandteile) immer wieder Thema im Austausch mit dem virtuellen Agenten: „Die Teilnehmer zeigen eine Orientierung am hybriden Status des Austausch und markieren diese Hybridität als Besonderheit“ (Krummheuer 2010: 325).

Gleichzeitig zeichnet sich der Austausch Krummheuer (2010: 324–325) zufolge durch eine unentschiedene Ambiguität aus. So wird der als hybride Einheit wahrgenommene Dialogpartner nicht konsequent unter dem Aspekt der Einheit betrachtet, sondern immer wieder werden auch die Unterschiedlichkeiten zwischen dem virtuellen Agenten und den Nutzer*innen betont. Gerade dadurch trete Krummheuer zufolge die Hybridität des Austauschs immer wieder hervor und werde immer wieder relevant gesetzt. Nicht nur die Nutzer*innen, sondern auch der virtuelle Agent stellt die Unterschiede häufig heraus und trägt insofern zu der ‚unentschiedenen Unentschiedenheit‘ mit bei, die die regelmäßige Thematisierung der Hybridität zur Folge hat. Dies zeigt sich besonders deutlich in von Krummheuer analysierten Erst-Begegnungen zwischen unerfahrenen Nutzer*innen und

dem virtuellen Agenten im Rahmen einer Veranstaltung. In diesen Situationen treten häufig Test-Situationen auf, in denen der virtuelle Agent als „fremdes und andersartiges Gegenüber“ angesprochen wird, dieser sich aber auch „selbst als ein ungewohntes Gegenüber dar[stellt]“ (Krummheuer 2010: 124).

Außerdem bestätigt Krummheuer (2010: 105) die Unterschiede in der Dialogführung, die Planbasiertheit des virtuellen Agenten im Unterschied zu Nutzer*innen, die es „gewohnt [sind], Gespräche situativ und interaktiv mit seinem Gegenüber auszuhandeln“. Diese Unterschiede führen zu einer „Kontextblindheit“ (vgl. Krummheuer 2010: 109–112), gerade wenn mehrere Nutzer*innen am Austausch mit dem virtuellen Agenten beteiligt sind. Unter Rückgriff auf die theoretische Rahmung von Goffman (1973) zeigt Krummheuer, dass das ‚Publikum‘, d. h. umstehende Personen, die im Rahmen der Veranstaltung den Austausch mit der Maschine verfolgen, wesentlich zur Teilnehmendenstruktur beiträgt und dem virtuellen Agenten – unentschieden – einerseits die Rolle eines „andersartigen“ Gesprächsbeteiligten zuschreibt, ihn gleichzeitig aber im Sinne Goffmans als „non-person“ behandelt, über welche in deren Anwesenheit ohne soziale Sanktionierung gesprochen werden kann. Auch anhand der Analyse ritueller Klammern (z. B. Begrüßungen und Verabschiedungen) zeigt sich, dass durch die vorgegebene Programmstruktur teilweise „der Eindruck eines wechselseitigen Aushandlungsprozesses [entsteht]“ (Krummheuer 2010: 160), der allerdings gerade durch seine Vorstrukturiertheit Limitierungen im Aktionsradius der Maschine verdeutlicht. Dadurch werden Deutungsrahmen für Situationen, Beziehungen und Äußerungen im Verlauf des Gesprächs brüchig und inkonsistent. Durch Nutzer*innen, die das Gerät aktiv an seine Grenzen bringen wollen und es herausfordern und testen, wird die technische Seite stärker betont als durch Nutzer*innen, die sich den Äußerungen des Geräts anpassen (vgl. Krummheuer 2010: 175).

Die Struktur des hybriden Austauschs bezeichnet Krummheuer (2010: 229) als „aufgebrochen“: Während im Zentrum eine „Abfolge von Aktivitäten“ steht, „in denen eine Handlung des Nutzers zu einer Veränderung des Zustands der Maschine [...] führt“, ist diese Aktivitätenfolge eingebettet in Umgebungshandlungen, die mit Arminen (2005) als „pre-“ und „post-monitoring of the action“ charakterisiert werden können, die aber von der Maschine nicht einbezogen werden (so auch Suchman 1987). Krummheuer weitet diese ‚Aufgebrochenheit‘ auch auf die sequenzielle Verknüpfung der Beiträge auf. Insbesondere nach Paarsequenzen, bei denen der zweite Beitrag von der Maschine produziert wurde, folge häufig kein dritter Zug, der eine Verstehenssicherung dokumentieren oder Korrekturen bzw. Reparaturen mit sich bringen würde (vgl. Krummheuer 2010: 229). Krummheuer entwickelt auf Basis dessen ein Modell für den hybriden Austausch, das zusammengefasst hervorhebt, dass „keine Interaktion zwischen System und einem isolierten Nutzer stattfindet“ (Krummheuer 2010: 231–232). Stattdessen werden

verschiedene Teilnahme- und Erwartungsstrukturen sowie die verschiedenen Vorgehensweisen zur Konstruktion von Sinn bei Mensch und Maschine abbildet.

2.2.4 Beteiligung maschineller Entitäten an der Praxis

Teil der praktisch-situativen Aushandlung des Charakters der Kommunikation ist neben dem Gesprächsbeteiligtenstatus auch die Zuschreibung von Handlungsträgerschaft („Agency“) – ein bereits mehrfach angerissener Themenbereich. Wurde die Diskussion um den Akteursstatus technischer Entitäten zunächst in Philosophie und Anthropologie geführt, trug sie in den 80er-Jahren wesentlich mit zur Begründung der Techniksoziologie bei. In dieser Diskussion treten deterministische Auffassungen von Technik mehr und mehr zurück, stattdessen werden Positionen dominant, die ein Verschwimmen klassischer ontologischer Grenzen von Technik und Sozialem annehmen (vgl. Rammert 1993: 10). Die ausschließliche Zuordnung von Handlungsfähigkeit zum Menschen scheint nicht mehr generell gültig zu sein (vgl. Rammert/Schulz-Schaeffer 2002a).

Die Akteur-Netzwerk-Theorie (ANT) schlägt in dieser Debatte vor, auch bei nicht-menschlichen Akteur*innen („actants“) eines Netzwerks Handlungsträgerschaft zu sehen (vgl. Latour 2005: 70–72) und Handlungen als Knotenpunkte des Netzwerks zu begreifen, in denen Handlungen von „Hybrid-Akteuren“ vollzogen werden – ein solcher werde z. B. beim Schusswaffengebrauch aus „Waffe und Schütze gebildet“ (Latour 2019: 428). Intention als Voraussetzung für Handlungen wird dabei abgelehnt: „An actant can literally be anything provided it is granted to be the source of an action“ (Latour 1996: 373). Krummheuer (2010: 44) führt in diesem Zusammenhang eine Arbeit Bruno Latours an, die die Durchführung von Experimenten zur Untersuchung von Hefe beschreibt – Latour konzeptualisiert dabei neben dem durchführenden Forscher auch die Hefe als Aktant (siehe Latour 2002). Die Ursprünge der ANT, die u. a. in den Laboratory Studies zu finden sind (siehe Lynch 1993; Knorr-Cetina 1995), bahnen zwar eine solche Auffassung an, indem sie Subjekten gleichermaßen wie Objekten, menschlichen und technischen Materialien eine Eingebundenheit in die vollzogenen Handlungen attestiert (vgl. Knorr-Cetina 1995). Auch die darauf aufbauenden Workplace Studies und das Forschungsfeld der Computer-Supported Cooperative Work (CSCW) verweisen auf diese praktische Dimension des Zusammenspiels von menschlicher und maschineller ‚Beteiligung‘ am Geschehen (vgl. Luff/Hindmarsh/Heath 2000; Hughes 2001; vom Lehn 2018). Die ANT jedoch hat sich hier radikaler positioniert und hat die Aufhebung einer ontologischen Scheidelinie zwischen sozialen und technischen Abläufen noch weitergetrieben. Diese sozialtheoretisch avancierte und

radikale Idee wurde für ihre Ambiguität und ihren unklaren konzeptuellen Status (siehe etwa Jones 1996: 296) ebenso kritisiert wie für die Ausblendung wesentlicher Differenzen in der Handlungsfähigkeit von Mensch und Maschine.⁴⁶ Erstens trage die Perspektive der ANT zu einer Diffusion des Akteursbegriffs bei, der die Analysen „intelligenter“ Technik erschwere (vgl. Braun-Thürmann 2002: 63), zweitens blende sie wesentliche Unterschiede im Handlungspotenzial von Menschen und technischen Akteuren aus.

Auch in der vorliegenden Arbeit soll mit einer weniger kategorischen Vorstellung der Verteilung von Agency gearbeitet werden. Dies soll v. a. dadurch gelingen, dass keine (letztlich auch in der ANT vertretene) Akteursperspektive eingenommen, sondern die Praxis in den Vordergrund gerückt wird.⁴⁷ Konsequenterweise muss in dieser auch die Verteilung von Agency als Gegenstand einer fortlaufenden Aushandlung betrachtet werden (vgl. Habscheid/Hector/Hrncal 2023; Waldecker/Hector/Hoffmann 2024). Dazu soll noch einmal auf die Konzeption Hirschauers (2004; 2016) eingegangen werden.⁴⁸ In dieser wird eine Auffassung vertreten, die die Praktiken in den Mittelpunkt rückt, an denen Körper und Dinge – auch technologische Artefakte – teilhaben können („partizipatorisch“), zu denen sie aber nicht immer („kontributorisch“) einen *aktiven* Beitrag leisten (Hirschauer 2004: 75). Dinge und insofern auch technische Artefakte wie Smart Speaker sind in dieser Sichtweise „materielle Partizipanden des Tuns“. Sie können nach Rammert/Schulz-Schaeffer (2002b: 56) als Teil eines „verteilten Handelns“ (man könnte mit Verweis auf spätere Debatten auch sagen: als Teil von Praxis) begriffen werden.⁴⁹ Technologische Artefakte können also am Vollzug der Praxis ‚beteiligt sein‘, in den Prozess der Ausführung oder als Ressource in die Ko-Operation und die entstehende Interaktion eingebunden werden und so auf verschiedene Weise *partizipieren*. Wie diese Partizipation gestaltet ist, ist Teil der fortlaufenden Praxis und wird unter allen Beteiligten ausgehandelt.

Denkbar wäre auch gewesen, zur Frage der Beteiligung von VUIs an Gesprächen (insbesondere in Mehrparteien-Konstellationen) mit Goffmans (1979) im *Footing*-Konzept integrierter Kategorisierung von *ratified participants*, *unaddressed recipients* und *bystanders* zu arbeiten – dieses Konzept wurde etwa von Matthewson (1992) produktiv auf pseudokommunikative Äußerungen von TV-Zuschauer*innen

⁴⁶ Für einen Überblick über wesentliche Strömungen dieser Debatte siehe Braun-Thürmann (2002: 35–65); siehe auch Rammert/Schulz-Schaeffer (2002b).

⁴⁷ Zu den praxeologischen Grundlagen siehe Kap. 2.1.

⁴⁸ Diese wurde bereits als praxeologische Grundlage eingeführt (siehe Kap. 2.1.1) und wird hier im Licht der Mensch-Maschine-Relation nur in aller Kürze aufgegriffen.

⁴⁹ Für eine Perspektivierung dieser Diskussion für die Analyse von Embodied Conversational Agents siehe Krummheuer (2010: 41–48).

angewendet, die sich an Fernsehfiguren gerichtet haben (siehe auch Alač et al. 2020: 28–29). Allerdings bringt ein solches Vorgehen auch Schwierigkeiten mit sich – v. a., weil VUIs eher nicht als *participants* im Sinne Goffmans konzeptualisiert werden können. Im Folgenden soll kurz diskutiert werden, wie weit ein solcher Ansatz reichen könnte und welche Probleme sich daraus ergäben.

Goffman (1979: 7–10) unterscheidet zwischen Teilnehmenden einer sozialen Begegnung, die als solche anerkannt sind und insofern z. B. in die rituelle Rahmung der Situation eingebunden werden (*ratified participants*). Diese können im Verlauf einer Unterhaltung direkt adressiert sein, können aber auch zuhörend am Gespräch teilnehmen ohne primär Angesprochene zu werden (*unaddressed recipients*). Davon abzugrenzen sind Individuen, die nicht an einer sozialen Begegnung teilnehmen, aber dennoch anwesend sind und die Unterhaltung zu Teilen mitverfolgen können (*bystanders*). Unternimmt man den Versuch, mit einem solchen Konzept dem Beteiligungsstatus einer Maschine habhaft zu werden, böte es sich zunächst an, die beteiligten Menschen einer Interaktion als *ratified participants* oder *unaddressed participants* zu verstehen, während der Smart Speaker überwiegend *bystander* zu sein scheint: Er ist in Hörweite, wird aber nur im Rahmen von „cross-play“ (Goffman 1979: 9) – d. h. einer über die Grenzen der sozialen Begegnung hinweg vollzogenen Konversation – angesprochen. Der Umstand, dass Smart Speaker sowohl von Nutzer*innen als auch von Wissenschaftler*innen sowie im journalistischen Diskurs fortgesetzt als „lauschende“ oder „zuhörende“ Geräte bezeichnet werden (vgl. Lau/Zimmerman/Schaub 2018; Pins/Alizadeh 2021; Lind 2021: 363–364; Waldecker/Volmar 2022: 171; Krotz 2022: 26), scheint diese Konzeptualisierung zu bestätigen.⁵⁰ In der Konzeptualisierung von *bystanders* bei Goffman (1979: 8) ist auch ein mögliches Abhören mitgedacht („[...] they may surreptitiously exploit the accessibility they find they have, thus qualifying as eavesdroppers, here not dissimilar to those who secretly listen in on conversations electronically“), was vor dem Hintergrund opaker Verwertung der aufgezeichneten Daten (vgl. Crawford/Joler 2018)⁵¹ auf den ersten Blick ebenfalls stimmig erscheint. Allerdings muss dazu gesagt werden, dass *eavesdropping* bei Goffman nicht als „Abhören“ im Sinne einer Überwachung zu verstehen ist, sondern vielmehr als fokussiertes, sinnverarbeitendes Zuhören, während die bei Smart Speakern gängigen Abhörpraktiken der Firmen viel schematischer ablaufen (siehe auch Hector 2025).

⁵⁰ Lind (2021: 362) wertet dazu korpuslinguistisch auf Basis des Deutschen Referenzkorpus (DeReKo) Belegstellen aus der medialen Berichterstattung für das Lexem ‚Alexa‘ aus und analysiert die ‚Alexa‘ zugeschriebenen semantischen Rollen. Mit Blick auf „hören“, „horchen“, „lauschen“, und „mithören“ entstehe durch die Verwendung dieser Begriffe eine „agensähnliche Rolle“ – Lind (2021: 364) verweist dabei auf den vergleichsweise hohen Intentionalitätsgrad des Zuhörens.

⁵¹ Siehe außerdem Kap. 3.2.6.

Bei genauerem Hinsehen erweist sich eine solche Konzeptionierung allerdings auch in anderer Hinsicht als ungeeignet. So scheint sie nicht nur wenig aufschlussreich zu sein, sondern auch problematisch im Hinblick auf die theoretischen Implikationen des Konzepts für die Untersuchung. Wenig Aufschluss über die Einbindung in die soziale Praxis und den Beteiligungsstatus gibt die Anwendung dieses Konzepts, weil sich bloß erneut bestätigen würde, was spätestens seit den Arbeiten von Suchman (2007) bekannt ist und sich auch in den hier vorgelegten Analysen noch einmal bestätigen wird: Maschinen sind keine interaktional gleichberechtigten Teilnehmenden an sozialen Interaktionen – die Folgen ihrer limitierten Wahrnehmungs- und Ausdrucksressourcen und grundlegend andere Prozessierung verunmöglichen dies. Mit einer Betrachtung nach Goffman käme man insofern zu dem Ergebnis, dass VUIs mal wie *ratified participants* behandelt werden, mal wie *bystanders* – zudem wäre ein schnelles Oszillieren zwischen diesen beiden Kategorien beobachtbar.⁵² Problematisch wäre die Anwendung des Konzepts in einer praxeologisch ausgerichteten Untersuchung, weil sie den VUIs Subjekthaftigkeit zuschreiben würde – und auch wenn Smart Speakern ein gewisser Grad an ‚Eigensinn‘ innezuwohnen scheint und Anwender*innen stellenweise formal Subjekthaftigkeit ‚zuschreiben‘, führt eine solche Unterscheidung konzeptionell zu einer Wiedereinführung der Dichotomie von Subjekt und Objekt, die ja gerade mit einer praxeologischen Perspektive und einer flachen Ontologie nach Schatzki (2016) suspendiert werden sollte (vgl. Schäfer 2020: 18). Insofern soll dieses Konzept, obwohl es möglicherweise naheliegend erscheint, nicht für die Analysen der vorliegenden Arbeit herangezogen werden – die zuvor eingeführten praxeologischen Perspektiven bilden hingegen einen konzeptuellen Rahmen, um sowohl sprachlich-interaktionale Praktiken in dyadischen VUI-Dialogen als auch soziale Praktiken, in denen VUIs ‚beteiligt‘ sind, analysieren zu können, ohne VUIs dabei essentialistisch von Menschen unterscheiden oder diesen implizit Subjekthaftigkeit zuschreiben zu müssen.

2.2.5 Zwischenfazit zum Verhältnis von Mensch und Maschine

Mit Krummheuer (2010) wurde die Auffassung eingeführt, dass es sich beim Austausch zwischen Menschen und Maschinen nicht um *soziale Interaktion* handelt, der auch in den folgenden Ausführungen gefolgt wird. Darauf aufbauend kann in der vorliegenden Arbeit der Austausch als *Dialog* zwischen ungleichen Entitäten konzeptualisiert werden. Die ‚Menschlichkeit‘ des maschinellen Gegenübers wird

⁵² Siehe dazu v. a. die Analysen in Kap. 6.2.

insofern situativ und abhängig von der Teilnehmendenkonstellation, vom sozio-kulturellen Kontext und den materiellen Gegebenheiten immer wieder neu ausgehandelt. Daran partizipieren maschinelle Entitäten im Vollzug der Praxis als „materielle Partizipanden des Tuns“ (Hirschauer 2004) – die Arbeit affirmiert insofern konzeptionell eine praxistheoretische Perspektive, grenzt sich dabei aber von der kurz andiskutierten Positionierung der ANT ab. Agency wird in diesem Sinne eine verhandelte und von den Beteiligten zugeschriebene, empirische Größe und es ist Aufgabe auch der vorliegenden Arbeit, die Aushandlungs- und Zuschreibungsprozesse näher zu analysieren (insoweit sie in sprachliche Praktiken eingebunden sind bzw. darin zum Ausdruck kommen). Dies gilt besonders vor dem Hintergrund des potenziellen Aufbaus von Erwartungen an die Umwelt, v. a. in Form konditioneller Relevanzen in den sequenziell verlaufenden Dialogen mit dem VUI. Auch Fälle von Initiierungen oder durch das VUI ausgelöste Expansionen eines VUI-Dialogs können entsprechende Erwartungen aufbauen. Gerade bezüglich dieser kann vermutet werden, dass sie als Form der eigenständigen Handlung behandelt werden. Den Geräten könnte insofern eine Handlungsträgerschaft und damit auch das Potenzial zur Beteiligung an anderen sozialen Praktiken wie Gesprächen zugeschrieben werden. Der Status der Gesprächsbeteiligung von VUIs ist insofern ebenfalls Gegenstand eines fortlaufenden Aushandlungsprozesses, der mit sprachlichen Praktiken hergestellt oder verworfen wird, wobei das *Gespräch* selbst als soziale Praktik diskutiert wurde. Dies soll im Rahmen der Analyse genauer beleuchtet werden.

2.3 Aneignung und Domestizierung

Neue Medientechnologien in der Wohnumgebung eines Haushalts treten in einen mehrschrittigen und teilweise reziproken Prozess mit den Haushaltsmitgliedern ein. In diesem Prozess machen die Nutzer*innen sich mit den Funktionen der Technologien vertraut, testen ihre Grenzen, weben sie in ihre Alltagspraktiken ein und *domestizieren* die Geräte auf diese Weise. Dieser Prozess vollzieht sich nicht zuletzt in den sprachlichen Praktiken der Nutzer*innen, insbesondere, wenn es um VUIs geht, die ohnehin durch die Produktion verbalsprachlicher und dialogischer Ausdrücke gesteuert werden. Unterschiedliche Strömungen im Schnittbereich von Kommunikations- und Medienwissenschaft, Soziologie und Sprachwissenschaft haben diesen Prozess genauer untersucht und Konzepte sowie Begriffe hervorgebracht, an die die vorliegende Arbeit anknüpfen will: Im Analyseteil werden sprachliche Praktiken herausgearbeitet und das Verhältnis von VUIs zu Gesprächen der Nutzer*innen untersucht. Beides ist Bestandteil der praktischen

Domestizierung bzw. Aneignung der Geräte durch ihre Anwender*innen und die Einhegung der Medientechnologien in die Praktiken der Haushaltsgemeinschaft.

Der Begriff der Domestizierung verweist bereits darauf, dass Medientechnologien in diesem Zusammenhang als etwas beschrieben werden, das von außen kommt, potenziell ‚wild‘ ist und gezähmt werden muss – und zugleich die Wohnumgebung mit der Außenwelt in teilweise opaken Prozessen verknüpft und zugänglich macht (vgl. Waldecker/Hector 2023). Im Anschluss an den Ethnologen Hermann Bausinger (1984) steht dabei die Widersprüchlichkeit des (häuslichen) Alltagslebens im Mittelpunkt: Dieses steht zwischen konservativer Veränderungsfeindlichkeit, die dem Individuum bisweilen zur Last wird, auf der einen Seite, sowie Rückzugsort und Vertrauensbasis auf der anderen Seite. Bausinger arbeitet ein in dieser Spannung stehendes Merkmal des Alltagslebens heraus, auf das die Begründer des Domestizierungsansatzes – im Rückgriff auf Anthony Giddens – mit „ontological security“ verweisen: das Vertrauen in die Konstanz der eigenen Identität, der Kontinuität des Lebens und der direkten Umgebung und insofern auch in die Beständigkeit des Haushalts (vgl. Giddens 2003[1984]: 50). In diese ontologische Sicherheit sind Medien als spezifische Form der Technologie eingewoben: Sie sind verknüpft mit alltäglichen Ritualen und auch wenn die ontologische Sicherheit zunächst durch die Medientechnologien hinterfragt wird (vgl. Silverstone/Hirsch 1992: 17), verlieren sie durch das Eingewoben-Werden ihre Bedrohlichkeit (vgl. Bausinger 1984: 349–350) – sie werden domestiziert. Mit diesem Prozess hat sich der Domestizierungsansatz in unterschiedlichen Projekten und Facetten beschäftigt (vgl. Hartmann 2013a: 39–52).

Zugleich sind im Bereich der Medienrezeptionsforschung ähnliche Fragestellungen untersucht worden. Dabei wurde, ebenfalls in den 90er-Jahren, zunehmend der Begriff der (kommunikativen) *Aneignung* in verschiedenen Disziplinen gebraucht – eine Übersicht über verschiedene Verwendungsweisen liefert Faber (2001: 27). Dieser Begriff betont dabei die aktive Rolle, die Nutzer*innen von Medientechnologien bei der Medienrezeption haben: Etwa Fernsehzuschauer*innen sind nicht einfach passive Konsument*innen, sondern vielmehr durch kognitive, emotionale, kommunikative und somit transformatorische Prozesse aktiv an der Rezeption beteiligt (vgl. Faber 2001: 28; siehe auch Ayaß 1993: 38–39). Diese Prozesse stellen eine natürliche Grenze des Einflusses von Medieninstitutionen und -konzernen dar: Partizipieren die Nutzer*innen nicht oder anders als vorhergesehen an den (von Institutionen bzw. Konzernen geplanten) Medienpraktiken, limitiert dies deren Macht, was für Fernsehen und andere Massenmedien (vgl. Keppler 1993) ebenso gilt wie für datengetriebene Medientechnologien wie Smart Speaker, wenngleich hier eine hintergründige Datenverarbeitung und -verwertung das Bild noch komplexer werden lässt (vgl. Habscheid et al. 2021; Waldecker 2022; Waldecker/Hector/Hoffmann 2024). Für Smart Speaker ist die

aktive Rolle der Nutzenden jedenfalls an der Oberfläche offensichtlich, weil der bidirektionale Charakter des Interfaces dialogische Eingaben erfordert. Gerade die Aneignungsforschung hat sich mit den verbalsprachlichen Äußerungen der Nutzer*innen beschäftigt, die Ausdruck und Vollzug einer Aneignung durch die Nutzer*innen sind: Durch sie binden Nutzer*innen die Medientechnologien und die von außen kommenden Inhalte an die Lebenswelt an und die Rezipient*innen erhalten „die Möglichkeit, [der] eigenen Beobachtung Relevanz zu verleihen“ (Faber 2001: 34). Eine Betrachtung durch die Linse der angewandten Sprachwissenschaft in Verbindung mit (mikro)soziologischen Ansätzen hat sich also als fruchtbar erwiesen, wie insbesondere die Ergebnisse des Projekts „Über Fernsehen sprechen: Die kommunikative Aneignung von Fernsehen in alltäglichen Kontexten“ zeigen (vgl. Holly/Püschel 1993a; Holly/Püschel/Bergmann 2001). Zuerst sollen also Ergebnisse aus dem Bereich der sprachwissenschaftlich-mikrosoziologischen Medienaneignungsforschung zusammengefasst werden, um sie für die Analyse von VUI-Dialogen anzuwenden. Im Anschluss wird das Domestizierungskonzept einschließlich einer Betrachtung seiner Anwendung und Kritik eingeführt.

2.3.1 Kommunikative Aneignung von Medien

Der Aneignungsbegriff hat sich in unterschiedlichen mit der Medienforschung befassten Disziplinen etabliert, um eine aktivere Rolle von ‚Rezipient*innen‘ bzw. ‚Nutzer*innen‘ unterschiedlicher Medientechnologien und -inhalte zu betonen (vgl. Hepp 2005: 67). Im deutschsprachigen Raum hat sich die Untersuchung dieser Medienaneignungsprozesse v. a. an der Schnittstelle von mediensoziologisch und sprachwissenschaftlich ausgerichteter Medienforschung entfaltet (vgl. Hartmann 2009b: 306) und im Kontext der Fernsehforschung, etwa im bereits erwähnten DFG-Projekt „Über Fernsehen sprechen: Die kommunikative Aneignung von Fernsehen in alltäglichen Kontexten“ in Chemnitz, Gießen und Trier, besondere Beachtung erfahren (siehe Holly/Püschel 1993a; Holly/Püschel/Bergmann 2001). In diesem Verständnis ist mit dem Begriff ‚kommunikative Aneignung‘ v. a. auf kommunikative Prozesse während der bzw. im direkten Anschluss an die Medienrezeption abgestellt worden. Ausgangspunkt war der Versuch eines Neuansatzes in der Medienrezeptionsforschung, der von reziproken Verhältnissen zwischen Zuschauer*innen und ‚Medientexten‘ ausgegangen ist (vgl. Holly 2001b: 17).

Für die Aneignungsforschung lassen sich unterschiedliche Forschungstraditionen bestimmen, auf denen Arbeiten aus diesem Bereich aufbauen. Die vorliegende Arbeit bezieht sich v. a. auf Arbeiten aus drei relevanten Strängen, die sich auch teilweise überlappen (im Überblick siehe Holly/Püschel 1993b: 8; Holly

2001b: 17–18): Erstens hat sie, wie die Domestizierungsforschung, Wurzeln in den British Cultural Studies (siehe etwa Hepp 1998), zweitens in einer wissenssoziologischen und damit verbundenen ethnomethodologisch-konversationsanalytischen Tradition (z. B. Keppler 1993; 1996; Ayaß 1993; 2001; Bergmann/Ulmer 1993) sowie drittens eine gesprächslinguistische Ausprägung (etwa Baldauf 2002). In den Cultural Studies wurde der Begriff in einem weiten Sinn gebraucht, um zu beschreiben, dass sich „Menschen eingebettet in kulturelle Kontexte in einem aktiven Prozess Medieninhalte zu Eigen machen“ (Hepp 2005: 69). In dieser weiten Verwendung sei, so Hepp weiter, jede Mediennutzung auch ein Medienaneignungsprozess. Im engeren Sinn hingegen, etwa bei Charlton (1993: 13) ausgeführt, wird mit der Aneignung auf eine spezifische Rezeptionsphase verwiesen (vgl. Winter 1995: 31; siehe auch Faber 2001: 27–28), in der Nutzer*innen eine produktive Auseinandersetzung bei und nach der Rezeption vollziehen. Die Aneignung kann, wie in der Arbeit von Charlton (1993: 13) aus einer sozialpsychologischen Perspektive, sogar den „Intradialog“ umfassen. Als *kommunikative* Aneignung werden hingegen spezifisch sprachliche bzw. interaktive Praktiken begriffen, wie dies etwa Püschel (1993) oder Hepp (1998) fokussieren. Gesprächspraktiken und die Rezeption von neuen (Medien)inhalten sind auch außerhalb der Fernsehsituation eng miteinander verwoben – durch sie verbinden die Nutzer*innen bzw. Rezipient*innen die Inhalte mit ihren Erfahrungen und Lebenswelten, kommen zur gemeinsamen Aushandlung von Bewertungen und wirken vergemeinschaftend und für Institutionen konstituierend (vgl. Ayaß 2012: 2–3), z. B. in Theaterpausengesprächen (vgl. Hrncal 2020: 288–289) oder bei der schulischen Diskussion von TV-Serien (vgl. Weiser-Zurmühlen 2021: 309). Püschel (1993: 115) unterscheidet am Beispiel des Fernsehens zwischen primären und sekundären Thematisierungen: Primäre Thematisierungen sind dabei das „Sprechen über Fernsehereignisse unmittelbar in der Rezeptionssituation“ (Püschel 1993: 115), während sekundäre Thematisierungen von der Rezeption unabhängiges, nachträgliches Sprechen darüber umfasst. Die Arbeit von Hepp (1998) zur Rolle von Gesprächen über das Fernsehen zeigt ebenfalls auf, dass die Aneignung sich zu einem erheblichen Teil konversationell vollzieht und Gespräche als „Katalysator“ (Hepp 1998: 209) für die Medienaneignung wirken: Sie helfen dabei, „gesehene Bezüge und Zusammenhänge zu intersubjektivieren, sich gegenseitig zu orientieren oder gemeinsame Lesarten von Fernsehsendungen auszuhandeln“ (Hepp 1998: 209–210).

Püschel (1993: 118) arbeitet außerdem heraus, dass sich in den primären Thematisierungen das Sprechen nicht als Gespräch im engeren Sinne darstellt, sondern eine weitaus größere Toleranz für Schweigephasen besteht und die Konzentration auf eine andere Aktivität – eben das Fernsehen – vordergründig sein kann; die Teilnehmenden befinden sich insofern in einem „open state of talk“ (Goffman 1981: 107), in dem Teilnehmende zwar berechtigt zu Äußerungen sind,

aber dazu und nicht einmal zu Reaktionen auf andere Äußerungen ‚sozial verpflichtet‘ sind (vgl. Goffman 1981: 134–135). Fernsehbegleitendes Sprechen ist dabei mit Verweis auf Henne/Rehbock (2001: 31) als „empraktisches Sprechen“ konzeptualisiert worden, das in die Ausübung einer anderen, außersprachlichen Handlung eingebunden ist bzw. dieser dient (vgl. Holly/Baldauf 2001: 45).⁵³ Konzeptionell an die Idee des ‚Empraktischen‘, an Goffman und an die genannte Arbeit von Püschel (1993) sowie an Studien von Brünner (1987) und Fiehler (1993) zum Sprechen als Nebenkommunikation anschließend benennt Baldauf (2001: 65–68) unterschiedliche Erscheinungsformen des fernsehbegleitenden Sprechens: „Freistehende Äußerungen“ sowie „Gesprächinseln“. Sie beschreibt anschließend deren kommunikative Strukturen, zu denen neben Interaktionen unter den Fernsehzuschauer*innen auch solche gehören, in denen die Zuschauer*innen die Fernsehfiguren zu adressieren scheinen.

Damit ist ein weiteres Phänomen angesprochen, das im Rahmen der Medienaneignungsforschung beleuchtet wurde: der Dialog zwischen Medieninhalten (z. B. Fernsehfiguren) und Rezipient*innen. Dazu wurde das ursprünglich von Horton/Wohl (1956) entwickelte Konzept der parasozialen Interaktion aufgegriffen und weiterentwickelt. Wie Ayaß (1993: 36) darstellt, hat das Konzept v. a. mit Bezug auf die „Schein-Reziprozität massenmedialer Inhalte“ Anwendung gefunden und ist dabei erneut v. a. in den Dienst solcher Argumentationen gestellt worden, die eine von passiven Medienrezipient*innen bzw. -konsument*innen ausgehen. Dies hat gerade nicht dazu beigetragen, Rezipient*innen als aktive Handelnde zu konzeptionalisieren – obwohl dies im ursprünglichen Konzept nach Horton/Wohl durchaus verankert gewesen sei. „Schein-Reziprozität“ hingegen, so argumentiert Ayaß (1993: 35–36), könne durch viele verschiedene Formen hergestellt werden, die aber alle als soziales Handeln verstanden werden müssten: „Sogar der einsame Zapper ‚handelt‘“ (Ayaß 1993: 36) – und ist insofern nicht in scheinbar (d. h. *para*-) soziale, sondern in soziale Vollzüge eingebettet (siehe auch Mikos 1996). Der Begriff *para*-sozial sei insofern irreführend – was hingegen eben nicht gegeben sei, sei eine ‚echte‘ Reziprozität und folglich ‚echte‘ Interaktivität, dafür aber kommunikative Praktiken, mit denen Schein-Reziprozität und insofern scheinbare Interaktivität hergestellt werde. Ayaß (1993: 36) schlägt konsequenterweise vor, mit dem Begriff der *Para*-Interaktion zu arbeiten, was mittlerweile trotz weiterer begrifflicher Debatten wohl als etabliert gelten kann (vgl. Luginbühl/Meer 2022).

⁵³ Der Begriff des „Empraktischen“ geht auf Karl Bühler zurück, der damit den „Einbau des Sprechens in anderes sinnvolles Verhalten“ (Bühler 1965[1934]: 52) bezeichnet; mit Verweis auf Kap. 2.1.3 unterbleibt eine ausführliche Diskussion an dieser Stelle.

Zudem führen Holly/Baldauf (2001: 44) den Begriff der Pseudokommunikation ein, den sie enggeführt für an Medienfiguren adressierte Äußerungen von Rezipient*innen beschreiben – Sprechhandlungen, in denen Rezipient*innen „reagieren, als könnten die Fernsehakteure sie hören und verstehen“ (Holly/Baldauf 2001: 59) – während Para-Interaktion das gesamte Spektrum von auf scheinbarer Reziprozität beruhenden Handlungen abdeckt und dabei häufig v. a. auf das Handeln der Medienfiguren fokussiert wurde (vgl. Luginbühl/Meer 2022). Pseudokommunikation soll hier also als eine mögliche Ausprägung von Para-Interaktion verstanden werden. Erkenntnisse aus dem bereits erwähnten Projekt zum fernsehbegleitenden Sprechen zeigen, dass pseudokommunikative Äußerungen nicht als „der Versuch, mit Fernsehakteuren zu kommunizieren“ (Baldauf 2001: 72) missverstanden werden dürfen. Stattdessen hat diese Form der ‚Beteiligung‘ im fernsehbegleitenden Sprechen vielfältige Funktionen, etwa solche „emotionaler Aneignung“ (Faber 2001: 36), aber auch „implizite Stellungnahme[n]“ (Baldauf 2001: 72) und Aspekte der Verarbeitung, der Gruppendynamik und der „positiven Selbstdarstellung“ (Klemm 2001: 107) spielen eine wichtige Rolle. Genau dieses Spannungsfeld will die Analyse von ohnehin gesprächsähnlichen VUI-Dialogen sowie ihre Einbindung in Gesprächspraktiken genauer beleuchten.

VUI-Dialoge können Merkmale von Para-Interaktion bzw. Pseudokommunikation aufweisen. Sowohl die VUIs wie auch die Anwender*innen produzieren durch die Stimmein- und ausgaben, in denen sie sich wechselseitig (implizit und explizit) adressieren. Allerdings besteht dabei auch ein entscheidender Unterschied im Gegensatz zu den bisherigen Begriffsbedingungen: Das VUI ermöglicht tatsächlich einen Rückkanal und weist eine Form von Dialogizität auf. Damit ist der Austausch zur Bedienung der Geräte nicht nur *scheinbar* reziprok, sondern tatsächlich von einer gewissen Wechselseitigkeit geprägt – wenngleich nach wie vor nicht von Interaktion im engeren Sinne gesprochen werden kann.⁵⁴ Bei VUIs ist die Para-Interaktion also von einer ganz anders gelagerten Operativität, bei Radio, Fernsehen und anderen massenmedialen Inhalten besteht dieser Kanal hingegen nicht. Als Medienfigur tritt außerdem spezifisch die „Persona“ des VUI durch die synthetisierte Stimme auf. Teilweise produzieren die Nutzer*innen von VUIs zwar auch Äußerungen, die zwar formal an den Smart Speaker bzw. das VUI adressiert sind, die Geräte wurden zuvor allerdings gar nicht aktiviert und können die Eingaben insofern gar nicht verarbeiten. Diese lateralen Adressierungen des VUI, die Funktionen außerhalb des VUI-Dialogs übernehmen, könnten möglicherweise als pseudokommunikative Praktiken verstanden werden. Das Konzept kann aber – aufgrund des synthetischen und funktionalen

⁵⁴ Siehe Kap. 2.2.1.

Charakters des Gegenübers und auch, weil durch eine erneute Aktivierung potenziell ein Kanal hergestellt werden könnte – nicht ‚einfach‘ übertragen werden. Vielmehr sollen diese sprachlichen Praktiken einer genauen Analyse unterzogen werden, u. a. um sie als Teil des Domestizierungsprozesses verständlich zu machen. Auf den theoretischen Hintergrund zu diesem Bereich wird im folgenden Abschnitt eingegangen.

2.3.2 Zum Domestizierungsansatz

Die Metapher der Domestizierung betont v. a. die Beziehungsarbeit zwischen dem „Öffentlichen“, den für die (noch näher zu erörternde) „ontological security“ (Giddens 2003[1984]: 50) bedrohlichen, zu zähmenden und wilden Medien auf der einen Seite und dem „Privaten“, der Wohnumgebung und der sozialen Konfiguration des Haushalts auf der anderen Seite. Gerade in den Texten der frühen britischen Domestizierungsforschung wird dies deutlich: Es ist der Anspruch der Autoren, ein Modell zu entwickeln „aimed towards understanding the nature of the relationship between private households and public worlds and the role of communication and information technologies in that relationship“ (Silverstone/Hirsch/Morley 1992: 13). Der Domestizierungsbegriff wie er in einer Arbeitsdefinition bei der Kommunikationswissenschaftlerin Maren Hartmann (2013a: 146) ausbuchstabiert ist, fokussiert stärker Prozesse der Aneignung von Medientechnologien *innerhalb* der Haushaltsgemeinschaft:

Die Domestizierung von Medien und Technologien umfasst den dynamischen und reziproken Prozess des Medien-Häuslich-und-Zu-eigen-Machens. In diesem Prozess werden den technischen Konsum-Objekten und dazugehörigen Inhalten anhand sozialer Aushandlungen Bedeutungen zugewiesen (u. a. als Teil der sogenannten moralischen Ökonomie). Kernkontext der Aushandlungen ist der Alltag der Mediennutzer, der Routinen, räumliche Anordnungen und ähnliches beinhaltet. Medien werden dabei als besondere technische Objekte definiert, da sie Objekt und Inhalt gleichermaßen umfassen. (Hartmann 2013: 146)

Hier scheint die Vorstellung von Medien als Objekte und Inhalte, die in die häusliche Umgebung und seine sozialen Strukturen eingegliedert und dort reziprok „zu eigen gemacht“ werden, zu dominieren. Gerade datenintensive Medientechnologien wie Smart Speaker tragen allerdings zugleich Inhalte aus der privaten Wohnumgebung nach außen (vgl. Waldecker/Hector 2023: 6). Dies macht eine erneute Öffnung des Blicks auch für die Dynamik zwischen Haushalt und Öffentlichkeit relevant und fordert dichotome Grenzziehungen zunehmend heraus (vgl. Waldecker/Hector 2023: 6).

Hartmann (2020: 48) unterscheidet zwischen einer britischen und einer norwegischen bzw. „nordischen“ Strömung des Domestizierungsansatzes, die sich unabhängig voneinander entwickelt haben und sich in einigen Punkten voneinander unterscheiden. Die Entwicklung des britischen Ansatzes war eingebettet in Arbeiten am von der Wissenschaftsförderung Großbritanniens geförderten „Programme on Information and Communication Technologies“ (PICT) und insbesondere in das darin integrierte und von Roger Silverstone geleitete Projekt „The Household Uses of Information and Communication Technologies“ (HICT). Das inhaltliche Zentrum des Ansatzes bildet die moralische Ökonomiestruktur des Haushalts (*moral economy of the household*), die als Ausgangspunkt der erwähnten ‚ontologischen Sicherheit‘ steht und in die Medientechnologien eingebunden und in ihr von den Haushaltsmitgliedern domestiziert werden und damit wechselseitig eine aktive Rolle in der Gestaltung von Medientechnologien und -inhalten einnehmen (vgl. Silverstone/Hirsch/Morley 1992: 17). Von diesem Zentrum aus entwickelt der Ansatz eine Reihe von Konzepten, sowohl zur Aneignung der Technologien wie auch zur Vermittlung dieser mit privaten und öffentlichen Sphären, die sich als heuristisch sehr produktiv herausgestellt haben (vgl. Berker et al. 2006: 4; Hartmann 2009a).

Ebenfalls in den 1990er-Jahren gewann an der norwegischen Universität Trondheim ein in der Technologie- und Wissenschaftsforschung gründender Forschungsstrang an Kontur, dessen erste Publikationen teilweise noch vor den britischen Studien veröffentlicht wurden und welcher die im erwähnten Projektzusammenhang des HICT entwickelten Konzepte auch nicht einbezog. Die Untersuchungen aus Norwegen sind weiter gefasst: Sie beschränken sich nicht auf den Haushalt und auch nicht auf Medientechnologien, außerdem betrachten sie die Zähmung stärker als gesellschaftlichen denn als individuellen Prozess (vgl. Sørensen 1994: 42; siehe auch Hartmann 2020: 48–49). Sie ähneln dennoch den britischen Domestizierungsstudien, erstens terminologisch, weil sie auch den Begriff der Domestizierung nutzen und als Technologieaneignung prozesshaft beschreiben (vgl. Sørensen 1994), andererseits, weil die norwegischen Studien ebenfalls empirisch – v. a. ethnografisch – arbeiten, den Fokus auf die Praktiken der Nutzer*innen legen und mit diesem Instrumentarium technische Aneignungsprozesse zu fassen vermögen. Hartmann (2020: 49) zufolge war diese Orientierung auf die Nutzer*innen sogar der Initialzündung für den weiteren Ausbau dieses Forschungsstrangs. Die von Hartmann (2020) als ‚nordisch‘ bezeichnete Domestizierungsschule, zu der etwa auch eine finnische Strömung aus dem Bereich der *design research* gerechnet werden kann,⁵⁵ zeichnet

55 Für einen Überblick über Arbeiten in diesem Feld und eine Diskussion zur Klassifizierung der Ansätze als „nordisch“ siehe Hartmann (2020: 50–52).

sich neben der Nutzer*innen-Zentrierung durch eine größere Offenheit aus, sowohl methodologisch als auch im Hinblick auf die untersuchten Technologien. So wurden von Aune (2007) z. B. Technologien der privaten Stromversorgung in den Blick genommen, zu diesen Studien zählen aber auch Untersuchungen zu Autos (Sørensen 1994), Glasfaserkabeln (Löfström 2007) oder E-Government-Anwendungen (Østergaard Madsen/Kræmmergaard 2015). Daran wird deutlich: Der norwegische bzw. nordische Ansatz ist insbesondere für eine Anwendung außerhalb des ‚klassischen‘ Haushalts offener.

Zunehmend ist die Frage gestellt worden, was einen Haushalt charakterisiert: Das „Domestische“ kann nicht nur durch das gemeinsame Bewohnen einer physischen Gebäudeeinheit bedingt sein, sondern der domestische Raum ist auch in den frühen Texten zur Domestisierungsforschung als Raum für ontologische Sicherheit im Sinne Giddens‘ und Vertrauen betrachtet worden, der sich – gerade durch die zunehmende Mobilität der Medientechnologien – von den physischen Wohneinheiten lösen kann (vgl. Silverstone 2006: 234); die Wohneinheit ist dann nur noch Ausgangspunkt und Ressourcenlager für die von dort ausgehende Mobilität (vgl. Bakardjieva 2006: 68–70). Zugleich kann die domestische Sphäre durch Medientechnologien auch in ein Café, in den Stadtpark oder in einen Jugend-Treff verlagert werden. Es lösen sich so naive Dichotomien zwischen Inner- und Außerhäuslichkeit, Privatheit und Öffentlichkeit auf, deren Dimensionsgrenzen nicht an Wohnungsmauern, sondern vielmehr an Praktiken der Bewohner*innen gekoppelt sind, die in diesen hausen und dynamische Beziehungen zwischen diesen Domänen konstituieren. Damit ist auf das Alltagsleben der Nutzer*innen verwiesen, das mit Bakardjieva (2006: 70) die Fokussierung der frühen Arbeiten auf den *Medienkonsum* aufricht und damit unterschiedliche Analyseperspektiven zulässt: Ihr zufolge ist der Begriff des Konsums zu sehr zugeschnitten auf eine Betrachtung aus Design- und Ökonomie-Perspektiven. Demgegenüber erlaubt der Alltagsbegriff, den sie in der Phänomenologie bei Schütz/Luckmann (2003) verwurzelt, eine umfassendere Betrachtung der Realitäten und darin vollzogenen Praktiken der Mediennutzung und -anwendung.⁵⁶

Diese Weiterentwicklungen haben in der britischen Entwicklungslinie erst in den frühen 2000er-Jahren stattgefunden, während sie in den norwegischen Texten schon von Beginn an impliziert waren (vgl. Sørensen 1994). Konsequenterweise herrscht darin auch von Anfang an eine größere methodologische Flexibilität (vgl. Hartmann 2020: 52). Obschon bereits die britischen Ausgangstexte keine methodologische Starre postulieren, sind gerade die Anwendungen des Domestisierungsansatzes um die 2000er von großen methodologischen Unterschieden (in-

⁵⁶ Siehe auch Kap. 4.

nerhalb des qualitativen Spektrums) und von Anpassungen ans Feld geprägt, so dass hier nur schwer ein Methoden-Nexus zu bestimmen ist. Die Studien aus Norwegen und Finnland tendieren zudem dazu, die den Technologien zugrundeliegenden Infrastrukturen stärker zu berücksichtigen und sich eher der Verknüpfung von Erfindung, Design, Umsetzung und Nutzung bestimmter Technologien zu widmen.

Die Grundzüge des Domestizierungsansatzes (d. h. die Dimensionen der Domestizierung, die „doppelte“ bzw. „dreifache“ Artikulation der Medien und die Idee der moralischen Ökonomie des Haushalts) können helfen, sprachliche Praktiken einzuordnen, soziotechnische Ordnungen an der sprachlichen Oberfläche verständlich zu machen und so Beteiligungen, Bewertungen und Statuszuschreibungen zu Smart Speakern, auch innerhalb der Haushaltsgemeinschaft, zu verstehen. Das Modell stellte sich im Rahmen einer Reflexion als „to be particularly useful for the empirical in-depth analysis of daily media consumption“ (Berker et al. 2006: 4) heraus. Bei der Untersuchung von Smart Speakern fallen der physisch verfasste Haushalt, d. h. die ummauerte Wohnumgebung mit ihren Bewohner*innen, mit den dort vollzogenen Alltagspraktiken und den von den gemeinsamen Bewohner*innen geteilten Wertvorstellungen und Positionen, auf den ersten Blick noch zusammen – Smart Speaker sind nicht mobil, sie werden nicht mitgenommen und in wechselnden physischen Zusammenhängen eingesetzt, sondern sind klar dem ‚zu Hause‘ in seiner physischen und sozialen Verfasstheit zugeordnet.⁵⁷ Erst eine Analyse mit Hilfe eines auf diese Domäne zugeschnittenen Modells kann kontrastiv zeigen, wo Außerhäuslichkeit und verschwimmende Dichotomien auch für die Nutzer*innen eine Rolle spielen und die „continuous negotiation“ von „inside und outside“ (Silverstone 2006: 233) sichtbar machen. Zweifelsohne ist diese strukturell nicht nur bei der über vernetzte Cloud-Systeme prozessierten Sprachverarbeitung relevant, sondern auch z. B. bei mobiler Smart-Home-Steuerung, der Anwendung von Geodaten für Wettervorhersagen oder der Verbindung von Streaming-Diensten mit den Smart Speakern. Doch erst durch den Gebrauch der auf das ‚zu Hause‘ zugeschnittenen Konzepte können die Aspekte besonders gut herausgestellt werden, in denen die Nutzer*innen selbst in ihren Domestizierungspraktiken die Grenzverschiebungen relevant machen und sie in ihren alltagspraktischen Umgang mit den Geräten und in den Austausch über die Geräte und damit verbundene Anwendungen einfließen lassen. Es können so wirtschaftliche und gesellschaftliche Prozesse verstanden werden – „through the messi-

⁵⁷ Damit soll nicht ausgeschlossen werden, dass auch in (teil-)öffentlichen Umgebungen und z. B. in beruflichen Kontexten Smart Speaker zum Einsatz kommen können; diese sind aber nicht Bestandteil der vorliegenden Untersuchung.

ness of domestic practices in everyday life“ (Livingstone 2007: 17). Dies gilt umso mehr, wenn diese Praktiken als Teil sozialer, räumlicher und somit interaktiver Ordnungen mit einer praxeologischen Brille an der sprachlichen Oberfläche beobachtet und verstanden werden sollen und muss insofern der erste Schritt als Bestandsaufnahme sein. Anschließend können die Befunde als Prozesse mit ökonomischer und gesellschaftlicher Tragweite ausgewertet werden – Letzteres ist jedoch nicht im Rahmen dieser Arbeit zu leisten.

Wesentlich zur Popularität des Domestizierungsansatzes beigetragen hat die Instruktivität der prozesshaft formulierten Dimensionen der Domestizierung. Diese wurden erstmals von Morley/Silverstone (1990) formuliert, im Folgenden aber immer wieder angepasst und lange nicht einheitlich gefasst (vgl. etwa Silverstone/Hirsch 1992; Silverstone/Haddon 1996; Berker et al. 2006; Hynes/Romes 2006; Sørensen 2006). Diese sollen nachfolgend kurz charakterisiert werden. Die Dimensionen sind allerdings nicht die einzige und möglicherweise nicht einmal die wichtigste Stütze des Domestizierungsansatzes: Vielmehr funktionieren sie nur flankiert durch die Vorstellung von der „doppelten“ bzw. „dreifachen Artikulation der Medien“ sowie durch das, was oben schon als Sphäre der ‚ontologischen Sicherheit‘ beschrieben wurde und konzeptuell bei Silverstone/Hirsch/Morley (1992) als „moral economy of the household“ gefasst ist. Diese beiden Konzept-Säulen charakterisieren die beiden Einheiten, die durch den Domestizierungsprozess miteinander vermittelt werden: die Verfasstheit des Haushalts auf der einen und die Verfasstheit der Medientechnologien auf der anderen Seite. Diese Charakterisierungen sind für die heuristische Anwendung des Ansatzes zentral, weil auch diese sich an dem Anspruch messen lassen müssen, *in situ* beobachtet zu werden.

2.3.2.1 Dimensionen der Domestizierung

Auch wenn sich die verschiedenen Zuschnitte der Dimensionen gerade in den frühen Arbeiten unterscheiden, lassen sich nach Hartmann (2013a: 21–24) vier Bereiche ausmachen: (a) Kommodifizierung bzw. Imagination, (b) Aneignung – mit den Unterkategorien Objektifizierung und Eingliederung – und (c) Umwandlung. Sie sind allerdings nicht in ihrer zunächst naheliegenden Temporalität zu lesen,⁵⁸ sondern vollziehen sich parallel und beeinflussen sich reziprok. Gleichwohl „ist eine gewisse Linearität nicht zu leugnen“ (Hartmann 2009a: 307). Die drei genannten Dimensionen greifen ineinander und zeichnen ein ganzheitliches Bild des Domestizierungsprozesses.

⁵⁸ In der ursprünglichen Konzeptualisierung der Dimensionen – noch als Phasen – waren diese durchaus linear zu verstehen (vgl. Silverstone/Hirsch/Morley 1992: 20–21), wovon jedoch später als Reaktion auf deutliche Kritik abgerückt wurde (vgl. etwa Silverstone/Haddon 1996).

Zur Dimension der Kommodifizierung (*commodification*) zählen alle Aspekte, die noch vor dem Kauf der untersuchten Medientechnologien liegen. Sie können insofern bei der Betrachtung der Domestizierungspraktiken nur mittelbar beobachtet werden: Überlegungen der Hersteller zu Design, Vermarktung und Vertrieb spielen hier eine besondere Rolle. Der für diese Dimension wesentliche Aspekt der Imagination bezieht sich primär auf die Vorstellungen der Designer*innen, Hersteller*innen und Produktmanager*innen von ihren Produkten und deren Nutzungsszenarien bei den Anwender*innen sowie die Platzierung auf dem Markt (vgl. Silverstone/Haddon 1996: 44–46). In der vorliegenden Studie wird diese Seite der Imagination nur gestreift, weil keine Daten dazu erhoben werden. Zugriff besteht allerdings auf Imaginationen der Anwender*innen, wenn diese mit den ersten visuellen, haptischen und (em)praktischen Eindrücken von den Geräten umgehen und diese in der Interaktion verbalisieren: Erwartungen und Einstellungen, die beim Entpacken und der ersten Inbetriebnahme geäußert werden und praktisch in den ersten Anwendungsversuchen zum Ausdruck kommen, werden an der sprachlichen Oberfläche sichtbar und können entsprechend analysiert werden. So entsteht retrospektiv ein Bild von den Imaginationen im Kontrast zu den tatsächlichen realweltlichen Erfahrungen mit den Geräten und ihren Einbindungen in die jeweiligen technologischen Gefüge der Haushalte.

In der zweiten Dimension, der Aneignung (*appropriation*)⁵⁹ wird unterschieden zwischen der Objektifizierung (*objectification*) und der Eingliederung oder ‚Veralltäglichung‘ (*incorporation*). Während bei der Objektifizierung der räumliche Aspekt im Vordergrund steht, ist in der Eingliederung die Zeitlichkeit, die „injection of media technological practices into the temporal patterns of domestic life“ (Silverstone 2006: 235), die entscheidende Größe. Silverstone (2006: 234–235) bezeichnet diese zusammen als Strategien der Domestizierung – Kommodifizierung und Umwandlung begreift er demgegenüber als Rahmung der Aneignung. Hartmann (2013a: 22–23) bereitet die Strategien später als zwei ‚Unterdimensionen‘ der Aneignung auf, wie sie auch hier verstanden werden sollen.⁶⁰ Zur Objektifizierung zählen Praktiken der Platzierung (vgl. Röser 2007: 21) und insofern die

59 Die bei Hynes/Romes (2006) dargestellte Terminologie, in der die Dimension der *appropriation* den Umgang mit Imaginativen zu fassen versucht, zeigt einerseits noch einmal deutlich die Verschränkung der verschiedenen Dimensionen miteinander, ist andererseits aber auch in anderen Arbeiten üblich, in denen die Kommodifizierung mit der Imagination deutlich von der Aneignung abgegrenzt wird; Hynes/Romes (2006: 127) beziehen sich hier insbesondere auf die erste Darlegung der Dimensionen von Silverstone/Hirsch/Morley (1992), in der die Dimension Kommodifizierung noch nicht enthalten war und v. a. in der Aneignung impliziert wurde.

60 Siehe auch Hartmann (2009b: 306).

Einbettung in die materielle, räumliche Umgebung des Haushalts. Die Dimension trägt wesentlich dem Umstand Rechnung, dass die Aneignung von Medientechnologien nicht losgelöst vom sie umgebenden physischen Haushalt stattfindet – und dass Objekte immer in bereits vorkonfigurierten räumlichen Umgebungen erscheinen (Silverstone/Hirsch/Morley 1992: 20). Dies ist keineswegs nur eine Frage der Ästhetik der Inneneinrichtung: Interaktive Verfahren und Ordnungen sind wechselseitig von der räumlichen Umgebung, in der sie stattfinden, geprägt und prägen die räumliche Umgebung (vgl. Schmitt/Hausendorf 2016); räumliche Arrangements können als interaktive Ressource fungieren und ein Faktor bei der inkrementellen Produktion von Teilnehmer*innen-Strukturen sein (vgl. Hausendorf/Mondada/Schmitt 2012). Sie geben insofern Aufschluss über die bereits weiter oben adressierte Frage der Handlungsträgerschaft von Technologien⁶¹ und helfen, Konfigurationen von Gesprächs,beteiligung‘ zu verstehen. Die Dimension der Objektifizierung vollzieht sich insofern nicht nur unter dem Aspekt der Platzierung und der materialen Arrangements, sondern auch in ihrer Anwendung und Nutzung *in* diesen Arrangements (vgl. Silverstone/Hirsch/Morley 1992: 20). Unter Einbezug von Mimik und Gestik ergeben sich analytische Räume, in denen der Prozess der Domestizierung mit Blick auf den Gesprächsbeteiligtenstatus der Geräte beschrieben und verstanden werden kann. Gerade vor dem Hintergrund der primär verbalen Steuerung dieser Technologien lassen sich so über die Analyse von Rederechtsaushandlungen (*conversational floor*) und von Adressierung von Evaluations- und Aufzeigeleistungen Erkenntnisse zu diesem Prozess gewinnen.

Dies überlappt sich mit dem Prozess der Eingliederung (*incorporation*), in dem das Gerät nicht nur in die räumlichen Arrangements im Haushalt, sondern auch in dessen etablierten Ablauf sozialer Praktiken, d. h. dessen Routinen (vgl. Schüttpelz/Meyer 2017: 158),⁶² eingebunden wird, die ja ebenfalls räumliche Arrangements prägen und von diesen geprägt werden (vgl. Silverstone/Hirsch/Morley 1992: 21). Die Medientechnologien werden dabei mit einer funktionalen Perspektive betrachtet:

Technologies are functional. They may be bought with other features in mind and indeed serve other cultural purposes in appropriation. They may indeed become functional in ways somewhat removed from the intentions of designers or marketers. They may have functions. Functions may change or disappear altogether [...]. To become functional a technology has to find a place within the moral economy of the household, specifically in terms of its incorporation into the routines of daily life. (Silverstone/Hirsch/Morley 1992: 21)

⁶¹ Siehe Kap. 2.2.4.

⁶² Siehe Kap. 2.1.1 zu Praxis und Ko-Operation.

Zu den Untersuchungsgegenständen zählen insofern einerseits die emergenten Routinen in der Anwendung des Geräts selbst, die durch neue sprachliche Praktiken sichtbar und verstehbar werden. Sprachliche Formen der tatsächlich beobachtbaren Kommandos an den Smart Speaker sind hier, gerade in ihrem Zusammenhang zu bestimmten Nutzungskontexten und -szenarien, von Interesse, ebenso wie die sich möglicherweise verändernde Gestaltung dieser im zeitlichen Verlauf oder vor Publikum. Analysiert werden vor diesem Hintergrund auch Reparatur-, Lern- und Anpassungsprozesse, die auch über die Annahmen der Nutzer*innen über die Geräte Aufschluss geben. Andererseits zählen dazu auch Fragen der Einbettung von Smart-Speaker-Anwendungen in laufende soziale Interaktionen und der Gesprächsorganisation, d. h. die Eingliederung in sprachlich hervorgebrachte interaktionale Ordnungen und damit verbundene Routinen, die diese herstellen. Diese können auch Bestandteil sozialer Ordnungen innerhalb des Haushalts sein, in denen z. B. ‚Nutzungshoheit‘, Besitzverhältnisse und situative Aushandlungen über die Gerätenutzung wie Porcheron et al. (2018) sie nachvollziehen, zum Ausdruck kommen.

Als Umwandlung (*conversion*) wird schließlich die Dimension der Domestizierung bezeichnet, die – in der ursprünglichen Formulierung der ‚Gründerväter‘ des Ansatzes – abhebt auf „the relationship between the household and the outside world“ (Silverstone/Hirsch/Morley 1992: 22). Wie bereits ausgeführt, ist diese Dichotomie zwischen Haushalt und „outside world“ überkommen und kann so nicht aufrecht erhalten werden, wie – als eine der stärksten Kritikerinnen dieser Konzeptualisierung – Bakardjieva (2005; 2006) ausführt. Silverstone (2006: 234) nennt in einer Konzeptionsreflexion verschiedene Aspekte: Rückanbindung an die Designer*innen, Vorführung, Darstellung und Bewertung im sozialen Umfeld, ggf. auch vor Besucher*innen, den daran anschließenden auch öffentlichen Diskurs und kulturelle Widerstände und Veränderungen. Hartmann (2013a: 23) fasst die Aspekte als „das Nach-außen-Tragen der Nutzung“ zusammen, und auch wenn erneut mit der Erwähnung eines „Außen“ ein unklares „Innen“ konstruiert wird, ist doch die Formulierung eingängig: Es sind die Momente, die als Teil des Domestizierungsprozesses auftreten, in denen sich die Haushaltsmitglieder nicht nur miteinander, in der eigenen Wohnumgebung und (vermeintlich) exklusiv mit den Medientechnologien und den dahinterstehenden Infrastrukturen befassen, sondern in denen sie diese mit anderen teilen, vorführen, bewerten, und somit zur Bildung eines gesellschaftlichen Diskurses beitragen sowie wechselseitig zur Überprüfung eigener Bewertungen gelangen. „Consumption is never a private matter“, schreibt dazu Silverstone (2006: 234) und auch wenn die Fokussierung auf den Medienkonsum abgelegt wurde, lässt sich auch über die Anwendung und alltagspraktische Nutzung von Medientechnologien sagen, dass diese keine pri-

vate Angelegenheit ist – dies gilt bei Smart Speakern noch viel mehr vor dem Hintergrund des Spannungsverhältnisses von privaten und öffentlichen Domänen.⁶³

2.3.2.2 Dreifache Artikulation: Medien als Objekte, Rituale und Inhalte

Die „doppelte Artikulation“ der Medientechnologien, der zweite vorzustellende Grundpfeiler des Domestizierungsansatzes neben den beschriebenen Dimensionen, ist seit den frühen Publikationen der britischen Domestizierungsforschung in diversen Arbeiten präsent (vgl. etwa Morley/Silverstone 1990: 36). Die doppelte Artikulation ist in Anlehnung an das sprachwissenschaftliche bzw. semiotische Verständnis von doppelter Artikulation zu begreifen: als Unterscheidung zwischen bedeutungs**tragenden** und bedeutungs**unterscheidenden** Einheiten. Mit dem von Silverstone/Hirsch/Morley (1991) hergeleiteten Verweis auf den Linguisten André Martinet und dessen Werk zur „double articulation linguistique“ (1949) wird der Begriff zugänglich: In André Martinets strukturfunktionalistischer Semiotik liegt die Bedeutung eines Wortes eben sowohl in den einzelnen Phonemen als signifikative Einheiten (erste Artikulation), wie auch in den durch sie gebildeten distinktiven Einheiten (zweite Artikulation). Dem entspricht – auch wenn diese Übertragung anfällig ist für Kritik (vgl. Hartmann 2006: 85) – das Gerät als Objekt auf der Ebene der ‚ersten Artikulation‘ und die vermittelten Inhalte als ‚zweite Artikulation‘ (vgl. Silverstone/Haddon 1996: 62).⁶⁴ Die beiden Artikulationen sind in der Semiotik Martinets zwei strukturell unterschiedliche, aber voneinander abhängige Ebenen des Sprachgebrauchs. Diese Interdependenz zwischen medientechnologischem Objekt und den vermittelten Inhalten, die zwar strukturell ver-

⁶³ Siehe Kap. 3.2.6.

⁶⁴ Bei Martinet (1963[1960]: 21–24) ist die *erste* Artikulation (bzw. *Gliederung*) die signifikative und die *zweite* die distinktive. Es überrascht, dass schon bei Morley/Silverstone (1990: 36) der ersten Artikulation, der morphematischen Ebene, die Medientechnologien als Objekt und der zweiten, phonematischen Ebene die Inhalte zugeordnet wurden. Zwar werden die Objekte als Bedeutungsträger verstanden, was ihre Zuordnung zur morphematischen Ebene plausibel macht, doch der Vergleich der Inhalte mit der semiotisch bedeutungsunterscheidenden Ebene ist vor dem Hintergrund der Arbitrarität der bedeutungsunterscheidenden Elemente und ihrer (prinzipiell beliebigen) Kombinierbarkeit nicht nachvollziehbar und legt den Schluss nahe, dass die Zuordnung vielmehr andersherum hätte erfolgen müssen: Die Objekte sind bedeutungs**unterscheidend** und bringen Inhalte auditiv oder visuell zum Ausdruck, erst die Inhalte selbst sind bedeutungs**tragend** (so auch bei Röser 2007: 22). Möglicherweise liegt der Ursprung bei Silverstone (1990: 178) selbst: Er schreibt „it [u. a. Fernsehen, T.H.] is both meaningful in itself and it is the transmitter of meanings“. Diese Passage könnte so interpretiert werden, dass die Objekte Bedeutungsträger sind, während die Bedeutung der Inhalte in diesen selbst liegt. Dieses Verständnis scheint zeichentheoretisch gesehen schief, setzt sich aber explizit in den weiteren Arbeiten zum Modell (etwa Hartmann 2006: 87) fort.

schiedene Ebenen der Mediennutzung darstellen, aber gleichzeitig nicht ohne einander zu fassen sind, stellt den Kernaspekt der Anwendung des semiotischen Konzepts im Domestisierungskontext dar (vgl. Hartmann 2006: 85–86; 2009a: 309).

Die erste Artikulation der Medien, verstanden als materielles, physisches Objekt, als gekaufte, gegenständliche Ware, ist eingängig. Sie verweist einerseits auf die bereits vorgestellte Dimension der Objektifizierung: Smart Speaker präsentieren sich als physischer, zunächst zu entpackender Artefakt mit Farbe und Form und rufen als solche ästhetische Bewertungen und Einstellungen hervor; ihre materiale Beschaffenheit wird z. B. durch die Anwendung von Touch-Technologien oder Lichtringen auch semiotisch genutzt und durch ihre Lokalisierung in der Wohnumgebung rufen sie interaktional-räumliche Orientierungen und Schemata auf. Durch die physische Präsenz des Geräts erhält es in Interaktionsordnungen ein körperäquivalentes ‚Dasein‘ und ist folglich auch körperlich ‚adressierbar‘: Ressourcen von Blick, Mimik, Gestik und Körperhaltung könnten von Interaktionsbeteiligten etwa genutzt werden, um das eigene Sprechen als ‚an einen Smart Speaker gerichtet‘ verständlich zu machen.

Demgegenüber ist die Interpretation dessen, was unter der ‚zweiten Artikulation‘ zu verstehen ist, schwankend und bis in die heutigen Arbeiten hinein unstet. Wie oben schon zum Zwecke der Veranschaulichung geschehen, bietet es sich auf den ersten Blick an, den ‚Objekten‘ kontrastiv die ‚Inhalte‘ gegenüberzustellen, die sie als Bedeutungsträger vermitteln; diese naheliegende Lesart ist auch in etlichen Studien zur Anwendung gekommen (vgl. etwa Röser 2007). Dies lässt sich auch mit den Ausführungen von Silverstone (1994: 123) begründen, der die Annahme vorschlägt, dass „the cultural value of such a machine as a television lies both in its meaning as an object [...] and in its content [...]“, sodass der Begriff des „Inhalts“ bereits gesetzt zu sein scheint. Zuvor geht Silverstone (1994: 123) jedoch auch auf eine andere Basis für die ‚zweite Artikulation‘ ein: „Television is a medium and its communications – its programmes, narratives, rhetorics and genres provide the basis for its second articulation“. So sind verschiedentlich auch nicht direkt die vermittelten Inhalte, sondern vielmehr die *Inhaltsgestaltung*, die ‚Software‘ oder die vollzogenen Narrative, die Sendepläne oder die „Programm-Narration“ (Hartmann 2013a: 26) herangezogen worden (vgl. auch Hartmann 2006: 87) – ausgelöst sicherlich auch durch eine Formulierung bei Silverstone/Haddon (1996: 74), in der die doppelte Artikulation – anders als bei Silverstone (1994) – explizit nicht als ‚Inhalt‘ („specific ‚programme‘ messages“), sondern als „symbolic environment to which the technology gives access“ verstanden wird – und, so muss konstatiert werden, in den folgenden Domestizierungsstudien und -bezügen, immer wieder unterschiedlich und teilweise diffus rezipiert wird.

Eine mögliche Interpretation ist, dass Silverstone und Haddon eine analytische Trennlinie zwischen der Objekt- und der Inhaltsseite aufstellen, ohne damit

die Thesen etwa Marshall McLuhans zu unterlaufen und eine singularistische Perspektive auf einzelne Inhalte bzw. Texte zu postulieren, die ohne Ansehen ihrer medialen Vermittlung interpretiert werden. Darüber hinaus scheint auch weder in den Arbeiten der frühen britischen Vertreter noch in den Weiterentwicklungen und Anwendungen des Domestizierungsansatzes ein starkes Interesse an einer Auseinandersetzung mit den einzelnen Inhalten in ihrer jeweiligen konkreten Ausprägung (einer konkreten Sendung oder Anwendung und deren visuelle, auditive Ressourcen) zu bestehen, sondern in der Tendenz werden unter der zweiten Artikulation vielmehr die Medientechnologien neben ihrer Eigenschaft als Objekt als *Potenzial* verstanden, überhaupt Inhalte zu vermitteln. Der Fokus liegt darauf, welcher Art diese Inhalte sein können, wie sie zwischen domestischer Sphäre und Außenwelt ‚vermitteln‘ können, wie sie programmatisch, narrativ, rhetorisch gestaltet, aufbereitet sowie durch die Objekte geprägt sind und im Haushalt rezipiert werden. Diese Sichtweise bezieht sich durchaus auf die Inhaltsseite der Medientechnologien, betrachtet aber weniger einzelne Kommunikate, sondern vielmehr den Charakter der potenziell vermittelbaren Inhalte und ihren historischen, gesellschaftlichen Kontext.

Der Vorschlag von Hartmann (2006: 96–97; 2013a: 117–123), die doppelte Artikulation auf eine dreifache Artikulation auszuweiten, in der Medien als Objekte, Inhalte und Kontexte bzw. Rituale verstanden werden, kann als Versuch gelesen werden, diese Undeutlichkeiten aufzuheben. Wie Hartmann allerdings selbst bemerkt, bleiben Unklarheiten in Bezug auf die Begriffe ‚Inhalt‘ und ‚Kontext‘: Unter Inhalten können weiterhin sowohl eine engere Auslegung (im Sinne einzelner Kommunikate) als auch eine weitere Auslegung als inhaltlicher Kontext (oder beides) verstanden werden; ferner ist der Begriff ‚Kontext‘ stellenweise unklar, da er sowohl für den konkreten sozialen, alltagspraktischen Kontext der Medienutzung als auch für den weiteren Kontext im Sinne eines gesellschaftlichen Bezugs der Inhalte und deren Formate zu gebrauchen wäre (siehe auch Livingstone 2007). Auf Hartmanns Arbeiten aufbauend legen allerdings Courtois/Verdegem/Marez (2013) eine sehr brauchbare Heuristik vor: Sie unterscheiden zwischen Medien als Objekten, als Texten und als „immediate sociospatial context“ (Courtois/Verdegem/Marez 2013: 424). Am Rande dieser Triangulation platzieren sie den ‚Kontext‘ im gesellschaftlichen Sinne, der zwar für alle drei ‚Artikulationen‘ der Medien eine Rolle spielt, aber nicht selbst eine darstellen soll (vgl. Courtois/Verdegem/Marez 2013: 425).

Mit dieser ‚dreifachen Artikulation‘ ausgestattet entsteht eine größere heuristische Klarheit, die auch Smart Speaker der Analyse in diesem Sinne zugänglich macht: Neben der bereits angerissenen Ebene von Smart Speakern als ‚Objekte‘ (s. o.) können so die von Smart Speakern in die Interaktion eingebrachten ‚Inhalte‘ umrissen werden – darunter fallen dann z. B. Nachrichten, Musikstreaming

oder interaktive Angebote.⁶⁵ Dazu in Beziehung zu setzen sind drittens der Kontext und die wechselseitige und ritualisierte Herstellung dessen – einerseits etwa durch die Anwendung oder Nicht-Anwendung von Begrüßungsformeln, das Betreiben von Face-Work (Goffman 1955; Brown/Levinson 1987) im Dialog mit dem Smart Speaker, andererseits durch die Einbettung des Smart Speakers in spezifische soziale Situationen und nicht zuletzt durch die abgerufenen Inhalte. Der gesellschaftliche Kontext (z. B. zu Fragen der Datenverwertung, des Privatheits- und Datenschutzdiskurses ebenso wie zu Genderfragen oder zur Verteilung von Arbeit und Lohn)⁶⁶ kann in diesem Rahmen analytisch abgegrenzt werden und wird nicht explizit beachtet, ist aber *in situ* nicht von den tatsächlichen Alltagspraktiken zu trennen.

2.3.2.3 Moralische Ökonomie des Haushalts

Einen dritten Stützpfeiler des Domestizierungsmodells bildet die „moral economy of the household“. Diesen Teil des Konzepts bauen Silverstone/Hirsch/Morley (1992), wie auch die Fokussierung auf den Haushalt, auf anthropologischen und historischen Studien auf, die den Haushalt als nicht nur als wirtschaftlich zusammengehörige Wohneinheit mit geteilten Mahlzeiten betrachten, sondern die Haushalte auch als „part of a transactional system of economic and social relations within the formal or more objective economy and society of the public sphere“ (Silverstone/Hirsch/Morley 1992: 14) sehen. Haushalte sind also eine Einheit innerhalb eines privaten, informellen, transaktionalen ‚Subsystems‘ der Bewohner*innen – ein *moralisches* Austauschsystem, das auch deswegen ‚Ökonomie‘ genannt wurde, weil so die Relation zur öffentlichen Ökonomie hervorgehoben wird (vgl. Hartmann 2009a: 309).

Ausgetauscht und ‚verhandelt‘ werden moralische Einheiten: Wertvorstellungen, Einsichten, Haltungen, Geschmäcker und Bewertungen über die den Haushalt umgebende Umwelt (vgl. Silverstone/Hirsch 1992: 16). Diese bilden über ein gemeinsames Erleben und Erfahren der Welt und daraus entstehende Narrative eine Häuslichkeit – einen stabilen ‚moralischen Kreislauf‘, der einerseits dynamisch ist und durch neue Erfahrungen und Erlebnisse in Schwung gehalten wird und über die Zeit auch Veränderungen an Haltungen und Wertvorstellungen hervorrufen kann. Andererseits ist die moralische Ökonomie zugleich Basis für die stabile und veränderungsresistente ontologische Sicherheit im Sinne Giddens‘ (2003[1984]: 50), die bereits Gegenstand war (s. o.) und die – Haushalt und Häuslichkeit zusammengefasst – ein ‚zu Hause‘ ausbilden (vgl. Bakardjieva 2005: 69–72).

⁶⁵ Siehe Kap. 3.2.3.

⁶⁶ Siehe Kap. 3.2.4 bis 3.2.6.

Die moralische Ökonomie bildet eine Integrität, die den Haushalt gegen neue Einflüsse verteidigen oder diese inkorporieren kann und ihn in Bezug zur Öffentlichkeit als Einheit entstehen lässt. Wird dieses häufig als sensibel beschriebenes Gefüge um noch nicht domestizierte Medientechnologien ergänzt – d. h. neue Verbindungen zur ‚Außenwelt‘ geschaffen, mit evtl. unbekannten Formaten und Inhalten – muss die moralische Ökonomie des Haushalts diese in ihren Kreislauf eingliedern, um die Stabilität aufrecht zu erhalten und ihrer Neigung zur Konstanz und zur konservativen Behandlung von Veränderungen beizukommen. Dies geschieht im Rahmen der Domestizierungspraktiken, mit denen die Haushaltsmitglieder die Medien durch Eingliederungs- und Anpassungsleistungen ‚domestizieren‘.

Die ‚Einheit‘ des Haushalts ist in diesem Modell auch als öffentliche, wirtschaftliche Einheit zu begreifen. Die frühen Domestizierungsstudien arbeiteten sich viel am Begriff des ‚Konsums‘ von Medien ab. Sie reihten sich damit ein in eine seit den 1980er-Jahren entstandene Strömung der Medienforschung, die Konsument*innen von Medien nicht als passive, schwammartig aufsaugende ‚Verbraucher‘ von Medien verstanden wissen wollte, sondern diesen eine eigene Handlungsfähigkeit und Autonomie zuschrieb.⁶⁷ Diese Rolle der Haushalte bleibt auch in der späteren Domestizierungsforschung erhalten – „Households, through their consumptions of goods and services, become actively engaged with the formal economy“ (Ward 2006: 148) –, wird allerdings im weiteren Verlauf der Entwicklung auch expandiert und von der Engführung auf die ökonomische Perspektive gelöst. Es wird dann nicht mehr nur der ökonomische Bezug zwischen Haushalt und Umwelt betrachtet, sondern generell die Zugehörigkeit des individuellen Haushalts zu öffentlichen Diskursen, an denen er als Einheit und mit seinen einzelnen Bewohner*innen beteiligt ist. Diese Expansion wird schon von Silverstone (2006: 238) in seiner Konzeptreflexion angelegt:

In one sense the notion of the moral economy is naively empirical. It asks the questions in what ways, if at all, households or families create for themselves private and personal cultures, which have consequences for the way in which the anonymous, homogenizing technologies and services of public and commercial life, are used and valued. (Silverstone 2006: 238)

Noch wichtiger ist aber für die analytische Arbeit der darauf folgende Teil, mit dem deutlich wird, dass der formulierte empirische Anspruch auch tatsächlich in der Untersuchung von Praxis einzulösen ist:

⁶⁷ Siehe auch Kap. 2.3.1.

[...] from a lateral perspective, the question arises too of how we can relate an understanding of patterns of information and communication technology use, resistance, participation and the rest to what we can understand. (Silverstone 2006: 238)

Es sind genau diese Muster der Nutzung, der Widerstände und Beteiligungsstrukturen (und mögliche andere Kategorien, die bereits angesprochen waren), mit Hilfe derer Aufschluss über Wertvorstellungen, entstehende Narrative und Haltungen gegeben werden kann. Deren Konstitution und Weiterentwicklung vollziehen sich in der alltäglichen Nutzung von Medientechnologien: Relevante Geschichten werden geteilt und zu Narrationen aufgebaut, Evaluationen ausgedrückt und verhandelt, Wertvorstellungen implizit und explizit in Praxis vollzogen und geäußert.

Smart Speaker können für die moralische Ökonomie des Haushalts und für dessen ontologische Sicherheit in besonderem Maße eine ‚Bedrohung‘ sein. Dazu trägt zum einen die Verwendung interaktionsähnlicher, gesprochener Sprache als Interface bei. Zum anderen entsteht durch eine teilweise unklare Datenverarbeitung der eigenen Stimme, durch die Anknüpfung an kommerzielle Dienste und Online-Shopping sowie durch die konstante Aufzeichnung der Wohnumgebung ein virulenter Datenschutz-Diskurs um Smart Speaker und andere Smart-Home-Anwendungen (siehe etwa Pridmore/Mols 2020; Mols 2021; Waldecker 2022; Hector et al. 2023).⁶⁸ In diesem positionieren sich die Anwender*innen von Smart Speakern und bauen mittels ihrer Nutzungspraxis und durch ihre eigenen praktischen Erfahrungen und Erlebnisse mit den Geräten sowie den Austausch mit anderen Haushaltsmitgliedern, eigene Bewertungen auf, die Teil von Wertvorstellungen werden können und in die moralische Ökonomie des Haushalts einfließen. Dieser ist (auf die eine oder andere Weise, selbst *ex negativo*) Bestandteil des öffentlichen Diskurses. Insofern sind Evaluationen der Nutzungspraxis mit Bezug auf den Datenschutz, Positionierungen der Haushaltsmitglieder aber auch – gerade im Hinblick auf die Domestizierungsdimension der ‚Umwandlung‘ – der Austausch mit Besucher*innen und die dabei entstehende Entwicklung eigener Positionen, für die Beschreibung der moralischen Ökonomie des Haushalts von besonderer Relevanz.

2.3.2.4 Kritik des Domestizierungsansatzes

Der Domestizierungsansatz ist unter verschiedenen Gesichtspunkten kritisiert worden,⁶⁹ denen teilweise mit der spezifischen Methodologie der vorliegenden Arbeit, teilweise mit im Gegenstandsbereich der Smart Speaker selbst liegenden Argumenten begegnet werden soll. So betonen Arbeiten, in denen die Dimensio-

⁶⁸ Siehe Kap. 3.2.6.

⁶⁹ Für eine Übersicht siehe Hartmann (2013a: 101–117).

nen des Domestizierungsansatzes überwiegend selbst angewendet werden, dass diese nicht linear zu begreifen sind und arbeiten sich an einer vermeintlichen Linearität von Kommodifizierung, Aneignung und Umwandlung ab; gleichzeitig verarbeitet eine Reihe von Arbeiten diese Dimensionen in der Tat zu einer linearen Abfolge. Die Kritik muss sich hier auf die Arbeiten beziehen, die von einer temporalen Abfolge der Dimensionen ausgehen. Als Umgang damit schlägt Ling (2004: 32) vor, den Ansatz immer wieder in der Empirie zu überprüfen und festzustellen, ob spezifisch temporale Muster auffällig sind. Ähnliches gilt für die immer wieder diskutierte Frage nach der Handlungsmacht der Nutzer*innen: Mit dem Begriff der ‚Domestizierung‘ scheint sich – so die Kritik – die Medienaneignung auf einen Prozess des Zähmens zu beschränken, der die reziproken Mechanismen dieses Vorgangs vernachlässigen könnte und die Distribution von Agency aus dem Blick verliert.⁷⁰ Dabei sind sowohl eine Überbetonung der Agency der Medientechnologien auf der einen (siehe etwa Feenberg 1999: 107; Helle-Valle/Slettemeås 2008) wie auch eine Überbetonung der Agency der Nutzer*innen auf der anderen Seite (vgl. Bassett 2007: 65–72) kritisiert worden. Für die vorliegende Arbeit bedeutet dies: Erstens muss auch dieser Kritikgegenstand Frage einer empirischen Überprüfung bleiben: Relevant sind die Strategien, mit denen Nutzer*innen *in situ* Handlungsträgerschaft aushandeln und dabei die beteiligten Geräte einbeziehen (vgl. Habscheid/Hector/Hrncal 2023). Zweitens wird mit dem bereits eingeführten praxeologischen Konzept der „materiellen Partizipanden des Tuns“ von Hirschauer (2004; 2016)⁷¹ eine praxeologische Perspektive eingenommen, die eine empirische Betrachtung dieser Fragestellung frei von abstrakten Festlegungen ermöglicht.

Bassett (2007: 63–67) ist es auch, die die bereits mehrfach formulierte Kritik, dass die soziale Praxis der Domestizierung nicht hinreichend Ausgangspunkt der Domestizierungsforschung ist, zugespitzt formuliert und die These aufstellt, dass die Betrachtung der *Routinen* des Alltags die Betrachtung der sozialen Praxis verdränge. Diese Kritik muss in zweierlei Hinsicht zurückgewiesen werden: Einerseits, weil Silverstone (2006: 231) kurz zuvor richtigerweise betont: „Domestication is practice“. Die Praxis der Domestizierung von Medientechnologien nachzuvollziehen und verstehbar zu machen, ist – jedenfalls in der vorliegenden Arbeit – gerade Gegenstand der Untersuchung. Zweitens sind, insbesondere in einem praxeologischen Verständnis, wie es hier skizziert wurde, insbesondere die Routinen des Alltags soziale Praxis (vgl. Schüttpelz/Meyer 2017: 158). Der Alltagsbegriff war ebenfalls

⁷⁰ Waldecker/Hector/Hoffmann (2024) sowie Habscheid/Hector/Hrncal (2023) diskutieren die Agency der Nutzer*innen von Smart Speakern vor dem Hintergrund der Domestizierungsprozesse.

⁷¹ Siehe dazu Kap. 2.1 bzw. 2.2.

Gegenstand vereinzelter Kritik: Die Aneignung neuartiger Medientechnologien und deren Einbettung falle eben *nicht* unter die Alltagspraxis (vgl. Ling 2004: 32) und im Speziellen nicht die erste Inbetriebnahme. Es kann an dieser Stelle nicht ausführlich auf Theorien des Alltags eingegangen werden, doch sprechen erstens die Argumente einer Phänomenologie des Alltags wie sie Schütz/Luckmann (2003) skizzieren durchaus dafür, dass auch solche Situationen zum „Alltagsleben“ zählen (‘alltätlich’ ist in diesem Sinne nicht mit ‘gewöhnlich’ gleichzusetzen) und zweitens wird abseits der begrifflichen Fragestellungen in der vorliegenden Untersuchung durch eine zweite Erhebungsphase mehrere Monate nach der ersten Inbetriebnahme der Smart Speaker sichergestellt, dass auch eine routiniertere und ‘gewöhnlichere’ Nutzung, die sich in übrige Alltagsabläufe der Haushalte einbettet, in den Blick genommen werden kann.⁷²

Der letzte Kritikpunkt, auf den hier (noch einmal zusammenhängend) eingegangen werden soll, betrifft die Frage nach dem *Wo* der Alltagsroutinen. Es wurde zurecht eingewendet, dass der Alltag keineswegs nur in der Wohnumgebung der Nutzer*innen stattfindet, sondern auch außerhalb. In diesem Zusammenhang wurde das Modell – wie bereits kurz erwähnt – auch für eine Dichotomie von „innen“ und „außen“ kritisiert; dieser Kritik ist, wie ebenfalls bereits ausgeführt, schon konstruktiv begegnet worden (vgl. Bakardjieva 2006; Hartmann 2013b; 2019). Darüber hinaus trifft sie im Falle der Anwendung auf den Gegenstandsbereich der Smart Speaker nur bedingt zu: Da diese Geräte für die stationäre Nutzung konzipiert sind und sich insbesondere die hier im Fokus stehende verbale Steuerung auch nahezu ausschließlich in der Wohnumgebung vollzieht, ist eine Domestizierungspraxis von Smart Speakern außerhalb dessen nur schwer vorstellbar – diese müsste mit der Außerhaussteuerung etwa von Smart-Home-Geräten verbunden sein, die dann aber in eine größere Medienökologie eingebettet wären. Das soll noch einmal nicht heißen, dass nicht gerade Smart Speaker durch ihre Funktionsweise zur Auflösung von Dichotomien von Öffentlichkeit und Privatheit beitragen – im Gegenteil (vgl. Waldecker 2022). Wie diese Auflösungsprozesse interaktiv verhandelt werden und im Prozess der Domestizierung – gerade in der Dimension der Umwandlung und in der Beschaffenheit der moralischen Ökonomie des Haushalts – von den Beteiligten relevant gesetzt werden, wäre aber im Rahmen empirischer Analysen zu zeigen. Auch sind Steuerungen von Smart-Home-Anwendungen über mobile Apps möglich, die auf eine verteilte Infrastruktur zurückgreifen, in die auch der Smart Speaker eingebunden ist (vgl. Goulden 2019; Strüver 2023a). Doch die Nutzungspraxis mit Smart Speakern, deren Domestizierung hier im Vordergrund steht, sind an die Wohnumgebung ge-

⁷² Siehe Kap. 5.

koppelt, sodass im Rahmen der Datenerhebung eine Fokussierung auf diese sinnvoll scheint. Inwiefern Außerhäusliches dann für den Domestizierungsprozess relevant wird, ist als empirisch zu beantwortende Frage mit in die Analysen zu tragen. Dieser vollzieht sich in den sprachlichen Praktiken der Nutzer*innen, mit denen Smart Speaker auf der einen Seite genutzt und zugleich in die kommunikativen Konfigurationen der Haushalte eingebettet werden.

2.3.3 Zwischenfazit Aneignung und Domestizierung

Wendet man die dargestellten, bisherigen Untersuchungen zur kommunikativen Aneignung und Domestizierung auf Smart Speaker an, tritt die sprachliche Dimension der Domestizierungs- bzw. Aneignungsprozesse auch aufgrund der Akustik als spezifisches Charakteristikum in der Bedienung dieser Interfaces besonders hervor (vgl. Merkle/Hector 2025; Habscheid/Hector/Hrncal 2025). Der Domestizierungsansatz wird dabei mit den Arbeiten zur sprachlichen Aneignung von Medientechnologien kombiniert und erweitert: Erstens vollzieht sich der in den Dimensionen und Konzepten abgebildete Domestizierungsprozess in den sozialen Praktiken der Haushaltsmitglieder und ist insofern an der sprachlichen Oberfläche beschreib- und verstehbar. Die sprachlichen Praktiken der Haushaltsmitglieder ziehen sich durch alle Dimensionen des Domestizierungsansatzes. Sie sind der Treibstoff, mit dem sie die Geräte in ihre interaktiven Ordnungen einbinden, sich zu ihnen positionieren und mit denen sie innerhalb der moralischen Ökonomie des Haushalts Smart Speaker als Teil des Wertesystems verhandeln und über den sie überhaupt zu entsprechenden Einsichten und individuellen Werturteilen gelangen. Wie immer wieder betont wird, nimmt der Domestizierungsansatz schon von seiner Verwurzelung her nicht nur die Praktiken innerhalb der häuslichen Umgebung selbst in den Blick, sondern ist auch an Prozessen an der Schnittstelle von Öffentlichkeit und Privatheit interessiert. Das macht ihn für die Analyse von Smart Speakern besonders interessant, da diese Geräte Teil einer fortlaufenden, diskursiven Aushandlung von Grenzen zwischen den beiden Domänen sind: Schon durch ihre Funktionsweise und ihre Datenverarbeitung und -verwertung machen sie Aspekte des Privaten zu Aspekten von (teils opaken) Öffentlichkeiten und zu Teilen eines sich radikal verändernden wirtschaftlichen Systems. Das ubiquitäre Konzept des „Überwachungskapitalismus“ (Zuboff 2018)⁷³ wird auf diese Weise einer empirischen Überprüfung auf Seiten der Anwender*innen unter-

⁷³ Siehe Kap. 3.2.6.

zogen. Durch die empirischen Methoden der angewandten Sprachwissenschaft⁷⁴ wird das Domestizierungsmodell um eine Analyseebene erweitert, in der die von den Nutzer*innen selbst relevant gesetzten Aspekte in den Mittelpunkt rücken.

Damit wird zweitens der Versuch unternommen, das Modell methodologisch auszubauen: In den bisherigen Untersuchungen zur Domestizierung standen v. a. Methoden der qualitativen Sozialforschung und insbesondere der Ethnografie im Mittelpunkt. Davon wird durch die Anwendung einer ethnomethodologisch-praxeologischen Gesprächsanalyse, wie sie in den folgenden Analysen vorzunehmen ist, zwar nicht radikal abgewichen; allerdings sind Methoden im Kernbereich der Gesprächsanalyse bisher v. a. im Rahmen der sprachwissenschaftlichen Aneignungsforschung angewendet worden und nicht so sehr in einer integrierten Betrachtungsweise mit dem in der Kommunikations- und Medienwissenschaft beheimateten Domestizierungsansatz zum Einsatz gekommen. Der darin liegenden Spannung, dass die klassischerweise verfolgten Erkenntnisinteressen des Domestizierungsansatzes nicht unbedingt deckungsgleich mit denen der empirischen Linguistik zu sein scheinen, wird in zweierlei Hinsicht begegnet: Erstens hat es in der angewandten Sprachwissenschaft durchaus bereits Tradition, linguistische Methoden und Analysen in den Dienst der Beantwortung von Forschungsfragen aus anderen Teildisziplinen zu stellen (vgl. etwa Meer/Pick 2019). Treibend für die Analyse ist nicht das Interesse an den linguistischen Strukturen selbst, sondern das Augenmerk liegt auf den Zusammenhängen von sprachlichen Praktiken und kommunikativer Funktion. Zweitens profitiert von dieser Verbindung insbesondere eine Medienlinguistik, die sich zur Aufgabe macht, den Zusammenhang von Medien und sprachlichen Formen aufzudecken. Gerade technologische Artefakte wie Smart Speaker, die konstitutiv einen sprachlichen Dialog mit ihnen evozieren, bedürfen einer Form der Medienrezeptionsforschung, die deren dialogische Anatomie in den Mittelpunkt stellt und die Medienaneignung entlang der sprachlichen Praktiken beleuchtet.

⁷⁴ Siehe dazu Kap. 4 und 5.

3 Forschungsüberblick und Gegenstandsbestimmung

Das folgende Kapitel soll den Forschungsstand zu Mensch-Maschine-Interaktionen in der Linguistik im Überblick zusammenfassen sowie im Speziellen und detaillierter den interdisziplinären Forschungsstand zu Smart Speakern unter verschiedenen Gesichtspunkten darstellen. Das Kapitel liefert daher zunächst einen phänomenorientierten Überblick über Forschung, die sich primär aus sprachwissenschaftlicher Sicht mit dem Austausch zwischen Mensch und Maschine beschäftigt hat (Kap. 3.1). Dabei werden vorwiegend nicht Smart Speaker, sondern andere Dialogsysteme wie Chatbots, Embodied Conversational Agents, humanoide Roboter und telefonbasierte Systeme beleuchtet. Daran anschließend wird der interdisziplinäre Forschungsstand zu Smart Speakern unter verschiedenen Aspekten zusammengefasst. Es wird dazu zunächst knapp die historische Entwicklung und die technische Verarbeitung und Prozessierung der heutigen Systeme nachgezeichnet (Kap. 3.2.1), ehe auf die Funktionsweise und die technischen Hintergründe selbst eingegangen wird (Kap. 3.2.2). Eine detaillierte informatische Beschreibung ist dabei entbehrlich und auch gar nicht im Rahmen dieser Arbeit zu leisten. Wesentlich ist aber ein überblicksartiges Verständnis der ablaufenden Prozesse. Anschließend an die technische Beschreibung werden kommunikations- und medienwissenschaftliche Arbeiten vorgestellt, in denen Smart Speaker ausgehend von ihren Funktionen als Intermediäre und als Plattformen konzeptualisiert werden (Kap. 3.2.3). Dieses Kapitel liefert insofern auch einen Überblick über den Funktionsumfang der Geräte, auch wenn die vorliegende Arbeit auf deren spezifische Interfaces fokussiert. Darauf folgend werden Studien zusammengefasst, die bereits zur Nutzung und Aneignung von Smart Speakern aus unterschiedlichen disziplinären Perspektiven und mit einer noch überschaubaren Bandbreite an verschiedenen Methoden geforscht haben, darunter zentral die schwerpunktmäßig linguistisch arbeitenden Untersuchungen (Kap. 3.2.4). Ein besonderes Augenmerk soll im Anschluss auf solche Arbeiten gelegt werden, die Smart Speaker unter Gender-Gesichtspunkten beleuchtet haben (Kap. 3.2.5). Dabei wird insbesondere auf die Wahl der Stimmfarbe und die ausgehandelte Rollenverteilung zwischen dem Smart Speaker und den Nutzer*innen abgehoben. Abschließend sollen die im Fokus der öffentlichen Aufmerksamkeit stehenden Themen Datenschutz, Datenverwertung und Privatsphäre anhand der dazu erschienenen einschlägigen Studien betrachtet werden (Kap. 3.2.6). Die zum Zeitpunkt der Studie gängigen und genutzten Smart Speaker werden in Kap. 3.3 genauer beschrieben, um einen Einblick in den Aufbau und die Charakteristika der einzelnen Systeme von Amazon, Google und Apple zu bekommen.

3.1 Mensch und Maschine in der empirischen Linguistik

Der Austausch zwischen Mensch und Maschine ist von der Sprachwissenschaft in mindestens zweierlei Hinsicht betrachtet worden: Einerseits ist die Subdisziplin der Computerlinguistik selbst wesentlicher Bestandteil der vorwiegend von der Informatik geprägten Entwicklung von Interfaces, schriftlich oder mündlich: Die Hervorbringung von *Speech Processing* und *Natural Language Processing* ist nicht vorstellbar, ohne eine Beteiligung der Linguistik. Andererseits, und dies soll Fokus der folgenden Ausführungen sein, auch wenn die Unterscheidung nicht bei allen Arbeiten trennscharf vollzogen werden kann, hat die Sprachwissenschaft empirisch den Austausch selbst untersucht, ohne dabei primär die technologische Weiterentwicklung als Ziel der Forschung zu positionieren. Phänomenbezogen besteht dabei eine ‚Verwandtschaft‘ von Smart Speakern, die selbst bislang nur begrenzt linguistisch untersucht wurden (siehe Kap. 3.2.4), mit anderen Dialogsystemen, insbesondere Chatbots, Embodied Conversational Agents und humanoiden Robotern (vgl. Lotze 2018: 29) sowie mit telefonbasierten Sprachdialogsystemen (vgl. Thar 2015). Die nachfolgenden Ausführungen sind ein Überblick über Erkenntnisse aus der explizit interaktionslinguistischen Untersuchung des Dialogs mit Maschinen, bevor im darauf folgenden Kapitel der interdisziplinäre Forschungsstand zu Smart Speakern einschließlich einer Übersicht über die Funktionsweise und den technischen Hintergrund präsentiert wird.

3.1.1 Chatbots

Als eine „neue Form der Dialogizität“ beschreibt die Sprachwissenschaftlerin Netaya Lotze den Dialog mit Chatbots und mit Embodied Conversational Agents (Lotze 2020: 363). Letztere basieren auf Chatbot-Systemen, sind aber umfangreich erweitert, insbesondere durch multimodale, personalisierte Avatare, über die auch eine basale Form von Emotion ausgedrückt werden kann (vgl. Lotze 2016: 39).⁷⁵ Aus einer linguistischen Perspektive setzt sich Lotze (2016) in einer umfangreichen Arbeit mit Chatbots⁷⁶ auseinander und zeigt vier methodische Zugänge zu diesen auf, die in unterschiedlichen Teilgebieten der Linguistik verwurzelt sind: Neben dem Zugriff, den die Konversationsanalyse ermöglicht, arbeitet sie mit dem Modell des „interaktiven Alignment“ aus der Psycholinguistik, dem Konzept des „Computer-Talk“ (Zoeppritsch 1985; Krause/Hitzenberger 1992; Fischer

⁷⁵ Siehe auch Krummheuer (2010: 81) und Kap. 3.1.2.

⁷⁶ Zu *Social Bots*, automatisierten Accounts auf Social Media-Plattformen, siehe Kaerlein (2018).

2006),⁷⁷ das an der Schnittstelle zwischen Linguistik und Informationswissenschaft angesiedelt ist, sowie dem eher ‚kernlinguistischen‘ Konzept der Dialogkohärenz. Lotze (2016: 90–91) unterscheidet weitere nicht-linguistische Ansätze, die relevant sind (insbesondere informatische Forschungen zu Design und Usability und soziologische bzw. psychologische Ansätze zur Untersuchung sozialer Folgen und Einflussfaktoren). Sie betont, dass die Analyse von Mensch-Maschine-Interaktion immer eine auf mehreren Ebenen interdisziplinäre Unternehmung ist (vgl. Lotze 2016: 86).

Lotze entwickelt in ihrer Arbeit einen Untersuchungsrahmen, der auf den Erkenntnissen der Konversationsanalyse basiert und als Grundlage für die weiteren Analysen dient: Wesentliche Parameter sind dabei die Dialoggliederung, die Bestimmung von Rahmen- und Adjazenzstrukturen, die Rolle von Störungen und Reparaturen sowie der Einfluss der Form der technischen Vermittlung (vgl. Lotze 2016: 106). Diesen Rahmen wendet sie in konversationsanalytischen Auswertungen, Analysen zur Dialogkohärenz und bei der Betrachtung mit der Perspektive des Computer Talk an. Allein ihre Erkenntnisse aus der konversationsanalytischen Betrachtung der Mensch-Maschine-Dialoge sind sehr reichhaltig. So stellt sie fest, dass Turns des Chat-Systems länger sind als die Nutzer*innen-Turns und auch mehr TCUs beinhalten (vgl. Lotze 2016: 234). Dieser Befund korrespondiert mit den Erkenntnissen von Cyra/Pitsch (2017a; 2017b), die anhand eines Embodied Conversational Agents (ECA)⁷⁸ die inkrementelle Erweiterung von Nutzer*innen-Turns beleuchten und dabei von der Feststellung ausgehen, dass längere und inkrementell expandierte Turns für die Dialogsysteme eine Herausforderung darstellen.⁷⁹ Für einen gelingenden Dialog sind also kürzere Beiträge vorteilhaft. Die konventionalisierte Rahmung der von Lotze untersuchten Chatbot-Dialoge, z. B. durch Begrüßung und Verabschiedung, wird eher von den Bots eingehalten, weniger aber durch die Nutzer*innen (vgl. Lotze 2016: 239–240; siehe auch Wuenderlich/Paluch 2017). Nutzer*innen hingegen arbeiten häufig mit Adjazenzellipsen, die einen semantischen Anschluss innerhalb des Dialogs herstellen (sollen); diese jedoch erkennen die analysierten Chatbots kaum und verwenden sie auch nicht selbst (vgl. Lotze 2016: 242). Im Vergleich mit anderen Strategien aus der computervermittelten schriftlichen Kommunikation wie der Verwendung von Emoticons und Emojis, Buchstabeniteration, kreativer Satzzeichengebrauch oder Reduktionsformen (siehe etwa Androutsopoulos/Busch 2020) stellt sie fest, dass diese in ihren Daten kaum vorhanden sind, und sieht eine Ursache dafür in der Standardnähe der Beiträge

⁷⁷ Ausführlicher dazu siehe Kap. 6.1.2.1.

⁷⁸ Siehe dazu Kap. 3.1.2.

⁷⁹ Die Studien fokussieren sich in einem „Wizard-of-Oz“-Szenario auf die Entstehung und Klassifikation der Expansionen sowie darauf aufbauende Vorschläge für Design-Lösungen.

des Chat-Systems, sodass „im Zuge dynamischer Anpassungsprozesse“ auch die User*innen solche sprachlichen Formen selten übertrügen (vgl. Lotze 2016: 251–252). Lotze (2016: 247) zeigt aber auch, dass die Orientierung an sprachlichem Material, das (den Auffassungen der Nutzer*innen nach) leicht von den Chat-Systemen verarbeitet (geparsed) werden kann, eine wesentliche Reparaturstrategie ist; es könnte entsprechend auch eine Strategie sein, Ausdrücke aus der zwischenmenschlichen Chat-Kommunikation in diesem Zusammenhang zu vermeiden – auch, weil der Konventionalisierungsgrad hierbei noch deutlich geringer ist als im standardsprachlichen Bereich.

Psycholinguistisch arbeitet Lotze mit dem Konzept des „interaktiven Alignment“ (Pickering/Garrod 2004), in dem die im Dialog verwendeten sprachlichen Konstruktionen mit den kognitiven Repräsentationen in Verbindung gebracht werden – die Wiederaufnahme sprachlicher Strukturen (u. a. auf syntaktischer, lexischer, prosodischer und multimodaler Ebene) ist ein Zeichen für ein solches Alignment, d. h. für die „Angleichung“ kognitiver Repräsentationen. Lotze (2016: 257–262) weist nach, dass in Dialogen mit Chatbots weder auf lexikalischer noch auf syntaktischer Ebene ein vergleichbar häufiges Alignment stattfindet wie in Mensch-Mensch-Interaktionen. Im Dialog mit Chatbots kommt Alignment weniger häufig zustande und ist wesentlich seltener über mehrere Dialogzüge persistent. Neuere Dialogsysteme zeigen zwar höhere Alignment-Werte als ältere Systeme, doch eine Reihe von Störfaktoren und Variablen scheint das Auftreten der Alignments in HCI weiterhin niedrig zu halten (vgl. Lotze 2016: 281–282). Im Hinblick auf Kohärenz und Kohäsion kann Lotze zeigen, dass die Chat-Systeme einen deutlich niedrigeren Anteil kohärenter Anschlüsse produzieren als in verglichenen Mensch-Mensch-Interaktionen, in denen sämtlich Kohärenz hergestellt wird. Dies ist umso bemerkenswerter, als die Anzahl der von den Systemen verwendeten Kohäsionsmittel sehr hoch ist und teilweise sogar höher ist als die der Nutzer*innen (vgl. Lotze 2016: 312). Fehlender Common Ground, fehlende semantische Referenzdefinitionen und nicht ausdefinierte thematische Progression in Frames und Skripts sowie mangelhafte Verarbeitung von Konjunkturen und turnübergreifenden Konnektiva sind dafür die Hauptursache – sie liegen alle auf der Ebene „Tiefenstruktur“ der Chat-Systeme (Lotze 2016: 314–315).

Netaya Lotze zeigt mit ihrer Monografie umfangreiche Anknüpfungspunkte für die Untersuchung von Mensch-Maschine-Dialogen auf: Unterschiede, die durch die verbale Bedienweise der Benutzerschnittstelle entstehen, sind ebenso relevant wie die Frage danach, wie die sprachlichen Besonderheiten im Dialog mit Chatbots sich in die Praxis der Nutzer*innen einbetten, inwieweit sie in die sprachlichen Register übergehen oder diese beeinflussen und zu einer Bewertung der Systeme insgesamt beitragen. Lotze (2018: 31) stellt fest: „Wenn [...] ein Mensch mit einem künstlichen Gegenüber in Interaktion tritt, ist das keine Kom-

munikation unter Gleichen“. Wie in Kap. 2.2 dargestellt, soll bei einem Austausch zwischen Mensch und Maschine wie Lotze ihn beschreibt, in der vorliegenden Arbeit nicht von „Interaktion“ die Rede sein. Denn auch wenn das Design der Dialogsysteme (durch die verbale Benutzerschnittstelle bei Smart Speakern noch einmal mehr) die „Illusion eines menschenähnlichen Gegenübers [unterstützt] und [...] dadurch soziale Wirkungen [evoziert]“ (Lotze 2016: 73; siehe auch Bucher 2014), zeigt sich auch auf sprachlicher Ebene die Bearbeitung wesentlicher Differenzen zwischen Mensch und Maschine. Lotze (2018: 30) unterscheidet passend zwischen „Dialogkompetenz und -performanz“ und hat in ihren Arbeiten gezeigt, dass bei Chat-Systemen durchaus ein Grad von Dialogperformanz vorliegt, sie aber „von einer eigenen Dialogkompetenz weit entfernt“ sind (Lotze 2018: 30). Vielmehr führen aufgebaute Illusionsformen eines menschlichen Gegenübers zu übersteigerten Erwartungen und in besonderer Weise zu Brüchen im Dialog (vgl. Lotze 2018: 46).⁸⁰ Als Perspektive verweist Lotze (2018: 46) weniger auf weitere Annäherung zwischen Mensch-Mensch- und Mensch-Maschine-Dialogen, sondern vielmehr auf das wechselseitige Erlernen einer „anwendungsadäquate[n] Interaktionsform“. Zu untersuchen, ob und wie sich dieser Prozess vollzieht, ist wesentlicher Teil des Ansinnens der vorliegenden Arbeit.

3.1.2 Embodied Conversational Agents

Eine erweiterte Form von Chatbots sind Embodied Conversational Agents (ECAs). Sie haben prinzipiell nach Cassell (2000: 29) folgende Eigenschaften:

- the ability to recognize and respond to verbal and nonverbal input
- the ability to generate verbal and nonverbal output
- the ability to deal with conversational functions such as turn taking, feedback, and repair mechanisms
- the ability to give signals that indicate the state of the conversation, as well as to contribute new propositions to the discourse

⁸⁰ Es soll dabei nicht übersehen werden, dass es auf Social Media-Plattformen wie Twitter durchaus gelingt, Bots einzusetzen, die von menschlichen Accounts nur schwer zu unterscheiden sind (siehe dazu Kaerlein 2018), was bereits eine ethische Debatte über die Pflicht zur Identifikation nicht-menschlicher Benutzerkonten ausgelöst hat (siehe Graff 2016; Williams 2018). Diese werden jedoch in durch die Plattformen präfigurierten und stark konventionalisierten Umgebungen eingesetzt, in denen interaktive Vorgänge grundlegend anders gesteuert werden (Aiello et al. 2012; Howard/Woolley/Calo 2018). In diesen ist, so argumentiert Kaerlein (2018), die Ununterscheidbarkeit von Mensch und Maschine gerade eine Folge der spezifischen medialen Bedingungen, unter denen die Kommunikation stattfindet.

Eine Übersicht zu ECA-Design-Projekten findet sich bei Cassell/Sullivan/Prevost (2000) und bei Cassell et al. (2000). ECAs wurden intensiv in der bereits andiskutierten Arbeit von Krummheuer (2010) untersucht. Sie geht insbesondere der Frage nach, ob der Austausch mit ECAs als Interaktion bezeichnet werden kann. Diese Frage verneint sie und charakterisiert den Austausch als hybrid und ambig zugleich.⁸¹ Krummheuer stößt dabei auch sprachlich auf interessante Strategien, auch wenn die Arbeit sich mit einem primär soziologischen Interesse den ECAs widmet. Ihre Befunde von 2010 und 2011 stehen im Einklang mit den Erkenntnissen von Lotze: Die Orientierung auf Konventionalisierungen und rituelle Klammern des ECA ist stärker als die des Menschen (vgl. Krummheuer 2010: 187). Sie kann ebenfalls an der Schnittstelle von „Spiel, Provokation und Test“ (Krummheuer 2010: 305) liegende Formen von unangemessenem und beleidigendem Verhalten feststellen, mit dem Nutzer*innen die technische Beschaffenheit des Geräts betonen und es ‚herausfordern‘. Diese Situationen werden vom Publikum in den Daten Krummheuers als unterhaltsam, aber auch als irritierend aufgefasst und teilweise mit Lachen bzw. Kichern kommentiert. Krummheuer betont hier, aber auch z. B. mit Bezug auf Störungen und Reparaturen, dass der Beitrag der Nutzer*innen in ihren Settings immer ein interaktional und situativ ausgehandelter Beitrag ist, der unter Beobachtung der Umstehenden zustande kommt, während der ECA diese Prozesse nicht beobachten kann – der direkte Dialog ‚nur‘ mit dem ECA ist also nur eine kleine Facette des Austauschs. Der ECA seinerseits simuliert eine Form „intersubjektiven Verstehens“ (Krummheuer 2010: 261), die aber (wie bei Chatbot-Systemen) auf Basis von Inputs und Skripts erfolgt und insofern keine interaktionale Aushandlung darstellt.

Die Mündlichkeit, die bei der Steuerung zum Einsatz kommt, bringt aber auch weitere Spezifika des Austauschs mit sich. So ist z. B. die Aushandlung des Rederechts und die Gestaltung von Sprecher*innen-Wechseln ebenfalls Teil des Austauschs. Krummheuer (2010: 209) beobachtet dabei v. a. eine Anpassung der Nutzer*innen, auf die längere Pausen und eine sehr geringe Überlappungsrate hindeuten. Auch machen die Nutzer*innen die Produktionsplanung des nächsten Beitrags gegenüber anderen Anwesenden *accountable* (vgl. Krummheuer 2010: 208), was ebenfalls für eine ‚Anpassung‘ der menschlichen Akteur*innen spricht.

In neueren Arbeiten zu ECAs argumentiert Cassell (2020), allerdings weniger linguistisch fundiert und eher aus einer Design-Perspektive, dass mit neueren Systemen durchaus eine Form von sozialer Beziehung zwischen Mensch und Maschine entstehen kann. Bei der gemeinsamen Bearbeitung von Aufgaben, im Beispiel von Cassell Lesen- und Schreiben-Lernen, habe sie beobachtet, „that they

⁸¹ Siehe Kap. 2.2.3.

[Conversational Agents, T.H.] do in fact improve relations between people and the systems“ (Cassell 2020: 17) und dass sie auch wechselseitige Erfolge bei der Bearbeitung der Lese- und Schreib-Aufgaben als gemeinsam bearbeitetes Projekt erzielen konnten. Allerdings fehlt eine sequentielle Betrachtung der Dialoge in alltäglichen Umgebungen; Cassell stützt sich auf Experimente aus Settings, die primär aus dem Bereich der Design-Evaluation stammen und auch einen vortheoretischen Begriff von „social bond“ verwenden. Gleichwohl bleibt die Frage, inwieweit eine Beziehung zu einem ECA eingegangen werden kann, relevant, wie auch Studien zeigen, die nicht auf schriftliche Chats oder ECAs schauen, sondern verkörperte, menschenähnliche Robotik in den Blick nehmen.

3.1.3 Humanoide Roboter

Die Ergebnisse von Lotze und Krummheuer zu Chatbots bzw. ECAs lassen sich gut an die auf sprachliche Aspekte fokussierenden Arbeiten zu Dialogen mit humanoiden Robotern anknüpfen. Auch hier zeigt sich, dass der Dialog mit einem Roboter strukturelle Unterschiede im Vergleich mit Mensch-Mensch-Interaktionen aufweist (vgl. etwa Habscheid et al. 2020: 174; Fischer 2011; Pitsch/Gehle/Wrede 2013). Um diese herauszuarbeiten, ist eine Reihe an Studien unterschiedlicher Art mit „museum-guide-robots“ durchgeführt worden. Solche gelten als eine der „principal areas applications considered in research on human-robot interaction“ (Kuzuoka et al. 2008: 201). Diese „real-world-settings“ ermöglichen es einerseits, Erkenntnisse über die Gestaltung von Dialogen mit Robotern zu gewinnen, die aus Design-Perspektive verwertet werden können. Sie ermöglichen aber gleichsam das Gewinnen von Erkenntnissen über zwischenmenschliche Interaktion (vgl. Pitsch 2015: 29–30). In diesen Settings stoßen die Roboter auf Herausforderungen, die in Laborbedingungen nicht gegeben sind: die Anbahnung eines Dialogs mit dem Roboter (siehe dazu Scheffler/Pitsch 2020) und situationsadäquate Eröffnungen (vgl. Gehle et al. 2017) sind hier ebenso zu nennen wie mehrere Dialogbeteiligte (vgl. Pitsch/Gehle/Wrede 2013), der angemessene Umgang mit Nicht-Verstehen und Reparaturen (vgl. Gehle et al. 2015) und die durchaus bedeutungstragende Bewegung innerhalb der Ausstellungsräumlichkeiten (vgl. Pitsch/Gehle/Wrede 2013: 2). Auch die bereits erwähnte „Kontextblindheit“ stellt die Dialogsysteme vor Herausforderungen und führt zudem zu Ängsten und Sorgen auf Seiten der Nutzer*innen die auch diskursiv tradiert werden (siehe Habscheid et al. 2020). Hierzu entwickeln Teile der genannten Studien in sog. *Wizard-of-Oz*-Experimenten (WOZ-Studien) Designvorschläge für eine Weiterentwicklung der Dialogmodelle. In dieser Form der Studie wird der Roboter-Prototyp im Hintergrund durch einen Menschen gesteuert, während den Teilnehmer*innen ‚vorgetäuscht‘ wird, mit einem voll funktionsfähigen

Roboter in Dialog zu treten. Dabei wurden v. a. Fragen der Gesprächseröffnung (Gehle et al. 2017; Kuzuoka et al. 2008) und der Anbahnung (Scheffler/Pitsch 2020) bearbeitet. Es konnte festgestellt werden, dass mit passenden Kombinationen von Pausen und *Restarts* die Responsivität und damit die Wahrscheinlichkeit für einen länger anhaltenden Dialog erhöht werden kann (vgl. Pitsch et al. 2009a). Außerdem wurde v. a. die Notwendigkeit multimodaler Rezeptions- und Ausdrucksmöglichkeiten in den Bereichen Blick, Raum und Körperbewegung festgestellt (Gehle et al. 2017; Scheffler/Pitsch 2020). Parallelen zu Chatbots und ECAs zeigen sich auch hier im Hinblick auf die Ritualisierung bzw. Konventionalisierung der Begrüßung, die seitens der humanoiden Roboter deutlich stärker hergestellt wird als von den Nutzer*innen, auch wenn solche Klammern in den hier untersuchten Settings das Zusammenbrechen des Dialogs unwahrscheinlicher und das Eintreten in *topic talk* wahrscheinlicher machen (vgl. Pitsch 2015: 254).

Die Situiertheit und grundlegende Offenheit menschlicher Interaktion als Schwierigkeit für den Roboter treten gerade dann als Problem auf, wenn mehrere Personen am Dialog mit dem Roboter beteiligt sind. In Studien, in denen der Roboter nicht in einem WOZ-Setting kontrolliert, sondern im Modus der ‚autonomen‘ Steuerung eingesetzt wurde, konnte auch hierfür besonders die Rolle von Blick, Kopfbewegungen und der Koordination der Personen untereinander als relevant ausgemacht werden. So orientieren sich Teilnehmer*innen an Dialogen mit dem Roboter beim Auftreten von Fehlern koordiniert zueinander, unterbrechen den Blick zum Roboter und suchen die Orientierung zum Museumspersonal (vgl. Gehle et al. 2015: 410).

Besonders relevant sind Arbeiten, in denen die Interaktionsdynamiken innerhalb der Nutzer*innen-Gruppe genauer beleuchtet werden. So zeigt sich, ähnlich wie die bereits erwähnte Studie von Porcheron et al. (2018) es auch für häusliche VUI festgestellt hat,⁸² dass Antworten an den Roboter, die Adressierung dessen und die Beteiligung am Dialog interaktional zwischen den Beteiligten ausgehandelt wird, der Roboter diese Vorgänge aber nicht registriert: „a time slot which appears to the robot as ‚silence‘ turns out in fact as a locus of high interactional activity and negotiation between the visitors“ (Pitsch et al. 2017: 396). Für die Aus handlung der Partizipation werden neben leise gesprochenen verbalen Ausdrücken auch Körperhandlungen (z. B. Berührungen an der Schulter für die Zuweisung des Rederechts im Dialog mit dem Roboter) relevant (Pitsch et al. 2017: 395; siehe auch Pitsch 2020: 141–142). Bei Nutzer*innen-Gruppen mit Kindern zeigt sich, dass Erwachsene häufig die Rolle eines „Participation facilitators“ einneh-

⁸² Diese Studie wird zusammen mit anderen Studien zu Smart Speakern in Kap. 3.2.4 genauer diskutiert.

men bzw. interaktional zugeschrieben bekommen und Kinder unterstützen, die Rolle der primären Dialogpartner*innen einzunehmen (Gehle et al. 2017; Pitsch 2020). Auch hier wird die Relevanz von Blick, räumlicher Positionierung der Beteiligten und insbesondere Kopfbewegungen herausgestellt (vgl. Pitsch/Gehle/Wrede 2013).

Die situationsabhängige, spontane, dynamische und interaktiv ausgehandelte Partizipation und Positionierung zum Roboter zeigt sich – auch unabhängig von diskursiven (teils verfestigten Einstellungen) zu humanoider Robotik – auch in den Daten von Habscheid et al. (2020). So kann es auch durch das anthropomorphe Design der humanoiden Roboter zu einer Form der „emotionalen Befremdung“ (Habscheid et al. 2020: 184) kommen, die aus der Situation heraus entsteht und von den Beteiligten verbal expliziert und somit accountable gemacht wird. Insgesamt lässt sich feststellen, dass der Dialog mit autonomen, humanoiden Robotern auf Seiten der menschlichen Beteiligten als hochgradig situiert, von der räumlichen Umgebung geprägt und interaktional sowohl verbal wie non-verbal ausgehandelt betrachtet werden muss.

3.1.4 Telefonbasierte Sprachdialogsysteme

Auch Sprachdialogsysteme, die etwa in Call-Centern eingesetzt werden und einen Mensch-Maschine-Dialog evozieren, sind sprachwissenschaftlich beleuchtet worden. Eine Arbeit an der Schnittstelle von Gesprächslinguistik und Design-Forschung zu telefonbasierten Sprachdialogsystemen legt Thar (2015) vor. Die Analyse stützt sich auf Tonaufzeichnungen von sechs verschiedenen Dialogsystemen aus dem kommerziellen Call-Center-Einsatz (vgl. Thar 2015: 18). Dabei wurden 40 Personen in zwei Testreihen aufgezeichnet, wie sie Gespräche mit einem telefonbasierten VUI führen; dabei ist zu bemerken, dass es sich um Daten aus einem experimentellen Setting handelt (vgl. Thar 2015: 25), die nicht spontan entstanden sind, sondern durch einen entsprechenden Versuchsaufbau elizitiert wurden – dies diente mit Blick auf das Ziel der Arbeit, Empfehlungen für die Optimierungen des Designs der VUIs zu generieren, der Vergleichbarkeit. Gleichwohl geht Thar (2015: 25) davon aus, dass „die relevanten Verhaltensmuster durch die Untersuchungssituation nicht beeinflusst wurden“.

Thar (2015: 17) nimmt an, dass Nutzer*innen entweder bereits „vorhandene Gesprächskompetenz“ nutzen oder ein „neues Regelwerk“ entwickeln – genau dieses Spannungsfeld will auch die vorliegende Arbeit für Smart Speaker beleuchten, wenn auch mit einer weniger auf die Kompetenz der Nutzer*innen als vielmehr auf den praktischen Vollzug scharfgestellten Perspektive, wobei diese auch

implizites Wissen über Gesprächsorganisation sichtbar machen kann.⁸³ Thar arbeitet ethnomethodologisch-konversationsanalytisch. Den Teil ihrer Untersuchung zu Interaktionsphänomenen gliedert sie entlang klassischer konversationsanalytischer Kategorien und greift dabei auf Forschungsergebnisse zum Sprecher*innen-Wechsel, zu Reparaturen, zur Sequenzialität sowie zu Gesprächsbeginn und -ende auf (vgl. Thar 2015: 19). Diese Untersuchung führt Thar (2015: 250–251) zu Erkenntnissen im Hinblick auf den Mensch-Maschine-Dialog, u. a. beobachtet sie Schwierigkeiten im Sprecher*innenwechsel durch *barge-ins*, d. h. die sofortige Unterbrechung der Äußerungsproduktion des VUI, sowie bei der Durchführung von Reparaturen, die, so konstatiert sie, „immer selbstdurchgeführt umgesetzt werden“ (Thar 2015: 250) müssen. Diese Befunde werden auch in der Konzeption der Analysen der vorliegenden Arbeit wieder aufgegriffen.⁸⁴ Weitere Analysen von Telefonaten führen Thar (2015: 271–273) schließlich zu acht „Optimierungsprinzipien“, die u. a. die Bereiche Gesprächsstrategien und -organisation, Rezeptionssignale, Reparaturen und Gesprächssorten umfasst. Sie stellt dabei prosodische Besonderheiten bei den Stimmeingaben fest, die die Nutzer*innen verfolgen (insbesondere eine langsame, überartikulierte Sprechweise), obwohl dies die fehlerfreie Spracherkennung beeinträchtigen kann. Folglich empfiehlt Thar, die Systeme so zu gestalten, dass sie diese Sprechweise zum Vorteil nutzen können. Weitere Empfehlungen betreffen primär die Ausrichtung des VUI-Designs an grundlegenden Prinzipien der Gesprächsorganisation sowie der Gattungsspezifika der Telefonate, in denen die Systeme zum Einsatz kommen. Insgesamt liefert die Arbeit mehrere Anknüpfungspunkte auch für die vorliegenden Analysen und wohl die einzige umfassende sprachwissenschaftliche bzw. gesprächsanalytische Studie zu telefonbasierten VUI-Systemen.

Zu telefonischen Sprachdialogsystemen aus medienwissenschaftlicher bzw. -historischer Sicht arbeitete auch Volmar (2019) (siehe auch Waldecker/Volmar 2022). Er zeichnet deren Entwicklung von der Einführung des Tastentelephons über das Aufkommen von Tonwahlverfahren in Call-Centern bis hin zu KI-basierter automatischer Spracherkennung zu Smart Speakern nach. Er argumentiert, dass die Entwicklung von häuslichen Sprachassistenzsystemen eng verbunden ist mit der des Call-Centers, in der Nutzer*innen bereits an den Umgang mit maschinellen, dialogisch und mündlich operierenden Gegenübern gewöhnt wurden. Der Gebrauch von Sprachassistenzsystemen erinnere an Praktiken der Call-Center-Nutzung. Folglich schlägt er vor, Smart Speaker als „call centers for the home“ zu konzeptualisieren (Volmar 2019: 72) und verweist auf sich ausdifferenzierende

⁸³ Siehe Kap. 2.1.4.

⁸⁴ Siehe insbesondere Kap. 6.1.3 und 6.1.4.

Motive für den kommerziellen Einsatz von Sprachdialogsystemen: So sei es Anliegen von Call-Center-Betreibern, die notwendigen Personalressourcen zu senken und die Effizienz zu steigern, demgegenüber könne die Auswertung der erhobenen Sprachdaten zu unterschiedlichen Zwecken eine Motivation der Dienstanbieter von Smart Speakern sein.

3.2 Smart Speaker

Stationäre Sprachassistentensysteme für den Privathaushalt sind noch nicht lange verfügbar: In Deutschland war der erste Smart Speaker von Amazon (der *Amazon Echo* mit dem Sprachdialogsystem *Alexa*) ab Anfang 2017 ohne Vorbestellung erhältlich, auf dem US-amerikanischen Markt ca. anderthalb Jahre früher. *Google Home* bzw. *Nest* und der *HomePod* der Firma Apple folgten Mitte 2017 bzw. Mitte 2018 (vgl. Hoy 2018: 82). Folglich beziehen sich bisher veröffentlichte Studien zu großen Teilen auf den angloamerikanischen Raum, in dem alle Geräte etwas früher marktüblich waren, und sind insgesamt noch sehr jung, wie auch die Geschichte der Sprachtechnologie zeigt, über die nachfolgend ein Überblick gegeben werden soll.

3.2.1 Historische Entwicklung der Sprachtechnologie

Die Historie des sprachlichen Austauschs von Menschen und Maschinen ist eng verwoben mit der Entwicklung von Rechenmaschinen und Computern im Allgemeinen. Sie ist darüber hinaus mit der großen Frage nach der Künstlichen Intelligenz (KI) verknüpft: Ab wann kann eine Maschine intellektuell einem Menschen gleichwertig sein? Diese Entwicklungen und Erzählungen sind so oft in der einschlägigen Literatur mit unterschiedlichen Schwerpunkten beschrieben worden, dass hier davon abgesehen werden soll, noch einmal die philosophischen Überlegungen zur Unterscheidung von menschlicher und maschineller „Intelligenz“ von René Descartes von 1637 oder Denis Diderot von 1769 darzulegen. Es soll an dieser Stelle ebenso auf eine Erläuterung des von Alan Turing um 1950 ersonnenen Turing-Tests zur Unterscheidung menschlichen und maschinellen Denkens verzichtet werden⁸⁵ wie auf eine weitere Erzählung der Erfindung von ELIZA (Weizen-

⁸⁵ Siehe dazu Turing (1950); eine Anwendung auf zeitgenössische Dialogsysteme findet sich etwa bei Kaerlein (2018: 19–21); siehe auch Nebel (2019) und explizit zur Geschichte der Stimmsynthese Borbach (2018). Turgings Beitrag, der auch als ein Grundstein für die Entwicklung von Digitalität betrachtet werden kann, und das darin enthaltene Experiment zur Unterscheidung

baum 1966) und der daraus resultierenden Kritik an der Vorstellung von der ‚starken künstlichen Intelligenz‘ (vgl. Weizenbaum 1977).⁸⁶ Auch der „Schachtürke“ des Wolfgang von Kempelen von 1789 und darauf aufbauende Wizard-of-Oz-Experimente, die auf einer Täuschung der Nutzer*innen basieren, sollen hier nicht detailliert beschrieben werden (siehe dazu Felderer/Strouhal 2004).⁸⁷

Damit will die Arbeit keineswegs in einen ahistorischen Blindflug geraten. Wolfgang von Kempelen, Konrad Zuse, Alan Turing, Joseph Weizenbaum und andere haben zweifelsohne bedeutende Grundsteine für eine Erforschung und Weiterentwicklung des Austauschs zwischen Mensch und Maschine gelegt, die auch für die Gegenwartsforschung von höchster Relevanz sind (vgl. etwa Baranovska/Höltgen 2018b: 11). Doch nicht nur sind diese technologischen Entwicklungen mit ihren sozialen Konsequenzen mittlerweile umfangreich dargestellt worden (siehe dazu etwa die Monographien von Heintz 1993 und Wilker 2002),⁸⁸ sie sind auch für die hier durchgeführten Analysen nur implizit relevant: Durch Smart Speaker und andere Voice User Interfaces (VUIs) – spätestens mit der Präsentation von Siri 2011 – ist der mündliche und (bedingt) konversationsartige Austausch mit Maschinen Realität geworden. Es lässt sich dabei beobachten: Der Austausch ist keine Täuschung wie bei Wolfgang von Kempelen mehr, bei der ein Mensch die Antworten vorsagt. Die Anwendung von auf Wahrscheinlichkeitsrechnungen beruhenden NLP-Prozessen macht die Funktionsweise der Systeme verlässlicher und er prägt ihre Anwendungsgebiete. Die Systeme sind deutlich komplexer und

von Mensch und Maschine sind wirkmächtig und können implizit für die folgenden Analysen eine Rolle spielen, werden darin aber nicht explizit, sodass eine ausführliche Darstellung hier unterbleiben soll.

86 Zweifellos hat ELIZA großen Einfluss auf die ethnomethodologische bzw. praxeologische Betrachtung von Mensch-Maschine-Dialogen sowie auf deren Erforschung im Allgemeinen gehabt – das zeigt nicht nur das gesteigerte Interesse Harold Garfinkels an ELIZA (vgl. Suchman 1987: 65; Lynch 2011: 932), sondern auch die weitere Beschäftigung damit in rezenten Forschungsarbeiten (Baranovska/Höltgen 2018a; Shrager 2021; Eisenmann et al. 2023).

87 Es unterbleibt hier auch ein in einschlägigen Arbeiten häufig zu findender Abgleich zwischen den jüngeren Entwicklungen der Realität und populären Fiktionen in älteren Filmen, etwa aus Filmen wie *2001: Odyssee im Weltraum* (1968) mit dem darin eingeführten Sprachroboter *HAL 9000* (siehe Baum 2018). Weitere aktuellere Beschäftigungen mit dieser Frage in Film und Literatur finden sich etwa in *Her* (2013), in der sich der Protagonist mit dem „Operating System“ Samantha auf eine Liebesbeziehung einlässt, in Romanen wie *Maschinen wie ich* (McEwan 2019), Erzählungen und Kurzgeschichten (Kehlmann 2021; Braslavsky 2019) sowie in der populärwissenschaftlichen Literatur (etwa Ramge 2018; Misselhorn 2021; Lenzen 2024).

88 Zum Turing-Test siehe auch Krummheuer (2010: 78–79) und Lotze (2016: 28); zur Erfindung und zu den Folgen von Joseph Weizenbaums ELIZA siehe ebenfalls Krummheuer (2010: 79–80) und Lotze (2016: 31–34). Für eine zusammenhängende, populärwissenschaftliche Darstellung siehe Drösser (2020: 15–30).

v. a. ‚gehaltvoller‘ als die Dialoge, die etwa mit ELIZA möglich waren: Über VUIs können Informationen abgefragt, Programme bedient und Haushaltsgeräte gesteuert werden. Der Austausch entspricht dennoch nicht den in der Popkultur gezeichneten Fiktionen vom Mensch-Maschine-Austausch der Zukunft: Sie werden zwar von den Herstellern als *Persona* mit Identitätsmerkmalen ausgestattet, entwickeln aber keinen eigenen Willen oder Gefühle wie Samantha in „Her“. Der Austausch mit VUIs ist bei den o. g. Systemen akustisch für kompetente Sprecher*innen gut von zwischenmenschlichen Interaktionen zu unterscheiden.⁸⁹ Smart Speaker in ihrer zuvor beschriebenen Erscheinungsform basieren durch ihre Digitalität, ihre im Hintergrund prozessierte Vernetztheit, ihre stochastische Funktionsweise und nicht zuletzt durch ihre Kommerzialisierung insofern auf ganz anderen technologischen und sozialen Grundlagen als ihre Vorläufer.

Die jüngere Geschichte der Sprachtechnologie setzt mit den Erfolgen bei den korpusbezogenen Verfahren der Sprachtechnologie in den 80er-Jahren ein (vgl. Carstensen et al. 2010: 21). Diese bedeuteten eine Abkehr von symbolischen Ansätzen mit Universalitätsanspruch und konzentrierten sich auf das *Training*, d. h. den Einsatz von für die Nutzung optimierten, vorbereiteten Datensätzen zur Modelloptimierung bei der Erkennung von Sprachmaterial. Wie Carstensen et al. (2010: 21) herausstellen, ist zu diesem Zeitpunkt (noch) nicht die stochastische Vorgehensweise, sondern der Einbezug eines Trainingsverfahrens zur Anpassung von Rechenmodellen für verschiedene (u. a. stochastische) Verfahren entscheidend. Die Abkehr vom Universalismus hin zur Anwendung der Modelle und Anpassung der Programme durch die Trainingsdaten auf einen bestimmten Einsatzzweck hin waren die wesentliche Veränderung im Vergleich zu vorherigen Versuchen. Gleichwohl wurden v. a. Ansätze weiterentwickelt, die dem, wie Drösser (2020: 12) es vergleicht, „statistischen Lernen“ eines Kindes im Erstspracherwerb ähneln. *Hidden-Markov-Modelle* zur Umwandlung von Laut in Text und *N-Gramme* zur Identifizierung der Laute und Wörter (siehe dazu Carstensen et al. 2010: 130–136; Pfister/Kaufmann 2017: 107–110 sowie Kap. 3.2.2) wurden mit mehr Rechenleistung ausgebaut, konnten über mehr Trainingsdaten verfügen und in den 90er-Jahren konnten die Verfahren für einfache Anwendungsbereiche (z. B. einfache Telefon-Dialogsysteme) kommerziell eingesetzt werden.

Die in den 80er- und 90er-Jahren zur Verfügung stehenden Daten und v. a. die Rechenkapazitäten reichten allerdings keineswegs aus, um damit algorithmi-

⁸⁹ Eine Ausnahme könnte das von Google 2018 vorgestellte System „Google Duplex“ sein, das jedoch nicht in der Breite verfügbar ist. Durch die jüngeren Entwicklungen im Bereich der Large Language Models (LLMs) ist möglicherweise ebenfalls eine zunehmende Einschränkung der Unterscheidbarkeit von zwischenmenschlichen Interaktionen und Interface-Dialogen zu treffen, wobei diese aber zum Zeitpunkt der Datenerhebung für die vorliegende Arbeit noch nicht in der Breite verfügbar waren; zur Relevanz LLM-basierter Systeme siehe Kap. 7.3.

sches *Deep Learning* wie es heute Standard im Natural Language Processing (NLP) ist, zu ermöglichen. Sie scheiterten noch an größeren Erkennungsaufgaben und hatten hohe Fehlerquoten. Spracherkennung beschränkte sich weiterhin auf einzelne Schlüsselwörter oder basale Eingaben und hatte nur unter wenigen Aspekten Ähnlichkeit mit menschlichen Konversationen. Dennoch entwickelten sich durch den zunehmenden Ausbau von Netzwerken und die damit grundlegend veränderte Menge und Verfügbarkeit von Daten Anwendungen im Bereich der KI weiter, wie z. B. siegreiche Schachcomputer (Deep Blue von IBM gewann 1997 gegen den damaligen Schachweltmeister).

Für qualitativ hochwertige Spracherkennung und -produktion reichte dies dennoch noch nicht aus, auch wenn 1990 mit der Spracherkennungssoftware DragonDictate erste Anfänge unternommen worden waren. Erst die deutlich umfangreicheren Rechenkapazitäten und die Weiterentwicklung von *Deep Learning* um die 2010er-Jahre, mit denen sogenannte neuronale Netze der Standard wurden, ermöglichten die Mustererkennung in weitaus größerem Stil, die für Bild-, Text- und Spracherkennung angewendet werden konnte. Hier konnte ein anderes IBM-Produkt, der Computer Watson, in der Quiz-Show *Jeopardy!* gewinnen. Watson war im Unterschied zu Deep Blue ein „aus Daten lernendes System“ (Ramge 2018: 40), das viel mehr Rechenkapazität aufwies und die Optimierung der in ihm eingebauten Modelle insofern in viel kürzerer Zeit ermöglichte. Zu einer erhöhten Präzision trug zudem das „Long Short-Term-Memory“-Verfahren bei, das zwar bereits seit Ende der 90er-Jahre vorgestellt wurde (vgl. Hochreiter/Schmidhuber 1997), aber wegen der in der Breite nicht zur Verfügung stehenden Hardware kaum zum Einsatz kam. Es setzte sich durch und führte zu entsprechenden Leistungen, z. B. Anfang der 2010er-Jahre in der Erkennung von Handschriften (vgl. Schmidhuber 2019). Die Forschung großer Tech-Firmen wie IBM, Apple, Facebook oder Microsoft führten ebenfalls zu einem Schub, der zwischen 2000 und 2010 gravierende Veränderungen im Bereich der Bild- und Spracherkennung auslöste. Dieser Schub trug auch zur Entwicklung der Software Siri bei, die 2011 von Apple vorgestellt wurde und den ersten IPA darstellt, der über ein VUI bedienbar und auf marktüblichen Produkten in der Breite verfügbar war. Konkurrenzprodukte von Google, Amazon und Microsoft folgten in den Jahren 2012 bis 2014 und dominieren seitdem den Markt für IPA und insbesondere für IPA auf eigenen Endgeräten (Smart Speaker).

Die Entwicklungen seit der Jahrtausendwende waren also ganz wesentlich davon geprägt, dass große Rechenkapazitäten Modelle Realität werden ließen, die zuvor bereits seit den 70er- und 80er-Jahren entwickelt worden waren. Sie waren außerdem wesentlich davon beeinflusst, dass das private Internet sich ausbreitete und die Wirtschaft ein Marktpotenzial in entsprechenden Anwendungen erkannte und entsprechend investierte. Mathematisch-informatisch-technische Ent-

wicklungen waren ebenfalls an diesen Prozessen beteiligt, doch lag hier bereits viel Pionierarbeit vor, auf die mit dem Aufkommen der technologischen Verbesserungen aufgebaut werden konnte.

3.2.2 Funktionsweise und technischer Hintergrund

Smart Speaker werden im Deutschen auch als Sprachassistenten, „Intelligente Persönliche Assistenten“ oder „Intelligente Lautsprecher“ bezeichnet. Der hier zu beschreibende Gerätetyp zeichnet sich dadurch aus, dass er als stationäres Sprachassistenzsystem fest in der häuslichen Wohnumgebung der Nutzer*innen installiert wird und, durch seine Größe und ein obligatorisches Stromkabel, nicht mobil ist.⁹⁰ Sie sind „standalone screenless devices“ (Porcheron et al. 2018: 1) und grenzen sich somit von anderen Sprachassistenzsystemen ab, die auch als „Conversational Agents“, „Voice Assistants“, „Sprachassistenten“ oder „Conversational User Interfaces“ bezeichnet werden. Solche rein softwarebasierten Instanzen umfassen zwar auch einen Großteil der oben beschriebenen technologischen Anwendungen. Sie können auch als Voice User Interface (VUI) stimmlich bedient werden, verfügen aber nicht unbedingt über eine eigene Hardware. Vielmehr sind sie als Software auf mehr oder weniger mobilen Geräten wie Smartphones, Tablets, Computern oder auch im Auto installiert. Sie binden dabei in die Verarbeitung der Anfragen und die Darstellung der Ergebnisse auch die Bildschirme mit ein. Sie signalisieren z. B. durch Symbole, dass die Anfrage verarbeitet wird, zeigen unterstützend zur mündlichen Ausgabe auch Text, Tabellen, Karten oder Bilder und ersetzen teilweise sogar die rein mündliche Ausgabe durch eine visuelle Darstellung. Bedford-Strohm (2017: 486) verweist zur Abgrenzung der Smart Speaker von diesen Geräten auch auf den Begriff der „Voice-First-Geräte“, die ausschließlich über die Stimme bedient werden; ein Eingreifen über den visuellen Kanal ist nur bedingt möglich.⁹¹ Der Begriff Smart Speaker soll im Folgenden beibehalten werden. Er verweist erstens über den Lautsprecher auf dessen physische Materialität (auch wenn diese noch mehr umfasst als den Lautsprecher, s. u.). Der Begriff erscheint außerdem passend, weil er mit dem englischen Begriff „Smart“ den im deutschsprachigen Diskurs umstrittenen Begriff der „Intelligenz“

⁹⁰ Auch die zugehörigen Smartphone-Apps orientieren sich bei der Zuordnung der Geräte an verschiedenen Wohnbereichen (Esszimmer, Wohnzimmer, Küche o. ä.) und rahmen sie damit als stationär und nicht für den tragbaren Einsatz vorgesehen.

⁹¹ Zu den Spezifika der einzelnen Assistenzsysteme und die Anbindung insbesondere an die Smartphone-App, über die eine gewisse visuelle Kontrolle über die Geräte hergestellt wird, siehe Kap. 3.3.

umschiff, und ferner, weil die Konnotation mit *Smart Home* passend erscheint, da Smart Speaker wesentlicher Bestandteil einer Smart-Home-Struktur sein können (vgl. Goulden 2019; Strüver 2023a).

Hardwareseitig bestehen diese Systeme v. a. aus drei Komponenten: mehreren, miteinander zusammenwirkenden Mikrofonen (sog. *Far-Field-Voice-Recognition*, siehe Drösser 2020: 71), einem Lautsprecher sowie einer kleinen Recheneinheit, die die Übertragung ins Internet leistet und einige wenige lokale Prozesse steuert. Teilweise verfügen die Modelle außerdem über Lichtringe oder -punkte, die in unterschiedlichen Farben nichtsprachliche Signale geben können (vgl. Bedford-Strohm 2017: 487; Pearl 2016: 219–221).⁹² Die einzelnen Komponenten, aus denen das stationäre Sprachassistenzsystem sich zusammensetzt, werden von verschiedenen Software-Anwendungen in Anspruch genommen (vgl. die Abbildung bei Zarcone/Leschansowsky 2023: 169). Diese ergeben erst im Zusammenspiel eine Einheit und sind im Hintergrund als unterschiedliche Prozesse zu begreifen, die auch verschiedene Teildisziplinen der Informatik und der Computerlinguistik berühren.

Die genannten Prozesse lassen sich im Wesentlichen mit Natale (2020: 5) in drei verschiedene Großbereiche unterteilen: *Speech Processing* (SP), *Natural Language Processing* (NLP) sowie *Information Retrieval* (IR).⁹³ Damit sind jeweils fachlich tiefgreifende informatisch-computerlinguistische Prozesse angesprochen, die im Rahmen dieser Arbeit nicht ausführlich beschrieben werden können und auch nicht müssen. Allerdings sollen die Funktionen, die diese Prozesse für den Austausch zwischen Mensch und Maschine übernehmen, hier kurz überblicksartig vorgestellt werden, um erstens die an die Smart Speaker adressierten Äußerungen der Nutzer*innen vor diesem Hintergrund besser interpretieren zu können und zweitens die Ausgaben des VUI einzuordnen.⁹⁴

Wesentlich für das Funktionieren aller folgenden Dienste ist die dauerhafte Verbindung zum Internet (vgl. Hoy 2018: 82). Lediglich die für die Aktivierung des Geräts notwendige *wake word*-Suche findet auf den lokalen Recheneinheiten der Smart Speaker statt (vgl. Pearl 2016: 127), was auch den VUI-Design-Empfehlungen etwa bei Pearl (2016: 153) entspricht: „recording everything the user says, even when not engaged with your app, and sending it to the cloud is not ethical“. Nach Selbstauskunft von Amazon und Google wird diesen Empfehlungen auch bei den Echo- und Google-Assistant-Geräten gefolgt (vgl. Amazon 2021; Google 2021);⁹⁵

⁹² Für eine genauere Beschreibung der in dieser Arbeit analysierten, aktuell marktüblichen stationären Sprachassistenzsysteme, siehe Kap. 3.3.

⁹³ Für eine Übersicht über das mit Smart Speakern zusammenhängende Fachvokabular siehe auch Kahle/Meißner (2020: 21).

⁹⁴ Eine auch für Laien verständliche Übersicht liefert das Sachbuch von Drösser (2020).

⁹⁵ Siehe dazu die Diskussion in Kap. 3.2.6.

gleichwohl findet ein dauerhaftes akustisches Scannen (ohne weitreichende Auswertung) statt, das die mündlich vollzogene Bedienweise ermöglicht (vgl. Hoy 2018: 82). Sobald das Aktivierungswort erkannt wird, wechselt das Gerät in den Listening-Modus und es findet eine Transkription und Auswertung der Daten über die cloudbasierten Online-Dienste der Plattformen statt. Die Verarbeitung wird erst dadurch möglich, dass das aufgenommene akustische Material mit Hilfe von *Speech Recognition*-Anwendungen analysiert wird, für die große Datenmengen und Rechenleistungen erforderlich sind (vgl. Crawford/Joler 2018; Kohne et al. 2020: 20). Für die Spracherkennung, d. h. die reine Transkription von Gesprochenem in Geschriebenes, werden probabilistische Modelle eingesetzt, die es ermöglichen, die Wahrscheinlichkeit auszurechnen, dass ein bestimmter auditiver Input einer bestimmten Abfolge von Buchstaben entspricht (für mathematisch-technische Grundlagen zu diesem Prozess siehe Kamath/Withaker 2019: 141–201): Der auditive Input wird dabei zunächst in eine digitale Darstellungsform gebracht und dann, grob gesagt, mit einem Ansatz der statistischen Spracherkennung analysiert, den Pfister/Kaufmann (2017: 328) wie folgt beschreiben: „Der Erkenner verfügt für jedes Wort des Vokabulars über eine statistische Beschreibung von dessen akustischer Realisierung. Es gilt dasjenige Wort als erkannt, dessen statistische Beschreibung am besten auf die zu erkennende Äußerung [sic] passt“. Diese Erkennung wird durch eine große Menge von Daten ermöglicht, mit denen eine bei Pfister/Kaufmann zuvor als „Erkennung“ bezeichnete mathematische Funktion „trainiert“ wurde. Durch die fortlaufend hinzukommenden Daten, die laut der Datenschutzvereinbarungen mit den Anwender*innen der Geräte auch entsprechend verwendet dürfen, werden die Funktionen immer weiter angepasst und verändern bzw. ‚verbessern‘ das neuronale Netzwerk durch die neuen Berechnungen (siehe auch Knaut 2018: 80; Kipp 2023). Dieser Prozess vollzieht sich zu Teilen ohne manuelles Eingreifen und wird primär aufgrund der statistischen Angaben in der Funktion realisiert (vgl. Kamath/Withaker 2019: 6–7). Dabei werden über sog. *N-Gramm*-Sprachmodelle nicht nur die Tonhöhen eines einzelnen Wortes ausgewertet, sondern auch die Wahrscheinlichkeit des Auftretens in Nachbarschaft zu anderen Wörtern; bei der Berücksichtigung von zwei Wörtern (in älteren Systemen) spricht man hier von *Bi-Gramm*-Sprachmodellen, doch der Prozess ist prinzipiell auf bis zu *n* Wörter erweiterbar (siehe dazu Pfister/Kaufmann 2017: 417–421). Zu nicht unerheblichen Teilen erfolgt das Training jedoch auch durch menschliches Nachhören der gespeicherten Aufnahmen, deren maschinelle Transkription dann ‚verbessert‘ wird, was die mathematischen Berechnungsverfahren wiederum verzeichnen und in ihre Transkriptionsverfahren einfließen lassen – dies hatten alle Betreiberfirmen nach investigativen Recherchen in öffentlichen Statements bekanntgegeben (vgl. Kremp 2019; siehe auch Waldecker/Volmar 2022: 164; Hector 2025: 71).

Die genauen technischen Details über den Prozess der *Speech Recognition* geben die Hersteller nicht öffentlich außerhalb von, für Laien schwierig nachvollziehbaren, Fachjournalen oder eigenen Webseiten⁹⁶ bekannt. In etlichen Publikationen zu Smart Speakern aus den Geistes-, Kultur- und Sozialwissenschaften und in der Presse wird bei der Beschreibung die genaue Funktionsweise ausgespart (und auch die vorliegende Arbeit geht diesbezüglich nicht in die Tiefe). Wie jedoch Drösser (2020: 93–99) bemerkt, kommen die erwähnten probabilistischen Erkennungsprogramme in den derzeit verfügbaren Sprachmodellen generell zum Einsatz; erst der Einsatz von *Deep Learning* und die Entwicklung der dafür notwendigen Rechenleistungen machten eine solche Art des *Speech Processings* überhaupt möglich. Sehr ähnlich funktioniert der Prozess auch bei der Sprachsynthese, also dem Erzeugen sprachlicher Laute durch die Geräte. Insbesondere die Prosodie spielt bei der Sprachsynthese eine entscheidende Rolle (vgl. Pfister/Kaufmann 2017: 262–264) und auch sie konnte durch neuronale Netzwerkfunktionen und Rechenleistungen verändert werden, indem sie die statistischen Wahrscheinlichkeiten, z. B. auch für noch unbekannte Lexeme, für eine bestimmte Sprechgeschwindigkeit oder Tonhöhe aus deren Vorkommen in anderen Zusammenhängen heraus errechnen (siehe auch Drösser 2020: 55–60). Grundlage sind dabei schriftliche Inputs, die aus dem zweiten großen Baustein heraus entstehen, den Smart Speaker vereinen: *Natural Language Processing* (NLP).

NLP bezeichnet den Prozess der *Verarbeitung* natürlicher Sprache, d. h. deren Erkennen und graduelles Verstehen (siehe dazu Goksel-Canbek/Mutlu 2016: 594). Der Prozess kann dabei in zwei Teilbereiche unterteilt werden: Natural Language Understanding (NLU) und Natural Language Generation (NLG) (vgl. Kohne et al. 2020: 42–43). Das Verstehen und Generieren verweist hier nicht auf die Transformation von Gesprochenem zu Geschriebenem: Im Bereich des NLP wird ausschließlich mit schriftlichen Daten gearbeitet, der Wechsel zwischen der mündlichen und der schriftlichen Darstellungsweise liegt ganz im Bereich des zuvor erläuterten *Speech Processing*. Der transkribierte Input wird nun (im Prinzip) genauso auf der Basis schriftlichen Inputs verarbeitet wie bei einem Chatbot oder anderen Conversational User Interfaces auch (vgl. Kabel 2020: 1–2; Bedford-Strohm 2017: 486–487). NLU „sucht“ nun in verschiedenen Schritten in dieser Äußerung (*utterance*) nach *Intents* und entsprechenden *Entities*: Erstere ist, mit Kohne et al. (2020: 94), die „Absicht des Benutzers“ oder, genauer genommen, das, was das Programm als mögliche Absicht bzw. aufzurufende Funktion identifiziert (z. B. „Wetter“, „Nachrichten“ oder „Bahnverbindung“). Letztere sind Spezifikationen, die in der erkannten Funktion abgefragt

⁹⁶ Siehe etwa Amazon Science (2023).

werden sollen (z. B. „Wetter in Siegen“, „Sportnachrichten“, „Bahnverbindung nach Berlin“). Komplexere Systeme ‚suchen‘ außerdem nach dem *Skill*, d. h. dem spezifischen Programm, mit dem die angeforderte Funktion ausgeführt wird. Diesen *Skills* ist eine *invocation phrase* zugeordnet, die das Programm öffnet und den *Intent* sowie die *Entity* damit verarbeitet. Innerhalb des *Intents* werden zudem funktionspezifische *Slots* vergeben, die die *Entities* genauer hinsichtlich der ‚Such-Erwartungen‘ seitens der Software klassifizieren; so würde in dem fiktiven Beispiel „Bahnverbindung nach Berlin“ im Bereich nach Berlin als *Slot* eine Ortsangabe ‚erwartet‘ und die Suche daraufhin eingegrenzt.

Für die Verarbeitung des Inputs auf diese Weise durchläuft das Programm verschiedene Schritte, die grob in Tokenisierung, Lemmatisierung, Wortartenklassifikation, Satzteilung und Wort-Vektorisierung untergliedert werden können (vgl. Kohne et al. 2020: 44–49). In der Tokenisierung werden, auch hier mit Hilfe algorithmischer bzw. statistischer Verfahren, die erkannten Buchstaben in einzelne Wörter oder Satzzeichen unterteilt (vgl. Kamath/Withaker 2019: 92–93). Die erkannten Wörter werden nun durch sog. *Stemming* bzw. Lemmatisierung auf ihre Grundformen reduziert, um sie um Wortendungen durch Deklination und Konjugation und andere morphologische Prozesse zu bereinigen, sodass die Variabilität reduziert und somit die statistische Treffsicherheit erhöht wird; die genauen syntaktischen Bezüge der Wörter untereinander, die durch diesen Prozess verloren gehen, werden dabei marginalisiert (vgl. Kohne et al. 2020: 45–46; Kamath/Withaker 2019: 92). Durch Wortartenklassifikation und Satzteilung wird der Input weiter für die Verarbeitung zugerichtet. Um schließlich die Inputs den ‚gesuchten‘ Stellen zuzuordnen (*Skills*, *Intents*, *Launch*, *Entities*, *Slots*) ist diese Vorbereitung unumgänglich, aber noch längst nicht ausreichend, um eine hohe Treffsicherheit bei dieser Zuordnung zu erzielen, insbesondere nicht, wenn nicht vorausgesetzt wird, dass die Stimmeingaben in ihrer sprachlichen Gestaltung variieren: NLP kann, nach heutigen Standards, ebenso *Utterances* wie „Wie wird das Wetter in Siegen?“ wie auch „Wird es heute in Siegen regnen?“ verarbeiten. Dazu werden die erkannten Wörter vektorisiert, d. h., die Algorithmusfunktion prüft die gemeinsamen Kontexte und die Häufigkeit, mit der diese in den jeweiligen Kontexten auftreten, als Beziehung zueinander. Dieser Vorgang wird bei Kohne et al. (2020: 49) als räumliche Beziehung oder Vektoren dargestellt: Je größer der Abstand, d. h. je geringer das Vorkommen in einem gemeinsamen Kontext, desto weniger wahrscheinlich ist es, dass die Worte synonym für eine bestimmte Suchstelle verwendet werden können. Die Berechnung der gemeinsamen Kontexte erfolgt dabei auf Grundlage von zuvor durch die neuronale Netzwerk-Funktion verarbeiteten Daten, die die Bedeutungsähnlichkeiten und Auftreten unterschiedlicher Wörter in der gleichen Anfrage bzw. dem gleichen Zusammenhang für die Funktion erkennbar macht, sodass die entsprechenden Parameter darauf einge-

stellt sind. Durch die fortlaufende Nutzung entstehen auch hier immer weiter Neuberechnungen: Einmal erkannte Zusammenhänge bzw. Kontexte werden immer stabiler erkannt, während Abweichungen von den einmal erkannten Mustern immer schneller ausgeschlossen werden.⁹⁷ Zudem können diese „Advanced Bots“ (Kabel 2020: 2) nicht nur den Kontext, sondern auch die Interaktionshistorie auswerten und stellen Wahrscheinlichkeitsbezüge zu in vorherigen *Utterances* geäußerten *Intents* oder *Entities* her.

Für die NLG kommt eine Kombination aus Textbausteinen und Variablen zum Einsatz. Während einige der Dialoge manuell einprogrammiert sind, sodass einer bestimmten Frage eine konkrete, voreingestellte Antwort zugeordnet wird (etwa bei Fragen nach dem Sinn des Lebens oder nach Charaktereigenschaften des VUIs), wird bei anderen, funktionell komplexeren Anwendungen die Antwort mit Informationen aus dem *Information Retrieval*-Prozess (s. u.) gespeist und dann in entsprechende Text-Vorlagen eingebaut.⁹⁸ Die beiden Prozesse des NLP (NLU und NLG) rahmen zeitlich gesehen den Prozess des *Information Retrieval* (IR), da die NLU zunächst abgeschlossen sein muss, bevor überhaupt IR betrieben werden kann, gleichzeitig die NLG erst erfolgen kann, wenn der IR-Prozess zu einem Abschluss gekommen ist. Der ganze Vorgang dauert meist nur sehr wenige Sekunden, dennoch verzahnen sich hier wie bereits angedeutet prinzipiell autonome Software-Prozesse miteinander. Der letzte hier zu beschreibende Prozess der IR ist insofern voraussetzungsreich, als er auf mit den Cloud-Systemen der Hersteller verknüpfte Datenbanken und Internet-Suchmaschinen angewiesen ist (vgl. Natale 2020: 12–13). Welche Daten dort an welcher Stelle konkret abgefragt werden, hängt davon ab, welche Stellen mit den jeweiligen *Skills* verknüpft sind und welche *Entities* erkannt wurden. Die Stellen, auf die dabei zugegriffen wird, können sehr unterschiedlich sein, doch im Wesentlichen haben auch hier der sprunghafte Anstieg der Rechenleistungen in den letzten Jahren klassische Datenbankabfragen ersetzt, IR ist so vielmehr definiert als „finding material [...] of an unstructured nature [...] that satisfies an information need from within large collections [...]“ (Manning/Raghavan/Schütze 2018: 1).

Bei den Suchanfragen wird nicht zwingend von strukturierten Daten ausgegangen, sondern von nicht für die automatisierte Abfrage vorbereiteten Daten:

⁹⁷ Siehe dazu auch das von Drösser (2020: 35) angeführte Beispiel einer NLP-Anwendung, die durch die Überschwemmung mit rassistischen und beleidigenden Daten so „trainiert“ wurde, dass die Funktion nach kurzer Zeit selbst stabil rassistische und beleidigende Ausdrücke ausgab.

⁹⁸ Für eine übersichtliche Darstellung der Abhängigkeitsverhältnisse von maschinellem Lernen und algorithmischen Funktionen, Sprachtechnologie und Suchanfragen als Anwendungsfall siehe auch Kabel (2020: 40–41).

„Die Suche [...] berücksichtigt dabei die Vagheit und Unvollständigkeit, die sowohl bei der Formulierung des Informationswunsches als auch bei der – ggf. automatischen – Interpretation des Inhalts der betrachteten Dokumente besteht“ (Henrich 2008: 15). Doch welche Schnittstellen genau mit den einzelnen *Skills* verknüpft sind, ist von Anwendung zu Anwendung und den damit verbundenen Datentypen unterschiedlich (vgl. Waitelonis 2018: 8): So ist etwa der auf Amazon-Geräten vorinstallierte Wetter-Skill zum Zeitpunkt der Erhebung mit der Plattform „AccuWeather.com“ verknüpft, von wo aus die benötigten Informationen in strukturierter Form abgefragt werden können. Dabei (und auch bei anderen Funktionen, z. B. Routenplanung oder Kartendiensten) werden auch der Standort des Geräts bzw. Geodaten einbezogen. Auch für andere Wissens- bzw. Informationsabfragen (zur Unterscheidung von Daten, Wissen und Information in diesem Kontext siehe Henrich 2008: 18–19) kann auf durch semantische Datenbanken strukturiertes Wissen z. B. in Form des Google Knowledge Graph oder von Wikidata zugegriffen werden (vgl. Fensel et al. 2020: 8–10), was der IR deutlich umfangreichere Informationen bereitstellt und diese auch deutlich präziser werden lässt. Auch wenn also die Methoden der IR in nicht-strukturierten Wissensbeständen ebenfalls zum Einsatz kommen, treiben die Hersteller die Entwicklung von strukturierten Daten als Grundlage für die Wissensabfragen voran.

Solche Knowledge Graphs sind „very [Herv. i. O.] large semantic nets that integrate various and heterogeneous information sources to represent knowledge about certain domains of discourse“ (Fensel et al. 2020: 6); es handelt sich also etwas einfacher gesagt um sehr große, automatisch verarbeitbare und semantisch miteinander vernetzte Wissensbestände, die es ermöglichen, einfache Fragen (z. B. nach dem Alter von Prominenten oder der Höhe von Bauwerken) allein auf Grundlage der strukturierten Daten des *semantic web* zu beantworten, ohne auf in Text- oder Bildform und nicht-annotiert vorliegende Internetdokumente zurückgreifen zu müssen, deren maschinelle Verarbeitung trotz ausgeklügelter IR-Methoden häufig noch an ihre Grenzen stößt. Die Ergebnisse der Knowledge Graph-Suche sind bei Suchmaschinen-Abfragen über Google in sog. „Knowledge Panels“ noch über der eigentlichen Ergebnisliste zu sehen und werden in die Sprachausgaben der Smart Speaker eingebunden. Der Grundstock der Wissensbestände in ihrer heutigen Form speiste sich aus Internetquellen wie der Online-Enzyklopädie Wikipedia bzw. Wikidata (vgl. Vrandečić/Kröttsch 2014),⁹⁹ Freebase und der Statistik-Basis des *CIA World Factbook* (vgl. Singhal 2012).

⁹⁹ Für das Online-Projekt Wikipedia dient mittlerweile die von den Wikipedia-Trägerorganisationen gegründete Plattform Wikidata als *semantic web*-Anwendung und zugrundeliegende Datenbasis. Der Google Knowledge Graph speist sich auch weiterhin aus den dort hinterlegten Eintragungen und Zu-

Diese Internetsuchfunktionsweise über Knowledge Graphs ist nicht nur für den Funktionsbereich des IR von Interesse: Durch ihre Vernetztheit können Knowledge Graphs auch Spracherkennungssoftware und NLP-Methoden unterstützen, indem sie Zusammenhänge und Kontexte, in denen Äußerungen auftreten, nicht nur – wie oben beschrieben – auf Basis von *Datenbeispielen* und Neuronalen Netzwerk-Funktionen erkennen, sondern auch auf Basis von semantischen Verknüpfungen. Insofern spielt die weitere Entwicklung von Knowledge Graphs besonders für Smart Speaker und andere VUI-basierte Anwendungen eine wesentliche Rolle und bildet z. B. im Fall von Google auch die Basis für die gesamte VUI-Anwendung *Google Home*.

In der Gesamtschau betrachtet sind die Systeme für die Geistes- und Sozialwissenschaften also durchaus schwer zugänglich, auch wenn jüngst Einführungen und sogar Lehrbücher in diesem Bereich vorbereitet werden (etwa Lotze i. V.). Als Hauptquelle für das Wissen über die Funktionsweise können einerseits Blogs und Dokumentationen der Hersteller identifiziert werden, in denen die Schnittstellen auch für Drittanbieter beschrieben werden, die die Entwicklung von Anwendungen für die Geräte planen (vgl. Strüver 2023a). Andererseits sind Analysen von Application Programming Interfaces (APIs) mit unterschiedlicher Ausrichtung ein emergentes, interdisziplinäres Forschungsfeld, das derzeit aus unterschiedlichen Richtungen her Untersuchungen zu Schnittstellen durchführt – etwa aus den Plattform- und Interface-Studies (vgl. Gerlitz et al. 2019; Helmond/Nieborg/van der Vlist 2019; Hind/Seitz 2024) sowie den eher informationswissenschaftlich orientierten Security Studies (siehe etwa Kumar et al. 2018; Ford/Palmer 2019; Iqbal et al. 2022). Burrell (2016) unterscheidet unterschiedliche Formen von Opazität: erstens durch technisches Unverständnis entstehende Unklarheiten, zweitens durch firmenseitige Geheimhaltung entstehende Opazität und drittens technisch bedingte Intransparenzen, die im Charakter etwa von bestimmten algorithmischen Methoden des Machine Learnings angelegt sind. In ihrem Zusammenspiel erschweren sie die analytische Zugänglichkeit KI-basierter Anwendungen für die Sozial- und Geisteswissenschaften (vgl. Rohlfing et al. 2021: 717).

3.2.3 Smart Speaker als Intermediäre und Plattformen

Smart Speaker ermöglichen die stimmbasierte Nutzung unterschiedlicher Funktionen, die über die Systemanbieter (Amazon, Google und Apple) internetbasiert

sammenhängen und fließt somit auch in die Ausgaben der Smart Speaker und anderer Voice Assistants ein.

bereitgestellt werden. Dazu zählen erstens systeminterne Basisfunktionen, z. B. die Nutzung von Weckruf- und Timer-Funktionen, die Ansage der Uhrzeit und des Datums oder die Durchführung einfacher Rechenoperationen. Zweitens ermöglichen sie über Drittanbieter den Zugriff auf im Internet von Dritten bereitgestellte Angebote, z. B. auf Kommunikationsanwendungen, Wetterinformationen, Wissensdatenbanken, Nachrichten, Streamingplattformen und Kartendienste einschließlich der Verkehrsmeldungen. Teilweise greifen die Systemanbieter dabei direkt auf die Inhalte der Drittanbieter zu, etwa im Fall von ‚einfachen‘ Informationsabfragen (Wetter, Lottozahlen, Sportergebnisse), teilweise werden die Inhalte aber auch durch die Anwendungen der Drittanbieter vermittelt, etwa bei der Wiedergabe von persönlichen Mitteilungen aus Messenger-Diensten, der tagesaktuellen Meldungen eines bestimmten Nachrichtenanbieters oder der Suche nach Musiktiteln auf spezifischen Streaming-Diensten wie Spotify; dabei bevorzugen die Systemanbieter über die Standard-Einstellungen häufig eigene Dienste (z. B. den Streaming-Dienst „AppleMusic“ im Fall des „HomePods“ von Apple), sind aber meist nicht darauf limitiert. Drittens erlauben Smart Speaker teilweise den Zugriff auf Online-Shopping-Angebote (z. B. auf Amazon) und bieten die Möglichkeit, Einkäufe einschließlich des Bezahlvorgangs abzuwickeln und zu routinisieren. Viertens sind die Steuerung und Verwaltung von Smart-Home-Anwendungen möglich (u. a. Licht und Heizung, Jalousien, Küchengeräte und Unterhaltungsmedien). Die Geräte sind dabei internetbasiert mit den Benutzer-Konten der Anwender*innen verknüpft – sowohl mit den Accounts bei den Systemanbietern (z. B. Amazon- oder Google-Konto), als auch mit den Accounts bei eventuellen Drittanbietern (z. B. Spotify-Konto). Über die Systemanbieter-Konten sind sie auch verknüpft mit den Smartphones der Anwender*innen, die die stimmliche Bedienung der Geräte ergänzen, sowie mit den Smart-Home-Geräten. Darüber hinaus sind sie mit den Schnittstellen bzw. Anwendungen der Drittanbieter verbunden (vgl. Hector 2025).

In der interdisziplinären Medienforschung – insbesondere in den Medienwissenschaften, der Kommunikationswissenschaft und Publizistik sowie den Science and Technology Studies (STS) – sind Smart Speaker je nach fokussierter Funktion unterschiedlich konzeptualisiert und entsprechend mit verschiedenen theoretischen und methodologischen Zugriffen untersucht worden. Dabei lassen sich erstens Strömungen ausmachen, die Smart Speaker als Intermediäre gefasst haben, während sie zweitens als (Teil) digitale(r) Plattformen beschrieben wurden. Dabei sind unter unterschiedlichen Gesichtspunkten auch die Spezifika und Folgen ihrer Interfaces beleuchtet worden (dazu zählt auch die vorliegende Arbeit) – auf

weitere Veröffentlichungen dazu wird auch im Folgekapitel zur Nutzung und Aneignung der Geräte eingegangen.¹⁰⁰

Ein eher kommunikationswissenschaftlich-publizistisch orientierter Forschungsstrang bezieht sich auf die Betrachtung von Smart Speakern als *Intermediäre*. Dieser Zugang konzentriert sich auf die Funktion von Smart Speakern als Vermittler zwischen bereitgestellten „Inhalten“ auf der einen Seite und den Nutzer*innen auf der anderen Seite, wobei mit „Inhalten“ redaktionell mehr oder weniger stark kuratierte, v. a. journalistische Formen gemeint sind, die über (Medien)intermediäre aufbereitet und vermittelt werden (vgl. Weidmüller et al. 2021: 2). Untersuchungen zu Smart Speakern als Intermediäre konzentrieren sich häufig auf deren Rolle bei der Wiedergabe von Nachrichteninhalten (vgl. Weidmüller et al. 2021) und Radioangebote (vgl. Gattringer/Handel 2021) sowie den Einfluss, den der Interface-Typ auf die Wahrnehmung der Informationen haben könnte (vgl. Weidmüller/Etzrodt/Engesser 2022; Mooshammer/Etzrodt 2022; Frehmann 2023). Dabei nehmen Weidmüller et al. (2021: 7) das Verhältnis von Anbietern bzw. Redaktionen journalistischer Inhalte mit den Gerätebetreibern (insbesondere Amazon, Google und Apple) und den Charakteristika der Geräte vor dem Hintergrund des Gebots der Transparenz im Medienstaatsvertrag (MStV) in den Blick und kommen zu dem Schluss, dass in wesentlichen Teilen keine Nachvollziehbarkeit über die Kriterien für Auswahl und persönliche Aufbereitung der Inhalte bestand und nur teilweise kenntlich gemacht wurde, welche Inhalte präsentiert wurden und woher sie stammen (vgl. auch Natale/Cooke 2021). Bei der Frage, inwieweit Nutzer*innen den wiedergegebenen Nachrichteninhalten vertrauen, zeigte sich in einer Interview-Studie von Weidmüller/Etzrodt/Engesser (2022) für bestimmte abgefragte Dimensionen der Vertrauenswürdigkeit ein Zusammenhang zwischen der stimmbasierten Bedienung der Interfaces und dem Grad, dem Nutzer*innen den wiedergegebenen Inhalten vertrauten. Zu ähnlichen Ergebnissen kommt auch Frehmann (2023) in ihrer auf Online-Umfragen basierenden Studie. Auch andere kommunikationswissenschaftlich ausgerichtete Untersuchungen liefern Hinweise darauf, dass die Vermittlung von Inhalten über VUIs einen Einfluss auf die Wahrnehmung dieser haben könnten – der hybride Charakter des Austauschs mit VUIs und der Status der ‚Gesprächsbeteiligung‘ (insbesondere in Mehrparteienkonstellationen) hatte dabei dem Anschein nach eine Auswirkung auf die Verarbeitung und Bewertung der Inhalte (vgl. Weidmüller 2022; Etzrodt 2022). Dies unterstreicht die Relevanz einer gesprächsanalytischen Untersuchung der ‚Gesprächsbeteiligung‘ von VUIs, die hier als Desiderat identifiziert werden kann.

100 Zur Interface-Forschung siehe Kap. 2.1.4.

Ein anderer interdisziplinärer Forschungszweig untersucht Smart Speaker als *Plattformen* sowie als Bestandteil digitaler *Plattformsysteme* und nimmt somit die Rolle und Motive der Systemanbieter in den Blick, die den Zugriff auf Smart-Speaker-Funktionen bereitstellen und Verbindungen zu Drittanbietern herstellen (vgl. Strüver 2023b). Digitale Plattformen und ihre jeweiligen Konfigurationen können zwar je nach Zweck sehr unterschiedlich aussehen, sie dienen jedoch immer dem Ziel, verschiedene Teilnehmende oder Gruppen von Teilnehmenden zusammenzubringen und dort einen mehr oder weniger stark präfigurierten und zielgerichteten Austausch zu ermöglichen (vgl. van Dijck 2013: 29; für einen Überblick siehe van der Vlist 2022: 32–40). Plattformen sind dabei keine neutralen Instanzen, sondern durch Daten, Algorithmen, Eigentumsrechte und Geschäftsmodelle beeinflusst (vgl. van Dijck/Poell/Waal 2018: 325; siehe auch Gillespie 2018). Deren Anbieter prästrukturieren nicht nur durch die Interface-Gestaltung die Art des möglichen Austauschs (siehe unten), sondern sie verfolgen auch das Ziel einer Monetarisierung der Teilnahme der Nutzer*innen auf unterschiedliche Weisen (vgl. Srnicek 2016), primär durch die Auswertung von Nutzungsdaten. Smart Speaker sind damit Teil einer erweiterten Wertschöpfungskette, der u. a. auch ihre Nutzer*innen angehören (vgl. Crawford/Joler 2018). Die Quelle für den Gewinn der Plattformen ist dabei zu einem Großteil personalisierte Werbung, ein Geschäftsmodell zahlreicher Internet-Firmen (vgl. van der Vlist 2022: 25). Dies geht von der Annahme aus, dass die Daten Nutzungsmuster offenbaren, die Aufschluss über die private Lebensführung der Nutzer*innen sowie die konsumierten Produkte geben, die wiederum zur Bewerbung anderer Produkte genutzt und an Drittanbieter verkauft werden können (vgl. Sadowski 2020: 116; Khan 2018: 118–119). Ein komplementärer Erklärungsansatz ist, dass die aufgezeichneten Stimmeingaben und Plattformzugriffe, die bei jeder Nutzung des Smart Speakers erfasst werden, außerdem als Bestandteil der Weiterentwicklung von datenintensiven Systemen zur automatischen Spracherkennung verwertet werden (vgl. Crawford/Joler 2018). In diesem Sinne sind Smart Speaker Bestandteil digitaler Plattform-Ökosysteme.¹⁰¹

Strüver (2023b) zeigt darüber hinaus, dass es bei Amazon und dem VUI Alexa *insbesondere* die Bedeutung des Smart Speakers im Plattformsystem ist, die das Betreiben der Smart-Speaker-Anwendung für die Systemanbieter attraktiv macht: Über die Nutzung der VUIs kann Amazon als Plattformbetreiber neue Erkenntnisse über die Anwender*innen und deren Nutzungsverhalten gewinnen, mögliche neue Anwendungen oder Angebote identifizieren und

¹⁰¹ Implikationen für die Themenbereiche Datenschutz, -verwertung und Privatsphäre, die auch über die Plattformisierung hinaus reichen, werden in Kap. 3.2.6 diskutiert.

diese zugleich an den Nutzer*innen fortlaufend testen und verbessern (siehe auch Marres/Stark 2020). Wie Strüver (2023b) weiter herleitet, ist es daher auch der Anspruch der Betreiber, eine möglichst positive Nutzungserfahrung zu ermöglichen, um die Test- und Verbesserungsmöglichkeiten verschiedener Dienste und den Einblick in die Nutzungsweisen der Anwender*innen über den Betrieb des Smart Speakers weiter aufrecht zu erhalten (siehe auch Dahlgren et al. 2021). Eine Strategie, die Stellung des VUIs bei den Nutzer*innen zu stabilisieren, ist die obligatorische Einbindung des VUI in Smart-Home-Anwendungen. Die Nutzung des Smart Speakers von Amazon zur Steuerung des Smart Home verspricht nicht nur eine reibungslose Einbindung neuer Smart-Home-Anwendungen in den Haushalt und dessen Smart-Home-Konfiguration (vgl. Strüver 2023a), sondern auch eine fortlaufend komfortable und ‚unsichtbare‘ Bedienung dieser Geräte (vgl. Strüver 2023b). Amazon stellt dazu eine möglichst große Kompatibilität mit Drittanbietern sicher und kann so als feste Instanz im Smart Home umfassend Informationen über die Nutzung der Dienste und mögliche Verbesserungspotenziale gewinnen. Zugleich werden zahlreiche Alltagspraktiken der Haushaltsmitglieder unter Beteiligung der zentralen Smart-Home-Steuerung vollzogen, was sie anfällig für die Beeinflussung und Normierung durch die Plattformbetreiber macht (vgl. Goulden 2019). Damit sind Smart Speaker nicht nur Bestandteil eines Plattform-Ökosystems, sondern auch selbst nicht-neutrale Plattformen innerhalb der Haushalte ihrer Anwender*innen.

3.2.4 Nutzung und Aneignung von Smart Speakern

Empirische Studien zur Nutzung und zur longitudinalen Perspektive auf die Nutzung und Aneignung von Smart Speakern sind bislang noch rar. In einer repräsentativen Befragung zwischen Oktober 2023 und September 2024 in Deutschland gaben 25 Prozent der Befragten an, einen Smart Speaker zu besitzen (vgl. Statista 2025b); dieser Wert ist in vergleichbaren Umfragen in den vergangenen fünf Jahren stabil. Im Folgenden sollen auf Grundlage empirischer Arbeiten die zuletzt beobachteten Nutzungsweisen, mögliche Einflussfaktoren und Aneignungs- bzw. Domestizierungsprozesse¹⁰² in den Haushalten herausgestellt werden, um dabei die mit der vorliegenden Arbeit zu füllende Forschungslücke noch präziser zu fassen. Dabei wird nicht übersehen, dass etliche Studien sich mit VUIs auseinandersetzen, aber nicht spezifisch Smart Speaker in den Blick nehmen. So beschäftigt sich etwa die Interview-Studie von Luger/Sellen (2016) mit „Conversational

¹⁰² Siehe Kap. 2.3.

Agents“. Es wurden dabei aber vornehmlich Nutzer*innen von Siri, Google Now und Cortana befragt, die diese nicht über Smart Speaker, sondern über andere Endgeräte wie Smartphones, Tablets und Computer bedienen (vgl. Luger/Sellen 2016: 5289). Die herausgearbeiteten primären Nutzungsszenarien wie z. B. eine Tastatureingabe zu vermeiden, weil man unterwegs ist (vgl. Luger/Sellen 2016: 5291), sind insofern nicht übertragbar. Gleichwohl könnte der festgestellte „gap between user expectation and system operation“ (Luger/Sellen 2016: 5295) durchaus auch bei Smart Speakern eine Rolle spielen (siehe dazu Strüver 2020).¹⁰³

Etliche Studien zu Smart Speakern haben zunächst Potenzialanalysen vorgenommen, die basalen Funktionen vorgestellt und mögliche Anwendungsfälle diskutiert (vgl. Goksel-Canbek/Mutlu 2016; Bedford-Strohm 2017; Hoy 2018). Eine auf Smart Speaker scharfgestellte Interview-Studie mit 19 Interviews von Nutzer*innen und unter Einbezug von 170 Protokollaufzeichnungen, die in der zugehörigen Smartphone-App hinterlegt werden,¹⁰⁴ legen Ammari et al. (2019) vor. 18 der 19 Befragten hatten einen Smart Speaker zu Hause und Anspruch der Studie war u. a. die Untersuchung der alltäglichen Nutzungskontexte. Diese sind vorhersehbarerweise abhängig von den konkreten Nutzer*innen, es zeichnet sich jedoch hier deutlich ab, dass ‚einfache‘ Anwendungsfälle mit routinisiertem Charakter (z. B. Abspielen von Musik, Timer beim Kochen, Erinnerungen oder Wetter-Abfragen) dominieren. Einen wesentlichen Anteil nimmt auch die Steuerung anderer Smart-Home-Anwendungen ein, die ebenfalls nach festen Mustern erfolgt, so z. B. das Licht ein- bzw. ausschalten oder die Heizung regulieren (vgl. Ammari et al. 2019: 15). Die Dominanz dieser Nutzungstypen bestätigt sich auch einer Befragungsstudie von Lopatovska et al. (2019). In neun verschiedenen Haushalten wurden über einen Zeitraum von jeweils vier Tagen insgesamt 136 Anfragen an einen Amazon *Echo* gestellt, von denen 86 Wetterabfragen, Musikwiedergabe und die Kontrolle anderer Smart-Home-Geräte waren.

Das Spektrum von Anwendungsfällen, das die ebenfalls interviewbasierte, qualitativ ausgerichtete Studie von Brause/Blank (2020) mit dem Konzept der „Use Genres“¹⁰⁵ aus der Domestizierungstheorie (siehe dazu Bakardjieva 2005) un-

103 Zum Aufbau von Nutzungserwartungen an Conversational Agents in der Werbung siehe auch die Analyse von Hennig/Hauptmann (2019).

104 Siehe dazu Kap. 3.3 sowie Habscheid et al. (2021).

105 „Use Genres“ sind ein Konzept aus der Domestizierungsforschung, mit dem Nutzungsweisen von Technologien aus ihrer konkreten, praktischen Situierung heraus beschrieben und verstanden werden (vgl. Bakardjieva 2006: 73). Solche ‚Gattungen‘ der Nutzung sind in diesem Sinne nicht etwas von den Herstellern oder durch die Technologie Angelegtes, sondern entstehen dynamisch und in Relation zu den Handlungen der Nutzer*innen (vgl. Brause/Blank 2020: 753).

terscheidet, ist deutlich weiter aufgefächert. Es gehören dazu „companionship, self-control and productivity, health care support, better sleep, peace of mind and improved accessibility“ (Brause/Blank 2020: 759) sowie Sicherheitsvorkehrungen und sogar Ausspionierungen in einer Ehe. Besondere Beachtung findet eine räumlich verteilte Nutzung, d. h. Anwendungsfälle, die über den direkten Austausch zwischen dem Menschen und der Instanz des Smart Speakers hinaus wirken – so z. B. die Steuerung von Smart Home-Anwendungen, die Folgen für den Status anderer, vernetzter Geräte haben, oder das Erhalten von Nachrichten auf dem Mobiltelefon (vgl. Brause/Blank 2020: 757). Die Autor*innen entwickeln die „Use Genres“ aus den Interviewdaten heraus und liefern insofern sehr wertvolle Hinweise auf mögliche Verwendungsweisen, die auch in den im Analyseteil dieser Arbeit untersuchten Daten eine Rolle spielen. Dass die Nutzer*innen selbst diese Kategorien relevant machen, macht die Bestandsaufnahme im Sinne einer ethnomethodologischen Vorgehensweise besonders wertvoll.¹⁰⁶

Die Inbetriebnahme sowie die Integration anderer Geräte und Konten sind nicht immer unproblematisch und störanfällig, wie auch anhand der später präsentierten Daten deutlich wird. In Interviews der verschiedenen Studien wird dargestellt, dass die Übersichtlichkeit in den zugehörigen Smartphone-Apps verschiedener Hersteller nicht unbedingt gegeben ist; ferner wird berichtet, dass bei mehreren Smart Speakern die räumliche Zuordnung Probleme verursacht und z. B. das Licht im falschen Raum ausgeschaltet wird (vgl. Ammari et al. 2019: 16). Die fortlaufende, routinisierte Nutzung über einen Zeitraum von drei bis sechs Monaten hinweg scheint jedoch insgesamt bei den genannten, ‚eingespielten‘ und in die Alltagspraktiken der Nutzer*innen übergegangenen Anwendungsfällen zu liegen, die primär eher ‚einfache‘ Nutzungsszenarien umfassen. In Befragungen zeigte sich zudem eine Tendenz, dass die Nutzungshäufigkeit des Smart Speakers über längere Zeit nach Anschaffung eher abnimmt (vgl. Lopatovska et al. 2019: 991; siehe auch Barthel/Helmer/Reineke 2023), andere nutzer*innenzentrierte Faktoren (insbesondere Alter und Technologieaffinität) schienen jedoch eher geringe Auswirkungen auf die Nutzung von Smart Speakern zu haben (vgl. Lopatovska et al. 2019: 990).

Die Tageszeit als Einflussfaktor auf die genutzten Anwendungen diskutieren Bentley et al. (2018) in einer empirisch breit angelegten Studie in 88 verschiedenen Haushalten, in denen über einen durchschnittlichen Beobachtungszeitraum von 110 Tagen insgesamt 65.499 Aktionen mit Smart Speakern ausgewertet wurden. Sie bestätigen, dass Musikwiedergabe, Wissensabfragen und Smart-Home-Anwendungen zu den am häufigsten genutzten Kategorien gehören. Sie bestätigen

106 Siehe dazu Kap. 4.1.

auf dieser Basis allerdings nicht die hohe relative Häufigkeit von Wetterabfragen. Auch wenn die Gesamtverteilung der genutzten Kategorien im Tagesverlauf recht konstant bleibt, beobachten die Autor*innen Höhen für Wetterabfragen am Morgen zwischen sechs und neun Uhr und für die Bedienung von Smart-Home-Elementen (insbesondere wohl Licht und Heizung) in den Abendstunden (vgl. Bentley et al. 2018: 7). Uhrzeitabfragen haben eine Spitze zwischen drei und neun Uhr am Morgen „likely as users lay in bed wondering how many hours they have left before they need to get up“ (Bentley et al. 2018: 7). Am Wochenende findet außerdem sehr viel häufiger eine Nutzung der Smart Speaker statt. Weitere demografische Einflussfaktoren wie Alter oder die Größe des Haushalts werden zwar auch in dieser Studie diskutiert, die Nutzungsszenarien bleiben aber im Wesentlichen konstant (siehe dazu und für die folgenden Angaben Bentley et al. 2018: 16–18). In der longitudinalen Betrachtungsweise zeigte sich, dass die Nutzungsszenarien überwiegend konstant verteilt blieben, dass aber Smart-Home-Anwendungen tendenziell zunehmen und die Musiksteuerung demgegenüber tendenziell abnimmt. Die Länge der Inputs (und somit ein Indikator für die Komplexität derer) nimmt über die Zeit zu, korreliert aber auch mit der Zunahme von Wissensabfragen, für die längere Inputs notwendig sind, insbesondere im Vergleich zu kurzen, routinierten Befehlen wie Uhrzeitabfragen. Die Daten dieser größeren Studie belegen die bereits zuvor geäußerten Vermutungen noch einmal eindrücklich: Zusammenfassend sind routinierte Anwendungsfälle mit kurzen und meist zweizügigen Ein-/Ausgabe-Sequenzen fest in den Alltag integriert und weisen sogar über den Tagesverlauf eine gewisse Stabilität auf. Neue Anwendungen werden selten ausprobiert, einzig die Steuerung von Smart-Home-Anwendungen nimmt nach der Anschaffung tendenziell weiter zu.

Die bisher genannten Studien bezogen sich auf den US-amerikanischen Raum. Zur Nutzung und Aneignung von Smart Speakern forschten für den deutschsprachigen Raum mit einem qualitativ-ethnografischen Untersuchungsdesign Pins et al. (2020), die u. a. Medientagebücher einsetzten, um die Nutzung von Smart Speakern empirisch zu erforschen. Sie bestätigen in kleinerem Rahmen die Ergebnisse zur Nutzung von Smart Speakern von Ammari et al. (2019) dahingehend, dass auch in deren Daten die ‚einfachen‘ Anwendungsfälle deutlich überwogen und kompliziertere Nutzungskontexte mit mehreren Zügen und neue Skills in vielen Haushalten nicht ausprobiert wurden. In zehn Interviews, die auf Basis der Grounded Theorie (Strauss/Corbin 2010) ausgewertet wurden, zeigte sich, dass die Smart Speaker „Bestandteil des häuslichen Lebensmittelpunktes“ wurden (Pins et al. 2020: 353), indem sie an zentralen Stellen und in gemeinsam von allen Haushaltsmitgliedern genutzten Räumen aufgestellt wurden; dort sind sie „als fester Bestandteil in ihrem Alltag eingebettet“ (Pins et al. 2020: 357). Außerdem zeigte sich eine Anpassungsleistung an die VUIs seitens der Befragten, die

angaben, die Formulierung der ‚richtigen‘ Stimmeingaben und eine bestimmte Sprechweise über die Zeit erlernt zu haben; in einem befragten Haushalt gibt es eine geteilte Liste mit Befehlen für bestimmte Zwecke (Pins et al. 2020: 354). Smart Speaker lassen sich zudem anpassen: Individuelle Befehle können programmiert und mit bestimmten Funktionen verknüpft werden (z. B. „Schicht im Schacht“ schaltet eine bestimmte Lampe aus). Pins et al. (2020: 356) konstatieren, dass „die Interaktion sich häufig auf kurze, geschlossene Sequenzen reduziert – hauptsächlich deshalb, weil Nutzer an komplexeren Interaktionen scheiterten“. Sie stellen eine Lücke zwischen einem prinzipiell breiten möglichen Nutzungsspektrum fest, das jedoch nur zu einem Bruchteil ausgeschöpft wird, und benennen als limitierenden Faktor klar die Sprachsteuerung, die erstens zu häufig keine gute Erkennung der Sprache liefert und zweitens zu wenig Freiheiten in der Auswahl der benötigten Inhalte lasse (vgl. Pins et al. 2020: 356; siehe auch Habscheid/Hector/Hrncal 2025).

Zum Potenzial der Protokolldaten der Assistenzsysteme arbeiten Habscheid et al. (2021) und stellen in ihrer qualitativ-explorativen Untersuchung heraus, dass dieser Datentyp zwar einerseits Aufschlüsse darüber geben kann, wie der Austausch zwischen Nutzer*innen von VUIs und den Plattformen konfiguriert ist. Es entsteht ein Nutzungsprotokoll, das bestimmte Datentypen enthält und zur Bearbeitung und Löschung vorsieht, während sich andererseits der Export des Protokolls einschließlich seiner Audio-Mitschnitte umständlich gestaltet. Für eine detaillierte Betrachtung der Einbettung von VUIs in die soziale Praxis greift eine ausschließlich auf diesen Protokollen basierende Analyse allerdings zu kurz: die Erfassung durch das Protokoll ist dafür zu bruchstückhaft (vgl. Habscheid et al. 2021; Hector 2025: 75). Diese Erkenntnisse werfen wichtige Anschlussfragen auf, zu deren Beantwortung die vorliegende Arbeit einen Beitrag leisten will: Mit welchen Praktiken passen sich die Nutzer*innen den Anforderungen der Smart Speaker an, welche übernehmen sie in ihren alltäglichen Gebrauch? Welche Strategien zur Fehlerbehebung sind empirisch beobachtbar? Welcher Status wird den Geräten in laufenden sozialen Interaktionen zugeschrieben? Die Studie von Pins et al. (2020) zeigt auf Basis der Interviews, dass genau diese Punkte bei der Bedienung der Smart Speaker eine Rolle spielen und die Technologie die Alltagspraktiken der Nutzer*innen verändern kann. Als explorative Studie lieferte sie insofern sehr wertvolle Impulse für die folgende Arbeit; der genaue Vollzug und die Einbettung in den Alltag jedoch kann nur auf Basis von Aufnahmen *in situ* untersucht werden.

Dieser Aufgabe gehen nur einige wenige Arbeiten nach, deren Befunde nachfolgend zusammengefasst werden sollen. Porcheron et al. (2018) explorieren in ihrer sehr einschlägigen Studie, auf welche Weise Smart Speaker in die sozialen und interaktionalen Zusammenhänge des Alltags eingebettet werden. Die Unter-

suchung findet auf Basis von ein-monatigen Audio-Aufnahmen in fünf Haushalten und insgesamt ca. sechs Stunden Audiomaterial statt, das mit Hilfe eines sog. Conditional Voice Recorders (CVR) zusammengetragen wurde.¹⁰⁷ Diesem Material nähern sie sich mit einer konversationsanalytischen und an die HCI und CSCW angebundenen ethnomethodologischen Perspektive. Dabei sind die Schlüsselergebnisse eher auf eine Design-Perspektive zugeschnitten und geben konkrete Ratschläge an zukünftige Design-Projekte von VUIs (vgl. Porcheron et al. 2018: 10), die für die vorliegende Studie weniger einschlägig sind; von besonderem Interesse sind aber die Detail-Analysen. In diesen wird unterschieden zwischen der Einbettung von Amazons *Echo*¹⁰⁸ in die häuslichen Alltagsaktivitäten auf der einen und der Einbettung in die sequenzielle Organisation auf der anderen Seite. Die Adressierung und Einbindung der Nutzung des Smart Speakers in den häuslichen Alltag erfolgt, so konstatieren Porcheron et al. (2018: 5), „with relative ease through everyday talk“, was v. a. auf der Annahme fußt, dass etwa das gemeinsame Abendessen in einer Familie ohnehin eine „multi-activity“-Situation ist, bei der sich konstant mehrere Handlungsstränge überlappen und insofern die Adressierung und Einbindung eines VUI einen von mehreren dieser Stränge darstellt und nicht heraussticht. Sie beobachten, dass bei der Aushandlung von Zugriffsrechten auf den Smart Speaker kompetitive Momente beobachtet werden können, die konversationellen Methoden der Rederechtsaushandlung gleichen. Das Herstellen von „accountability“ (Garfinkel 1967: 34) erfolgt, ohne eine hohe Spezifität für genau diese Art der Darstellungen, der „accounts“ (Garfinkel 1967: 2), feststellen zu können, die sich auf Äußerungen zu Adressierungen eines Smart Speakers beziehen: „such utterances are treated in similar kinds of ways to the ways that all social actions are treated: as accountable to the situation they are in“ (Porcheron et al. 2018: 6).¹⁰⁹

Deutlich spezifischere Eigenschaften scheint die Einbindung der Geräte Porcheron et al. (2018) zufolge nicht in die Aktivitäten im Allgemeinen, sondern in die sequenzielle Gestaltung des *turn-by-turn talk* aufzuweisen. Für die Gestaltung der Anfrage an den Smart Speaker (etwa eine Frage oder ein Befehl) ist zunächst auffällig, dass in den Porcheron et al. vorliegenden Daten die Äußerung des Aktivierungsworts als eine selbstständige *turn-constructural unit* betrachtet wird,

107 Ein Conditional Voice Recorder ermöglicht es, Audioaufnahmen von VUI-Dialogen zu erheben, auf denen nicht nur die Nennung des Aktivierungsworts und der unmittelbar folgende Teil, sondern auch eine vordefinierte Zeitspanne vor- und nachher aufgezeichnet wird, siehe dazu Kap. 5.2.

108 Andere Hersteller bzw. Modelle werden nicht berücksichtigt.

109 Zu „accountability“ bzw. „accounts“ als ethnomethodologische Grundbegriffe siehe Kap. 4.1 sowie Bergmann/Meyer (2021b).

nach der ein Sprecher*innenwechsel stattfinden kann. Die kurze Pause zwischen Aktivierungswort und Anfrage wird mehrfach für andere Äußerungen genutzt, die keinen Input an den Smart Speaker darstellen; diese Beobachtung erinnert an die von Pitsch et al. (2017) vorgelegten Befunde zur interaktionalen Dynamik bei der Bedienung von Robotern, in denen der Interaktionsraum ebenfalls auf Phasen der Interaktion *vor* dem Roboter ausgedehnt wurde, in denen die Nutzer*innen den Roboter selbst aber nicht adressieren. Die Pause kann außerdem genutzt werden, um die Steuerung des Smart Speakers von einer anderen Person zu übernehmen, d. h. selbst einen Input an das VUI zu formulieren, bevor der ursprünglich aktivierende Interaktant dies kann (Porcheron et al. 2018: 7). Darüber hinaus scheint gemeinsames Herstellen von Stille für die Produktion einer Anfrage ein übliches Verfahren zu sein, bei der der *turn-by-turn talk* vorübergehend ausgesetzt und nur der Input formuliert wird. Dies trägt auch zu einer besseren Hörbarkeit der Ausgabe bei. Bei Fehlverarbeitung der Anfrage lässt sich ferner eine Art „sequenzielle Kollaboration“ beobachten, in der verschiedene Teilnehmer*innen durch prosodische oder lexikalische Variation versuchen, die gewünschte Funktion aufzurufen (vgl. Porcheron et al. 2018: 8–9).

Im Umgang mit den Outputs konzentrieren sich Porcheron et al. (2018: 7–9) auf Störungen der Geräte. So werden, ähnlich wie in zwischenmenschlichen Interaktionen, zu lange Pausen als Hinweise auf Probleme interpretiert, wobei die Toleranz für Zögerungen im Austausch mit VUIs deutlich höher zu sein scheint. Gibt das VUI eine unerwünschte Ausgabe zurück oder formuliert, dass die Eingabe nicht verstanden wurde, kann ein „mismatch“ (Porcheron et al. 2018: 8) zwischen der Fehlermeldung des Geräts und der interaktionalen Bearbeitung dieser beobachtet werden. Letztere konzentrieren sich als Reparaturversuche immer wieder auf prosodische und lexikalische Variationen, wobei die prosodischen Strategien im Rahmen der Arbeit von Porcheron et al. (2018) nicht betrachtet werden. Insgesamt kommen sie zu dem Schluss, der Austausch mit VUIs sei „fundamentally different from human interaction“ (Porcheron et al. 2018: 9). Der Unterschied ergebe sich primär daraus, dass sich der sequenzielle Verlauf des Dialogs nicht aus der Beziehung der jeweiligen Sequenzteile aufeinander und der darin liegenden Interpretation der vorherigen Äußerungen entfaltet, sondern vielmehr durch die Eingabe-Ausgabe-Struktur vorgegeben ist (vgl. Porcheron et al. 2018: 9).

Die Arbeit von Porcheron et al. (2018) liefert, auch wenn sie die Ableitung von Design-Empfehlungen aus diesen Analysen fokussiert, entscheidende Anknüpfungspunkte für weitere Arbeiten, die die Nutzung von Smart Speakern im praktischen Vollzug beleuchten. So beschreiben Beneteau et al. (2019: 4–5) auf Basis einer Audio-Erhebung bei 10 Familien sechs verschiedene Reparaturstrategien: prosodische, artikulatorische, syntaktische und semantische Anpassungen sowie gesteigerte Lautstärke und Wiederholungen. Die Autor*innen dieser Unter-

suchung analysieren drei Fallbeispiele mit einer ebenfalls konversationsanalytischen Mentalität, um diese Strategien zu illustrieren, und empfehlen zukünftigen VUI-Designs, Fehler und Reparaturen als Fokus der Entwicklung festzulegen (vgl. Beneteau et al. 2019: 10), sowohl in den Bereichen Code-Switching und diskursiver Ordnung wie auch in der Teilnehmenden-Struktur. So raten die Autor*innen dazu, seitens des VUI eine stärkere Dialogsteuerung und präzisere Nachfragen zu realisieren und dabei das Primat der Natürlichsprachigkeit abzulegen (um z. B. nach Keywords zu fragen, die vom VUI einfacher verarbeitet werden können) (vgl. Beneteau et al. 2019: 11). Auch empfehlen die Autor*innen, dass das VUI aktiv die Teilnehmenden-Struktur der Dialogbeteiligten abfragen und steuern könnte, um die Diversität der angewendeten Reparaturstrategien zu erhöhen; diese Empfehlung leitet sich aus der Beobachtung ab, dass bei Beteiligung mehrerer Personen an Smart-Speaker-Dialogen auch eine größere Diversität im Hinblick auf die Anwendung von Reparaturstrategien bestehe (siehe auch Beneteau et al. 2020a; Beneteau et al. 2020b).

Habscheid (2022) diskutiert auf der Grundlage des auch in der vorliegenden Arbeit verwendeten Datenkorpus aus Video-Aufzeichnungen von Inbetriebnahmesituationen und Audio-Aufzeichnungen von der regelmäßigen Nutzung der Systeme¹¹⁰ die wechselseitigen Anpassungen von Nutzer*innen und VUI aneinander. Dabei fokussiert er auf Situationen des Wechsels zwischen soziotechnischem Dialog und „Meta-Interaktionsraum“ (Habscheid 2022: 168), wobei in Letzterem die Nutzer*innen über die VUIs und den Gebrauch reflektieren, während im ersten der Austausch mit dem VUI im Vordergrund steht. Er stellt heraus, dass Sequenzialität – über Adjazenzpaare hinaus verstanden als „as a fundamental resource of meaning constitution“ (Habscheid 2022: 191) – eine Herausforderung für die Geräte darstelle. Die indexikalische soziale Interaktion der Nutzer*innen zu erfassen bleibt für VUIs bisher eine Grenze der Leistungsfähigkeit und insofern die Anpassungsleistung der Nutzer*innen elementar für das Funktionieren des soziotechnischen Dialogs – soziale Interaktion bleibe eine Simulation der Teilnehmenden und dies aufzudecken ist eine der wesentlichen Vorteile einer praxeologischen Analyse (siehe auch Alač et al. 2020; Hector 2022; Habscheid/Hector/Hrncal 2025).

Eine praxeologische Perspektive wird auch bei Habscheid/Hector/Hrncal (2023) angewendet, um einen gesprächsanalytisch fundierten Beitrag zur Debatte um Agency menschlicher und nicht-menschlicher Entitäten zu leisten. In den qualitati-

¹¹⁰ Die Daten wurden mit einem von Porcheron et al. (2018) adaptierten CVR erhoben, siehe dazu ausführlicher Hector et al. (2022) sowie Kap. 5.

ven Analysen, die sich wiederum auf Beispiele aus dem bereits zuvor erwähnten Datenkorpus aus Inbetriebnahme- und Nutzungssituationen beziehen, kann gezeigt werden, dass Agency zwischen Nutzer*innen und VUIs auch als „dynamic practical accomplishment“ (Habscheid/Hector/Hrncal 2023: 23) betrachtet werden kann. Agency und ihre Aushandlung ist dabei aber kein Nullsummenspiel: Die Steigerung der Agency der Nutzer*innen muss nicht mit einem Verlust an Agency auf Seiten des VUIs einhergehen und umgekehrt (vgl. Habscheid/Hector/Hrncal 2023: 24). Eine einzelne Situation kann sowohl Verlust wie auch Steigerung von Agency auf jeder der beiden Seiten bedeuten – zugleich ist die Handlungsmacht der Nutzer*innen sowohl untereinander als auch mit dem des VUIs verbunden, wenn auch nicht determiniert. Der Blick auf den lokalen, situationalen Kontext und die soziale Konfiguration der Beteiligten untereinander ist aber eine wesentliche Einflussgröße. Herausgefordert wird diese allerdings durch die teilweise intransparente und im Hintergrund stattfindende Verwertung der Daten, wie auch Waldecker/Hector/Hoffmann (2024) herausarbeiten. Die Nutzer*innen sind dann zwar *in der Nutzungssituation* möglicherweise mit Agency ausgestattet und beherrschen etwa den konversationellen Floor (vgl. Waldecker/Hector/Hoffmann 2024: 11). Das Interface suggeriert dies zusätzlich über Möglichkeiten der Löschung einzelner Aufnahmen aus dem Protokoll (vgl. Habscheid et al. 2021: 19). Die Nutzer*innen sind aber zugleich nicht mehr Teil des praktischen Vollzugs der Auswertung – Grenzen zwischen Öffentlichkeit und Privatheit werden in diesem Sinne neu verhandelt (vgl. Waldecker/Hector/Hoffmann 2024: 2). Die Autor*innen konstatieren zwei verschiedene Typen von „Agencies“: Einerseits sind die Nutzer*innen „able to retain control or a sense of superiority even in complicated situations“ (Waldecker/Hector/Hoffmann 2024: 13). Andererseits ist schon die Untersuchung von Agency deutlich erschwert, wenn es um Datenpraktiken geht, die im Zuge der Nutzung und weit darüber hinaus im Hintergrund stattfinden. Auch hier konnte eine praxeologische Analyse helfen, diese Differenzen aufzudecken und mit empirischem Material zu unterfüttern.

3.2.5 Smart Speaker und Gender

Ein wachsendes Forschungsfeld beschäftigt sich interdisziplinär mit Smart Speakern vor dem Hintergrund der sozialen Konstruktion von Gender. Ausgangspunkt ist in vielen Arbeiten der Umstand, dass die meisten VUIs und insofern auch Smart Speaker per Default eine weiblich klingende Stimme eingestellt haben und die zugehörigen Interface-Personae zudem weiblich gelesen sind (etwa Alexa und Siri). Dies ist aus verschiedenen Disziplinen kritisch diskutiert und analysiert worden (vgl. Natale/Cooke 2021: 1010; Woods 2024: 98–130). Bereits in einer frühen Arbeit, erst ein

Jahr nach dem Start von Apples Sprachassistent Siri, kommt Both (2012: 130) zu dem Schluss, „dass Siri überwiegend auf weiblich konnotierte Register zurückgreift“ (siehe auch Both 2011). Both nimmt die „Geschlechter-Performanz auf der Ebene des Dialogs“ (Both 2012: 129) in den Blick und untersucht sprachliche Register.¹¹¹ Im Hinblick auf die mit einem „weiblichen Stil“ programmierte Dialogizität und die weiblich gestaltete Anthropomorphisierung der VUI scheinen die Ausführungen zunächst überzeugend und liefern wertvolle Hinweise etwa auf einen „passiven Eindruck“, eine kooperative Haltung und Bescheidenheit und Gefühlsbetontheit als (vermeintlich) weibliche Merkmale. In einer späteren Arbeit konstatiert Both (2014: 109): „Siri’s conversational style draws on a stereotypical female image of altruistic and cooperative behavior“. Boths Darstellung zu den weiblichen sprachlichen Registern oder dem ‚konversationellen Stil‘ entbehrt jedoch einer soliden theoretischen Basis zu femininen Stereotypen. Es mangelt ferner an einer dialogischen bzw. interaktionistischen Perspektive, in der ‚Doing Gender‘ als gemeinsam hergestellte Bezugskategorie der Beteiligten betrachtet wird (West/Zimmermann 1987): Zwar argumentiert Both, dass die sprachlichen Register „eingeschrieben“ sind, doch ob diese tatsächlich von den Programmierer*innen auf der einen und von den Nutzer*innen auf der anderen Seite als Gender-Merkmal relevant gemacht werden, kann auf diese Weise nicht ermittelt werden und muss Gegenstand weiterer Forschungen sein.

Gleichwohl scheint nicht nur auf Basis dieser Studie einiges darauf hinzudeuten, dass durch die Auswahl der weiblich klingenden Stimme Geschlechterstereotype manifestiert und evtl. sogar ausgeweitet werden, wie frühe Arbeiten aus einer feministischen Perspektive (Draude 2006; Gustavsson 2005; Weber/Bath 2007) schon für ähnliche digitale Systeme feststellen. Es besteht das Potenzial, dass genderspezifische Stereotype zur Aufteilung von Arbeiten im Haushalt und Care-Arbeit sich durch die Zuordnung einer weiblichen Stimme, aber auch durch die Gestaltung von Werbung reetablieren (vgl. Hennig/Hauptmann 2019; Strengers/Kennedy 2020; Sadowski/Strengers/Kennedy 2021): „the 21st-century smart home maintains current gendered stereotypes by promising wife or housekeeper functionality“ (Strengers/Nicholls 2018: 75). Diese auch von Both (2012; 2014) untersuchte „Einschreibung einer geschlechtsspezifischen Arbeitsteilung“ (Both 2012: 132) verweist darauf, dass die weiblich gelesene Personifikation der Geräte mit den übernommenen Aufgaben im Service-, Haushalts- und Care-Bereich

¹¹¹ Auch wenn Both (2012) diesen Begriff hier verwendet, kann hier nicht von sprachlichen Registern im Sinne der linguistischen Anthropologie gesprochen werden, wie sie etwa Agha (2004) versteht; diese werden auch nicht methodisch untersucht (siehe dazu Kap. 7). Both selbst verwendet mit Verweis auf Lübke (2005) auch den Stilbegriff, ohne sprachwissenschaftliche Implikationen hier zu reflektieren.

korrespondiert, die traditionell von weiblichen Personen ausgeführt werden (vgl. Gustavsson 2005). Sie sind insofern als servile, hilfsbereite und zurückhaltende Assistentinnen inszeniert, die sich nur zu den ihr übertragenen Aufgaben äußern, sich aber darüber hinaus nicht ins Geschehen involvieren und insofern an Hauspersonal erinnern (siehe auch Strüver 2020: 1–10; Habscheid et al. 2021; Dickel/Schmidt-Jüngst 2021; Waldecker/Volmar 2022: 172). So manifestieren und erweitern sich, so die Argumentation, durch die Nähe zu menschlichen, weiblichen Stimmen, Gender-Stereotype (vgl. Natale/Cooke 2021: 11) und gegenderte Normen und Phantasien (vgl. Strengers/Sofoulis 2024; Woods 2024: 98–130).

Natale/Cooke (2021: 1010) führen aus, dass diese Anthropomorphisierung im Dienst von kapitalistischen Interessen steht (siehe auch Natale 2023). Um Sprachassistenten marktfähig zu machen, muss die neue Technologie in bekannte Repräsentationen und Stereotypen eingebettet werden. Dabei greifen die Designer*innen auf diese Kategorien zur Aktivierung von Identitätsmarkern und sozialen Skripts zurück, die einen Roboter „abbilden“ können (vgl. Clark/Fischer 2022), auch wenn sie diesen Prozess selbst nicht unbedingt benennen können (vgl. Sweeney 2021: 153–155; siehe auch Young 2019: 117). Er entfaltet seine Wirkung auch erst im Zusammenspiel zwischen den Nutzer*innen und den stereotypisch weiblichen Charakteristika (Name, Aufgabenbereich, Konversationsstrategien, Stimme), die als eine Art weiblich gelesene Verkörperung eines Algorithmus betrachtet werden können (vgl. Phan 2017: 31). Dass die Interessen des Marktes mit diesen Verfahren durchaus vertreten werden könnten, zeigen Experimente, in denen Nutzungspräferenzen für männlich oder weiblich klingende Stimmen im Dialog mit VUIs getestet wurden; dabei wiesen die Proband*innen geschlechtsunabhängig eine Präferenz für weibliche Stimmen auf (vgl. die Übersicht von Sweeney 2021: 154). Solche Studienergebnisse, die bereits seit den 90er-Jahren bekannt sind, finden Eingang in die Design-Praktiken, tradieren sich dort und führen zu einer Art „cultural ‚common sense‘ design practice, obscuring their linkages to historically specific and socially-produced systems of oppression“ (Sweeney 2021: 155). Die Folge kann somit eine Dekontextualisierung und Depolitisierung der historisch-kulturellen Realität von Haushaltsdiensten sein (vgl. Phan 2019: 4).¹¹² Auf die Gefahren eines so möglicherweise verursachten gesellschaftlichen Schadens weisen Loideain/Adams (2020) hin und diskutieren potenziell anwendbare

112 Für eine historische Verlängerung der Geschichte von Dienern und Dienstboten bis ins Digitale siehe Krajewski (2010), der in seiner Kultur- bzw. Mediengeschichte des Dieners argumentiert, dass – nach dem Verschwinden menschlicher Diener und der vermehrten Übertragung von Aufgaben an Maschinen im 20. Jahrhundert – derzeit ein Transformationsprozess stattfindet, in dem „den von ‚unbeseelten‘ Dingen verrichteten Diensten derzeit wiederum ein Subjektcharakter zurückerstattet wird“ (Krajewski 2010: 19).

Regulierungsverfahren zur Begrenzung dessen durch EU-Recht und konkret durch die DSGVO (siehe auch Zarcone/Leschanowsky 2023: 172–173).

Betrachtungen aus einer solchen kritischen Perspektive zum Design von VUIs bzw. Smart Speakern bleiben nicht bei Fragen von Gender stehen, sondern nehmen auch andere Domänen wie Rasse, sozialen Status, Bildungsniveau u. a. in den Blick (vgl. Strengers/Nicholls 2018; Phan 2019: 21–23; Schiller/McMahon 2019; Natale/Cooke 2021). Außerdem ist der Phänomenbereich etwa von den Feminist Technology Studies auch auf die Marketing- und Werbekonzepte für Smart-Home-Anwendungen und Technologie-Einsatz in häuslichen Umgebungen und deren Wechselspiel mit Gender Scripts in den Blick genommen worden (vgl. Chambers 2020; Sadowski/Strengers/Kennedy 2021; Woods 2024: 102–105). Eine Inhaltsanalyse von Werbematerial und Marktberichten zu Smart-Home-Anwendungen kam zu dem Ergebnis, dass häufig Frauen nicht als kompetente Anwenderinnen dieser Geräte dargestellt bzw. inszeniert werden, aber als diejenigen, die durch ihre Integration in den Alltag entlastet werden (vgl. Chambers 2020: 314). Tatsächlich steigen vielmehr die gesellschaftlichen, kulturell geprägten Erwartungen an Care-Arbeit und das häusliche Umfeld (vgl. Woods 2024: 119).

Mit Blick auf den linguistischen Fokus der Untersuchung können nicht alle diese Aspekte detaillierter behandelt werden. Als Anschlussfrage an diesen Diskurs ergibt sich jedoch unbedingt diejenige nach der kommunikativen Konstruktion von Gender im Austausch mit den Smart Speakern. „Voice is a key marker for identity and can be inflected with assumed racial, gendered, and other meanings“ (Woods 2024: 129). Auf die Stimme soll insofern ganz besonders da eingegangen werden, wo sie sprachlich hervortritt, etwa bei der Auswahl der Klangfarbe im Zuge der Inbetriebnahme. Dabei sind sowohl die direkten Dialoge mit dem VUI wie auch die vorbereitende und anschließende zwischenmenschliche Interaktion zwischen ko-präsenten Personen relevant. Es könnten unter diesen Gesichtspunkten sprachliche Verfahren identifiziert werden, mit denen die durch das Design eingeschriebenen Gender-Stereotype bestätigt, verworfen oder anderweitig thematisiert werden. So soll der bisher skizzierte Diskurs auf einer empirischen Basis mit einer praxeologischen Perspektive angereichert werden, um so die Gender-Konstruktionen *in situ* zu beschreiben.

3.2.6 Datenschutz, Datenverwertung und Privatsphäre

Ein weiterer Diskurs dreht sich um Fragen nach dem Datenschutz, der Datenverwertung durch die Anbieter und der Privatsphäre der Nutzer*innen von Smart Speakern. Dabei sind verschiedene Aspekte zu unterscheiden: Einerseits wird die Auswertung und weitere Verarbeitung der von den Nutzer*innen bereitgestellten

Daten durch die Hersteller kritisch betrachtet und die (u. a. rechtliche) Zulässigkeit der Verwertung diskutiert (siehe etwa Turow 2021). Andererseits wird hinterfragt, inwieweit in die Privatsphäre der Nutzer*innen von Smart Speakern durch die Geräte eingegriffen wird und inwieweit diese auch nutzer*innseitig (wissentlich oder unwissentlich) zugunsten einer komfortablen Nutzung der Dienste aufgegeben wird (siehe etwa Lutz 2023). Es entsteht somit ein Spannungsfeld zwischen einem erbetenen Mithören auf der einen und einem Eingriff in die Privats- oder sogar Intimsphäre der Haushalte auf der anderen Seite (vgl. Habscheid et al. 2025b).

Zur rechtlichen Zulässigkeit der Verwertung der Daten kommt auf Basis der aktuellen Rechtslage, insbesondere im Hinblick auf die Datenschutzgrundverordnung (DSGVO), ein Gutachten der Wissenschaftlichen Dienste des Deutschen Bundestags (2019) jedenfalls für den Smart Speaker von Amazon zu dem Schluss, dass die Verarbeitung zu den angegebenen Zwecken zulässig ist. Die Gutachter*innen prüften dabei die Voraussetzung der Einwilligung in die Verarbeitung, die notwendige Bereitstellung von Pflichtinformationen und die mögliche Weitergabe an staatliche Stellen. Sie bezogen sich in ihrem Gutachten insbesondere auf die Zulässigkeit der Transkribierung der mündlichen Eingaben, nachdem diese von verschiedener Seite in Zweifel gezogen worden war (vgl. Wissenschaftliche Dienste des Deutschen Bundestages 2019: 5). Die Autor*innen stellen zwar fest: „Amazon dürfte seiner Pflicht zur Informationsvermittlung [...] in hinreichendem Maß nachgekommen sein.“ (Wissenschaftliche Dienste des Deutschen Bundestages 2019: 9), weisen aber auch darauf hin, dass die Daten potenziell zu sehr verschiedenen Zwecken, auch außerhalb der von Amazon selbst bereitgestellten Plattformen, genutzt werden könnten. Dies würde weitere Informationspflichten nach sich ziehen. Als problematisch im Blick auf die geltende Rechtslage wird in dem Papier hervorgehoben, dass für die Nutzer*innen erkennbar sein muss, wann überhaupt Daten zur weiteren Verarbeitung gesammelt werden. Stimmdateien gelten dabei im Sinne der DSGVO als besonders schutzwürdig und werden in der Kategorie der biometrischen Daten behandelt (vgl. Zarcone/Leschansky 2023: 172). Dies sei, so das Gutachten der Wissenschaftlichen Dienste weiter, insbesondere für Dritte relevant, die sich nicht im Klaren darüber sind, dass ein Smart Speaker potenziell das Gesprochene aufzeichnet; ferner sei der Minderjährigenschutz möglicherweise nicht gewährleistet. Beide Nutzer*innengruppen können nicht ohne Weiteres von der Datenerhebung ausgenommen werden (vgl. Wissenschaftliche Dienste des Deutschen Bundestages 2019: 8–9).

Zum letztgenannten Punkt äußerte sich auch die Datenethikkommission der Bundesregierung (2019) in ihrem Bericht. Dieser hebt insbesondere darauf ab, dass nicht nur intransparent ist, wie die erhobenen Daten verarbeitet werden,

sondern auch, wann Daten erhoben werden und ob die Anwesenden darum wissen. Sie empfehlen insofern „[b]indende technische Vorgaben zur Abschaltbarkeit von Mikrofon und Internetverbindung sowie eine Sichtbarmachung, ob das Mikrofon an- oder ausgeschaltet ist“ (Datenethikkommission der Bundesregierung 2019: 101). Dazu gehöre eine dem Medium und der Situation angemessene Transparenzpflicht, z. B. über ein akustisches Signal. Ferner wird eine Reduktion der Online-Verarbeitung der Daten sowie eine Regulierung dieser Vorgänge empfohlen (vgl. Datenethikkommission der Bundesregierung 2019: 101; 118).

Von verschiedener Seite wird hinterfragt, ob tatsächlich – wie es den Angaben der Hersteller entspreche – eine Speicherung und Verwertung in der Cloud nur dann erfolgt, wenn das Aktivierungswort erkannt wird. Diese Aufzeichnungen werden in der Smartphone-App als ‚Protokolldaten‘ hinterlegt, sodass sie für die Nutzer*innen über ein zweites, mit dem Smart Speaker verbundenes Gerät nachvollziehbar werden (vgl. Habscheid et al. 2021). Auf Basis einer Netzwerktraffic-Analyse zeigen etwa Ford/Palmer (2019), dass zwar die Übertragung definitiv unterbrochen wird, wenn das Mikrofon geräteseitig abgeschaltet ist. Zugleich deuten die Daten durchaus darauf hin, dass bei eingeschaltetem Mikrofon auch Daten transportiert werden, deren Übertragung an die Cloud nicht dokumentiert wird. Die beständige Internetverbindung, der für die Inanspruchnahme der Funktionalitäten benötigte „Always-On“-Status und die unklare Weiterverarbeitung der Daten in den Cloud-Servern erschweren eine Kontrolle erheblich (vgl. Gray 2016; Apthorpe et al. 2017). Da auch eine Verifizierung der tatsächlichen Aufnahme des Aktivierungsworts in den Datencentern der Hersteller stattfindet, kann nicht einmal sicher ausgeschlossen werden, dass nicht auch ein Teil der sog. Hotword-Detection ebendort stattfindet. Wird eine Löschung vorgenommen, bestreitet Amazon zwar, dass die Datensätze weiterhin verfügbar sind, gibt jedoch auch an, andere Datensätze teilweise länger zu speichern, um einen reibungslosen Ablauf der genutzten Dienste trotz der Löschung gewährleisten zu können. Dabei wird nicht genauer spezifiziert, welche Datensätze hier gemeint sind (vgl. Amazon 2021). Für Drittanbieter-Daten gibt Amazon an, dass diese nicht gelöscht werden können (vgl. Amazon 2021). Die Frage nach der Datenspeicherung ist bei anderen Herstellern ebenso unklar.

Opazität besteht insofern nicht nur beim *Ob*, sondern auch hinsichtlich der Frage, *wie* die Daten ausgewertet werden. Über die genauen Prozesse der algorithmischen Auswertung der Aufnahmen werden seitens der Hersteller keine Informationen bereitgestellt, Amazon beschreibt aber in knapper Form, dass die Daten für eine Verbesserung genutzt werden, und zwar mit Hilfe von „maschinellern Lernen, einem branchentypischem Prozess, bei dem Menschen einen sehr kleinen Anteil von Anfragen überprüfen, um Alexa zu helfen, die zutreffende In-

terpretation der Anfragen zu verstehen“ (vgl. Amazon 2021).¹¹³ Dieser Passus scheint die Reaktion auf einen medial geführten Diskurs zur Datensicherheit zu sein, bis zu dessen Aufkommen diese Information nicht an die Besitzer*innen der Geräte weitergegeben wurde. Mit Blick auf Datenschutz lässt sich somit auch die Frage stellen, wer Zugriff auf die Daten und welche Bestandteile davon hat: Die Angaben der Hersteller, dass es sich dabei um vollständig anonymisierte Stimmbeispiele handelt, können wohl nicht ganz unbegründet in Zweifel gezogen werden (vgl. Zuboff 2018: 302). Auch der Schutz vor unbefugtem Zugriff auf die Daten scheint nicht durchgängig gewährleistet zu sein und die Hersteller könnten hierbei ihre Sorgfaltspflichten vernachlässigt haben; zu diesem Schluss kommen jedenfalls Berichte von Investigativjournalist*innen und Whistleblowern (vgl. Petereit 2021a; 2021b), denen zufolge der unautorisierte Zugriff auf Sprachaufnahmen über Jahre hinweg für eine vierstellige Anzahl von nicht befugten Amazon-Mitarbeiter*innen gegeben war. Teilweise herrsche ein „Datenchaos“, Angaben über Speicherort und -art könnten nicht rekonstruiert werden und insofern seien auch die Grundsätze der DSGVO nicht einzuhalten.

Mit einer größeren gesellschaftlichen Perspektive scheint es jedoch zu einfach zu sein, die großen Tech-Firmen ‚lediglich‘ für Missstände im Hinblick auf den Datenschutz und die Intransparenz zu kritisieren. Der Umgang mit *Big Data* und Sprachaufnahmen als Bestandteil davon kann auch als gesamtgesellschaftliche Herausforderung betrachtet werden, bei der sich wirtschaftlich betrachtet die klassischen Rollen von Produzent*in und Konsument*in verschieben und die bisherigen Marktmechanismen und Verfahren zu dessen Regulation versagen. In diesem Prozess sind Amazon, Google und Apple mehr als Dienstleister, die sich nicht an rechtliche Vorgaben halten: Sie sind, um es mit der Wirtschaftswissenschaftlerin Shoshana Zuboff zu sagen, Treiber des „Überwachungskapitalismus“. Mit diesem totalitaristischen Wirtschaftskonzept zeichnet Zuboff (2018) die Folgen der Digitalisierung und der damit einhergehenden Verdatung des privaten Alltags als eine Art real werdende Dystopie. Menschliches Verhalten werde zu „Verhaltensdaten“ (siehe hierfür und für die folgenden Begriffe Zuboff 2018: 22), die nicht alle zur Verbesserung des maschinellen Lernens und somit der angebotenen Dienstleistungen dienen, sondern die als sog. „Verhaltensüberschuss“ in die Entwicklung von „Vorhersageprodukten“ einfließen, die das menschliche Verhalten prädiktiv „erahnen“. Diese Vorhersageprodukte können in einem von Zuboff als „Verhaltensterminkontraktmarkt“ bezeichneten Handelsverkehr verkauft werden

113 Beim Hersteller Google handelt es sich dabei eigenen Angaben zufolge um 0,2 Prozent, bei Amazon und Apple weniger als ein Prozent (siehe ZEIT Online 2019); diese Angaben dürften allerdings angesichts der nicht unerheblichen Bedeutung menschlicher Beteiligung am Prozess des maschinellen Lernens zweifelhaft sein (siehe auch Kremp 2019).

und werden wiederum dadurch besonders wertvoll, dass sie treffsicher sind. Die Erhöhung der Treffsicherheit gebe letztlich Anreiz zur *Beeinflussung* des menschlichen Verhaltens: „Ergebnis dieses Wandels ist, dass automatisierte Maschinenprozesse unser Verhalten nicht nur *kennen*, sondern auch in einer wirtschaftlichen Größenordnung *auszuformen* vermögen [Herv. i. O.]“ (Zuboff 2018: 33). Die Art der Verhaltensdaten und des daraus generierten Datenüberschusses können vielfältig sein (etwa Such- und Einkaufsverhalten im Internet, Kommunikationsverhalten, Nutzungszeiten und -orte usw.). Darunter fallen auch sprachliche Daten bzw. Konversationen, die die Smart Speaker mitschneiden und zu Sprachüberschuss werden. Dieser Sprachüberschuss und der Wettbewerb, in dem die Anbieter aktuell stehen, führen dazu – so die Argumentation Zuboffs (2018: 307–309) –, dass die Art und Qualität der angebotenen Dienstleistungen hinter die „Jagd nach Verhaltensüberschuss“ zurücktritt und primär Potenziale erschlossen werden, um noch weiteren Sprachüberschuss generieren zu können.

Dieser Argumentation und dem Konzept des Überwachungskapitalismus muss man nicht folgen; eine ausführliche Diskussion der Szenarien von Zuboff ist an dieser Stelle untunlich (siehe aber Morozov 2019). Unbestreitbar ist allerdings, dass die gesellschaftlichen und ökonomischen Verhältnisse durch die Dominanz der hier untersuchten Anbieter und auch vermittelt der hier untersuchten stationären Sprachassistenzsysteme in Bewegung geraten sind (siehe etwa Sadowski 2020). Wie Woods (2018) herleitet, begünstigen sich die im Kapitel zuvor ausgeführten Gender-Stereotype und das Anliegen der Datengewinnung seitens der Großkonzerne gegenseitig. Die Dominanz der Datafizierung sozialer Praktiken und die Anwendung von Künstlicher Intelligenz in allen Lebensbereichen prägen und normieren wirtschaftliche und gesellschaftliche Strukturen des 21. Jahrhunderts. Die Konzentration dieser Prozesse auf kommerziell betriebene Plattformen wie Amazon, die zu Infrastrukturen des täglichen Lebens werden (Dolata/Schrape 2018; Dolata 2019; Goulden 2019; Plantin/Punathambekar 2019; Strüver 2023a), führt zu Veränderungen in der Wirtschaftsordnung (vgl. Srnicek 2016; Montalban/Frigant/Jullien 2019). Diese Entwicklungen werden unterschiedlich bewertet: Während etwa Hepp (2020) die hier angesprochenen Prozesse mit dem Konzept der Mediatisierung in einer durchaus als optimistisch zu bezeichnenden Diskussion von Vor- und Nachteilen bespricht, fürchten andere Autor*innen die Manifestierung von gesellschaftlichen Ungleichheiten, die Verschlechterung von Arbeits- und Lebensverhältnissen und eine erodierende Staatlichkeit angesichts der Macht durch die gesammelten und für verschiedene Zwecke auswertbaren Daten (siehe etwa Crawford/Joler 2018; Couldry/Mejias 2019).

Vor diesem Hintergrund ist auch der Schutz der Privatsphäre nicht mehr nur eine Frage danach, ob sich die Tech-Firmen an geltendes Recht halten, sondern auch, welche Implikationen es für die Nutzer*innen hat, wenn Smart Speaker Au-

dioaufzeichnungen intimer Details aus dem Privatleben an einen proprietären und opak agierenden Cloud-Dienst weiterleiten (vgl. Waldecker/Volmar 2022). Konkret für Smart Speaker wurden bereits Resignation (vgl. Lau/Zimmerman/Schaub 2018: 18), Überforderung sowie die Entwicklung eines „pragmatischen Fatalismus“ (Waldecker/Martin/Hoffmann 2025) oder Zynismus (Lutz/Newlands 2021) seitens der Nutzer*innen beobachtet, der auch in anderen digitalen Zusammenhängen an der Schnittstelle von Intimität, Privatheit und Öffentlichkeit festgestellt werden konnte (siehe etwa Schmidtke/Englert/Waldecker 2019). Dies zeigt an, dass ein Großteil der Gesellschaftsmitglieder den Entwicklungen möglicherweise kritisch, aber nicht aktiv widerständig gegenübersteht. Nutzer*innen, die generell größere Bedenken hinsichtlich der Privatsphäre haben, neigen auch zu größeren Bedenken im Hinblick auf die Überwachung durch Smart Speaker und Sicherheitsschwächen von Smart Speakern, z. B. Datenmissbrauch (vgl. Mols 2021: 163–165). Nutzer*innen, die mehr über die Geräte wussten und im Umgang mit ihnen geübt waren, waren im Hinblick auf Sicherheitsaspekte und die Datenverwertung innerhalb der Plattformlogiken ebenfalls besorgter (vgl. Mols 2021: 163–165). Im Einklang damit halten Lutz/Newlands (2021: 154) auf Grundlage ihrer Befragung britischer Smart-Speaker-Anwender*innen fest, dass diese die größte Sorge im Hinblick auf eine Hörbarkeit der Daten für Dritte haben (z. B. Vertragsarbeiter*innen der Systemanbieter). Zwar scheinen Nutzer*innen als Reaktion darauf unterschiedliche Strategien für „Privacy Work“ zu entwickeln (vgl. Brause/Blank 2023; siehe auch Chalhoub/Flechais 2020), darunter auch die Unterbrechung laufender Gespräche bei einer versehentlichen Aktivierung des Smart Speaker, was wiederum ein Hinweis auf die Relevanz einer empirischen Untersuchung sprachlicher Praktiken in VUI-Dialogen ist. Zugleich zeigen Lutz/Newlands (2021: 155), dass die von ihnen befragten Nutzer*innen einen hohen Grad an Pragmatismus bzw. Zynismus erkennen lassen und an einer intensiveren Auseinandersetzung mit Datenschutzeinstellungen nicht interessiert sind bzw. keinen Aufwand dafür betreiben wollen. Waldecker (2022: 156) konstatiert in einer kritischen Diskussion des Konzepts der *Datensouveränität*: „De facto haben sich die Nutzenden von Smart Speakern in gewissem Rahmen mit der Auswertung ihrer Daten abgefunden“, ¹¹⁴ und nimmt an, dass die Nutzer*innen

114 Die Studie von Waldecker (2022) fasst Teilergebnisse des Projekts „Un/erbetene Beobachtung in Interaktion: ‚Intelligente Persönliche Assistenten‘ (IPA)“ im Sonderforschungsbereich „Medien der Kooperation“ zusammen. In diesem Projekt ist auch die vorliegende Arbeit entstanden. Die genannte Untersuchung stützt sich auf Interviews in denselben Haushalten, in denen auch für die vorliegende Arbeit gemeinsame Datenerhebungen und interdisziplinäre -auswertungen durchgeführt wurden.

eine umfängliche Souveränität weder erlangen können, noch danach zu streben scheinen.

Notwendig für eine weitere Exploration der hier aufgezeigten brisanten Spannungsfelder ist eine empirische Grundlage, die die *Praxis* ins Zentrum der Untersuchungen rückt. Es muss betrachtet werden, wie *in situ* Smart Speaker eingesetzt werden, wie dabei Datenschutz und -verarbeitung sowie Kategorien von Intimsphäre, Privatheit und Öffentlichkeit sprachlich-kommunikativ reflektiert werden und welche Veränderungen im Alltag der Nutzer*innen sich beobachten lassen. Die linguistische Betrachtungsweise, in der der soziale Status der Smart Speaker in den Blick rückt, trägt somit dazu bei, die soziale Wirklichkeit im Umgang mit den skizzierten gesellschaftlichen Diskursen analysierbar zu machen. Dabei wird über die sprachlichen ein Zugriff auf soziale Praktiken ermöglicht und es können die Annahmen der Anwender*innen über die Funktionsweise der Geräte, ihre Einstellungen zu Datenschutz, -verwertung und Privatsphäre so beobachtet werden, wie sie selbst von den Sprecher*innen relevant gesetzt werden. Die mit dieser Brille opak bleibenden Datenpraktiken im „Back-End“ von Smart Speakern, die Auswertung und mittel- bis langfristigen Folgen für die Weiterentwicklung von Algorithmen werden dabei als Domäne betrachtet, zu deren weiterer Untersuchung die Analysen des „Front-Ends“ einen Beitrag leisten können.

3.3 Untersuchte Smart Speaker-Modelle

Die drei am Markt gängigsten Smart Speaker – Amazon Echo, Google Home und Apple HomePod – sollen nachfolgend kurz vorgestellt werden. Dabei soll die äußere Gestaltung, die Bedienweise und Funktionalität der Geräte im Mittelpunkt stehen. Zu dieser gehören notwendigerweise auch stimmliche Eigenschaften des Voice User Interfaces (VUI). Ferner werden die zugehörigen Smartphone-Anwendungen beschrieben. Amazon Echo ist mit weitem Abstand der bekannteste Smart Speaker: 2024 hatten 72 Prozent der Haushalte, die einen Smart Speaker besitzen, ein solches Modell (hierfür und für die folgenden Angaben vgl. Statista 2025a). 22 Prozent der Nutzer*innen verwendeten ein Produkt von Google (Google Home und Google Nest mit jeweils 15 bzw. 8 Prozent). 15 Prozent verwendeten den HomePod von Apple. Die Beschreibung der Smart Speaker-Modelle geht zunächst vom Amazon Echo aus. Anschließend werden Google Home und Apple HomePod v. a. in ihrer Differenz zum Amazon-Produkt vorgestellt.

3.3.1 Amazon Echo

Der Amazon Echo ist in verschiedenen Produktvarianten erschienen. Seit dem Marktstart waren bis zum Zeitpunkt der Datenerhebung vier Generationen mit jeweils zwei verschiedenen Ausführungen erhältlich: Amazon Echo sowie eine weitere Ausführung als „Echo Dot“, die stets mit etwas Verzögerung zur Veröffentlichung der neuen Generation als alternative Produktvariante erschien und insgesamt etwas kleiner und im Hinblick auf den Lautsprecher etwas weniger leistungsstark ist, in der Funktionalität jedoch keine Einschränkungen hat. Die Abb. 1–3 zeigen die ersten drei Generationen des Amazon Echo, wobei die erste Generation (Abb. 1) in der größeren Echo-Variante dargestellt ist, während Generation 2 und 3 in der Echo Dot-Produktvariante abgebildet sind:



Abb. 1: Amazon Echo, 1. Generation; Bild: „Frmorrison“ / CC BY-SA 3.0.



Abb. 2: Amazon Echo Dot, 2. Generation; Bild: Raimond Spekking / CC BY-SA 4.0.



Abb. 3: Amazon Echo Dot 3. Generation mit aktiviertem Lichtring; Bild: „hamburgfinn“ / Pixabay.

Hardwareseitig besteht das System v. a. aus drei Komponenten: mehreren miteinander zusammenwirkenden Mikrofonen (sog. *Far-Field-Voice-Recognition*, siehe Drösser 2020: 71), einem Lautsprecher sowie einer kleinen Recheneinheit, die die

Übertragung ins Internet leistet und einige wenige lokale Prozesse steuert. Über die eingebauten LED-Leuchten sendet das Gerät in unterschiedlichen Farben nichtsprachliche Signale (vgl. Pearl 2016: 219–221; Bedford-Strohm 2017: 487). Die Abb. 4–6 zeigen die LED- und Mikrofon-Einheit eines Echo Dot sowie die zugehörigen Recheneinheiten und den Lautsprecher.

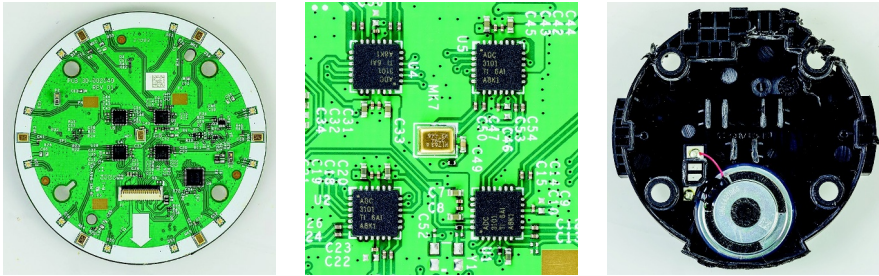


Abb. 4–6: Sechs Lautsprecher und LED-Einheiten, Recheneinheiten und die Unterseite eines Echo Dot mit Lautsprecher-Einheit. Bilder: Raimond Spekking / CC BY-SA 4.0.

Diese einzelnen Komponenten, aus denen sich der Amazon Echo zusammensetzt, werden von verschiedenen Softwareanwendungen in Anspruch genommen. Diese ergeben erst im Zusammenspiel eine Einheit und sind im Hintergrund als unterschiedliche Prozesse zu begreifen, die auch verschiedene Teildisziplinen der Informatik und der Computerlinguistik berühren. Auf Abb. 3 ist der aktivierte Lichtring zu erkennen, der rund um die obere bzw. untere Kante des mit Stoff verkleideten, zylinder- bzw. kugelförmigen Korpus des Geräts gespannt ist. Dieser signalisiert in blauer Farbe, dass die Aufzeichnung aktiv ist, die aufgenommenen Audiodateien also zur Auswertung in die Cloud übertragen werden (Listening-Modus). Ist der Lichtring ausgeschaltet (und passt sich dadurch optisch der Farbe des Smart Speakers an), findet zwar bei hergestellter Internetverbindung und Stromversorgung die Suche nach dem Aktivierungswort statt, aber es erfolgt keine regelmäßige Übertragung in die Cloud.¹¹⁵ Der Lichtring kann bei der 3. Generation des Geräts noch andere Farben annehmen (siehe dazu Albrecht 2020: 35): Leuchtet der Lichtring etwa orange, wird eine Internetverbindung gesucht, ein „pulsierender gelber Lichtring“ steht für bereitstehende Benachrichtigungen. Ein roter Lichtring zeigt an, dass das Mikrofon abgestellt und somit die Suche nach dem Aktivierungswort unterbrochen wurde. Dies kann über eine der vier

¹¹⁵ Siehe Kap. 3.2.6 für weitere Auswertungen zu diesem Sachverhalt.

Tasten auf der Oberseite des Geräts erfolgen (siehe Abb. 2 und 3). Zwei weitere Tasten können zur Lautstärkeregelung (mit Plus- und Minus-Symbol) genutzt werden. Eine vierte „Aktionstaste“ dient der Aktivierung des Geräts ohne verbale Nennung des Aktivierungsworts (vgl. Albrecht 2020: 35–36).

Die Positionierung der Tasten und des Lichtrings variieren etwas in der kugelförmigen 4. Generation des Amazon Echo (siehe Abb. 7). Außerdem ist das Gerät mit einem anderen Lautsprecher ausgestattet und hat einen geringfügig anderen Klang, der sich jedoch Reviewer*innen wie Kawalkowski (2020) zufolge kaum vom Vorläufermodell unterscheidet. Die Recheneinheit wurde geringfügig beschleunigt, aber die Funktionen sind sonst nahezu identisch, was sich durch die Verarbeitung der Daten in der Cloud erklärt: Die Hardware ist für das Leistungsspektrum nur noch ein nachgelagerter Bestandteil (siehe auch Stiftung Warentest 2021: 28; Freeman-Mills 2021).



Abb. 7: Amazon Echo Dot, 4. Generation;
Bild: Public Domain / Unsplash, Bearbeitung:
Sina van Oostrum.

Alle Amazon Echo-Geräte sind mit dem VUI Alexa ausgestattet, das die akustische Steuerung des Geräts über Stimmein- und ausgaben ermöglicht. Die Stimmfarbe des Geräts lässt sich nicht in den werkseitig verbauten Funktionen anpassen, das Aktivierungswort hingegen kann geändert werden (möglich sind auch „Amazon“, „Computer“ oder „Echo“). Auch wenn Alexa im alltäglichen Sprachgebrauch häufig synonym mit dem Smart Speaker verwendet wird (vgl. Albrecht 2020: 7), ist der Sprachassistent auch auf anderen Endgeräten verwendbar, z. B. auf Smartphones, Smart Speakern von anderen Produzenten oder Tablets (vgl. Albrecht 2020: 11). Außerdem lassen sich Drittanbieter-Programme, *Skills*, auf den Geräten installieren, die dann teilweise auch andere VUIs einprogrammiert haben, sodass sich die Stimmfarbe innerhalb der Programme verändern lässt. Die synonyme Verwendung von Alexa als Bezeichnung für den Smart Speaker ist also in doppel-

ter Hinsicht nicht akkurat: Einerseits kann die Software Alexa auf anderen Geräten verwendet werden, andererseits können auf dem Smart Speaker andere Anwendungen genutzt werden.¹¹⁶

Weitere Produktvarianten der Amazon Echo-Reihe sind der Amazon Echo Show und der Echo Spot, die jeweils einen Monitor verbaut haben, auf dem Uhrzeit, Wetter, Text- und Videomaterial und andere Anwendungen in Kombination mit der stimmlichen Schnittstelle auch visuell präsentiert werden (hierfür und für das Folgende siehe Albrecht 2020: 18–19). Ermöglicht werden auch Videotelefonate. Geräte mit Bildschirm sind seit 2017 auf dem US-Markt erhältlich und machten global gesehen 21 Prozent der verkauften Amazon-Echo-Geräte aus (vgl. Canals 2019). Die Geräte werden in die weiteren Ausführungen und in die Analyse nicht einbezogen. Es ist davon auszugehen, dass die Möglichkeit der Wahrnehmung visueller Inhalte, der Steuerung über den Bildschirm (d. h. nicht über den Smartphone-Bildschirm, der ein externes Gerät erforderlich macht) und das gänzlich anders gestaltete Zusammenspiel optischer und akustischer Signale eine Vergleichbarkeit mit Voice-First-Devices (Bedford-Strohm 2017) sehr erschwert. Sicherlich wäre ein Vergleich, der in zukünftigen Arbeiten angestellt werden könnte, aufschlussreich im Hinblick auf die Spezifika, die die stimmliche Benutzerschnittstelle mit sich bringt; dieser soll jedoch im Rahmen dieser Arbeit nicht geleistet werden (siehe aber Hector et al. 2025).

Bei den hier im Fokus stehenden Geräten erfolgt wie erwähnt die Einrichtung und ein Großteil der Einstellungen über die Alexa-Smartphone-Applikation, die ebenfalls als Alexa bezeichnet wird und über die gängigen App-Stores auf Android- und Apple-Geräten (Smartphones und Tablets) installiert werden kann. Über die App kann das Gerät bedient und, selbst ohne einen Smart Speaker, der volle Funktionsumfang in Anspruch genommen werden – Stimmeingaben können auch über das im Endgerät verbaute Mikrofon erteilt werden. Die Anwender*innen konzentrieren sich, gerade in den Einrichtungssituationen, immer wieder auf die App, außerdem verweist das VUI bei bestimmten verbal geäußerten Stimmeingaben (etwa zu Smart Speaker-Einstellungen) selbst verbal auf die Anwendung, wenn die Anfrage nicht über das VUI verarbeitet werden kann. Die App ist, abhängig von der verwendeten Version und dem Endgerät, jeweils geringfügig unterschiedlich aufgebaut (insbesondere zwischen Android- und Apple-Geräten treten Unter-

116 Bis zur 3. Generation war als weitere Produktvariante der Echo Plus erhältlich. Er ähnelte den Geräten der Echo-Reihe, hatte also keinen Bildschirm verbaut, war dafür aber hardwareseitig für die Smart Home-Steuerung optimiert und mit einem Temperatursensor ausgestattet (vgl. Albrecht 2020: 17). In der 4. Generation ist die „Plus“-Linie eingestellt, dafür werden mittlerweile Echo Show Geräte mit integriertem Bildschirm vertrieben.

schiede auf). Sie umfasst aber im Wesentlichen die gleichen Funktionen, die exemplarisch in den Abb. 8–10 in der Android-Version von 2021¹¹⁷ dargestellt sind (für eine detaillierte Funktionsbeschreibung der App siehe Albrecht 2020: 36–49; Habscheid et al. 2021).

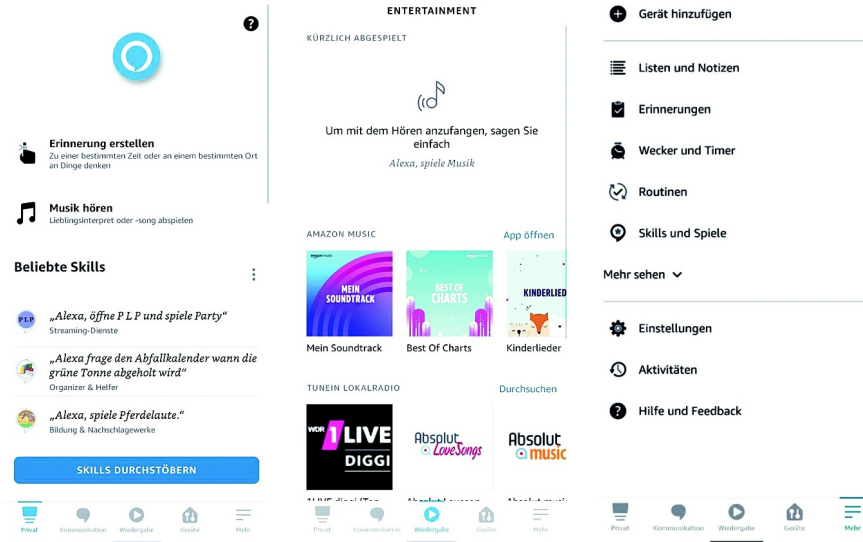


Abb. 8–10: Alexa-Smartphone-App, Funktionsbereiche „Privat“, „Wiedergabe“ und „Einstellungen“; Bilder: eigene Aufnahme / T.H.

Im Bereich „Privat“ (in anderen App-Versionen auch „Startseite“) findet sich die Möglichkeit zur Sprachsteuerung über die App im oberen Bereich, der blaue Button mit dem Alexa-Logo in Form einer angedeuteten Sprechblase aktiviert die Sprachsteuerung (Abb. 8). Darunter finden sich personalisierte Vorschläge für weitere Anwendungen, Skills oder neue Funktionsweisen. Über die Toolbar im ganz unteren Teil der App können nun andere Funktionsbereiche der App ausgewählt werden. Im Menüpunkt „Kommunikation“ (hier nicht abgebildet) finden sich Steuerungselemente für Anrufe und Textnachrichten, die über die Amazon-Geräte vermittelt werden. Diese werden hier protokolliert und der Zugriff auf die

¹¹⁷ Dieser Zeitpunkt wurde gewählt, damit die Darstellung hier möglichst ähnlich zur Darstellung auf den Endgeräten der im Rahmen der Studie aufgezeichneten Nutzer*innen ist (die Aufzeichnungen erfolgten zwischen Ende 2020 und Mitte 2022, siehe Kap. 5).

jeweiligen Kontakte hergestellt. Im Bereich „Wiedergabe“ (Abb. 9) werden Musik, Hörbücher, Radio und ähnliche akustische Medieninhalte gesteuert. Playlisten und Radiosender stehen voreingestellt zur Verfügung, können aber auch manuell angelegt werden. Die Verwaltung von Streaming-Diensten wie Spotify oder Deezer kann über dieses Menü erfolgen. Im Bereich „Geräte“ (hier nicht abgebildet) erfolgt die Steuerung der Smart Speaker, die der App hinzugefügt wurden. Außerdem können hier auch Smart Home-Elemente verwaltet werden. Der Funktionsbereich „Mehr“ (Abb. 10), der bei einigen anderen App-Versionen oben links angeordnet ist, enthält weitere Funktionen zur Steuerung des Smart Speakers über die App, z. B. Erinnerungen, Wecker, das Anlegen von Routinen (die vordefinierte Abfolge von bestimmten Funktionen, ggf. zu bestimmten Zeiten) und die Verwaltung weiterer Skills. Im Bereich „Aktivitätsverlauf“ bzw. im Bereich „Datenschutz“ sind als Protokolldaten die ausgeführten Aktionen der Nutzer*innen hinterlegt (siehe Abb. 11).



Abb. 11: Alexa-Smartphone-App, Funktionsbereich „Aktivitätsverlauf“; Bild: eigene Aufnahme / T.H.

Bei mündlicher Steuerung und entsprechender Einstellung werden auch die aufgezeichneten Audio-Daten hier abgespeichert (vgl. Habscheid et al. 2021: 19) und können über den blauen Wiedergabebutton angehört werden, außerdem wird der verstandene Wortlaut transkribiert und die Stimmausgabe (ausschließlich schriftlich) dokumentiert. Die App ermöglicht also die Nutzung wesentlicher Funktionen auch über die App (teilweise mit integrierter stimmlicher Bedienung). Das Funktionsspek-

trum geht über das des VUIs hinaus: Für bestimmte Einstellungen und Anwendungen, gerade, wenn sie eher prinzipiellen Charakter haben, kann nur die App verwendet werden. So können z. B. die Adresse des Standorts, die WLAN-Verbindung, Zehldaten für Online-Shopping und andere Daten, die für das Prozessieren bestimmter Anwendungen erforderlich sind, nicht über das VUI, sondern nur über die App geändert werden. Wie die Anwendung der App in die stimmliche Nutzung eingebunden ist, soll in den Analysen mitbetrachtet werden (insbesondere in den Inbetriebnahme-Situationen), wobei trotz der hohen Relevanz der App weiterhin ein Fokus auf der Anwendung der stimmlichen Benutzerschnittstelle liegen soll.

3.3.2 Google Home und Google Nest

Die Smart Speaker von Google sind im Wesentlichen sehr ähnlich aufgebaut wie die Produktreihe von Amazon. Firmierte die Serie zunächst unter dem Markennamen „Google Home“ wurde sie 2019 in „Google Nest“ umbenannt, ohne dass daraus wesentliche Änderungen am Funktionsumfang erfolgt wären.



Abb. 12: Google Nest-Reihe: Aktivierter Google Home Mini; Bild: Andrea Marchitelli, Wikimedia Commons / CC BY-SA 4.0.



Abb. 13: Google Nest-Reihe: Google Home Smart Speaker, Google Home Hub und Google Home Mini; Bild: Y2kcrazyjoker4, Wikimedia Commons / CC BY-SA 4.0.

Die drei Geräte (siehe Abb. 13) – Google Home bzw. Google Nest Audio, Google Home Hub bzw. Google Nest Hub und Google Home Mini bzw. Google Nest Mini – sind jeweils die erste bzw. zweite Generation, wobei sich im Leistungsspektrum nur wenig verändert hat: Ähnlich wie beim Amazon Echo findet die Hauptanpassung der Funktionen, vermittelt über eine App, in der Cloud statt. Die Google-Reihe zeigt, anders als die Echo-Reihe, die Aktivität nicht durch einen Lichtring, sondern durch eine Leiste von Lichtpunkten auf dem Display an (siehe Abb. 12). Diese können weiß und orange leuchten, nacheinander aufblinken und pulsieren, wobei Anzahl und Bewegung der Punkte codiert sind – so bedeuten etwa vier orange Punkte, dass das Mikrofon ausgeschaltet ist, während nacheinander aufleuchtende orange Punkte einen Software-Prozess (z. B. ein Update) anzeigen (vgl. Google Support 2021). Auch während der Ersteinrichtung werden über die Punkte Statusmeldungen des Geräts übermittelt. Ferner ist, in Abgrenzung zu Amazon Echo, die Oberfläche des Geräts teilweise über Touch-Technologie funktionalisiert. So kann durch Berührungen an den Rändern des Oberflächengewebes die Lautstärke eingestellt werden. Auch hierbei geben die weißen Punkte Hinweise auf die Höhe der eingestellten Wiedergabelautstärke. Die „Google Home“-App (siehe Abb. 14 für die Startseite) umfasst ebenfalls die Steuerung der wesentlichen Funktionen des Smart Speakers, wobei auffällig ist, dass dieser stärker an das Smartphone und den Google-Account der Nutzer*innen angebunden ist. So verweisen mehrere Links innerhalb der App auf das Benutzer*innen-Konto bei Google, in dem Einstellungen zum „Google Assistant“ vorgenommen werden können, das sowohl über die Smart Speaker als auch über das Smartphone oder andere Endgeräte verwendet wird. Dies wird durch das Profilbild des*der Google-Nutzer*in oben rechts deutlich, das zum Aufrufen wesentlicher Einstellungen (z. B. Stimmprofil des VUI¹¹⁸ oder Eingabe persönlicher Präferenzen bei der Internetsuche) angetippt werden muss.

Außerdem zeigt sich, in Abgrenzung zur Alexa-App, eine stärkere Fokussierung auf einen Haushalt, dem verschiedene Geräte hinzugefügt werden können. Die in Abb. 14 prominent platzierte Überschrift „Neustadt“ ist der Name des Haushalts. Dieser Haushalt kann teilweise losgelöst vom Google-Konto verwaltet werden (über die Schaltfläche „Einstellungen“, die in der Mitte der Startseite platziert ist, ist er jedoch letztlich ebenfalls immer wieder damit verknüpft). Im Vergleich mit der „Alexa“-App besteht bei dieser zwar auch eine Verknüpfung mit einem Amazon-Konto und die Repräsentation eines ‚Haushalts‘ mit mehreren Geräten. Gleichwohl ist aber eine deutlich we-

118 Anders als bei Amazon-Produkten kann beim Stimmprofil für verbale Ausgaben des Smart Speakers von Google zwischen zwei verschiedenen Profilen ausgewählt werden.

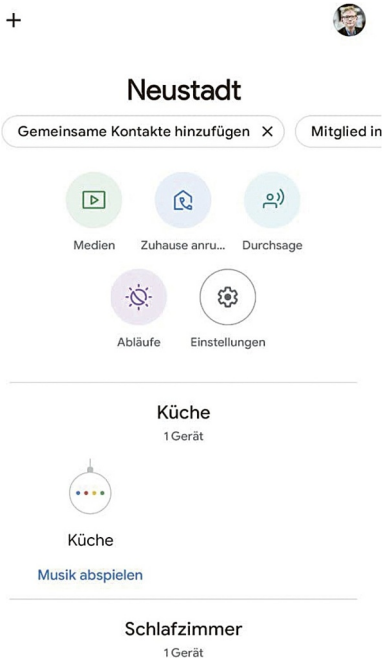


Abb. 14: Startseite der Google-Home-App; Bild: Screenshot / T.H.

niger starke Anbindung an das Amazon-Konto innerhalb der Anwendung sichtbar, im Vordergrund stehen die Funktionen des Smart Speakers. Dies ist auch dadurch erklärbar, dass die Amazon-Reihe nicht mit mehreren strukturell unterschiedlichen Instanzen umgehen muss: Während das VUI von Google sowohl auf dem Smartphone wie auch auf dem Smart Speaker ansprechbar ist und hier in einem integrierten Ansatz ermöglichen will, die gleichen Präferenzen beizubehalten, fokussiert sich Amazon auf die einheitliche Steuerung des Smart Speakers selbst bzw. der App-Instanz des Smart Speakers. Als letzte Unterscheidung zum Amazon-Produkt lässt sich bei Google ferner eine weniger starke Personalisierung feststellen. Das Aktivierungswort von Google Home- bzw. Google Nest-Produkten lautet „Okay Google“ und „Hey Google“ (ist also nicht mit einem onymischen Ausdruck personalisiert) und es zeigt sich auch in den Werbestrategien der Hersteller weniger stark der Aufbau einer „Persona“ (vgl. Dickel/Schmidt-Jüngst 2021).¹¹⁹

¹¹⁹ Siehe auch Kap. 6.1.1.

3.3.3 Apple HomePod

Der HomePod von Apple ist seit 2017 auf dem Markt. 2020 erweiterte Apple – analog zu den Serien von Amazon und Google – das Produktangebot um einen HomePod Mini (siehe Abb. 15). Der HomePod ist demgegenüber 2021 eingestellt worden. Das Produkt von Apple ist, ähnlich wie die Google-Reihe, mit dem auch auf dem Smartphone oder Tablet verfügbaren Apple-VUI („Siri“) ausgestattet. Über dieses VUI findet, wie auch bei Google, eine Verknüpfung der Einstellungen von Smartphone und Smart Speaker statt. Die Steuerung erfolgt ebenfalls über eine App, die „Apple Home“-App, die auf dem Smartphone oder Tablet installiert ist. Über diese können auch Smart Home-Elemente gesteuert werden. Die Verknüpfung erfolgt hier mit der Apple-ID, dem Pendant zum Google- oder Amazon-Account.



Abb. 15: Apple HomePod Mini; Bild: Arne Müsseler / CC BY-SA 3.0 de, Bearbeitung: Sina van Oostrum.

Der Smart Speaker von Apple ist Stiftung Warentest (2021: 28) zufolge mit einem leistungsfähigeren Lautsprecher ausgestattet als die Produkte der anderen Hersteller, während das VUI (Reaktion auf das Aktivierungswort, „Sprachmelodie, Betonung und Phrasierung“) schlechter als das von Amazon bewertet wurde. Der HomePod lässt keine Drittanbieter-Skills zu. Verknüpfen lässt er sich lediglich mit dem Musik-Streamingdienst von Apple selbst und darüber hinaus nur mit Apple-Anwendungen. Dabei ist die Funktionalität der Anwendungen im Vergleich zur Bedienung über das Smartphone – ähnlich wie bei der Nutzung von Google-Anwendungen über den Smart Speaker – teilweise eingeschränkt.

Wichtig ist zu betonen, dass die technologischen Entwicklungen im Bereich der Smart Speaker sehr schnell voranschreiten. Während der ca. sechs Jahre, die die Geräte jetzt auf dem deutschen Markt erhältlich sind, wurden mehrere Generationen, Produktvarianten und Hardware- wie Software-Updates entwickelt. Zum Zeitpunkt der Veröffentlichung dieser Arbeit werden die gängigen Produkte möglicherweise

bereits anders aussehen, neue Funktionalitäten ermöglichen, während andere wegfallen, neue Marken ausbilden und andere einstellen. Durch die Volatilität dieses Sektors und die noch fehlende Etablierung einheitlicher Standards ist es unmöglich, hier den Anspruch einer vollständig aktuellen Beschreibung des Smart Speaker-Felds zu erfüllen. Insofern hat sich die Beschreibung auf die Geräte konzentriert, die zum Zeitpunkt der Erhebung marktgängig und v. a. in den untersuchten Haushalten verfügbar waren (auf diese wurde auch bei den Abbildungen ein Schwerpunkt gelegt). Neue Versionen und Software- wie Hardware-Updates größerer Art, die nach der Datenerhebung erfolgten, werden hier nicht besprochen, da sie für die Nutzer*innen nicht relevant waren und während der Analyse keinen Einfluss mehr auf die erhobenen Daten hatten.¹²⁰

¹²⁰ Siehe aber Kap. 7.3.

4 Methodologische Einordnung

Eine praxeologisch ausgerichtete Arbeit scheint zunächst auf das Problem zu stoßen, dass eine an Praktiken ausgerichtete Methodologie sich noch nicht fest etabliert hat und die Diskussion dazu erst in den letzten Jahren beginnt (vgl. Schäfer/Daniel 2015: 40). Wie eine besondere Herausforderung kann zudem die Verbindung einer praxistheoretischen Grundierung mit den sprachwissenschaftlichen Erkenntnisinteressen dieser Arbeit wirken, die methodologische Implikationen einer ganz anderen Tradition mit sich bringen. Doch bei genauerem Hinsehen zeigt sich gerade die Verbindung von Praxeologie als theoretischer Ausgangslage mit einem sprachwissenschaftlich-empirischen Zugang als besonders fruchtbar, wenn für die methodologische Situierung als Bindeglied die Ethnomethodologie eingesetzt wird (vgl. auch Hector 2022). Diese ist nicht nur eine wichtige Ressource praxeologischer Theoriebildung (vgl. Hillebrandt 2014: 43), sondern hat schon früh Thesen aufgestellt, die bei der Entwicklung einer Methodologie zur Analyse sprachlicher als Bestandteil sozialer Praktiken hilfreich sein werden. Außerdem, und das ist am wichtigsten, liefert sie „bewährte Analyseverfahren“ (Schäfer/Daniel 2015: 42), die bereits in den 60er- und 70er-Jahren in der Sprachwissenschaft aufgegriffen wurden und sich etablieren konnten, was nahelegte, deren Mentalität zu folgen, u. a. mit der Sequenzanalyse als zentralem Verfahren der Datenauswertung.

Das Verhältnis von Praxeologie und Ethnomethodologie sowie der Weg Letzterer in die Linguistik hinein soll im Folgenden kurz vorgestellt werden, um den Zusammenhang zwischen einer soziologischen Forschungsrichtung und sprachlichen Analysen deutlich zu machen und die Entscheidung, diesen methodologischen Prinzipien zu folgen und sie in den Dienst einer praxeologischen Untersuchung zu stellen, zu legitimieren (Kap. 4.1). Es werden anschließend auch die darauf aufbauenden methodologischen Ausgangspunkte vorgestellt. Damit ist ein Bündel von Strömungen gemeint, das Imo (2013) als „Sprache-in-Interaktion“ zusammenfasst – die Konversations- bzw. Gesprächsanalyse, die Interaktionale Linguistik sowie die Gesprochene-Sprache-Forschung (Kap. 4.2). Darauf aufbauend wird Ethnografie als Forschungsstil in der Gesprächsanalyse diskutiert, der die vorliegende Arbeit maßgeblich geprägt hat und Perspektiven über die Gesprächsanalyse hinaus bietet (Kap. 4.3). Abschließend wird die Rolle von Multimodalität in der sprachlichen Praxis diskutiert (Kap. 4.4), um so die methodologische Positionsbestimmung zu komplettieren. Im Zwischenfazit (Kap. 4.5) werden die einzelnen Teile noch einmal mit Blick auf die vorliegende Arbeit reflektiert.

4.1 Ethnomethodologie als praxeologische Methodologie

Zu betonen ist zunächst noch einmal die immer wieder herausgestellte Nähe der Arbeiten des Begründers der Ethnomethodologie, Harold Garfinkel, zu den späteren sprachphilosophischen Arbeiten Ludwig Wittgensteins (vgl. Bergmann 2007b: 54; 2011: 229; Auer 2013: 132), die beide dadurch charakterisiert sind, dass sie den alltäglichen *Gebrauch* sprachlicher oder nicht-sprachlicher Ressourcen zur Durchführung sozialer Handlungen in den Mittelpunkt stellen und die Betrachtung auf neuropsychologischer bzw. kognitiver Ebene ablehnen. Die sozialen und mithin sprachlichen Praktiken können nicht unabhängig vom alltäglichen *Vollzug* dieser betrachtet werden: Experimentelle Untersuchungsdesigns können darüber – diese Überzeugung teilt der späte Wittgenstein mit Garfinkel – keinen Aufschluss geben. Harold Garfinkel verweist also, ganz ähnlich wie Wittgenstein, auf den *Gebrauch* der Zeichen und misst der Sprache eine entsprechend hohe Bedeutung bei. Die Ähnlichkeit der praxeologischen und ethnomethodologischen Wurzeln, die sich in den Arbeiten von Wittgenstein und Garfinkel zeigen, ist bemerkenswert, denn „obwohl sie wissenschaftsgeschichtlich nichts miteinander zu tun haben“ (Auer 2013: 132), liefern sie beide die Grundlage für bis heute produktive Paradigmen der Sozialtheorie, die immer wieder aktualisiert und neu zugerichtet wurden. Auch Meyer (2015: 91) stellt heraus, dass die Ethnomethodologie als „praxistheoretischer Ansatz“ gesehen werden muss: „Praktiken werden in der Ethnomethodologie sowohl sozialen Strukturen als auch sozialen Handlungen bzw. Akteuren (und ihren Intentionalitäten und Identitäten) vorgeordnet“ – damit folgt sie ganz dem Ziel, „Praxis allen anderen [...] Erklärungsgrößen vorzuordnen“, das Schüttpelz/Meyer (2017: 158) für eine praxistheoretische „Konstitutionsanalyse“ in ihrem bereits erwähnten Glossar zur Praxistheorie formulieren. Die Ethnomethodologie eignet sich damit bestens als Verbindungsstück zwischen den dargelegten praxissoziologischen Grundlagen und der sprachwissenschaftlichen Beschreibungsebene, auf der sprachliche *als* soziale Praktiken herausgearbeitet werden sollen. Die Verschränkung von Praxeologie und Ethnomethodologie wird dabei auch in den folgenden Ausführungen immer wieder durchscheinen.

Die Ideen Harold Garfinkels und die v. a. durch sein 1967 erschienenes Hauptwerk *Studies in Ethnomethodology* begründete Forschungsrichtung grenzte sich anfänglich v. a. gegen den Strukturfunctionalismus Talcott Parsons ab, bei dem Garfinkel selbst studiert hatte. In der Linie der phänomenologischen Soziologie nach Alfred Schütz war er jedoch mit diesem der Auffassung, dass in den Arbeiten Parsons zur sozialen Ordnung die subjektiven Perspektiven der einzelnen (und immer einen konkreten situativen Zusammenhang eingebetteten) Individuen zu kurz kommen und durch ein Verständnis von Wissenschaftsrationalität

überschrieben werden.¹²¹ Garfinkel führt stattdessen eine alltägliche Rationalität ein, die von der wissenschaftlichen Rationalität strikt zu trennen ist. In der Untersuchung der alltäglichen Rationalität werden die konkreten sozialen und alltäglichen Situationen ernst genommen, in die gesellschaftliche Ordnungsprinzipien hinein diffundieren, in denen sie aber auch erst wirkmächtig und real werden und eine Weiterentwicklung und ein Zurückdiffundieren in die Ordnungsprinzipien erfahren (vgl. Bergmann 2007b: 53). Nur diese an der Oberfläche beobachtbaren sozialen Handlungen können herangezogen werden, um die Praktiken zu beschreiben, mit denen soziale Ordnung immer wieder situativ neu hergestellt und ausgehandelt wird.

Damit ist die Grundidee der Ethnomethodologie nach Garfinkel (1967) bereits angerissen: Es sollen die Praktiken und Vorgehensweisen, d. h. die *Methoden*, beschrieben werden, die die Mitglieder einer Gesellschaft (*Ethnos*) nutzen, um ihr eigenes Handeln erkenn- und verstehbar zu machen und dabei gleichsam die Wirklichkeit der Situation, in der sie sich befinden, zu strukturieren (vgl. Bergmann 2007b: 51). Diese Praktiken und Vorgehensweisen können sich übersituativ stabilisieren und somit gesellschaftliche „Rollen“ bilden, die wiederum ebenfalls zu stabileren Gefügen konglomerieren können und somit „Institutionen“ ausbilden (vgl. Patzelt 1987: 10). Damit folgt auch die Ethnomethodologie einer flachen Ontologie im Sinne Schatzkis (2016), weil auch hier zwischen den ‚kleineren‘ und den ‚größeren‘ Einheiten ein konstitutiver Zusammenhang besteht, nicht aber eine Kausalitäts- oder Abhängigkeitsbeziehung zwischen beiden. Die Unterscheidung zwischen Mikro- und Makroebenen bleibt insofern aus Sicht der überwiegenden Zahl der Ethnomethodolog*innen irrelevant (vgl. Bergmann 2011: 231). Die Praktiken bzw. *Ethno-Methoden* sind die entscheidenden Einheiten, die ‚größere‘ Einheiten konstituieren, während die größeren reflexiv auf die Ethno-Methoden zurückwirken. Die Wirklichkeit und ihre Geordnetheit wird ethnomethodologisch insofern als „fortwährendes Resultat situativer, aber methodischer Alltagspraktiken“ (Bergmann 2011: 229) untersucht. Mit dem Konzept der „accountability“ (Garfinkel 1967: 34) wird dieser zentrale ethnomethodologische Gedanke auf den Punkt gebracht: die These, dass diese methodisch eingesetzten Praktiken, die einer bestimmten sozialen Wirklichkeit folgen und in einer spezifischen sozialen Situation angemessen sind, gleichsam diejenigen sind, die diese Situation erst als solche erkenn- und darstellbar machen (vgl. Bergmann 2007a: 125–126). Die Ethnomethodologie geht also davon aus, dass „die Tätigkeiten, mit denen Gesell-

¹²¹ Das Verhältnis von Harold Garfinkel zu seinem Lehrer Talcott Parsons, von dem er sich zugleich mit Vehemenz abgrenzt, ist vielschichtig und kann hier nicht zufriedenstellend dargestellt werden. Zur Beziehung der beiden siehe Heritage (1984b); zur wissenschaftlichen Auseinandersetzung zwischen Alfred Schütz und Parsons siehe Schütz/Parsons (1977).

schaftsmitglieder die Szenerien organisierter Alltagsangelegenheiten schaffen und bewältigen, identisch mit den Vorgehensweisen sind, mit denen Mitglieder diese Szenerien zurechenbar (*account-able*) machen“ (Garfinkel 2020[1967]: 35),¹²² und dass dies als eine „endlose fortlaufende, kontingente Leistung“ (Garfinkel 2020[1967]: 35) anzusehen ist. So ist das Verhältnis von Subjekt bzw. subjektivem Handlungsplan, Umständen und sozialen Regeln als prozessual zu begreifende Eigenrationalität des Alltäglichen charakterisiert (vgl. Bergmann 2007a: 127).

Da nach Auffassung der Ethnomethodologie in allen gesellschaftlichen Lebensbereichen Ethno-Methoden zum Einsatz kommen und diese durch ihre inhaltliche „Ladung“ (siehe Patzelt 1987: 10) spezifisch für den Wirklichkeitsbereich sind, in dem sie sich vollziehen, ist kein Phänomengebiet wichtiger als ein anderes. Überall müssen die situativen Gegebenheiten, der soziale Kontext und die Hintergründe der handelnden Personen ernst genommen werden – Garfinkel/Sacks (1979: 138) bezeichnen dies als „ethnomethodologische Indifferenz“. Wie Patzelt (1987: 11–13) ausführt, ist es aufgrund der gesellschaftlichen Pluralität, auf die Garfinkel und Sacks mit diesem Konzept eine Antwort finden, unmöglich, Zusammenhänge allgemein und unspezifisch zu erfassen. Stattdessen hat sich ausgehend von der „Allgemeinen Ethnomethodologie“ eine Reihe ethnomethodologischer Subdisziplinen gebildet, die in der Soziologie sehr produktiv waren und sind.¹²³ Diese arbeiten alle strikt empirisch mit ‚ungestellten‘ bzw. ‚natürlichen‘ Daten,¹²⁴ häufig mit Video- oder Audioaufnahmen als Fixierungen des Prozesses, und nehmen in jedem Fall den realen zeitlichen Verlauf in den Blick. Ethnografische Methoden wie die teilnehmende Beobachtung sind ebenso verbreitet, auch wenn die technische Fixierung der Abläufe als Ton- und Videoaufnahmen und die damit verbundene Nachvollziehbarkeit der realen zeitlichen Abläufe vorzuziehen ist (vgl. Bergmann 2001: 922). Das Datenmaterial wird induktiv bearbeitet und die Analysekategorien ohne Vorannahmen erst aus dem Material heraus entwickelt.

Trotz dieser methodischen Gemeinsamkeiten ethnomethodologischer Untersuchungen ist, auch als Reaktion auf eine Methodenkritik, der Formulierung eines klaren Regelwerks von Garfinkel stets eine Absage erteilt worden, mit dem Verweis auf ein „unique adequacy requirement“ (Garfinkel 2002: 175–176): Die Methoden müssen immer den gegebenen Umständen angepasst werden. Damit

¹²² Hier wird auf die von Schüttpelz/Rawls/Thielmann (2020) herausgegebene Übersetzung ins Deutsche verwiesen.

¹²³ Für Beispiele aus sehr unterschiedlichen Bereichen wie ethnomethodologische Studien in der Kriminologie, einer „Ethnomethodologie des Lesens“ oder einer ethnomethodologischen Wissenschaftsanalyse siehe Patzelt (1987: 12).

¹²⁴ Zum Natürlichkeitspostulat siehe Kap. 5.3.

nahm Garfinkel auch Stellung zur dahingehend geäußerten Kritik, dass die methodische Ausleuchtung von Ethno-Methoden ja selbst eine Methode ist, die es ja gerade zu hinterfragen galt, aber in sich schon eingeschriebene Vorannahmen enthalte (siehe dazu Bergmann 2007b: 57). Eine methodologische Offenheit wird auch in der aufkeimenden Methodologie-Diskussion in der Praxissoziologie eingefordert, die insofern auch eine Rückbesinnung auf ethnomethodologische Prinzipien darzustellen scheint: Eine unflexible Methodentreue wird ebenso abgelehnt, demgegenüber eine Entwicklung von Methoden entlang empirischer Untersuchungen vorgeschlagen (vgl. Schäfer/Daniel 2015: 52), was genau dem Vorschlag eines „unique adequacy requirement“ entspricht, für das ein „möglichst breites Spektrum an methodischen Herangehensweisen“ (Schäfer/Daniel 2015: 52) zur Verfügung stehen sollte. Der theorie- und methodenbildende Anspruch wurde auch in der Ethnomethodologie äußerst gering gehalten – besonders mit dem Anspruch, keine Vorprägungen durch vorgefertigte Konzepte zuzulassen. Dieses Prinzip zieht sich durch die Praxeologie ebenso wie durch die ethnomethodologischen Subdisziplinen, wenngleich es dort gelegentlich bei genauerer Betrachtung auch zugunsten einer stärkeren Methodenlehre verletzt wurde.

4.2 Sprache-in-Interaktion

Eine der Strömungen, die sich als ethnomethodologische Subdisziplin entfaltete, ist die ethnomethodologische Konversationsanalyse.¹²⁵ Von dieser inspiriert und von einer sprachwissenschaftlichen Weiterentwicklung der ursprünglich soziologischen Erkenntnisinteressen geprägt, entwickelte sich zudem eine Reihe weiterer Forschungsansätze, die Imo (2013: 59) zusammen mit der Konversationsanalyse unter dem Oberbegriff „Sprache-in-Interaktion“ zusammenfasst. Diese Traditionen greifen auch Forschungslinien und Sprachbeschreibungstheorien auf, die sich unabhängig davon bereits in der Sprachwissenschaft entwickelt hatten und teilweise eine Wiederbelebung durch neue Verbindungen erfuhren. Die Ansätze lassen sich nicht trennscharf voneinander abgrenzen – die nachfolgende Gliederung orientiert sich an der terminologischen Handhabung bei Imo (2013: 59), auch wenn klar ist, dass zwischen den Teilbereichen nicht nur wissenschaftshistorische Entwicklungen verbindend sind, sondern auch aktuell Übergänge im Hinblick auf die Bearbeitung gemeinsamer Erkenntnisinteressen bestehen. Für die vorliegende Arbeit sind Prin-

¹²⁵ Ihre ethnomethodologischen Wurzeln sind, je nach Ausprägung, mehr oder weniger präsent. Der Zusatz „ethnomethodologisch“ ist insofern nicht in jedem Fall gerechtfertigt und auch von den Autor*innen selbst nicht verwendet, weshalb er auch im Folgenden getilgt wird.

zipien der ethnomethodologischen Konversationsanalyse und der damit eng verbundenen Gesprächsanalyse sowie das Forschungsprogramm der Interaktionalen Linguistik entscheidend. Zurückgegriffen wird auch auf Teile der älteren Gesprochene-Sprache-Forschung. Diese Strömungen und Programme sollen als Teil der methodologischen Einordnung kurz vorgestellt werden. Es werden deren Entwicklungslinien nachgezeichnet und ihre Relevanz für die vorliegende Arbeit begründet. Weitere, mehr oder minder scharf abgegrenzte Strömungen und Schattierungen im Umfeld von Sprache-in-Interaktion, wie z. B. die Dialoganalyse (Fritz/Hundsnurscher 1994b),¹²⁶ die Gesprächslinguistik (Brinker/Sager 2010; Mroczynski 2014) oder die Konstruktionsgrammatik (Ziem/Lasch 2013), teilen zwar die analytische Arbeit an Audio- bzw. Videoaufnahmen von gesprochensprachlichen Daten, einen empirischen Anspruch und eine Engführung auf sprachwissenschaftliche Interessen. Sie unterscheiden sich aber auch deutlich, zum einen in der Ausprägung ihres Anwendungsbezugs und mit Blick auf die sprachanalytische Ebene der Untersuchungen (vgl. Imo 2013: 71), zum anderen grenzen sie sich von der Konversations- bzw. Gesprächsanalyse und der Interaktionalen Linguistik insbesondere dadurch ab, dass sie nicht die sprachliche Praxis bzw. den sprachlichen Zuschnitt auf die Interaktion in den Mittelpunkt rücken, sondern auf andere Aspekte fokussieren. Sie sind insofern für die Bearbeitung der Fragestellungen nicht geeignet, auch wenn sie vereinzelt Impulse liefern können; auf sie wird daher in der nachfolgenden Darstellung nur peripher eingegangen.

4.2.1 Konversationsanalyse

Wie Patzelt (1987: 11) anmerkt, handelt es sich bei der Konversationsanalyse zwar einerseits um eine der „speziellen Ethnomethodologie[n]“, indem sie Sprachverwendung ins Zentrum rückt und die Rolle sprachlicher Zeichen bei der Herstellung sozialer Wirklichkeit untersucht, damit aber andererseits nahezu alle Lebensbereiche tangiert und deshalb zwar eine „spezielle“ Ethnomethodologie ist, gleichzeitig aber nicht in einem thematisch gebundenen Feld Halt macht (vgl. Patzelt 1987: 12). Die ethnomethodologische Konversationsanalyse ist auch insofern eine etwas anders gelagerte Teildisziplin, als sie nach ihrer Gründung v. a. von Harvey Sacks, Gail Jefferson und Emanuel Schegloff als eigenständiges Teilgebiet

¹²⁶ Es ist bezeichnend für die Unschärfe der verschiedenen Ansätze, dass das erste Kapitel dieses Sammelbandes eine Einführung in die ethnomethodologische Konversationsanalyse von Jörg Bergmann ist. In den weiteren Kapiteln werden, u. a. von Johannes Schwitalla, Heinrich Löffler und Jacques Moeschier, weitere Ausdifferenzierungen dieser Ansätze eingeführt, denen hier nicht im Detail nachgegangen werden soll.

äußerst produktiv war und bis heute ist und sich zunehmend als eigene Forschungsrichtung verstand. Sie grenzte sich mit ihrem Fokus auf die inhärente Struktur von Gesprächen zunehmend von der Ethnomethodologie ab. Von der Konversationsanalyse haben sich weitere Forschungszweige wie die Interaktionale Linguistik (Selting/Couper-Kuhlen 2001a) inspirieren lassen. Sie wurde zunehmend methodisch wie theoretisch autonom von der Ethnomethodologie Garfinkels, was auch zu methodologischen Auseinandersetzungen und Spannungen zwischen Ethnomethodologie und Konversationsanalyse führte (vgl. Bergmann 2001: 921; 2007a: 134). Die Konversationsanalyse brach so auch mit einigen ethnomethodologischen Grundsätzen – so sollten in der konversationsanalytischen Datenauswertung gerade keine Hintergrundinformationen einbezogen werden, sondern nur das für die Analyse relevant gemacht werden, was auch die Sprechenden selbst in der Konversation relevant setzen (Bergmann 2001: 922). Während die Ethnomethodologie im Sinne Garfinkels die Einmaligkeit und Situiertheit der Ereignisse betont und Zuschreibungen übergeordneter Kategorien ablehnte, versuchte die Konversationsanalyse abstraktere Mechanismen der Interaktionsordnung herauszuarbeiten (vgl. Pollner 1991: 373). Für diesen Versuch, allgemeinere Verfahren zu finden, wurde sie seitens der Ethnomethodologie kritisiert; im Gegenzug bemängelten Konversationsanalytiker*innen die im engeren Sinne ethnomethodologischen Ansätze als zu partikularistisch und argumentierten mit der geringen Aussagekraft der Ergebnisse, die eine Überbetonung der Einzigartigkeit von Situationen impliziere.

Die Entwicklung der ethnomethodologischen Konversationsanalyse war einer der wesentlichen Faktoren und leistete Pionierarbeit für die weitere Etablierung sequenzanalytischer Verfahren in Soziologie und Sprachwissenschaft (vgl. Ploder/McElvenny 2022: 292). Mit ihrer Abstraktion von der konkreten Situation und mit ihrer Engführung auf das *Sprachliche* zur Herstellung von Sinn und Ordnung in der sozialen Interaktion war die Konversationsanalyse durchaus in der Lage, wesentliche Prinzipien sprachlicher Interaktion zu entdecken und auf unterschiedliche Kontexte anzuwenden. So wurden etwa Verfahren des Turn-Takings, d. h. des Sprecherwechsels, systematisiert (vgl. Sacks/Schegloff/Jefferson 1974), selbst- und fremdinitiierte Reparaturleistungen klassifiziert (vgl. Schegloff/Jefferson/Sacks 1977; Schegloff 1997a) sowie insbesondere von Schegloff (1968; 2007) die Bedeutung des sequenziellen Ablaufs von Gesprächen und Adjazenzpaaren (z. B. Frage – Antwort) herausgestellt (siehe auch Imo 2013: 74–75).¹²⁷ Letztere war als Strukturmerkmal nicht nur für kleinere Einheiten wie Begrüßungen und Verab-

¹²⁷ Auch Birkner (2020a: 16–17) nennt diese drei Themen als Gegenstandsbereiche der frühen Konversationsanalyse.

schiedungen (vgl. Schegloff/Sacks 1973) zentral, sondern auch für größere Einheiten wie Erzählungen (vgl. Quasthoff 1980a; 2001), Beschwerden (vgl. Günthner 1999a) oder Witze (vgl. Kotthoff 2006a; Norrick 2001). Diese Arbeiten waren möglich aufgrund der von Sacks (1984: 21) zusammengeführten Erkenntnisse, dass soziale Aktivitäten formell beschrieben werden können, dass sie außerdem methodisch und rekurrent auftreten und dass sie insofern generalisierbare Eigenschaften aufweisen, die es herauszuarbeiten gelte. Dazu sei es nötig, die Daten anzuschauen, ohne Vorannahmen oder konkrete Fragestellungen an sie heranzutragen, idealerweise nicht einmal über den Kontext der Daten genau Bescheid zu wissen, um dann aus den Daten heraus die Aspekte herauszufiltern, die für die Teilnehmer*innen selbst relevant sind und die sie sich gegenseitig aufzeigen und sie so gleichzeitig den Analytiker*innen zugänglich machen (vgl. Sacks/Schegloff/Jefferson 1974: 728). Dieses Konzept ist als *display*-These ein Eckpfeiler der konversationsanalytischen Forschungsarbeiten und bezieht sich nicht nur auf die reflexive Herstellung von Kontexten, sondern auch für das gegenseitige Aufzeigen von Sinn und Ordnung sowie das Verstehen dieser Ordnung. Das *display*-Konzept ist unmittelbar mit dem Konzept des *doing* verwoben, mit dem die Regelmäßigkeit und Konventionalisierung bestimmter kontextualisierender Verfahren betont wird, die ‚auf dem *display*‘ angezeigt werden (vgl. Birkner et al. 2020: 18). Dies kann von kleineren, aber wiederkehrenden sozialen Ereignissen bis hin zu gesellschaftlichen Rollenbildern – etwa *doing gender* (vgl. West/Zimmermann 1987) – reichen (siehe dazu auch Schegloff 1997a).¹²⁸ Ein anderer wesentlicher Eckpfeiler ist das Konzept des Adressatenzuschnitts, auch Recipient-Design, ebenfalls eingeführt in dem bereits erwähnten Aufsatz von Sacks/Schegloff/Jefferson (1974: 727). Dieses Konzept umfasst eine ganze Reihe von gesprächsorganisatorischen Erscheinungen, die dazu beitragen, einen Turn, d. h. einen Redezug, immer auf die Gesprächspartner*innen zuzuschneiden – bzw. auf die Bilder, die sich die Sprecher*innen von den Empfänger*innen machen.¹²⁹ Diese Anforderung an die Sprecher*innen verlangt von diesen, den sozialen Kontext, in dem sich diese befinden, zu erkennen und gleichzeitig, fortlaufend mit zu produzieren und damit auch die Rolle, die das Gegenüber in diesem Kontext hat. Dieses auch als „Adressatenzu-

¹²⁸ Zu Kritik und Umgang mit der *display*-These siehe Kap. 4.3.

¹²⁹ Damit ist, wie Hitzler (2013: 112) ausführt, ein sprecher*innenspezifisches Konzept umrissen, nicht eine „analytische Maxime“ – auch wenn es in den meisten Arbeiten so behandelt wird und damit auch gegen den bereits angerissenen ethnomethodologischen Anspruch der Freiheit von abstrakten methodischen Prinzipien verstößt.

schnitt“ bekannte Konzept ist Ausgangslage für viele konversationsanalytische Untersuchungen und wird auch in den vorliegenden Analysen eine Rolle spielen.

Die ethnomethodologische Fokussierung auf die sprachlichen Ressourcen, die bei der Herstellung sozialer Interaktion genutzt wurden, war anfänglich nicht nur dem Interesse an einem fundamentalen Aspekt des Sozialen geschuldet, sondern auch eine Form von analytischer Bequemlichkeit (vgl. Day/Wagner 2008: 44): Tonbandaufnahmen waren leicht zu beschaffen und gut zu analysieren und die Transkriptionen gaben viel Analysematerial her. Ton- und Videoaufnahmen dokumentieren das Gesprochene unvoreingenommen im Sinne einer „registrierende[n] Konservierung“ (Bergmann 1985: 305) und ermöglichen damit eine detaillierte Rekonstruktion der Aufzeigehandlungen von Gesprächsteilnehmer*innen. Von dort aus konnten konversationsanalytische Untersuchungsansätze, die sich trotz des ethnomethodologischen Anspruchs der Reduktion generalisierbarer Konzepte entwickelt hatten, in benachbarte Disziplinen diffundieren. Sie wurden so etwa in der empirischen Linguistik aufgegriffen und führten zur Entwicklung von noch stärker auf Sprache fokussierten Ansätzen, z. B. die Interaktionale Linguistik (siehe übernächster Abschnitt).

4.2.2 Gesprächsanalyse

Der Begriff *Gesprächsanalyse* wird erstens nahezu synonym für „Konversationsanalyse“ verwendet – als Übersetzung des Ausdrucks „conversation analysis“ – und ist mit diesen Konzepten eng verwandt (vgl. Imo 2013: 71–76). Auch in gesprächsanalytischen Arbeiten ist ein starker Begriff von Empirie gesetzt und bildet die Sequenzanalyse das grundlegende Verfahren zur Datenauswertung. Allerdings ist mit der Übertragung in den deutschsprachigen Raum bei Arbeiten aus dem Bereich der Gesprächsanalyse in den meisten Verwendungsweisen eine methodische Öffnung einhergegangen, etwa wenn ethnografisches Wissen in die Analysen einbezogen wird (siehe Deppermann 2000), kognitive Ansätze ergänzend verfolgt werden oder andere qualitative Forschungsmethoden komplementär hinzugezogen werden (vgl. Deppermann 2008: 10). Er steht also dafür, die bereits erwähnte „analytische Mentalität“ (Schenkein 1978: 1) der Konversationsanalyse beizubehalten, behält sich als Forschungsparadigma aber gleichzeitig Abweichungen vor und kann damit auch als eine Rückbesinnung zum „unique adequacy requirement“ (Garfinkel 2002: 175–176) gesehen werden, das in der Ethnomethodologie von Garfinkel postuliert worden war. Genau auf diese Form der Analyse stützt sich auch die vorliegende Arbeit, die in der strikten Befolgung konversationsanalytischer Maximen keinen Mehrwert für die Bearbeitung der eingangs gestellten Leitfragen sieht. Während die Maximen sich für die Erforschung konversationeller Universa-

lien etabliert hatten (im Überblick Schegloff 2006), steht in der vorliegenden Arbeit gerade nicht die Erarbeitung sprachlicher Universalien, sondern die ihre Anwendung unter spezifischen Kooperationsbedingungen im Fokus. Insofern scheint es geradezu geboten, Hintergrundwissen über die Sprecher*innen, ihre Vorherfahrungen und ihre Beziehung zueinander, über die lokalen, materiellen Gegebenheiten der Haushalte und über die Funktionsweise der VUIs wo erforderlich hinzu-zuziehen und die Analyse davon zu informieren.¹³⁰

Mit *Gesprächsanalyse* wird darüber hinaus zweitens ein weiteres, spezifisches Forschungsprogramm verstanden, das die ethnomethodologische Konversationsanalyse adaptiert und sie um Fragen der Sprechakttheorie und der Textlinguistik erweitert: „Eine umfassende Gesprächsanalyse hat [...] beide Ereignisse des sozialen Ereignisses ‚Gespräch‘ zu berücksichtigen und zu beschreiben: das Handlungsergebnis und den Handlungsvollzug“ (Brinker/Sager 2010: 20). Diese Weiterentwicklung der Konversationsanalyse war insbesondere in den 70er- und 80er-Jahren produktiv und versuchte zwischen der ethnomethodologischen Konversationsanalyse und der philologischen Perspektive zu vermitteln. Sie kann insofern auch als soziologisch interessierter Arm der Gesprochene-Sprache-Forschung betrachtet werden, der für das Forschungsprogramm der Ethnomethodologie empfänglich war und sich etwas später entwickelt hat, gleichzeitig jedoch nicht so sehr in der sozialwissenschaftlichen Tradition verwurzelt war, dass er den Anschluss etwa an die Arbeiten von Dell Hymes oder John Gumperz (s. o.) gefunden hätte. Dieses Programm ist für die vorliegende Arbeit nur eingeschränkt relevant.

4.2.3 Interaktionale Linguistik

Als sprachwissenschaftliche Wendung des konversationsanalytischen Ansatzes ist die *Interaktionale Linguistik* zu verstehen, die sich um die Jahrtausendwende etablierte und insbesondere methodische Erkenntnisse aus der Linguistik der späten 90er-Jahre bündelte (Selting/Couper-Kuhlen 2000; 2001a; 2001b). Die Interaktionale Linguistik nahm Grundgedanken der ethnomethodologischen Konversationsanalyse auf, verschob aber das Erkenntnisinteresse der Konversationsanalyse: Im Forschungsprogramm, wie es von Selting/Couper-Kuhlen (2001a) umrissen wurde,

130 Ferner wird der Terminus Gesprächsanalyse auch als Sammelbegriff für verschiedene Ansätze zur Untersuchung von Gesprächen verwendet und schließt damit die bis hierher vorgestellten Forschungslinien ein (vgl. Imo/Lanwer 2019: 18). Als Sammelbegriff für all diese Forschungslinien soll jedoch der bereits eingeführte, von Imo (2013) geprägte Begriff „Sprache-in-Interaktion“ oder im Folgenden auch der Begriff der ‚Gesprächsforschung‘ verwendet werden, um methodologische Implikationen anderer Begriffe zu vermeiden.

geht es zwar auch, aber nicht primär um die Herstellung von Sinn und Ordnung in der sozialen Interaktion und das Sprachliche als Mittel zum Zweck, sondern vielmehr um die Frage, welche konkreten sprachlichen Muster sich im Sprachgebrauch herausbilden, empirisch nachweisbar sind und von den Sprecher*innen auf die Interaktion zugeschnitten eingesetzt werden. Wie diese sprachlichen Muster durch ihren Interaktionszuschnitt entstehen und geformt werden und gleichzeitig wie die Interaktionen durch entstandene Muster geformt werden, sind die Kernfragen der Interaktionalen Linguistik (vgl. Barth-Weingarten 2008: 78), die insofern einen deutlich fokussierteren Blick auf die sprachlichen Details einnimmt, sich aber, wie Barth-Weingarten (2008: 78) ausführt, auf allen sprachwissenschaftlichen Beschreibungsebenen bewegen kann.

Innerhalb des sprachwissenschaftlichen Forschungsdiskurses setzte die Interaktionale Linguistik besonders dem „written language bias“ (Linell 2005) ein konsequent auf Empirie und Mündlichkeit eingestelltes Programm entgegen. Die Interaktionale Linguistik verfolgt somit, mit Couper-Kuhlen/Selting (2017: 3), „an interest in spoken language and a desire to see it studied in its natural habitat: in social interaction“. Durch die starke Verbindung zwischen sprachlicher Form und sozialer Interaktion, auf die in diesem Sinne alle sprachlichen Äußerungen zugeschnitten sind, bleibt auch eine starke Verbindung zwischen Konversationsanalyse und Interaktionaler Linguistik bestehen; die Interaktionale Linguistik ist „heavily indebted in conversation analysis“ (Couper-Kuhlen/Selting 2017: 3). So sehen die Begründerinnen dieses Forschungsprogramms die Interaktionale Linguistik auch als „Interface“ zwischen Sprachwissenschaft und Interaktionsanalysen, wobei sie sich disziplinar klar der Linguistik verschreiben (Selting/Couper-Kuhlen 2001a: 260). Zugleich betonen sie die Interdisziplinarität und schlagen vor, sich nicht um eine detaillierte Abgrenzung zu bemühen, sondern sich auf das Zusammenwirken der Disziplinen zu konzentrieren: „Rather than trying to define borders, the field of Interactional Linguistics should be looked upon as an interdisciplinary space [...]“ (Couper-Kuhlen/Selting 2017: 18).

Sprache wird als zugeschnitten auf ihren Verwendungszweck betrachtet, ihr ‚natürlicher Lebensraum‘ sind kontinuierlich neu entstehende kommunikative Probleme, zu deren Lösung kompetenten Sprecher*innen ein bestimmtes Repertoire an sedimentierten Verfestigungen zur Verfügung steht (vgl. Couper-Kuhlen/Selting 2017: 3). Die Lösungen wiederum werden durch die Sprechenden an die Individualität und Singularität jeder kommunikativen Konstellation, der sozialen Hintergründe und der materialen Gegebenheiten angepasst und haben somit zugleich auch immer Züge von Emergenz (vgl. Barth-Weingarten 2008: 82; siehe auch Günthner 2018b). Die Interaktionale Linguistik nimmt mit Auer (2000) insofern die Prozessualität der gesprochenen Sprache ernst (und das nicht nur, wie Auer in den Mittelpunkt stellt, im Bereich der Syntax). Dies hat zur Konsequenz, dass etwa ein

strukturalistischer Satzbegriff als Segmentierungskategorie zu kurz greift und andere Regelwerke, wie sie normativ in Grammatiken niedergeschrieben sind, auf der empirischen Ebene nicht nur nicht ‚eingehalten‘ werden, sondern systematisch Regeln, die erst durch den Gebrauch entstehen, übersehen. Grammatik ist insofern ein „ständig sich veränderndes Inventar an Orientierungsmustern“ (Imo 2013: 79), das sich im Spannungsfeld von situativer Emergenz auf der einen Seite und grammatischer, musterhafter Verfestigung auf der anderen Seite bewegt (siehe auch Günthner/Bücker 2009).

Als Wegbereiter und Grundlage der Interaktionalen Linguistik können andere soziologische Arbeiten gesehen werden, die einen Fokus auf die sprachliche Beschreibungsebene legten (vgl. Imo/Lanwer 2019: 59). Dazu zählen die Kontextualisierungstheorie, die Arbeiten von John Gumperz und Dell Hymes zur Interaktionalen Soziolinguistik und zur Ethnographie der Kommunikation (Gumperz/Hymes 1972; Hymes 2004; Gumperz 1982; 1999), die ebenfalls einen gebrauchsbasierten Ansatz verfolgten, sich gleichermaßen gegen die abstrakten, kontextlos gelösten Sprachstrukturen richteten und linguistische Formen in ihrem lokalen, aber auch ihrem sozialen und kulturellen Diskurs berücksichtigen wollten (vgl. Günthner 2008: 53). Besonders prominent und fruchtbar erwies sich das Konzept der Kontextualisierungshinweise (*contextualization cues*), die von Gumperz (1982) und für das Deutsche von Auer (1986) in den Blick genommen wurden und zeigen, dass auf allen sprachlichen Ebenen, von der Prosodie bis zur Stilistik, indexikalische Zeichen auftreten können, die als relationale, an einen sozialkulturellen Kontext gebundene, inferentielle Äußerungen zu verstehen sind und gleichzeitig diesen Kontext auch immer wieder konstituieren. Diese Grundannahmen sind nicht nur für die Entwicklung des „Forschungsprogramms Interaktionale Linguistik“ (Selting/Couper-Kuhlen 2001a) wesentlich gewesen (siehe auch Couper-Kuhlen/Selting 2017: 8–10), sondern sie sollen auch in den Analysen der vorliegenden Arbeit fruchtbar gemacht werden. Sie sind Teil einer anthropologischen Linguistik, die Sprache als Teil kultureller Praktiken begreift und deren Rolle bei der Entstehung und Entwicklung von sozialen und kulturellen Identitäten, in der Kulturverständigung und in der alltäglichen Situierung empirisch untersucht (vgl. Günthner 2008: 53; Foley 1997: 3).

Die Interaktionale Linguistik hat, so kann konstatiert werden, viele, weit verzweigte Wurzeln. In den Analysen wird nicht streng allen methodischen Geboten der Interaktionalen Linguistik gefolgt – dazu bietet sie sich auch mit der praxeologischen Fundierung dieser Arbeit nicht optimal an. Gleichwohl sollen wesentliche Prinzipien, insbesondere ein sequenzanalytisches Vorgehen und ein Interesse an Zusammenhängen zwischen sprachlicher Form und kommunikativer Funktion, übernommen werden. Dies kann aber nicht isoliert von anderen methodologischen Paradigmen geschehen, auf die nachfolgend noch einzugehen ist.

4.2.4 Gesprochene-Sprache-Forschung

Von der philologischen Linie der Forschung her beeinflusst waren unterschiedliche Arbeiten, die als *Gesprochene-Sprache-Forschung* zusammengefasst werden können (vgl. Imo/Lanwer 2019: 19–20). Auch auf diese wird in Arbeiten der Interaktionalen Linguistik regelmäßig zurückgegriffen, teilweise können sie auch zur Interaktionalen Linguistik hinzugerechnet werden. Sie wurzeln in einem Aufsatz des Germanisten Otto Behagel (1927), der in einer Festschrift die Domänen der Mündlichkeit und der Schriftlichkeit voneinander abgrenzte und auf wesentliche Unterschiede aufmerksam machte. Auch wenn dies in der philologischen Forschung zunächst keine große Rolle spielte, wurden diese Gedanken wieder aufgegriffen und Versuche einer Einordnung der gesprochenen Sprache in die Sprachgesamtheit (vgl. Moser 1969), Typisierungen (vgl. Schröder/Steger 1981) und syntaktische Beschreibungen (vgl. Eggers 1962) unternommen. In dieser linguistischen Auseinandersetzung, unbeeinflusst von den zeitlich parallel stattfindenden Prozessen, die zur Ausbildung der Ethnomethodologie, der Konversationsanalyse oder der Interaktionalen Soziolinguistik führten, rang man um die Kriterien, mit denen die Mündlichkeit in Abgrenzung von der Schriftlichkeit überhaupt bestimmt werden sollte (etwa Koch/Oesterreicher 1985). Zur Gesprochene-Sprache-Forschung in dieser Tradition können in jüngerer Zeit etwa die Arbeiten von Schwitalla (2012; 2018) oder der Band von Ágel/Hennig (2006) gezählt werden, deren Untersuchungen nicht so sehr von den soziologischen Forschungstraditionen beeinflusst waren, sondern stärker in der Germanistik verwurzelt sind. Die Empirie ist hier häufig nicht leitend für die Analyse, sondern Fragestellungen der Typisierung und Charakterisierung gesprochen-sprachlicher Funktionen.

An gesprochener Sprache interessiert und gebrauchsbasiert arbeitet auch die *Funktionale Pragmatik*, die Gesprächsforschung mit Kontexten und gesellschaftlichen Zusammenhängen über die einzelne Situation hinaus verbunden hat (vgl. Meer/Pick 2019: 31). Im Mittelpunkt der Untersuchungen der u. a. von Konrad Ehlich und Jochen Rehbein begründeten Funktionalen Pragmatik stehen der Zusammenhang von Sprache und Handeln bzw. Handlungen und (gesellschaftlichen) Zwecken des in Gespräche eingebundenen Handelns (vgl. Ehlich 2019: 304). Damit konnten insbesondere über Gespräche in gesellschaftlichen Institutionen (u. a. Schule und Hochschule, Gericht, Sozialberatung, aber auch Familie) und somit über den Zusammenhang von gesellschaftlichen Zwecken und damit verbundenen Handlungsmustern Erkenntnisse gewonnen werden (vgl. Meer/Pick 2019: 32). Dafür wurde ein spezifisches Set an Begriffen und Methoden entwickelt (einführend Ehlich 2007), das teilweise große Ähnlichkeit mit anderen Ansätzen aus der Gesprächsforschung aufweist und diesen insbesondere durch die starke Handlungs- und Anwendungsorientierung ähnelt. Forschungsergebnisse aus den

Forschungslinien der Gesprochene-Sprache-Forschung und der Funktionalen Pragmatik werden zwar in den Analysen dieser Arbeit teilweise herangezogen, deren Erkenntnisinteressen und methodologische Vorgehensweisen aber nicht geteilt. Im Mittelpunkt der vorliegenden Arbeit steht eine induktive Vorgehensweise und die sequenzanalytische Datenauswertung, was (mit guten Gründen) in den genannten Strömungen nicht der Fall war.

4.3 Sprache und Ethnografie

Gesprächsanalytische Arbeiten im weiter oben vorgestellten, ersten Sinne (als Fortsetzung der Konversationsanalyse) zeigten sich etwas offener für methodologische Erweiterungen aus dem Bereich der Ethnografie, ganz besonders durch den wegweisenden Aufsatz von Deppermann (2000). Damit sind sie etwas anders nuanciert als Arbeiten der ‚strengen‘ Konversationsanalyse am anderen Ende eines ausdifferenzierten Spektrums von Arbeiten, die nicht immer eindeutig klassifiziert werden können und müssen. Dennoch war eine solche Öffnung nicht selbstverständlich: Der Anspruch der Konversationsanalyse ist es, nur die Informationen für die Analyse zu verwenden, die die Sprecher*innen auch im Gespräch selbst für die Organisation der Konversation relevant machen – Schlussfolgerungen sollen also ausschließlich anhand der Gesprächsaufzeichnungen gezogen werden. Keine zusätzlichen Informationen sollen in die Analysen einfließen. Diese Regeln werden auch in der vorliegenden Arbeit ganz bewusst nicht befolgt. Denn schon aus den theoretischen Verortungen wurde deutlich, dass dies einer praxeologisch ausgerichteten Arbeit mit einem Interesse an einer ganzheitlichen Betrachtungsweise nicht gerecht werden kann: Es wird in dieser Arbeit explizit nicht nur nach den Prinzipien gefragt, mit denen eine Interaktion bzw. ein VUI-Dialog organisiert werden, sondern auch danach, wie interaktionale und soziotechnische Arrangements entstehen, wie sprachliche Praktiken Ausdruck einer sozialen Vollzugswirklichkeit an der Oberfläche sind und wie sich das kommunikative Repertoire der Interagierenden entwickelt. Ohne Hintergrundinformationen über die Sprecher*innen und deren Beziehungsgeflecht lassen sich diese Praktiken und Prinzipien nicht erfassen. Es wird mit Deppermann (2000: 99) davon ausgegangen, „dass das Erfassen von Sinn und Ordnung im Gespräch keine Frage eines einfachen Ablesens oder Hörens ist. Das Verständnis von Sinn und Ordnung bedarf vielmehr der Interpretation“. Interpretationen, so konstatiert Deppermann, werden in der strengen Konversationsanalyse für einen Teil der untersuchten Fragestellungen nicht hinreichend berücksichtigt. Zwar wird, z. B. von Schegloff (1997b), die Relevanz von Kontextwissen durchaus einbezogen, dies soll jedoch lediglich aus den Daten heraus rekonstruiert werden. Depper-

mann (2000: 100) bezweifelt demgegenüber, dass dies möglich ist, aus drei Gründen: Erstens, weil die Aufzeigeleistungen nicht immer für Gesprächsteilnehmer*innen wie auch für Forscher*innen eindeutig sind, zweitens, weil nicht trennscharf unterschieden werden kann zwischen Kontextaspekten, die im Gespräch relevant sind, und solchen, die dies nicht sind, und drittens, weil Aufzeigeleistungen nicht ohne Vorwissen *erkannt* werden können.

Deppermann steht mit dieser Kritik nicht allein, schon Gumperz (1982; 1999) und Gumperz/Hymes (1972) argumentieren so bei der Begründung der Interaktionalen Soziolinguistik und der Ethnographie des Sprechens (siehe auch Hymes 2004). Sie argumentieren, dass das Erkennen von Kontextualisierungshinweisen nur möglich ist, wenn auch Hintergründe der Teilnehmendenkonstellation sowie kulturelles Vorwissen einfließen.¹³¹ Auch die Forschungen an der Schnittstelle von Linguistik und Anthropologie, z. B. von Duranti (1997; 2008; 2009) oder Foley (1997), die sich der Konstruktion von sozialen und kulturellen Identitäten über sprachliche Praktiken widmen und die Interaktionale Linguistik wesentlich beeinflusst haben, arbeiten mit einem hohen Maß an ethnografischer Anreicherung der linguistischen Daten.

Der überwiegende Teil der *medienlinguistischen* Arbeiten mit gesprächslinguistischen Anteilen aus den letzten Jahren berücksichtigt Hintergrundinformationen, auch wenn der ‚Grad der Ethnografisierung‘ der Arbeit unterschiedlich ausgeprägt ist: Er reicht von knappen Vorabinformationen zu Beziehung und Interaktionsgeschichte, wie sie exemplarisch bei König (2019: 41–45) zu finden sind, bis hin zu umfangreichen Ethnografien, wie etwa bei den Arbeiten von Dang-Anh (2017; 2019) oder Huynh (2020), in denen Interviews, Situationsbeobachtungen oder Feldnotizen eine monatelange Annäherung und ein tiefgehendes Verständnis der kommunikativen Praktiken ermöglichten. Diese verschiedenen Umgänge mit dem Einbezug der Ethnografie sollten nicht gegeneinander ausgespielt werden, sondern sie sind im Sinne des „unique adequacy requirement“ (Garfinkel 2002: 175–176) Möglichkeiten, auf die Fragestellung und den zu untersuchenden Gegenstandsbereich angepasst zu reagieren.

Ethnografie etabliert sich in der empirischen, angewandten Linguistik also zunehmend als Bestandteil der Forschung, wie auch neuere Lehrbücher zur angewandten Sprachwissenschaft (etwa Meer/Pick 2019) oder zur Konversationsanalyse (etwa Birkner et al. 2020) zeigen. In der Ethnologie, Anthropologie und Soziologie hat ethnografisches Arbeiten demgegenüber bereits eine lange Tradition

¹³¹ In diesem Stil arbeiten u. a. auch Kallmeyer/Keim (1994) und Kallmeyer (1995), die eine Verbindung von ethnografischen Beobachtungen und sprachlichen Analysen durchführen, um städtische Sozialkonfigurationen zu untersuchen.

zur Erforschung sozialer Praktiken. Sie ist hier methodologisch schon länger etabliert und hat ihre Wurzeln in der älteren ethnologischen Kulturanalyse, der Chicagoer Schule sowie der phänomenologischen Alltagssoziologie nach Schütz (vgl. Breidenstein et al. 2013: 13; Ayaß 2016: 335–336). Diese Ansätze eint eine methodische Vielfalt, denn „Ethnografie ist keine Methode“ (Breidenstein et al. 2013: 34), sondern vereint vielmehr ein flexibles Methodenspektrum zu einem Forschungsstil, der sich durch einen „feldspezifischen Methodenopportunismus“ (Breidenstein et al. 2013: 34) auszeichnet und insofern weder auf ein bestimmtes Vorgehen noch einen bestimmten Datentyp festgelegt ist. Ethnografisch zu arbeiten, bedeutet, sich den Bedingungen im Feld anzupassen. Es wird versucht, „die Komplexität des untersuchten Felds zu erfassen“ (Ayaß 2016: 337). Das setzt eine zweite Bedingung ethnografischer Forschung voraus: Ethnografische Arbeit zeichnet sich dadurch aus, dass Forscher*innen ins ‚Feld‘ gehen und dort unmittelbare und vergleichsweise lange Erfahrungen im direkten Kontakt mit den Beforschten machen (vgl. Breidenstein et al. 2013; siehe dazu auch Charmaz 2006). Es entsteht, bei allen Arten qualitativer Sozialforschung, aber bei ethnografischen Verfahren im Besonderen, eine andere Beziehung und ein längerer Kontakt zwischen den Forscher*innen und den Beforschten bzw. dem beforschten Feld (vgl. Scholl 2016: 25). Teilnehmende Beobachtung und Protokolle darüber sind üblicherweise Kern einer ethnografischen Vorgehensweise, aber auch Interviews, Aufnahmen, gesammelte Dokumente und Artefakte und auch die eigene Erprobung und Erfahrungen von Praktiken des Feldes können Bestandteile einer Ethnografie sein (vgl. Ayaß 2016: 337). Einwirkungen der Forscher*innen auf das Feld, also deren „Reaktivität“, werden durch reflexive Verfahren, mit denen die Forscher*innen ihr Agieren im Feld in die Analyse einfließen lassen, bearbeitet. Es entsteht so eine „*intellektuelle Distanz* [Herv. i. O.]“ (Breidenstein et al. 2013: 109) zu den Erlebnissen und Erfahrungen; diese Verfahren gehören zur Anatomie der ethnografischen Arbeit (siehe auch Reichertz 1992).

Ethnografisches Arbeiten betont außerdem eine Art Transkription des sozialen Geschehens. Das Geschehen selbst kann, aufgrund seiner Situiertheit, seiner Zeitlichkeit und Flüchtigkeit, seiner Materialität und seiner Singularität, nicht erfasst und untersucht werden, es kann nur greifbar gemacht werden, wenn es darstellbar gemacht wird, und erst durch die Darstellbarkeit wird das Offensichtliche überhaupt sichtbar. Allein dieser Darstellungsprozess ist wesentlicher Bestandteil der reflexiven Leistungen der Ethnograf*innen. Ethnografische Darstellung, die klassischerweise schriftlich erfolgt, aber darauf nicht beschränkt ist (vgl. Dwelling/Prus 2012: 194–196), ist von Einfachheit im Stil geprägt und soll v. a. eine leichte Zugänglichkeit zu den gemachten Beobachtungen ermöglichen. Ethnografisches Arbeiten ist als „offene Forschung“ verstanden explorativ (vgl. Dwelling/Prus 2012: 53–54), d. h. hypothesengenerierend, und geht – ganz ähnlich wie auch

die vorgestellten methodologischen Konzepte der Konversationsanalyse – ohne vorab gebildete Vorannahmen an die Daten heran. Diese würden den Blick verstellen und mit *a priori* angenommenen Kategorien die Analysen zu sehr verzerren. Ethnografie bzw. ethnografische Methoden eignen sich also als Forschungsstil bestens zur Analyse sozialer Praktiken, wie auch die von Schäfer/Daniel/Hillebrandt (2015) zusammengestellten Arbeiten von Lengersdorf (2015), Both (2015) und Porsché (2015) deutlich zeigen. Der Stil erlaubt es, methodisch flexibel auf das beobachtete Geschehen reagieren zu können; er kann unterschiedliche Schwerpunkte setzen und eine ganzheitliche Betrachtung von *in situ* verfertigtem Geschehen ermöglichen, und so praxeologisch die Lebenswirklichkeit einfangen.

In Verbindung mit dem sprachwissenschaftlichen Erkenntnisinteresse bedeutet das: Ethnografie ist erstens ein *Forschungsstil*, dessen methodologische ‚Denkweise‘ auf die Linguistik übertragen werden soll. Er muss sich den linguistischen Zielen nicht unterordnen, sondern soll, an den Methodenzwang des Gegenstands angepasst (vgl. Amann/Hirschauer 1997) – also ausgerichtet auf die Beschreibung des *Sprachlichen* –, eine Haltung sein, mit der sprachliche als soziale Praktiken erfasst werden können. Zweitens ist Ethnografie eine *Notwendigkeit*, um überhaupt sprachliche Praktiken als Bestandteil sozialer Praktiken einfangen zu können; dazu wurde bereits auf den deutlichen Appell von Deppermann (2000) und dessen breite Rezeption in der Sprachwissenschaft, auch über die Gesprächsforschung hinaus, hingewiesen. Drittens ist in ethnografischen Arbeiten eine Vielzahl von Methoden angewendet worden, mit denen unterschiedliche Erfahrungen gemacht werden konnten und die für verschiedene Zwecke mehr oder minder geeignet waren. Auf diese Erfahrungen soll bei der Auswahl der Methoden ebenfalls zurückgegriffen werden, auch wenn sie sich zugleich an den interaktional-linguistischen Notwendigkeiten orientieren muss. Das bedeutet auch, dass bei Sprache und Gesprächen als Forschungsgegenstand Abwägungen zwischen einem intensiven, partizipativen Feldkontakt und z. B. einer in solchen Settings unvermeidbaren Beeinflussung auch der sprachlichen Ausdrücke getroffen werden müssen. Dieses als Beobachterparadoxon (vgl. Labov 1972) formulierte Phänomen wird, auch bei einer Relativierung des etwa von Potter/Wetherell (1987) erhobenen Anspruchs „natürlicher Daten“,¹³² einige ethnografische Methoden ausschließen. Im Gegenzug kommen für eine umfassende Beschreibung sprachlicher Praktiken, die die soziale und materiale Situiertheit ernst nimmt, besonders Methoden in Betracht, die das Geschehen aufzeichnen und es so einer feingranularen Untersuchung zugänglich machen, die durch bloße Beobachtung

132 Für eine kritische Betrachtung dazu siehe etwa ten Have (1999) und sehr differenziert Gerwinski/Linz (2018). Siehe auch die Ausführungen in Kap. 5.3.

nicht ermöglicht wird. Nur so können parallel stattfindende Bestandteile einer sprachlichen Praktik (etwa Verbalisierung und kinesische Bestandteile), die körperlich-räumliche Organisation und Konfiguration von Interaktionen und der Einbezug der materialen Umgebung hinreichend erfasst werden.

4.4 Multimodalität

Goodwin (2012) zeigt, dass sprachliche Äußerungen eingebunden sind „in eine soziale Ökologie der Praktiken zur Bedeutungsherstellung“ (Goodwin 2012: 418), die alle Mitglieder einer Gesellschaft umgibt und an der wir uns alle ebenso orientieren wie wir sie gleichermaßen mitgestalten. Diese Ökologie umfasst zwar sprachliche Ausdrücke und auch wenn diese unbestritten einen hohen bedeutungsstragenden Anteil haben, so argumentiert Goodwin (2000: 1490), dass Sprache nicht isoliert betrachtet werden kann, sondern einen Analyseansatz benötigt, „that takes into account the simultaneous use of multiple semiotic resources by participants“; dazu gehören verschiedene Zeichensysteme, die situative und sozialräumliche Umgebung sowie die sequenzielle Organisation selbst. Goodwin wendet sich damit gegen einen Dualismus von Text und Kontext und plädiert für eine ganzheitliche Betrachtungsweise der Interaktion und ihrer Beteiligten.¹³³ Goodwin postuliert, dass die Kategorien Text und Kontext nicht von Analytiker*innen *a priori* gesetzt werden können, sondern dass diese Unterscheidung Gegenstand einer gemeinsamen Hervorbringung und interaktiven Aushandlung ist (vgl. Goodwin 2000: 1519; 1981; 1984).

Der Mainstream der Sprachwissenschaft hat diese Untersuchungsbereiche, auf die Goodwin bereits in seinen frühen Arbeiten eingeht, erst um die Jahrtausendwende für sich entdeckt: Aspekte von Gestik, Mimik, Körperlichkeit und Raum sind erst spät in den Blick der Linguistik gerückt (siehe etwa Mondada 2007). Die technischen Entwicklungen im Bereich der Datenerhebung und -aufbereitung machten es möglich, auch Videoaufnahmen von Interaktionen zu erheben und den Aufwand

¹³³ Schüttpelz (2020: 139) betont die Zusammengehörigkeit des (nach Goodwin 2018) *akkumulativen* Charakters der sprachlichen Interaktion mit deren *altruistischen* Eigenschaften: „Vielleicht [...] kann man sagen, dass ‚Sprache‘ nichts anderes ist als menschlicher ‚Altruismus‘, oder im Gegenzug, dass sich der ‚Altruismus‘ des Menschen zuerst und zuletzt in seiner ‚Sprachlichkeit‘ verwirklicht“. Es ist dieser Altruismus, die Möglichkeit einer Assistenz zum Gelingen einer erfolgreichen Interaktion, der sich ganz besonders aus der sozialen Ökologie speist: Sprachliche Erscheinungen sind ein Teil davon, ebenso die Interaktionsgeschichte, die gemeinsamen Beziehungen und die räumliche Umgebung, die über Gestik und Mimik in die Interaktion einbezogen werden kann (vgl. Schüttpelz 2020: 138–142).

für ihre Verarbeitung erheblich zu reduzieren (vgl. Deppermann/Schmitt 2007: 15), und führten so zu einer Ausdehnung der Analysebereiche auch in den Bereich der Gestik, Mimik und des interaktionalen Raums. Die Mechanismen der Organisation sprachlicher Interaktion wurden in der Tradition der frühen konversationsanalytischen Arbeiten zu großen Teilen anhand von Telefongesprächen erschlossen (vgl. Sacks/Schegloff/Jefferson 1974; Schegloff/Jefferson/Sacks 1977; Jefferson 1978). Es konnten auf diese Weise grundlegende Erkenntnisse über die konversationelle, sequenzielle Organisation der Interaktion gewonnen werden, die Rolle anderer Ausdrucksressourcen konnte dabei allerdings aufgrund der Charakteristika von Telefongesprächen kaum bearbeitet werden. Dabei soll nicht übersehen werden, dass Gestenforschung bereits weitaus früher begann, etwa mit den wegweisenden Arbeiten von Kendon (1972; 1980; 2004), Kendon/Sebeok/Umiker-Sebeok (1981) oder auch Ekman/Friesen (1969); diese nehmen jedoch keine an der Praxis orientierte Perspektive ein, gehen nicht sequenzanalytisch vor und sind insofern weniger an der interaktiven Hervorbringung sozialer (und insofern kommunikativer und sprachlicher) Praktiken interessiert.¹³⁴ Einschlägigere Vorarbeiten im Bereich der Multimodalität leistete insbesondere Charles Goodwin; sein Zusammendenken von Sprache und materialer und räumlicher Umgebung (Goodwin 1994; 1981; 1996; 2000; 2003; 2007; 2012) war eine wesentliche Vorbereitung, bevor der *multimodal turn* insbesondere den empirischen Zweig der Linguistik erfasste.¹³⁵ Dieser wirkte sprachwissenschaftlich prägend: Die Bedeutung der Multimodalität in der Interaktion wird in heutigen Arbeiten kaum noch unterschätzt. Gleichwohl gibt es eine fortlaufende Debatte darüber, was genau unter Multimodalität zu verstehen ist. Ein einheitliches Verständnis dieses Terminus ist bislang nicht gefunden, wie auch Deppermann (2013: 2) betont und eine Verwendung des Begriffs angepasst auf den jeweiligen Kontext (hier: Konversationsanalyse, Gleiches gilt aber auch für die Adaption an andere Subdisziplinen) anmahnt:

[...] „[M]ultimodality“ is a label which is already worn out and has become most fuzzy by its use in various strands of semiotics, discourse and media analysis. So, there may be the danger that under the heading of ‚multimodal interaction‘ a lot of things will gather which do not have to do much with CA. (Deppermann 2013: 2)

¹³⁴ Theoretische Grundlagen für deren Erforschung im linguistischen Kontext lieferte bereits Bühler (1965[1934]) in seiner Sprachtheorie.

¹³⁵ Wie Selting (2013: 590) betont, nahmen gelegentlich auch andere Konversationsanalytiker*innen zusätzliche Aspekte abseits des verbal-vokalen Ausdrucks in den Blick, so etwa Schegloff (1984) oder Lerner (2003), doch wurden diese hier eher als Mittel zum Zweck einbezogen und nicht systematisch in ihrer Relevanz für die Interaktionsorganisation untersucht.

Der vorliegenden Arbeit liegt ein Verständnis von Multimodalität zugrunde, das sich an der Pluralität und Gleichzeitigkeit der verkörperten Ausdrucksressourcen orientiert. Mondada (2021: 44) konstatiert, dass deren Ausgangspunkt in der Erkenntnis ethnomethodologischer Arbeiten über die Körperlichkeit des Sprechens liegt, die die „artificiality of separating language and the body in action“ hervorheben. Die Organisation der Interaktion beruht nicht nur auf sprachlichen Verfahren (in denen Lexik, Syntax, Prosodie usw. kombiniert werden), sondern auch auf Gestik, Blick, Körper(haltung), räumlichen Bewegungen und der physischen Umgebung und ist somit inhärent *multimodal* (vgl. Mondada 2021: 45).¹³⁶ Entscheidend ist, und so soll Multimodalität verstanden und in den Analysen bearbeitet werden, woran die Gesprächsbeteiligten *in der Praxis* die Sinnkonstitution festmachen und welche Ressourcen sie dafür einsetzen. Dabei kann in einem praxeologischen Sinne sowohl der eigene Körper wie auch die materiale und imaginäre Umgebung relevant werden und auch die Vergangenheiten und imaginierten Zukünfte dieser (vgl. Goodwin 2018), auf die ebenfalls mit diesen Ressourcen verwiesen werden kann.¹³⁷ Keine der Modalitäten kann dabei *a priori* als dominant gegenüber einer anderen Modalität aufgefasst werden (vgl. Stivers/Sidnell 2005: 2) – dies kann nur empirisch gezeigt werden:

When people interact within embodied social frameworks that are structured and changed through their shifting co-presence, analysts should attend to what the participants themselves are treating as important. (Streeck/Goodwin/LeBaron 2011: 12)

In einem methodologisch ganzheitlichen Ansatz, wie er bisher skizziert wurde, sollen die verschiedenen Ausdrucksressourcen also gleichermaßen Berücksichtigung finden – je nachdem, in welchem Maße die Beteiligten selbst Sinn an diese koppeln. Ausgehend von der bisherigen Multimodalitätsforschung und den dabei berücksichtigten Modalitäten ist insbesondere auf folgende Aspekte zu achten, die schon in den bisherigen Ausführungen immer wieder genannt wurden: Körperlichkeit, Räumlichkeit und Simultaneität sowie das Wechselspiel mit Sprache (siehe dazu etwa Fricke 2013; 2015).¹³⁸ Auf diese drei Aspekte wird nachfolgend kurz eingegangen.

136 Für dieses in Studien zur sozialen Interaktion überwiegende Verständnis von Multimodalität als Kombination verschiedener Ausdrucksressourcen in der Praxis der Interaktionsorganisation siehe Mondada (2014b). In Abgrenzung zu disziplinär anders gelagerten Verständnissen, etwa in der Computerwissenschaft oder der Semiotik, siehe Mondada (2021: 45).

137 Siehe dazu auch das Konzept der „Deixis am Phantasma“ (vgl. Bühler 1965[1934]) und neuere Adaptionen, z. B. von Stukenbrock (2015) zur Deixis in der Face-to-Face-Interaktion; siehe auch Heller (2018) sowie Selting (2013: 590).

138 Nach Fricke (2012: 47–49), die ein stärker semiotisch orientiertes Multimodalitätsverständnis vertritt, ist Multimodalität im engeren Sinn erst dann gegeben, wenn sich *ein* Zeichensystem auf mehreren wahrnehmbaren Kanälen darstellt, z. B. bei redebegleitenden Gesten. Um dadurch ent-

Körperlichkeit: Interaktion ist nach Hausendorf (2015: 52) durch die „Inanspruchnahme humanspezifischer Körperlichkeit“ materialisiert. Damit sind zum einen die sprachlichen Äußerungen selbst gemeint, die zu ihrer Produktion wie Rezeption auf einen menschlichen Körper angewiesen sind – in ihrer einfachsten Form sind dafür keine weiteren Hilfsmittel notwendig. Darüber hinaus sind aber andererseits nicht nur die verbalsprachlichen Äußerungen, sondern auch ganze Körper mit ihrer Materialität an der Sinnkonstitution beteiligt. Dazu kann zunächst die Gestik gezählt werden, mit deren Bedeutung in der Interaktion und praktischer Sinnkonstitution sich schon früh Charles Goodwin auseinandersetzt und die auch Jürgen Streeck eingehend untersucht hat.¹³⁹ Streeck (2016: 77) sieht in sprachbegleitenden Gesten Praktiken des Beschreibens, d. h. „habitualisierte[] Methoden [Herv. i. O.]“, die wiederkehrend zur Lösung bestimmter kommunikativer Aufgaben, d. h. in der kommunikativen Praxis, eingesetzt werden, die aber kontextlosgelöst nicht als konventionalisierte Formen beschrieben werden können. Zeigegesten können in hohem Maße in eine Lokaldeixis unterschiedlicher Art eingebunden sein (vgl. Uhlmann 2013; Stukenbrock 2015; 2016). Streeck (2013: 679) lenkt in seiner „Praxeologie der Gestik“ die Aufmerksamkeit weg von der Beziehung, die verbale, visuelle, kognitive oder andere Teile untereinander haben; eine Praxeologie der Gestik ist stattdessen „interested in the ways in which gestures participate in the bringing about of shared practical understandings“ (Streeck 2013: 679) – Gestik wird als Praxis, Gesten werden als Praktiken betrachtet. Insbesondere bei bildhaften Gesten wird, so Streeck (2013: 681), der Vorteil einer praxeologischen Beschreibung deutlich: Es gibt allenfalls in Ansätzen ein kodifiziertes, kanonisiertes Repertoire von bildhaften Gesten, in der Mehrzahl der Fälle „when a specific gestural depiction is needed, it is improvised. The depiction may or may not include established depictive forms“ (Streeck 2013: 679). Diese Betrachtung

stehende Abgrenzungsproblematiken zu vermeiden, soll den Abschattungen von Multimodalität (engerer/weiterer Sinn), wie sie auch Imo/Lanwer (2019: 248) einführen, an dieser Stelle nicht gefolgt werden, sondern all diese Fälle als multimodale Interaktion begriffen werden. Sprachliche Interaktion ist mithin *immer* multimodal – auch sprachliche Äußerungen kombinieren verschiedene Ausdrucksressourcen (etwa Lexik und Prosodie, um nur die auffälligsten zu nennen).

139 Erneut sollen die Arbeiten von Adam Kendon, auf die bereits einführend verwiesen wurde, nicht übersehen werden, auch weil Kendon (2004) eine durchaus differenzierte Klassifizierung verschiedener Gestentypen vorlegt. Doch werden diese in Ermangelung einer interaktional oder praxeologisch orientierten Untermauerung in den Analysen nicht fruchtbar gemacht und auf eine genauere Darlegung insofern an dieser Stelle verzichtet; für eine fruchtbare Anwendung von Kendons Arbeiten in der Linguistik siehe etwa Huynh (2020), die eben diese Unterfütterung selbst als Gegenstand der Arbeit im Kontext der Darstellung und interaktional-praktischen Aus handlung von Emotionen leistet.

tungsweise soll helfen, die *in situ* vollzogenen Gesten zu erfassen und den praxeologisch-theoretischen Anspruch der Arbeit auch methodologisch einzulösen.

Körperlichkeit bleibt allerdings nicht auf Verbalsprache und Gestik beschränkt, auch wenn diese besondere Aufmerksamkeit erfahren haben. Mit einem kompatiblen theoretischen Hintergrund wurde auch zu anderen körperlichen Ressourcen gearbeitet, z. B. wurde bereits früh die Rolle von Blickverhalten in der sozialen Interaktion erkannt (vgl. Kendon 1967) und in seiner Rolle für die Turn-Koordination (vgl. Goodwin 1980; Mondada 2007) oder für den Spracherwerb (vgl. Tomasello 1988) untersucht und weitere Arbeiten darauf aufgebaut, die auch das Zusammenspiel dieser beiden Ressourcen in den Blick nehmen (vgl. Sidnell 2006; Mondada 2007; Streeck 2014). Methodologisch wurde die Einführung auf den Blick v. a. durch den Einsatz von Eye-Tracking-Technologien bereichert, Brône (2018) legen in ihrem Band eine Reihe aktueller Forschungen vor, die z. B. auch die Rolle von Blick beim Turn-Taking (vgl. Auer 2018) und bei Mensch-Computer-Dialogen berücksichtigt, wie auch die Arbeiten von Gehle et al. (2015) oder Pitsch et al. (2009b; 2017) zeigen. Darüber hinaus ist, wie Streeck/Goodwin/LeBaron (2011) argumentieren, Körperlichkeit in der Interaktion auch über Gestik und Blickverhalten hinaus eine entscheidende Voraussetzung für eine gemeinsame Ausrichtung aufeinander. Diese bildet, zeitlich entkoppelt von den unmittelbaren Äußerungen und Bewegungen der Gesprächsteilnehmer*innen, als „embodied participation framework“ (Streeck/Goodwin/LeBaron 2011: 2; siehe auch Goodwin 2002) eine Grundlage für die Gesprächsorganisation und Aushandlung der Teilnahme am Gespräch, auch für – insbesondere von Stukenbrock (2015; 2018 u. a.) untersuchte – multimodale deiktische Praktiken, die erst dadurch verständlich werden, dass Körperhaltung, Blick und räumliche Umgebung einbezogen werden; „embodied interaction“ unterscheidet sich somit in seiner Gesamtheit von spezifischen Gesten und Blick und nimmt die übergreifende Struktur der Körperlichkeit mit in den Blick:

One phenomenon that quickly emerges from records that preserve not only the talk but also the bodies of actors, is that action is built through the mutual elaboration of diverse semiotic resources with quite different properties, each of which, including language, can make only a partial, incomplete contribution to the action in progress. (Streeck/Goodwin/LeBaron 2011: 3)

Ein weiterer Hinweis auf die Relevanz dieser körperlichen Praktiken ist in der Forschungstradition der Ethnografie zu finden, die sich bereits früh mit Körpern beschäftigt und deren Rolle bei dem Vollzug sozialer Praktiken ernst genommen hat. Im Rahmen verschiedener Studien haben ethnografische Arbeiten herausgestellt, dass körperlicher Ausdruck als „culturally shaped and socially acquired“ (Meyer 2013: 235) betrachtet werden muss.

Räumlichkeit: Eng verbunden mit den körperlichen Ressourcen, der Körperhaltung und dem Blickverhalten ist die räumliche Umgebung, teilweise ist auch von „visuospatial modalities“ die Rede (Stivers/Sidnell 2005: 2), was die Verbindung von Sichtbarkeit und Räumlichkeit betont. Dabei soll Stukenbrock (2013a: 402) gefolgt werden, wenn sie die These aufstellt, dass eine Gemeinsamkeit der Ansätze zur Betrachtung von Raum und Interaktion die (überwiegend) geteilte Annahme ist, „that space is not a given entity, but an interactive achievement of the participants“ und insofern nicht vom „interactional space“, sondern vom „interactive achievement of space“ die Rede sein sollte (siehe auch Mondada 2013; Hausendorf 2013). Darüber hinaus kann der Raum nicht nur ein gemeinsames „achievement“ sein, sondern auch eine interaktional genutzte Ressource zur Herstellung von „co-orientation, co-ordination and co-operation“ (Hausendorf 2013: 278). Diese wird von Sprecher*innen zwar systematisch eingesetzt, deren Bedeutung wurde aber im Vergleich zu visuellen und verbalen Ressourcen lange übersehen (vgl. Hausendorf 2013: 278). Um den Raum als Ressource nutzen zu können, d. h. ihn als „räumliche Praktik“ zu funktionalisieren, muss er wechselseitig von den Interagierenden als solcher verfertigt werden (siehe dazu auch Stukenbrock 2012). Daher ist die räumliche Umgebung sowohl interaktionales „achievement“ wie auch Ressource, interaktionaler Raum ist ein „shared ,here‘ for perception, movement and action“ (Hausendorf 2013: 277) und – so Hausendorf weiter – gebunden an das „emplacement“ der Gesprächsteilnehmenden, d. h. leiblich in den Vollzug der Praxis eingebunden. Erst in diesem Raum können Körperlichkeit und Körperhaltungen zu Körperformationen und somit zu interaktionalen Arrangements werden.

Simultaneität: Körperlichkeit und Räumlichkeit sind nur insofern mögliche, konstituierende Bestandteile von Multimodalität, als sie *parallel* zu anderen Modalitäten auftreten. Das bedeutet nicht, dass die Modalitäten durchgängig zueinander synchron sein müssen. Zwar sind sie es teilweise, z. B. bei der hohen Synchronität von prosodischen Eigenschaften wie Akzent und gestischen Eigenschaften wie Signifikanzpunkten, die rhythmischen „Taktstockgesten“ eigen sind (vgl. Müller 1998: 103; Loehr 2007). Doch, wie Levinson (2006: 53) ausführt, ist Temporalität allein nicht ausreichend:

[...] synchrony alone will not do the trick of hooking up the bits in the different signal streams: gestures, facial expressions, nods, and the like can come earlier or later than the words they go with. If temporal binding is not sufficient, what will do the trick?

Nach Levinson ist es nicht eine Frage des Ablesens, der genauen zeitlichen Parallelität oder Sequenzialität von bestimmten Ausdrücken, sondern auch hier eine Frage der Sinnkonstitution. In der Praxis können visuelle, geräuschhafte, räumliche oder körperliche Praktiken entsprechend nicht nur untermalend gebraucht

werden, sondern dafür eingesetzt werden, zwei kommunikative Probleme gleichzeitig zu lösen – z. B. metakommunikative Aufgaben wie Turn-Taking parallel zu bearbeiten (vgl. etwa Mondada 2007). Somit laufen nicht nur intrapersonelle, sondern auch die interpersonelle Koordination in hohem Maße multimodal ab. Im Sinne einer „opportunistischen und gefräßigen Praxis“ (Goodwin 2018: 445) können alle in einer bestimmten situativen Umgebung zur Verfügung stehenden Ressourcen – verbal-vokale ebenso wie visuelle, körperliche und räumliche – funktionalisiert werden. Es ist insofern auch methodologisch zentral, nicht nur eine sich temporal-sequenziell entfaltende Praxis im „Nacheinander“ anzunehmen, sondern auch ein „Miteinander“, so wie auch Charles Goodwin in seiner Methodologie ausgeht von „social events as dynamic, emergent, intertwined totalities, not as aggregations of individual acts that are sequentially performed“ (Schüttpelz/Meyer 2018: 174).

Ergänzend zu diesen drei für die vorliegende Arbeit wesentlichen methodologischen Dimensionen von Multimodalität soll noch einmal auf das Konzept der verbalen, vokalen visuellen Praktiken und Ressourcen von Selting (2013) verwiesen werden,¹⁴⁰ das insbesondere hilfreich ist, um zu unterscheiden zwischen den rein visuellen Ausdrucksmöglichkeiten (Blick, Gesichtsausdruck, Gestik, Haltung, Einbezug von Objekten usw.) bzw. verbalen (d. h. auf allen linguistischen Beschreibungsebenen sprachlichen) und vokalen (d. h. insbesondere stimmlich-qualitativen und prosodischen) Ressourcen auf der einen und ihrer Anwendung in einem bestimmten Kontext „as they are deployed for the signalling of interactional meaning“ (Selting 2013: 591) auf der anderen Seite. Dieser Begriff der Praktik reduziert die Rolle der sozialkulturellen Hintergründe (ohne sie auszublenden) und stellt stattdessen auf die Interaktionsorganisation auf der mikrostrukturellen Ebene scharf, was für die Beschreibung der sprachlichen Interaktion ebenfalls eine notwendige Voraussetzung ist.

4.5 Zwischenfazit

Es soll also in diesem Sinne mit einer aus der ethnomethodologischen Konversationsanalyse entsprungenen und von der Interaktionalen Linguistik informierten Gesprächsanalyse gearbeitet werden, d. h. primär mit Sequenzanalysen von Beispielen aus dem erhobenen Korpus, die – wo von den Daten und den Unterfragestellungen her passend – um Aspekte der Ethnografie sowie qualitativ-interpretative Videoanalysen (siehe Schmitt 2013; Tuma 2018) erweitert wird. Es

¹⁴⁰ Siehe auch Kap. 2.1.3.

wurde deutlich, dass für die Analysen die „Prägnanz und die Wucht der visuellen Anteile an der Interaktionskonstitution“ (Schmitt/Hausendorf 2016: 9) berücksichtigt werden muss, soweit dies im Rahmen der erhobenen Daten möglich ist. Eine multimodale Herangehensweise ist – wie Schmitt/Hausendorf (2016: 9) betonen – Voraussetzung für die Untersuchung von Praktiken in einem visuell-räumlichen, interaktionalen Setting, bei dem u. a. auch die Konfiguration soziotechnischer Arrangements im Raum in einer laufenden Interaktion untersucht werden soll, um der „sichtbaren Verankerung verbaler Strukturen in räumlich-situativ strukturierten Handlungszusammenhängen“ (Schmitt/Hausendorf 2016: 10) Rechnung zu tragen.

Tuma (2018: 426) unterscheidet bei qualitativ-interpretativen Videoanalysen, die der ethnomethodologischen Konversationsanalyse in ihren Grundannahmen folgen, die „linguistische Multimodalitätsforschung“ sowie die „Video-Interaktionsanalyse“, wobei sich die vorliegende Arbeit nicht zum Ziel gesetzt hat, eine durchgehend multimodale Beschreibung zu liefern, sondern diese einzubeziehen, wo es im Hinblick auf die Erhebungsmodalitäten und die Fragestellungen sinnvoll und möglich ist. Entsprechend werden multimodal ausgerichtete Analysen vornehmlich zu Situationen angefertigt, in denen die Ersteinrichtung des Smart Speakers im Fokus steht – hierzu liegt Videomaterial vor.¹⁴¹ Videoaufnahmen, die auch die kontinuierliche Nutzung von VUIs dokumentieren, wären wünschenswert, konnten jedoch im Rahmen dieser Arbeit nicht erhoben und ausgewertet werden.¹⁴² Um sprachliche Praktiken umfassend einzufangen, bietet sich gerade eine Kombination der sprachwissenschaftlichen Einführung mit ethnografisch integrierten Bestandteilen besonders an (vgl. Deppermann 2000). So soll ein ganzheitliches Bild der Praxis gewonnen werden, durch das in den einzelnen Untersuchungen möglichst viele der für die sprachlichen Praktiken relevanten Aspekte aufgedeckt werden können. Eine solche Vorgehensweise erweist sich gerade für medienlinguistische Fragestellungen als tauglich, wie bereits in exemplarischen praxeologisch grundierten Analysen von Videoaufnahmen der Smart Speaker-Ersteinrichtung gezeigt werden konnte (vgl. Hector 2022; Habscheid et al. 2021; Habscheid/Hector/Hrncal 2023).

Die Methodologie dieser Arbeit ist insgesamt qualitativ-explorativ und somit hypothesengenerierend ausgerichtet. Ziel ist es, Phänomene zu entdecken, die für die untersuchten Teilnehmer*innen von Bedeutung sind, mit denen sie Sinn und Ordnung in VUI-Dialogen und Gesprächen unter Einbezug eines VUIs herstellen. Dazu stehen sequenzanalytische Detailbetrachtungen einzelner Ausschnitte im Vordergrund, die ein verstehendes Betrachten von VUI-Dialogen in ihrem Kontext

¹⁴¹ Siehe Kap. 5.2.

¹⁴² Zu sich daraus ergebenden Limitationen und Folgefragen siehe Kap. 7.

ermöglichen. Ein Anspruch auf Repräsentativität besteht insofern nicht und war auch nie das Ziel der Untersuchung. Im Sinne qualitativer Forschung ist aber sehr wohl die „Relevanz der gefundenen Strukturen“ (Scholl 2016: 28) sicherzustellen und durch eine Diversität im Hinblick auf (*a posteriori* gebildete) Strukturen und Materialien sowie durch Abstraktion und Reflexion der Forschungspraxis zu festigen. Gütekriterien für die Arbeit sollen im Rahmen des Untersuchungsdesigns festgelegt werden – sie können für qualitative Arbeiten nicht auf der abstrakten Ebene vordefiniert, sondern nur „unter Berücksichtigung der jeweiligen Fragestellung, Methode, der Spezifik des Forschungsfelds und des Untersuchungsgegenstands“ (Steinke 2007: 323) diskutiert werden. Auch Deppermann (2008: 105) stellt fest: „Die Güte einer Untersuchung ist nicht an punktuellen Merkmalen festzumachen“. Orientierung können bei dieser im Folgekapitel zu führenden Diskussion die von Steinke (2007) formulierten Kriterien geben, ebenso wie drei von Deppermann (2008: 105–110) genannte Aspekte: „Qualität der Datenmaterialien“, „Durchführung und Präsentation der Gesprächsanalyse“ sowie die „Frage der Generalisierbarkeit“. Wie das bisher zur methodologischen Einordnung Gesagte, heruntergebrochen auf ein konkretes Untersuchungsdesign und die Auswahl bzw. Zurichtung von bestimmten Methoden, in die konkrete Forschungspraxis und die Analysen einfließt, soll im nachfolgenden Kapitel bearbeitet werden.

5 Methodik: Projektkontext, Untersuchungsdesign, Erhebung, Datentypen und -aufbereitung

Vor dem Hintergrund der theoretischen und methodologischen Annahmen ist evident, dass die Untersuchung auf Basis realer (d. h. nicht erfundener), möglichst authentischer Interaktionsdaten von der Nutzung von Smart Speakern im Privathaushalt zu erfolgen hatte. Startpunkt für die Erhebung war die Identifikation zweier potenziell aufschlussreicher Situationen: die Ersteinrichtung eines Smart Speakers im Privathaushalt sowie die alltägliche Nutzung der Geräte und ihre Einbettung in die häusliche Praxis. Während Erstere auf Video aufgezeichnet wurden, konnte die Zweite mit Hilfe einer speziellen Technologie, eines „Conditional Voice Recorders“ (CVR),¹⁴³ erhoben werden. Die Videos umfassen dabei jeweils einen Zeitraum von ca. 20 bis 25 Minuten, die Audio-Aufnahmen ca. sechs Minuten. Die entstandenen Datentypen sowie das Vorgehen bei ihrer Erhebung sollen im Folgenden genauer erläutert werden, um das Untersuchungsdesign transparent zu machen. Zuvor soll aber der Kontext des Forschungsprojekts, in dessen Rahmen die Datenerhebung für diese Arbeit stattgefunden hat, vorgestellt werden.

5.1 Projektvorstellung und Untersuchungsdesign

Die vorliegende Dissertation entstand im Rahmen des von der DFG geförderten Sonderforschungsbereichs 1187 „Medien der Kooperation“ an der Universität Siegen,¹⁴⁴ genauer im Teilprojekt B06 „Un/erbetene Beobachtung in Interaktion: ‚Intelligente Persönliche Assistenten‘“. Die Daten, die in den nachfolgenden Kapiteln ausgewertet werden, entstanden im Rahmen dieses Projekts, in dem auch primär der Verfasser für die Erhebung und Aufbereitung der Daten zuständig war. Andere Projektmitglieder, die mit den linguistischen Fragestellungen des Projekts befasst waren, waren Stephan Habscheid (Projektleitung) und Christine Gebhard (geb. Hrncl). Die Projektgruppe wurde insbesondere in der Transkription der Aufnahmen von studentischen Mitarbeiter*innen unterstützt.¹⁴⁵ Neben dem lin-

¹⁴³ Siehe dazu Porcheron et al. (2018), Hector et al. (2022) sowie ausführlicher Kap. 5.3.

¹⁴⁴ Siehe <https://www.mediacoop.uni-siegen.de/de/>.

¹⁴⁵ Für Arbeiten an den Transkripten danke ich an dieser Stelle noch einmal Viviane Börner, Franziska Petri, Franziska Niersberger-Gueye, Sarah Diehl, Marie Dienst, Katharina Meiers, Chris Dangelmeier, Leonie Tittel, Johanna Klein und Finnja Strunk.

guistischen Teil hat das Projekt auch in einem mediensoziologischen Teil geforscht. Dabei wurde mit denselben Haushalten gearbeitet, die auch der linguistische Teil untersucht, und die Erhebung nahezu vollständig gemeinsam geplant und durchgeführt. Die Mediensoziolog*innen führten in diesen Haushalten jeweils zu zwei Zeitpunkten – einmal kurz nach der Ersteinrichtung der Smart Speaker und einmal mit einem Abstand von mehreren Monaten – Interviews durch, die in Teilen auch die vorliegenden Untersuchungen informieren und im Sinne eines ethnografisch angereicherten Vorgehens Hintergrundinformationen liefern.

Das Untersuchungsdesign des Projekts sowie die Rahmung durch das Forschungsprogramm des Sonderforschungsbereichs (SFBs) hatten erstens Auswirkungen auf die Gestaltung der vorliegenden Arbeit, die sich in den theoretischen und methodologischen Ausführungen wiederfinden lassen (siehe auch Habscheid et al. 2025b): Die praxeologische Orientierung, das beschriebene Medienverständnis sowie der empirische Anspruch der Arbeit sind wesentlich von programmatischen Arbeiten im SFB informiert. Zweitens wurde durch die im Projekt angelegte Vorgehensweise nicht determiniert, aber präkonfiguriert, welche Situationen überhaupt zur Erhebung und Auswertung in Frage kommen würden. Die Bestimmung der Erhebungssituationen erfolgte zunächst tentativ auf Grundlage einer theoretischen Beschäftigung mit Smart Speakern und VUI-basierten Systemen: Dabei waren zwei Situationen als potenziell ertragreich identifiziert worden: Typ 1 war die Erstinstallation im Haushalt, in der die Nutzung erstmals ausprobiert und gemeinsam reflektiert werden würde, Typ 2 das gemeinsame Ausprobieren von Smart Speakern von Haushaltsmitgliedern untereinander sowie mit Gästen. Die Gestaltung der Gesprächsdynamiken sollte insbesondere durch Aufzeichnungen untersuchbar gemacht werden, in denen mehrere Personen an der Nutzung der Geräte beteiligt sind. Diese Planungen wurden im Rahmen einer Pilotstudie (siehe Hector/Hrncal 2020) einer Überprüfung unterzogen, um Modifikationen vornehmen und das Vorgehen bei der Erhebung praktisch planen zu können. Beide Situationstypen wurden als prinzipiell sehr aufschlussreich bewertet; zugleich wurde deutlich, dass es in den Datenerhebungssituationen notwendig sein würde, die Studienteilnehmer*innen detaillierter zu instruieren (vgl. Hector/Hrncal 2020: 8). Dabei waren v. a. zwei Aspekte wichtig: Zum einen sollte deutlich werden, dass nicht die Erwartung besteht, dass die Ersteinrichtung für eine Kamera im parainteraktiven Stil (Horton/Wohl 1956; Luginbühl 2019) „vorgeführt“ und interessant gemacht werden muss. Zweitens stellte sich heraus, dass die Darstellungen auf den Smartphone-Bildschirmen für den Verlauf der Interaktion essenziell waren, sodass sie – wo möglich – miterhoben werden sollten. Ferner wurde festgestellt, dass Hinweise zum Bildausschnitt und zur Bedienung der Kameras notwendig wären (vgl. Hector/Hrncal 2020: 9).

5.2 Datentyp I: Videoaufzeichnungen der Smart-Speaker-Ersteinrichtung

Der erste Datentyp, die *Erstinstallation* des Smart Speakers in den Haushalten, war der Startpunkt für eine längere Erhebungsphase mit den Haushalten. Zu diesem Zeitpunkt wurden auch die erwähnten Interviews geführt auch eine zweite, noch genauer zu beschreibende Phase zur Erhebung von Daten aus der alltäglichen Nutzung (siehe Kap. 5.3) begann zeitgleich. Um diese Aufnahmen anfertigen zu können, war es notwendig, Haushalte zu identifizieren, die kurz vor der Anschaffung eines Smart Speakers standen – d. h. idealerweise bereits eine Kaufentscheidung getroffen, aber noch nicht umgesetzt hatten – oder diesen bereits gekauft, aber noch nicht zu Hause eingerichtet hatten. Die Teilnahme an der Studie, die mit einer kleinen Aufwandsentschädigung honoriert wurde, sollte so wenig wie möglich als Motivationsfaktor für die Kaufentscheidung dienen, um sicherzustellen, dass die Sprecher*innen auch Nutzer*innen von Smart Speakern sein könnten, wenn sie nicht Teilnehmer*innen der Studie geworden wären. Trotz des hypothetischen Charakters dieser Bedingung zeigte sich in den Daten (z. B. in den Angaben über Kaufentscheidungsmotivation und Nutzungsanlässe), dass diese Entscheidung richtig war. Die Ersteinrichtungssituation schien unter verschiedenen Gesichtspunkten im Vorfeld als potenziell ergiebig: Erstens war zu erwarten, dass Fragen des Verhältnisses von Interaktion, Raum und Architektur bzw. sozialer Beziehung (vgl. Schmitt/Hausendorf 2016) verhandelt werden würden, gerade bei der Platzierung des Geräts. Außerdem war anzunehmen, dass erste Einstellungen, z. B. zum Datenschutz, zu potenziellen Anwendungskontexten und zu interaktionalen Verfahren der Einbindung in laufende Gespräche und somit auch der ‚Beteiligung‘ an der sozialen Praxis, ebenso sichtbar würden wie Vorgänge des Einlernens in und des Übernehmens von gesprochensprachlichen Formen zur Bedienung der Geräte.

Eine der Erkenntnisse aus der Pilotstudie konnte im Hinblick auf die Erhebung der Ersteinrichtungsvideos nicht umgesetzt werden. Dort heißt es: „Insgesamt empfiehlt sich die Anwesenheit der Forschenden, die die technische Seite betreuen und einen geeigneten Bildausschnitt auswählen können“ (Hector/Hrncal 2020: 8). Durch die Kontaktbeschränkungen während der Covid-19-Pandemie, die in die Hauptzeit des Erhebungszeitraums fiel, konnte dies nicht wie geplant umgesetzt werden. Der im Projekt gefundene Umgang damit hatte jedoch forschungspraktisch gesehen auch Vorteile: Dadurch, dass persönliche Besuche bei den Haushalten aufgrund des Infektionsschutzes nicht möglich waren, wurden auch Haushalte aus Orten in die Studie aufgenommen, die ansonsten aufgrund des Reiseaufwands zu weit entfernt gewesen wären.

Mit Bezug auf die „Natürlichkeit“ der Daten können die Umstände insgesamt ebenfalls positiv bewertet werden. Gerwinski/Linz (2018: 107) identifizieren erstens die „Legitimität“ bzw. den Grad der Arrangiertheit der aufgezeichneten Situation und die Invasivität des Aufnahmeverfahrens sowie zweitens dessen Berücksichtigung in der Auswertung der Daten als methodische Herausforderungen was die Natürlichkeit angeht.¹⁴⁶ Zum ersten Punkt kann gesagt werden, dass der „Arrangiertheitsgrad“ der Aufnahmesituation sinkt, wenn die Forscher*innen selbst nicht anwesend sind: Die zeitliche Flexibilität ist höher und die Situation bekommt einen weniger experimentellen Charakter, die Äußerungen sind insgesamt weniger stark elizitiert als wenn eingebettet in einen Besuch vor Ort – der „Sprechanlass“ (Schank 1979: 74) wird vom Besuch der Forscher*innen hin zum tatsächlichen Einrichten des Smart Speakers verschoben. Damit soll nicht gesagt werden, dass andere Vorgehensweisen nicht legitim sind, und häufig sind Grenzziehungen zwischen elizitierten und spontan entstehenden Gesprächsdaten ohnehin schwierig oder unmöglich (vgl. ten Have 1999: 49). Gleichwohl kann hier konstatiert werden, dass die Abwesenheit der Forscher*innen die Legitimität einer konversationsanalytischen Betrachtung des aufgezeichneten Materials eher erhöht als geschwächt hat. Zweitens ist evident, dass das Aufnahmeverfahren weniger invasiv ist, wenn keine zusätzlichen Personen im Raum sind. Die Situation ist nicht davon beeinflusst, dass Gäste vor Ort sind und entsprechende Handlungen (z. B. Getränke anbieten, Plaudern usw.) vollzogen werden. Auch ohne die Anwesenheit ist die Situation davon geprägt, dass die Aufzeichnungen stattfinden (vgl. Labov 1972 und Ehlich 2007 sowie für eine Betrachtung der Prägung der Ersteinrichtungssituation eines Smart Speakers durch die Aufzeichnung zu Forschungszwecken Hector 2022). Doch um dem Anspruch gerecht zu werden, „die Sichtbarkeit der Aufnahmesituation durch ein möglichst wenig invasives Datenerhebungsdesign zumindest zu minimieren“ (Gerwinski/Linz 2018: 112), ist es vorteilhaft, nicht als Forscher*in vor Ort zu sein. Im Rahmen der Datenauswertung kann v. a. durch die kontinuierliche Reflexion über die mögliche Beeinflussung durch die Erhebung und durch Auswertungen – etwa Hector (2022) oder Hector et al. (2022) – sichergestellt werden, dass dies nicht aus dem Blick gerät, sondern vielmehr nutzbar gemacht wird und Einzug in Fragestellungen und Analysen hält.

Um die Erhebung zu realisieren, wurden Smartphones, die als Kameras dienten, sowie Audio-Aufnahmegeräte mit der Post an die zuvor über verschiedene Kanäle (sowohl Kaltakquise über Flyer und Plakate wie auch Erhebung im sozialen Umfeld der Forscher*innen) akquirierten Teilnehmer*innen geschickt. Beilie-

146 Siehe dazu ausführlich Gerwinski/Linz (2018: 107–116).

gend war auch eine ausführliche schriftliche Anleitung zur Installation der Aufnahmegeräte mit einer Abbildung exemplarischer Sitzkonstellationen sowie Formulare für die Einwilligung¹⁴⁷ und für die Aufwandsentschädigung. Darüber hinaus fand in sieben von acht Haushalten vor der Einrichtung (nach Eintreffen des Pakets) ein telefonisches oder videobasiertes Gespräch zwischen mindestens einem Haushaltsmitglied sowie den Erheber*innen statt, insbesondere zur technischen Instruktion im Hinblick auf die Aufnahmetechnik – allerdings nicht während der Inbetriebnahme selbst. Trotz des instruktiven Charakters dieser Bestandteile ist ethnomethodologisch interessant, wie die Teilnehmer*innen die Kameras platzieren und was sie selbst für relevante Ausschnitte halten.

Die Videoaufzeichnungen ermöglichen zusammengefasst einen – auch für multimodale Analysen zugänglichen – Blick darauf, wie das Gerät in Augenschein genommen, bewertet und imaginiert wird und wie diese ersten Eindrücke, Bewertungen und Imaginationen ebenso wie räumliche Arrangements praktisch realisiert werden und zusammen mit der Einbindung in die bereits installierte, ggf. (noch) nicht vollständig eingerichtete, möglicherweise störanfällige mediale Infrastruktur des Haushalts und seiner Mitglieder vollzogen werden. Sie geben aber keinen Einblick in die Nutzung der Geräte, wie sie nach der Ersteinrichtung erfolgt, sich ggf. verstetigt, in soziale Praktiken einpasst und dort weiter verhandelt wird und sich – so jedenfalls kann angenommen werden – auch verändert. Darum wurde der im nächsten Kapitel vorzustellende Audio-Datentyp erhoben.

5.3 Datentyp II: Audioaufzeichnungen der Smart-Speaker-Nutzung

Die Untersuchung der Nutzung von Smart Speakern über einen längeren Zeitraum hinweg in einer privaten Wohnumgebung „steht vor der Herausforderung, dass dazu Sprachdaten erforderlich sind, die nicht nur die Anwendung des Smart Speakers selbst, sondern auch die Kontexte der Anwendung dokumentieren [...]“ (Hector et al. 2022: 1). Zwar zeichnen die Geräte auch selbst in der Cloud die an sie gerichteten Anfragen auf und machen sie über die Smartphone- oder Tablet-Applikation für die Nutzer*innen zugänglich.¹⁴⁸ Wie Habscheid et al. (2021: 44–45)

¹⁴⁷ Die genutzte Einwilligungserklärung wurde im Rahmen des SFB-Teilprojekts, in dem diese Arbeit entstand (siehe Kap. 1), gemeinsam mit der Datenschutzstelle der Universität Siegen sowie dem Leibniz-Institut für Deutsche Sprache (Mannheim) entwickelt. Letzteres war notwendig, weil ein Teil der Daten in das Archiv für Gesprochenes Deutsch (AGD) bzw. in das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) eingeflossen ist (siehe Kap. 5.5).

¹⁴⁸ Für eine genauere Beschreibung siehe Kap. 3.3.

darlegen, ist in den Protokolldaten die Nutzung im Kontext ihrer „Einbettung in soziale Interaktion und deren praktische bzw. diskursive Kontexte dagegen nur im Ausnahmefall und unvollständig erfasst“. Wie die Analyse der Protokolldaten zeigt, sind auch sequenziell mit der Smart Speaker-Anfrage verbundene Äußerungen häufig abgeschnitten. So können Dynamiken der Vorbereitung der Nutzung, der ko-operativ verfertigten Anbahnung einer Stimmeingabe, der Aushandlung über die Nutzungshoheit einerseits und Bewertungshandlungen, interaktive Bearbeitungen von Äußerungen des Smart Speakers und die ‚Verwertung‘ der vom Smart Speaker präsentierten Informationen und die Aushandlung derselben nicht berücksichtigt werden. Es ergäbe sich so ein höchst lückenhafter Blick auf die Verfahren der sprachlichen Aneignung bzw. Domestizierung der Medientechnologie im Haushalt und Fragen nach dem Gesprächsbeteiligtenstatus könnten ebenso wenig profund adressiert werden, weil eine sequenziell-praxeologische Analyse auf Grundlage dieser Daten nicht durchführbar wäre.

Es zeigt sich, dass für die Erhebung gesprächsanalytisch auswertbarer Daten eine andere Lösung gefunden werden musste. Bereits Porcheron et al. (2018) waren auf diese Herausforderung gestoßen und hatten zu diesem Zweck einen sogenannten „Conditional Voice Recorder“ (CVR) entwickelt. Dieser ahmt in seiner Funktionsweise in Teilen einen Smart Speaker nach: Er nimmt die akustischen Signale in der Umgebung, in der er platziert ist, auf und scannt sie auf das Aktivierungswort („Alexa“, „Hey/Okay Google“ bzw. „Hey Siri“). Außerdem speichert das Gerät temporär im Zwischenspeicher drei Minuten¹⁴⁹ Audiomaterial, das gelöscht und überschrieben wird, wenn kein Aktivierungswort fällt, aber gespeichert wird, sofern beim Scan ein Aktivierungswort erkannt wird. Sodann werden weitere drei Minuten aufgezeichnet und gespeichert, sodass sich Audioaufnahmen von ca. sechs Minuten ergeben. Wird ein weiteres Aktivierungswort während der Aufnahmezeit erkannt, verlängert sich die Aufnahmezeit entsprechend um weitere drei Minuten. Dazu ist der CVR aus verschiedenen Komponenten zusammengesetzt: Einem Raspberry Pi, Modell 3, einem damit verbundenen USB-Konferenzmikrofon, einem USB-Stick, auf dem die Aufnahmen gespeichert werden, einer SD-Karte, auf der das Betriebssystem gespeichert ist. Zusätzlich sind LED-Leuchten verbaut, die den Status des CVR anzeigen (alle Leuchten blinken: „Start“; grün leuchtet: „Aufnahme läuft, wird aber nicht gespeichert“; grün und rot leuchten: „Aufnahme läuft und wird gespeichert“; gelb leuchtet: Verarbeitungsfehler). Er sieht damit einem Smart Speaker – schon aufgrund seiner technischen Eigenschaften – nicht nur von seiner Funktionsweise her, sondern auch optisch (vgl. Abb. 16) im Hinblick auf Form und Größe ähnlich, wodurch er sich

¹⁴⁹ Die Länge des gespeicherten Audiomaterials ist variabel einstellbar.



Abb. 16: Conditional Voice Recorder (CRV), bestehend aus Konferenzmikrofon (rechts) sowie Raspberry Pi 3 mit USB-Stick (links); Bild: Sina van Oostrum.

unauffällig in der Wohnumgebung und der Nähe des Smart Speakers platzieren lässt (vgl. Merkle/Hector 2025).

Der von Porcheron et al. (2018) entwickelte und erfolgreich eingesetzte CVR war Grundlage für eine Implementierung und Weiterentwicklung des CVR im Zusammenhang des Forschungsprojekts. Nicht nur musste das Gerät entsprechend auf lokal zur Verfügung stehender Hardware auf Grundlage einer GitHub-Dokumentation¹⁵⁰ nachgebaut werden,¹⁵¹ es ergab sich auch die Notwendigkeit einer Weiterentwicklung der bestehenden Software: Die vorliegende Programmierung – eine auf Python-Basis geschriebene und mit der als open source bereitgestellten Machine-Learning-Plattform TensorFlow¹⁵² realisierte Hot-Word-Erkennung,¹⁵³ bei der eine Bibliothek von Trainingsdaten von SnowBoy zum Einsatz kam (vgl. Porcheron 2019) – war lediglich auf das Erkennen des Aktivierungsworts „Alexa“ ausgerichtet. Da jedoch alle drei der marktüblichen Hersteller berücksichtigt werden sollten, war diese Konfiguration nicht ausreichend. Die Weiterentwicklung wurde in Zusammenarbeit mit einer externen Firma realisiert,¹⁵⁴ die auch eine Dokumentation der vollzogenen Schritte veröffentlichte (siehe Kernel Concepts 2022). Eine umfangreiche Be-

¹⁵⁰ Siehe Porcheron (2019).

¹⁵¹ Für hilfreiche Hinweise danke ich Martin Porcheron von der Swansea University (Computational Foundry), vormalig Mixed Reality Lab, University of Nottingham. Für Hilfe beim Nachbau danke ich dem Fabricator Lab der Universität Siegen, insbesondere Fabian Vitt.

¹⁵² Siehe <https://www.tensorflow.org/>.

¹⁵³ Siehe <https://github.com/Kitt-AI/snowboy>.

¹⁵⁴ Für die gute Zusammenarbeit danke ich der Firma Kernel Concepts sowie insbesondere Simon Budig.

sprechung der CVR-Weiterentwicklung, eine Reflexion der Forschungspraktiken als Datenpraktiken sowie eine Diskussion der Folgen dieser Erhebungstechnologie für die Erhebung und Auswertung auch im Sinne einer sich in die Forschungsergebnisse selbst einschreibenden Forschungspraxis findet sich bei Hector et al. (2022).

Der Einsatz des CVR als Erhebungstechnologie ermöglichte eine Aufzeichnung in zwei Zeiträumen von jeweils drei bis vier Wochen: einer ersten Erhebungsphase unmittelbar nach der Ersteinrichtung (Phase I) sowie einer zweiten Erhebungsphase ca. drei bis vier Monate später (Phase II). So können wir ein umfangreiches Bild über die Smart-Speaker-Nutzung in seiner Einbettung in den Vollzug von sozialen und kommunikativen Praktiken in den Haushalten bekommen. Auch wenn dabei vermieden werden konnte, als Erheber*in selbst anwesend zu sein, und somit die Situation weniger stark durch die Erhebung beeinflusst war, zeigt sich auch in den vom CVR erhobenen Datensätzen, dass die Aufzeichnung nicht unbemerkt blieb (vgl. Hector et al. 2022) und die Verwendung des Smart Speakers und seine Einbindung in die Alltagspraktiken bis zu einem gewissen Grad auch ohne die Anwesenheit der Forscher*innen verändert hat. Die Erhebung von Audio-Daten mit dem CVR war zudem mit weiteren Einschränkungen verbunden.¹⁵⁵ So ist die Zuverlässigkeit der Wake-Word-Erkennung des CVR deutlich niedriger als die der kommerziell vertriebenen Smart Speaker (vgl. Hector et al. 2022), sodass eine *vollständige* Dokumentation der Nutzung der Geräte nicht gewährleistet ist. Dies ist in den Auswertungen zu berücksichtigen und insbesondere bei der Formulierung von Tendenzen zu reflektieren.

Betrachten wir dazu noch einmal die von Gerwinski/Linz (2018: 107) identifizierten methodischen Herausforderungen: Legitimität bzw. „Grad der Arrangiertheit“, Invasivität und Auswertungssensitivität.¹⁵⁶ Der Grad der Arrangiertheit ist bei diesem Datentyp sehr gering: Durch die über mehrere Wochen andauernde Aufzeichnung durch den CVR werden nur selten gesondert Situationen zur Smart Speaker-Nutzung und somit zur CVR-Aufzeichnung geschaffen. Gesonderte „Sprech-anlässe“ im Sinne Schanks (1979: 74) entstehen in den Haushalten nicht oder nur sehr selten (z. B. bei Nachfragen durch die Erheber*innen zum technischen Verlauf der Aufzeichnungen oder bei Fehlfunktionen). Über den CVR-Einsatz gelingt es also, Situationen zu erfassen, in denen – jedenfalls ausweislich der fehlenden Reflexion in den Daten – auch ohne die Beobachtung durch den CVR ein Smart Speaker zum Einsatz käme.

¹⁵⁵ Zu den detaillierten Auswirkungen auf die Erhebung in den einzelnen Haushalten siehe Kap. 5.5.

¹⁵⁶ Siehe Kap. 5.2.

Das Verfahren ist allerdings gerade durch seine lange Dauer durchaus invasiv. Auch wenn Maßnahmen zum Schutz der Daten unternommen wurden – so hat der CVR keine Verbindung zum Internet, speichert die Aufzeichnungen ausschließlich lokal und das gefundene Datenerhebungsverfahren ermöglicht den Nutzer*innen eine Prüfung vor der Datenweitergabe an das Forschungsteam –, bleibt ein materieller Gegenstand in der privaten Wohnumgebung, der mithört. Die Daten werden außerdem, anders als die Aufzeichnungen des Smart Speakers, nicht nur einem anonymen Komplex von Cloud-Diensten zugänglich gemacht, sondern eben auch einem mehr oder weniger bekannten Forschungsteam (siehe dazu Hector 2022; Hector et al. 2022). Teile der zum Datenschutz ergriffenen Maßnahmen, z. B. die erwähnten LED-Leuchten (s. o.), verstärken den Effekt noch, indem sie zusätzlich leuchten, blinken und auf sich aufmerksam machen. Dabei entsteht ein Widerspruch zwischen dem Bestreben, die Aufzeichnung möglichst unauffällig und integriert in den Alltag zu gestalten, zugleich aber im Sinne einer Datenschutzethik auf die Aufzeichnung aufmerksam zu machen.

Vor dem Hintergrund der Debatte um die Natürlichkeit der Daten¹⁵⁷ und der von Gerwinski/Linz (2018: 108) geforderten Sensitivität diesbezüglich bei der Auswertung wurden bei der Erhebung mehrere Maßnahmen ergriffen, um die Beeinflussung der Aufnahmen durch die CVR-Technologie zu erkennen und einen Umgang damit entwickeln zu können. Erstens wurden die Weiterentwicklung und Anwendung der Technologie sowie die damit in Verbindung stehende Forschungspraxis selbst reflektiert (siehe Hector et al. 2022). Dabei wurden nicht nur die interaktiven Reflexionen der Aufgezeichneten, die sich in den Daten finden lassen, besprochen, sondern auch ausführlich das Vorgehen bei der Auswertung thematisiert und dabei die Praktiken herauskristallisiert, die bei der Forschung vollzogen wurden, um einen Umgang mit den Datentypen zu finden (vgl. Hector et al. 2022). Zweitens ist die Beobachtung durch einen zusätzlichen, von den Forscher*innen installierten ‚Mithörer‘ ein interessanter Untersuchungsgegenstand, auch in seiner Ähnlichkeit, aber auch Andersartigkeit zu Smart Speakern. Explizit mit den Charakteristika des CVR in Abgrenzung zum Smart Speaker setzen sich weitere Arbeiten auseinander, die das bereits mehrfach erwähnte Spannungsfeld zwischen Komfort und Beobachtung, in dem sich Smart Speaker bewegen, weiter beleuchten (Merkle/Hector 2025; Hector 2022) und gleichzeitig für die mitlaufende Erhebungstechnologie und den potenziellen Einfluss auf die in der vorliegenden Arbeit verwendeten Daten sensibilisieren. Drittens lassen sich für diese Arbeit sowohl ethnografische Erkenntnisse wie auch Einsichten aus den Interviews nutzbar machen, die mit den gleichen Haushalten geführt worden sind.

¹⁵⁷ Siehe die Diskussion zum Ende von Kap. 4.3.

Die ethnografischen Informationen ergeben sich insbesondere aus den recht umfangreichen Vor- und Nachgesprächen zur Anbahnung und Abwicklung der Erhebung, etwa kurze Erzählungen über die beabsichtigte und tatsächliche Nutzung, auffällige Begebenheiten im Rahmen der Anwendungen, Erkenntnisse zur soziodemografischen Struktur des Haushalts sowie zur Beziehungskonstellation. Ferner ergeben sich diese aus den Personendatenformularen, die für jeden Haushalt mit erhoben wurden. Diese geben nicht nur Aufschluss über die Sprachbiografie, sondern auch über die Erfahrungheit und Technikaffinität und werden in den Analysen herangezogen. Die Interviews werden für diese Arbeit aufgrund des Schwerpunkts und der beschriebenen methodologischen Verortung nicht systematisch ausgewertet.¹⁵⁸ Gleichwohl werden Informationen aus den Interviews auch in den Analysen nicht ausgeblendet. Dies betrifft insbesondere Berichte über die sozialen Konfigurationen sowie die Nutzungsdynamiken innerhalb des Haushalts, Metareflexionen über die Anwendung des Smart Speakers sowie die Problembehandlung bei auftretenden Störungen. Im Sinne eines konversationsanalytischen Vorgehens sollen sie für die Kategorienbildung nicht primär sein, doch sofern Hintergrundinformationen aus den Interviews einbezogen werden, wird dies – wie andere genannte Aspekte – als ethnografische Erweiterung kenntlich gemacht und reflektiert. Dies steht im Einklang mit einem sprachfokussierten, aber nicht kontextblinden Vorgehen, wie es Deppermann (2000) beschreibt und wie es auch in Kap. 4.3 als Arbeitsgrundlage hergeleitet wurde.

5.4 Datenaufbereitung

Die von den Anwender*innen selbst mit bereitgestellten Aufnahmegeräten (Kameras bzw. Smartphones) erhobenen *Videoaufzeichnungen* liegen als MP4-Dateien vor. Sie wurden anonymisiert bzw. in den Transkripten möglichst sinnerhaltend pseudonymisiert. Die Videos wurden zudem für die Abbildung von Videostills in der vorliegenden Arbeit leicht verfremdet, um die Wahrscheinlichkeit einer Wiedererkennbarkeit von Personen so weit wie möglich zu reduzieren. Die Transkription erfolgte gemäß der Transkriptionskonventionen des Gesprächsanalytischen Transkriptionssystems 2 (GAT 2, Selting et al. 2009) sowie – mit leichten Abwandlungen – den darauf aufbauenden multimodalen Erweiterungen nach Mondada (2022).¹⁵⁹

¹⁵⁸ Eine interviewbasierte Studie zum Status der ‚Gesprächsbeteiligung‘ von Smart Speakern aus dem Projektkontext, die sich diesen Fragen nicht linguistisch, sondern mediensoziologisch nähert, findet sich bei Englert/Hoffmann/Waldecker (2022).

¹⁵⁹ Siehe Kap. 5.6 für eine Erläuterung und eine Übersicht über die Transkriptionskonventionen.

Teilweise erfolgten die Transkriptionen im Rahmen einer Kooperation mit dem Forschungs- und Lehr-Korpus Gesprochenes Deutsch (FOLK) am Leibniz-Institut für Deutsche Sprache, wo die Ersteinrichtungen aus vier von acht teilnehmenden Haushalten¹⁶⁰ archiviert werden konnten.¹⁶¹ Diese können nach vorheriger Registrierung über die Datenbank für Gesprochenes Deutsch (DGD) im virtuellen Korpus „Sprachassistenten-Einrichtung“ zur wissenschaftlichen Nutzung abgerufen und einschließlich des Transkripts angesehen werden.¹⁶² In Tab. 1 findet sich außerdem die Sprechereignis-ID für die im FOLK hinterlegten Ersteinrichtungen. Werden in den Analysen Ausschnitte aus diesen Daten besprochen, so wird stets ein Direktlink zum entsprechenden Ausschnitt angegeben.

Die mit dem CVR erhobenen *Audioaufzeichnungen* wurden ebenfalls vollständig inventarisiert. Die Inventarisierung umfasst alle Datenpunkte einschließlich einer rekonstruierten Angabe zu Datum und Uhrzeit der Aufzeichnung, die aufgrund der technischen Beschaffenheiten des CVR eine Herausforderung für die Erhebung war – u. a. aufgrund der fehlenden Internetverbindung war nach einmaliger Unterbrechung der Stromversorgung keine gesicherte Datums- und Uhrzeitangabe mehr möglich (vgl. Hector et al. 2022: 11). Letztlich können diese Angaben nicht als zuverlässig gelten; allerdings sind die Abstände der Aufnahmen untereinander von dieser Problematik nicht betroffen, sodass eine Zuordnung der Aufnahmen zum zeitlichen Verlauf der Nutzung des Geräts zweifelsfrei möglich ist. Durch die Inventarisierung lassen sich – trotz des qualitativ-explorativen Charakters der vorliegenden Arbeit – kleinere, quantitative Auswertungen mit dem Datenkorpus auf Grundlage gebildeter Kollektionen anstellen.¹⁶³ Diese sollen die Analysen im Sinne eines komplementären Vorgehens, wie es etwa Kendrick (2017) vorschlägt, gelegentlich ergänzen.

Die CVR-Aufnahmen liegen als WAV-Dateien vor und wurden ebenfalls nach GAT 2 auf der Stufe eines Basistranskripts (vgl. Selting et al. 2009: 369) transkribiert. Die Daten wurden außerdem vollständig anonymisiert bzw. pseudonymisiert. Die Benennung der Beispiele, die im Analyseteil vorgestellt werden, ergibt sich aus den Haushalten (01 bis 08), dem Situationstyp („EE“ für Ersteinrichtung, „CVR“ für CVR-generierte Aufnahme), einer ggf. nachfolgenden Spezifikation (Phase 1 oder 2 der CVR-Erhebung) sowie einer daran anschließenden Nummer,

¹⁶⁰ Für eine tabellarische Übersicht über alle teilnehmenden Haushalte siehe Kap. 5.5.

¹⁶¹ Für die erfolgreiche Zusammenarbeit möchte ich an dieser Stelle Silke Reineke und Evi Schedl (Leibniz-Institut für Deutsche Sprache, Mannheim) danken.

¹⁶² URL: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayShareObject&shareID=Qhy5TO&objectType=meta_1 (zuletzt geprüft am 03.03.2025).

¹⁶³ Siehe Kap. 5.5 zur Auswahl von drei Fokushaushalten sowie Kap. 6.1 zur Bildung einer Kollektion dyadischer VUI-Dialoge in den Fokushaushalten.

die sich bei den Ersteinrichtungen auf das entsprechende Segment im aufnahme-spezifischen Gesprächsinventar bezieht und bei den CVR-Aufnahmen fortlaufend je Aufnahme vergeben wird. Jedes Beispiel erhält zudem zur einfacheren Bezugnahme und Wiedererkennbarkeit einen Titel, der sich aus einem oder mehreren Schlagwörtern aus dem Verlauf der Aufzeichnung ergibt.

5.5 Untersuchte Haushalte und Auswertung

Insgesamt wurden in acht Haushalten Daten aus den zuvor beschriebenen Situationstypen erhoben. Dabei sind insgesamt sechs Videos von Ersteinrichtungen sowie insgesamt 226 CVR-Aufzeichnungen entstanden. Die Ersteinrichtungen sind im Durchschnitt rund 19 Minuten lang und umfassen eine Gesamtdauer von einer Stunde und rund 53 Minuten. Das CVR-Audiomaterial umfasst insgesamt 30 Stunden und rund 58 Minuten. Eine Übersicht über die erhobenen Daten in allen Haushalten liefert Tab. 1.¹⁶⁴

Wie aus den Zahlen deutlich wird, unterscheidet sich der Umfang des erhobenen Materials teilweise erheblich. Diese Unterschiede entstanden einerseits durch erhebungstechnische Herausforderungen, so wurde z. B. die CVR-Erhebung in der Phase I in Haushalt 2 unterbrochen und die erhobenen Daten vernichtet, weil der im Haushalt lebende Hund den CVR gegen Ende des Aufnahmezeitraums von der Anrichte gestoßen hatte.¹⁶⁵ Außerdem erwies sich die Aufzeichnung mit dem CVR in anderen Haushalten als instabil – so wurden teilweise deutlich weniger Daten erhoben, als erwartbar gewesen wäre (z. B. in der zweiten Phase von Haushalt 3, in dem eine Nachbefragung ergab, dass der Smart Speaker öfter genutzt wurde, als hier dokumentiert ist) –, dies kann nur über Abbrüche des CVR während der Erhebungszeiträume erklärt werden. Darüber hinaus hatten die Haushalte 4 und 5 bereits einen Smart Speaker, sodass die Ersteinrichtung und die daran anschließende „Phase I“ nicht aufgezeichnet werden konnten. Haushalt 6 hatte ebenfalls bereits einen Smart Speaker, es konnte aber die Ersteinrichtung eines Zweitgeräts gefilmt werden.

164 Die darin enthaltenen und alle folgenden personenbezogenen Angaben wurden vollständig pseudonymisiert; dies umfasst neben Vor- und Nachnamen und Ortsnamen auch weitere Äußerungen der Teilnehmenden, die eine Wiedererkennung ermöglichen würden (z. B. spezifische Ausbildungsstätten oder Arbeitsplätze, Tiernamen oder Spitznamen).

165 Der relevante Zeitraum der ersten Wochen nach der Ersteinrichtung war zu diesem Zeitpunkt schon vorüber, sodass von einer Wiederholung der Phase I abgesehen wurde; es lagen aus diesem Haushalt ersatzweise Aufnahmen aus Besuchssituationen vor.

Tab. 1: Übersicht über erhobene Daten in allen Haushalten; blau hervorgehoben: Fokushaushalte.

Pseudonym	Nr.	Dauer Ersteinrichtung (hh:mm:ss)	Anzahl CVR- Aufnahmen Phase I	Dauer CVR- Aufnahmen Phase I (hh:mm:ss)	Anzahl CVR- Aufnahmen Phase II	Dauer CVR- Aufnahmen Phase II (hh:mm:ss)	Gerät	FOLK Sprechereignis- ID
Faßbender	01	00:24:21	11	01:11:31	8	01:11:01	Amazon Echo Dot, 3. Gen.	FOLK_E_ 00465_SE_01
Soellner	02	00:19:04	-	-	3	00:29:51	Amazon Echo Dot, 4. Gen.	FOLK_E_ 00466_SE_01
Würz	03	00:23:07	20	04:20:03	2	00:04:41	Google Home Mini	
Ruhlange	04	-	-	-	21	02:17:13	Amazon Echo Dot, 3. Gen.	
Riker	05	-	-	-	52	08:14:03	Apple HomePod	
Reschke	06	00:03:58	-	-	6	00:36:24	Apple HomePod	
Waldes	07	00:21:27	52	06:44:50	1	00:06:03	Apple HomePod Mini	FOLK_E_ 00467_SE_01
Matthäi	08	00:21:44	9	01:02:46	41	04:40:01	Google Home Nest	FOLK_E_ 00484_SE_01
SUMME	Σ	01:53:39	92	13:19:10	134	17:39:17		

Für die Aufdeckung relevanter Phänomene und die Konzeption der Analyse wurde das gesamte erhobene Material aus allen Haushalten mehrfach gesichtet und ausgewertet. Die Auswahl der Beispiele für die Detailanalysen konzentriert sich davon ausgehend auf drei *Fokushaushalte* Faßbender (1), Waldes (7) und Matthäi (8). Sie sind in Tab. 1 blau hinterlegt. In der Kollektion der Fokushaushalte stehen somit insgesamt 122 CVR-Aufnahmen mit einer Gesamtdauer von 14 Stunden und 56 Minuten zur Verfügung sowie drei Video-Aufnahmen der Inbetriebnahme-Situationen mit einer Gesamtdauer von einer Stunde und rund acht Minuten. Eine kontrollierte Reduktion des gesamten zur Verfügung stehenden Materials war notwendig, um einen gleichmäßigen Analysestandard zu sichern. Die drei Haushalte können im Sinne einer ethnografischen Anreicherung etwas genauer vorgestellt und das Datenmaterial in der Tiefe ausgewertet werden, was für acht Haushalte nicht ohne Qualitätsverluste zu leisten gewesen wäre. Die Erhebung in den drei ausgewählten Haushalten verlief ‚idealtypisch‘, d. h., alle drei Datentypen konnten erhoben werden und Besonderheiten wie eine fehlende Erhebungsphase, die Einrichtung eines Zweitgeräts oder das gänzliche Fehlen der Ersteinrichtung treten in den Haushalten 1, 3, 7 und 8 nicht auf. Für die Beschränkung auf die Haushalte 1, 7 und 8 sprach zudem, dass die Ersteinrichtung bei diesen Haushalten nicht allein vollzogen wird, wie es in Haushalt 3 der Fall ist, sowie darüber hinaus die angestrebte Varianz der Hersteller: Die drei Haushalte decken Geräte von allen drei Herstellern (Amazon, Google und Apple) ab und wurden auch deswegen ausgewählt. Zudem war in diesen drei Haushalten ein kontinuierlicher Feldzugang gesichert, was z. B. für Rückfragen zu Einverständniserklärungen und situativen Rahmungen sehr hilfreich war. In anderen Haushalten war dies nach Ende der Aufzeichnungsphasen teilweise nicht mehr der Fall.

Die Konzentration auf drei Haushalte ermöglichte auch, die qualitativen Analysen um Aussagen über die Häufigkeit bestimmter Phänomene zu ergänzen. Es ist wichtig zu betonen, dass dabei gleichwohl keine formale Kodierung des Datenmaterials im Sinne eines systematischen und fortlaufend weiterzuentwickelnden Kodierschemas erfolgt ist. Darauf ist nicht aus grundsätzlicher Ablehnung verzichtet worden – Stivers (2015) argumentiert, dass es Verfahren des Kodierens gibt, die durchaus auch die für konversationsanalytische Untersuchungen notwendige Kontextsensibilität berücksichtigen können; auch Meiler/Siefkes (2023: 320) oder Pitsch (2023) schlagen Mixed-Methods-Ansätze in der empirischen Linguistik vor. Allerdings lag der Gewinn von Erkenntnissen, die sich auf der Basis von Kodierung gewinnen lassen könnten, nicht im Zentrum der Erkenntnisinteressen dieser Arbeit. Dagegen sprach erstens der qualitativ-explorative und insofern entdeckende Charakter der Studie, der sprachliche Praktiken im sozialen Kontext verstehen will und eine daran anknüpfende Kodierung nicht erfordert und auch nur unter großen Einschränkungen überhaupt ermöglicht. Zweitens

konnte bei der Untersuchung von VUIs in privaten Wohnumgebungen nur auf einen dünnen Forschungsstand zurückgegriffen werden, sodass es angesichts der Neuheit der beobachteten Praktiken erforderlich erscheint, diese zunächst über Sequenzanalysen zu erschließen und somit detailliert qualitativ-explorativ zu erfassen und beschreibbar zu machen. Eine gleichzeitige Anwendung beider Verfahren hätte insofern auch im Rahmen der vorliegenden Arbeit keiner von beiden Vorgehensweisen vollständig gerecht werden können. Weil Häufigkeitsverteilungen gleichwohl von Interesse auch für Anschlussstudien sein können, wurde hierfür ein Verfahren gewählt, bei dem einzelne Phänomene im Hinblick auf ihr Auftreten quantifiziert wurden; damit in Verbindung stehende Aussagen müssen jedoch stets als streng limitiert auf die vorliegende Kollektion gesehen werden. Sie beziehen sich zudem auf eine Kollektion dyadischer VUI-Dialoge, auf deren Zusammensetzung noch näher einzugehen ist. Quantitative Aussagen zur Häufigkeit und Verteilung bestimmter Phänomene im Korpus gelten also als erste Hinweise (nicht hingegen als hinreichende Bedingung) auf die Etablierung sprachlicher Praktiken.

Die drei ausgewählten Haushalte sind studentische Wohngemeinschaften in Nordrhein-Westfalen, deren Bewohner*innen zwischen 20 und 31 Jahre alt sind – eine genaue Beschreibung dieser drei Haushalte folgt im weiteren Verlauf des Kapitels. Die soziodemografisch ähnlichen Eigenschaften ergaben sich primär aus dem Vorgehen bei der Akquise im studentischen Milieu der Universität heraus, allerdings sind Daten aus Haushalten mit anderen Altersausprägungen im übrigen Korpus vorhanden: So sind die Haushalte 2 und 4 jeweils ein zusammenlebendes Paar (je ein Mann und eine Frau im Alter zwischen 20 und 30 Jahren), Haushalt 3 und 6 jeweils ein verheiratetes Ehepaar (in Haushalt 3 ein Mann und eine Frau zwischen 60 und 70 und in Haushalt 6 ein Mann und eine Frau zwischen 25 und 35 Jahren) und in Haushalt 5 ein allein wohnender Mann zwischen 30 und 40 Jahren mit regelmäßigem Besuch von seinem Lebensgefährten, ebenfalls zwischen 30 und 40 Jahren.¹⁶⁶ Die Daten aus den Haushalten 2 bis 5 werden teilweise vorgestellt, um die Analyse um bestimmte Aspekte zu bereichern. Die Haushalte werden jedoch nicht detailliert eingeführt und in deutlich geringerem Umfang berücksichtigt. Die Analysen konzentrieren sich also bei der Auswahl der vorgestellten Auszüge auf Daten aus diesen drei Haushalten, werden aber teilweise um einzelne Aufnahmen aus den anderen Haushalten ergänzt. Dies ist insbesondere dann der Fall, wenn auftretende Phänomene sich an Ausschnitten aus anderen Haushalten besonders gut illustrieren lassen.

¹⁶⁶ Zum Schutz der personenbezogenen Daten werden keine präziseren Angaben über das Alter der Studienteilnehmer*innen gemacht.

In den Fokushaushalten hat die Erhebung zwischen Dezember 2020 und Juli 2022 stattgefunden, wie Tab. 2 genauer aufschlüsselt:

Tab. 2: Übersicht Fokushaushalte und Erhebungszeiträume.

Pseudonym	Nr.	Gerät	Ersteinrichtung	CVR Phase I	CVR Phase II
Faßbender	01	Amazon Echo, 3. Gen.	04.12.2020	04.12.2020 – 20.02.2021 ¹⁶⁷	28.06.2021 – 07.09.2021
Waldes	07	Apple HomePod	22.10.2021	22.10.2021 – 19.11.2021	17.06.2022 – 08.07.2022
Matthäi	08	Google Nest	19.12.2021	19.12.2021 – 18.01.2022	05.06.2022 – 03.07.2022

Die Haushalte sollen nachfolgend etwas detaillierter vorgestellt werden; zudem werden die Personenkonstellation und der Ablauf der Ersteinrichtung des Smart Speakers erläutert, da auf diese Daten im Analyseteil wiederholt referiert wird.

5.5.1 Haushalt 1: Amazon: Echo Dot mit „Alexa“ (Faßbender)

In Haushalt 1 leben Lukas Faßbender (25 Jahre, LF) und Alex Kripp (27 Jahre, AK) in einer Wohngemeinschaft in Münster. Beide studieren derzeit einen Master-Studiengang und sind im studentischen Umfeld aktiv: Sie sind z. B. eingebunden in die akademische Selbstverwaltung der dortigen Universität und haben üblicherweise regelmäßig befreundete Studierende zu Gast. Sie sind deutsche Muttersprachler, beide mit einer Sprachbiografie im westlichen Nordrhein-Westfalen. Die Ersteinrichtung vollziehen Lukas und Alex im Dezember 2020 im Wohnzimmer ihrer Wohnung. Zur Zeit der Einrichtung bestimmen die Covid-19-Pandemie und die damit verbundenen Kontaktbeschränkungen das soziale Leben: Wie auch in den anderen Haushalten konnte die Ersteinrichtung nicht vor Ort begleitet werden. Die Kameraausschnitte sind entsprechend – unter Anleitung des Erhebers über ein Video-Telefonat vorab – selbst gewählt. Dabei antizipieren die Bewohner die voraussichtliche Position des Smart Speakers sowie ihre eigene Körperpositionierung in Beziehung zu der des Geräts. Der Smart Speaker wird vorläufig neben dem Fern-

¹⁶⁷ Aus technischen Gründen kam es dabei zu einer Unterbrechung des Aufnahmezeitraums von ca. einer Woche.

seher auf einer TV-Bank positioniert. Er steht damit im gemeinsam genutzten Wohnzimmer, das auch von seiner Einrichtung her auf die Mediennutzung ausgerichtet ist: Ein großes Sofa steht dem Fernseher frontal gegenüber, die TV-Bank unter dem Fernseher dient der Ablage von Peripheriegeräten für den TV sowie zur Strukturierung von notwendigen Kabelanschlüssen. Während der Aufnahmesituation platzieren sich Lukas und Alex jeweils seitlich auf den weiterhin zentral stehenden Fernseher und den darunter platzierten Smart Speaker, die nun beide im Zentrum der Raumarchitektur verortet sind (siehe Abb. 17 und 18).

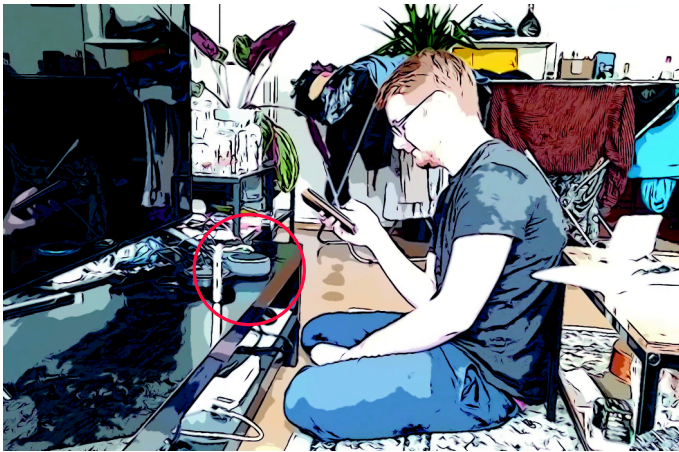


Abb. 17: Lukas (hier im Bild) und Alex (nicht im Bild) richten den Smart Speaker ein (Perspektive 1).



Abb. 18: Alex (links) und Lukas (rechts) richten den Smart Speaker ein (Perspektive 2).

Alex sitzt während der Ersteinrichtung auf dem Sofa und ist insofern etwas erhöht und distanziert, Lukas hingegen sitzt auf dem Fußboden vor der TV-Bank und dem Smart Speaker; er ist der primäre Nutzer und vollzieht die notwendigen

Schritte zur Ersteinrichtung des Geräts. Diese räumliche Konfiguration (siehe Abb. 19) spiegelt sich mehrfach in den interaktionalen Abläufen und unterschiedlichen Graden der Dialogbeteiligung bzw. des Involvements.¹⁶⁸

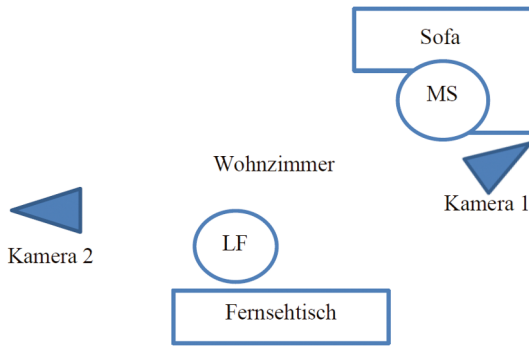


Abb. 19: Positionierung von Teilnehmenden und Kameras im Raum während der Ersteinrichtung in Haushalt 1, Darstellung von oben; Bild: Leibniz-Institut für Deutsche Sprache Mannheim (mit frdl. Genehmigung).

Lukas und Alex geben an, dass sie im Zeitraum der ersten Erhebungsphase aufgrund der Maßnahmen zur Eindämmung der Pandemie außergewöhnlich wenig Besuch bekommen haben und sich teilweise und insbesondere über die Zeit zwischen Mitte Dezember 2020 und Mitte Januar 2021 nur wenig in der Wohnung aufgehalten haben, weil sie ihre Familien an ihren Heimatorten besuchten. Dies zeigt sich auch in den Daten, die der CVR aufgezeichnet hat: In den Aufnahmen aus der ersten Aufnahmephase wird der Smart Speaker vornehmlich für das Abspielen von Musik verwendet. Die genauen Kontexte bleiben dabei, trotz der Aufzeichnung von insgesamt sechs die Stimmeingaben umgebenden Minuten, im Detail überwiegend unklar. Teilweise scheint die Musik sportliche Übungen zu begleiten oder in Haushaltstätigkeiten der Wohngemeinschaft eingebunden zu sein. In der zweiten Phase hat sich der Standort des Smart Speakers verändert. Dieser ist nun nicht länger im Wohnzimmer, sondern in der Küche platziert. Diese Veränderung ergab sich, weil die beiden häufigsten Anwendungsfälle – Musikhören und die Timer-Funktion – parallel zu Aktivitäten in der Küche (Kochen, Reinigungsarbeiten nach der Zubereitung von Speisen und nach dem Essen) stattfinden. Dies kann auf den Aufnahmen nachvollzogen werden – der Großteil der Aufzeichnungen aus der zweiten Erhebungsphase entsteht bei Smart Speaker-

¹⁶⁸ Siehe Kap. 6.2.

Nutzung parallel zu Küchenaktivitäten, häufig der gemeinschaftlichen Zubereitung von Speisen.

5.5.2 Haushalt 7: Apple: HomePod mit „Siri“ (Waldes)

In Gelsenkirchen leben Konrad und Till Waldes (KW und TW, beide 21 Jahre alt), die in einer Wohngemeinschaft leben. Sie sind Zwillingenbrüder und es ist ihre erste eigene Wohnung. Der Umgang zwischen den beiden ist sehr vertraut. Sie studieren im Bachelor und haben regelmäßig Besuch, was jedoch im Rahmen der Aufnahmen nicht zu hören ist. Sie sind deutsche Muttersprachler mit einer Sprachbiografie im Rheinland bzw. im Bergischen Land. Sie nehmen die Ersteinrichtung im Oktober 2021 in ihrem gemeinsam genutzten Wohnzimmer vor. Die Covid-19-Pandemie hat zwar immer noch Auswirkungen auf das soziale Leben – so geben Konrad und Till, ähnlich wie Lukas und Alex aus Haushalt 1, an, dass sie unter anderen Umständen häufiger Besuch gehabt hätten. Allerdings lassen die Umstände zu, dass die Erhebungstechnik (d. h. v. a. Conditional Voice Recorder und Kamera) persönlich durch den Erheber in der Wohnung der Teilnehmer vorbeigebracht werden können. Um dennoch eine vergleichbare Situation bei der Ersteinrichtung herzustellen, verlässt der Erheber die Wohnung anschließend und lässt Konrad und Till auch die Erhebungstechnik nach kurzer mündlicher Erläuterung selbstständig aufbauen. Erst mehrere Stunden nach Fertigstellung der Inbetriebnahme holt er diese wieder ab.



Abb. 20: Konrad (links) und Till (rechts) richten den Smart Speaker ein (Perspektive 1).



Abb. 21: Konrad (rechts) und Till (links) richten den Smart Speaker ein (Perspektive 2).

Zunächst sitzen Till und Konrad gemeinsam vor dem Smart Speaker auf dem Boden und entpacken ihn dort (siehe Abb. 20). Sie wechseln allerdings die Position, als das Gerät an den Strom angeschlossen werden muss, und begeben sich näher an ein Regal, in das dieses gestellt wird (siehe Abb. 21 und Abb. 22). Im Zuge dieses Wechsels nehmen sie auch neue Positionen für die weitere Einrichtung ein: Till bleibt auf dem Sofa sitzen, während Konrad vor dem Regal auf dem Boden hockt und erst wieder aufsteht, als der Vorgang der Inbetriebnahme für vollständig abgeschlossen erklärt wird. Diese Positionierungen spiegeln den Grad der Involviertheit bei der Ersteinrichtung und der späteren Anwendung: Während Konrad dicht beim Gerät bleibt und dies sich in körperlicher Reichweite befindet, nimmt Till auf dem Sofa Platz, von wo aus er den Smart Speaker seitlich zu sich hat, ihn aber nicht mehr ohne Umstände erreichen könnte. Die Einrichtung erfolgt über das Smartphone von Konrad, der klar der primäre Anwender in dieser Situation ist und auch anschließend in den CVR-Aufnahmen der nahezu ausschließliche Nutzer des Smart Speakers ist. Dazu geben die Bewohner an, dass Till noch zusätzlich einen Smart Speaker von Amazon in seinem eigenen Zimmer stehen hat und ferner, dass Konrad deutlich häufiger außer Haus ist.

Die in den Daten der CVR-Erhebung am häufigsten dokumentierte Nutzung ist auch hier die Steuerung von Musik. Allerdings zeigen sich auch einige andere Anwendungsfälle (z. B. Nutzung einer über das VUI generierten Einkaufsliste, Wissensabfragen – teilweise verbunden mit dem Musikhören – und Unterhaltungsanwendungen). Die Äußerungen des VUI sind dabei teilweise in Interaktionen zwischen Konrad und Till eingebunden, teilweise dokumentieren die Äußerungen ausschließlich den Austausch zwischen dem VUI und Konrad allein und

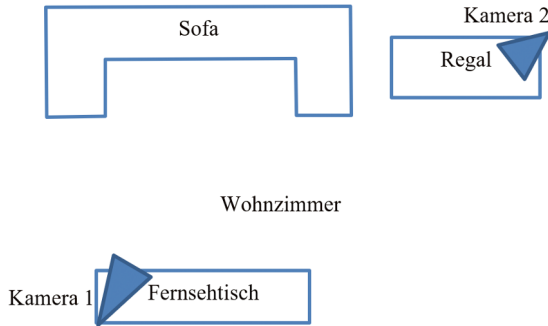


Abb. 22: Positionierung der Kameras im Raum während der Ersteinrichtung in Haushalt 7, Darstellung von oben; Bild: Leibniz-Institut für Deutsche Sprache, Mannheim (mit frdl. Genehmigung).

in einem Fall den Austausch zwischen dem VUI und Till allein. Gäste sind auf einer Aufzeichnung zu hören.

5.5.3 Haushalt 8: Google Nest mit „Google Home“ (Matthäi)

Die Wohngemeinschaft von Samuel Matthäi (SM), Lara Schiffer (LS) und Robin Lorentz (RL) aus Dortmund nutzt den Smart Speaker Google Nest. Samuel ist 30 Jahre alt und studiert mit dem Abschlussziel Lehramt; er arbeitet seit mehreren Jahren als Vertretungslehrkraft im Inklusionsbereich. Robin ist 23 Jahre alt, er studiert im Master ebenfalls mit dem Abschlussziel Lehramt. Lara ist 24 Jahre alt und studiert im Master. Sie beschreiben das Verhältnis zueinander als vertrauensvoll und freundschaftlich. Samuel und Robin haben regelmäßig ihre Lebenspartnerinnen zu Besuch, die teilweise auf den ausgewerteten Aufnahmen zu hören sind; die Lebenspartnerin von Robin, Alexandra Wormsberge (AW), ist auch während der Ersteinrichtung anwesend, aber kaum von der Aufnahme erfasst; sie beteiligt sich auch nicht an der Ersteinrichtung des Google Nest, die am gemeinsamen Küchentisch sitzend durchgeführt wird (siehe Abb. 25).

Die Einrichtung findet vollständig in der abgebildeten Konstellation statt (siehe Abb. 23–24), es wird währenddessen Bier getrunken und die Stimmung ist ausgelassen. Wie auch auf den Aufnahmen zu hören ist, die während der Ersteinrichtung gemacht werden, ist auch hier das Sozialleben noch von den pandemiebedingten Kontaktbeschränkungen im Winter 2021/22 geprägt, sodass – so geben die Haushaltsmitglieder an – selten andere Besucher*innen empfangen werden. Der Smart Speaker wird in der Küche platziert und steht nach Angaben von Samuel lange Zeit auf dem Küchentisch, wo er auch eingerichtet wurde, obwohl diese Positionierung

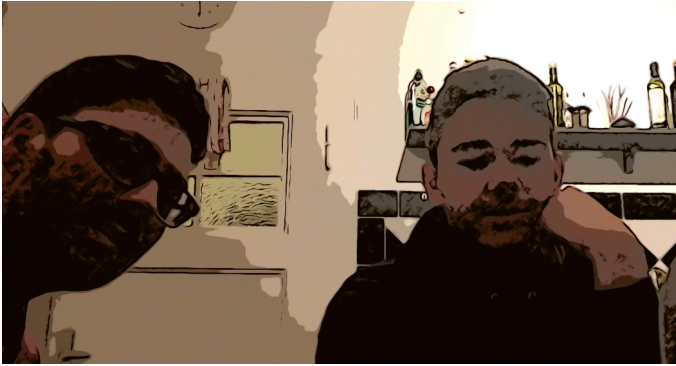


Abb. 23: Samuel (links), Robin (rechts) und Lara (nicht im Bild) richten den Smart Speaker (ganz rechts) ein (Perspektive 1).



Abb. 24: Lara (links), Samuel (Mitte) und Robin (rechts, halb im Bild) richten den Smart Speaker ein (Perspektive 2).

nicht optimal sei, weil er beim Essen oder anderen am Küchentisch zu verrichtenden Aktivitäten gelegentlich störe. Es sei aber auch kein besserer Ort gefunden worden.

Die Beteiligung der drei Bewohner*innen an der Ersteinrichtung ist ausgeglichen, wobei Robin sich als primärer Anwender positioniert, indem er sein Smartphone als erstes mit dem Smart Speaker verknüpft. Samuels Smartphone kann ebenfalls kurz darauf verknüpft werden. Lara ist hingegen bei den im weiteren Verlauf der Aufnahme vollzogenen Tests des Smart Speakers noch abgelenkt, weil es sich als Herausforderung erweist, ihr Smartphone mit dem Smart Speaker zu verbinden (die Gründe dafür bleiben unklar).

Auch die Mitglieder dieser Wohngemeinschaft sind zum Zeitpunkt der Ersteinrichtung im Dezember 2021 aufgrund der bevorstehenden Weihnachtsfeier-

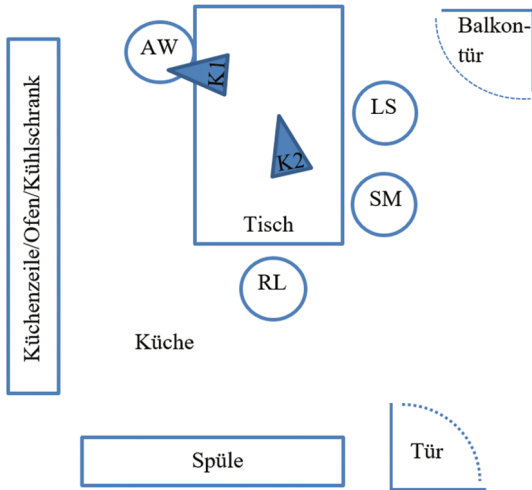


Abb. 25: Positionierung von Teilnehmenden und Kameras im Raum während der Ersteinrichtung in Haushalt 8, Darstellung von oben; Bild: Leibniz-Institut für Deutsche Sprache Mannheim (mit frdl. Genehmigung).

tage pandemiebedingt sehr zurückhaltend im Hinblick auf persönliche Kontakte und haben während der ersten Aufzeichnungsphase keinen dokumentierten Besuch, phasenweise sind sie selbst verreist. Im Verlauf der zweiten Aufzeichnungsphase im Sommer 2022 sind sie hingegen alle drei sehr regelmäßig zu Hause und empfangen alle zu unterschiedlichen Situationen Besuch, der auch auf den CVR-Aufnahmen zu hören ist; entsprechende Genehmigungen zur Verwendung der Daten liegen vor.

5.6 Transkriptionskonventionen

Wie beschrieben wurden die Audioaufnahmen nach dem Transkriptionsstandard GAT 2 (vgl. Selting et al. 2009) bzw. im Falle der Videoaufzeichnungen von den Ersteinrichtungssituationen im multimodalen Standard nach Mondada (2022) transkribiert.¹⁶⁹ Nachfolgend findet sich eine Übersicht über die dabei verwendete

¹⁶⁹ Die multimodalen Transkriptionskonventionen finden sich in der Übersicht bei Mondada (2022). Eine ausführliche Herleitung entlang der Charakteristika multimodaler Interaktionsanalyse mit entsprechenden Beispielen findet sich bei Mondada (2018). Mondada kombiniert die von ihr entwickelten Konventionen für multimodale Transkription mit den konversationsanalytischen Transkriptionssymbolen nach Jefferson (2004). Allerdings lassen sich auch der für

ten Transkriptionskonventionen, die Selting al. (2009) entnommen und für die multimodalen Transkriptionskonventionen eng an Mondada (2022) orientiert ist. Generell gilt für die Transkription, dass die Segmentierung der Transkription des Gesprochenen entlang der von den Transkribent*innen erkannten Intonationsphrasen erfolgt, die sich durch ein Bündel von Merkmalen (u. a. Tonhöhen-sprünge, Akzentuierung, Rhythmus, Dehnung und Pausen) identifizieren lassen (vgl. Selting et al. 2009: 370). Die einzelnen Segmente werden nummeriert und durch Sprecher*innen-Siglen in Großbuchstaben den Sprecher*innen zugeordnet; dabei werden Pausen nicht zugeordnet. Kommentarsegmente werden mit einem kleinen ‚k‘ gekennzeichnet.

Sequenzielle Struktur/Verlaufsstruktur

[]	Überlappungen und Simultansprechen
[]	

Ein- und Ausatmen

°h / h°	Ein- bzw. Ausatmen von ca. 0.2–0.5 Sek. Dauer

Pausen

(.)	Mikropause, geschätzt, bis ca. 0.2 Sek. Dauer
(-)	kurze geschätzte Pause, ca. 0.2–0.5 Sek. Dauer
(--)	mittlere geschätzte Pause, ca. 0.5–0.8 Sek. Dauer
(0.5)	gemessene Pausen, ca. 0.5 bzw. 2.0 Sek. Dauer

Sonstige segmentale Konventionen

und_äh	Verschleifungen innerhalb von Einheiten
äh öhäm	Verzögerungssignale, sog. „gefüllte Pausen“
:	Dehnung, Längung, ca. 0.2–0.5 Sek.
::	Dehnung, Längung, ca. 0.5–0.8 Sek.
:::	Dehnung, Längung, ca. 0.8–1.0 Sek.

Lachen und Weinen

haha hihi	silbisches Lachen
((lacht))	Beschreibung des Lachens
<<lachend>>	Lachpartikeln in der Rede, mit Reichweite
<<:-)> soo >	„smile voice“

gesprächslinguistische Arbeiten etablierte GAT2-Standard mit den multimodalen Konventionen im Standard nach Mondada kombinieren, wie Rothe (2022: 148–149) erläutert.

Rezeptionssignale

hm ja nein	einsilbige Signale
hm_hm ja_a	zweisilbige Signale
?hm?hm	mit Glottalverschlüssen, meistens verneinend

Sonstige Konventionen

((hustet))	para- und außersprachliche Handlungen und Ereignisse
<<hustend>>>	sprachbegleitende para- und außersprachliche Handlungen und Ereignisse mit Reichweite
()	unverständliche Passage ohne weitere Angaben
(solche)	vermuteter Wortlaut
(also/alo)	mögliche Alternativen
((unverständlich, ca. 3 Sek.))	unverständliche Passage mit Angabe der Dauer
((...))	Auslassung im Transkript

Akzentuierung

akZENT	Fokusakzent
ak!ZENT!	extra starker Akzent

Tonhöhenbewegung am Ende von Intonationsphrasen

?	hoch steigend
,	mittel steigend
—	gleichbleibend
;	mittel fallend
.	tief fallend

Multimodale Transkriptionen von Gestik und Blick

Gesten und Beschreibungen von verkörperlichten Handlungen werden zwischen zwei identischen Symbolen notiert (vgl. Mondada 2022).¹⁷⁰ Sie werden synchron zu dazugehörigen Gesprächspassagen platziert und mit Sprecher*innen-Siglen in

170 Abweichend vom Transkriptionsstandard nach Mondada (2022) wird dabei aus Gründen der Vereinfachung davon abgesehen, für jede*n Teilnehmer*in und jeden Aktionstyp jeweils ein eigenes Symbol zu verwenden. Die Zuordnung ist bei Transkripten der hier vorliegenden Komplexität über die klein gedruckten Sprecher*innen-Siglen leicht möglich. Stattdessen werden den Aktionstypen gleichbleibende Zeichen zugeordnet, die nachfolgend aufgeschlüsselt werden.

Kleinbuchstaben versehen. Dabei werden im Einzelnen folgende Symbole verwendet:

*	*	Gestik
+	+	Blick
\$	\$	Mimik
§	§	Körperbewegung
Δ	Δ	Objektmanipulation
*--->		Die beschriebene Handlung setzt sich in Folgezeilen
---->*		fort, bis dasselbe Symbol erneut auftritt.
>>		Handlung beginnt vor dem Ausschnitt
--->>		Handlung endet nach dem Ausschnitt
RH		Rechte Hand
LH		Linke Hand

Bei der Anwendung der Transkriptionsstandards wurden die verbalen Äußerungen und non-verbale Laute (Abfolgen von Tönen sowie Musik oder andere Medieninhalte) der Smart Speaker ebenfalls transkribiert. Bei der Wiedergabe längerer verbaler Medieninhalte (z. B. radioähnliche Nachrichten oder Podcasts) im Hintergrund wurde im Sinne der Übersichtlichkeit teilweise eine detaillierte Transkription getilgt und stattdessen im Kommentar Inhalt und Charakter der Wiedergabe zusammengefasst. Die Vergabe der Siglen für die Smart Speaker (AL für Alexa, GA für Google Assistant und SI für Siri) folgt den zu unterschiedlichen Graden personalisierten „Stimm“-Personae der Smart Speaker.¹⁷¹ Durch die Transkription der nicht-menschlichen Teilnehmenden kommt es zu kleineren Auffälligkeiten im Bereich der Akzentdarstellung: Durch die insgesamt akzentuierte Sprechweise der VUIs aller drei untersuchten Hersteller sind in den Äußerungen häufiger Nebenantizente vermerkt, als sie üblicherweise in gesprochensprachlichen Äußerungen menschlicher Teilnehmer*innen auftreten; es werden dabei teilweise auch mehrere Fokusakzente notiert (vgl. Selting et al. 2009: 373). Stellenweise sind die Grenzen der Intonationsphrasen aufgrund der Prosodie kaum bestimmbar; hier werden syntaktische und pragmatische Geschlossenheit mit berücksichtigt (vgl. Selting et al. 2009: 370). Ferner ist im Duden (2021) die Schreibsilbentrennung für Alexa mit Ale|xa angegeben, weil einzelne Vokalbuchstaben nicht getrennt werden. Die Prosodie der Sprechsilbe stellt sich jedoch klar mit drei akzentuierbaren Vokalen dar, die – folgt man dem Konstituentenmodell nach Pike/Pike (1947) – als a|lex|a dargestellt werden können, wobei „e“ den Silbennukleus bildet (siehe auch Uhmman 1991; Ramers 2015: 113–114).

171 Siehe dazu Kap. 3.3.

6 Analyse sprachlicher Praktiken

Die Analyse ist zweiteilig aufgebaut: Zunächst wird die Dialogorganisation zwischen genau einem Mensch und einem ‚beteiligten‘ Smart Speaker mit VUI, d. h. in dyadischen Konstellationen, unter verschiedenen Aspekten der Gesprächsorganisation betrachtet (Kap. 6.1). Einleitend zu dem Kapitel werden auch die Entwicklung und Auswahl der Kategorien begründet. Aufbauend auf diesen Befunden soll in Kap. 6.2 untersucht werden, wie der Smart Speaker in Mehrparteien-Settings, d. h. in laufende Mensch-Mensch-Interaktionen, eingebunden und wie er am Vollzug der Praxis ‚beteiligt‘ wird. Dabei wird erstens das bereits diskutierte Modell von Hirschauer (2004; 2016) zur praktischen Partizipation am verteilten Handeln genutzt. Gespräche verstehe ich dabei als soziale Praktik, die (sehr variabel gestaltbares) *Sprechen* voraussetzt, während Sprechen nicht notwendigerweise ein Gespräch konstituiert (vgl. Goffman 1979: 6–7). Dabei wird auch gezeigt, wie der Dialog mit Smart Speakern in soziale Praktiken eingeht, und genauer untersucht, wie sich dies an der sprachlichen Oberfläche zeigen kann. Den Ausgangspunkt bildet der bereits diskutierte Aufsatz von Porcheron et al. (2018), dessen Ergebnisse sich zwar auf ein sehr kleines Korpus stützen und vorläufig sind, die aber zugleich die Analyse der vorliegenden Arbeit sowohl methodologisch wie auch im Hinblick auf die Kategorienbildung informiert hat. In der Analyse der Mehrparteienkonstellationen soll anhand von Video- und Audioaufnahmen, in denen mehrere menschliche Teilnehmer*innen beteiligt sind, sequenzanalytisch nachvollzogen werden, welche Verfahren zur Einbettung und zum Übergang genutzt werden und ob bzw. wie Smart Speaker dabei zu ‚Beteiligten‘ an der sozialen Praxis und an Gesprächen gemacht werden – und umgekehrt, wie Menschen zu Beteiligten an Dialogen zwischen Mensch und Smart Speaker gemacht werden. Dieser Teil greift entsprechend sowohl auf Erstinstallations- wie auch auf CVR-Daten aus dem Korpus zu, beschränkt sich aber auf Aufzeichnungen, in denen mehr als ein*e menschliche*r Teilnehmer*in zu sehen bzw. zu hören ist.

6.1 Organisation dyadischer VUI-Dialoge

Bei der iterativen Sichtung und Inventarisierung des Materials, der Bildung von Kollektionen sowie der Diskussion innerhalb des Projektteams, aber auch bei anderen Datensitzungen¹⁷² – wichtigen Schritten der Gesprächsanalyse, die nicht

¹⁷² Dazu zählen u. a. Workshops innerhalb des SFB 1187 sowie Datensitzungen im Rahmen des Doktorand*innen-Netzwerks Gesprächsforschung (vgl. Rogowska/Schneider 2022: 256–257).

nur vorbereitende Arbeit leisten, sondern Bestandteil der Analyse sind (vgl. Depermann 2008: 35–36) – konnten früh zwei Erkenntnisse gewonnen werden. Erstens unterscheiden sich dyadische Konstellationen wesentlich von Mehrparteienkonstellationen, v. a. zeigen sich bestimmte Phänomene der Beteiligung und Einbettung in die soziale Praxis ausschließlich in Aufzeichnungen mit mehreren beteiligten Sprecher*innen, während demgegenüber Charakteristika von VUI-Dialogen in Aufzeichnungen, auf denen nur ein*e einziger*r Sprecher*in und das VUI zu hören sind, deutlicher hervortreten. Daraus ergab sich die bereits eingeführte Zweiteilung der Analyse, die zunächst Charakteristika von VUI-Dialogen herausarbeiten sollte, die dann in Mehrparteienkonstellationen als Befunde vorausgesetzt werden können. Zweitens zeigt sich mit großer Deutlichkeit, dass VUIs und Nutzer*innen den VUI-Dialog entlang gesprächsorganisatorischer Probleme organisieren. Die Nutzer*innen passen dabei bekannte sprachliche Strategien zur Lösung dieser Probleme auf den Kontext eines VUI-Dialogs an, die VUIs zeigen ebenfalls eine Orientierung an solchen Kategorien. Die Analyse im ersten Analyseteil orientiert sich entsprechend an klassischen Kategorien der Konversationsanalyse und konkret den von Schegloff (2006) aufgeführten grundlegenden Problemen der Gesprächsorganisation – Schegloff (2006: 73) spricht von „generic organizations of practice“ –, die auch in der konversationsanalytischen Forschung viel Beachtung fanden und jüngere Teildisziplinen wie die Interaktionale Linguistik informiert und zu weiterer Forschung angeregt haben. Der universale Anspruch der Kategorien, den Schegloff (2006: 70–71) postuliert, lädt außerdem geradezu zu einer Überprüfung anhand einer anderen Form der Dialogizität ein – ein solches Vorgehen wendet auch Thar (2015: 214) in ihrer Untersuchung zu telefonbasierten Sprachdialogsystemen in Call-Centern produktiv an. VUI-Dialoge brechen gerade dadurch gewohnte Prinzipien auf, dass eine*r der an der Verfertigung des Dialogs ‚Beteiligten‘ ein VUI ist und insofern grundlegend anders mit der Situationalität und prinzipiellen Offenheit der Praxis umgeht – nämlich planbasiert und algorithmisch. Dieses Aufbrechen der dialogischen Bedingungen lässt mit Garfinkel (1967: 35) krisenhafte Situationen entstehen, deren genaue Analyse den Blick auf interaktionale Dynamiken und deren Bedingungen freilegen kann (vgl. Pitsch 2015: 229–230; 2023: 148).

Dieses Kapitel wird sich insofern auf mit dem CVR erhobene Daten mit genau einem*einer menschlichen Sprecher*in aus der ersten und zweiten Analysephase nach der Ersteinrichtung stützen.¹⁷³ Dazu wurde eine weitere Kollektion gebildet, die sich aus den Daten in den Fokushaushalten speiste. In diese Kollektion wurden nur Aufnahmen aufgenommen, auf denen weitere Sprecher*innen neben

173 Zum Datentyp der CVR-Daten siehe Kap. 5.3.

den Anwender*innen des VUI entweder nicht zu hören sind oder nicht (auch nicht mittelbar über die Beteiligung an der vor oder nach dem VUI-Dialog aufgenommenen Praxis) in die Bedienung des VUI involviert waren.¹⁷⁴ Diese Kollektion umfasst 63 Aufnahmen mit einer Gesamtdauer von sieben Stunden und 29 Minuten Audiomaterial. Diese Aufnahmen umfassen 143 Invokationen, von denen 10 im Hinblick auf die Aktivierung über das Aktivierungswort nicht erfolgreich waren, d. h. eine Aktivierung des VUI nicht stattfindet. Damit ergeben sich 133 VUI-Dialoge, in denen der Austausch zwischen Anwender*innen und VUIs beobachtbar ist. Die Analyse erfolgte – wie im methodologischen Teil der Arbeit genauer diskutiert – mit einer qualitativ-explorativen Ausrichtung. Um das Datenmaterial genauer zu beschreiben, die präsentierten Praktiken innerhalb der Kollektion in ihrer Quantität einordnen zu können sowie um die qualitativen Analysen anzureichern und die Auswahl für die Diskussion innerhalb dieses Kapitels zu begründen, wurden bei einigen Phänomenen Aussagen über die Häufigkeit getroffen. Angaben über die Häufigkeit vorgestellter Phänomene beziehen sich auf diese 63 Aufnahmen mit 143 Invokationen bzw. 133 VUI-Dialogen. Es wurden aber auch andere Aufnahmen aus den Fokushaushalten und in einzelnen Fällen auch aus anderen Haushalten analysiert, wenn dies für die Illustration eines Phänomens zweckdienlich war, und zudem wurde das gesamte Korpus für die Konzeption der Arbeit herangezogen. Lediglich die Häufigkeitsangaben beziehen sich auf die genannte Kollektion.¹⁷⁵

Die Auswertung brachte drei Kategorien hervor, die zu drei der „generischen Probleme“ der Gesprächsorganisation nach Schegloff passten. Für diese „Probleme“ muss Schegloff zufolge jede Sprache kommunikative Lösungen finden – sie unterscheiden sich zwar sprach- und kulturspezifisch in hohem Maße, ihre Mechanismen gehören aber zugleich immer zu deren spezifischem kommunikativem Repertoire (vgl. Schegloff 2006). Für VUI-Dialoge stachen besonders heraus: (1) die sequenziell in der Zeitlichkeit verlaufende Ablauf- und Handlungsorganisation, (2) die Organisation des Sprecher*innen-Wechsels (Turn-Taking) sowie (3) der Umgang mit Störungen und entsprechende Reparaturpraktiken. Jedem dieser

174 Damit lässt sich eine Beteiligung nicht gesichert ausschließen – auch ohne eine akustische Dokumentation der Anwesenheit könnten über Blick, Gestik und andere Ausdrucksressourcen andere Menschen am Vollzug des VUI-Dialogs beteiligt sein. Wenn sie allerdings über sechs Minuten hinweg nicht hörbar sind und die Anwesenheit von den Anwender*innen nicht relevant gemacht wird, ist – selbst wenn andere Personen anwesend sind – zumindest von einer ‚Beteiligung‘ mit einem niedrigen Aktivitätsniveau im Sinne Hirschauers (2016) auszugehen.

175 Die Häufigkeitsangaben für das gesamte Korpus zu erstellen hätte eine systematische Kodierung des umfangreichen Datenmaterials notwendig gemacht, was im Rahmen der Arbeit nicht parallel zu den folgenden Auswertungen leistbar gewesen wäre.

Probleme ist ein Unterkapitel gewidmet. Zwei weitere der von Schegloff (2006: 79–83) genannten Probleme (das Problem der Lexemauswahl sowie das der übergreifenden Interaktionsorganisation) werden hier nicht näher beleuchtet, sie werden im ersten Fall durch die Analysen in den übrigen drei Kapiteln abgedeckt (insbesondere in Kap. 6.1.1.3) oder, im zweiten Fall, in den Folgekapiteln relevant. Ergänzt wird hingegen ein Unterkapitel, das sich mit Anredeformen in VUI-Dialogen befasst. Anredeformen und andere Formen der Adressierung sind ebenfalls umfangreich Gegenstand sprachwissenschaftlicher Untersuchungen gewesen und sind seit den frühesten konversationsanalytischen Arbeiten relevant, wie Schegloff (1968) selbst beschreibt. Sie sind ebenfalls ein zentrales Phänomen in VUI-Dialogen: Zur Konstitution eines Dialogs, d. h., um überhaupt einen Austausch zu ermöglichen und ein Interface zu erzeugen, muss das VUI über die Produktion einer Anredeform in den Listening-Modus gebracht werden. Diese Voraussetzung bringt spezifische sprachliche Praktiken hervor und ist folgewirksam auch für die anderen Praktiken, daher wurde sie den Ausführungen zu den drei übrigen Problemen vorangestellt.

6.1.1 Anredeformen im VUI-Dialog

Anredeformen sind nominale oder pronominale sprachliche Ausdrücke, die Sprecher*innen als Ressource für explizite bzw. direkte Adressierungsverfahren zur Verfügung stehen (vgl. Hartung 2001: 1350; siehe auch Kühn 1995: 49). Die Adressiertheit einer Äußerung lässt sich in sozialer Interaktion auch durch Körperhaltung, Blick oder die Gestaltung der Äußerung zum Ausdruck bringen (vgl. Hartung 2001: 1348–1349). Diese Ressourcen könnten auch für VUI-Dialoge eine Rolle spielen und im weiteren Verlauf der Arbeit werden einige Analysen in dieser Hinsicht präsentiert – gerade mehrfach adressierte Äußerungen (siehe Kühn 1995) werden dabei noch eine Rolle spielen. Gerade die vokativische, onymische Anrede am Beginn des Austauschs ist allerdings so auffällig, dass sie auch am Beginn des Analyseteils und im Fokus dieses Kapitels steht, das daran anschließend auch durch das VUI produzierte Anreden beleuchten soll. Dabei wird – hier und im weiteren Verlauf der Arbeit¹⁷⁶ – gezeigt, dass die onymische Anrede nicht nur die Adressierung bzw., im Fall des VUIs, *Aktivierung* oder *Invo-kation* leistet, sondern auch anderweitig strukturierend für den Austausch zwischen Anwender*innen und VUI wirkt. Wie Albert/Hamann (2021: 1) in ihrer im

¹⁷⁶ Siehe insbesondere die Formulierung einer Basis-Sequenzstruktur für VUI-Dialoge in Kap. 6.1.2.

Grenzbereich von Kommunikationswissenschaft, HCI-Forschung und Sozialpsychologie zu verortenden Studie zu Aktivierungsworten bei VUIs mit Verweis auf Arbeiten u. a. von Goodwin und Schegloff formulieren, sind „first utterance[s] in a conversation [...] an important interactional resource for contextualization, for projecting upcoming action, and for establishing the fundamental interactional roles of speaker and recipient“. Auch Ergebnisse einer Befragungsstudie von Jung/Kim (2019) deuten darauf hin, dass Nutzer*innen mit dem Aktivierungswort mehr als nur eine technische Form, sondern auch eine kontextualisierende Funktion verbinden – für sich (als Kontrollmechanismus) und für andere Personen, die an der laufenden Situation beteiligt sind (als Aufzeigeleistung).

Nicht zufällig ist einer der ersten konversationsanalytischen Beiträge überhaupt eine Beschäftigung mit der Gesprächseröffnung (Schegloff 1968): Gesprächseinstiege verlaufen in hohem Maße rituell und haben ein hohes generelles Aussagepotenzial über Gesprächsmechanismen. Das wirft die Frage auf, ob Kontextualisierung, Handlungsprojekt und Rollenaushandlung auch im VUI-Dialog durch die Invokation vermittelt entstehen oder ob sie hier andere Funktionen einnehmen können. Mittlerweile gibt es außerdem die technische Option, dass VUIs selbst eine Dialogsequenz initiieren. Dies ist jedoch im vorliegenden Korpus nur in einem einzigen Fall dokumentiert, alle anderen VUI-Dialoge werden von den Nutzer*innen initiiert – und dazu findet notwendigerweise immer eine onymische Anrede in Form eines Aktivierungsworts statt. Zugleich zeigt sich, dass das Aktivierungswort nicht nur zu Beginn der Dialoge geäußert wird, sondern auch bei Re-Aktivierungen und Fortsetzungen zum Einsatz kommt.¹⁷⁷

Konversationsanalytisch ist der Gebrauch der Aktivierungswörter in VUI-Dialogen sowie auch weiter gefasst Adressierungsleistungen im Austausch zwischen Mensch und VUI nur wenig untersucht. Die bereits erwähnte Arbeit von Albert/Hamann (2021) ist eine der wenigen Studien, die sich explizit damit auseinandersetzen. Formen der Anrede ohne diesen Bezug wurden aus Sicht der empirischen Sprachwissenschaft hingegen durchaus untersucht. Es lassen sich nominale und pronominale Anredeformen unterscheiden: Nominale Anredeformen sind syntaktisch meist freistehende, optionale Nominalphrasen, die z. B. ein Name, aber auch ein Titel oder eine soziale Rolle sein können. Diese können erstens zu Beginn einer Konversation stehen und das kommunikative Problem der Zugänglichkeit bzw. der Herstellung eines geteilten Aufmerksamkeitsfokus zur Aufnahme eines weiteren Austauschs lösen; diesen Typ bezeichnet Schegloff (1968:

¹⁷⁷ Abzugrenzen sind diese Adressierungsleistungen von anderen Formen der Personenreferenz, etwa Selbstreferenzierungen und Referenzen auf dritte an- und abwesende andere Personen (vgl. Stivers/Enfield/Levinson 2007: 1–2; Schegloff 1996: 481–482).

1080) als „*summonses*“. Zweitens können sie, im weiteren Verlauf von Interaktionen geäußert, eine Vielzahl weiterer Funktionen erfüllen. Pronominale Anredeformen sind demgegenüber häufig nicht freistehend, sondern „*bound forms of address*“ (vgl. Braun 1988: 11), obschon diese Unterscheidung, wie Braun (1988: 11) selbst an Beispielen illustriert, nicht durchgängig ist und auch gebundene nominale Anreden sowie ungebundene pronominale Anreden (z. B. freistehendes „*Sie!*“) möglich sind. Die folgenden Analysen konzentrieren sich zunächst auf onymische und insofern nominale Anredeformen, die Anwender*innen gebrauchen, um das VUI zu aktivieren. Sie werden also insbesondere in ihrer Verwendung als Anredepraktiken in *Summons-Answer-Sequenzen* untersucht. Daran anschließend werden Anreden in den Blick genommen, die das VUI produziert – zunächst pronominale, anschließend auch onymische Anredeformen.

Zuvor soll ein knapper Überblick über bisherige Studien zu Anredeformen gegeben werden. Dabei sind verschiedene Typen Arbeiten zu unterscheiden: erstens solche, die sich insbesondere aus soziolinguistischer Perspektive mit pronominalen Anreden beschäftigen und dabei Fragen von Beziehungsgestaltung, Macht und Höflichkeit in den Blick nehmen (Brown/Gilman 1968; Ervin-Tripp 1972; Fasold 1990: 1–38; Besch 1996; für einen Überblick siehe Braun/Kohz/Schubert 1986). Zweitens wurden umfangreich sprachstrukturell ausgerichtete, sowohl sprachspezifische wie auch -vergleichende Studien unternommen, die sich insbesondere auf das pronominale Anredesystem der jeweiligen Sprachen beziehen, für das Deutsche etwa Bayer (1979), Amendt (1995) und Besch (1996); einzelsprachlich ausgerichtete Arbeiten zu anderen Sprachen versammelt Winter (1984), z. B. zum Arabischen, Rumänischen, Russischen und anderen Sprachen; kontrastive Studien stammen u. a. von Vorderwülbecke (1976), Kuglin (1977), Kohz (1982) oder Hess-Lüttich (1993). Darüber hinaus wurden umfangreich auch *onymische* bzw. nominale Anredeformen untersucht, die weniger auf das pronominale System abzielen, sondern Personenreferenzen sowohl diachron wie synchron untersuchen (vgl. Macha 1997; Besch 2004; Christen 2006; Nübling/Fahlbusch/Heuser 2012).

Bei den bisher genannten Arbeiten kommen allerdings die praktischen Anwendungskontexte und -formen von Anredeformen nicht in den Blick. Ebenso sind empirische Untersuchungen zu interaktionalen Verwendungsweisen onymischer Anredeformen in der bisherigen Forschung besonders dünn repräsentiert (vgl. Günthner 2016: 410); zu diesem Bereich existiert bislang nur ein überschaubares Feld von Arbeiten. Dazu zählt die Betrachtung namentlicher Anrede bei der *turn allocation* (vgl. Lerner 2003).¹⁷⁸ Rendle-Short (2010) beschäftigt sich mit der Positionierung von

178 Zu Turn-Taking in Dialogen mit einem VUI siehe Kap. 6.1.3.

onymischen Formen in Interviews; Norrick/Bubel (2009) können zeigen, dass onymische Adressierungen als Humorressource gebraucht werden – demgegenüber zeigen Butler/Danby/Emmison (2011), dass die gleichen Formen auch für die Anzeige von Nicht-Übereinstimmung und potenziellen sequenziellen Inkongruenzen genutzt werden; zur Vielzahl der Gebrauchskontexte arbeitet auch Clayman (2010). Günthner (2016) demonstriert eindrücklich, dass der Gebrauch onymischer Anredeformen „[a]ufgrund ihres Potenzials zur Aufmerksamkeitsfokussierung, Beziehungskonstitution und Indizierung personifizierter Zuwendung“ (Günthner 2016: 433) als Ressource im Beziehungsmanagement dient (siehe auch Bauer 2016 für ähnliche Befunde zu onymischen Adressierungen in der Chat-Kommunikation).

Anredeformen sind kontextsensitiv: Sie werden durch den Kontext beeinflusst, in dem sie geäußert werden, und zeigen den Kontext an, in dem sie sich bewegen. Sie sind Teil der zwischenmenschlichen Kommunikation; diese „verweist nicht nur indexikalisch auf kulturelle Konventionen, Habitualisierungen und Erwartungen, sondern konstituiert diese aktiv mit“ (Günthner 2017: 24). Sie sind im Sinne Goffmans (1974: 262) „Beziehungszeichen“, die auf der einen Seite einzelne Individuen in ein soziales Geflecht einordnen, auf der anderen Seite aber auch dieses soziale Geflecht kontextsensitiv rekontextualisieren. Darum sind gebrauchsbasierte kulturvergleichende Studien zu Anredeformen besonders aufschlussreich: weil sie über den praktischen Vollzug der Beziehungsarbeiten Aufschluss geben können (vgl. Günthner 2017; Günthner/Zhu 2017). Einen stärker gebrauchsbasierten Ansatz fordert konsequenterweise auch de Stefani (2012) mit der Begründung einer „interaktionalen Onomastik“, die konversationsanalytisches Vorgehen in die Namensforschung überträgt und die sozialen Verwendungskontexte von (hier: Orts-)Namen berücksichtigt; in dieser Weise arbeiten auch Günthner (2016; 2017; 2018a; 2020) und Günthner/Zhu (2015; 2017). Berücksichtigt man den Befund, dass „onymische Praktiken [...] sich im Prozess der Interaktion in enger Abstimmung und Koordination mit den Handlungen des Gegenübers [entfalten]“ (Günthner 2020: 208) und dass diese Gespräche strukturieren und organisieren (vgl. Clayman 2012a: 1854), so ist es weniger verwunderlich, dass die von Nutzer*innen geäußerten namentlichen Adressierungen, die an die Smart Speaker gerichtet sind, durchaus kontextsensitiv im Hinblick auf Tonhöhe, Betonung und Stimmmodulation variieren (vgl. Albert/Hamann 2021: 4; Cohn/Zellou 2021), obwohl die VUIs diese Form der Variation zum derzeitigen Entwicklungsstand nicht verarbeiten können. Nutzer*innen neigen hier aber dazu – das suggeriert das empirische Material –, diese Variationsleistungen neben anderen Charakteristika zwischenmenschlicher Interaktion auf die Dialoge mit Smart Speakern anzuwenden. Dies unterstreicht auch die eingangs erwähnte kontextualisierende Funktion der Anredeformen, die für VUI-Dialoge untersucht werden soll.

6.1.1.1 Summons-Answer-Sequenzen

Empirisch zeigt sich mit Blick auf die praktische Gestaltung von Anreden im Zusammenhang mit VUIs sehr deutlich die Relevanz sogenannter „Summons-Answer“-Sequenzen, die jedem Austausch mit dem VUI vorangehen, meist vor jeder Paarsequenz aus Stimmein- und Stimmausgabe.¹⁷⁹ Clayman (2012a: 1856) empfiehlt, Anreden nicht als „address terms“, sondern vielmehr „in terms of the actions [...] that they implement“ zu bezeichnen; insofern wird hier nun zunächst auch von „Summons“ die Rede sein. Aus der frühen Konversationsanalyse ist zu Summons insbesondere erneut die Arbeit von Schegloff (1968) zu nennen, der sich mit der sequenziellen Struktur von Gesprächseröffnungen beschäftigt hat. Als Summons werden dabei „attention-getting devices“ (Schegloff 1968: 1080) beschrieben, die jedoch, so Schegloff weiter, unterschiedlicher Art sein können und den ersten, semantisch leeren Paarteil einer paarsequenziellen Struktur bilden. Mit dem ersten Paarteil wird lediglich die Aufmerksamkeit erlangt: Neben technischen Funktionen wie einem Telefonklingeln können dies insbesondere onymische Praktiken, spezifische Phrasen (z. B. „Entschuldigen Sie!“) oder auch körperliche Praktiken sein (vgl. Schegloff 1968: 1080). Explizit sind hier mit dem Klingeln eines Telefons oder dem Klopfen an der Tür (vgl. Hopper 1992: 56) Geräusche Bestandteil des sequenziellen Interaktionsverlaufs. Im Falle von VUIs sind insbesondere onymische Adressierungen relevant, die als Aktivierungswort oder Invokation (auch „Hot Word“ oder „Wake Word“) bezeichnet werden. Obschon die Aktivierung auch über die Smartphone- bzw. Tablet-App sowie bei einigen Modellen auch taktil über einen Aktivierungsknopf am Gerät selbst erfolgen kann,¹⁸⁰ kann über diese Form der Aktivierung nur wenig gesagt werden, weil diese im vorliegenden Korpus nicht systematisch erfasst sind: Durch die Form der Datenerhebung, die sich bei der longitudinalen Betrachtung auf die Anwendung des Aktivierungsworts als Trigger für die Aufzeichnung stützt,¹⁸¹ werden solche Aktivierungen nicht systematisch erfasst. In den Ersteinrichtungsaufzeichnungen wird die Aktivierung aufgrund des fortlaufenden Dialogstatus und der ablaufenden Einrichtungsskripte nur an spezifischen Stellen realisiert – man könnte von einem „continuing state of incipient talk“ (Schegloff/Sacks 1973: 325) sprechen.

¹⁷⁹ Siehe Kap. 6.1.2.

¹⁸⁰ Siehe dazu Kap. 3.3. Einige Modelle ermöglichen auch die Nutzung einzelner Funktionen über Berührungen: So ist es etwa beim Google Home Mini möglich, die Wiedergabelautstärke (sowohl der VUI-Stimme wie auch der abgespielten Medieninhalte) über eine Berührung an der Oberfläche zu steuern; eine Berührung an der linken Seite bedeutet dabei leiser, an der rechten Seite lauter. Der Smart Speaker ratifiziert diese Weise der Lautstärkeregelung sowohl durch ein bestätigendes Geräusch wie auch durch eine leuchtende Anzeige auf der Oberfläche.

¹⁸¹ Siehe Kap. 5.3.

Wird sie doch realisiert, handelt es sich eher um Dialoge zum Testen und Einlernen, in denen ebenfalls das Aktivierungswort als Summons verwendet wird, nicht die App oder die taktile Form der Aktivierung; aus diesen Dialogen werden Beispiele gezeigt.

Der zweite, durch das Summons konditionell relevant gesetzte Paarteil ist die „Answer“: die Replik auf die explizite Adressierung. Stukenbrock (2013b: 231) spricht von „Fokussierungsaufforderung“ und „Fokussierungsbestätigung“. Hopper (1992: 57) bezeichnet Summons-Answer-Sequenzen als „marked case“ in persönlichen Begegnungen. Die Realisation einer Summons-Answer-Sequenz setzt einen gewissen Grad an Asymmetrie voraus, der durch das Summons aufgelöst wird, d. h., eine*r der Beteiligten sieht den*die andere*n (noch) nicht und weiß noch nicht über das bevorstehende Kommunikationsereignis Bescheid. Nur eine*r der Beteiligten weiß, dass eine wechselseitige Wahrnehmung – als Voraussetzung für eine Interaktion – geplant ist und fordert den*die andere Beteiligte*n mittels des Summons dazu auf, für eine Konversation bereitzustehen. So wird das kommunikative Problem der „Verfügbarkeit“ gelöst, das Voraussetzung für einen Austausch ist (vgl. Schegloff 1968: 1089; siehe auch Day/Wagner 2008: 36–38). Schon der Beginn einer Ko-Operation wird also ko-operativ von den beiden Beteiligten hervorgebracht, in dem systematisch die Answer-Sequenz Bestandteil dessen ist. Die Summons-Answer-Sequenz wird, das zeigen die bisherigen Ausführungen, also niemals ‚allein‘ realisiert, sondern von beiden Beteiligten gemeinsam hervorgebracht, ein Blick auf den praktischen Vollzug ist bei der analytischen Betrachtung dessen also inhärent (vgl. Goodwin 1980: 280).

Der Begriff des Summons bzw. der Summons-Answer-Sequenz wird nicht systematisch für Mensch-Maschine-Dialoge angewendet, findet aber bei den konversationsanalytischen Arbeiten aus diesem Bereich etwa bei Pitsch (2015) sowie Opfermann/Pitsch (2017) Erwähnung. Lotze (2016: 244) stellt fest, dass in Mensch-Maschine-Dialogen Paarsequenzen wie Summons-Answer-Sequenzen „funktional nicht immer ineinander[greifen]“, stattdessen beobachtet sie „Inkohärenzen oder Quasi-Kohärenzen“, die sie primär mit pragmatischen Kategorien wie der „Auswahl eines unangemessenen Sprechakts“ (Lotze 2016: 244) erklärt. Ausführlich widmet sich Krummheuer (2010: Kap. 6) den Gesprächseröffnungen und konstatiert: „Ähnlich wie die Eröffnungen von Telefongesprächen wird auch der hybride Austausch durch eine ‚summon-answer‘-Sequenz [sic] eröffnet“ (Krummheuer 2010: 161). Ihr zufolge sind Summons-Answer-Sequenzen kausal dafür verantwortlich, dass der Einstieg in den hybriden Austausch „kein Problem“ darstellt (Krummheuer 2010: 162). Es ist bemerkenswert, dass Hopper (1992: 54) gerade die (durch die Summons-Answer-Sequenz aufzulösende) Asymmetrie hervorhebt, ist dieser Begriff doch seit Suchman (1987) als wesentliches Charakteristikum von Mensch-Maschine-Dialogen bekannt. Diese besteht im Falle der hier untersuchten

VUI-Dialoge gerade darin, dass die menschlichen Beteiligten jederzeit den Smart Speaker ‚wahrnehmen‘ können (auch visuell), während der Smart Speaker explizit in den entsprechenden Modus gebracht werden muss, um eine Aufzeichnung sensorisch zu ermöglichen – und kann als das gesehen werden, was Suchman (1990: 43) als „extreme form of resource-limited communication“ beschreibt (siehe auch Suchman 2021: 76).¹⁸²

Es stellt sich nun mit Blick auf die alltagspraktische Einbindung die Frage, wie der Summons dabei von den Nutzer*innen realisiert wird, wie er sich als Gesprächsbestandteil manifestiert und ob er auch hier ‚problemlos‘ eingesetzt werden kann. Die Ergebnisse von Albert/Hamann (2021) deuten darauf hin, dass dabei eine Orientierung an ‚gewohnten‘ konversationellen Mustern erfolgt – dies soll genauer untersucht werden. Darüber hinaus soll die Realisation der Answer-Sequenz in den Blick genommen werden: (Wie) wird diese realisiert? Wie wird sie wiederum sequenziell durch die Nutzer*innen eingebunden und wie trägt dies zum Einstieg in eine Konversation bei? Ferner sollen abschließend Fälle betrachtet werden, in denen das VUI den Dialog initiiert, und auf die sprachliche Umsetzung dabei geachtet werden.

Bevor ein erstes Beispiel vorgestellt wird, soll kurz auf die Aktivierungsworte selbst eingegangen werden. Ein Aktivierungswort erfordert eine hohe phonologische, onomastische Spezifität, um Äußerungen außerhalb des Kontexts eines VUI-Dialogs und somit Fehlaktivierungen zu vermeiden; so gilt in der HCI-Forschung: „The WUW [Wake-up-Word, T.H.] must be spoken in the context of requesting attention, i. e. alerting context and should not be recognized in any other context.“ (Képuska 2011: 239).¹⁸³ Die Aktivierungsworte für die jeweiligen VUIs lauten „Alexa“ für die Produkte von Amazon, „Hey Google“ und „Okay Google“ für die Produkte von Google sowie „Hey Siri“ für die Smart Speaker von Apple.¹⁸⁴ Dabei fällt zunächst auf, dass das Aktivierungswort „Alexa“ aus einem einzigen Lexem besteht, das sich als Onym deuten lässt, während „Hey Google“, „Okay Google“ und „Hey Siri“ jeweils zweigliedrig sind. Dabei lässt sich „Hey“ auch klar als „aufmerksamkeitsheischende“ (Neuland/Könning 2021: 90) Sum-

182 Siehe Kap. 2.2.2.

183 Dass diese Spezifität Herausforderungen für nicht englischsprachige Sprecher*innen mit sich bringt und insbesondere für Sprachen mit phonologisch, syntaktisch und pragmatisch anderen Sprachstrukturen, zeigen Bleakley et al. (2021).

184 Die Aktivierungsworte sind in einem beschränkten Rahmen individuell adjustierbar. Dies wurde jedoch in den Haushalten, aus denen die vorliegenden Daten stammen, unterlassen, weil dies nicht kompatibel mit der Erhebungstechnologie gewesen wäre, die auf dieselben Aktivierungsworte reagiert (vgl. Hector et al. 2022).

mons-Interjektion klassifizieren,¹⁸⁵ die auch als Gruß verstanden werden kann – was nicht verwunderlich ist, da es regelmäßig zu semantischen Überlappungen zwischen Summons und Gruß kommt (vgl. Schegloff 1968). Das gemeinhin als Gesprächspartikel klassifizierte Lexem „Okay“ hingegen zeichnet sich zwar durch eine sehr große semantische Vielfalt aus (vgl. Metcalf 2011: 9; Kessler 2010; Betz et al. 2021), ist aber bisher vorwiegend in anderen Funktionen beschrieben worden, etwa als Responsiv (vgl. Zifonun/Hoffmann/Strecker 1997: 63; Betz/Deppermann 2021), zum Vollzug von Pre-Closings (vgl. Schegloff/Sacks 1973: 327; Mondada/Sorjonen 2021), um Themenbereiche, einschließlich ungelöster Probleme, abzuschließen (vgl. Bangerter/Clark/Katz 2004: 3) und um Verstehen zu dokumentieren (vgl. Deppermann 2018b: 66–71).¹⁸⁶ Cirko (2016: 30) benennt „okay“ in Kombination mit „ja“ auch als „Aufmunterungssignal“ in einer kontrastiven Betrachtung von Prüfungssignalen, und Zifonun/Hoffmann/Strecker (1997: 376) erwähnen, dass „okay“ turneinleitend auftritt, gerade am „handlungsstrukturellen Übergang“ bzw. auch zur Markierung thematischer Übergangspunkte (vgl. Beach 1993: 334; Levin/Gray 1983). Beach (1993: 337) verweist dabei auf eine „dual function“ von turnintialem „okay“, mit dem sowohl auf den vorherigen Zug geantwortet als auch der nächste Zug vorbereitet werden kann. Damit hat „okay“ durchaus auch einen gewissen initialen Charakter. Als Summons oder Gruß, d. h. mit einem deutlich höheren Grad an Initialität und aufmerksamkeitsgewinnender Funktion, ist „okay“ bisher hingegen nicht beschrieben worden.

Ebenfalls auffällig ist, dass „Alexa“ und „Siri“ zusätzliche Onyme sind, die über den Markennamen des Herstellers hinausreichen, während „Google“ keine Unterscheidung zwischen Firmenname und VUI-Benennung vornimmt. Die Adressierung erfolgt somit bei Google mit eindeutigem Bezug zur populären Suchmaschine, während bei Amazon und Apple die VUI-Bezeichnungen „Alexa“ und „Siri“ eingezogen sind. Die Personifizierung, die durch die Onymizität entsteht, vergleichen Natale/Cooke (2021: 1009) mit den Metaphern grafischer Interfaces (etwa Desktop, Ordner oder Fenster). Die ‚persönlichen Assistenten‘ existierten so betrachtet, so die Argumentation von Natale/Cooke (2021: 1009), nur auf semiotischer Ebene – sie sind eine Repräsentation und strukturieren das Interface und somit die Eingaben der Nutzer*innen (siehe auch Natale 2020). Damit verbunden ist die Konstruktion einer „Persona“ für das VUI, zu der neben dem Namen weitere Faktoren gehören (vgl. Lee/Kim/Lee 2019).¹⁸⁷ Der Grad der Persona-Gestaltung bei Google ist allerdings

¹⁸⁵ Zu den Funktionen von „Hey“ als Summons siehe ausführlich Haviland (2015).

¹⁸⁶ Einen Überblick über Studien zu „okay“ liefern Herzberg (2016) und Betz et al. (2021); weitere Studien zu „okay“ als Verstehensprozessmarker besprechen Bangerter/Clark/Katz (2004: 3).

¹⁸⁷ Bucher (2014) diskutiert die Persona-Entwicklung von Social Bots vor dem Hintergrund der Gestaltung von parasozialen Beziehungen. Der Begriff ist in der HCI-Forschung gängig, um

schon im Hinblick auf das Aktivierungswort deutlich geringer ausgeprägt als bei anderen Systemen. Auch wenn diese Unterschiede in keinem der nachfolgenden Beispiele von den Beteiligten relevant gesetzt werden,¹⁸⁸ könnten insbesondere die unterschiedlichen Onymizitäts- und Invokationspotenziale der verschiedenen Aktivierungsworte auch Einfluss auf die Kontextualisierungsfunktion haben – dies wird im Rahmen der Analysen mit zu prüfen sein.

Betrachten wir nun zwei verschiedene Typen von Summons-Answer-Sequenzen im vorliegenden Korpus und insbesondere die sprachliche Gestaltung des Summons. Diese ist entweder, wie im folgenden Beispiel, durch eine Pause vom folgenden Turn abgesetzt:

Beispiel (1): Wecker (01-CVR01-09)

003 LF: aLEXa?
 004 (1.4)
 005 LF: stell ein wecker AU:F (-) fünfzehn
 minuten.
 006 (1.4)
 007 AL: FÜNFzehn minuten ab jEtzt;

Nach der Invokation durch den Anwender Lukas (LF, 003) wartet dieser 1,4 Sekunden ab. Erst im Anschluss erfolgt die Formulierung einer Stimmeingabe. Er geht damit anders vor als der Anwender Konrad (KW) im folgenden Ausschnitt:

Beispiel (2): Pause (07-CVR01-14)

128 KW: hey siri (0.2) PAUse.
 129 (1.9)
 130 SI: im MOment wird,
 131 KW: [(bald geht wieder die POST.)]
 132 SI: [(unverständlich)) wiedergegeben;]
 133 k: ((Musik stoppt))

menschliche Subjekte von maschinellen Dialog-Agenten zu unterscheiden, die teilweise die Illusion einer Person erwecken sollen (vgl. Lotze 2016: 73–74).

188 Ein Beispiel, in dem Nutzer*innen im Zuge der Ersteinrichtung ausführlich Vergleiche anstellen, findet sich in Kap. 6.2.2. Auch hier werden aber keine Unterschiede in der Adressierung der Smart Speaker deutlich.

Konrad formuliert seine Invokation prosodisch innerhalb einer einzigen Intonationsphrase. Während sich der Sprecher aus Beispiel (1) mit seiner Pause zumindest andeutungsweise an den Strukturen einer Summons-Answer-Sequenz orientiert und die Pause ein deutlicher Hinweis darauf ist, fällt diese Orientierung in Beispiel (2) weg. Die mit 1,4 Sekunden vergleichsweise lange Pause (vgl. Levinson 2016) in Beispiel (1) könnten einerseits ein ‚Überbleibsel‘ aus der Summons-Answer-Struktur sein: An die Stelle, an der konversationell üblicherweise eine Ratifizierung des Summons und somit eine Bestätigung des gemeinsamen Aufmerksamkeitsfokus durch das Gegenüber produziert wird, tritt hier eine Pause, in der technisch der Listening-Modus aktiviert werden kann. Möglich wäre auch, dass ein visuelles Signal abgewartet wird, das der hier verwendete Amazon Echo Dot in Form eines an der Oberfläche des Geräts angebrachten Leuchtrings aussendet, der sich blau verfärbt.

Demgegenüber weicht der Sprecher aus Beispiel (2) vom Muster der „Summons-Answer-Sequenz“ ab: Nach dem Summons wird keine Answer-Sequenz abgewartet, sondern in der gleichen Intonationsphrase die Folgeäußerung produziert. Seitens der VUIs wird diese Folgeäußerung bei den Modellen von Amazon und Google nicht explizit erwartbar gemacht, die Erwartbarkeit des nächsten Zuges nach einem Summons ist technischerseits jedoch gleichwohl angelegt: Folgt keine verbale Äußerung oder wird eine solche nicht erkannt, beenden die Geräte den Zuhörmodus nach einer festgelegten Zeitspanne wieder. Dieser sogenannte „no-speech-timeout“ (Pearl 2016) oder „no-input-timeout“ (Skantze 2021: 13) erfolgt bei Amazon nach acht¹⁸⁹ und bei Google nach fünf Sekunden.¹⁹⁰ Der Smart Speaker von Apple ist das einzige Modell, das auch das Ausbleiben einer Folgeäußerung nach dem Summons konversationell relevant setzt und durch eine Rezeptionspartikel ‚einfordert‘, was die Benutzer Konrad und Till im folgenden Beispiel reflektieren:

Beispiel (3): Reaktion (07-CVR01-03)

005 KW: ((räuspert sich))
 006 (1.0)
 007 KW: °h hey SiRi?
 008 (2.5)
 009 SI: mh_HM?
 010 (1.1)
 011 KW: hh° <<lachend> (spie,) > SPIEle: (.) get
 into it (.) vo::n (-) doja cat.

¹⁸⁹ Siehe Amazon (2023).

¹⁹⁰ Siehe Google (2023).

012 (2.9)
 013 SI: ich spiele get into it yuh von DOja cat
 jetzt.
 014 (1.2)
 015 k: ((Klirren))
 016 KW: voll FUNny wenn man nicht (-) direkt
 spricht-
 017 hast du geHÖRT grade?
 018 TW: JA:.
 019 KW: hast du gehört was die geMACHT hat?
 020 (0.7)
 021 TW: ich SPIELE,
 022 KW: NEI:N davor,
 023 (0.3)
 024 TW: NEIN.
 025 KW: ich hab grad ihren NAMen gesagt,
 026 TW: und dann hat die [geSAGT-]
 027 KW: [und hab dann NICH]
 sofort reagie:rt.
 028 weil ich nicht wusste ob sie schon ZUHört.
 029 (0.3)
 030 TW: JA,
 031 (0.4)
 032 KW: [ich mach das NOCHmal,]
 033 k: [((“Get into it” von Doja Cat setzt ein))]
 034 (0.8)
 035 KW: hey SIri?
 036 k: ((Musik spielt, pausiert kurz und beginnt
 erneut; gleichzeitig gelegentliches leises
 Klappern und Rascheln, 7.0 Sek.))
 037 KW: die hat das NICHT gehört.
 038 HEY siri,
 039 TW: <<creaky> (ach GEIL,>)
 040 k: ((Klappern)) [((Musik endet))]
 041 KW: [PAUSe?]
 042 (2.9)
 043 k: ((Musik setzt wieder ein))
 044 (0.8)

045 KW: hey SIri?
 046 (1.3)
 047 KW: STOP.
 048 [(1.9)]
 049 k: [((Musik endet))]
 050 KW: jetzt SCHAU ma.
 051 wie FUNny.
 052 TW: [(STOP.)]
 053 k: [((Klappern))]
 054 (2.2)
 055 KW: warte äh (.) a_SO.
 056 (0.3)
 057 KW: hör ZU,
 058 HEY siri,
 059 (2.5)
 060 SI: mh_HM?
 061 KW: °h °h °h PLAY;
 062 (0.7)
 063 k: ((Musik setzt ein))
 064 KW: hast du geHÖRT?
 065 (0.7)
 066 TW: JA:.

Das Beispiel zeigt zunächst auf, dass die Pause von 2,5 Sekunden (008) zu einer „Antwort“-Sequenz führt, die für „Fokussierungsbestätigungen“ (Stukenbrock 2013b: 231) typisch ist: eine semantisch leere, zweisilbige Partikel mit steigendem Intonationsverlauf („mh_HM?“, 060). Das Beispiel weist allerdings darüber hinaus auf einige weitere interessante Punkte im Zusammenhang mit der Adressierung der VUIs hin. Auffällig ist nämlich, dass der Nutzer Konrad die Fokussierungsbestätigung seitens des VUI nicht für erwartbar hielt, wie sich erstens daran zeigt, dass er die Folgeäußerung lachend produziert (011), und zweitens dadurch expliziert wird, dass er im weiteren Verlauf seinem Mitbewohner Till davon berichtet und dies als „FUNny“ (016, 051) bewertet. Auch liefert er bei der Erklärung an seinen Mitbewohner, der das Geschehen nicht aufmerksam verfolgt hatte, einen „account“ (vgl. Heritage 1988; Garfinkel 1967: 36) dafür, nicht sofort gesprochen zu haben, sondern die längere Pause abgewartet zu haben („weil ich nicht wusste ob sie schon ZUHört.“, 028). Dies verweist darauf, dass die Herstellung einer gemeinsamen Aufmerksamkeitsfokussierung von Anwender und VUI in der Tat ein kommunikatives Problem im Austausch mit dem VUI darstellt, für das die explizite

Bestätigung der Aufmerksamkeitsfokussierung durch die Partikel eine sprachliche Lösung darstellt, die jedoch von Konrad nicht realisiert wurde. Nachdem sich bei der anschließenden Wiederholung zur Demonstration für Till zunächst Schwierigkeiten ergeben – läuft gerade Musik, wird die Fokussierungsbestätigung nicht produziert (vgl. 035–056) –, stoppt der Nutzer die Musik. Er stellt anschließend sicher, dass auch Till seine Aufmerksamkeit auf das Geschehen richtet („hör ZU“, 057) und reproduziert anschließend die Summons-Answer-Sequenz.

Ferner zeigt das Beispiel die Ambivalenz des Gesprächsbeteiligtenstatus des VUI auf: Im Innenverhältnis ihres sequenziellen Vollzugs orientieren sich Konrad und das VUI an den Strukturen einer Summons-Answer-Sequenz. Dabei ist zu bemerken, dass auch diese von zwischenmenschlichen Summons-Answer-Sequenzen durchaus abweicht: Eine Pause von 2,5 Sekunden wäre hierbei durchaus markiert und außerdem nicht regelhaft genau 2,5 Sekunden lang, sondern – sensitiv gegenüber dem Kontext, der momentanen Beschäftigung der anderen Person, den kognitiven Leistungen, den interaktiven, räumlich-akustischen Bedingungen – tendenziell kürzer oder auch länger. Im Verhältnis zwischen Konrad und Till wird außerdem diese Summons-Answer-Sequenz Gegenstand eines anderen Gesprächsstrangs, in den sich die anschließende Wiederholung einbettet. Damit schreibt Konrad dem VUI einen Gesprächsbeteiligungsstatus zweiter Ordnung zu, die Partizipation am Gespräch ist nicht mehr gleichwertig im Sinne eines zwischenmenschlichen Gesprächs, sondern vielmehr ist das VUI mit seinen technischen Fähigkeiten Objekt einer Vorführung geworden. Auch im folgenden Beispiel zeigt sich dieser Vorführungscharakter, in dem Konrad sein Wissen über die Spezifika der Summons-Answer-Sequenz bei dem VUI von Apple an seinen Mitbewohner Till weitergibt:

Beispiel (4): Intercom (07-CVR01-08)

017 KW: hey SIri,
 018 (0.6)
 019 KW: INtercom:-
 020 (0.4)
 021 KW: äh: ich bin zuHAUse;
 022 (2.0)
 023 SI: oke: inte_com an alle in mein HAUS.
 024 k: ((Stille mit vereinzelt Signaltonen,
 ca. 3 Sek.))
 025 TW: oh;
 026 (2.3)
 027 SI/KW <<Wiedergabe der aufgezeichneten Äußerung
 von KW aus Z. 21; p> äh: ich bin

zuHAUse;>
 028 (1.9)
 029 TW: hey SIri,
 030 (0.7)
 031 TW: intercom: an alle in mein HAUS?
 032 (1.7)
 033 SI: oke: inte_com an alle in mein HAUS;
 034 (0.3)
 035 SI: [((iMessage Senden-Ton))]
 036 k: [((iPhone SMS-Ton))]
 037 (0.2)
 038 TW: das is ja COOL;
 039 KW: (ja jetzt hast du EInfach nur;)
 040 TW: ich hab jetzt einfach nur geATmet;=ne,
 041 (0.8)
 042 KW: nein;
 043 (1.3)
 044 TW: ach SO;
 045 (0.3)
 046 TW: ah:: das heißt das geRÄT-
 047 KW: [(du kannst EInfach-)]
 048 TW: [(is praktisch)] mein HAUS;
 049 (0.7)
 050 KW: nein du kannst einfach (.) nein nein also
 du musst das gAr nich SAgen;
 051 TW: das heißt [intercom] an ALle,
 052 KW: [((unverständlich))]
 053 (0.4)
 054 KW: du musst einfach nur sagen INtercom;
 055 [(ja du KANNST-)]
 056 TW: [hey SIri,]
 057 (0.8)
 058 TW: INtercom?
 059 (-) guten MORgen;
 060 (0.2)
 061 SI: oke: (.) SPRICH,
 062 (0.2)
 063 SI: ((Intercom-Aufnahme-Ton))
 064 (0.2)
 065 TW: guten MORgen;

066 (1.3)
 067 SI: ((iMessage Senden-Ton))
 068 (0.8)
 069 KW: was untersch (.) der UNterschied is bei
 [äh-]
 070 k: [((iPhone SMS-Ton))]
 071 KW: bei SI:ri,
 072 TW: oh ich hab eine NACHricht bekommen;
 073 KW: bei der siri is du musst nicht WARten,
 074 (0.7)
 075 KW: wie bei der [aLEXa;]
 076 TW: [a:h;]
 077 KW: du kannst einfach nur (.) du kannst
 einfach LOSsprechen;
 078 du kannst auch einfach SAgen (.) warte
 kurz,
 079 (1.0)
 080 KW: hey SIri,
 081 an alle (.) es ist Abendessen;
 082 (1.8)
 083 SI: ((iMessage Senden-Ton))
 084 (3.0)
 085 TW: [und jetzt (.) JA-]
 086 SI: [((SMS-Ton))]
 087 TW: wahnsinn (.) das ist der HAMmer;
 088 (0.5)
 089 TW: das geht bei der alexa so nämlich NISCH;

Zunächst steht das Feature „Intercom“ von Apple in Rede, das den Versand einer Audio-Nachricht an alle im Haushaltsnetz angeschlossenen Endgeräte ermöglicht.¹⁹¹ Dabei diskutieren Konrad und Till die Frage, wie die Reihenfolge von Invokation („Hey Siri“), Intent („Intercom“) und Entity (jeweilige spezifische Nachricht) hier zu realisieren ist.¹⁹² Konrad präsentiert diese Funktion scheinbar erstmals gegenüber Till (020–024) und formuliert einen entsprechenden Intent sowie eine direkt daran anschließende Nachricht, die aufgezeichnet und verschickt werden soll („äh: ich bin zuHAUse;“, 021). Siri beantwortet dies mit „oKE: inte_com an alle in mein hAU.“

¹⁹¹ Siehe dazu Apple (2023).

¹⁹² Siehe 3.2.2 für Erläuterungen zum technischen Hintergrund.

(023), wobei das VUI zunächst die Partikel „oKE:“ als Turnübergabebestätigung einsetzt, um anschließend den verstandenen Intent zu präsentieren. Im Anschluss daran wird die aufgezeichnete Audio-Nachricht aufgezeichnet und – auf der Aufnahme deutlich leiser hörbar – wiedergegeben (027). Aus dieser Sequenz leitet Till in der Folge eine Formulierung ab, wie er selbst die Spracheingabe zu formulieren habe, und probiert dies sogleich aus (029–031). Er produziert allerdings nicht direkt im Anschluss an die Stimmeingabe zur Aktivierung der Funktion auch eine entsprechende Audionachricht – dies wird von Konrad korrigiert und anschließend erneut probiert (054–068), wobei Till hier noch immer eine für das VUI „zu lange“ Zeitspanne zwischen der Stimmeingabe zur Funktionsaktivierung und der aufzuzeichnenden Nachricht vergehen lässt, sodass eine verbale Bestätigung der Funktionsaktivierung seitens des VUI und die Aufzeichnung der zu versendenden Audionachricht sehr nahe beieinander liegen (059–061) und Till die Nachricht noch einmal wiederholt (065). Daraufhin wendet sich das Gespräch spezifischer dem Summons-Answer-Verfahren zu. Konrad erklärt: „bei der siri ist du musst nicht WARten.“ (073) und „du kannst einfach nur (.) du kannst einfach LOSsprechen.“ (077). Er demonstriert anschließend erneut die Intercom-Funktion, wobei er die Stimmeingabe zur Aktivierung durch den abgekürzten Ausdruck „an alle“ ersetzt. Auffälligerweise produziert er zwischen Invokation und dem Intent gleichwohl eine (kurze) Pause – obwohl genau dessen Obsoleszenz präsentiert werden sollte. Er produziert die Äußerungen aber innerhalb einer Intonationsphrase, wie bereits in Beispiel (1).

Das Reden darüber, wie das VUI anzusprechen ist und wie anschließend die Stimmeingabe formuliert werden kann, zeigt, dass eine Orientierung an den Prinzipien der Gesprächsorganisation erfolgt, etwa an bekannten Mustern zur Gestaltung von Summons-Answer-Sequenzen, die eine Fokussierungsbestätigung erforderlich machen und insofern eine Pause relevant setzen. Dass diese weggelassen werden kann, kann zum Gegenstand einer metasprachlichen Aushandlung werden, wie hier bei Till und Konrad. Dabei greifen die Anwender auch auf das Wissen zur Bedienung anderer Modelle zurück, bei denen dieses Vorgehen von den Beteiligten als nicht funktional klassifiziert wird (075, 089). Obwohl mit dieser Pause also einerseits auf konversationelle Verfahren der gemeinsamen Aufmerksamkeitsherstellung verwiesen ist, zeigt sich darin auch die vorübergehende Verfestigung und parallele Auflösung einer sprachlichen Praktik, die darauf aufbaut: die kurze Pause nach der Invokation, die bereits in Beispiel (1) und (2) illustriert wurde. Dafür, dass diese bereits in Auflösung begriffen ist, spricht die metasprachliche Bearbeitung von Konrad und Till auf eine Weise, die erkennen lässt, dass die beiden es als positiv bewerten, wenn diese wieder weggelassen werden kann.

Der Umstand, dass dieses Gespräch unmittelbar unter demonstrierendem Einbezug des VUI geführt wird, zeigt, dass das Gerät klar in seinem maschinellen

Charakter behandelt wird. Die Anrede an das VUI wird vorgeführt (080), so wie die richtige Bedienung eines Geräts erklärt wird. Dabei wird der Summons in die – in erster Ordnung an Till gerichtete – syntaktische Struktur angegliedert, sodass deutlich wird, dass sich der Austausch mit dem VUI in den von Konrad und Till gestalteten *turn-by-turn talk* einbindet. Gleichwohl zeigen sich auf der sprachlichen Oberfläche Parallelen zwischen der Bedienung des Geräts und der Adressierung des Mitbewohners mit entsprechenden Konsequenzen:

Beispiel (5): Red mit Alexa (07-CVR01-10)

```

015 k:  ((lauter Türknall))
016      ((Haushaltsgeräusche 59.0 Sek.))
017 KW: HEY siri:.
018      (0.8)
019 KW: spiele FOUR five seconds von rihanna.
020      (0.8)
021 TW: <<von weit weg rufend> WAS?>
022      (1.7)
023 SI: ich SPIELE jetzt four five sEconds,
024      von rhianna und kanye WEST und paul
          mccartney.
025 KW: isch spiel ÄHH:,
026      [ich äh red mit alexa (.) mit SIri:.]
027 k:  [((Musik setzt ein))                ]
028      ((Musik spielt; gelegentliche
          Haushaltsgeräusche 170 Sek.))

```

Konrad nutzt den Smart Speaker zum Abspielen von Musik (017–019) Nachdem er die Invokation sowie die folgende Stimmeingabe geäußert hat – er produziert diese beiden Schritte hier in zwei getrennten Intonationsphrasen – reagiert darauf Till (021) und zeigt damit seine Interpretation der lautlichen Äußerungen, deren genauen Wortlaut er nicht verstanden hat, als an ihn adressiert an. Eine Klärung dieses Missverständnisses ist notwendig: Konrad benennt, nach einer Pause von 1,7 Sekunden sowie nach der Äußerung des VUI, den eigentlichen Adressat der Äußerung (025–026). Ähnlich aufgebaute Beispiele liegen im Korpus noch zwei weitere Male vor.

Es soll nicht bestritten werden, dass Nutzer*innen sich bei der Kommunikation mit einer Maschine auch auf diese orientieren und dabei auch anzeigen, dass es sich beim adressierten Kommunikationspartner um eine Technologie handelt (vgl. Guzman 2019). Gleichwohl erregt die menschliche Stimme anderer ko-

präsender Sprecher*innen offensichtlich die Aufmerksamkeit. Bei der Bedienung von nicht über die Stimme bedienten Entertainment-Medien hätte das Anschalten und Einstellen von Musik sicherlich nicht die Aufmerksamkeit von Till auf sich gezogen. Der Inhalt der Äußerung kann vermutlich aufgrund der Entfernung der Sprecher nicht verstanden, aber allein der Stimmklang kann hier also als Summons verstanden werden – möglicherweise in Verbindung mit den prosodisch durchaus distinkten Strukturen eines solchen (vgl. Pillet-Shore 2018), den Konrad ja auch produziert, aber an das VUI richtet. Dadurch fallen auf der Oberfläche das Mensch-Mensch-Gespräch und der Mensch-Maschine-Dialog kurzzeitig zusammen und erfordern eine Klärung.

Auffällig ist in dem Beispiel zudem ein weiterer Aspekt, der im Zusammenhang mit der Adressierung von Bedeutung ist: Konrad führt am Schluss der Äußerung eine selbstinitiierte Selbstreparatur durch („mit alexa (äh) mit Siri“, 026), die entsteht, weil Konrad zunächst das VUI als „alexa“ bezeichnet. Dies könnte damit erklärt werden, dass in seinem Elternhaus, in dem er ca. ein halbes Jahr vor Erstellung dieser Aufnahmen noch lebte, ein Smart Speaker von Amazon mit Alexa als VUI genutzt wurde. Auch der gezogene Vergleich zwischen Alexa und Siri im Beispiel (4) („Intercom“) könnte so erklärbarer werden: Auch mit dem VUI „Alexa“ bestehen entsprechende Erfahrungen. Deonymische Tendenzen für das Lexem „Alexa“ können daraus an dieser Stelle entsprechend nicht abgeleitet werden. Bemerkenswert ist, dass Konrad in beiden Fällen nicht mit dem eigentlichen Namen der Hardware referenziert („Echo“ oder „HomePod“), sondern mit dem Namen der VUI-Persona, der hier als Pars pro Toto für das gesamte Gerät gelesen werden kann.

6.1.1.2 Durch das VUI produzierte Anredeformen

Die Pronominalform der zweiten Person ist in den Daten im Korpus ausschließlich in der T-Form dokumentiert. Dies wird seitens des VUI im Rahmen der Ersteinrichtung so eingeführt:

Beispiel (6): Was kannst du? (07-EE-15)¹⁹³

400 SI: lass uns LOSlegen;
 401 (.)
 402 SI: sag.

¹⁹³ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00467_SE_01_T_01_DF_01&cID=c400&wID=&textSize=400&contextSize=12.

403 (0.25)
 404 SI: hey SIri (.) was kAnnst du?
 405 §(0.46) §
 tw: §legt rechte Hand auf rechtes Knie§
 406 KW: +°h +
 kw: +wendet Blick von Handy ab zu SI+
 407 (0.49)
 408 KW: hey SIri (.) was kAnnst du?

Die Ersteinrichtungssituationen zeichnen sich – dies wird auch in weiteren Analysen noch deutlich werden¹⁹⁴ – dadurch aus, dass das VUI an einigen Stellen vorübergehend die Dialogführung übernimmt und adhortative Formen produziert (400).¹⁹⁵ In dieser Phase fordert das VUI den Nutzer dazu auf, eine Frage zu formulieren (in den Eröffnungsskripten von Amazon und Google werden hier ähnliche Wiederholungsaufforderungen geäußert, allerdings mit der Aufforderung, Befehle nachzusprechen). Der Nutzer Konrad folgt dieser Aufforderung im vorliegenden Beispiel (408) und ko-operiert entsprechend mit dem Skript: Es wird ein Einlernen und gemeinsames Beginnen vollzogen. Das VUI schlägt dabei die direkte Anrede in der T-Form vor¹⁹⁶ – bei Zifonun/Hoffmann/Strecker (1997: 317) auch als „Balanceform“ in Abgrenzung von der „Distanzform“ bezeichnet. Das VUI verwendet diese auch selbst – in keinem der im Korpus erfassten Aufnahmen findet sich die V-Form, weder seitens der Nutzer*innen noch seitens der VUIs. Das mag überraschen, ist doch generell bei einem „unbekannten, unvertrauten, sozial distanten“ (Zifonun/Hoffmann/Strecker 1997: 317) Gegenüber die Distanzform üblich, auch wenn hier zweifelsohne Sprachwandelprozesse zu Verschiebungen in der Verwendung der T- und V-Form geführt haben. Mit Blick auf die Domestizierung des Geräts ist die Stelle von besonderer Bedeutung, weil der Smart Speaker dadurch sprachlich eine Nähe und Informalität vermittelt: Sowohl der vergemeinschaftende Charakter des Adhortativ als auch die T-Form stellen ein gemeinsames Projekt zwischen VUI und Anwender*innen her. Dieses ‚Einlernen‘ kann als wesentlicher Schritt im Rahmen der Domestizierung von VUIs verstanden werden – insbesondere im Rahmen dessen, was bei Silverstone/Hirsch/Morley (1992: 21–22) als „Incorporation“, d. h. die Eingliederung in die Praxis, bezeichnet wird. Die Übernahmen entsprechender vorgeschlagener sprachlicher Formen sind grundlegende Schritte, um dieses wechselseitige Verständnis, das

¹⁹⁴ Siehe Kap. 6.2.2.1.

¹⁹⁵ Siehe auch Beispiel (52) („Hallo, ich bin dein Google Assistant“).

¹⁹⁶ Zu formellen und informellen Anreden siehe Brown/Gilman (1968) und Besch (1996).

auch als eine Form von Ko-Operation im Sinne Goodwins verstanden werden kann, herzustellen. Die Frage der Anredeform ist dabei für die Beziehungsgestaltung von entscheidender Bedeutung: Anreden und Grüße sind unter Konversationspartnern wichtige erste Merkmale zur Aushandlung der Beziehung und können mit Goffman (1973: 64) als *presentational rituals* bezeichnet werden, die Aussagen zur sozialen und hierarchischen Beziehung der Kommunikationsbeteiligten enthalten und sich sprachlich in Form von „grammatical encodings of relative social status“ (Brown/Levinson 1987: 179) darstellen. Die vom VUI vorgeschlagene T-Form in Verbindung mit der Ratifizierung und Übernahme dessen im Rahmen der Wiederholung (408) sind also ein ko-operativ verfertigter Schritt zur ‚Eingliederung‘ des Geräts in den Haushalt und in die vollzogenen sprachlichen Praktiken. Äußerungen, in denen durch die Hörerdeixis (Zifonun/Hoffmann/Strecker 1997: 317–318) die kommunikative Rolle des Adressaten zum Ausdruck gebracht wird, greifen also in den vorliegenden Daten erstens konsequent auf die T-Form zurück. Dies weicht von den Chatbots ab, die Lotze (2021) im Hinblick auf die Adressierungen erforscht; sie stellt bei drei von vier untersuchten Systemen die Nutzung der Sie-Form fest, allerdings nutzen alle Anwender*innen konsequent die Du-Form (vgl. Lotze 2021: 320).

Im Hinblick auf die pronominalen Adressierungen ist deutlich geworden, dass sich die VUIs der verschiedenen Hersteller der Du-Form bedienen. In den vorliegenden Daten sind Äußerungen der VUIs, die eine Anredeform enthalten, tendenziell kein Bestandteil der Smart Speaker-Funktionen im engeren Sinne, sondern treten dann auf, wenn das VUI durch die Verwendung der Hörerdeixis dazu auffordert, geäußerte Stimmeingaben zu korrigieren, in denen einzelne Bestandteile nicht erkennbar oder anwendbar waren, dazu einlädt, neue Funktionen auszuprobieren, oder die Äußerung anderweitig expandiert, so wie im folgenden Beispiel:

Beispiel (7): Schneegestöber (01-CVR01-01)¹⁹⁷

071 [(16.7)]
072 k: [((vereinzelte Hintergrundgeräusche))]
073 LF: aLEXa.
074 (1.2)
075 LF: wie ist das wetter DRAUßen?
076 (1.2)
077 AL: in amelsbüren beträgt die temperatur

197 Für eine Betrachtung desselben Beispiels unter dem Aspekt der Sequenzorganisation siehe Kap. 6.1.2.1.3.

aktuell minus ZWEI grad celsius bei
 bewölktem himmel,
 078 (0.5)
 079 AL: der wetterbericht für heute nacht sagt
 schneegestöber vorAUS mit einer
 tiefsttemperatur von minus vier grad;
 080 ich hoffe (.) du hattest [einen schönen
 SONN.]
 081 LF: [aLEXa (.) stop-
]
 082 (166.4)

Die nach der Invokation vollzogene Eingabe-Ausgabe-Sequenz ist nach der Vorstellung des Wetterberichts (077–079) abgeschlossen. Das VUI expandiert hier allerdings und bringt eine „Hoffnung“ zum Ausdruck,¹⁹⁸ die es als initiativen und empathischen ‚Gesprächsbeteiligten‘ erscheinen lassen könnte. Lukas bricht allerdings hier die Äußerungen des VUI unmittelbar nach der Du-Anrede durch das VUI ab (080), das Höflichkeitsritual wurde daher nicht abschließend produziert und der Vorgang legt den Schluss nahe, dass Lukas nicht in einen ‚höflichen‘ Austausch mit dem VUI treten und von diesem entsprechend auch nicht mit über die Funktion hinaus expandierten Äußerungen adressiert werden möchte. Dafür spricht auch ein sehr ähnlich aufgebautes Beispiel mit dem gleichen Sprecher, in dem das VUI eine Funktionserweiterung vorstellt bzw. dazu ansetzt:

Beispiel (8): Wenn du mehr Zeit benötigst ... (01-CVR01-08)¹⁹⁹

009 [(4.9)]
 010 k: [((vereinzelte leise
 Hintergrundgeräusche))]
 011 LF: aLEXa?
 012 (0.9)
 013 LF: stell den timer auf ZWANZIG minuten.
 014 (1.0)

198 Eine Reflexion darüber, ob das VUI tatsächlich etwas „hoffen“ kann und inwieweit es sich hierbei um eine formelhafte Höflichkeitsgeste handelt, sowie über die Intentionen der Hersteller, die mit der Einspeisung solcher Äußerungen in die VUIs verbunden sind, unterbleibt an dieser Stelle; siehe dazu ausführlicher Kap. 6.1.2.1.3. Die Diskussion wird in Kap. 7 wiederaufgegriffen.

199 Für eine Diskussion dieses Beispiels mit einem Fokus auf die Sequenzorganisation siehe Kap. 6.1.2.1.3 sowie unter dem Gesichtspunkt des Turn-Takings siehe Kap. 6.1.3.2.

015 AL: zwAnzig minUten (.) ab JETZT.
 016 übrigens (.) wenn du MEHR zeit benötigst==
 017 [=Sage einfa-]
 018 LF: [alex a STOP.]
 019 [(70.5)]

Vergleichbar zu der Situation in Beispiel (7) („Schneegestöber“) äußert das VUI eine Expansion über die eigentliche Funktion hinaus (016); die Sequenz zwischen der Stimmeingabe (Timer-Stellen, 013) und der darauf folgenden Stimmausgabe (015) war bereits funktional abgeschlossen. Dies markiert auch die turninitiale und prosodisch durch eine Mikropause abgesetzte platzierte Produktion von „übrigens“ (016), das hier mit Egbert (2003: 199) eine „graduelle Themenentwicklung“ einleitet; syntaktisch wird diese Überleitungsfunktion durch die Positionierung im Vor-Vorfeld unterstützt (vgl. Auer 1997: 59). Die so eingeleitete Expansion wird jedoch kurz nach der im anschließend projizierten Konditionalsatz platzierten Du-Anrede („wenn du MEHR zeit benötigst“, 016) abgebrochen (018). Auch hier lässt sich schließen, dass der Anwender die Einladung zur Erweiterung der Funktion nicht hören, sondern den Austausch auf die Dyade beschränken wollte.

Auch die im folgenden Beispiel produzierte Du-Anrede aus demselben Haushalt ist nicht Bestandteil der eigentlichen Funktion, sondern ein Hinweis auf eine Störung:

Beispiel (9): Du hast keine Benachrichtigungen (01-CVR01-03)

037 LF: h°
 038 aLEXa?
 039 (1.4)
 040 LF: wie ist der TImerstand?
 041 (1.4)
 042 AL: du hast KEIne [benachrichtigungen-]
 043 [(((Signalton setzt ein))]]
 044 (0.2)
 045 AL: (((Signalton))
 046 (0.5)
 047 LF: [alex a STOP-]
 048 AL: [(((Signalton))]
 049 (4.7)
 050 LF: <<leise sprechend> K0misches ding ey->
 051 °h
 052 (2.8)

Gefragt nach dem Timerstand eines laufenden Timers (040) antwortet das VUI, dass keine Benachrichtigungen vorliegen, was jedoch nicht der gewünschten Information entspricht, was der Anwender Lukas auch zum Ausdruck bringt (050) und dabei den Verstärker „ey“ aus dem jugendsprachlichen Sprachrepertoire nutzt (vgl. Neuland 2008: 137).

Insgesamt zeigt sich die Tendenz, dass Du-Ansprachen auf Expansionen oder Fehlfunktionen oder Fehlverständnis hinweisen können, während sie als Teil des ‚üblichen‘ Funktionsspektrums – gerade in den häufiger genutzten Funktionen wie Smart Home-Steuerung, Musikhören usw. – nicht auftreten: In 15 im Korpus dokumentierten Fällen,²⁰⁰ in denen das VUI eine Hörerdeixis produziert, ist nur ein Fall als Bestandteil des nutzer*innenseitig vorgesehenen Dialogs zu verstehen:

Beispiel (10): Du hörst gerade (07-CVR01-40)

013 KW: hey SIri?
 014 (0.8)
 015 KW: von WEM ist des lie:d?
 016 (1.4)
 017 SI: du hörst gerade <<deutsch ausgesprochen>
 BRighter> than the sun von colbie caillat-

Die übrigen Fälle sind Expansionen, Ausdruck von Fehlfunktionen oder ‚Rückfragen‘ wie im folgenden Beispiel.²⁰¹

Beispiel (11): Lautstärke (08-CVR01-08)

106 LS: oKAY google,
 107 [fünf lauten (.) STÄRken (.) leiser.]
 108 k: [((Radio pausiert))]
 109 RL: (xxxx xxx)
 110 (2.0)
 111 LS: [((stöhnt))]
 112 GA: [entSCHULdigung.]

200 Die Suche wurde dabei vom Teilkorpus der dyadischen Dialoge auf das Gesamtkorpus ausgeweitet, weil es sich zwar um ein Problem der dyadischen Dialogorganisation zwischen Mensch und VUI handelt, das sich allerdings auch in Dialogen mit mehreren beteiligten Anwender*innen zeigt.

201 Eine ausführliche Analyse des Beispiels unter dem Aspekt der Beteiligtenkonstellation findet sich in Kap. 6.2.3.

- 113 die HÖCHste lautstärke ist zehn.
 114 auf welche stufe möchtest du die
 LAUTstärke stellen?
 115 (0.2)
 116 RL: DREI.

Das VUI greift also v. a. dann auf eine Hörerdeixis zurück, wenn etwas außerhalb der Nutzung des vorgesehenen Dialogablaufs im engeren Sinne relevant wird, um die Nutzer*innen darauf aufmerksam zu machen, notwendige Informationen einzuholen oder sich in einer bestimmten Weise zu ‚präsentieren‘. Das Nutzungsverhalten des Nutzers Lukas aus Beispiel (8) („Wenn du mehr Zeit benötigst ...“) zeigt zudem, dass es zwar funktional möglich ist, die Nutzung des Geräts fortzusetzen, ohne auf die Expansionen einzugehen, dass diese jedoch eine gewisse Intrusivität aufzuweisen scheinen.

Onymische Anredeformen durch das VUI können ebenfalls auftreten, einer der im Korpus belegten Fälle stammt aus Haushalt 8 (Google Home):

Beispiel (12): Guten Morgen (08-CVR01-08)

- 015 RL: oKAY google,
 016 (0.5)
 017 k: ((Räuspern))
 018 RL: guten MORgen,
 019 (1.8)
 020 GA: guten MORgen.
 021 [ROBin.]
 022 RL: [((unverständlich 1.0 Sek))]
 023 (unverständlich)) cool der weiß sogar
 wer ich BIN.
 024 k: ((ausgelassen, bis Z. 33))
 025 (0.7)
 026 RL: wer BIN ich?
 027 (1.2)
 028 GA: dein NAME ist RObin.
 029 (1.0)
 030 RL: ((lacht)) der kann auch verstehen (.) wenn
 ich sage (.) wo ist MEIN handy,

Am Morgen nach der ersten Inbetriebnahme des Geräts, die am Abend zuvor stattgefunden hatte, begrüßt der Nutzer Robin (RL) seine Mitbewohnerin und sie

tauschen sich kurz über die Tagesplanungen aus (001–014, ausgelassen).²⁰² Anschließend erfolgt nach einer Pause die Adressierung des VUI, das hier ebenfalls begrüßt wird.²⁰³ Dies kann im Kontext des fortgesetzten Testens und Ausprobierens gesehen werden – eine Routine in der Nutzung hat sich zu diesem Zeitpunkt (nach einem Tag) noch nicht eingestellt. Auf die Grußformel antwortet das VUI mit einem Gegengruß und einer onymischen Anrede, wie sie in Begrüßungen durchaus üblich ist (vgl. Hartung 2001: 1350), von der sich Robin beeindruckt zeigt („cool der weiß sogar wer ich BIN.“, 023). Er äußert dies, wie Habscheid (2022) an einem ähnlichen Beispiel herleitet, in einem „Meta-Interaktionsraum“, in dem die Äußerungen nicht an das VUI, sondern an ein anderes Gegenüber gerichtet sind.²⁰⁴ Später knüpft Robin in einem neuen Befehl an diese Identifikationsleistung (034) an und demonstriert dies noch einmal. So unmittelbar aufgefordert produziert das VUI eine onymische Anrede, die sich durch die erbetene Identifikation erklärt, die im Kontext des (gemeinsamen) Ausprobierens zu verstehen ist. Dass es Robin hier um eine Gesamtschau von Identifikationsfunktionen geht, wird auch durch die Äußerung von Robin zur Handysuche (038) deutlich, bei dem die Identifikation durch den Fokussakzent auf dem Personalpronomen noch einmal unterstrichen wird.

Abgesehen von den Eröffnungspassagen, in denen das VUI die Dialogführung zu übernehmen scheint (wie sich durch sequenzeröffnende Äußerungen, Adhortative und Imperative zeigt), tritt es in den im Korpus dokumentierten Daten nur ein Mal auf, dass das VUI den Austausch initiiert. Der einzige Fall zeigt sich in folgendem Beispiel:²⁰⁵

Beispiel (13): Lieferstatusupdate (04-CVR02-17)

001 k: ((Hintergrundgeräusche bis Z. 22))
 002 AL: ((Zweigliedriger Ton mit steigender
 Tonhöhenbewegung))
 003 (5.5)
 004 SR: aLEXa?
 005 (0.5)

202 Eine ausführlichere Analyse dieses Beispiels unter Einbezug der daran beteiligten Nutzerin Lara (LS) findet sich in Kap. 6.2.3.

203 Wir kommen auch auf den Umstand, dass Robin hier einen Gruß an das VUI richtet, noch einmal zurück, siehe Kap. 6.2.3.

204 Ausführlicher dazu siehe Kap. 6.2.

205 Dieses Beispiel entstammt nicht aus einem der drei Fokushaushalte, sondern aus Haushalt 4, siehe Kap. 5. Es wird hier präsentiert, weil es auch unter Einbezug der übrigen Haushalte 2, 3, 4, 5 und 6 die einzige Dokumentation einer Initiierung durch das VUI darstellt.

006 SR: was ist meine beNACHrichtigung;
 007 k: (1.9)
 008 AL: eine NEUe benAchrichtigung;
 009 (1.1)
 010 AL: von AMAzon shopping.
 011 (0.9)
 012 AL: eine LIEferung,
 013 (.)
 014 die HERDabdeckplatte beInhaltet,
 015 (.)
 016 soll heute ANkommen.
 017 (1.1)
 018 AL: war dieses LIEferstatusupdate hILfreich?
 019 (0.7)
 020 SR: ja;
 021 DANke.
 022 (2.1)
 023 AL: vIelen dank für dein FEEDback.
 024 k: ((Hintergrundgeräusche und
 Geschirrkloppern, ca. 2.5 Min.))

Das VUI spielt zum Beginn des hier gezeigten Ausschnitts ein Geräusch ab: Es produziert einen zweigliedrigen, melodischen Ton (002), der sich aus der Interface-Forschung heraus mit Blattner/Sumikawa/Greenberg (1989) als „Earcon“ konzeptualisieren lässt (siehe auch Schoon/Volmar 2012: 12). Earcons werden von Blattner/Sumikawa/Greenberg (1989: 11) als „audio messages used in the user-computer interface to provide information and feedback to the user about computer entities“ beschrieben. Sie sind semiotisch gesehen ikonische Zeichen mit einem vergleichsweise niedrigen Ikonizitätsgrad und lassen sich als „musikalisches Motiv“ klassifizieren (vgl. Oswald 2012: 260). Anders als die von Gaver (1986) beschriebenen „Auditory Icons“ basieren sie nicht auf einer Ähnlichkeit zu natürlich vorkommenden, materialen Vorgängen (z. B. Zerknüllen von Papier als akustisches Zeichen für die Verschiebung einer Datei in den Papierkorb), sondern arbeiten mit melodischen Mustern und Kompositionen. Die steigende Tonhöhenbewegung kann als eröffnender und kreativer Charakter dieses Geräuschs verstanden werden (vgl. Blattner/Sumikawa/Greenberg 1989: 30–31), der die Aufmerksamkeit auf sich ziehen und letztlich eine Realisationsform von „Summons“ bilden kann. Auf diesen Summons reagiert Sam nach ca. 5,5 Sekunden (004), allerdings muss er, um auf diesen Summons überhaupt antworten zu können, das Gerät erneut in den Listening-Modus bringen und entsprechend eine Art ‚Gegen-Summons‘ äußern. Auch die dann geäu-

Berte Frage („was ist meine beNACHrichtigung;“, 006) ist noch Bestandteil dieses ‚Gegen-Summons‘, denn dass die von Sam geäußerte Sequenz sich nun überhaupt auf das vom VUI produzierte Summons bezieht, muss zunächst präzisiert werden. Es wird diese Präzisierung durch das VUI anschließend ausdrücklich ratifiziert und nochmals klargestellt, dass nun die entsprechende Benachrichtigung vorgetragen wird (008–010). Aus der Summons-Answer-Sequenz wird also in diesem Fall eine Summons-Summons-Answer-Ratifizierung-Sequenz.

6.1.1.3 Zwischenfazit Anredeformen

Die untersuchten Anredeformen zeigen eine starke Orientierung an bisher beschriebenen Praktiken zur Realisierung von Summons-Answer-Sequenzen, die der Herstellung eines Kanals für den Dialog und einer gemeinsamen Aufmerksamkeitsfokussierung dienen. Dabei hatten schon frühere Beschreibungen von Summons deren eher technischen Charakter im Blick (z. B. Telefon- oder Türklingeln), der durch die Invokation ebenfalls hervortritt, prosodisch aber noch weitergehend moduliert werden kann. Die Summons-Answer-Sequenz wird von den VUIs bei der Ersteinrichtung als Praktik präsentiert, um den VUI-Dialog zu beginnen. Die Nutzer*innen werden sodann beim Einlernen begleitet. Das VUI von Apple zeigt durch ein Rezeptionssignal an, dass die Invokation erkannt wurde, was bei anderen Modellen visuell in Form des leuchtenden Lichtrings bzw. der leuchtenden Punkte signalisiert wird. In einem Fall ist eine Dialoginitiiierung durch das VUI dokumentiert, dabei kommt ein Geräusch als Summons zum Einsatz (ähnlich wie ein Telefonklingeln), darauf folgt ein mehrteiliger Dialog. So orientieren sich Anwender*innen und VUI einerseits an bekannten Praktiken für die Organisation einer Gesprächseröffnung.

In Formen der expliziten Adressierung im Austausch mit dem VUI – in den detailliert betrachteten Summons-Answer-Sequenzen ebenso wie in Anredeformen der Nutzer*innen – manifestieren sich andererseits aber auch Brüche des Gesprächsteiligtenstatus von VUIs: Die Notwendigkeit, den Summons für jede Stimmeingabe bzw. jede neue Funktion ebenfalls erneut zu produzieren, verweist darauf, dass ein kontinuierlicher Gesprächsvollzug nicht gegeben ist. Außerdem scheinen Anredeformen, die durch das VUI produziert werden, tendenziell ein Hinweis darauf zu sein, dass der Dialog von der ‚typischen‘ zweiteiligen Struktur (Stimmeingabe – Stimmausgabe) abweicht, auf die im nächsten Kapitel näher einzugehen ist. Die Nutzer*innen brechen solche Expansionen in den analysierten Beispielen ab.

Zusammengenommen sind in allen betrachteten Bereichen konversationelle Verfahren beobachtbar, die für Mensch-Mensch-Interaktionen ausführlich beschrieben wurden; diese werden jedoch aufgrund ihrer Technizität teilweise neu funktionalisiert (z. B. der onymische Summons vor jedem Dialogbeginn), aktiv

konfiguriert (z. B. durch die Einführung der Du-Anrede in der Ersteinrichtung) oder dienen zur Bearbeitung und Ausweitung von kommunikativen Projekten außerhalb des konversationell erwartbaren Rahmens.

6.1.2 Sequenzialität und Sequenzorganisation

Es ist eine der zentralen Thesen und Befunde jahrzehntelanger konversationsanalytischer Forschung, dass Sprache im konversationellen Gebrauch sequenziell organisiert ist, und es war ihre Leistung, diesen Umstand in den Mittelpunkt der Analysen zu rücken (vgl. Stivers 2012: 191). Das „Sequence-organizational‘ Problem“ (Schegloff 2006: 73) ist eines der bereits erwähnten sechs „generic organizations of practice“, die als Organisationsprinzipien überall da auftreten, wo sprachliche Interaktionen vollzogen werden. Konstitutiv für das Problem ist demnach die Frage nach der Kohärenz aufeinanderfolgender Redeeinheiten und entsprechend entstehender Handlungsverläufe und der „Nature of that Coherence“ (Schegloff 2006: 73). Schegloff (1990: 51) bezeichnet Sequenzen als ‚Klumpen‘ („clumps“), in denen Konversationen sich vollziehen und gestalten.

In Anlehnung an Schegloff (2007: 2) und darauf Bezug nehmende Darstellungen (Imo/Lanwer 2019: 171; Birkner 2020b: 237) sollen nachfolgend zwei Aspekte unterschieden werden: Erstens die Konsequenzen einer für die mündliche Äußerungsproduktion grundlegenden Eigenschaft, der Zeitlichkeit. Durch ihre Zeitlichkeit tragen mündliche Äußerungen sowohl ein retraktives wie auch ein prospektives Potenzial in sich (vgl. Deppermann 2008: 49). Gesprochensprachliche Einheiten sind nacheinander positioniert und stehen in Verbindung zueinander. Eine einzelne Äußerung entfaltet ihre Bedeutung also aus dem Potenzial der relativen Positionierung zu anderen Äußerungen – nur durch sie können Beteiligte den Äußerungen Sinn zuschreiben. Dieser Aspekt, auf den ich im Folgenden mit *Sequenzialität* referieren will, umfasst also sowohl den Vollzug von Sprecher*innenwechseln, Äußerungspositionierungen in längeren Gesprächen, den Ablauf von Handlungen und andere Bereiche gesprochensprachlicher Interaktion (vgl. Imo/Lanwer 2019: 171).

Zweitens wird in diesem Kapitel der „Aufbau von Abschnitten, die als Handlungseinheit zu sehen sind“ (Birkner 2020b: 237), betrachtet. Die konversationsanalytische Forschung hat ausführlich untersucht, wie kleinere Handlungseinheiten – z. B. Begrüßungen, Einladungen, Gesprächsbeendigungen (vgl. Schegloff/Sacks 1973; Sacks/Schegloff/Jefferson 1974; Sacks 1995: 175) – und größere Handlungseinheiten – etwa Narrationen (vgl. Quasthoff 1980a), Vorwürfe (vgl. Günthner 1999b), Bewerbungsgespräche (vgl. Birkner 2001) oder Bewertungsinteraktionen (vgl. Hrncal 2020) – organisiert sind (vgl. Schegloff 2007: 2). Auf diese

handlungsorganisatorischen Aspekte soll mit dem Begriff der *Sequenzorganisation* Bezug genommen werden. Unter den Aspekt der Sequenzorganisation fällt auch die genauere Untersuchung von „Paarsequenzen“, die als ein zentraler Baustein der Sequenzorganisation bereits früh entdeckt wurden (vgl. Schegloff/Sacks 1973). Im Folgenden sollen zunächst die Sequenzorganisation in VUI-Dialogen zwischen einem*r Anwender*in und einem VUI beleuchtet werden, um anschließend allgemeinere Merkmale der Sequenzialität in solchen VUI-Dialogen zu explorieren.

6.1.2.1 Sequenzorganisation in VUI-Dialogen

Paarsequenzen stellen eine „Basiseinheit“ (Birkner 2020b: 243) zwischenmenschlicher Interaktion dar; sie gehören zum „Kerninventar der Konversationsanalyse“ (Imo/Lanwer 2019: 177). Paarsequenzen in der menschlichen Interaktion sind konversationsanalytisch ausführlich untersucht und in ihren Eigenschaften beschrieben worden (vgl. Stivers 2012). Als Basiseigenschaften lassen sich festhalten, dass Paarsequenzen aus mindestens zwei Paarteilen bestehen, die von verschiedenen Sprecher*innen produziert werden und zeitlich prototypischerweise ‚benachbart‘ zueinander stehen (wobei diese Nachbarschaft nicht immer gegeben sein muss und etwa durch Einschubsequenzen unterbrochen werden kann, vgl. Schegloff 2007: 97–114):

The simplest and minimal form of a sequence is two turns long: the first *initiating* [Herv. i. O.] some kind of action trajectory – such as requesting, complaining, announcing, and the like; the second responding to that action in either a compliant or aligning way [...] or in a misaligning or noncompliant way [...]. (Schegloff 2006: 74–75)

Der erste Teil macht somit einen bestimmten zweiten Paarteil konditionell relevant, d. h. für beide Teilnehmer*innen erwartbar (vgl. Schegloff 1968; Auer 2020: 107). Es ergeben sich daraus ‚typische‘ paarsequenzielle Muster zum Vollzug von Handlungen, etwa Grüßen (Gruß und Gegengruß), Fragen (Frage und Antwort) oder Einladen (Einladung und Annahme/Ablehnung).²⁰⁶ Handlungsverläufe müssen allerdings nicht immer ‚paarig‘ sein. Sie können auch aus mehreren solcher Paarsequenzen zusammengesetzt oder mehrteilig sein und im dritten Zug etwa eine Ratifizierung oder Evaluation beinhalten. Außerdem können sogar viele verschiedene Handlungsschritte notwendig sein, etwa um eine Geschichte zu erzählen. Solche größeren Einheiten können Paarsequenzen als Bestandteile beinhalten (vgl. Quasthoff 2001: 1297), die als Bausteine dienen, müssen dies aber nicht. VUI-Dialoge wei-

²⁰⁶ Für eine tabellarische Übersicht über Beispiele typischer Paarsequenzen siehe Stivers (2012: 192).

sen einen paarsequenziellen Kern auf, der jedoch sowohl durch Prä- wie auch Postsequenzen und Einschübe erweitert werden kann, wie nachfolgend zu zeigen ist.

6.1.2.1.1 Zur Basis-Sequenzstruktur des VUI-Dialogs

Dyadische VUI-Dialoge sind hochgradig entlang der mit ihnen zu vollziehenden Handlungseinheiten aufgebaut. Ihre Struktur folgt im Prinzip dem Muster einer Paarsequenz, weist dabei allerdings einige Besonderheiten auf, die an folgendem Beispiel demonstriert werden sollen:

Beispiel (14): Doja Cat (07-CVR01-58)

033 KW: hey SIri spiele woman von doja cat.
 034 (0.6)
 035 [(1.7)]
 036 k: [((Musik wird sehr leise))]
 037 SI: hier ist WOman von <<deutsche Aussprache>
 doja> CAT.
 038 (1.4)
 039 k: ((neues Lied "Woman" von Doja Cat spielt;
 KW singt phasenweise mit; vereinzelt
 Haushaltsgeräusche und Schniefen 170.0
 Sek))

Der Anwender Konrad (KW) präsentiert in diesem Ausschnitt die Intention, Musik über den Smart Speaker abzuspielen. Die vollzogene Handlung ist also die Instruktion zu einer Musikwiedergabe – an deren praktischer Verfertigung partizipiert auch der Smart Speaker. Als Basisstruktur für die dafür notwendige Dialogsequenz scheint sich auf den ersten Blick eine einfache Abfolge zu ergeben: ‚Stimmeingabe‘ – ‚Stimmausgabe‘. Zunächst produziert Konrad die Stimmeingabe – sie ist im gegebenen Beispiel besonders deutlich als Befehl anhand des Imperativs in Verbspitzenstellung zu erkennen. Dadurch wird auf der sprachlich-sequenziellen Oberfläche ein Folgezug des VUI konditionell relevant gesetzt, d. h. für den Anwender erwartbar: Eine Antwort bzw. die Umsetzung der imperativisch formulierten Stimmeingabe werden sequenziell projiziert.

Das Grundgerüst aus ‚Stimmeingabe – Stimmausgabe‘ ist jedoch bei detaillierter Betrachtung des Ausschnitts um zwei Aspekte zu erweitern. Vor der Äußerung der eigentlichen Stimmeingabe äußert Konrad eine Invokation, die zusammen mit der Reaktion des VUI als eine Art Präsequenz betrachtet werden kann (vgl. Gülich/Mondada 2008: 56): ‚Invokation – Listening-Modus‘. Dabei ist der zweite Teil

dieser Präsequenz – der tatsächliche Statuswechsel des Smart Speakers in den Listening-Modus – nicht akustisch dokumentiert. Die untersuchten Smart-Speaker-Modelle wenden verschiedene visuelle Verfahren an, um den Listening-Modus anzuzeigen. So verfärbt sich beim Amazon Echo Dot etwa der Lichtring blau, beim Google Home Mini werden vier Punkte auf der Oberfläche sichtbar.²⁰⁷ Ob diese visuellen Signale von den Nutzer*innen beachtet werden, war mit der gewählten Erhebungsmethode nicht zu erfassen (die CVR-Daten lassen darauf keine Rückschlüsse zu, lediglich die Ersteinrichtungsdaten können diesbezüglich herangezogen werden).²⁰⁸ Die Nutzer*innen produzieren die Invokation – wie auch im vorliegenden Beispiel (vgl. 033) – nicht immer als intonatorisch getrennte Einheiten, sondern können diese auch intonatorisch zusammenziehen, sodass auf der sprachlichen Oberfläche der Vollzug der Präsequenz nicht ablesbar ist – retrospektiv kann ihr Gelingen allerdings sowohl für die Nutzer*innen wie auch analytisch festgestellt werden, weil sonst keine Bezugnahme des VUI erfolgen könnte (036). Die Invokationspräsequenz ist für VUI-Dialoge zentral, weil sie – abgesehen von Ausnahmen wie VUI-Initiierungen oder Expansionen (s. u.) – *immer* zuerst stattzufinden hat, bevor eine Stimmeingabe überhaupt zielführend geäußert werden kann. Das Gelingen einer Eingabe ist also abhängig vom Gelingen der vorherigen Präsequenz.

Außerdem ist der zweite Teil der Eingabe-Ausgabe-Paarsequenz zweiteilig: Er besteht aus einer verbalen Äußerung und der Realisierung der darin enthaltenen Aufforderung. Die Äußerung des VUI hat dabei die kommunikative Funktion eines ‚Scharniers‘: Einerseits kündigt es prospektiv die Wiedergabe des Titels an, unter Verwendung der Lokaldeixis („hier“, 036), die die Titelwiedergabe als Leistung des Geräts projiziert. Zugleich wiederholt es retraktiv Teile der Äußerung von Konrad (Interpretin und Titel des gewünschten Liedes), sodass im Falle von Missverständnissen die Möglichkeit zur Beeinflussung noch vor Start der Musikwiedergabe bestünde. Die Wiedergabe ist hier also zweiter Teil eines zweiten Paarteils einer Eingabe-Antwort-Sequenz. Der erste Bestandteil der Sequenzstruktur ist allerdings nicht in allen Fällen realisiert und nicht obligatorisch, wie das folgende Beispiel zeigt:

Beispiel (15): Nächster Titel (07-CVR01-52)

016 KW: hey siri nächster Titel-
 017 k: ((Lied endet und neues Lied setzte ein))

In diesem Beispiel wird sprachlich nur der erste Teil einer Paarsequenz realisiert: die Aufforderung, den nächsten Titel abzuspielen (016). Eine verbale Erwiderung des

²⁰⁷ Siehe dazu Abb. 3 und 12 (Kap. 3.3.1 bzw. 3.3.2).

²⁰⁸ Siehe dazu Kap. 6.2.3.8.

VUI wird nicht verbalisiert, die Stimmeingabe wird allerdings offensichtlich prozessiert und die Aufforderung umgesetzt – der nächste Titel wird abgespielt (017). In anderen Fällen hingegen kann der zweite Teil der Paarsequenz keine Realisierung über die Antwort hinaus beinhalten, weil das Liefern der Antwort die praktische Umsetzung der Eingabe ist. Dies trifft insbesondere auf Eingaben vom Typ einer Frage zu:

Beispiel (16): Wie viele Tage (07-CVR01-38)

024 KW: hey SI:ri?
 025 (1.1)
 026 KW: wie viele Tage noch bis zum: ersten januar zweitau:
 senddreißig-
 027 (1.3)
 028 SI: bis daHIN dauert es noch (.)
 zweitausendneunhundertsechsunachtzig
 tage;

Nach dem Summons (024) stellt Konrad an das VUI eine Ergänzungsfrage (026), die auf eine ‚Menge‘ gerichtet ist und eine genaue Angabe erfordert (die Anzahl der Tage bis zu einem bestimmten Datum). Das VUI Siri beantwortet die Frage (028) nach einer kurzen Pause. Es folgt keine weitere Aktion, die auch seitens Konrad nicht angefordert wurde. Im übrigen Verlauf der Aufnahme ist es still und es ist daher nicht zu klären, wie sich die Abfrage in den Vollzug der Praxis einfügt oder wie die Information verarbeitet wird – sie hat jedenfalls keine Auswirkungen auf nachfolgende Dialoge mit dem VUI.

Die Basis-Sequenzstruktur für von Nutzer*innen initiierte VUI-Dialoge²⁰⁹ folgt also zwei verschiedenen Mustern, die sich je nach Art der zu prozessierenden Stimmeingabe und der damit verbundenen Ziele unterscheiden:

- (A) Invokation – (Listening-Modus –) Stimmeingabe – (Stimmausgabe/Scharnier –) praktische Umsetzung.
- (B) Invokation – (Listening-Modus –) Stimmeingabe – Stimmausgabe.

Dabei ist das Muster (A) v. a. bei imperativisch formulierten Eingaben feststellbar, Muster (B) hingegen bei Fragen, in denen die Beantwortung einer Frage im Zentrum des Dialogs steht. Diese Muster scheinen im dyadischen Korpus eine hohe Stabilität aufzuweisen. Abweichungen entstehen, wenn Einschübe oder Expansio-

²⁰⁹ In einem Fall ist im Korpus auch eine Initiierung des VUI-Dialogs durch den Smart Speaker dokumentiert, siehe dazu Beispiel (13) („Lieferstatusupdate“) in Kap. 6.1.1.2.

nen auftreten, auf die im Folgekapitel der Fokus gerichtet werden soll. Der Befund steht im Einklang mit bisherigen Arbeiten zum Austausch zwischen Mensch und Maschine: Arminen (2005: 203) erarbeitet mit Bezug auf Suchman (1987) ein Schema für eine „Basic Sequence“, die „for any human-computer interaction“ generalisiert werden könne und wie folgt zusammengefasst wird: „any single user’s action which causes a change in the system state emerges on the interface as a shift in the display“. Nutzer*innenseitig vollzieht sich also eine Abfolge von Optionen-Auswahl, Entscheidung/Handlung und Evaluation, während sich für den Computer eine Abfolge von „DISPLAY 1“, durchgeführter Handlung und „DISPLAY 2“ ergebe (Arminen 2005: 203). Das entsprechend dreizügige Modell dieser Sequenzmodellierung bezieht sich jedoch erstens auf visuelle Interfaces und hat zweitens nicht den sprachlichen Vollzug des Austauschs von Mensch und Maschine im Blick, sondern relativ abstrakte Handlungsschritte, die allein nicht ausreichen, um VUI-Dialoge zu beschreiben. Das hier postulierte Schema für die Basis-Sequenzstruktur von VUI-Dialogen (mit den beiden Subtypen A und B) bezieht die Prozessierung über den akustischen Kanal mit ihren Konsequenzen und die sprachliche Realisation mit ein und versteht sich insofern nicht als Gegenentwurf, sondern als komplementär zum Muster nach Arminen (2005).

6.1.2.1.2 Einschübe

Das aufgezeigte Basismuster der im Korpus dokumentierten VUI-Dialoge ist v. a. zu ergänzen, wenn (aus unterschiedlichen Gründen) Eingaben nicht in der erwarteten Weise prozessiert werden können. Dies tritt v. a. dann auf, wenn das VUI noch auf weitere Angaben wartet, etwa weil es, wie im folgenden Beispiel, einen Teil der Äußerungen nicht protokolliert hat:

Beispiel (17): Timer vier Minuten (01-CVR01-11)

```

072 LF: [aLEXa?           ]
073 k:  [((TV wird pausiert)) ]
074      (0.9)
075 LF: TImeR (0.3) vier minuten.=
076 AL: =für WIE lange?
077      (1.3)
078 LF: <<artikulierte sprechend, h> h° (-) alexa:
      (.) tImeR vier miNU:ten.>
079      (1.0)
080 AL: vier minuten (-) ab JETZT.
081      (3.9)

```

Eine mögliche Erklärung für die Notwendigkeit der Reparatur ist, dass das VUI bereits nach der Aussprache des *Intents*, d. h. der die aufzurufende Funktion spezifizierenden Äußerung,²¹⁰ den Listening-Modus nach der kurzen Zögerung durch Lukas verlassen²¹¹ und die entsprechende *Entity* für den *Slot* der Zeitangabe nicht mehr verarbeitet hat. Eine andere denkbare Fehlerquelle wäre, dass die Zeitangabe durch Fehler in der Speech Recognition nicht verarbeitet werden konnte. So oder so erkennt das VUI zwar, dass ein Timer gestellt werden soll, es scheint aber nicht prozessiert zu werden, für wie lange – so wird eine ‚Rückfrage‘ des VUI (076) produziert, was wiederum eine Antwort des Anwenders konditionell relevant setzt. Das Basis-Sequenz-Muster (in diesem Fall das Muster A) wird hier also binnenerweitert und sieht für dieses Beispiel wie folgt aus (Einschub hervorgehoben):

Invokation – (Listening-Modus –) Eingabe – *Rückfrage* – *Invokation* – (Listening-Modus) – Nutzer*innen-Antwort – Stimmausgabe/Scharnier – praktische Umsetzung.

Für einen solchen Einschub soll noch ein zweites, scheinbar sehr ähnliches, im Detail aber anders gelagertes Beispiel angeführt werden, das noch ausführlich diskutiert wird: Reparaturleistungen, die durch das VUI ‚initiiert‘ werden.²¹²

Beispiel (18): Timer (01-CVR02-08)

029 LF: aLEXa?
 030 [(1.3)]
 031 k: [((TV läuft))]
 032 LF: ZWEI minUten;
 033 (0.8)
 034 k: ((zweigliedriger Ton mit fallender
 Tonhöhenbewegung))
 035 (1.0)
 036 LF: aLEXa,
 037 (1.4)
 038 LF: TImeR auf zwei minUten.
 039 (1.2)
 040 AL: zwei minuten (-) ab JETZT,

²¹⁰ Siehe Kap. 3.2.2.

²¹¹ Möglicherweise aufgrund einer fälschlichen Identifikation eines *transition relevance place* (siehe Kap. 6.1.3.1).

²¹² Siehe insb. Kap. 6.1.4.1 für eine Diskussion zur ‚Initiierung‘ von Reparaturpraktiken.

Der Anwender Lukas adressiert dabei das VUI mit der Äußerung „ZWEI minU-ten;“ (032), äußert dabei allerdings keinen *Intent*, sondern gibt lediglich den *Slot* an, d. h. die Spezifikation, die dem *Intent* zuzuordnen ist.²¹³ Das VUI markiert konsequenterweise eine unklare Bezugsstelle mit einem Ton, der als „Earcon“ konzeptualisiert werden kann.²¹⁴ Die Tonfolge – leicht dissonant mit einer absteigenden Tonhöhenbewegung – scheint einen beschließenden Charakter zu haben, sie verweist darauf, dass etwas *nicht* oder *nicht mehr* funktioniert (vgl. Blattner/Sumikawa/Greenberg 1989: 30; Grond/Hermann 2012: 57; Oswald 2012: 254).

Das Geräusch kann mithin als eine nonverbale Form eines Paarteils verstanden werden. So wird es von Lukas gedeutet, wie sich in Segment 036 zeigt: Lukas interpretiert das Earcon nach einer Pause von ca. anderthalb Sekunden als Markierung einer Reparaturbedürftigkeit und wiederholt den Befehl. Dabei schließt er das Aktivierungswort mit ein und repariert die Stelle durch das Hinzufügen des *Intents* in der ersten Position der Äußerung („Timer“) sowie die lokal gebrauchte Präposition „auf“, die nun zwischen *Intent* und *Slot* steht. Die Reparatur ist geglückt, wie die Folgeäußerung des VUI (040) zeigt. Das Sequenzmuster lässt sich also wie folgt ausdrücken (Einschub hervorgehoben):

Invokation – (Listening-Modus –) Eingabe – *Reparaturanzeige* – Invokation – (Listening-Modus –) *Reparatur* – Stimmausgabe/Scharnier – praktische Umsetzung.

Zwischen der Eingabe und dem Ausgabescharnier zeigt sich also eine für Einschübe (Binnenexpansionen) sensitive Stelle. Diese kann zur Spezifikation oder Reparatur der Eingabe genutzt werden – unter Beibehaltung des Sequenzmusters, das sich insofern als stabil erweist. Eine Erweiterung des Basismusters wäre allerdings auch im Anschluss möglich, wie nachfolgend zu zeigen ist.

6.1.2.1.3 Expansionen

Als grundlegende Eigenschaft der Sequenzorganisation wurde schon in frühen konversationsanalytischen Arbeiten die „three-part-structure“ (Sacks/Schegloff/Jefferson 1974: 722) herausgearbeitet: Die paarig organisierten Austauschsequenzen enden eben nicht nach zwei Zügen (z. B. Einladung – Annahme), sondern sie haben danach Raum zur weiteren, freieren Entfaltung (Bewertung, Ratifizierung,

²¹³ Es unterscheidet sich damit auch vom zuvor gezeigten Beispiel, in dem der *Intent* prozessiert und der *Slot* nicht verstanden wurde, was dem VUI die Möglichkeit gab, eine Rückfrage zu formulieren. Ob das VUI also auf die sequenzielle Position innerhalb eines Dialogs als relevante Stelle zugreifen kann oder nicht, hängt auch davon ab, ob es den Inhalt der vorherigen Äußerung ganz oder teilweise prozessieren konnte.

²¹⁴ Vgl. Kap. 6.1.1.2.

Widerspruch o. a.). Die Expansion von Paarsequenzen um einen „dritten Teil“ (Imo/Lanwer 2019: 178) stellt – möglicherweise gerade aufgrund der Offenheit – bislang insgesamt eine nur wenig bearbeitete Forschungslücke in der Sprachwissenschaft dar (vgl. Imo/Lanwer 2019: 178; Schegloff 2007: 115–168), dazu zählen auch dritte Züge in soziotechnischen Dialogen (vgl. Habscheid 2022). Dritte Züge als Antworten auf den zweiten Zug sind also deutlich variabler und nicht durch konditionelle Relevanz i. e. S. verbunden, auch ist die Frage der Abgrenzung noch nicht hinreichend geklärt: „Wo die Grenzen zu ziehen sind zwischen einem dritten Teil einer dreiteiligen Sequenz und dem Beginn einer neuen Sequenz muss noch erforscht werden“ (Imo/Lanwer 2019: 178). Dies lässt sich in Dialogen zwischen Anwender*innen und Smart Speakern zunächst formal eingrenzen: ‚Dritte Züge‘ sind solche, in denen das VUI Äußerungen formuliert, die über die beschriebenen Muster hinausreichen, d. h., nicht als Teil der Antwort oder der praktischen Umsetzung verstanden werden können, die aber zugleich unterbrechungsfrei und ohne neue Invokation an vorherige Äußerungen des VUI anschließen. In zwei im Teilkorpus der dyadischen Dialoge dokumentierten Fällen zeigt sich, dass so gerahmte dritte Züge in VUI-Dialogen auftreten können. In den beiden nachfolgenden Beispielen sind Expansionen belegt, die durch das System ‚offeriert‘ werden. Zunächst wird dazu noch einmal das bereits vorgestellte Beispiel (8) betrachtet.²¹⁵

Beispiel (8): Wenn du mehr Zeit benötigst ... (01-CVR01-08)

```

009      [(4.9)                ]
010 k:   [((vereinzelt leise
        Hintergrundgeräusche))]
011 LF:  aLEXa?
012      (0.9)
013 LF:  stell den timer auf ZWANzig minuten.
014      (1.0)
015 AL:  zwAnzig minUten (.) ab JETZT.
016      übrigens (.) wenn du MEHR zeit benötigst==
017      [=Sage einfa-]
018 LF:  [alexa STOP. ]
019      (70.5)

```

²¹⁵ Siehe dazu die Analyse in Kap. 6.1.1.2 unter dem Aspekt der Anrede. Das Beispiel wird mit dem Fokus auf Überlappungen im Turn-Taking noch einmal aufgegriffen (siehe Kap. 6.1.3.2).

078 (0.5)
 079 AL: der wetterbericht für heute nacht sagt
 schneegestöber voRAUS mit einer
 tiefsttemperatur von minus vier grad;
 080 ich hoffe (.) du hattest [einen schönen
 SONN.]
 081 LF: [aLEXa (.) stop-
]
 082 (166.4)

Hier wird die Expansion produziert, um die ‚Hoffnung‘ zum Ausdruck zu bringen, dass der Tag des Anwenders schön war. Diese Äußerung lässt sich wohl am ehesten mit dem Konzept der „phatischen Kommunikation“ bzw. „phatic communion“ verstehen, wie es von Malinowski (1946[1923]) ursprünglich entwickelt und von Jakobson (1960[2007]) in die Sprachwissenschaft übertragen wurde, wo es vielfältig aufgegriffen, angewendet und weiterentwickelt wurde (vgl. etwa Senft 2009; Habscheid 2018a; 2018b). Mit diesem Konzept werden „sprachliche Praktiken, die augenscheinlich zweckfrei und unter dem Aspekt der Bedeutung irrelevant sind“ (Habscheid 2018b: 178), als sozial relevant eingestuft, sie dienen – technisch gesehen und in der Terminologie von Jakobson (1960[2007]: 165–166) – der Etablierung, Erhaltung und Beendigung des „Kontakts“ und dazu, die „Funktionstüchtigkeit des Kanals zu überprüfen“ (Jakobson 1960[2007]: 165–166). Senft (2009: 228–231) schreibt solchen Äußerungen darüber hinaus auch soziale Bindungsfunktionen zu; wie Habscheid (2018b: 180) ausführt, wird phatische Kommunikation teilweise auch als „beziehungsorientierte[] Kommunikation“ verstanden. Beziehungsorientierte Kommunikation muss nicht „echt“ und „tiefgefühl“ sein (Holly 2001a: 1386) – es genügt, so Holly weiter, dass sie „widerspruchsfrei inszeniert“ wird.

Der Ansatz des VUI, phatische Kommunikation zu betreiben, ist insofern also bemerkenswert, denn er kann einerseits als solche „Inszenierung“ betrachtet werden, die andererseits gerade nicht widerspruchsfrei bleibt. Als Inszenierung von Beziehungskommunikation kann sie verstanden werden, weil die Äußerung nicht der Verfertigung von Handlungen auf ein gemeinsames Ziel von Anwender*in und VUI hin dient, sondern – gerade in Verbindung mit der persönlichen Anrede und der Verwendung der Sprecherdeixis – der Etablierung und möglichen Aufrechterhaltung einer Art sozialen Beziehung.²¹⁸ Die Widerspruchsfreiheit muss allerdings

²¹⁸ Das Konzept der Parasozialen Beziehungen – ursprünglich entwickelt bei Horton/Wohl (1956) und im Zuge der Medienaneignungsforschung aufgegriffen und weiterentwickelt, u. a. bei Frey (1996) und Gleich (1996) – soll hier jedoch keine Anwendung finden. Es wurde medienpsychologisch auf menschliche Medienfiguren (u. a. Seriencharaktere, Moderator*innen und später

in zweierlei Hinsicht bezweifelt werden: Erstens konzeptionell, weil das VUI zwar gezielt eine Persona aufbaut und phasenweise konversationell an der Praxis teilnimmt, aber keinen Begriff von Vergangenheit und Zukunft hat und insofern auch nicht etwas ‚hoffen‘ – oder, was hier vielmehr ausgedrückt wird, ‚wünschen‘ – kann, und zweitens auf der Ebene des praktischen Vollzugs, weil der Anwender Lukas die Äußerung unmittelbar nach ihrem Beginn abbricht (081). Es bleibt anzunehmen, dass mit dieser Form der Erweiterung der zuvor beschriebenen paarigen Sequenz v. a. kommerzielle Interessen der Anbieter verbunden sind, die die Nutzer*innen auf verschiedene Weisen an ihre Plattformen zu binden suchen und dazu verschiedene Verfahren einsetzen und erproben – auch parainteraktive Anreden und Wünsche.

Aufgrund der Limitierungen des akustischen Kanals ist phatische Kommunikation im Anschluss an dyadische Sequenzen eine der wenigen Möglichkeiten, dies überhaupt zu realisieren – andere Verfahren, die grafische Interfaces bieten (unauffällige Einblendungen z. B. individueller Grüße und Empfehlungen), fallen dabei weg. Insofern hilft hier eine Perspektivverschiebung: Was sich auf der sprachlichen Oberfläche und in der situativen sozialen Praxis als Expansion zur Integration von phatischer Kommunikation in den Austausch präsentiert, dient – so ist anzunehmen – der individuelleren Gestaltung der Nutzer*innen-Erfahrungen und der Nutzer*innen-Bindung. Dieser Versuch schlägt jedoch fehl und wird durch den Nutzer abgebrochen. Damit bewirkt sie jedoch einen umgekehrten Effekt: Es wird geradezu unterstrichen, dass die Maschine programmierte, sequenziell unangebundene Äußerungen produziert. Die Funktion phatischer Kommunikation speist sich in Mensch-Mensch-Dialogen aus möglichen Folgeäußerungen des Gegenübers, die die widerspruchsfreie Inszenierung der Beziehungskommunikation komplettieren und die „Funktionstüchtigkeit des Kanals“ (Jakobson 1960[2007]: 166) bestätigen – gerade dies geschieht hier nicht.

Die beiden hier gezeigten Fälle von dritten Zügen nach Paarsequenzen sind nur zwei mögliche Expansionsformen – im ersten Fall stellte das VUI eine bisher ungenutzte Funktion vor, im zweiten Fall wurde eine Form phatischer Kommunikation seitens des VUI betrieben. Es könnte noch weitere Expansionsformen geben, die allerdings im vorliegenden Korpus nicht dokumentiert sind. Wie die

auch auf Influencer*innen auf Social Media) angewendet, mit denen Rezipient*innen aktiv eine Beziehung gestaltet haben – die konzeptionellen Unterschiede zur synthetisierten Stimme eines VUI scheinen zu groß (siehe aber Schneider 2021, die in Befragungen mit Smart Speaker-Nutzer*innen durchaus Elemente von gestalteten Beziehungen zu Smart Speakern herausarbeiten kann). Parainteraktive und pseudokommunikative Elemente lassen sich allerdings in VUI-Dialogen durchaus finden (siehe Kap. 2.3 für die theoretischen Grundlagen und v. a. 6.2 für weitere Analysen diesbezüglich; die Diskussion wird auch in Kap. 7 noch einmal aufgegriffen).

Analysen zeigen, sind Expansionen generell aufschlussreich, wenn es um Charakteristika von VUI-Dialogen und Differenzen zu sozialer Interaktion geht. Das Interface macht durch die Erweiterung der zweiteiligen Basis-Sequenzstruktur auf sich aufmerksam und bringt die Anwender*innen in die ungewohnte Rolle, auf Anfragen des VUI zu reagieren. Dabei wird deutlich, dass das VUI die Expansionen nicht in den situationalen Verlauf einbetten kann – zwar ist es angebunden an den vorherigen Austausch zwischen Anwender*in und VUI, aber nicht an den übrigen Vollzug der Praxis und nur sehr bedingt zugeschnitten auf die Rezipient*innen, sodass es nicht verwundert, dass Expansionen dieser Art in den dokumentierten Beispielen unterbrochen werden.

6.1.2.2 Sequenzialität in VUI-Dialogen

An die Ausführungen zur Sequenzorganisation anschließend soll nun ein allgemeinerer Blick auf Sequenzialität in VUI-Dialogen geworfen werden, d. h. insbesondere die Relevanz der Positionierung bestimmter Äußerungen innerhalb einer Abfolge von Äußerungen. Dies eingehender zu beleuchten, begründet sich u. a. darin, dass „Dialogizität“ – d. h. mit Imo (2016: 338) die Einstellung, „Sprache grundsätzlich von ihrem tatsächlichen oder möglichen Potential ihrer dialogischen Wirkung her zu denken“ – sich u. a. auf das Prinzip der Sequenzialität stützt (vgl. Linell 1998: 84).²¹⁹ Die Dialoggestaltung ist also wesentlich von Adressat*innenorientierung geprägt. Ob es sich dabei aus konversationsanalytischer Perspektive um Recipient-Design (Sacks/Schegloff/Jefferson 1974: 727) im engeren Sinne handelt, wird später zu diskutieren sein. Entscheidend ist für die Sequenzialität in VUI-Dialogen, dass bei näherer Betrachtung rekurrent beobachtbare sprachliche Praktiken sichtbar werden, in denen sich durch- aus spezifische Charakteristika auf einer anderen Ebene ausmachen lassen können: nicht sprachstrukturell im Sinne einer spezifischen Syntax oder eines selektierten Lexikons,²²⁰ sondern sequenzielle Strategien der Nutzer*innen zum Umgang mit ihrem ‚restringierten‘ Gegenüber (vgl. Fischer et al. 2011). Dies steht im Einklang mit den bereits vorgestellten Konzepten von Suchman zur Einführung von Mensch-Maschine-Interaktion als „resource-limited communication“ (Suchman 1990: 43).²²¹ Zentral für diese Strategien ist die Ausbildung neuer Formen von Dialogizität (vgl. Lotze 2020: 363; Opfermann/Pitsch 2017: 312; Habscheid 2022: 174). Aus Analysen der Gesprächsorganisation mit dem konversationellen Agenten Max²²² leitet Krummheuer (2010: 226–227) ab, dass diese sich durch Diskontinuität sowie technische Determiniertheit

²¹⁹ Siehe auch Kap. 2.2.2.

²²⁰ Dies könnte allerdings ebenfalls mit Einschränkungen zutreffen, siehe Kap. 6.1.3.

²²¹ Siehe auch Kap. 6.1.1.1.

²²² Siehe Kap. 3.1.2.

und Trägheit auszeichnet. Die sequenzielle Kohärenz dabei ist regelmäßig „aufgebrochen“ (Krummheuer 2010: 229). Diese ‚Aufgebrochenheit‘ zeigt sich etwa im folgenden Beispiel:

Beispiel (19): Wie laut ist es? (07-CVR01-55)

```

048 KW: wie LAUT ist es jetzt im moment;
049      [(1.7)           ]
050 k:  [((Musik endet)))]
051 SI: die LAUTstärke ist bei zweiunddreißig
      proZENT;
052      [(1.4)           ]
053 k:  [((Musik setzt wieder ein)))]
054 KW: hey siri LAUTstärke achtunddreizig
      prozent.
055      [(22.4)         ]
056 k:  [((Musik wird erst sehr leise/verstummt
      und dann lauter als zuvor;spielt
      fortlaufend)))]

```

Nachdem der Nutzer Konrad eine Abfrage der Lautstärke vervollständigt hat (048), fordert er darauf aufbauend eine Einstellung der Lautstärke auf einen bestimmten Wert, woraufhin diese angepasst wird. Die beiden Sequenzen sind unmittelbar thematisch miteinander verknüpft und die Information aus 051 war notwendig, um in 054 eine sinnvolle Angabe zur Lautstärkeregulierung machen zu können. Gleichwohl stellt sich der Vorgang in zwei separaten Sequenzen dar, die der zuvor eingeführten Basis-Sequenzstruktur folgen und in denen keine Bezugnahme aufeinander markiert wird. Um aufbauend auf der Information zur aktuellen Lautstärke aus der ersten Sequenz eine Erhöhung einzustellen, formuliert Konrad dies als ‚neue‘ Sequenz mit einer erneuten Invokationspräsequenz. Abgesehen von der Wiederholung des Lexems „Lautstärke“, mit der Konrad an die Äußerung des VUI anschließt, wird kein Kohärenzmarker produziert; gerade die Wiederholung (und nicht etwa die Produktion einer Pronominalform) betont allerdings die Nicht-Kohärenz zwischen den beiden Sequenzen, ebenso wie die erneute Invokation, die diese voneinander abgrenzt. Die eigentliche Mehrzügigkeit des Austauschs geht dabei sprachlich verloren und er zerfällt in zwei Teile, die im Gegenteil verbal und intonatorisch voneinander abgegrenzt sind. Sofern das VUI also keine Expansion anbietet, ist diese nutzer*innenseitig nur auf die hier beschriebene Weise zu realisieren; es zeigt sich darin sowohl die von Krummheuer beschriebene Aufgebrochenheit wie auch die Trägheit des Austauschs.

Auch im nachfolgenden Beispiel zeigt sich die Trägheit von VUI-Dialogen, die durch die fehlenden sequenziellen Verknüpfungen entsteht:

Beispiel (20): Mix der Woche (01-CVR01-05)

009 AK: aLEXa?
 010 (1.3)
 011 AK: spiel meinen mix der WOche aus spotify?
 012 [(1.4)]
 013 k: [((leises Hintergrundgeräusch))]
 014 AL: dein mIx der woche von SPotify;
 015 AK: aLEXa?=
 016 =LAUter?
 017 (0.5)
 018 k: ((„Splitter von Granaten“ von Adam Angst
 wird abgespielt))
 019 AK: aLEXa?
 020 (.) nächster SONG.
 021 (0.3)
 022 AK: und LAUter bitte:-
 023 (1.2)
 024 k: ((“Splitter von Granaten“ von Adam Angst
 wird lauter abgespielt))
 025 AK: aLEXa?
 026 nächstes LIED;
 027 k: ((„Splitter von Granaten“ von Adam Angst
 wird weiter abgespielt))
 028 ((„Wir schießen auf Karriere“ von Revolte
 Springen wird abgespielt))

Der Ausschnitt umfasst insgesamt vier basisstrukturelle Sequenzen, in denen das Abspielen, die Einstellung der Lautstärke sowie die Auswahl eines anderen Titels vollzogen werden. Die erste Sequenz (009–014) startet die Musikwiedergabe, die zweite Sequenz (015–018) formuliert die Anforderung, die Lautstärke zu erhöhen. Diese wird jedoch noch vor Beginn der Musikwiedergabe abgeschlossen, sodass sie nicht umgesetzt wird, zumal das comparative Potenzial der Eingabe („LAUter“, 016) uneingelöst bleibt, weil nicht klar ist, im Verhältnis zu welchem Positiv es steht (es erfolgte keine aktuelle Wiedergabe, die als solcher betrachtet werden könnte). In der dritten Sequenz formuliert AK die Aufforderung, den (in der aufgerufenen Playlist) nächsten Titel abzuspielen, und verbindet dies in einer Expansion

mit der Reformulierung der Eingabe zur Regulierung der Lautstärke („und LAUter bitte:-“, 022). Allerdings hat dies keinen Wechsel des Lieds zur Folge, sondern ausschließlich die Anpassung der Lautstärke. Dieser Sequenz-Teil scheint aus nur einer Eingabe zu bestehen, weist aber zwei ‚Aufgaben‘ für das VUI auf, die in zwei Turns produziert werden, zwischen denen 0,3 Sekunden Pause und eine klare prosodische Abgrenzung bestehen, obschon AK sie mittels eines Konnektors miteinander verbindet. Es ist möglich, dass aufgrund dieser Eingabenformulierung nur eine der beiden Aufgaben verarbeitet wird. In der vierten Sequenz schließlich wird die Eingabe zum Wechsel des Lieds reformuliert und der nächste Titel wiedergegeben (028); die beiden kommunikativen Projekte konnten abgeschlossen werden.

Es sind in diesem Beispiel mehrere Merkmale, die die ‚Aufgebrochenheit‘ sequenzieller Strukturen zeigen: zum einen die viermalig wiederholte Invokation (009, 015, 019, 025), die immer wieder eine neue Basis-Sequenzstruktur eröffnet. Außerdem zeigen sich mehrfache Wiederholungen von einzelnen Lexemen („LAUter“, 016 und 022) oder syntaktischen sowie prosodischen Strukturen („nächster SONG“, und „nächstes LIED“, 020 und 026). Die Äußerungen von AK nehmen nicht auf verbal produzierte Äußerungen des VUI Bezug, sondern sind alleinstehend, syntaktisch offen, ohne Kohärenzmarker und auf die Ausführung einer Handlung des adressierten Gegenübers gerichtet. Sie entfalten so im Zusammenspiel mit der TCU-Gestaltung auch sequenziell ihre Wirkung als ‚Kommando‘ oder ‚Befehle‘.

6.1.2.3 Zwischenfazit Sequenzialität und Sequenzorganisation

Die Aufgebrochenheit sequenzieller Strukturen (Krummheuer 2010: 229)²²³ ist zugleich die Folge der hohen Stabilität der zuvor beschriebene Basis-Sequenzstruktur. Diese ist zunächst eine technische Notwendigkeit, doch trotz thematisch miteinander verbundener Eingaben und möglicher VUI-Ausgaben ist in den dyadischen Konstellationen im Korpus keine Entfaltung einer sequenziellen Struktur dokumentiert, die über die beschriebene Sequenzstruktur²²⁴ hinausreicht (abgesehen von den beschriebenen Beispielen von Postexpansionen). Zwar sind Aneinanderreihungen solcher Sequenzen dokumentiert, die aber – wie beschrieben – sprachlich keine sequenzielle Kohärenz markieren und keine ‚Gesprächsentfaltung‘ darstellen, bei denen Sprecher*innen wechselseitig Bezüge aufeinander herstellen. In den hier dokumentierten Fällen beginnt der Austausch zwischen den Nutzer*innen und dem VUI nach jedem Ablauf einer Basis-Sequenzstruktur von vorn, unabhängig davon, wo im Verlauf des Dialogs diese positioniert ist. Dies hängt auch mit den Nutzungssituationen und -typen zusammen: Im Teilkorpus der dyadischen VUI-

²²³ Siehe auch Kap. 2.2.3.

²²⁴ Kap. 6.1.2.1.1.

Dialoge dokumentiert und hier vorgestellt sind v. a. Nutzungen für Informationsabfragen, Musikwiedergabe und Alarmfunktionen. Im erweiterten Korpus lässt sich darüber hinaus noch Smart Home-Steuerung als ein ähnlich aufgebauter Nutzungstyp ausmachen. In Mehrparteien-Konstellationen kommen auch andere Anwendungsfälle zum Einsatz, diese stehen aber im Kontext von Test- und Vorführungssituationen. Die Relevanz der Basis-Sequenzstruktur lässt sich jedoch auch dort feststellen.

Es zeigt sich also eine Sequenzorganisation, die sprachlich in einem Höchstmaß auf die (vermuteten oder tatsächlichen) Verarbeitungsfähigkeiten der VUIs sowie auf die zu vollziehenden Handlungen ausgerichtet ist. Dies bestätigt noch einmal die Annahme, dass sich die sequenzielle Dialoggestaltung zwischen VUI und Anwender*innen substantiell verändert, wenn Mehrparteien-Interaktionen den VUI-Dialog umgeben und wenn die Anwendung nicht zur routinisierten Durchführung von Handlungsabläufen geschieht, sondern auch als Aktivität selbst ihre Spuren in der Praxis hinterlässt. Auf diese Phänomene gehe ich in Kap. 6.2 ausführlich ein; zuvor sollen aber die basalen Prinzipien der Dialogorganisation auf Grundlage der direkten Dialoge zwischen Anwender*innen und VUI weiter beschrieben werden. Dazu wird im nächsten Schritt das Turn-Taking eingehender betrachtet.

6.1.3 Turns und Turn-Taking zwischen Mensch und VUI

„So far it seems to be the case that wherever investigators have looked carefully, talk in interaction is organized to be done one speaker at a time“ (Schegloff 2006: 71) – auch wenn es dabei Ausnahmen gibt, kann diese Beobachtung als der zentrale Ausgangspunkt für das Problem des Sprecher*innenwechsels gesehen werden. Damit wird Turn-Taking, d. h. die Übergabe des Rederechts zwischen den beteiligten Sprecher*innen, als eine zentrale kommunikative Herausforderung in Gesprächen offensichtlich, die in der sprachlichen Praxis bewältigt werden muss. Dies gelingt kompetenten Sprecher*innen überwiegend „mit schlafwandlerischer Sicherheit und Präzision“ (Auer 2020: 106). Außerdem verläuft er in der Mehrzahl der Fälle schnell: Bei einer Untersuchung mit zehn verschiedenen Sprachen konnten Stivers et al. (2009: 10589) zeigen, dass der Abstand zwischen einem Turn und dem nachfolgenden Redezug sprachübergreifend am häufigsten zwischen 0 und 200 Millisekunden beträgt. Die Sprachen waren dabei so ausgewählt, dass sie eine hohe sprachstrukturelle, kulturelle und geografische Diversität abbildeten. Die Abweichungen zwischen den Einzelsprachen waren insgesamt nur sehr gering; dies stützt die Hypothese, dass Turn-Taking in sich endogen entfaltenden, nicht institutionell überformten Konversationen eine sprachliche Universalie darstellt

(vgl. Stivers et al. 2009: 10591; Pika et al. 2018). Für das Englische zeigt sich ein Modalwert von ca. 200 Millisekunden (vgl. Levinson/Torreira 2015: 7; Levinson 2016). Die im vorigen Kapitel diskutierte Sequenzialität in Gesprächen trägt dazu bei, den Sprecher*innen-Wechsel so reibungslos zu organisieren: Teilweise ergeben sich nächste Sprecher*innen aufgrund eines mehr oder weniger ausgeprägten Grades an sequenzieller Determiniertheit (vgl. Auer 2020: 107).

Der kanonisch gewordene und bereits mehrfach erwähnte Text von Sacks/Schegloff/Jefferson (1974) liefert für das Turn-Taking die bis heute in der konversationsanalytischen Forschung zentralen Fragen und Begriffe. Die bisher beschriebenen Beobachtungen und Befunde legen nahe, dass es Koordinationseinheiten gibt, an denen sich die Sprecher*innen orientieren und mit denen sie sich gegenseitig den Aufbau ihrer eigenen Äußerungen verdeutlichen; dies ist auch in der jüngeren Gesprächsforschung Konsens (vgl. Auer 2020: 113–114). Die zentralen Aspekte, unter denen dieses System zu betrachten ist, sowie ein zugehöriges Begriffsinstrumentarium haben Sacks/Schegloff/Jefferson (1974) formuliert: Erstens unter dem Aspekt der *turn-construction*, d. h. entlang der Frage, welche Einheiten „oberhalb der Wortebene und unterhalb der Sequenz- und Themenebene“ (Deppermann/Proske 2015: 18) einen Turn, d. h. einen Sprecher*innen-Beitrag, ausmachen und wie er zusammengesetzt ist. Diese Einheiten bestimmen und bezeichnen Sacks/Schegloff/Jefferson (1974: 701) als „turn-constructive unit“ (TCU). Eine TCU ist als Einheit dann abgeschlossen, wenn im Verlauf der Äußerungsproduktion Punkte erreicht werden, an denen das Rederecht von anderen Gesprächsteilnehmer*innen übernommen werden könnte – diese Punkte werden als *transition-relevance places* (TRPs) bezeichnet. Damit ist allerdings nicht gesagt, dass der Turn zwangsläufig dort enden muss, er kann auch von dem*der gleichen Sprecher*in fortgesetzt werden. Außerdem kann eine Übergabe des Rederechts auch ohne einen Abschluss der TCU durch einen TRP stattfinden – etwa durch kompetitives Turn-Taking (vgl. Couper-Kuhlen/Selting 2017: 106) oder im Fall unterschiedlich aufgebauter Ko-Konstruktionen, in denen ein Turn von zwei oder mehr Sprecher*innen gemeinsam produziert bzw. vervollständigt wird (vgl. Günthner 2015).

In dieser Arbeit an der Definition spiegelt sich die praxeologische Ausrichtung der Forscher*innen in der frühen, ethnomethodologischen Konversationsanalyse: Nicht etwa syntaktische Geschlossenheit, prosodische, gar semantische oder andere linguistische Merkmale sind entscheidend für die Einheitenbildung (sie sind vielmehr konstitutive Charakteristika der Einheiten), sondern „mögliche Übergabepunkte“ (Auer 2020: 113), die genuin auf die Orientierung an den anderen Gesprächspartner*innen ausgerichtet sein *müssen* und auch von den Gesprächsteilnehmenden gemeinsam als „cooperatively achieved objects“, als „the result of interaction between the interlocutors“ (Couper-Kuhlen/Selting 2017: 36) hervorgebracht werden. Dabei spielen neben syntaktischer, prosodischer und se-

mantischer Abgeschlossenheit innerhalb einer Einheit (vgl. Ford/Thompson 1996) ganz besonders multimodale Aspekte eine Rolle (vgl. Goodwin 1979; 1981), insbesondere der Blick ist – mit gewissen Limitierungen (vgl. Lerner 2003) – auch für die Auswahl der nächsten Sprecher*innen zentral.

Damit ist auch der zweite Aspekt angesprochen, unter dem das Turn-Taking-System zu untersuchen ist: der *turn-allocation*, d. h. nicht ‚nur‘ der Konstruktion von Redezugeneinheiten, sondern auch der Zuweisung des Rederechts für die Konstruktion des oder der nächsten Einheiten. Diese beiden Aspekte sind sehr eng miteinander verwoben – entstehen doch die Einheiten erst durch das Allokationspotenzial an bestimmten Stellen im inkrementellen Verlauf der Konversation. Für den Sprecher*innenwechsel nach einer TRP unterscheiden Sacks/Schegloff/Jefferson (1974: 703) zwischen Fremd- und Selbstwahl: Ist ein Turn zu Ende, kann der*die Sprecher*in eine*n nächste*n Sprecher*in bestimmen (Fremdwahl). Möglich ist aber auch, dass keine Wahl getroffen wird – sodann kann jede*r andere Gesprächsbeteiligte das Rederecht ergreifen (Selbstwahl). Geschieht auch dies nicht, kann der*die bisherige Sprecher*in weiter sprechen (vgl. Imo/Lanwer 2019: 175; Auer 2020: 113–114; Sidnell 2012: 314–315). Als zentrale Charakteristika des Turn-Takings beschreiben Sacks/Schegloff/Jefferson (1974: 699–700) die „twin features“: Das System des Sprecher*innenwechsels ist einerseits „context-free“, weil es als Repertoire von Möglichkeiten unabhängig von der konkreten Sprechsituation zur Verfügung steht. Andererseits aber ist es in hohem Maße „context-sensitive“, weil jede Konstellation von Teilnehmenden, jede lokale, sozialräumliche Konfiguration, mithin jede Gesprächssituation einzigartig ist und die Teilnehmenden das Repertoire auf verschiedene Gegebenheiten anpassen können (siehe auch Lerner 2003).

Sacks/Schegloff/Jefferson (1974: 701) weisen darauf hin, dass in anderen „speech-exchange systems“ andere Systematiken des Sprecher*innenwechsels gegeben sein könnten, etwa durch institutionelle Zuweisungen des Rederechts. Clayman (2012b: 151) nennt hierfür Interviews, zeremonielle Vorgänge und Debatten als Beispiele und bemerkt die beschränkende und spezifizierende Funktion, die solche institutionellen Settings haben können. Allerdings grenzt er diese scharf ab von „ordinary conversation“; diese sei die tatsächlich dominierende Form sprachlichen Austauschs im täglichen Leben (Clayman 2012b: 151). Ob dieser Dualismus allerdings auch empirisch Bestand hat, lässt sich nicht nur vor dem Hintergrund soziologischer Debatten zu Konzeptionierungen des „Alltags“ hinterfragen,²²⁵ sondern auch mit Blick auf den Gegenstandsbereich der in Smart Speakern verbauten VUIs: Diese sind zwar Bestandteil

²²⁵ Siehe ausführlich dazu die grundlegenden Konzeptionierungen der Phänomenologie des Alltags bei Schütz/Luckmann (2003) sowie Kap. 2.3; für eine kurze, sprachwissenschaftlich zugechnittene Übersicht siehe Lehr et al. (2001).

des Alltags (bzw. werden dazu gemacht, indem sie in die und mit den alltäglichen Praktiken im häuslichen Umfeld verwoben werden). Zugleich sind sie aber durch ihre Technizität von zwischenmenschlicher Alltagskommunikation zu unterscheiden und scheinen konzeptionell institutionell überformten kommunikativen Arrangements zu ähneln.²²⁶ Die Technizität des VUI als ‚Gesprächspartner‘ könnte also auf die Formen, mit denen die Gesprächspraxis hervorgebracht wird, einen ähnlichen Einfluss haben wie institutionelle Konfigurationen.²²⁷

Diese Ausführungen zeigen schon an, dass bei VUI-Dialogen andere Verfahren des Turn-Taking zum Einsatz kommen (müssen). Diese sind durch die technischen Notwendigkeiten der VUIs sowie durch die Kommunikationsanlässe geprägt. Nichtsdestoweniger eignet sich die Turn-Taking-Universalie auch zur Untersuchung von Beteiligten mit nicht voll konversationsfähigen bzw. nicht-menschlichen Partizipanden, wie etwa Mondémé (2022) am Beispiel von Mensch-Tier-Interaktionen erörtert. Tomasello (1988: 76) weist in einer Studie zum frühkindlichen Spracherwerb darauf hin, dass in Turns organisierte Abfolgen („episodes of behavioral turn-taking“) noch nicht zwingend ‚Konversationen‘ entstehen – vor diesem definitorischen Kurzschluss gewarnt zu sein ist auch bei den folgenden Ausführungen und Analysen zu Turn-Taking in VUI-Dialogen zentral, denn die Anwendung und Beleuchtung gesprächsorganisatorischer Prinzipien soll keine Gleichsetzung von VUI-Dialogen mit Gesprächen bedeuten.

Für das VUI-Design benennt Skantze (2021: 1–2) verschiedene Herausforderungen, die unmittelbar mit dem Turn-Taking in Verbindung stehen. Dazu zählen erstens überlappende Äußerungsproduktionen von Nutzer*innen und VUI (vgl. Roddy/Skantze/Harte), zweitens verzögerte Antworten durch das VUI, die wiederum zu ungünstig platzierten Folgeturns durch die Nutzer*innen führen können (Kuzuoka et al. 2008; Schlangen/Skantze 2009; Skantze 2021), und drittens Mehrparteieninteraktionen als Herausforderungen für VUIs im Allgemeinen (vgl. Bohus/Horvitz 2009; 2010; Pitsch/Gehle/Wrede 2013; Johansson/Skantze 2015). Es ist nicht überraschend, dass diese Herausforderungen sich grob in den Punkten spiegeln, in denen sich in Mensch-Mensch-Interaktionen Muster etabliert haben, um sie entweder zu vermeiden (vgl. Ford/Thompson 1996; McFarland 2001; Clark/Fox Tree 2002; Clark 2005; de Ruiter/Mitterer/Enfield 2006; Torreira/Bögels/Levinson 2015) oder sie, für den Fall, dass sie doch auftreten, ohne Gesichtsbedrohung und längere Verzögerung lösen zu können (vgl. Schegloff 2000b; Lerner 2003). Neben verbalen, d. h. auf Ebene der Syn-

²²⁶ Siehe Analyse zu Beispiel (7) („Schneegestöber“) in Kap. 6.1.2.1.3.

²²⁷ Sprecher*innen haben allerdings die Option, die institutionellen Rederechtssysteme zu brechen (ggf. mit entsprechenden Sanktionen), was sie als Beschränkung durch Regeln und Normen verständlich werden lässt; dies ist ein wesentlicher Unterschied zu Limitierungen in der Rede-rechtsverteilung, die durch technische Voraussetzungen entsteht.

tax, Pragmatik und Semantik bzw. Lexik angesiedelten Verfahren, zählen dazu auch prosodische Lösungen (insbesondere Intonation, Intensität, Stimmqualität und Atmung) in ihrem *Zusammenspiel* mit anderen Verfahren (vgl. Selting 1996). Darüber hinaus wurde mit zunehmender Forschungstätigkeit in diesem Bereich die Bedeutung von Körperhaltung, Blick und Gestik immer deutlicher (exemplarisch sei hier verwiesen auf die Arbeiten von Goodwin 1980; Mondada 2007; Goodwin 2007; Mondada/Oloff 2011; Auer 2018).

Im VUI-Design kommen zur Ermöglichung eines ‚reibungsfreien‘ Austauschs zwischen Anwender*innen und VUI einerseits Verfahren zum Einsatz, die das Turn-Taking ermöglichen oder vereinfachen, wie z. B. die bereits ausführlich diskutierte Anwendung von Aktivierungswörtern,²²⁸ die als explizites Turn-Taking-Signal fungiert (vgl. Cha et al. 2020). Gleichzeitig werden aber auch verschiedene Modelle zur Erkennung von TRPs eingesetzt: Skantze (2021: 12) unterscheidet zwischen „silence-based“, „IPU-based“²²⁹ und „continuous models“. Während bei erstgenannten lediglich Stillephasen, in denen kein Ton aufgezeichnet wird, herangezogen werden, sollen IPU-basierte Modelle auch menschliche Verfahren zur Turn-Allokation einbeziehen, z. B. Erkenntnisse zu Prosodie und – gerade für humanoide Roboter – Körperhaltung und Blick. Dazu werten sie die zuvor als Einheit identifizierten Äußerungen einzelner Sprecher*innen aus, um entsprechende Merkmale zu erkennen. Auf diesem Prinzip basieren auch *continuous models*, die zusätzlich nicht an den Grenzen von Sprecher*innenbeiträgen, sondern inkrementell während der Gesprächsbeitragsproduktion immer wieder neue Entscheidungen über Beginn und Ende eines Turns treffen. Die markt gängigen Modelle von Smart Speakern mit ihren jeweiligen VUIs arbeiten auf der Grundlage von *continuous models*, die auch Turn-Übergabe-Markierungen erfassen sollen.²³⁰ In diesen soll das VUI u. a. anwenderseitige turninitiale oder -interne Zögerungen bzw. Pausen, verschiedene Sprecher*innen und sogenanntes *barge-in* – Unterbrechungen durch die Nutzer*innen im laufenden Turn (vgl. Heins et al. 1997) – erkennen.²³¹

In der konversationsanalytischen Literatur, die sich auf zwischenmenschliche Interaktionen bezieht, sind also drei miteinander verbundene Schwerpunkte zur Untersuchung von Turn-Taking auszumachen. Unter dem Aspekt der *turn-construction* sind dies:

²²⁸ Siehe Kap. 6.1.1.

²²⁹ Die Abkürzung *IPU* steht für „Inter-pausal units“ und meint zusammenhängende Äußerungseinheiten einzelner Sprecher*innen, die nicht oder nur durch Rezeptionssignale unterbrochen werden (vgl. Skantze 2021: 4).

²³⁰ Siehe etwa Natarajan/Mandal/Ström (2023).

²³¹ Entsprechende Features wurden für Amazon bereits 2020 angekündigt, siehe Natarajan/Mandal/Ström (2023).

- a. die Bestimmung und Aushandlung von TRPs durch die Beteiligten,
- b. turninterne Eigenschaften von TCUs, sowie
- c. unter dem Aspekt der *turn-allocation* die Verteilung von Rederecht mit ihren verbalen und non-verbalen Verfahren.

In der Literatur, die sich primär mit Mensch-Maschine-Dialogen befasst, wird Turn-Taking demgegenüber v. a. als ein Bündel von kommunikativen Herausforderungen in Interface-Dialogen beschrieben (in den wissenschaftlichen Arbeiten aus der Computerlinguistik und HCI-Forschung mit entsprechenden Designanliegen und -vorschlägen verknüpft), wobei diese mit den zuvor genannten Aspekten eng verbunden sind:

- a. die Erkennung von TRPs in Abgrenzung zu anderen sprachlichen Phänomenen (wie z. B. Zögerungen ohne Turn-Abschluss),
- b. Überlappungen bzw. Unterbrechungen,²³²
- c. Zögerungen und entsprechende Schwierigkeiten in den Folgeturns.

Die nachfolgenden Analysen zeigen – wiederum ausgehend von den gesprächsorganisatorischen und somit konversationsanalytisch untersuchten Aspekten –, wie diese sprachlich in VUI-Dialogen zur Bewältigung der kommunikativen Herausforderungen zum Einsatz kommen und inwieweit sie dabei suffizient sind oder sich ggf. neue sprachliche Praktiken auffinden lassen. Zuvor wird jedoch anhand eines Beispiels noch einmal aufgezeigt, was banal erscheinen mag, doch zugleich zwischen den Kategorien liegend nicht übersehen werden darf: der enge Zusammenhang der Organisation von Sequenzen (wie im vorherigen Kapitel beleuchtet) und des Sprecher*innenwechsels. Wie oben herausgearbeitet, zeichnen sich VUI-Dialoge durch einen hohen Grad an handlungssequenzieller Determiniertheit der Äußerungen aus. Dies soll nachfolgend illustriert werden:

Beispiel (21): Lewis Capaldi (07-CVR01-29)

- 003 KW: äh:: hey siri spiele someone you LO:ved
von lewis capaldi;
004 (2.4)
005 SI: ich spiele someone you loved von lewis
capaldi JETZT-
006 (1.3)

²³² Mehrparteieninteraktionen erweisen sich wie bereits erwähnt ebenfalls als Herausforderung für VUIs (vgl. Porcheron et al. 2018; Beneteau et al. 2020a).

007 k: ((„Someone You Loved“ von Lewis Capaldi
wird abgespielt 175.2 Sek.))

Der Dialog verläuft nach dem Muster einer oben aufgezeigten Basis-Sequenzstruktur: Nach der Invokation und dem Herstellen des Listening-Modus folgt die Eingabe (003) sowie ein Antwort-Scharnier (005) und die praktische Umsetzung (007). Betrachtet man diesen Ausschnitt unter dem Aspekt des Sprecher*innenwechsels, fällt auf, dass dieser sich seitens beider Beteiligten genau an den Grenzen der vollzogenen Handlungsschritte zum Abspielen des gewünschten Musiktitels orientiert, auch wenn dies nicht bei der Invokations-Präsequenz zutrifft, weil diese seitens des VUI nonverbal bleibt. Eine sequenzielle Entfaltung über diesen zweizügigen Dialog hinaus entsteht auch hier aufgrund der hohen Zweckbezogenheit des VUI-Dialogs nicht.²³³ Im Gegensatz zu einem zwischenmenschlichen Gespräch mit freier Themenentfaltung sind die Art der Äußerung und sogar notwendige Informationsbestandteile (auszuführende Aktion, Titel, Interpret) determiniert und dies limitiert den möglichen Interpretationsrahmen für TRPs. Zwar hat das VUI zu erkennen, dass trotz möglicher syntaktischer Geschlossenheit nach „spiele“ oder nach „someone you loved“ (005) noch weiter expandiert wird (siehe auch Clayman 2012b: 155), doch es ist klar, auf welche Bestandteile eines Sprachbefehls (*Intents*, *Entities* und *Slots*) es ‚wartet‘,²³⁴ wodurch mögliche TRPs reduziert werden und das Turn-Taking durch die Technizität seitens des VUI konfiguriert wird. Umgekehrt ist der Anwender Konrad bei Beginn des Dialogs bereits mit der Technizität des Gegenübers vertraut und weiß, dass im VUI keine freie Themenentfaltung vorgesehen ist; sein Reden ist intentional, zweckgerichtet und an den maschinellen Verarbeiter adressiert, dessen Verarbeitungseigenheiten Konrad aus zahlreichen vorherigen Anfragen dieser Art kennt. Entsprechend eindeutig ist auch, wann Konrad den Sprecherwechsel vorsieht, nämlich nach der vollständigen Produktion der Stimmeingabe seinerseits. Die Kontingenz der Rederechtsübernahmen ist hier also im Vergleich zu einem nicht formalisierten Alltagsgespräch mit freier Entfaltung erheblich reduziert. So fällt im Beispiel auch die Möglichkeit einer vorzeitigen Übernahme des Rederechts weg, etwa ab dem Punkt, an dem alle relevanten Informationen zur Umsetzung der Instruktion vorliegen (der Titel des abzuspielenden Liedes genügt). Nachfolgend werden die konversationelle Verfertigung von TRPs zwischen VUIs und Nutzer*innen, auftretende Überlappungen sowie die Charakteristika der Eingabe-TCUs untersucht. Untunlich scheint hingegen eine genauere Beleuch-

233 Siehe dazu Kap. 6.1.2.

234 Siehe auch Kap. 3.2.2.

tung der VUI-TCUs sowie der maschinellen Verfahren der TRP-Bestimmung, die sehr eng mit den Designvorstellungen der Entwickler*innen verbunden ist, die hier jedoch nicht im Fokus stehen sollen. Ebenfalls nicht im Detail beleuchtet werden an dieser Stelle Verfahren zur Turn-Allokation (vgl. Lerner 2003). Dies ergibt sich aus dem Fokus auf die dyadischen VUI-Dialoge, Mechanismen der Rederechtsverteilung treten aber bei der Betrachtung der Mehrparteienkonstellationen wieder auf.

6.1.3.1 Transition-relevance Places (TRP)

Im Fokus stehen nun also die Bestimmung und (zwischen Anwender*innen und VUI) ko-operative Herstellung von TRPs, die als eine der zentralen Herausforderungen im mündlichen Austausch mit Maschinen beschrieben wird (vgl. Skantze 2021: 1). Dazu soll zunächst noch einmal auf Beispiel (21) („Lewis Capaldi“) eingegangen werden: Nach der Äußerung der Eingabe durch den Anwender Konrad (003) folgt eine Pause von 2,4 Sekunden. Das ist deutlich länger als die für das Englische festgestellten 200 Millisekunden für das Turn-Taking in zwischenmenschlichen Interaktionen (s. o.), das üblicherweise bei Frage-Antwort-Sequenzen auftritt und auch bei der Gestaltung von VUIs als Orientierungswert gilt (vgl. Skantze 2021: 3). Nimmt man einen hohen Grad an Verfestigung der Paarsequenz an, könnte diese Dauer (je nach Kontext) möglicherweise bereits markiert sein.²³⁵ Der Anwender Konrad zeigt dies jedoch nicht an, was nahelegt, dass die Latenz aus seiner Sicht unauffällig ist. Dafür spricht auch, dass (ohne dass hier eine genaue quantitative Auswertung stattfinden soll) andere Datensätze im Korpus ähnliche Latenzen aufweisen und dies entsprechend in der regelmäßigen Anwendung von VUIs nicht auffällig ist.²³⁶

Die maschinelle TRP-Erkennung kann bei einem interaktionsanalytischen Vorgehen nicht Teil der Analyse sein. Sie lehnt sich zwar technisch gesehen an das Turn-Taking-System zwischenmenschlicher Interaktionen an, funktioniert aber je nach System unterschiedlich und auf Basis jeweils verschiedener *turn-*

²³⁵ Diese Aussage kann nur mit aller Vorsicht getroffen werden. Aus den eingangs zitierten Untersuchungen, die zeigen, dass sehr schnelles Turn-Taking prinzipiell *möglich* ist und auch praktiziert wird, kann nicht geschlussfolgert werden, dass generell Äußerungen, bei denen der Abstand länger ist, markiert sind. Die soziospatiale Umgebung, der Gesprächskontext, der Grad der sequenziellen Verfestigungen und situativ möglicherweise andere Faktoren verhindern solche Feststellungen. Untersuchungen haben außerdem gezeigt, dass in zwischenmenschlichen Interaktionen konditionelle Relevanz nicht unbedingt ko-okkurent mit Adjazenz ist, sondern zweite Paarteile auch noch lange Zeit nach der Äußerung des ersten Paarteils eingelöst werden können. Entsprechend kann hier nur gesagt werden, dass diese Antwort deutlich über der notwendigen Planungszeit bei zwischenmenschlichen Interaktionen liegt und markiert sein *könnte*.

²³⁶ Zur Pause als prosodisches Gliederungssignal siehe Gülich (1970) sowie Grammis (2023).

taking-cues, die neben pragmatischen Merkmalen auch paraverbale Aspekte wie die Intonationskontur oder Atmen einbeziehen können, aber nicht müssen (vgl. Skantze 2021: 5–6). Es kann als Konsens in der HCI gelten, dass prosodische den verbalen *cues* deutlich unterlegen sind (vgl. Skantze 2021: 9)²³⁷ und dass die maschinelle Erkennung eines TRP aufgrund ihrer Programmiertheit deterministischer erfolgt als die menschliche Erkennung. Bei der Analyse soll allerdings konsequent die Teilnehmer*innenperspektive eingenommen werden, sodass die Gründe für die ko-operative Herstellung eines TRP auch analytisch nachrangig sind – vielmehr steht die Interpretation der Teilnehmer*innen im Fokus.

Für diese ist die Frage der Latenz zentral. Das folgende Beispiel knüpft noch einmal an die Frage der vorgesehenen Dauer zwischen zwei Turns sowie deren Potenzial für die Identifikation von Turn-Grenzen an:

Beispiel (22): Spotify (01-CVR01-04)

- 010 LF: aLEXa?
 011 (1.4)
 012 LF: <<creaky> a::h.>
 013 (2.7)
 014 LF: spiel SPotify;
 015 (3.6)
 016 LF: aLEXa?
 017 (1.0)
 018 LF: spiel SPotify;
 019 (1.2)
 020 AL: mUsik von spotify wird ABgespielt;
 021 (1.3)
 022 k: („What it feels like“ von Sons of Maria
 wird abgespielt, ca. 129 Sek.)

Während bereits in den obigen Ausführungen deutlich wurde, dass für menschliche Beteiligte die Frage der Latenz extrem kontextabhängig, situativ und teilweise auch kulturspezifisch ist (vgl. Stivers et al. 2009), ist das VUI durch einen gewissen Grad an Unflexibilität gekennzeichnet. Der bereits zuvor erwähnte „no-

²³⁷ Siehe aber Ekstedt/Skantze (2022) mit Ergebnissen zu Prosodie in selbstlernenden konversationellen Interface-Modellen, die nahelegen, dass diese Prosodie (insbesondere phonologische Eigenschaften, aber auch Lautstärke und Stimmhöhe) einbeziehen.

input-timeout“ (Skantze 2021: 13)²³⁸ führt in diesem Beispiel dazu, dass der Wechsel vom Anwender Lukas zum VUI nicht stattfinden kann: Lukas äußert erst das Aktivierungswort, zögert dann aber mit insgesamt ca. 4,1 Sekunden stiller Pause und einer gedehnten Zögerungspartikel (012). Die bis zur Äußerung der eigentlichen Stimmeingabe (014) verstrichene Zeit scheint also zu lang zu sein, um eine Übergabe noch zu ermöglichen – das Gerät ist nicht mehr im Listening-Modus. Weder wird der Moment nach der Invokation (010) als TRP behandelt (auch von Lukas nicht), noch wird die Hesitationspartikel, die Lukas produziert (012), durch das VUI als solche erkannt. Diese kann in zwischenmenschlichen Interaktionen als Markierung für „turn-yielding“ (Duncan 1972) fungieren (vgl. Keseling 1989) und könnte hier anzeigen, dass eine Äußerung und ein entsprechender TRP noch folgen werden. So aber schaltet das Gerät zurück in den Modus vor der Invokation. In der Folge muss der Nutzer diese sowie die anschließende Eingabe wiederholen (016–018). Die Basis-Sequenzstruktur erweist sich auch hier als stabile Einheit des VUI-Dialogs.

6.1.3.2 Überlappungen

Sprachliche Praktiken, die in zwischenmenschlichen Interaktionen zum Einsatz kommen können, um etwa Zeit für die Planung des Turns zu gewinnen, ohne das Rederecht aufgeben zu müssen, können also auf VUI-Dialoge nicht übertragen werden. Schwierigkeiten im Timing zeigen sich auch in folgendem Beispiel:

Beispiel (23): Soll es heute noch regnen? (07-CVR01-07)

```
005 KW: hey Siri,
006      (2.0)
007 KW: soll es heute no[ch REGne ]n?
008 SI:      [ ``'mh_MH: , ]
009      (1.5)
010 SI: es sieht heute NICHT nach regen aus.
```

Das VUI von Apple mit Siri produziert nach der Invokation ein Rückmeldesignal im Sinne einer Summons-Answer-Sequenz (Schegloff 1968).²³⁹ Das passiert auch im vorliegenden Beispiel (vgl. 008), ca. 2,5 Sekunden nach der Invokation (005). Der Anwender Konrad hat allerdings zwischenzeitlich bereits in dieser Zeitspanne (nämlich nach ca. 2,0 Sekunden) begonnen, die Stimmeingabe zu artiku-

²³⁸ Siehe Kap. 6.1.1.1.

²³⁹ Siehe Kap. 6.1.1.1.

lieren (007). So überlappen sich das Rückmeldesignal, das noch auf die zuvor geäußerte Invokation bezogen war, und der Sprachbefehl miteinander, was allerdings nicht zu weiteren gesprächsorganisatorischen Schwierigkeiten führt, von Konrad nicht relevant gemacht wird und die Ausgabe des VUI nicht beeinflusst.

Die Überlappung bleibt auch im folgenden Beispiel für ein *barge-in* im Fokus. Als *barge-ins* werden anwender*innenseitige Äußerungen verstanden, die sich gezielt mit den Äußerungen des VUI überschneiden (vgl. Heins et al. 1997). Die Äußerungen der Nutzer*innen werden so kommunikativ als Unterbrechung der Äußerungsproduktion des VUI funktionalisiert:

Beispiel (24): Musik auf Spotify (08-CVR02-02)

005 SM: oKE google,
 006 (2.3)
 007 SM: öffne SPOTify;
 008 (1.8)
 009 GA: entSCHULDige;
 010 [(.) auf diesem gerät] kann ich leider
 keine [APPS öffnen;]
 011 SM: [<<behaucht>OH::: >]
 012 [oKE google-]
 013 (0.5)
 014 SM: spiele SPOTify Ab;
 015 (1.7)
 016 GA: alles KLAR;
 017 muSIK auf spOtify;
 018 (0.2)
 019 GA: BITte sehr;

Nach der Invokation formuliert Samuel die Stimmeingabe „öffne SPOTify“ (007). Diese wird durch das VUI verstanden, kann aber nicht im engeren Sinne der Befehlsformulierung umgesetzt werden – wie sich später zeigt, kann das Gerät zwar Musik aus Spotify wiedergeben, aber die App selbst nicht ‚öffnen‘. Das VUI produziert vor der genaueren Erklärung dazu, was nicht realisiert werden kann (das Öffnen der App), eine exklamative Entschuldigung (009). Diese scheint für Samuel zu projizieren, dass die Stimmeingabe pragmatisch gesehen nicht erfolgreich sein wird; es ist für ihn hinreichend, um die Stimmausgabe (011) zu unterbrechen. Zunächst produziert er eine behauchte Interjektion (011), die zeitlich mit der Äußerung des VUI zusammenfällt, anschließend eine erneute Invokation, die sich mit

Der Dialog läuft zunächst nach dem beschriebenen Basis-Sequenzmuster ab.²⁴² Nach dem Stimmausgabescharnier und der (stillen) praktischen Umsetzung (015) folgt eine Expansion in einem neuen Segment. Der TRP (015) wird dabei seitens des VUI nicht als solcher positioniert. Stattdessen produziert das VUI eine Expansion (016–017), die über den Dialog hinausreicht und Züge einer freien Themenentfaltung annimmt. Lukas bricht diesen an einer Stelle ab, an der kein TRP erreicht war und die Äußerung weder intonatorisch noch syntaktisch geschlossen war. Das VUI produziert den Turn auch während der Produktion des *STOP*-Befehls (018) von Lukas noch weiter, unterbricht dann aber und stellt die Äußerung ein. Es erfolgt im gegebenen Auszug keine Nachbearbeitung (z. B. verbalisierte Rederechtszuweisung, Entschuldigung o. Ä.).

Das Turn-Taking verläuft, so kann konstatiert werden, in *barge-ins* teilweise abweichend von Praktiken des Turn-Taking, die für Interaktionen zwischen Menschen und insbesondere für Überlappungen untersucht wurden (insbesondere Schegloff 2000b). Zwar stoppt das VUI die Äußerungsproduktion, wenn eine überlappende Invokation erkannt wird, was die Präferenz für die unmittelbare Vermeidung von gleichzeitigem Sprechen innerhalb einer Konversation (vgl. Schegloff 2000b: 4) bestätigt. Doch die Frage, wer stoppt, die bei Schegloff als wesentlicher Teil der Lösung zur Behebung des gleichzeitigen Sprechens betrachtet wird und somit konstitutiv für die sprachlichen Praktiken zum Umgang mit Überlappungen ist, ist nicht Gegenstand einer konversationellen Aushandlung zwischen VUI und Anwender*in. Diese ist determiniert: Das VUI stoppt bei Erkennung des Aktivierungsworts. Entsprechend kommen beschriebene Praktiken und ihre Voraussetzungen hier nicht zur Anwendung – so kommt es nicht zu schnellen Turn-Finalisierungen, Verzögerungen oder kompetitiven Überlappungen, sodass die Anwender*innen ihre Äußerungen unbeeinflusst produzieren können.

6.1.3.3 Charakteristika von TCUs der Anwender*innen

Die Eingaben der Nutzer*innen können unterschiedliche Formen aufweisen, die nachgerade den sequenziellen Verlauf prägen und Annahmen über die sequenziellen ‚Fähigkeiten‘ des VUI spiegeln. Dazu zählen alleinstehende substantivische und adjektivische Äußerungen wie im folgenden Beispiel:

Beispiel (25): Leiser und Pause (07-CVR01-19)

003 KW: hey siri LEIser.

004 k: ((Musik wird leiser und spielt

²⁴² Siehe dazu Kap. 6.1.2.1.1.

fortlaufend))

005 (30.6)

006 k: ((weitere Songs werden simultan
abgespielt, evtl. auf einem zusätzlichen
Gerät 8.0 Sek.))

007 ((Musik spielt, TV läuft 47.0 Sek.))

008 KW: hey Siri: leiser.

009 k: ((Musik wird leiser))

((Musikwiedergabe ausgelassen, 119 Sek.))

021 TW: hey siri PAUse.

022 (0.5)

023 TW: ((hustet))

024 k: ((Musik setzt aus))

Das Beispiel zeigt noch einmal, dass bei routinierten Anwendungen wie etwa bei der Musikwiedergabe das VUI in seiner Eigenschaft als Steuerungsinstrument zum Einsatz kommt: Es wird stark routinisiert gesprochen. Die TCUs, die die Anwender Konrad und Till in den Ausschnitten produzieren, bestehen aus einem einzelnen Lexem (Adjektiv im Komparativ, 003 Substantiv, 021) und einer Folge-Aktion ohne verbale Bestätigung, es zeigt sich besonders deutlich der indexikale Zeichencharakter, der auch im VUI-Dialog in die Gestaltung der Turns einfließt (vgl. Bergmann/Meyer 2021a: 45–46). Die sprachlichen Formen basieren auf dem eingeübten Wissen um die Operationen zur Steuerung der Geräte und können als Anpassung an die Limitierungen des Interfaces verstanden werden: Sie sind effizient, weniger fehleranfällig bei der technischen Verarbeitung und nehmen auf die bereits erwähnte phatische Sprachfunktion nach Jakobson (1960[2007]) ebenso wie auf „Facework“ im Sinne Goffmans (1955) oder auf sprachliche Verfahren zur Herstellung von Höflichkeit wie etwa Indirektheit (vgl. Brown/Levinson 1987) keinen Bezug (siehe auch Merkle/Hector 2025). Die sprachlichen Formen erinnern an Kommunikation im Rahmen stark institutionalisierter Präkonfigurationen, in denen die Handlungsausübung gegenüber den kommunikativen Verfahren stark dominiert (vgl. Drew/Heritage 1992) und wie sie z. B. im medizinischen Kontext bei Operationen (vgl. Bezemer et al. 2011; Mondada 2014a) und Untersuchungen (vgl. Lindström et al. 2017: 312) oder in Fahrschulstunden (Deppermann 2018a) auftreten (siehe auch Goffman 1979: 6–7).

Wie wir schon in anderen Beispielen gesehen haben, können dabei auch kleine, syntaktisch offene Phrasen zum Einsatz kommen, die anhand des folgenden Auszugs noch einmal illustriert werden sollen:

Beispiel (26): Timer löschen (07-CVR01-53)

009 KW: hey SIri,
 010 (0.2)
 011 KW: Timer auf (.) zwanzig minuten.
 012 [(0.8)]
 013 k: [((Musik endet))]
 014 k: ((Klopfen))
 015 (0.6)
 016 SI: zwanzig minuten (-) LOS gehts,
 017 k: ((Musik setzt wieder ein;
 Haushaltsgeräusche 5.0 Sek.))
 018 KW: hey SIri timer löschen;
 019 (0.8)
 020 k: [((Musik pausiert kurz))]
 021 SI: [er wurde abgeBROchen-]

Der Nutzer hört über den Smart Speaker von Apple Musik und stellt parallel dazu einen Timer (011). Die Syntax, mit der diese Eingabe formuliert wird, richtet sich mehr an den Verarbeitungsanforderungen des VUI aus als an standardsprachlich üblichen Verfahren: Durch die Platzierung des Lexems „Timer“ in der ersten Position wird die technische Anforderung des Systems bedient, den *Intent* zu finden. Die *Entity*,²⁴³ d. h. die gewünschte Spezifikation, folgt unmittelbar danach („zwanzig minuten“, 011) – diese projiziert den gewünschten Zustand der angegebenen Funktion. Damit ist die Äußerung durch die unterbleibende prädikative Verknüpfung zwar syntaktisch nicht geschlossen, enthält aber alle notwendigen Angaben im Hinblick auf die Verarbeitung der Eingabe. Ähnlich verhält es sich auch bei der nachfolgenden Eingabe (018), in der Konrad den Timer abbricht und dafür wiederum das Substantiv „timer“ produziert, die *Entity* hier aber in Form eines deontischen Infinitivs produziert, wie sie Deppermann (2018a) in Instruktionen in Fahrschulgesprächen feststellt und dies im Zusammenhang mit einer „epistemically, and therefore, deontically [...] superior position“ betrachtet, die auch im Fall der VUI-Dialoge gegeben zu sein scheint: Jede Notwendigkeit zur Explikation der Eingaben entfällt, auch aufgrund der Kontextblindheit des VUIs, demgegenüber die durchzuführenden Handlungen auch nicht im Sinne Garfinkels (1967) „accountable“ gemacht werden müssen.

243 Siehe Kap. 3.2.2.

Syntaktisch etwas komplexer, aber pragmatisch sehr ähnlich ist die Verwendung von imperativischen V1-Sätzen wie im folgenden Beispiel:

Beispiel (27): ‚Together‘ von Sia (07-CVR01-19)

```

011 KW: hey SIri,=
012      =spiele toGEther von sia.
013      (2.2)
014 SI: ich SPIEle together-=from the motion
      picture musIc.
015      (0.3)
016 SI: [von SIa.                ]
017 k:  [((Lied setzt ein und spielt
      fortlaufend))]
```

Der Anwender Konrad verwendet dabei nach der Invokation (011) einen Imperativ in V1-Stellung (012). So entsteht eine syntaktisch geschlossene Einheit, die sich zugleich an den technischen Konfigurationen des VUI ausrichtet: Das Prädikat „spiele“ bildet dabei den *Intent* zur Musikwiedergabe. Deppermann (2021: 222) verweist auf die hohe Bedeutung der sequenziellen Positionierung bei der Verwendung von Imperativen und führt aus, dass sich die tatsächliche Handlungsrelevanz von Imperativen nicht ohne ihre sequenzielle Position bestimmen lässt. Imperative gelten laienlinguistisch häufig als unhöflich und gesichtsbedrohend, sind jedoch in einer Vielzahl unterschiedlicher Erscheinungsformen und -kontexten in Alltagshandlungen vertreten, in denen sich Verfahren für deren Anwendung etabliert haben (vgl. Sorjonen/Raevaara/Couper-Kuhlen 2017: 1). Dies aufnehmend lässt sich sagen, dass Imperative in VUI-Dialogen ein weiterer Fall sind, in dem Imperativformen in den Vollzug der Alltagspraxis eingebunden werden – mit der Besonderheit, dass sie sich nicht an ein menschliches Gegenüber richten. Etablierte sprachliche Praktiken im Umgang mit Imperativformen in Mensch-Mensch-Interaktionen – etwa Imperativformen von Modalpartikeln begleiten zu lassen, um diese in die soziale Praxis einzubetten und den Nutzen des vorgeschlagenen Handlungsschritts für den*die Aufgeforderte*n deutlich zu machen (für das Schwedische siehe Lindström et al. 2017: 320, für das Dänische siehe Heinemann/Steensig 2017: 169) – kommen dabei nicht zu Anwendung. In erster sequenzieller Position nach der Invokation weisen sie im vorliegenden Korpus immer den Charakter einer spontanen Aufforderung auf, die sich durchaus – und damit für das Gros der linguistisch untersuchten Imperativformen eher untypischerweise (vgl. Auer 2017: 412) – auf das Ausführen einer Aktion zum (vordergründig) einseitigen Vorteil der Sprecher*innen richtet. Sprecher*innen verbalisieren in dyadischen VUI-Dialogen auch keinen Kon-

text einer *gemeinsamen* Handlung, wie es Zinken/Deppermann (2017) für in die laufende Praxis eingebettete Imperative in multimodalen Analysen feststellen können. Damit sind die imperativischen Formen Ausdruck der Technizität des Gegenübers und verweisen auf die asymmetrische Beziehung und die starke Routinisierung der Dialoge (siehe auch Rauniomaa 2017: 342; Auer 2017: 413–415).

Fünftens und letztens lassen sich unterschiedliche Fragetypen identifizieren. In Anlehnung an die morpho-syntaktische Unterscheidung bei Graf/Spranz-Fogasy (2018: 25) und Spranz-Fogasy (2010) werden nachfolgend Ergänzungs- und Entscheidungsfragen²⁴⁴ unterschieden (siehe auch Imo 2013: 143), wobei bei Ergänzungsfragen (W-Fragen) üblicherweise ein Fragewort in turninitialer Position steht, während Entscheidungsfragen syntaktisch als V1-Gestalten präsentiert werden. In der Kollektion der dyadischen VUI-Dialoge sind auf 133 erfasste VUI-Dialoge sieben Ergänzungsfragen und zwei Entscheidungsfragen dokumentiert.

Ergänzungsfragen zielen in den dokumentierten VUI-Dialogen auf die Gewinnung von Informationen ab – neben im Korpus ebenfalls belegten ‚Standard‘-Fragen wie „Wie ist das Wetter?“ oder „Wie viel Uhr ist es?“ (siehe obige Beispiele) können sich diese auch auf laufende Vorgänge beziehen,²⁴⁵ wie das folgende Beispiel illustriert:

Beispiel (28): Timerstand (08-CVR02-41)

- 010 RL: hey GOOgle;
 011 (1.8)
 012 RL: wie lang ist der TImeR (noch)-
 013 (1.5)
 014 GA: DU hast zwei timer-
 015 der timer für FÜNFundzwanzig minuten läuft
 noch sEchs minuten und neunundzwanzig
 sekunden-
 016 (0.2)
 017 GA: und der für FÜNFundvierzig minuten ist
 angehalten dreIzehn minuten und dreißig
 sekunden restlaufzeit.

Der Anwender Robin fragt nach der Dauer des Timers. Damit bezieht er sich auf eine Funktion des VUI, die zum Zeitpunkt der Äußerung aktiviert ist. Das VUI ‚be-

²⁴⁴ Bei Hentschel/Weydt (2013: 376) ist von „Ergänzungs-“ und „Bestimmungsfragen“ die Rede.

²⁴⁵ Siehe auch Beispiel (19) („Wie laut ist es?“) in Kap. 6.1.2.2.

richtet‘ über die Timer-Funktion, wobei zwei parallele Funktionsinstanzen (Timer) gestellt sind, von denen einer noch aktiv läuft (015) und einer angehalten wurde (017). Die Identifikation der beiden Timer-Instanzen erfolgt über ihre Laufzeit („timer für FÜNFundzwanzig minuten“, 015, „der für fünfundvierzig miNUTen“, 017). Entscheidungsfragen treten demgegenüber eher im Kontext von Test- und Vorführsituationen auf und sind auf die Funktionen des VUI selbst bezogen. Lediglich Informationsabfragen zum Wetter wie das folgende Beispiel weisen auch in dyadischen Konstellationen Entscheidungsfragen auf:

Beispiel (29): Regenwahrscheinlichkeit (07-CVR01-56)

```

006 KW: hey SIri:,
007 KW: soll es heute noch REGnen?
008      [(1.2)                ]
009 k:  [((Schrittgeräusche enden))]
010 SI: es könnte am abend ab neunzehn uhr REGnen;
011      (2.4)
012 KW: °h
013      (0.6)
014 KW: HEY siri,
015 KW: WIE hoch ist die regenwahrscheinlichkeit
      für heute abend.
016      (1.1)
017 SI: es könnte am abend ab neunzehn uhr REGnen;
018      (0.7)
019 KW: (HM hm.)

```

Auffällig ist allerdings, dass der Nutzer die Antwort scheinbar nicht zufriedenstellend findet, was durch die Wiederholung der Eingabe mit einer abgewandelten Fragestellung deutlich wird, die auf eine präzisere Antwort zielt. Hier wandelt Konrad die Entscheidungs- zur Ergänzungsfrage um, erhält allerdings dasselbe Ergebnis. Bestimmungsfragen mit Lokalinterrogativa wie „wo ist mein Handy“ oder „wo ist der nächste Edeka“ treten in der Kollektion dyadischer VUI-Dialoge überhaupt nicht auf, sie werden nur in den Mehrparteieninteraktionen in Vorführsituationen genutzt.

6.1.3.4 Zwischenfazit Turns und Turn-Taking

Die Beispiele konnten zeigen, dass in dyadischen Dialogen zwischen VUIs und ihren Anwender*innen TRPs nicht wie in zwischenmenschlichen Interaktionen

ausgehandelt werden. Die Rederechtsverteilung ist vielmehr durch die starke Aufgaben- und Zweckorientierung sowie die technischen Grenzen bzw. antizipierten Möglichkeiten der Geräte limitiert. Das Turn-Taking in dyadischen Dialogen mit VUIs ist insofern in hohem Maße von deren pragmatischem Potenzial gekennzeichnet: Die auszuführende Handlung steht bei der Gestaltung im Mittelpunkt, andere sprachlich ausgedrückte Aspekte zwischenmenschlicher Kommunikation und (etwa phatische Aspekte und Bestandteile des Face-Work) treten demgegenüber in den Hintergrund. Beschriebene Verfahren zur Vermeidung gleichzeitigen Sprechens scheinen zwar als Orientierung zu dienen, werden aber beobachtbar nicht durchgängig appliziert. Insbesondere im Fall von *barge-ins* unterbrechen Nutzer*innen systematisch laufende Äußerungen als Teil der Interface-Operation. Die sprachliche Gestaltung ist in dieser Hinsicht auch eine Folge davon, dass keine sozialen Rollen ‚gewahrt‘ werden müssen; die Interaktionen zeigen die Nutzer*innen als einseitig bestimmende Instanzen. Auch darin kommt die höhere Priorisierung der auszuführenden Handlungen gegenüber der Einhaltung gesprächsorganisatorischer Prinzipien zum Ausdruck. Anders könnte es sich noch einmal in Mehrparteieninteraktionen verhalten: Versteht man das VUI mit Hirschauer (2016) als einen materiellen „Partizipanden“ ohne eigenes Aktivitäts- und Rollenpotenzial (wobei auch dies nur mit Einschränkungen gilt), könnte dies Auswirkungen auf die Verteilung sozialer Rollen in Mehrparteienkonstellationen haben.

Die bisherigen Analysen erlauben, eine nicht abschließend zu verstehende Typologisierung von TCU-Gestaltungsverfahren der Eingaben aufzustellen, die sich aus den Daten im Korpus heraus ergibt:

- (1) alleinstehende Nomen oder Adjektive (Letztere v. a. im Komparativ),
- (2) syntaktisch offene Phrasen,
- (3) darunter insbesondere deontische Infinitive,
- (4) Imperativsätze,
- (5) Ergänzungs- und Entscheidungsfragen.

Im Kontext von Einrichtungs- und Testsituationen sind auch andere nutzer*innenseitige Eingaben beobachtbar, die nicht in dieser Typologie abgebildet sind. Allerdings sind diese Situationen nicht aussagekräftig für die routinierten Nutzungspraktiken und die tatsächlich vollzogene Dialogorganisation zwischen Anwender*innen und VUI. Initiale Äußerungen der VUIs sind ansonsten lediglich im erweiterten Korpus an einer Stelle dokumentiert – es treten dann auch andere Eingaben auf, z. B. Antwortpartikeln (insbesondere „ja“ und „nein“) oder alleinstehende Adverbien zur VUI-Steuerung (z. B. „weiter“). Ferner konnten notwendige Reparaturen als Auslöser etwa für die Produktion von *barge-ins*, aber auch für andere „Turbulenzen“ (Auer 2020: 215) bei der Produktion und Übergabe der Turns (z. B. die Entstehung eines *no-response-timeout* oder von Überlappungen

durch Probleme bei der Umschaltung zwischen den verschiedenen Modi des Geräts) beobachtet werden. Sie spielen auch eine Rolle bei der Abweichung von beschriebenen Sequenzmustern im vorherigen Kapitel. Entsprechend werden diese nachfolgend ausführlicher beleuchtet.

6.1.4 Reparaturen

Reparaturmechanismen sind ein weiteres der „generic organizations of practice“ (Schegloff 2006: 71). Sie sind die Lösung für das allgemeine Problem, dass im Verlauf einer Interaktion Störungen auftreten können – aus unterschiedlichen Gründen und mit unterschiedlichen Bewältigungsstrategien. Reparaturen gehören – spätestens seit dem Aufsatz der konversationsanalytischen Pionier*innen (Schegloff/Jefferson/Sacks 1977), in dem dieser Begriff eingeführt wird, und weiter angeregt durch weitere Arbeiten Schegloffs (1979; 1987; 1992) – zu den intensiver bearbeiteten Feldern konversationsanalytischer bzw. interaktionslinguistischer Forschung. Reparaturen sind Bauer (2020: 334) zufolge eine „Komponente der universellen Grundausrüstung zur Sicherung des gegenseitigen Verstehens und der Progression des Gesprächs“; Reparaturmechanismen sind strukturelle Universalien der Gesprächsorganisation – d. h. in dieser Hinsicht unabhängig von ihrem Kontext –, zugleich aber in ihrer Form anpassungsfähig an situative und kulturelle Gegebenheiten, d. h. „kontextsensitiv“ (vgl. Egbert 2009: 10).

Das große Interesse konversationsanalytischer bzw. interaktionslinguistischer Arbeiten an Reparaturen (Uhmann 1997; Egbert 2009; Sidnell 2009: 110–138; Kitzinger 2012; Pfeiffer 2015; Couper-Kuhlen/Selting 2017: 112–209; für eine Übersicht siehe Bauer 2020: 335) hat seine Gründe nicht zuletzt darin, dass gerade durch Störungen im interaktionalen Ablauf die Prinzipien, nach denen dieser funktioniert, sichtbar werden. Die Organisationsprinzipien sprachlicher Interaktion sind mit Schegloff (2006: 70) die „infrastructure that supports the overall or macrostructure of societies“ – und so wie auch andere Infrastrukturen werden sie als solche am sichtbarsten, wenn sie ganz oder teilweise außer Betrieb geraten (vgl. Star/Bowker 2005: 231). Die Ursachen für Störungen, die Identifikation der zu reparierenden Stellen sowie die Mechanismen der Bearbeitung liefern entscheidenden Aufschluss darüber, wie Interaktionen ohne Störung strukturiert sind und was sequenzielle Abläufe behindern kann.

Reparaturen sind mit Bauer (2020: 333) „metakommunikative Praktiken“ als eine Handlungsoption zum Umgang mit Störungen. Es handelt sich dabei um eine zweiteilige Handlungsfolge: Initiierung und Durchführung der Reparatur werden voneinander unterschieden. Die Frage, in welcher Form sich solche metakommunikativen Handlungen des Reparierens im Dialog mit Smart Speakern vollziehen,

drängt sich bei der Betrachtung sprachlicher Praktiken im Dialog mit Smart Speakern geradezu auf, sind die Geräte doch störanfällig. Störungen und anschließende Reparaturen im Dialog mit Smart Speakern sind – trotz der Andersartigkeit des Austauschs²⁴⁶ – als Gesprächspraktiken zunächst durchaus vergleichbar: Aus unterschiedlichen Gründen führte die Äußerung nicht zum beabsichtigten Ergebnis und verhindert so ein ‚störungsfreies‘ Verlaufen der vollzogenen Praxis. Zugleich jedoch unterscheiden sich der Verlauf eines VUI-Dialogs und der Ablauf einer zwischenmenschlichen Interaktion auch an dieser Stelle: Wie bereits gezeigt,²⁴⁷ sind VUIs nicht in gleichem Maße in der Lage, auf den inkrementellen Charakter im Sprecher*innenwechsel zu reagieren. Übergaberelevante Punkte (TRPs) werden – wie die vorherigen Analysen teilweise bereits zeigen konnten – mit einer weniger hohen Wahrscheinlichkeit korrekt erkannt und nach Unterbrechungen kann der *turn-by-turn talk* nicht umstandslos fortgesetzt werden. Dies beeinflusst auch Reparaturmechanismen, wie in den nachfolgenden Analysen gezeigt werden soll.

In der Literatur werden vier Typen von Reparaturen unterschieden. Die Typologie baut auf zwei Unterscheidungsmerkmalen auf: erstens, wer die Reparatur initiiert (der*die Produzent*in der Problemstelle oder ein*e andere*r Teilnehmer*in), und zweitens, wer sie durchführt (auch hier kommen der*die Produzent*in der Problemstelle ebenso infrage wie ein*e andere*r Teilnehmer*in). Damit ergibt sich folgende schematische Darstellung (vgl. Tab. 3), angelehnt an Egbert (2009: 21) sowie Imo/Lanwer (2019: 176):

Tab. 3: Typologisierung von Reparaturen.

	Selbstdurchgeführt	Fremddurchgeführt
Selbstinitiiert	Selbstinitiierte Selbstreparatur	Selbstinitiierte Fremdreparatur
Fremdinitiiert	Fremdinitiierte Selbstreparatur	Fremdinitiierte Fremdreparatur

Neben der Frage der Initiierung spielt auch eine Rolle, *wann* im Verlauf der Konversation eine Reparatur initiiert bzw. produziert wird. Egbert (2009: 21) integriert dies in ihr Modell zur Unterscheidung verschiedener Reparaturtypen: Die Reparatur kann im Falle von Selbstinitiierungen noch in derselben TCU durchgeführt werden (selbst- oder fremddurchgeführt). Ist sie hingegen fremdinitiiert,

²⁴⁶ Siehe dazu insbesondere die in Kap. 3 vorgestellten Arbeiten von Krummheuer (2010) und Suchman (1987, 2007).

²⁴⁷ Siehe Kap. 2.2.2 und 6.1.2.

kann die Reparatur frühestens in der darauffolgenden TCU produziert werden. Selbstinitiierte Selbstreparaturen können auch in der dritten Position produziert werden (Third Position Repair, vgl. Schegloff 1987). In der vierten oder einer späteren Position können auch fremdinitiierte Fremdreparaturen nach einem weiteren Sprecher*innenwechsel auftreten (vgl. Schegloff 1992: 1323; Egbert 2009: 158–163). Dabei nimmt die Wahrscheinlichkeit einer Reparatur kontinuierlich ab, der „repair-initiation-opportunity-space“ (Schegloff/Jefferson/Sacks 1977: 375) wird „dünn“, je weiter sich die Problemquelle durch wiederholtes turn-taking entfernt“ (Egbert 2009: 163).

In die nachfolgenden Analysen von Reparaturen soll neben der Frage der Turn-Organisation ferner die Frage der Präferenzstrukturen einbezogen werden, die als möglicher Erklärungsansatz für das Auftreten sprachlicher Formen herangezogen werden kann. Wie Schegloff/Jefferson/Sacks (1977) schon in ihrem frühen Aufsatz zu Reparaturen feststellen, besteht seitens der Sprecher*innen eine hohe Präferenz für Selbstreparaturen. Die höchste Präferenz besteht für selbstinitiierte Selbstreparaturen, während für fremdinitiierte Fremdreparaturen die geringste Präferenz besteht (vgl. Pomerantz 1984; Selting 1987).²⁴⁸ Egbert (2009: 163) weist mit Verweis auf eine Arbeit von Kasper (1984) zu Reparaturen im schulischen Kontext darauf hin, dass diese Präferenzen im institutionellen Kontext überformt und sogar gänzlich aufgelöst werden können. Insofern ist es naheliegend, dass auch der soziotechnische Charakter des Dialogs zu einer Veränderung der Präferenzorganisation im Zusammenhang mit Reparaturen führen könnte.

6.1.4.1 Reparaturen in Mensch-Maschine-Dialogen

Reparaturmechanismen lassen sich – zunächst als vorläufiges Modell – auf Dialoge mit Smart Speakern projizieren. Einige Studien nehmen eine solche Anwendung konversationsanalytischer Grundlagen auf Smart-Speaker-Dialoge auch bereits vor, v. a. im Kontext von designorientierten Studien aus dem Schnittstellenbereich von HCI und Sprach- und Sprechwissenschaft, deren primäres Ziel es ist, Empfehlungen für eine Verbesserung der Geräte zu formulieren (vgl. Beneteau et al. 2019). Auch die bereits besprochene Studie von Porcheron et al. (2018) betrachtet Reparaturstrategien der Nutzer*innen. Dabei unterscheiden die Autor*innen zwischen der Anzeige und Identifikation eines Fehlers auf der einen und der Reparatur desselben auf der anderen Seite (vgl. Porcheron et al. 2018: 8–9). Wiederholungen, prosodische Veränderungen und lexikalische Anpassungen bzw. allgemeiner Reformulierungen werden hier als zentrale Reparaturstrategien der Nutzer*innen

²⁴⁸ Zum Präferenzkonzept, das auf Annahmen des Face-Work nach Goffman (1955; 1973) aufbaut, siehe Heritage (1984b) und Levinson (2000: 345–360) sowie Pomerantz/Heritage (2012).

herausgearbeitet. Ausgehend von der detaillierten Betrachtung zweier Beispiele wird in der Studie mit Blick auf das VUI-Design empfohlen, Reparaturvorgänge und den Status des Geräts transparenter zu machen und so z. B. in bisher beobachtbaren Phasen längerer Stille, die als Fehlerindikator von den Nutzer*innen gedeutet werden, eine Information über die derzeit laufenden Berechnungen oder Informationen einzuarbeiten (vgl. Porcheron et al. 2018: 10). Darauf aufbauend arbeitet auch die Untersuchung von Reineke/Helmer (2024), in der die Autorinnen einen mikro-longitudinalen mit einem konversationsanalytischen Ansatz kombinieren, verschiedene Praktiken auf Seiten der Nutzer*innen heraus, mit denen Reparaturen vollzogen werden. Die Autorinnen beobachten ganze und teilweise Wiederholungen der ursprünglichen Stimmeingaben, jeweils mit Modifikationen im prosodischen Bereich, z. B. Lautstärke, Sprechtempo, Akzentuierung, Intonierung, Pausen (vgl. Reineke/Helmer 2024: 9). Teilweise drückt die Analyse der Reparaturen aus, wo die Nutzer*innen die Problemquelle vermuten. Insgesamt konstatieren Reineke/Helmer (2024: 10) aber auch, dass aufgrund der der fehlenden Sensitivität für die lokale Sequenzhistorie im Austausch fragwürdig ist, ob überhaupt von konversationellen Reparaturen gesprochen werden kann. Sie argumentieren allerdings für die Verwendung des Begriffs als Analysefolie, weil sie Reparaturen aus der zwischenmenschlichen Interaktion durchaus ähneln und weil er die sprachlichen Praktiken der Nutzer*innen zur Reparatur sinnvoll einholen könne (vgl. Reineke/Helmer 2024: 10).

Durchaus auch konversationell ausgerichtet war die (allerdings auf HCI-Design ausgerichtete) Untersuchung von Pins et al. (2020: 9). In dieser identifizieren die Autor*innen den fehlenden Rückkanal – „Lösungs- und Feedback-möglichkeiten [sic] (bspw. durch Rückfragen)“ – als einen möglichen Ausgangspunkt für weitere Designanstrengungen. Garg/Cui/Kapadia (2021: 351:42) empfehlen „assisted self-repair“ auf Basis einer Untersuchung von Interviews mit Google Home-Nutzer*innen sowie Logfile-Daten, in denen insbesondere die Aneignungsprozesse von Kindern in den Blick genommen wurden. Sie empfehlen, dass die Smart Speaker nicht – wie Beneteau et al. (2019) es vorschlagen – selbst Angebote zur Reparatur machen, sondern stattdessen zunächst in Form von „unrelated responses“ (Garg/Cui/Kapadia 2021: 351:21) Antworten auf die *verstandenen* Fragen geben, um über diese Antworten anzuzeigen, welche Probleme bei der Verarbeitung vorgelegen haben könnten; Cho/Rader (2020: 33:12) hatten diesen Begriff in ihrer HCI-Studie zu Google Home eingeführt und halten diese Form der Antwort für hilfreicher als einen steuernden Eingriff.

Solche Diskussionen zeigen einmal mehr, dass eine genauere, empirische, kontextbasierte Betrachtung von VUI-Dialogen auch als Grundlage für die Weiterentwicklung der Geräte hilfreich sein kann und für den Gegenstandsbereich der Reparaturen auch erfolgreich angewendet wurde, um Designempfehlungen aus

einer linguistisch informierten Perspektive abgeben zu können. Sie untermauern die Relevanz der folgenden Analysen. Um als Grundlage für solche Studien zu dienen, muss sie jedoch für sich stehen und – ähnlich wie konversationsanalytische Forschung selbst – unabhängig von Designintentionen und praxeologisch die sprachlichen Prinzipien und Verfahren aufdecken, mit denen sich die Nutzung von VUIs in einem sozialen Kontext vollzieht.

Es kann auf einige Untersuchungen zugegriffen werden, die dies für andere Systeme bereits teilweise leisten. Für Embodied Conversational Agents betrachtet Krummheuer (2010: 236–261) in ihrer bereits eingeführten Arbeit Fehlschläge im Austausch mit dem virtuellen Agenten Max. Diese, so fasst sie zusammen, sind – anders als Reparaturen in zwischenmenschlichen Interaktionen – keine interaktiven Aushandlungsleistungen. Max ist aus diesen Reparaturvorgängen außerdem häufig „ausgeschlossen“ (Krummheuer 2010: 260) und die Reparatur vollzieht sich im „Meta-Interaktionsraum“ (Habscheid 2022: 168). Gleichwohl können erfolgreiche fremdinitiierte Selbstreparaturen beobachtet werden (vgl. Krummheuer 2010: 242–246); scheitern diese, bricht jedoch auf der Ebene des Praktischen die Fiktion einer Interaktion zusammen.

Bemerkenswert ist die Beobachtung, dass Bestandteile der praktischen Vollzüge im Zusammenhang mit der Reparatur nicht primär an den Smart Speaker adressiert sind, sondern auch an andere Zuschauer*innen und für diese ‚aufgeführt‘ werden. Krummheuer (2010: 244) spricht hier von „doing locating the trouble-source“, die z. B. multimodal durch ein Schieflegen des Kopfes, eine körperliche Annäherung an das Gerät oder mimische Ausdrücke realisiert werden, während parallel dazu Rückfragen formuliert werden. Die hohe Relevanz multimodaler Interaktionsbestandteile bei Fehlschlägen und die systematische, schrittweise Orientierung auf die anderen Personen und Rückorientierung auf die Maschine in solchen Situationen zeigen auch Gehle et al. (2015; 2017).

Diese Befunde werden in die folgenden Analysen von Reparatursequenzen einbezogen, mit dem Ziel, sprachliche Praktiken zu identifizieren, die dabei zum Einsatz kommen (können). Dazu wird hilfswise zunächst mit einem Modell gearbeitet, das auf der gezeigten Unterscheidung von Egbert aufbaut (vgl. Tab. 3). Bei einer Anwendung dessen auf Smart Speaker-Dialoge (vgl. Tab. 4) lassen sich insgesamt acht mehr oder weniger wahrscheinliche Fälle identifizieren, wobei jeweils unterschieden werden muss, wer im Dialog das Reparaturum²⁴⁹ produziert: das VUI

249 Damit ist die als „repairable“ (Schegloff/Jefferson/Sacks 1977: 363) identifizierte, als problematisch markierte Bezugsstelle (vgl. Selting 1987: 131) gemeint; demgegenüber ist vom „Reparans“ als der Äußerung die Rede, die das Reparaturum ersetzt durch das „was nun gelten soll“ (Bauer 2020: 347), d. h. die wesentliche Komponente der „Reparaturdurchführung“ (Pfeiffer 2015: 9).

oder der*die Anwender*in. Betrachten wir zunächst den ersten Fall, die Produktion eines Reparandums durch die Anwender*innen:

Tab. 4: Typologisierung von Reparaturen, Fallschema A.

	Selbstreparatur	Fremdreparatur
Selbstinitiiert	Fall AI: Anwender*in repariert eigene, an das VUI adressierte Äußerung	Fall AII: Anwender*in zeigt reparaturbedürftige Stelle in der eigenen Äußerung an, VUI repariert sie
Fremdinitiiert	Fall AIII: VUI zeigt reparaturbedürftige Stelle an, Anwender*in repariert sie	Fall AIV: VUI zeigt reparaturbedürftige Stelle an und repariert sie auch

Die Fälle AII und AIV sind aufgrund der bislang nicht programmierten Initiativkompetenz des Geräts im Hinblick auf solche konversationellen Phänomene unwahrscheinlich und im Korpus nicht nachweisbar, wobei sich mit Blick auf AIV Zweifelsfälle ergeben.²⁵⁰ AI und AIII hingegen scheinen denkbare Ausprägungen von Reparaturen im Dialog mit Smart Speakern zu sein, wobei wir sehen werden, dass AIII und BIII – die Frage also nach der Zuschreibung des Reparandums zu einem der beiden Beteiligten und der Identifikation dessen innerhalb des sequenziellen Verlaufs – nicht immer eindeutig geklärt werden kann.

Das Schema ist etwas anders aufgebaut, wenn die reparaturbedürftige Stelle durch das VUI produziert wird (vgl. Tab. 5). Der Fall BI scheint dabei äußerst unwahrscheinlich zu sein: Dass das VUI eine reparaturbedürftige Stelle selbst erkennt und repariert, ohne die Zuhilfenahme des menschlichen Gegenübers, ist nach dem, was wir in den bisherigen Kapiteln zu konversationellen Fähigkeiten der hier untersuchten VUIs gesehen haben, kaum vorstellbar.²⁵¹ Auch der Fall BII, bei dem das VUI den*die Anwender*in auf ein Reparans in der eigenen Äußerung hinweist und zur Reparatur einlädt, ist vor dem Hintergrund mangelnder metasprachlicher Fähigkeiten der Geräte unwahrscheinlich. Gleichwohl werden wir sehen, dass – rich-

²⁵⁰ Siehe dazu das Beispiel (36) („Timer-Namen“).

²⁵¹ Die bereits erwähnte Google Duplex-Technologie (siehe Kap. 3.2.1) arbeitet unter Zuhilfenahme von reparaturanzeigenden Diskursmarkern und selbstinitiierten Selbstreparaturen in der gleichen TCU. Dies dient primär der Simulation eines humanen Gegenübers in Telefongesprächen und kann somit als Beispiel für „Repair as a Vehicle for Other Actions“ (Couper-Kuhlen/Selting 2017: 115–116) gesehen werden, in denen nicht im engeren Sinne Reparaturleistungen vollzogen werden. Solche Funktionen sind nicht Teil der Geräte in der vorliegenden Untersuchung.

Tab. 5: Typologisierung von Reparaturen bei VUI, Fallschema B.

	Selbstreparatur	Fremdreparatur
Selbstinitiiert	Fall BI: VUI repariert eine eigene Äußerung	Fall BII: VUI zeigt reparaturbedürftige Stelle an, Anwender*in repariert sie
Fremdinitiiert	Fall BIII: Anwender*in zeigt reparaturbedürftige Stelle an, VUI repariert sie	Fall BIV: Anwender*in zeigt reparaturbedürftige Stelle an und repariert sie auch.

tet man den Blick auf die performative Seite eines VUI-Dialogs – es Turn-Abfolgen gibt, in denen ein solches Muster jedenfalls von den Beteiligten zugeschrieben wird. Die Fälle BIII und BIV treten häufiger auf, es ist jedoch, wie wir sehen werden, nicht trivial, diese sowohl voneinander wie auch von einigen Fällen im Schema A abzugrenzen, weil die Identifikation von Reparandum und Reparans nicht unbedingt eindeutig sein muss.

Wir können also einige Fälle ausmachen (vgl. Tab. 6), die wir uns nachfolgend genauer ansehen werden (sie werden in der Darstellung blau hervorgehoben), während einige Fälle schon auf Grundlage einer Datensichtung und theoretischer Überlegungen ausgeschlossen werden können (sie werden mit Asterisk markiert). Die Darstellung zeigt, dass die nicht-auftretenden Fälle in zwei Bereichen zu finden sind: Bei den eigenen Äußerungen fehlt dem VUI Kapazität zur Selbstinitiierung einer Reparatur, bei den Äußerungen der Anwender*innen fehlt die Kapazität zur Produktion einer Fremdreparatur. Demgegenüber sind VUIs in der Lage, eigene Äußerungen zu reparieren (wenn die Initiative dazu von den Anwender*innen ergriffen wird), und können außerdem Äußerungen der Anwender*innen als problematisch markieren, nicht aber eine Reparatur dazu eigenständig produzieren. Dies scheint – auf Basis der Vorüberlegungen – generell zu gelten; wir werden nachfolgend (wie bereits angekündigt) sehen, dass hierbei Ausnahmen und Zweifelsfälle bestehen können. Es sollen nun zunächst für die verbleibenden vier Kategorien einfachere Fälle vorgestellt werden, um davon ausgehend die Zweifelsfälle auch besser einordnen zu können.

Tab. 6: Typologisierung von Reparaturen, integrierte Darstellung beider Fallschemata (blau hervorgehoben sind näher zu beleuchtende Falltypen).

Selbstreparatur		Fremdreparatur	
Reparandum produziert durch Anwender*in (Fallschema A)			
Selbstinitiiert	Fall AI: Anwender*in repariert eigene, an das VUI adressierte Äußerung	* Fall AII: Anwender*in zeigt reparaturbedürftige Stelle in der eigenen Äußerung an, VUI repariert sie	
Fremdinitiiert	Fall AIII: VUI zeigt reparaturbedürftige Stelle in der Anwender*innen-Äußerung an, Anwender*in repariert sie	* Fall AIV: VUI zeigt reparaturbedürftige Stelle in der Anwender*innen-Äußerung an und repariert sie auch	
Reparandum produziert durch VUI (Fallschema B)			
Selbstinitiiert	* Fall BI: VUI repariert eine eigene Äußerung	* Fall BII: VUI zeigt reparaturbedürftige Stelle in der eigenen Äußerung an, Anwender*in repariert sie	
Fremdinitiiert	Fall BIII: Anwender*in zeigt reparaturbedürftige Stelle in der VUI-Äußerung an, VUI repariert sie	Fall BIV: Anwender*in zeigt reparaturbedürftige Stelle in der VUI-Äußerung an und repariert sie auch.	

6.1.4.2 Fallschema A: Reparaturbedürftige Äußerungen der Anwender*innen

Dazu eignet sich insbesondere das Schema AI: die selbstinitiierte Selbstreparatur, bei der sowohl das Reparandum wie auch das Reparans durch den*die Anwender*in produziert werden.

Beispiel (30): Spiel Spotify (01-CVR02-01)

007 LF: aLEXa?
008 (1.2)
009 AK: SPIEL de:n-
010 (0.6)
011 AK: spiel SPotify.
012 (1.9)
013 AL: HIER ist spotify.

Das Beispiel stammt aus Haushalt 1 aus der zweiten CVR-Erhebungsphase. Der Anwender Lukas äußert die Stimmeingabe „spiel SPOTify“ (011), die das VUI ratifiziert. Allerdings hatte Lukas zuvor unmittelbar nach der Adressierung einen bestimmten Artikel produziert („de:n-“, 009). Schon die Dehnung des Konsonanten zeigt an, dass eine Planungsschwierigkeit vorliegen könnte. Lukas bricht die Äußerung anschließend ab – sie ist weder pragmatisch noch prosodisch oder syntaktisch geschlossen – und initiiert damit eine Reparatur. Der Sprecher ersetzt daraufhin das Reparandum „de:n“ mit einer konkreten Anwendung, die als Reparans fungiert („SPOTify“, 011). Er übernimmt dabei die vorherige syntaktische Struktur aus 009 und orientiert sich dabei an der Grenze der Intonationsphrase. So gelingt es, die Initiierung, Identifizierung, Lokalisation und die Durchführung der Reparatur in einem Schritt zu vollziehen (vgl. Bauer 2020: 353). Das VUI kann diese Art der Reparatur offenbar verarbeiten, denn ein adäquater nächster Turn wird produziert, der von Lukas nicht beanstandet wird; gleichzeitig ist sie sehr ähnlich zum Reparaturablauf in Konversationen zwischen zwei Menschen.

Ein weiteres Beispiel für eine selbstinitiierte Selbstreparatur, das allerdings konzeptuell etwas anders gelagert ist, findet sich in Haushalt 8:

Beispiel (31): Siri/Google (08-CVR01-10)

```

001 SM: HEY siri;
002     (2.5)
003 SM: äh::m.
004     (0.6)
005 SM: NICH;
006     (0.6)
007 SM: HEY google;
008     (4.0)
009 SM: <<stimmhaft, nasal>h°>
010     (1.0)
011 SM: ÄHM <<stimmhaft>h°>
012     (0.2)
013 SM: LAUTstärke fünfzig prozent;
014     (1.3)
015     ((Musik wird laut abgespielt 1.3 Sek.))
016     ((Musik wird sehr leise abgespielt 4.4
017 SM: oke GOOgle;
018     (2.1)
019 SM: oke GOOgle;

```

- 020 (1.8)
 021 SM: LAUTstärke fünfzig prozent;
 022 (1.3)
 023 ((Musikwiedergabe mit Lautstärke wie in
 014, 35.0 Sek.))

Der Sprecher Samuel adressiert den Google-Assistant mit der Invokationsphrase „HEY siri“ (001).²⁵² Dieser reagiert allerdings nicht und Samuel äußert, nach der Produktion zweier reparaturanzeigender Diskursmarker (003 und 005) und längeren stillen Pausen, erneut eine Invokation (007). Dabei ersetzt er das Reparandum „siri“ durch das Reparans „google“, übernimmt aber die in die Summons-Sequenz integrierte Interjektion „hey“. Auffällig ist, dass anschließend eine längere Pause von über vier Sekunden entsteht, in denen der Sprecher Atemgeräusche produziert (009–011), die als Lachpartikeln gedeutet werden können (vgl. Jefferson 1979: 90; Werner 1983: 231; Merziger 2005: 66). Erst im Anschluss daran wird die Stimmeingabe produziert (013), doch dies wird – möglicherweise aufgrund der längeren Pause zwischen Summons-Answer-Sequenz und dem Folgeturn²⁵³ – offenbar nicht mehr durch das VUI prozessiert. Nach einer weiteren Pause wiederholt der Sprecher den Summons (017) und nimmt dabei erneut eine Reparatur vor: Nun wird die Anredeinterjektion „Hey“ (007) zum Reparandum und durch die im Google-Assistant alternativ mögliche Interjektion „oke“ ersetzt (017).

Es kann bereits hier diskutiert werden, ob es sich bei den in diesen Beispielen gezeigten Reparaturen (001/007 bzw. 007/017) um *selbstinitiierte* Selbstreparaturen handelt. Das VUI hat zwar jeweils keine verbale Reaktion gezeigt, aber das *Nicht-Gelingen* der Summons-Answer-Sequenz kann auch nicht als rein einseitiger Vorgang betrachtet werden (vgl. Schegloff 1968: 1083). Ausgelöst wurde die Reparatur also durch das *Ausbleiben* einer Reaktion des maschinellen Gegenübers. Bauer (2020: 382) klassifiziert genau dies als Form der unspezifischen Fremdinitiationspraktik zur Initiierung einer fremdinitiierten Fremdreparatur. Doch im Lichte des soziotechnischen Charakters dieses Dialogs muss diese Eingruppierung für den hier betrachteten Fall überdacht werden: Der Smart Speaker war zum Zeitpunkt des Ausbleibens der Reaktion noch nicht im Listening-Modus und hat insofern noch keinen Input aufgenommen und verarbeitet. Er wurde zwar von Samuel als maschineller Partizipand an der vollzogenen Praktik betrachtet, hat

²⁵² Samuel gibt an, dass er neben dem Google Assistant auch ein Smartphone von Apple besitzt, das er gelegentlich über das Apple-VUI „Siri“ steuert.

²⁵³ Siehe Kap. 6.1.1.

aber eben nicht „initiativ“ ein Nicht-Verstehen oder eine Reparaturbedürftigkeit der vorherigen Phrase signalisiert. Vielmehr hat der Anwender selbst aus der fehlenden Reaktion geschlossen, dass Reparaturbedürftigkeit vorliegt. Ähnlich gelagert ist der Fall auch im folgenden Beispiel, in dem der Smart Speaker adressiert wird, um ein bestimmtes Lied abzuspielen:

Beispiel (32): Billie Eilish (07-CVR01-24)

007 k: ((Musik stoppt))
 008 ((neues Lied setzt ein))
 009 KW: öhm HEY siri,
 010 spiele everything I WANTED von billie
 eilish.
 011 k: ((Musik spielt leise weiter, ca. 6 Sek.))
 012 KW: HEY siri,=
 013 =spiele billie Eilish everything I wanted.
 014 (2.4)
 015 SI: ich spiele everything i wanted von billie
 eilish JETZT.
 016 (0.9)
 017 k: ((Lied setzt ein))

Nach der Adressierung folgt nach sechs Sekunden Pause eine vollständige Wiederholung der an den Smart Speaker adressierten Äußerung, wobei die syntaktische Position von Titel und Interpret getauscht und die Präposition „von“ (010) getilgt wird. Das gesamte Segment 010 wird somit zum Reparandum, das Reparans ist mithin das gesamte Segment 013.

Diese Beispiele demonstrieren also zwei Dinge: Erstens zeigt sich in Beispiel (31) („Siri/Google“) bereits ein erster Zweifelsfall im Hinblick auf die Initiation der Reparatur – dies wird im weiteren Verlauf des Kapitels noch ausführlicher diskutiert. Zweitens wird in Beispiel (32) („Billie Eilish“) deutlich, dass eine große Einheit wiederholt wird, um die Reparatur zu realisieren. Darüber hinaus entstehen in den Beispielen schnell kaskadierende Reparaturleistungen. In diesen wird das Reparans der vorangegangenen Reparatur zum Reparandum der Folge-Reparatur (siehe Beispiel 31, „Siri/Google“).

Betrachten wir im Fallschema A den Fall AIII, die durch das VUI initiierte Reparatur der Äußerung der Anwender*in:

Beispiel (17): Timer vier Minuten (01-CVR01-11)²⁵⁴

075 LF: [aLEXa?]
 076 k: [((TV wird pausiert))]
 077 (0.9)
 078 LF: TImEr (0.3) vier minuten.=
 079 AL: =für WIE lange?
 080 (1.3)
 081 LF: <<artikulierte Sprechend, h> h° (-) alexa:
 (.) tImEr vier mINU:ten.>
 082 (1.0)
 083 AL: vier minuten (-) ab JETZT.
 084 (3.9)

Nach der Stimmeingabe des Anwenders (075) markiert das VUI eine Stelle der Äußerung („vier Minuten“) als problematische Bezugsstelle. Dabei fragt das VUI präzise nach der temporalen adverbialen Bestimmung in der elliptischen Konstruktion von Lukas. Ausgehend davon kann angenommen werden, dass das VUI zwar die gewünschte Funktion verstanden, die Spezifikation zur Dauer allerdings nicht als solche identifiziert hatte. Auffälligerweise produziert Lukas daraufhin die gesamte Stimmeingabe neu und schließt dabei auch das Aktivierungswort mit ein. Die Akzentuierung liegt dabei nun auf der Zeiteinheit („mINU:ten“, 078), die er durch eine Dehnung zusätzlich betont. Das Beispiel ist recht eindeutig zu klassifizieren: Die Reparatur wird klar durch das VUI initiiert, das Reparans durch Lukas produziert.

Allerdings stellen sich andere Fälle von Reparaturen weniger schematisch klar dar. Im folgenden Beispiel wird die Reparatur nicht durch eine Äußerung des VUI repariert, sondern ergibt sich aus dem, was das VUI im Anschluss „tut“ bzw. abspielt:

Beispiel (33): Lied-Ende (07-CVR01-49)

006 KW: hey SiRi,
 007 k: [((Klackern))]
 008 KW: [Titel noch]ma:l abspielen.
 009 (1.0)
 010 k: ((neues Lied setzt ein 4.3 Sek.))

254 Für eine Diskussion dieses Beispiels unter dem Aspekt der Sequenzialität siehe Kap. 6.1.2.

- 011 KW: hey siri VOR (.) riger titel.
 012 k: ((Musik endet))
 013 ((“Courage To Change” von Sia setzt ein
 und spielt fortlaufend; vereinzelte
 Haushaltsgeräusche 48.0 Sek.))

Der Anwender richtet an das VUI die Stimmeingabe mit der Aufforderung zur nochmaligen Wiedergabe (008). Dies jedoch scheitert, ein neues Lied wird stattdessen abgespielt und für ca. vier Sekunden wiedergegeben. Daraus ergibt sich die Notwendigkeit zur Reparatur der Äußerung: Statt des Befehls, den Titel „nochma:l“ (008) abzuspielen, muss nun in der Abfolge der Liedtitel ein Lied ‚zurückgesprungen‘ werden. Statt „nochmal“ wird hier nun also „VOR (.) riger“ (011) an die entsprechende Stelle zur Spezifikation des Titels produziert, zudem wird die zweite Äußerung ohne Verbalphrase produziert und infolgedessen auch in veränderter Satzstellung: Anstelle der Aufforderungsform mit Infinitiv (vgl. Zifonun/Hoffmann/Strecker 1997: 141; zu deontischen Infinitiven in der Interaktion siehe Deppermann 2006) steht nur noch der Titel mit seiner Spezifikation deutlich erkennbar im Nominativ. Dies zeigt, dass sich diese sprachliche Praktik bereits zu verfestigen begonnen hat, denn neben dem Wegfall des direktiv gebrauchten Infinitivs ist hier auch die Verwendung des Akkusativs getilgt, den dieser ansonsten als Objekt in der semantischen Rolle des Patiens erfordern würde (vgl. Zifonun/Hoffmann/Strecker 1997: 1310). Damit, so kann argumentiert werden, rückt das VUI als Agens in den Hintergrund: Weder dessen Handlung wird in Form einer Verbalphrase ausgedrückt noch seine semantische Rolle als Agens in Form der Kasuszugehörigkeit adressiert. Dass dies als Reparatur produziert wird, unterstreicht den Charakter der Verfestigung noch zusätzlich: Die verknappte Form wird im Fall des ‚Zurückspringens‘ – anders als beim Wiederholen – offenbar für aussichtsreicher gehalten.

Es lässt sich also anhand der Analysen von Reparaturen unter zwei Aspekten auch etwas zum Beteiligungsstatus des VUIs festhalten: Erstens scheint das vom Nutzer zugeschriebene Aktivitätsniveau gerade in der zweiten Stimmeingabe sehr gering – ein Ausdruck einer semantischen Rolle unterbleibt durch die Nominalphrase vollständig, der Smart Speaker tritt als Träger der Handlung in der Äußerung von Konrad gänzlich in den Hintergrund. Zweitens ist die ‚Beteiligung‘ des VUIs mit Hirschauer (2016: 49) passiv, aber nicht inhibitiv. Dies korreliert ebenfalls mit der verkürzten und um die Verbalkomponente gekürzten zweiten Formulierung der Stimmeingabe. Eine eindeutige Zuordnung in das Reparaturschema ist damit zwar noch immer nicht möglich, doch ermöglichte erst dieses Modell, die Unschärfe in der Zuordnung und der Unauflösbarkeit dieser Kategorien herauszuar-

beiten. Auch das nachfolgende Beispiel zeigt eine gewisse Unschärfe bei der Frage, wer die Reparatur initiiert:

Beispiel (34): Spiel Spotify! (01-CVR02-04)

173 LF: aLEXa?
174 (1.5)
175 LF: spiel MUSik.=
176 LF: =<<behaucht, p>(ja,>>
177 AL: HIER ist ein [rAdiosender (.) der
] dir vielleicht gefällt.
178 LF: [<<p> JA: den hab ich
gesu:cht;>]
179 AL: POP[hits (.) auf a]mazon music.
180 LF: [alexa STOP.]
181 LF: aLEXa?
182 (1.0)
183 LF: spiel SP0tify.
184 (1.1)
185 AL: ((beginnt Musik zu spielen))
186 AL: hier ist SP0tify.
187 k: ((Musik endet, Stille ca. 6.6 Sek.))
188 LF: aLEXa?
189 (0.8)
190 LF: <<akzentuiert> 'spIel ``!SP0T!ify.>
191 (1.3)
192 AL: HIER ist sp0tify;
193 (8.2)
194 LF: aLEXa;
195 spiel das LUMpenpack von spotify.
196 (1.2)
197 AL: THIS is das lUMpenpack von spotify.
198 ((spielt „Hauch mich mal an“ von Das
Lumpenpack ab))

Dabei richtet Lukas die Stimmeingabe „spiel MUSik“ (175) an das VUI, das mit der Empfehlung eines Radiosenders reagiert (177). Noch bevor das VUI diese Empfehlung komplett äußern kann, äußert Lukas den Stop-Befehl, um die Anforderung, Musik zu spielen, mit einer präziseren, auf eine einzelne Anwendung (Spotify) zugeschnittenen Formulierung zu wiederholen (183). Nachdem diese nicht bearbei-

tet wird, wiederholt Lukas den Befehl lexikalisch und syntaktisch identisch, lediglich seine Akzentuierung ändert sich und auch der Tonhöhenverlauf innerhalb des Segments verläuft nun mit stärkeren Ausschlägen (190). Erneut erfolgt keine hörbare Reaktion seitens des VUI, woraufhin nach ca. sieben Sekunden Pause Lukas die Spezifik des Befehls noch weiter erhöht und auch einen Interpreten ergänzt, der mit der Anwendung Spotify wiedergegeben werden soll (195); daraufhin wird ein Titel der Band „Das Lumpenpack“ abgespielt (198). Lukas repariert also in mehreren Schleifen seine ursprüngliche Stimmeingabe; dabei lässt sich feststellen, dass er die Spezifik der Anfrage in zwei Schritten steigert: von dem allgemeinen Befehl „spiel muSIK“ (175) zur Angabe einer Anwendung, mit der dies realisiert werden soll („spiel spotify“, 183 und, leicht variiert, 190) hin zur Präzisierung des Interpreten („spiel das LUMpenpack von spotify“, 195). Die Stimm-eingabe bleibt syntaktisch mit dem Gebrauch der Imperativform von „spielen“ in Verbspitzenstellung identisch, der Sprecher tauscht und ergänzt jedoch einzelne Lexeme innerhalb dieser Struktur; zudem verändert er insbesondere die Akzentuierung und den intrasegmentalen Tonhöhenverlauf. Ausgelöst werden die drei Reparaturen bzw. Reparaturversuche jeweils durch unerwünschte oder ausbleibende Äußerungen des VUI.

Es kann diskutiert werden, ob es sich bei den gezeigten Beispielen im engeren Sinne um Reparaturen handelt und wie diese zu klassifizieren sind. Beispiel (34) kann – so wie hier zunächst platziert – als fremdinitiierte Selbstreparatur gelesen werden. Die Produktion des Reparandums würde dann dem Anwender zugewiesen und das Abspielen des falschen Liedes als Reparaturinitiiierung gelesen werden. Dafür spricht die angedeutete Reparaturbeziehung zwischen den Segmenten 175 und 183 bzw. zwischen 183, 190 und 195, die bei letzteren als projektionserhaltende Substitution bzw. Insertion gedeutet werden können (vgl. Pfeiffer 2015: 102–105), bei ersteren eine projektionsverändernde Tilgung in Verbindung mit einer Substitution (vgl. Pfeiffer 2015: 122–123). Einzuwenden wäre allerdings, dass das VUI keineswegs verbal oder nonverbal expliziert, dass eine Reparaturnotwendigkeit vorliegt und nicht auf ein Reparaturandum hinweist. Möglich wäre insofern auch eine Lesart als *selbst*initiierte Selbstreparatur, in der das Abspielen des falschen Titels (oder eine andere, als nicht gewünscht markierte Äußerung des VUI) nicht als Reparaturinitiiierung verstanden würde. Dann allerdings läge die Initiiierung bei den Anwender*innen, was mit Blick auf den sequenziellen Verlauf nicht richtig ist: Die Notwendigkeit zur Reparatur zeigt sich erst, als das Spielen des nicht gewünschten Titels beginnt.

Erscheint dies zunächst als modellhaft-begriffliche Frage, kommt diesem Komplex allerdings mit Blick auf die Fragestellung nach dem Gesprächsbeteiligungstatus und den sprachlichen Praktiken, mit denen dieser realisiert wird, eine beträchtliche Bedeutung zu: Verstehen wir die Aktion eines Smart Speakers als

Initiierung einer Reparatur innerhalb eines sequenziellen Gefüges, wäre dies durchaus ein anderer Status von ‚Beteiligung‘ und somit auch Agency, als wenn der alleinige Initiator der Reparatur der Anwender wäre. Bei der Klärung hilft eine praxeologische Perspektive auf den Vollzug. Dabei zeigt sich, dass die Reparatur ko-operativ vollzogen wird: Das VUI ist – ganz eindeutig – an der Reparatur ‚beteiligt‘. Es trägt allerdings – um hier erneut das Schema von Aktivitätsniveaus nach Hirschauer (2016: 49)²⁵⁵ anzuwenden – weniger eine aktiv-proaktive Rolle, sondern vielmehr eine durchaus proaktive (nicht inhibitive), aber passive Rolle: Es wird „etwas an[ge]stoßen“, ein (anderes) „Handeln ,veranlass[t]“ (Hirschauer 2016: 49), was durch den Anwender aufgegriffen und dann aktiv als Reparaturnotwendigkeit interpretiert wird. Dabei greift der Anwender also nicht nur seine eigene vorhergehende Äußerung auf, sondern auch die Involvierung bzw. Nicht-Involvierung des VUI, die für die Anwender*innen nicht vorhersehbar sind.

Ein letztes Beispiel aus dem Bereich der Äußerungsreparaturen, bei denen die Reparanda im Bereich der Äußerungen der Anwender*innen liegen, soll zeigen, wie die Technizität des Austauschs auch Reparaturvorgänge prägt:

Beispiel (35): Timer -Namen (01-CVR01-10)

115 LF: aLEXa?
 116 (0.2)
 117 LF: timer AU:F-
 118 (1.4)
 119 LF: ä::hm ZWANzig minuten.
 120 (3.8)
 121 LF: aLEXa?
 122 (1.0)
 123 LF: timer auf ZWANzig minuten.
 124 (1.7)
 125 AL: zwAnzig minuten ab JETZT;
 126 [(39.5)]
 127 k: [((leise Musikwiedergabe, vereinzelt
 Hintergrundgeräusche))]
 128 LF: aLEXa?
 129 (1.5)
 130 LF: TImeR mInus fünf minuten;
 131 (1.7)

255 Siehe Kap. 2.2.4.

132 AL: zweiter timer (.) fünf minuten (.) ab
 JETZT==
 133 =möchtest du diesem timer einen namen
 GEben?
 134 (0.2)
 135 LF: alexa (.) lösche TImeR eins und zwei;
 136 (1.3)
 137 AL: du hast keinen TImeR mit der bezeichnung
 eins zwei,
 138 aber es gibt TImeR für fünf minuten und
 zwAnzig minuten.
 139 (0.5)
 140 LF: aLEXa?
 141 (0.8)
 142 LF: LÖSCH: (.) tImeR (.) für fünf minuten;
 143 (1.8)
 144 AL: fünf miNUten timer wurde abgebrochen.
 145 (0.6)
 146 LF: alexa LÖSCH tImeR für zwanzig minuten.
 147 (1.3)
 148 AL: zwanzig minuten TImeR wurde abgebrochen;
 149 (0.2)
 150 LF: aLEXa?
 151 (0.2)
 152 LF: TIme:r fünfzehn minUten;
 153 (0.5)
 154 AL: fünfzehn minuten ab JETZT;

Zunächst stellt Lukas einen Timer auf dem Smart Speaker ein (115–125); ein Vorgang, der nicht ohne erneute Reparatur zu Beginn abläuft. Beispiele dieser Art wurden bereits betrachtet, einschließlich des Phänomens, dass eine Zögerung zu Beginn – hier eine stark gedehnte Zögerungspartikel („äh::m“, 119) und die längere Pause zwischen *Intent* und *Slot* (1,5 Sekunden) – dazu geführt haben könnten, dass das VUI den Listening-Modus beendet hat. Nach ungefähr 40 Sekunden innerhalb der Timerlaufzeit will Lukas den Timer verkürzen. Dies steht im Einklang mit seiner eigenen Zögerung zu Beginn des Ausschnitts, in dem er bei der Spezifikation der Dauer unsicher zu sein schien. Mit dem Befehl „timer MINus fünf minuten“ (130) fordert er das VUI auf – das bestätigt auch der weitere Verlauf des Ausschnitts –, fünf Minuten von der ursprünglich eingestellten Zeit zu subtrahieren. Das VUI interpretiert die Anfrage allerdings als Versuch, einen

zweiten Timer zu stellen (132), und fragt nach, ob die Timer benannt werden sollen (möglicherweise damit diese in Abfragen zum Timerstand oder bei Abbruchnotwendigkeiten leichter identifiziert werden können). Statt dieser Aufforderung nachzukommen, produziert Lukas eine Stimmeingabe, um beide Timer zu löschen. Bis hierher wird der Ablauf nicht als Reparatur gewertet, denn bisher weisen die Äußerungen noch keine Störungen auf der Ebene des konversationellen Ablaufs auf. In den nun folgenden Äußerungen allerdings kommt es genau dazu: Die von Lukas genutzten Bezeichnungen „eins und zwei“ (135) werden durch das VUI als eine Bezeichnung für einen einzigen Timer gedeutet. Der Konnektor „und“ wird dabei in der Wiederholung getilgt (137) – und dies wird durch das VUI als reparaturbedürftige Stelle markiert, jedenfalls wenn es zu einer Folgeaktion durch den Smart Speaker kommen soll, denn eine Aktion wird hierbei nicht durchgeführt. Lukas interpretiert dies auch als Aufforderung zur Reparatur und produziert – und dies ist der bemerkenswerte Teil dieser Analyse – zwei projektionserhaltende Substitutionen (142 und 146), bei denen „eins und zwei“ durch die Spezifizierung „für fünf minuten“ und „für zwanzig minuten“ ersetzt wird. Dabei übernimmt Lukas den Reparaturvorschlag des VUI („für fünf minuten und zwanzig minUten“, 138), sodass sich der Verlauf hier – so könnte argumentiert werden – im Grenzbereich zur fremdinitiierten Fremdreparatur bewegt; Bauer (2020: 388) spricht von einer „Übergangsform“. Neben der Verschleierung und der Face-Arbeit, die Bauer als Begründung für diese Vorgehensweise vermutet, ist jedoch die Prägung durch die Technizität einzubeziehen, die nicht nur die Aufteilung der Reparatur auf insgesamt drei Sequenzen notwendig macht (weil die Zusammenfassung zu einer einzigen Stimmeingabe zunächst scheiterte, vgl. 135), sondern auch die explizite verbale Bestätigung der auszuführenden Aktion durch den Nutzer. Wenn die Reparaturennotwendigkeit gerade aus der Unklarheit über die genau durchzuführende Aktion resultiert, kann die Reparatur nicht selbst durchgeführt werden, und – so steht weiter zu vermuten – da der Austausch mit VUIs sich tendenziell aktionsbezogen darstellt, sind eindeutig fremdinitiierte Fremdreparaturen, die ohne die Ratifizierung durch die Anwender*innen auskommen (Fall BIV), allenfalls in diesem „Übergangsbereich“ (Bauer 2020: 388) zu finden.

6.1.4.3 Fallschema B: Äußerungen des VUI

Im vorhergehenden Kapitel konnte bereits festgestellt werden, dass die Zuordnung, ob die reparaturbedürftige Äußerung durch das VUI oder durch den*die Anwender*in produziert wird, nicht immer eindeutig zu klären ist. Betrachten wir dazu noch einmal einen Auszug, der Beispiel (34) („Spiel Spotify!“) nicht unähnlich ist:

Beispiel (36): Radio/Spotify (01-CVR02-02)

004 LF: aLEXa?
 005 (1.2)
 006 LF: <<all> spiel muSIK;>
 007 (0.7)
 008 AL: HIER ist ein rAdiosender-=
 009 [=der dir vielLEI-]
 010 LF: [alexa STOP.]
 011 (0.5)
 012 LF: alexa spiel SP0tify.
 013 [(4.7)]
 014 k: [((vereinzelte Hintergrundgeräusche))]
 015 LF: aLEXa?
 016 (0.7)
 017 LF: spiel SP0tify;
 018 (1.0)
 019 AL: HIER ist sp0tify;

Lukas produziert als imperativisch formulierte Stimmeingabe die Aufforderung, Musik zu spielen (006), woraufhin das VUI einen Radiosender vorschlägt (008). Obwohl das Beispiel aus der zweiten Phase der Erhebungen stammt – zu diesem Zeitpunkt der Smart Speaker also bereits fünf bis sechs Monate in Gebrauch ist –, scheint dies also ein wiederholt auftretender Fehler zu sein.²⁵⁶ Auf der sprachlichen Oberfläche ist die Reparatur als Selbstreparatur der Stimmeingabe von Lukas zu klassifizieren: Zwar produziert das VUI eine von Lukas als falsch interpretierte Äußerung, was durch den sofortigen Stop-Befehl mit *barge-in* (010) deutlich wird. Doch Lukas repariert nicht die Äußerung des VUI, sondern die *eigene* Äußerung (012) durch eine projektionserhaltende Substitution (vgl. Pfeiffer 2015: 102). Man könnte also von einer ko-operativ erzeugten, bidirektional gerichteten Reparatur sprechen: Das Reparandum ist auf der sprachlichen Ebene das Lexem „muSIK“ (006); die Reparatur zielt also konversations- bzw. sequenzanalytisch betrachtet darauf ab, das Lexem „muSIK“ (006) zu reparieren und es zu ersetzen. Gleichzeitig zielt die Äußerung darauf ab, eine andere Reaktion des VUI hervorzurufen, ist also pragmatisch auch mit den Wiedergaben dessen verbunden. Lukas

²⁵⁶ Auch in der ersten Phase der Erhebung findet sich ein solches Beispiel im Korpus, in dem der Befehl, „Musik zu spielen“, mit dem Hinweis auf den Radiosender durch das VUI beantwortet wird, was eine Reparatur zur Folge hat.

nimmt zwar keinen direkten Bezug auf die Äußerungen des VUI – diese initiieren aber mittelbar erst die Reparatur. Dies erzeugt ein konversationelles Spannungsfeld zwischen der fremdinitiierten Selbstreparatur auf der einen und der Gerichtetheit dieser Reparatur auf die Äußerungen des Geräts auf der anderen Seite. Das folgende Beispiel soll dieses Spannungsfeld noch einmal illustrieren:

Beispiel (37): Hangover (07-CVR01-27)

- 003 KW: hey SIri spiele happier than ever von
billie eilish-
- 004 k: ((Stille, leises Klappern 2.3 Sek.))
- 005 SI: ich habe HANGOver von billie eilish auf
apple mUsic nicht gefUnden-
- 006 KW: hey siri spiele happier than Ever (.) von
billie eilish;
- 007 (2.3)
- 008 SI: ich spiele jetzt (happiest) von billie
Eilish;
- 009 (2.3)
- 010 SI: ((spielt Happier Than Ever von Billie
Eilish, 149.4 Sek.))

Das Grundmuster dieser Reparaturleistung ist bereits aus den in Kap. 6.1.4.2 präsentierten Beispielen bekannt: Der Anwender Konrad äußert eine Stimmeingabe (003), das VUI interpretiert diese falsch (005). Die Stimmeingabe wird anschließend repariert (006); hier lässt sich die Reparatur als „Wiederholung ohne Revision der projizierten syntaktischen Struktur“ (Pfeiffer 2015: 97) identifizieren. Auch hier richtet sich die Reparatur aber darauf, eine andere Stimmausgabe des VUI zu evozieren. Das Reparaturandum ist in diesem Sinne nicht ein Bestandteil der Äußerung von Konrad, sondern die gesamte Äußerung des VUI, die wiederum eine spezifische Problemstelle hervorhebt, die sich aus der Abweichung von der Stimmeingabe ergibt („HANGOver“, 005). Die Äußerung von 005 fordert eine Reparatur geradezu ein (006), sodass die Lesart als (durch die Äußerung des VUI) fremdinitiierte Selbstreparatur sequenziell plausibler erscheint, doch auch hier ist pragmatisch gesehen die Reparatur darauf gerichtet, eine andere Stimmausgabe des VUI zu evozieren.

Abschließend soll die potenzielle Verwobenheit der zu reparierenden Stellen und der Reparaturen noch einmal anhand einer längeren Reparatursequenz vorgestellt werden:

Beispiel (38): Schottischer Dolch (03-CVR01-10)²⁵⁷

013 BW: oKAY google?
 014 (2.5)
 015 BW: WIE hEIßt der schOttische dOlch?
 016 (2.4)
 017 GA: tut mir LEID.
 018 (0.2)
 019 GA: ich habe dich nicht verSTANDen;
 020 (0.7)
 021 BW: okay GOOgle?
 022 (1.7)
 023 BW: wie (.) hEIßt (.) dEr (.) schOttische
 DOLCH?
 024 (2.7)
 025 GA: entSCHULdigung.
 026 (2.0)
 027 ich habe dich nicht verSTANDen;
 028 (2.3)
 029 BW: Okay;
 030 GOOgle?
 031 (1.5)
 032 BW: WIE heißt==
 033 =ein (.) dOlch (.) in (-) SCHOTTland?
 034 (2.6)
 035 GA: auf englisch heißt DAS-
 036 a GERman in scOtland.
 037 (2.0)
 038 BW: so ein QUATSCH.

Die Nutzerin Beate repariert im vorliegenden Beispiel konversationell gesehen mehrfach ihre eigenen Äußerungen: Die Stimmeingabe „WIE hEIßt der schOttische dOlch?“ (015), die als Abfrage von Hintergrundwissen verständlich wird, wenn man einbezieht, dass die Nutzerin zuvor eine Fernsehdokumentation u. a. zur Geschichte Schottlands geschaut hat, wird zunächst unter Beibehaltung der

²⁵⁷ Eine Analyse unter anderen Gesichtspunkten findet sich auch bei Hector/Hrncal (2024); das Beispiel stammt nicht aus den Fokushaushalten 1, 7 und 8, sondern aus Haushalt 3, illustriert aber besonders eindrücklich die kaskadierenden Reparaturfolgen und wird daher hier dennoch präsentiert.

syntaktischen Struktur ein erstes Mal wiederholt (023). Anschließend wird sie unter Veränderung der syntaktischen Struktur ein zweites Mal wiedergegeben (032–033). Bei der zweiten Wiederholung entkoppelt Beate das Subjekt „dOlch“ von der Attribuierung „schottisch“. Letztere wird stattdessen als lokale präpositionale Bestimmung ans Ende der Äußerung ‚geschoben‘ und nicht mehr adjektivisch, sondern als Nominalphrase ausgedrückt. Daraufhin interpretiert das VUI die Anfrage als Übersetzungswunsch (035–036), was Beate negativ ratifiziert (038).

Diese Lesart, die am Fallschema A orientiert ist, lässt sich jedoch auch vollständig umkehren: Dann würden auch hier nicht die Äußerungen von Beate (015, 023) als reparaturbedürftig verstanden, sondern die Äußerungen des VUI (017–019, 025–027); sodann wäre das Beispiel eines aus dem Fallschema B. Denkbar sind auch weiter verflochtene Lesarten: Ist zunächst die Äußerung von Beate (015) die durch das VUI als problembehaftet markierte Stelle, wird dies in der Folgeäußerung von Beate ‚zurückgespielt‘ und die Äußerung des VUI wird als die problematische Äußerung markiert, die sofort wieder durch das VUI als Reparaturum gerahmt wird.

6.1.4.4 Zwischenfazit Reparaturen

Auch wenn sich eine Reihe von Studien insbesondere aus dem Bereich der HCI damit beschäftigt hat, wie Reparaturnotwendigkeiten in Dialogen mit Maschinen entstehen, welche Reparaturstrategien angewendet werden und wie sie sich auch in sozialen Gruppen etablieren (für eine Übersicht siehe Beneteau et al. 2019: 2–4), ist bislang nicht darauf abgestellt worden, was dies für die praktische Gestaltung der Konversationen bedeutet.

Reparaturen gelten als „Paradebeispiel dafür, dass die interaktive Sinnproduktion in Gesprächen auch im Extremfall eines ‚Unfalls‘ geordnet und systematisch abläuft bzw. wiederhergestellt wird“ (Stoltenburg 2012: 2). Die oben gezeigten Beispiele lassen jedoch nur zwei Rückschlüsse zu: Entweder stimmt die von Stoltenburg mit Bezug auf die Arbeiten von Sacks, Schegloff, Jefferson und Goodwin formulierte Behauptung nicht, dass auch bei ‚Unfällen‘ in Gesprächen alles „geordnet“ und „systematisch“ abläuft. Nicht nur kann in den beiden zuletzt gezeigten Beispielen kein ‚Erfolg‘ im Hinblick auf das im Dialog vollzogene „kommunikative Projekt“ (Linell 2012) erreicht werden, auch in den zuvor gezeigten Ausschnitten ist eben die sequenzielle Ordnung gefährdet: Die Teilnehmendenstruktur und die ‚Personenfiktion‘ kann nicht länger aufrecht erhalten werden, die Äußerungen werden seitens der Nutzer*innen auf nahezu allen sprachlichen Beschreibungsebenen an die Skript-Erwartungen des Gegenübers angepasst – kaum anders als schon von Suchman (1987) beschrieben. Es braucht zur Reparatur teilweise einen hohen sprachlichen Aufwand und eine Wiederholung vollständiger Sequenzen, die schon in ihrer

Ursprungsformulierung durchaus auf erwartete Erwartungen ausgerichtet waren. Eine Ausnahme stellen lediglich die selbstinitiierten Selbstreparaturen dar – was nur konsequent ist, weil hier das VUI gar nicht als Partizipand an der Reparatur einbezogen wird.

Nun wäre zu fragen, ob dies nicht eben eine Form von Recipient-Design ist, wie es ja in der Konversationsanalyse als ein anderes Universalium jeder Konversation positioniert wird: der Annahme von Sacks/Schegloff/Jefferson (1974: 727), dass konversationelle Äußerungen immer so gestaltet sind, dass sie „display an orientation and sensitivity to the particular other(s) who are co-participants“. Deppermann (2014: 16) beschreibt Recipient-Design als „one of the most important grounds for situated choice of grammatical constructions and lexical items in talk“. Doch ist dies zwar – und hier kommen wir zurück auf die bereits diskutierte Dialoggestaltung²⁵⁸ – mit einem Bachtin’schen Dialogbegriff sicherlich Teil der „other-orientation“ (Linell 2009: 13), einem wesentlichen Element des dort angewendeten Dialogizitätsverständnisses: die Ausrichtung auf ein Gegenüber welcher Art auch immer (siehe Bachtin 1981; 1986: 64–71) und die damit verbundene Adressatenorientierung, die – liest man Linell (2009: 167) in einer weiteren Auslegung – durchaus auch Adressierungen von Menschen an Maschinen und umgekehrt umfassen könnte: „Every act is addressed to somebody, whether this addressee is individual or collective, real or imaginary, being another person (or group) or an aspect of one’s own self“. Recipient-Design aber – so führt Deppermann (2014) weiter aus – ist als Konzept deutlich enger gefasst und nicht auf eine dialogische Orientierung im Reden reduzierbar. Dies richtet sich in Abgrenzung zu den vorgestellten Auffassungen von ‚Dialog‘ nicht auf ein sehr weites Feld von „others“, sondern auf die (konkreten) Beteiligten in einer interaktionalen Begegnung und wird gemeinsam hervorgebracht. Als Beispiele führt Deppermann (2014: 18) u. a. Höflichkeitsdesign und „shared knowledge“ sowie die interaktive Herstellung eines „common ground“ an. In den Reparaturen zeigt sich, dass – auch wenn von den Teilnehmenden sicherlich phasenweise der Versuch unternommen wird, sich an diesen Konzepten zu orientieren – die Orientierung daran schnell aufgegeben und die Ko-Produktion dieser Merkmale nicht fortgesetzt wird, während dies bei zwischenmenschlichen Interaktionen entweder nicht vorkommt oder mindestens schwere Irritationen hervorruft und langfristig beziehungsbedrohend wirken kann (vgl. Brown/Levinson 1987; Locher/Watts 2005; Ehrhardt 2018).

Die andere mögliche Schlussfolgerung (mit Bezug auf die oben zitierte Äußerung von Stoltenburg), die hier vertreten werden soll, ist also: Die eingangs zi-

²⁵⁸ Siehe Kap. 6.1.2 sowie Kap. 2.2.2 und 2.2.3.

tierte These Stoltenburgs zur Geordnetheit von Sinnproduktion in Gesprächen stimmt, doch es handelt sich bei VUIs *nicht* um stabile Interaktionspartner*innen bzw. Gesprächsteilnehmer*innen.²⁵⁹ Am Beispiel der Reparaturen zeigt sich gleichwohl, was auf Basis der Analysen von Anredeformen, Sequenzialität und Turn-Taking bereits angenommen werden konnte: Das VUI und insbesondere die damit verbundene „Persona“ (z. B. Alexa) wird sprachlich durchaus – im Rahmen enger Grenzen und v. a. auf bestimmten formalen Ebenen – als Konversationspartner*in behandelt und folgt auch im sprachlich-praktischen Vollzug Regeln, die für zwischenmenschliche Interaktionen beschrieben worden sind. Diese bleiben jedoch nicht stabil: Im Falle von Störungen zeigt sich deren schnelle und häufige Verletzlichkeit. Anhand des „generic problems“ (Schegloff 2006: 77–79) der Reparaturleistungen kann also erneut illustriert werden, dass VUIs nur vorübergehend wie Interaktionsbeteiligte behandelt werden: Der Status wechselt und wird immer wieder praktisch neu ausgehandelt – sie sind am Vollzug der sozialen Praxis ‚beteiligt‘, die sich auch des konversationellen Charakters der Geräte als Ressource bedient. Sie werden dadurch aber nicht zu Gesprächsteilnehmer*innen.

Auch mit Blick auf die damit verbundenen sprachlichen Praktiken hilft der Blick auf den praktischen Vollzug der Smart Speaker-Nutzung: In den Analysen wurde von zwei konzeptionellen Seiten her kommend – dem Fallschema A, in dem Problemstellen als Hervorbringung der Anwender*innen betrachtet wurden, und Fallschema B, in dem sie dem VUI zugeordnet wurden – festgestellt, dass eine enge Verwobenheit der Smart Speaker-Äußerungen mit den Äußerungen der Anwender*innen besteht. Die Reparaturen werden zwar konversationell von den Nutzer*innen vollzogen – die VUIs zeigen keine oder nur schwache Orientierungen auf eine (potenzielle) Reparatur. Initiiert werden die Reparaturen somit von den Nutzer*innen – aber in Reaktion auf die Äußerungen des VUI. Dies hat auch Konsequenzen auf der Ebene der sprachlichen Realisierung der Reparaturen: Die ungleichen Wahrnehmungsmodalitäten, das vergleichsweise niedrige Inkrementalitätspotenzial der Smart Speaker und die daraus folgende notwendige Produktion und Rezeption von syntaktisch und pragmatisch abgeschlossenen Sequenzen führen häufig zu Wiederholungen, in denen prosodische, lexikalische und syntaktische Elemente substituiert werden. Etwa das Reparaturverfahren der Substitution ist hier im Vergleich zu Reparaturen bei zwischenmenschlichen Interaktionen (siehe Pfeiffer 2015) deutlich unpräziser; sprachliche Formen wie die Tilgung der dem Smart Speaker zugeschriebenen semantischen Rolle des Agens (Beispiel 33, „Lied-Ende“) unterstreichen diesen Eindruck. Weiteren Aufschluss zu dieser Frage sowie zum Aspekt der Gesprächsbeteiligung wird die gesprächsorga-

²⁵⁹ Hier bestätigen sich die theoretischen Überlegungen aus Kap. 2.2.1 in der Empirie.

nisatorische Einbindung des Smart Speakers in andere laufende Konversationen geben, die – besonders mit Blick auf Fehlschläge – in Kap. 6.2.4 genauer untersucht wird.

6.1.5 Zwischenfazit zu dyadischen VUI-Dialogen

Die Analysen zeigen, dass sich sowohl die Nutzer*innen wie auch die VUI-Dialoggestaltung an den „generic organizations of practice central to the conduct of interaction“ (Schegloff 2006: 71) orientieren, d. h. kommunikativen Problemen, die durch die sprachlichen Äußerungen gelöst werden sollen. VUI-Dialoge sind insofern sequenziell organisiert und es lassen sich zwei Muster einer Basis-Sequenzstruktur feststellen, die Nutzer*innen und VUI gemeinsam verfertigen und von denen ausgehend Erweiterungen in Form von Einschüben und Expansionen stattfinden können. Bei Äußerungen, die konversationelle Implikaturen enthalten und für menschliche Interpretation insofern Kohärenz aufweisen, können VUIs aufgrund dieser Strukturorientierung hingegen überfordert werden: Implementiert sind für die Anwender*innen vorhersehbare Sequenzstrukturen, die bei der Verarbeitung von Äußerungen, die im sequenziellen und situierten Kontext verständlich werden, nicht greifen (vgl. auch Habscheid 2022: 191). Es zeigt sich dabei, dass die Nutzer*innen im Verlauf der Domestizierung des VUIs Strategien zu entwickeln scheinen, mit denen sie die Technizität der VUIs bearbeiten und in die konversationellen Muster einpassen. Damit entsteht eine spezifische Form von Recipient-Design, wie es Sacks/Schegloff/Jefferson (1974: 727) in ihrem konversationsanalytischen Klassiker beschreiben:

With ‚recipient design‘ we intend to collect a multitude of respects in which talk by a party in a conversation is constructed or designed in ways which display an orientation and sensitivity to the particular other(s) who are the co-participants. (Sacks/Schegloff/Jefferson 1974: 727)

Im Unterschied zu sozialer Interaktion, für die Recipient-Design hier beschrieben wurde, sind allerdings in VUI-Dialogen einige Voraussetzungen nicht gegeben: Eine zeitlich parallel verlaufende Produktion und Rezeption sowie eine Reziprozität in der Wahrnehmung besteht, so kann konstatiert werden, in VUI-Dialogen nicht; dies steht im Einklang mit Befunden von Krummheuer (2010), die am Beispiel von Embodied Conversational Agents zeigen konnte, dass die Voraussetzungen für Interaktion im Sinne Goffmans (1983: 3) nicht durchgängig gegeben sind. Insbesondere der geteilte Aufmerksamkeitsfokus und die wechselseitige Wahrnehmungswahrnehmung – zwei Goffman zufolge wesentliche Grundpfeiler sozialer Interaktion und folglich auch für die Gesprächsgestaltung zentrale Vorausset-

zungen – müssen immer wieder neu hergestellt werden und bleiben nicht konstant bestehen. Recipient-Design als Gestaltungsprinzip sprachlicher Äußerungen basiert allerdings auf fortlaufend (und nicht nur aus der Interaktionshistorie, sondern auch aus situativ) aktualisierten Annahmen über das Gegenüber. Dabei handelt es sich um einen inkrementellen Prozess. Zwar passen auch die VUIs bis zu einem gewissen Grad die Stimmausgaben auf Grundlage von Annahmen über das (menschliche) Gegenüber an, doch sind diese Anpassungen limitiert und haben keine unmittelbaren Auswirkungen auf die konkrete sprachliche Gestaltung der synthetisierten Äußerungen. Zugleich sind die Voraussetzungen für die Entwicklung der Annahmen und Ziele im Dialog mit VUIs anders gelagert als in zwischenmenschlichen Interaktionen: Die „Intransparenz des Fremdseelischen“, von der Deppermann/Blühdorn (2013: 9) sprechen, ist aufgebrochen, denn über die Verarbeitungsfähigkeiten und mögliche Funktionen geben Anleitungen, Erklärvideos und Dialoge in Ersteinrichtungen²⁶⁰ sowie fortgesetzte Tests konkreten Aufschluss. Zwar sind die genauen technischen Abläufe bis zu einem gewissen Grad opak, aber anders durchdringlich als ein subjekthaftes Gegenüber. Es kann insofern von einer speziellen, einseitigen Form von „Adressatenzuschnitt“ (Deppermann/Blühdorn 2013: 8) bei Stimmeingaben gesprochen werden, bei dem sich Nutzer*innen in der praktischen Gestaltung ihrer Äußerungen an Annahmen über die Fähigkeit des VUIs in Verbindung mit ihren kommunikativen Zielen orientieren.²⁶¹

Das hat Folgen für die sprachlich vollzogene soziale Praxis: Es zeigt sich, dass die Aufmerksamkeitsfokussierung über wiederholte Summons-Answer-Sequenzen immer wieder neu hergestellt werden muss. Der sequenzielle Ablauf ist in hohem Maße durch die beschriebene Basis-Sequenzstruktur konfiguriert (wenngleich nicht determiniert). Das Turn-Taking weist Besonderheiten auf, die auf ungleiche Voraussetzungen bei der Bestimmung von TRPs zurückzuführen sind, die nicht wie in zwischenmenschlichen Interaktionen ‚ausgehandelt‘ werden. Reparatursequenzen umfassen größere Einheiten, weil diese Struktur immer wieder neu gestartet werden muss und nur eingeschränkt projiziert bzw. retraktiert werden kann. Damit ein kommunikatives Projekt gelingt, müssen die Nutzer*innen also ihre sprachlichen Eingaben so an das VUI anpassen, dass auf diese Spezifika des Dialogs Rücksicht genommen wird und dies mit ihren Erwartungen kombinieren (vgl. Habscheid 2022: 192), während VUIs nur in geringem Maße Anpassungsleistungen voll-

²⁶⁰ Siehe Kap. 6.2.2.1.

²⁶¹ Zum Verhältnis zwischen Vorannahmen und kommunikativen bzw. interaktiven Zielen im Recipient-Design siehe Hitzler (2013: 127).

ziehen. Nutzer*innen können zudem bei Problemen die kommunikativen Projekte sanktionsfrei beenden.

Im Austausch zwischen Mensch und Maschine kann also eine ‚neue‘ oder jedenfalls von Mensch-Mensch-Interaktionen teilweise abweichende Form von Dialogizität (vgl. Lotze 2020; Habscheid 2022) entstehen. Ein spezifisches sprachlich-strukturelles Register („Computer Talk“), wie es zunächst von Zoeppritz (1985) vorgeschlagen und anschließend von Krause/Hitzenberger (1992) aufgegriffen wurde, konnte in verschiedenen Untersuchungen nicht nachgewiesen werden – die dazu einschlägigen Studien trägt Lotze (2016: 160–161) zusammen (siehe auch Lotze 2025): Weder konnte eine Konsistenz bei den sprachlichen Äußerungen einzelner oder eine Übereinstimmung bei denen verschiedener Sprecher*innen gezeigt werden, noch waren homogene sprachliche Eigenschaften oder Form-Funktions-Gefüge nachweisbar (vgl. Fischer 2006: 74). Stattdessen zeigten sich in den verschiedenen Untersuchungen stark voneinander abweichende Ergebnisse (vgl. Fischer 2006: 5). Fischer (2006: 149) hebt hervor, dass der Austausch zwischen Mensch und Maschine in seiner sprachlichen Gestaltung von zu vielen verschiedenen Faktoren sowohl auf Seiten der Maschinen wie auch des Menschen beeinflusst wird – und eine Aushandlung darstellt, die nicht erlaubt, ein sprachliches Register für diese Situation im engeren Sinne festzustellen. Auch die bisher vorgebrachten Analysen liefern darauf keine Hinweise.

Gleichwohl postuliert Fischer, dass sich Nutzer*innen (kognitiv) an dem orientieren, was sie vermuten, was das maschinelle Gegenüber ‚kann‘ bzw. wo Herausforderungen liegen: „the peculiarities of human-robot dialogue are best understood as users’ strategies to deal with what they understand the challenges of the situation to consist in [...]“ (Fischer 2011: 31; siehe auch Ernst 2017). Nutzer*innen übertragen also nicht ‚einfach‘ Strategien von der Mensch-Mensch-Interaktion auf den Austausch mit der Maschine – so argumentiert Fischer in Abgrenzung zu Nass/Moon/Green (1997) und Nass/Moon (2000), die eine vorbewusste oder gedankenlose („mindless“) Übertragung sozialer Regeln auf die Mensch-Maschine-Interaktion konstatieren.²⁶² Diese Argumentation bringen die Autor*innen als Gegenvorschlag zu Anthropomorphisierungen vor, ausgehend von einer Konzeptualisierung von Anthropomorphisierungen als „belief [of individuals, T.H.] that computers are essentially human“ (Nass/Moon 2000: 93) – dies sei zwar nicht der Fall, allerdings, so argumentieren Nass/Moon weiter, hätten Nutzer*innen für dialogische Situationen mit nicht-menschlichen Akteur*innen eben keinen Erfahrungsschatz und bedienen sich insofern an Skripten der sozialen Interaktion, die in zwischenmenschlichen

²⁶² Für eine differenzierte Kritik dieses Paradigmas und daran anschließende konzeptionelle Vorschläge auf Basis umfassender Forschungsergebnisse siehe auch Lotze (2025).

Begegnungen eingeübt wurden. Nicht nur scheint das Argument des fehlenden Erfahrungsschatzes angesichts der weiten Verbreitung entsprechender Geräte nicht mehr gültig. Hinzu kommt, dass solche Argumentationsmuster aus dem „Computers are Social Actors“ (CASA)-Forschungsparadigma nur auf die Dialoge selbst fokussiert sind und die lokale und situationale Entfaltung der sozialen Praxis ausblenden, in der Äußerungen weitaus mehr als nur Ausdruck einer gedankenlosen Übertragung sein können (vgl. Reeves/Porcheron 2023: 585). Vielmehr, so argumentieren Clark/Fischer (2022) am Beispiel von humanoiden Robotern, behandeln Nutzer*innen die maschinellen Gegenüber als Darstellungen oder Abbilder (englisch *depictions*) von sozialen Akteur*innen – in spezifischen Situationen. Kurz gesagt: Die Nutzer*innen präsentieren auf der sprachlichen Oberfläche den Status eines ‚Beteiligten‘, weil sie – ähnlich wie bei Handpuppen, Fernsehfiguren (vgl. Keppler 1996) oder anderen interaktiven *Darstellungen* von Menschen – geradezu dazu einladen, diesen interaktiven Gestaltungsspielraum zu nutzen. Dies findet jedoch nur im Rahmen der durch diese Fiktion gesetzten Grenzen und zudem mit Bezug zur situativen Umgebung statt. Letzteres wird in Mehrparteienkonstellationen besonders deutlich, wenn formal an ein VUI adressierte Stimmeingaben doppelt funktionalisiert werden und nicht nur an das VUI selbst, sondern auch an die übrigen Beteiligten gerichtet sind und in dem Wissen produziert werden, dass diese die Eingabe hören können (vgl. Habscheid 2022: 192). Der Adressatenzuschnitt hat dann mehrere Dimensionen – und das ‚Funktionieren‘ im Kontext eines VUI-Dialogs muss nicht die wichtigste davon sein, wenn dieser eingebettet in einen sozialen Kontext mit mehreren Beteiligten stattfindet. Auf diese Thematik soll detailliert im zweiten Analyseabschnitt eingegangen werden.

6.2 VUI-Dialoge in Mehrparteienkonstellationen

Im Zentrum der bisherigen Analysen standen dyadische VUI-Dialoge, die nur in Ausnahmefällen mit zwischenmenschlichen Interaktionen verwoben waren. Diese Analysen wurden u. a. mit dem Zweck durchgeführt, eine Grundlage zu schaffen, auf der Mehrparteienkonstellationen eingehender betrachtet werden können. Damit ist nämlich sowohl für soziale Interaktionen wie auch für den Dialog zwischen Anwender*innen und VUIs die sprachliche Gestaltung basaler Organisationsprinzipien beschrieben. Bei Letzteren muss die Einschränkung berücksichtigt werden, dass diese explorativen Charakter haben – gleichwohl lassen sich die Einsichten aus den durchgeführten Untersuchungen heranziehen, um nun ‚gemischte‘ Situationen zu beleuchten.

In Mehrparteienkonstellationen treten Mensch-Mensch-Interaktionen und VUI-Dialoge parallel, sich überkreuzend und potenziell überlagernd auf. Einsichten zu diesem Zusammenspiel von zwischenmenschlichen Interaktionen und VUI-Dialogen in Mehrparteienkonstellationen liefern insbesondere Porcheron et al. (2018).²⁶³ Während die Einbettung in *Alltagsaktivitäten* den Beteiligten mehr oder weniger mühelos gelinge, sei die konkrete Integration der Bedienung des Smart Speakers über das VUI in *Alltagsgespräche*, den fortlaufenden *turn-by-turn talk*, eine Herausforderung für die Nutzer*innen – etwa im Hinblick auf Adressierungen, die Ko-Produktion von Stille zur Formulierung einer Eingabe, die das VUI verarbeiten kann, ebenso wie im Umgang mit den Ausgaben des VUI. Des Weiteren beobachten Porcheron et al. (2018) Umformulierungen (sowohl syntaktisch wie lexikalisch) als Praktiken zur Reparatur von Störungen, ebenso wie *gemeinsame* Bearbeitungen solcher. Solche Reparaturstrategien beschreiben auch Beneteau et al. (2019) – verbunden mit der Feststellung, dass in diesen Mehrparteienkonstellationen die Reparatur teilweise so schwierig ist, dass reparaturbedürftige Stellen zum Ausschluss des VUI als ‚Gesprächspartner‘ führen können.

Die Betrachtung der Mehrparteienkonstellationen erfolgt im Folgenden explorativ unter der Fragestellung, wie Smart Speaker am Vollzug der Praxis ‚beteiligt‘ werden. Im Hinblick auf die Konzeptualisierung von Beteiligung stütze ich mich auf die bereits ausführlich dargelegte Idee der materiellen Partizipanden des Tuns nach Hirschauer (2004; 2016).²⁶⁴ Demzufolge können Dinge und entsprechend auch Geräte als materielle Partizipanden Teil eines Praxisvollzugs werden: Um etwa eine Wettervorhersage über einen Smart Speaker zu hören, gehört das VUI ebenso zur Praxis wie eine Zahnbürste zum Zähneputzen oder ein Telefon zum Telefonieren (vgl. Hirschauer 2016: 51). VUI-Dialoge sind dabei, wie wir gesehen haben, durch zwei Merkmale charakterisiert: Der Vollzug gestaltet sich erstens sequenziell im Rahmen der *Basis-Sequenzstruktur*²⁶⁵ und zweitens akustisch.²⁶⁶ Wie Arbeiten zu Interfaces etwa von Brennan (1990) oder Arminen (2005)

²⁶³ Siehe dazu ausführlich Kap. 3.2.4.

²⁶⁴ Siehe dazu Kap. 2.1.1.

²⁶⁵ Siehe Kap. 6.1.2.1.1.

²⁶⁶ Zu den Begriffen „akustisch“ und „mündlich“ siehe Hector (2022: 199). Demzufolge ist der Begriff *mündlich* auf „von Menschen mit dem Mund produzierte Äußerungen“ bezogen, während *akustisch* sich auf jede „Ausbreitung und Wahrnehmung von Schall“ beziehe: „Die Verwendung des Begriffs ‚mündlich‘ ist daher vom Menschen produzierten Äußerungen vorbehalten; die Verwendung des Begriffs ‚akustisch‘ zeigt im Gegensatz dazu die maschinelle Beteiligung an. ‚Akustisch‘ wird insofern auch verwendet, wenn vom Zusammenspiel mündlicher (d. h. menschlicher) Äußerungen auf der einen Seite mit vom Smart Speaker gesendeten akustischen Signalen bzw. akustischen sprachlichen Zeichen auf der anderen Seite die Rede ist, etwa bei akustisch prozessierten Bedienweisen“ (Hector 2022: 199).

zeigen, weist zwar auch die Bedienung visueller Interfaces durch ihre sequenzielle Strukturiertheit Ähnlichkeiten zu zwischenmenschlichen Konversationen auf. Zugleich prägen aber die Eigenschaften gesprochener Sprache, die Auer (2000: 44–46) herausstellt – Flüchtigkeit, Irreversibilität und eine Produktions-Rezeptions-Synchronisierung –, auch das sequenzielle Gestaltungspotenzial akustisch produzierter Äußerungen. Auer (2000: 47) exemplifiziert dies anhand der „syntaktische[n] Grundoperationen“ der Projektion und der Retraktion, doch auch andere Merkmale der sequenziellen Gestaltung gehen auf diese Eigenschaften zurück. So ist etwa die Länge von Stimmeingaben zeitlich bis zu einem gewissen Grad limitiert, weil eine fehlerfrei produzierte und übertragene Äußerung aufgrund der hohen Störanfälligkeit gerade in Mehrparteienkonstellationen mit zunehmender Zeit unwahrscheinlicher wird und auch die menschliche Fähigkeit zur Planung und Produktion spontaner Äußerungen (ohne Abbrüche, die der Smart Speaker als Turn-Übergabe identifizieren würde) begrenzt ist. Die akustische Prozessierung hat außerdem zufolge, dass die Bedienung des Interfaces für ko-präsente Beteiligte hörbar wird und sich die geplante Bedienung so über einen gewissen Zeitraum ankündigt – vor der eigentlichen Stimmeingabe steht die Invokation, teilweise auch das kurze Warten auf einen Wechsel des VUI in den Listening-Modus.²⁶⁷ In einem visuellen Interface können eine Berührung oder ein Klick hingegen unvermittelt erfolgen und benötigen keine Invokation, außerdem ist schon die Sehfläche etwa auf Smartphones im Regelfall nicht für andere Beteiligte zugänglich, sondern muss zugänglich gemacht werden, was von verschiedenen kommunikativen Verfahren begleitet werden kann (vgl. Oloff 2019; König/Oloff 2019; Keppler 2019).

Die Kombination aus Sequenzialität und Akustizität rückt den VUI-Dialog phasenweise in die Nähe von Gesprächen (vgl. Englert/Hoffmann/Waldecker 2022), was auch in Werbespots insinuiert wird (vgl. Dickel/Schmidt-Jüngst 2021) und wie sich auch beim Abgleich der dyadischen VUI-Dialoge mit gesprächsorganisatorischen Prinzipien zeigt: Zwar weisen VUI-Dialoge gewisse Abweichungen auf, ähneln aber durchaus z. B. institutionell überformten, empraktischen Interaktionen. Abgesehen von ersten Erkenntnissen aus o. g. Arbeiten, die v. a. annehmen lassen, dass Smart Speaker-Nutzung in Mehrparteienkonstellationen spezifische Merkmale aufweisen könnte (Porcheron et al. 2018; Beneteau et al. 2019; Beneteau et al. 2020a; Etzrodt 2022), sind die wenigen bisherigen Arbeiten mit gesprächsanalytischem Fokus aber auf dyadische Konstellationen bezogen. Wenn also vom ‚Beteiligungsstatus‘ die Rede ist, ist damit die ‚Beteiligung‘ an der sozialen Praxis im Sinne Hirschauers auf der formalen und funktionalen Ebene der Äuße-

267 Siehe dazu Kap. 6.1.1.1.

rungen sowohl der Anwender*innen als auch der ko-präsenten Sprecher*innen und der VUIs gemeint. Zur begrifflichen Klärung hilft hier Goffmans bereits erwähnte Unterscheidung zwischen „talk“ und „a talk“ (Goffman 1979: 6) – in der Übersetzung „Reden“ und „Gespräch“ (Knoblauch/Leuenberger/Schnettler 2017: 44).²⁶⁸ Nicht jedes Reden konstituiert Goffman zufolge ein Gespräch, Reden kann auch Teil anderer Praktiken sein. In VUI-Dialogen vollziehen sich dem Charakter nach ebenfalls nicht-gesprächshafte Praktiken,²⁶⁹ wie wir sie bereits kennengelernt haben (z. B. Timerstellen, Musikauswahl und -steuerung, Informationsgewinnung) – und zwar auf konversationell gestaltete Weise und durch sprachliche Äußerungen, die an der Oberfläche stellenweise Gesprächen ähneln.

Im Kontext der Medienrezeptionsforschung sind Reden und Gespräche u. a. beim Fernsehen und ‚mit‘ Fernsehfiguren als Para-Interaktion beschrieben worden.²⁷⁰ Während die Kommunikation ‚mit‘ einem Fernseher nur in eine Richtung möglich ist (vgl. Holly/Baldauf 2001: 60), sind VUIs in dieser Hinsicht bidirektional: Die Beteiligung der Anwender*innen an Konversationen ist nicht nur möglich, sondern vorgesehen und notwendig, um die auszuführenden Handlungen erfolgreich zum Abschluss zu bringen.²⁷¹ Während also ein Teil der Mechanismen dem fernsehbegleitenden Sprechen an der Oberfläche ähneln könnte – dies wird in den folgenden Analysen genauer beleuchtet –, sind VUI-Dialoge grundsätzlich anders gelagert: VUIs sind nicht die Medieninhalte, sondern akustisch verfertigte Schnittstellen zur *Bedienung* des Geräts, über das dann bestimmte Medieninhalte wiedergegeben werden können, die sich teilweise (z. B. bei Wettervorhersagen) in die VUI-Struktur einpassen.

Die vorliegenden Daten zeigen Mehrparteieninteraktionen in beiden Datentypen: erstens in den Situationen der ersten Inbetriebnahme und Einrichtung durch zwei bzw. drei Mitglieder des Haushalts und zweitens in den CVR-Aufnahmen,

²⁶⁸ Siehe auch Kap. 2.2.1.

²⁶⁹ Goffman (1979: 14–16) spricht von „nonlinguistic contexts“, der in der Übersetzung (Knoblauch/Leuenberger/Schnettler 2017: 56) als „nichtsprachlicher Kontext“ wiedergegeben wird, doch der Begriff scheint (im Original und in der Übersetzung) etwas unscharf: Goffman betont in seiner Argumentation nicht die Freiheit des Äußerungskontexts von jeder Sprachlichkeit, sondern den Umstand, dass – mit Schatzki (1996) – soziale Praktiken vollzogen werden, die nicht primär ein Gespräch sind. Ein nicht den Vollzug eines Gesprächs, sondern auf die Durchführung der anderen Praktik gerichtetes Reden kann und wird in vielen Fällen die soziale Praktik aber begleiten.

²⁷⁰ Siehe dazu Kap. 2.3.1.

²⁷¹ Einschränkung muss gesagt werden, dass es durchaus möglich ist, einen Smart Speaker ausschließlich über das grafische Interface einer Smartphone-App zu steuern und nicht von den VUIs Gebrauch zu machen. Solche Nutzungsfälle sind jedoch in der vorliegenden Arbeit nicht untersucht worden.

wenn (anders als in Kap. 6.1) mehrere Personen (Haushaltsmitglieder oder Gäste) an der Nutzung des Smart Speakers ‚beteiligt‘ sind. Nachfolgend werden zunächst VUI-Dialoge präsentiert, die sich begleitend zu Mensch-Mensch-Interaktionen vollziehen. Diese werden im Hinblick auf ihre jeweilige Entfaltung und Verwobenheit sequenziell analysiert und so verschiedene Typen der Einbindung herausgestellt (Kap. 6.2.1). Daran anschließend werden kurz verschiedene Situationen charakterisiert und exemplifiziert, in denen in den vorliegenden Daten auf unterschiedliche Weise VUIs temporär zu ‚Beteiligten‘ an Gesprächen werden (Kap. 6.2.2): In den gezeigten Situationen scheinen sich, so die Arbeitshypothese, zwei Arten des Redens miteinander zu verbinden: das Reden zur Verfertigung eines VUI-Dialogs mit dem Reden im Gespräch über die Smart Speaker. Schließlich werden die sprachlichen Praktiken herausgearbeitet, deren Einsatz konstitutiv für die (Nicht-)‚Beteiligung‘ durch die untersuchten Äußerungen ist (Kap. 6.2.3). Die Befunde werden abschließend zusammengefasst und kurz diskutiert (Kap. 6.2.4).

6.2.1 Sequenzielle Einbindung von VUI-Dialogen

Das folgende Kapitel zeigt, wie soziale Interaktionen und VUI-Dialoge miteinander in Beziehung stehen bzw. in Beziehung gebracht werden (oder eben nicht), wenn sie in unmittelbarer zeitlicher Nähe bzw. gleichzeitig auftreten. Dabei ergeben sich aus den untersuchten Daten heraus verschiedene Typen: von unkommentierten, uneingeleiteten VUI-Dialogen, die während einer sozialen Interaktion auftreten, bis hin zu VUI-Dialogen, die sich mit Sequenzen aus der sozialen Interaktion überkreuzen. Die sequenzanalytische Perspektive wird hier dazu genutzt, die Trennung von sozialer Interaktion und VUI-Dialog zu hinterfragen und die praktische Herstellung dieser Kategorien nachzuvollziehen.

6.2.1.1 Uneingeleitete Einbindung

Das folgende Beispiel zeigt zunächst einen Fall aus der Wohngemeinschaft von Konrad (KW) und Till (TW) aus Haushalt 7:

Beispiel (39): Filmbericht (07-CVR01-54)

- 016 KW: also dadurch war für sie klar das dauert
ZEHN minuten dann steht ihr das es_ah_k
auf der tür,
017 TW: JA_a,
018 KW: vor der TÜR,

019 (0.8)
 020 KW: und wird äh MITbekommen dass-
 021 (0.7)
 022 KW: ä:hm grad en Officer in gefAhr ist-
 023 und (.) dann (-) KOMmen die halt einfach;
 024 TW: hm_MH,
 025 (1.0)
 026 KW: U:ND-
 027 wenn das es_eh_KA das macht,
 028 werden die die nicht einfach erSCHIEssen.
 029 (0.3)
 030 TW: näh STIMMT,
 031 KW: WEIsste,
 032 [WEIL-]
 033 TW: [JA.]
 034 KW: die wissen ja oke wir können nicht
 einfach_n (.) Officer erschießen,
 035 TW: HM_mh,
 036 KW: und dann fängt auch m (-) MAson (-) mäi
 mäi maiden;
 037 (0.7)
 038 KW: auch Ä:HM;
 039 (1.1)
 040 KW: Ä:HM-
 041 (0.2)
 042 KW: nich sein ZIEL er[füllen.]
 043 k: [((Weckton setzt ein))]
 044 (0.3)
 045 KW: HEY siri stop.
 046 [(0.8)]
 047 k: [((Weckton endet))]
 048 KW: ÄHM;
 049 (0.2)
 050 KW: dass WER einfach,
 051 (0.4)
 052 KW: ähm die erSCHIEssen kann das wär dann
 nicht mehr möglich,
 053 TW: mh_HM,
 054 (0.4)
 055 KW: der hat aber noch weitere FÜNF polizisten

oder so aufm bÜckel;

056 TW: AH_ja.

Konrad erzählt Till in dem Ausschnitt den Handlungsverlauf eines Films, den er kürzlich gesehen hat. Die Darstellung von Konrad weist typische Merkmale einer größeren Diskurseinheit – ähnlich zu einer Erzählung – auf, in der der *turn-by-turn talk* für einen längeren Zeitraum suspendiert und das Rederecht an einen „primären Sprecher“ (Quasthoff 1980b: 127) übertragen wird, während die Beteiligten einer oder mehreren anderen Personen interaktiv hervorgebracht die Rolle von Zuhörer*innen zuschreiben, was wiederum mit erwartbaren Rezeptionsäußerungen verknüpft ist (vgl. Quasthoff 1981; 2001). Die Sprechenden bringen so die Rollen ‚Erzähler*in‘ und ‚Zuhörer*in‘ wechselseitig hervor und verfertigen so eine „interaktive Gesprächsstruktur“ (Quasthoff 2001: 1300). Solche sequenziell organisierten, „übersatzmäßige[n] Diskurseinheiten“ (Quasthoff 2001: 1297; siehe auch Rehbein 1984; Hoffmann 1984; Hausendorf/Quasthoff 1996) weisen regelmäßig Formen auf, mit denen die Beteiligten gemeinsam aus dem *turn-by-turn talk* in den Vollzug der Erzählung überleiten (und zurück). Während Konrad in der Rolle als primärer Sprecher die Filmhandlung erzählt bzw. von seinen Seheindrücken berichtet, bestätigt Till diese Rolle durch regelmäßige *continuers* (vgl. Quasthoff 1981: 299–301; siehe auch Schegloff 1982; Couper-Kuhlen/Selting 2017: 512–514): Er äußert Partikeln wie „JA_a“ (017) und „hm_HM,“ (024) und andere kurze Äußerungen (vgl. 030), die den Fortgang des Berichts bzw. der Erzählung durch Konrad ratifizieren.

Während dieser Diskurseinheit ertönt der Signalton des zuvor über den Smart Speaker eingestellten Weckers (042). Der Weckton überlappt kurz mit einer Äußerung von Konrad (042). Konrad hält das Wecksignal mit einer kurzen Spracheingabe an, die prosodisch aus einer einzigen Intonationsphrase und neben der Invokation nur aus dem freistehenden „Stop“-Befehl besteht. Die Unterbrechung wird weder zuvor noch danach zwischen Konrad und Till relevant gesetzt. Konrad setzt die Erzählung nach Produktion einer Zögerungspartikel und einer kurzen Pause fort (048–049), wobei die Zögerungspartikel nicht für die Unterbrechung spezifisch zu sein scheinen, da Konrad diese bereits unmittelbar vor dem Beginn des Wecktons durch das VUI ebenfalls produziert hatte (038–041). Die Beteiligungskonstellation zwischen Konrad und Till mit den etablierten Rollen als Erzähler und Zuhörer wird, wie die folgenden Äußerungen der beiden zeigen, durch die VUI-Unterbrechung nicht bedroht: Konrad schließt mit pronominalen Wiederaufnahmen an seine vor der Unterbrechung geäußerten Teile der Erzählung an (etwa „die“, 052) und Till bestätigt auch weiterhin die Rolle als Zuhörer durch Rezeptionssignale in Form von Partikeln (053, 056). Die Bedienung des VUI, die sich mit der Interaktion zwischen Till und Konrad überlagert und kurzzeitig überlappt, wird also zwar zeitlich in den sequenziellen Verlauf eingebunden und

findet ebenfalls auf dem akustischen Kanal statt (dazu siehe auch Hector 2022: 199–200). In der stabilen Beteiligungskonstellation der aktuell durchgeführten primären Aktivität von Konrad und Till (der Erzählung) ist dies allerdings für die Sprecher*innen ein Nebenschauplatz – der Umstand, dass Konrad, der die Stimm-eingabe produziert, begünstigt die unproblematische Einbindung der Insertion in den *turn-by-turn talk*. Das VUI ist, mit Hirschauer (2016: 51), einer der „Träger“ der Praxis, wenn es um die Praktik des ‚Alarmierens‘ geht und unterbricht kurzzeitig die Praktik des Erzählens oder Berichtens – wird dadurch aber nicht zum Beteiligten dieser nebenher vollzogenen Praktik. Die ‚Einbindung‘ erfolgt auf der akustischen Oberfläche und zeigt sich durch die kurze Unterbrechung, die Tiefenstrukturen der vollzogenen Praktiken überlappen sich jedoch nicht.

Ein ähnliches, aber im Detail gleichwohl anderes Bild ergibt sich im folgenden Beispiel aus Haushalt 1, der Wohngemeinschaft von Alex (AK) und Lukas (LF):

Beispiel (40): Kartoffeln (01-CVR01-03)²⁷²

034 LF: h°
 035 ((schnieft zweimal))
 036 (1.6)
 037 LF: h°
 038 aLEXa?
 039 (1.4)
 040 LF: wie ist der TImerstand?
 041 (1.4)
 042 AL: du hast KEIne benachrichtigungen-
 043 ((Signalton))
 044 (0.2)
 045 AL: ((Signalton))
 046 (0.5)
 047 LF: [alexa STOP-]
 048 AL: [((Signalton))]
 049 k: ((vereinzelt Poltergeräusche, ca. 4.7
 Sek.))
 050 LF: <<leise sprechend> K0misches ding ey->
 051 °h
 052 [(2.8)]

272 Ein Teil dieses Ausschnitts wurde unter dem Aspekt der Anrede als Beispiel (9) („Du hast keine Benachrichtigungen“) diskutiert.

- 053 k: [((Hintergrundgeräusche))]
 054 LF: wie sieht_s mit den karTOffeln aus?
 055 AK: (ja/was?)
 056 LF: wie sieht_s mit den karTOffeln aus?
 057 (1.3)
 058 AK: gucke GRAdE,
 059 (1.9)
 060 AK: joa?

Der Timer, nach dessen aktuellem Stand Lukas hier fragt (040), wurde im Kontext der Lebensmittelzubereitung eingestellt. Dies ergibt sich aus der Aufnahme, die unmittelbar vor diesem Ausschnitt aufgezeichnet wurde, und in der zu hören ist, wie der Timer auf 20 Minuten gestellt wird. Der Zusammenhang zwischen den beiden Aufnahmen wird deutlich, weil ein Abstand von ca. 20 Minuten dokumentiert ist. Die Geräuschkulisse (Küchengeräusche, insbesondere Zerkleinern von Lebensmitteln, Geschirrkloppern und Spülen) lässt darauf schließen, dass ein länger andauernder Vorgang (z. B. Kochen oder Schmoren) begonnen wurde. Anschließend wird in der Zwischenzeit offenbar – so zeigen Teile der Aufnahme, aus der der obige Ausschnitt entnommen wurde – Fernsehen geschaut.²⁷³ Die Abfrage von Lukas erfolgt kurz bevor der Signalton zum Ende des Timers dann tatsächlich nach 20 Minuten einsetzt (043) und Lukas diesen beendet (047). Die dabei entstehende, dicht aufeinanderfolgende Wiedergabe der Sprachausgabe „du hast KEine benachrichtigungen“ (042), die zudem nicht zur gestellten Frage nach dem Timerstand passt, sowie die Wiedergabe des Signaltons verursachen eine schnelle und durchaus komplizierte Überlagerung von Vorgängen. Diese kommentiert Lukas mit einer entsprechenden Äußerung zum Charakter des VUI („KOMisches ding ey-“, 050). Unmittelbar im Anschluss fragt Lukas seinen Mitbewohner Alex nach dem Zustand der Kartoffeln (054–056), mit dem der gestellte Timer mutmaßlich in Verbindung steht, was Alex mit entsprechenden Rückmeldungen beantwortet (058–060).

Es ergibt sich hier also keine Einbindung in einen *turn-by-turn talk*, sondern vielmehr die Einbindung einer der Funktionen des VUI in die Praktik der Lebensmittelzubereitung. Dabei kann auch hier davon gesprochen werden, dass das VUI einer der „Träger“ (Hirschauer 2016: 51) dieser Praktik ist: Es weist auf das Ende einer bestimmten Zubereitungsphase hin. Diese hatte Lukas schon selbst antizipiert (wie sich an seiner Nachfrage zeigt) und der unklare Verlauf des VUI-Dialogs könnte

273 Durch gelegentliches Atmen wird auch belegt, dass wenigstens einer der Anwender auch währenddessen im Raum ist, obgleich damit noch nicht gesagt ist, dass auch tatsächlich auf den zu hörenden TV-Inhalt geachtet wurde, was aber für die Analyse auch nicht zentral ist.

sodann die Rückfrage bei seinem Mitbewohner ausgelöst haben, um eine Verifikation über den tatsächlichen Stand zu erhalten, die das VUI bzw. generell eine Timer-Funktion (ähnlich zu einer Eieruhr) nicht liefern kann. Das VUI wird daran allerdings übergangslos nicht weiter beteiligt: Schon in der nicht eindeutig adressierten Äußerung „KOMisches ding ey-“ (050) von Lukas ist von dem VUI als „ding“ die Rede, was erstens semantisch darauf hinweist, dass Lukas den Smart Speaker als Sachgegenstand betrachtet. Zweitens folgen im Anschluss keine Rückbezüge auf den Dialog mit dem VUI. Die „Trägerschaft“ des VUIs beschränkt sich also – vergleichbar zu anderen Gegenständen – auf einen sehr kleinen, klar abgegrenzten Ausschnitt im Zuge der Lebensmittelzubereitung. Es findet keine Relevanzherstellung im folgenden *turn-by-turn talk* zwischen Lukas und Alex statt und auch im ersten Beispiel lässt sich fragen, ob überhaupt von einer ‚Einbindung‘ gesprochen werden kann. Zwischen den Gesprächen und den anderen Praktiken werden in Beispiel (39) („Filmbericht“) gar keine Bezüge hergestellt. In Beispiel (40) („Kartoffeln“) wird über die Beteiligung an der Lebensmittelzubereitung ein impliziter Bezug auch zum Gespräch zwischen Lukas und Alex hergestellt. Dieser ist jedoch eng beschränkt auf den Ausschnitt an der Lebensmittelzubereitung, an der der Smart Speaker ‚beteiligt‘ war (in Form der Zeitmessung und der entsprechenden Erinnerung). Die Partizipation an dieser Stelle löst keine Folgebeteiligung an anderen Bestandteilen der Lebensmittelzubereitung aus und wird auch im Gespräch zwischen Lukas und Alex nicht noch einmal relevant. Die Nutzung des Smart Speakers bedurfte in diesen Beispielen auch nicht eines *accounts*, wie es in anderen Teilnehmendenkonstellationen der Fall ist – in Beispiel (39) auch deswegen nicht, weil die Nutzung durch den Alarmton hervorgerufen wird und insofern keinen *account* erfordert.

6.2.1.2 Prä-hoc-Einbindungen

Anders sind Fälle gelagert, in denen die Nutzung des Smart Speakers und damit auch der VUI-Dialog in den Fokus der Aufmerksamkeit rückt. Das folgende Beispiel zeigt eine mögliche Strategie, um einen solchen Übergang zu realisieren:

Beispiel (41): Super-Alexa-Modus (01-CVR02-06)

115 LF: <<pp> so.>
 116 (2.4)
 117 LF: aLEXa?
 118 (1.2)
 119 LF: spiel muSIK;
 120 (1.6)
 121 AL: hier ist ein R[Adiosender;]

122 LF: [alexa STOP;]
 123 (0.6)
 124 LF: aLEXa?
 125 (0.4)
 126 LF: spiel;
 127 (0.3)
 128 LF: SPotify;
 129 (1.2)
 130 AL: hier ist SPotify;
 131 (0.5)
 132 k: ((„Hectic“ von Enter Shikari wird erst
 laut und dann leiser abgespielt))
 133 LF: aLEXa;
 134 k: ((Musik wird leiser))
 135 LF: LEIser;
 136 k: [((Musik spielt weiterhin leiser))]
 137 [(1.5)]
 138 k: [((andere Melodie wird lauter
 abgespielt))]
 139 [(1.8)]
 140 KS: alexa LEIser;
 141 k: [((Musik wird leiser))]
 142 AK: [°h ha HA.]
 143 k: ((Musik wird in vorheriger Lautstärke
 weiter abgespielt))
 144 KS: alexa lAut [(.) stär (.) ke SECHS.]
 145 k: [((Musik wird leiser))]
 146 (1.2)
 147 k: [((Musik wird lauter abgespielt))]
 148 AK: [((is_es) ((unverständlich))]
 149 KS: (kommt drauf AN;)
 150 alexa lautstärke DREI.
 151 (1.4)
 152 k: [((Musik wird leiser abgespielt))]
 153 AK: [hat die ((unverständlich))]
 154 LF: so;
 155 (1.8)
 156 KS: **sag ma kennst du den Super alexa modus?**
 157 (1.7)
 158 LF: (den) WAS?

159 (0.7)
 160 KS: <<p> WArte.>
 161 alexa lautstärke SE[CHS.]
 162 k: [((Musik wird leiser))]
 163 (1.2)
 164 k: ((Musik wird lauter))
 165 KS: aLEXa?
 166 k: [((Musik wird leiser))]
 167 [(0.5)]
 168 KS: h0ch (.) h0ch (.) rUnter rUnter (.) lInks
 rEchts (.) lInks rEchts (.) be: a: (.)
 START,
 169 (1.4)
 170 AL: sUpEr alexa modus wird aktiVIERT.
 171 (0.6)
 172 AL: hiHI.
 173 (0.4)
 174 AL: zEhn extra leben adDIERT.
 175 (0.6)
 176 AL: speed level DREI freigeschaltet.
 177 (0.6)
 178 AL: alle power UPS verfügbar.
 179 AK: (die kann ja doch) [((unverständlich)).]
 180 AL: [FEHLfunktion.]
 181 AK: die kann auch in (DEU[TSCHland) sich
 selbst ((unverständlich))]
 182 AL: [powerups haben keine
 P0wer.]
 183 (0.4)
 184 KS: ((Lachen))
 185 AL: ABbruch,
 186 (0.2)
 187 k: [((Musik wird lauter))]
 188 LF: [mensch ACHim ich bin heute morgen]
 [((unverständlich ca. 1 Sek.))]
 189 KS: [alexa lautstärke DREI;]
 190 k: [((Musik wird leiser))]
 191 [(0.3)]
 192 KS: man kAnn (-) bei (ikea draußen) en GRILL
 kaufen;

Der Ausschnitt aus Haushalt 1 findet parallel zum gemeinsamen Kochen statt, zu dem auch mehrere Gäste anwesend sind. Der WG-Bewohner Lukas steuert zunächst die über den Smart Speaker wiedergegebene Musik, wobei er im VUI-Dialog mehrere Basis-Sequenzen²⁷⁴ produziert (vgl. 117–135). Nach mehrfachen Versuchen, die richtige Musik in der angemessenen Lautstärke wiederzugeben, beteiligt sich schließlich der befreundete Gast Kurt (KS) an der Bedienung des VUI (ab 140). Er übernimmt die Rolle der Steuerung – die Äußerung von Invokationen und Spracheingaben – von Lukas, ohne dass dies relevant gemacht wird. Zugleich beteiligt sich auch Alex als mithörender Ko-Rezipient der VUI-Dialoge: Er produziert Lachpartikeln (142) und kommentiert teilweise den Verlauf der Dialoge (vgl. 148, 153). Zwischen die Basis-Sequenzen zur Steuerung des VUI wird auf diese Weise eine Frage-Antwort-Sequenz zwischen Alex und Kurt eingeschoben (148–149).²⁷⁵ So ist die Steuerung des VUI bereits von der ausschließlich durch Lukas vollzogenen Auswahl von Musiktiteln und Lautstärkeregelung übergegangen in eine gemeinsam vollzogene Praxis.

Kurt führt dieses gemeinsame Ausprobieren des Smart Speakers weiter. Mit dem gesprächsorganisatorisch gebrauchten Imperativ „sag ma“ (156) initiiert er eine „kontextuell anknüpfende“ (Proske 2017: 90) Frage (siehe auch Auer/Günthner 2003). Diese ist an Lukas gerichtet und thematisiert eine spezifische Funktion von Amazon-Alexa, den „Super-Alexa-Modus“ (156). Bei der Funktion handelt es sich um ein sogenanntes „Easter Egg“, eine nicht in den offiziellen Anleitungen dokumentierte und primär der Unterhaltung dienende Funktion.²⁷⁶ Der „Super-Alexa-Modus“ adaptiert scherzhaft einen „Cheat“ – eine in Spielen versteckte Funktion, um sich einen Vorteil verschaffen zu können.²⁷⁷ Der Cheat stammt ursprünglich aus dem Videospiel *Teenage Mutant Ninja Turtles* (vgl. Baumann 2023). Der eigentlich als Controller-Tastenfolge zu lesende Code „Hoch, hoch, runter, runter, links, rechts, A, B, Start“, den Kurt verbalsprachlich äußert (vgl. 168), hat – nach dem Erfinder des Videospiels benannt – als „Konami-Code“ spiel- und genre-

274 Siehe Kap. 6.1.2.1.1.

275 Der genaue Wortlaut der Äußerungen von Alex lässt sich wegen der lauten Musikkwiedergabe sowie der Hintergrundgeräusche und der Überlagerungen nicht eindeutig klären. Allerdings lassen einzelne verständliche Lexeme, die sequenzielle Position der Äußerungen und die Prosodie darauf schließen, dass es sich hier um eine Interaktion mit Kurt handelt, in der die Performanz des Smart Speakers evaluiert wird.

276 Siehe aber Weinel/Griffiths/Cunningham (2014: 140–141) und Lakier/Vogel (2022) zu vielschichtigeren Analysen von Easter Eggs in Softwareanwendungen.

277 Zum kulturhistorischen Zusammenhang von Easter Eggs und Cheats siehe Consalvo (2007: 18–20).

übergreifend Bekanntheit erlangt und ist in verschiedenen Videospielen ein bekannter Cheat (vgl. Böhm 2020). Im Smart Speaker von Amazon löst die Wortkombination außer der zu hörenden Sprachausgabe (170–185) keine weitere praktische Umsetzung oder andere Funktion aus.

Der tatsächlichen Produktion der Spracheingabe für diese Funktion gehen mehrere Prä-Sequenzen voraus, die Kurt initiiert: Zunächst kündigt er die Demonstration der Funktion an, indem er sich erkundigt, ob Lukas diese kennt (156). Als dieser verneint, produziert er erneut einen gesprächsorganisatorischen Imperativ („WARte.“, 160). Mit diesem wird einerseits die Progression unterbrochen – was notwendig ist, um die Spracheingabe für das VUI gut vernehmbar produzieren zu können –, zugleich wird eine „Folgehandlung projiziert“ (Proske 2017: 84). Kurt stellt zudem die zuvor abgesenkte Lautstärke wieder höher ein (161–162), wodurch der folgenden Sequenz zusätzlicher Aufführungscharakter verliehen wird.²⁷⁸

Die Vorführung des Easter Eggs zeigt, dass das gemeinsame Ausprobieren und Vorführen in den Mittelpunkt gerückt wird. Dabei führt interessanterweise der Gast Kurt den Haushaltsmitgliedern Funktionen ihres eigenen Geräts vor, mit dem er möglicherweise selbst Erfahrung gesammelt hat oder über das er jedenfalls Kenntnis im Hinblick auf das Easter Egg erlangt hat. Der Übergang von der uneingebetteten und nicht kommentierten Lautstärkesteuerung durch Lukas hin zur Vorführsituation erfolgt inkrementell: Die akustische Steuerung des Interface ermöglicht es, diese mitzuverfolgen und im Rahmen sequenzieller Gesprächsorganisationsprinzipien ‚einzugreifen‘ bzw. Folge-Spracheingaben zu produzieren, die sich auf die vorherigen Steuerungsversuche beziehen. Die Kommentierungen von Alex zeigen, dass dabei mit dem Übergang hin zur Vorführsituation auch ein Übergang in ein gemeinsames *Gespräch* erfolgt. Der Übergang zeichnet sich durch die Produktion einer Prä-Sequenz (156–160) in Verbindung mit Vorankündigungen (161) und gesprächsorganisatorischen (156, 160) Äußerungen aus: Diese Äußerungen, hier v. a. von Kurt produziert, lassen die Nutzung des Smart Speakers vorübergehend im Mittelpunkt stehen; dafür sprechen auch die nicht vollständig verständlichen Äußerungen von Alex (153, 179, 181). Damit kann das VUI auch Eingang in das Gespräch finden und ist an diesem als Partizipand ‚beteiligt‘. Die tatsächlichen und potenziellen Äußerungen sowie praktische Umsetzungen

²⁷⁸ Eine hörbare Reaktion ist in der Aufnahme nur von Kurt (184) dokumentiert. Die anderen Haushaltsmitglieder sowie die Gäste sind im Anschluss deutlich leiser auf der Aufnahme zu hören und möglicherweise schon im Nebenzimmer oder jedenfalls weit entfernt vom Aufnahmegerät, sodass nicht ausgeschlossen werden kann, dass eine kurze Reaktion produziert wird (evtl. auch multimodal), die aufgrund der Aufnahmequalität nicht analysiert werden kann.

sind Redegegenstand und werden zunehmend in den Verlauf des Gesprächs integriert. Gleichwohl wird damit das VUI noch nicht zum Interaktanten in der Interaktion zwischen Lukas, Alex und Kurt: Auch wenn es praktisch am Vollzug des Gesprächs ‚beteiligt‘ ist, bleibt diese Partizipation auf die Bestandteile der sozialen Praxis beschränkt, in denen es entweder zur Musikkwiedergabe genutzt oder vorgeführt wird. Die Einbindung in das Gespräch wird also nicht kontinuierlich, sondern v. a. punktuell im Kontext der Demonstration vollzogen.

6.2.1.3 Post-hoc-Einbindungen

In den folgenden Beispielen startet der VUI-Dialog ohne dokumentierte Einleitung, wird aber *anschließend* auf unterschiedliche Weisen in die soziale Praxis eingebunden und *accountable* gemacht, etwa wie im folgenden Beispiel aus Haushalt 7:

Beispiel (42): Erraten (07-CVR01-14)

```

004 KW: ((singt Refrain mit 12.0 Sek.))
005      ((singt lauter 1.2 Sek.))
006      (4.6)
007 KW: HEY siri?=
008      =nächster SONG.
009      (0.8)
010 KW: und LAUter?
011 k:  ((Musik stoppt))
012      ((“Say Something” von Justin Timberlake
        setzt ein))
013 KW: DAS is cool ne?
014      kennst_e DAS?
015      (0.4)
016 KW: hey sirI: (.) LAUter?
017 k:  ((Musik wird lauter))
018 KW: das wird dir geFALL_n.
019      (18.0)
020 KW: du musst erRATen wer_s is;

```

Über einen längeren und auf der Aufnahme dokumentierten Zeitraum hinweg hört der Anwender Konrad Musik und singt teilweise mit. Er wechselt dann zu Beginn des obigen Ausschnitts das Lied und reguliert die Lautstärke (007–010). Anschließend spricht er seinen ko-präsenten Mitbewohner Till an – dass dieser

tatsächlich anwesend ist, ergibt sich aus Gesprächen im weiteren Verlauf der Aufnahme. Er referiert mit dem vagen Verweis „DAS“ (013) auf die Musikwiedergabe, genauer auf den wiedergegebenen Titel. Er fragt ihn auch, ob er „DAS“ kenne (014), worauf jedoch keine Antwort dokumentiert ist, und nimmt an, dass ihm „das“ (vermutlich der Musiktitel, dessen Interpreten Till im weiteren Verlauf der Aufnahme erraten soll, vgl. 020) gefällt (018). Zwischen den einzelnen Teilen der informellen Musikempfehlung adressiert Konrad erneut das VUI, um die Lautstärke zu erhöhen. Im Kontext der diskursiven Einheit der Empfehlung wird diese Lautstärkerhöhung also verständlich als Versuch, Till ein für diese Darbietung noch besseres bzw. lauterer Klangerlebnis zu ermöglichen. Der Dialog mit dem VUI wird in einen Kontext gestellt und so *accountable*.

Auch der folgende Ausschnitt dokumentiert einen VUI-Dialog aus Haushalt 7:

Beispiel (43): Colbie Caillat (07-CVR01-40)

003 KW: hey SIri von wem ist das lied?
 004 (1.8)
 005 SI: das IST brighter than the sUn von colbie
 caillat-
 006 (0.9)
 007 TW: (HÄ/HM)?
 008 (0.4)
 009 TW: hab_s nicht verSTANDen;
 010 TW: HEY-
 011 ((hustet))
 012 (0.5)
 013 TW: hey SIri?
 014 (0.8)
 015 TW: von WEM ist des lied?
 016 (1.4)
 017 SI: du hörst gerade <<deutsch
 ausgesprochen<BRiGhter>> than
 the sUn von colbie caillat-
 018 (1.1)
 019 TW: (colbi caillat) ((unverständlich));
 020 (6.6)
 021 [(0.6)]
 022 TW: [(NIE von gehört.)]
 023 (1.4)
 024 KW: nein colbie CAillat.

007 hm hm (DIE war leer.)
 008 GA: ich kEnne da einen ASsistenten persönlich;
 009 der dir widerSPRECHen würde;
 010 (2.0)
 011 LS: mh_HÄ hä hä witzig.
 012 k: ((LS und AW lachen))
 013 ((Räuspern eines MS))
 014 ((LS und AW lachen))

Der Ausschnitt entstammt aus einer Aufnahme aus Haushalt 8, in dem Robin (RL), Samuel (SM) und Lara (LS) kurz vor Beginn der Aufzeichnung die Ersteinrichtung des Google Home abgeschlossen haben. Die Sprecher*innen testen noch gemeinsam verschiedene Funktionen des Smart Speakers und adressieren teilweise verschiedene Stimmeingaben an das VUI. In diesem Kontext sind auch Aussagesätze als Stimmeingabe dokumentiert, die keine eindeutigen Aufforderungen oder Fragen beinhalten, so auch die Äußerung von Robin, die das VUI von Google mit der Feststellung ‚konfrontiert‘, das VUI Alexa von Amazon sei „WITZiger als du“ (004). Nachdem das VUI eine Sprachausgabe produziert hat (008), die auf Robins Eingabe antwortet, nimmt Lara eine – möglicherweise ironisch zu verstehende – Evaluation dieser vor (011), die wiederum zum Lachgegenstand wird (014). Letzteres lässt die ironische Deutung wahrscheinlicher werden. Die Produktion der Stimmeingabe ist zwar zuvor nicht durch andere Äußerungen angekündigt worden – allerdings kann festgestellt werden, dass in Situationen des ausgiebigen Testens, wie sie in den Daten v. a. im Kontext der Ersteinrichtung auftreten, mehrfach Aneinanderreihungen von VUI-Dialogen dokumentiert sind, die für sich genommen nicht *prä-hoc* eingebunden, aber *post-hoc* kommentiert und evaluiert werden.

6.2.1.4 Überkreuzungen

Die bisher gezeigten Einbindungen in die zwischenmenschlichen Interaktionen erfolgten jeweils zeitlich eindeutig bestimmbar *vor* oder *nach* der Bedienung des VUI. Das folgende Beispiel zeigt, dass sich diese beiden Sequenzen zeitlich auch überkreuzen können:

Beispiel (45): Können wir ein Spiel spielen? (07-CVR01-13)

018 TW: wollen wir ein bisschen FERNsehen
 schauen,
 019 KW: (gleich vielleICHT;)

020 TW: <<ohne auslautenden Vokal> OK.>
 021 (6.0)
 022 KW: hey siRI:?
 023 TW: <<all>möchtest du> was ABhaben?
 024 KW: können wir ein SPIEL spielen?
 025 k: ((Musik pausiert 1.0 Sek.))
 026 ((Musik wird wieder lauter))
 027 KW: (SCHA:de;)
 028 Ä:H ne danke ((xxx xxx XXX xxx;))
 029 (0.6)
 030 KW: hey SIri:?
 031 (1.0)
 032 KW: [ich hätt gern ein SPIEL.]
 033 k: [((Musik endet))]
 034 (3.7)
 035 SI: hier ist musik aus deiner media[THEK.]
 036 KW: [NE:.]
 037 TW: NÖ.
 038 (0.5)
 039 k: ((“Heaven Can Wait” von LSD setzt ein und
 spielt fortlaufend))
 040 TW: interessant DAS kann sie nIcht;

Der Ausschnitt aus Haushalt 7 setzt ein, nachdem – wie aus den vorherigen Aufnahmen hervorgeht – Essen geliefert und ausgepackt wurde. Der Vorschlag von Till, fernzusehen (018), kann also als Begleitaktivität zum Essen verstanden werden. Konrad lehnt diesen Vorschlag zunächst ab und wendet sich stattdessen einem anderen Gerät aus dem Bereich der Unterhaltungselektronik zu, dem Smart Speaker, den er nach einem Spiel fragt (024). Zwischen der Invokation und der Produktion der Stimmeingabe fragt Till an Konrad gerichtet, ob er etwas (von seinem Essen) abhaben möchte (023). Eine Antwort auf diese Frage wird konditionell relevant, allerdings hat die von Konrad ausgesprochene Invokation bereits stattgefunden und lässt sich nicht mehr ohne weitere Äußerungsproduktionen abbrechen. Konrad produziert entsprechend zunächst die Stimmeingabe und wartet das Ergebnis ab – in dem Fall wäre aufgrund der Entscheidungsfrage eine Stimmausgabe erwartbar. So vervollständigt der Nutzer nicht nur ‚seine‘ Teile der Basis-Sequenzstruktur, sondern ermöglicht durch das Nicht-Einlösen der konditionellen Relevanz, die Till eröffnet hat, auch, dass das VUI entsprechend den komplementären Part liefern kann und dabei gehört wird. Allerdings geschieht außer einer kurzen Absenkung der Lautstärke nichts, was Konrad kurz evaluiert

(027). Erst nach dieser evaluierenden Post-Expansion der Basis-Sequenzstruktur greift Konrad die Äußerung von Till (aus 023) wieder auf – nach einer turninitialen Hesitationspartikel (vgl. Schwitalla 2002: 261–263) beantwortet er die Frage von Till (028).

Das Beispiel kann somit als Beleg für die Stabilität der konditionellen Relevanzen innerhalb der Basis-Sequenzstruktur des Austauschs zwischen Anwender*in und VUI gelesen werden, und zwar auch dann, wenn in der zwischenmenschlichen Interaktion neue konditionelle Relevanzen entstehen – diese können ohne hier angezeigte Schwierigkeiten zeitlich nach hinten ‚verschoben‘ werden. Insbesondere die Stelle zwischen Invokation und Stimmeingabe, in der das VUI entweder bereits im Listening-Modus ist oder diesen aktiviert, ist störanfällig, denn die Produktion der Stimmeingabe hat noch nicht hörbar für die anderen Beteiligten begonnen und das laufende Turn-Taking ist noch nicht suspendiert, was aber wiederum Voraussetzung für die störungsfreie Produktion der Stimmeingabe ist (vgl. Porcheron et al. 2018: 7). Dies kann ein Faktor sein, der dazu führt, dass die Basis-Sequenzstruktur für die VUI-Eingabe hier jedenfalls von Konrad priorisiert gegenüber der Äußerung von Till behandelt wird und sich die Sequenzverläufe zeitlich überkreuzen.

6.2.2 VUIs im Zentrum sozialer Praxis

Wie in den bisherigen Beispielen deutlich wurde, wird ein VUI nur durch das zeitgleiche Auftreten von VUI-Dialog und laufender Interaktion zwischen zwei oder mehr menschlichen Beteiligten noch nicht zum Gesprächsteilnehmer. In den Daten zeigen sich aber auch Fälle, in denen sich das Gespräch über den Smart Speaker mit der Nutzung des VUI verbindet. Wie einleitend zu Kap. 6.2 dargestellt, ist die Nutzung von VUIs zugleich von empraktischem und konversationellem Charakter. Das Gespräch *über* den empraktisch-konversationell eingebundenen Gegenstand führt offenbar wiederholt, so die Arbeitshypothese, zu Gesprächen *mit* dem Smart Speaker. Dies kann als weiterer Beleg für die Gefräßigkeit (Goodwin 2018: 445) der sozialen Praxis gesehen werden: Praxis ist demzufolge „semiotically opportunistic, with the ability to incorporate voraciously whatever local materials might be used to construct the action required at just this moment“. Spezifischer auf Gespräche angewendet heißt das, dass konversationell geformtes akustisches Material in die Praxis einbezogen, transformiert, umgebildet und als Substrat weiterverwendet wird, sodass formell VUIs kurzzeitig wie Gesprächsteilnehmer erscheinen können, während auf der funktionalen Ebene eine Reihe anderer Praktiken vollzogen werden, z. B. Demonstrieren, Testen, das Betreiben humorvoller Interaktion und nicht

zuletzt das Aneignen und Domestizieren des Geräts mit seinem spezifischen Interface (siehe auch Klemm 2001).

Davon ausgehend stellen sich zwei Fragen: Erstens, in welchen sozialen Situationen konkret schreiben Sprecher*innen VUIs diesen Status zu, und zweitens, wie werden die ‚Beteiligung‘ oder Nicht-Beteiligung und oszillierende Stadien sprachlich (auch multimodal) inkrementell vollzogen? Zunächst sollen unterschiedliche Situationstypen vorgestellt werden, in denen in den vorliegenden Daten entsprechende Selbst- und Fremdzuschreibungen vorgenommen werden – sie zeigen sich entlang der Praktiken, die mit solchen Zuschreibungen vollzogen werden. Diese Situationen haben gemeinsam, dass in allen Fällen die Bedienung des Smart Speakers Teil des laufenden Gesprächs zwischen den menschlichen Beteiligten wird. Das Bedienen, Einrichten, Vorführen, Testen und Bewerten der Geräte schiebt sich ‚in den Mittelpunkt‘ der sozialen Praxis und wird zum Hauptgegenstand des Gesprächs.

6.2.2.1 Ersteinrichtungen des Smart Speakers

Zur genaueren Charakterisierung der Teilnehmendenkonstellationen in den Situationen der Inbetriebnahme wurde zunächst überprüft, ob und wie die Teilnehmenden selbst die Situation als solche erzeugen, wie sie die Inbetriebnahme beginnen und wen sie dabei als Partizipanden relevant setzen. Die Rahmung einer Situation und etwa zu Beginn ausgetauschte Grüße oder die Herstellung einer Positionierung der ‚Beteiligten‘ zueinander können Aufschluss darüber geben, wer bzw. was als ‚beteiligt‘ gelten kann. In den Situationen der Ersteinrichtung zeigt sich vor Beginn der eigentlichen Installation, dass „den Smart Speaker einrichten“ als gemeinsame Unternehmung rituell gerahmt wird (siehe auch Goffman 1974), etwa im folgenden Beispiel aus Haushalt 1:

Beispiel (46): So jetzt läuft alles (01-EE-01)²⁷⁹

002 LF: \$SO: video läuft,\$
 ak: \$läuft zur Tür---\$
 003 (0.51)
 004 AK: <<p> sekunDE,>
 005 (0.52)

²⁷⁹ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00465_SE_01_T_01_DF_01&cID=c2&wID=&textSize=900&contextSize=32.

006 AK: <<p> +AH §(.) ich wollt mir (noch nen)+
 kaffee holen,>§
 ak: +blickt in die Kamera-----+
 lf: §kommt hinter Kamera hervor----->
 >-----§

007 LF: §DANN hol dir noch schnell n kaffee;
 008 AK: JA das is (doch) ne gute idee-
 009 (0.27)§
 lf: §läuft zum anderen Ende des Raumes, neben
 das Sofa----->
 >-----§

010 LF: ((lacht))
 011 LF: [((lacht))]
 012 k: [((MS verlässt den Raum, 1.86 Sek.))]
 013 LF: Δ°h-----Δ
 lf: Δgreift Handy, dreht es in der HandΔ
 014 Δ(22.8)
 Δ
 lf: Δmanipuliert Handy, kniet sich mit linkem
 Bein auf Sofa und wendet
 sich Stativ zuΔ

015 LF: Δ°h Δ
 lf: Δbefestigt Handy in StativΔ
 016 (-)
 017 k: ((Auslöseton der Kamera))
 018 (-)
 019 k: Δ((Quietschen))Δ
 lf: Δdreht an HandyΔ
 020 h°

021 LF: §so. §
 ak: §steht von Sofa auf§
 022 Δ(2.62) Δ
 lf: Δläuft zu Alexa, bückt sich und greift
 nach DiktiergerätΔ

023 LF: Δh:m- Δ
 lf: Δnimmt Diktiergerät in die Hand und
 beginnt zu tippenΔ
 024 Δ(16.44) Δ
 lf: Δtippt auf Diktiergerät und drückt auf
 Knopf von AlexaΔ

025 LF: §so. §
 ak: §betritt den Raum mit Tasse in der Hand§
 026 (0.92)
 027 LF: Δjetz läuft ALles, Δ
 lf: Δlegt Diktiergerät neben AlexaΔ
 028 (1.73)
 029 LF: §Δun:d wir können STARten; Δ §
 lf: Δgreift nach Handy auf TischΔ
 ak: §läuft an Tisch vorbei zum Sofa§
 030 AK: §Δ(bin ich) ja ma geSPANNT;Δ
 §
 ak: §beugt sich zum Tisch um Tasse
 abzustellen§
 lf: Δgreift nach Handy Δ

Lukas und Alex haben die Videoaufnahme für die Untersuchung gestartet (002) und sowohl die Kamera wie auch die übrigen Aufnahmen sowie den Smart Speaker in der Nähe des zukünftigen Aufstellorts platziert. Lukas bestätigt, dass nun die Bedingungen erfüllt sind, um die Ersteinrichtung des Smart Speakers zu starten, was dies als gemeinsam und am Stück durchzuführende Aktivität rahmt. Dies wird durch Alex noch einmal bestätigt, der sich „noch schnell“ einen Kaffee holen will, bevor sie damit tatsächlich beginnen (006) und Lukas ratifiziert diese Situationsrahmung (007), er wiederholt dabei die situationsdifferenzierende Wortkombination „noch schnell“, mit der impliziert wird, dass dies ‚vor Beginn‘ einer dann nicht mehr zu unterbrechenden Aktivität steht. Auch die Äußerungen von Lukas und Alex nach Alex' Rückkehr weisen darauf hin, dass nun eine Aktivität gestartet wird (027–029), an der die beiden gemeinsam teilnehmen; Alex drückt Spannung aus (030). Eine ähnliche Rahmung zeigt sich auch bei der Ersteinrichtung in Haushalt 7, an der Konrad und Till beteiligt sind:

Beispiel (47): Let's go let's fetz (07-EE-01)²⁸⁰

002 KW: Δ((räuspert sich)) (.) let_s GO.
 003 (0.55) Δ
 kw: Δbetritt das Wohnzimmer mit Tasse in der
 linken Hand und schließt mit der rechten
 die TürΔ

²⁸⁰ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00467_SE_01_T_01_DF_01&cID=c2&wID=&textSize=500&contextSize=16.

004 TW: let_s go letz_FETZ,
 005 §(0.4)
 006 KW: ja; hh°
 007 (2.88)§
 kw: §durchquert den Raum und bleibt stehen----->
 >-----§
 008 KW: ALso-
 009 (0.46)
 010 KW: §((schmatzt)) ich würde SA:gen;
 §
 kw: §macht einen Schritt nach vorne Richtung
 Sofa§
 011 (0.49)
 012 KW: °hh
 013 (2.14)
 014 KW: ((schmatzt)) könnTEST du-
 015 (1.54)
 016 KW: *KOMmen h°? *
 kw: *dreht Handfläche nach vorne und bewegt
 Finger nach oben*
 017 (0.26)

Zu Beginn der Aufzeichnung produziert Konrad den adhortativ gebrauchten Ausdruck „let's GO“ (002), mit dem eine gemeinsam durchzuführende Aktivität angekündigt wird. Till erweitert dies zur Phrase „let's go, let's fetz“ (004), mit der die Durchführung der gemeinsamen Aktivität ratifiziert wird. Mit der Produktion des Diskursmarkers „ALso-“ (siehe dazu Alm 2007: 97–107; Dittmar 2010; Konerding 2004) sowie der Phrase „ich würde SA:gen;“ (010) projiziert Konrad eine daran anschließende Aktivität. Er fordert dabei, noch bevor er in diese genauer einsteigt, Tills Involviertheit an dieser ein (014–016). Zusammengenommen wird also auch hier die Ersteinrichtung als Aktivität gerahmt, die den vollen Fokus der Aufmerksamkeit und die (körperliche und mentale) Präsenz der Beteiligten erfordert.

In Haushalt 8 zeigt sich die Involviertheit weniger versprachlicht, wird jedoch ebenfalls deutlich, wenn die räumliche Anordnung der beteiligten Sprecher*innen in ihrem Verhältnis zum Smart Speaker berücksichtigt wird.²⁸¹ Die Beteiligten sitzen um den Küchentisch herum, der Smart Speaker steht dabei in der Mitte aller

²⁸¹ Siehe Abb. 23 (Kap. 5.5.3).

Beteiligten und frontal vor Robin (RL), der die Ersteinrichtung mit seinem Smartphone vollzieht. Ähnlich sind die räumlichen Konstellationen der beiden anderen Haushalte,²⁸² in denen ebenfalls der Smart Speaker in der Mitte zwischen den beteiligten Haushaltsmitgliedern steht, sodass dieser von allen Beteiligten sowohl visuell wie auch akustisch potenziell gut wahrgenommen werden kann. Darüber hinaus verwendet auch in Haushalt 8 der Anwender Robin nach dem Auspacken des Google Assistants einen aktivitätseinleitenden Ausdruck, gefolgt von einer handlungsplanerischen Äußerung:

Beispiel (48): Los geht's (08-EE-03)²⁸³

- 090 $\Delta(1.05)$ Δ
- rl: Δ dreht Anleitungsheft in beiden Händen
haltend auf Rückseite und wieder zurück Δ
- 091 RL: Δ LOS geht_s; Δ
- rl: Δ nimmt Anleitungsheft in linke Hand und
greift mit rechter nach Papphalterung des
Ladeadapters Δ
- 092 $\Delta(0.28)$ Δ
- rl: Δ dreht Papphalterung in mehrere
Richtungen Δ
- 093 RL: okay DANN-
- 094 Δ gehen wir doch ma NACH,
 Δ
- rl: Δ legt Papphalterung vor sich auf dem Tisch
ab Δ
- 095 RL: Δ ANleitung vor; Δ
- rl: Δ greift Anleitung auch mit rechter Hand
und zieht Heft dahinter hervor Δ

Es kann also davon ausgegangen werden, dass die vollzogene soziale Praktik auch aus Perspektive der Beteiligten in allen untersuchten Fällen die Erstinstalla-

²⁸² Für Abbildungen siehe Kap. 5.5.3.

²⁸³ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c85&wID=&textSize=200&contextSize=4.

tion des Smart Speakers ist und dass diese bis zu ihrem Abschluss im Mittelpunkt der Aufmerksamkeit stehen wird – damit wird für diesen ein hoher Grad von Beteiligung an der sozialen Praxis projiziert, wenn auch nicht explizit. Dabei muss erwähnt werden, dass die Erzeugung dieses Grads an Involviertheit auch durch die Durchführung der Untersuchung entsteht – den notwendigen Aufbau von Kameras und Aufzeichnungsgeräten und die Instruktion, die Einrichtung möglichst mit mehreren Beteiligten zu vollziehen.²⁸⁴ Erstens zeigen sich aber auch in den CVR-Aufnahmen entsprechende Rahmungen (s. u.) und zweitens ist zumindest vorstellbar, dass auch ohne die Spezifität der Teilnahme an einer Studie die Teilnehmer*innen um die Einrichtung eines neuen Geräts in der Wohnumgebung herum eine soziale Situation konstruieren würden. Dies steigert die „Legitimität“ (Gerwinski/Linz 2018: 107) der hier arrangierten und aufgezeichneten Aufnahmesituation – es kann davon ausgegangen werden, dass dies „auch dann stattgefunden hätte, wenn keine Aufnahme durchgeführt worden wäre“ (Sager 2001: 1023), wenn auch unter anderen Rahmenbedingungen und ggf. in einer leicht veränderten Teilnehmendenkonstellation. Das Natürlichkeitspostulat in der Gesprächsanalyse kann also mit Einschränkungen als erfüllt angesehen werden – die Daten wurden nicht im engeren Sinne elizitiert. Der hier gegebene Datentyp verdeutlicht aber noch einmal, dass solche Kriterien als Spektrum begriffen werden müssen (siehe auch Kendrick 2017: 3).

Die Teilnehmer*innen richten ihre Aufmerksamkeit in den folgenden Minuten (überwiegend) auf den Smart Speaker und die notwendigen Peripherie-Geräte, insbesondere Smartphones, sowie Anleitungen, Kabel und Verpackungsmaterial. Anders als etwa ein Schuhverkauf oder eine ärztliche Untersuchung – Beispiele für „nonlinguistic contexts“ bei Goffman (1979: 15) – sind jedoch Praktiken, an denen ein Smart Speaker ‚beteiligt‘ ist, nicht „nonlinguistic“ (und auch nicht ‚nicht-konversationell‘, zu diesem Begriff s. o.), sondern gerade davon gekennzeichnet, dass sie sich konversationell vollziehen – eben aufgrund der spezifischen Eigenschaften dieses Geräts und seines Interfaces, das konversationell und auf dem akustischen Kanal prozessiert (siehe auch Hector 2022). In den Ersteinrichtungssituationen tritt zudem die Arbeit am Aufbau einer Personenfiktion seitens des VUI deutlich hervor, wie etwa das folgende Beispiel aus Haushalt 7 zusätzlich illustriert.

²⁸⁴ Siehe Kap. 5.

Beispiel (49): Willkommen beim HomePod (07-EE-14)²⁸⁵

386 (1.15)
 387 SI: +halLO- +
 kw: +blickt TW an und dann zurück auf Handy+
 388 [ich] bin +SIri, +
 389 KW: [(oKAY.)]
 tw: +blickt KW an+
 390 (0.23)
 391 SI: willkommen beim \$HOME+pod.\$ +
 kw: \$lächelt--\$
 kw: +blickt kurz TW an+
 392 (0.43)
 393 SI: +du SIEHST es zwar nicht + (.)
 aber ich \$wInke
 gerade; \$
 tw: +blickt auf Handy und zurück zu KW+
 kw: \$stützt linke
 Hand mit Kaffeetasse auf linkem
 Oberschenkel ab\$
 394 TW: ((lacht))
 395 KW: [+(~~lacht~~)] +
 kw: +schweift mit Blick über TW zu HomePod+
 396 SI: [wenn du meine +Aufmerksamkeit] +
 +möchtest +
 (.) sag EINFach,
 kw: +blickt zurück auf Handy+
 tw: +blickt zurück auf Handy+
 397 (0.22)
 398 SI: hey SIri?
 399 (0.92)
 400 SI: lass uns LOSlegen;

285 Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00467_SE_01_T_01_DF_01&cID=c389&wID=&textSize=600&contextSize=20.

- 724 AL: hier sind erst einmal VIER dinge zum
 aus\$probieren-
 §
 lf: \$wippt mit Kopf auf und
 ab\$
- 725 (0.88)
- 726 AL: du kannst jederzeit STOP sagen-
 727 (0.96)
- 728 AL: leGEN wir los;
 729 LF: STOP-
- 730 AL: um mit mir ZU-
- 731 AL: [SPREchen (.) nenne einfach,]
 732 AK: [+(lacht, 1.43 Sek.))]

Die Begrüßung ist hier sehr ähnlich aufgebaut wie bei der zuvor gezeigten Ersteinrichtung des Apple-Geräts: Nach einer Grußform und Selbstvorstellung, die hier noch um eine genauere Erläuterung der nun folgenden Schritte ergänzt wird (718–726), folgt eine adhortative Form (728), die hier nicht mit der flektierten Imperativform von „lassen“, sondern mit dem Konjunktiv I von „loslegen“ gebildet wird. Der Adhortativ zeigt sich auch hier als Form der „enge[n] Handlungskoordination zwischen Sprecher und Anwesenden“ (Zifonun/Hoffmann/Strecker 1997: 140). Allerdings wird der anschließend durch das VUI eingeleitete erste Schritt des Handlungsablaufs von Lukas durch die Äußerung „STOP-“ unterbrochen (729); dies muss als Referenz auf den Hinweis des VUI gelesen werden, dass „jederzeit STOP“ gesagt werden könne (726). Alex ratifiziert diese Brechung durch lautes und fortgesetztes Lachen als Humor (732) und präsentiert sich damit als formal sekundärer, auf der Ebene der sozialen Beziehungen jedoch als primärer Adressat der Äußerung von Lukas. Das VUI hingegen setzt das vorgesehene Ersteinrichtungsskript fort und liefert keine Reaktion auf den Stop-Befehl; stattdessen wird mit der Aufforderung zur Wiederholung eines Befehls (ähnlich wie im zuvor präsentierten Einrichtungsdialog aus Haushalt 7) fortgesetzt. Ähnlich gerahmt werden auch die ersten Dialoge mit dem Google Nest in Haushalt 8:

Beispiel (51): Hallo, ich bin dein Google Assistant (08-EE-05)²⁸⁸

457 RL: dein google assistant wird dir ANTworten
wenn du hey [(google)-]

458 GA: [HALlo;]

459 +ich bin dein GOO +gle +assistant; +
sm: +wendet Blick auf GA+ +blickt RL an+
rl: +rollt Augen zu GA und zurück-----+

460 (0.3)

461 GA: ich bin \$+hier + um dir zu HELfen;>
>-----\$
sm: +blickt GA an+
rl: \$nickt, zeigt mit rechtem Daumen
hoch und senkt Hand\$

462 (0.21)

463 SM: \$[((lacht))]--\$
sm: \$senkt den Kopf\$

464 GA: [in der goo]gle HOME app kannst du dich
über meine funktionen informieren;

465 (0.9)

466 RL: HMhm;

467 SM: °h

468 LS: *oKAY;*
ls: *nickt*

Auffällig ist hier, dass die Nutzer*innen die Äußerungen des Google Nest ratifizieren (vgl. 465–467) – damit bestätigen sie, primär sich untereinander, dass sie die Äußerungen des VUIs zur Kenntnis genommen haben. Die rituelle Rahmung der Situation findet also nicht nur seitens der Nutzer*innen statt, sondern ist auch VUI-seitig programmiert und Bestandteil der parainteraktiven Elemente der VUIs. Die beiden Smart Speaker von Apple und Amazon stellen sich namentlich vor – siehe Beispiel (49) (388) und Beispiel (50) (720). Das ist insbesondere im Kontrast zu Beispiel (51) auffällig, weil das VUI herstellerseitig keinen eigenen Personennamen hat und anders als Alexa und Siri nur mit dem Firmennamen angesprochen

²⁸⁸ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c403&wID=&textSize=300&contextSize=8.

wird (vgl. 458).²⁸⁹ Alexa produziert zusätzlich eine weitere rituelle Äußerung zur Etablierung des Beginns sozialer Beziehung („SCHÖN (-) dich kennenzulernen“, 722).

Die an der Ersteinrichtung ‚beteiligten‘ Sprecher*innen bestätigen diese rituellen Rahmungen an dieser Stelle nicht explizit – sie stellen sich nicht selbst vor oder erwidern die entsprechenden Grüße. Allerdings gehen die drei untersuchten Haushalte sehr unterschiedlich mit den Begrüßungsdialogen um: In Haushalt 7 produzieren Konrad und Till lediglich Lachen und verhalten sich ansonsten kompliant mit dem Ersteinrichtungsdialog. In Haushalt 8 ist zwar auch keine Bestätigung der Grußformeln erkennbar, aber immerhin produzieren Robin und Lara Rezeptionssignale, teilweise auch multimodal (465, 467). Demgegenüber ratifizieren die beiden Sprecher in Haushalt 1 nicht nur nicht die Teilnahme des Geräts an der Interaktion zwischen ihnen, sondern verhalten sich zudem nicht-kompliant im folgenden Einlern-Dialog: Das VUI produziert die Ankündigung, ‚Dinge‘ (gemeint sind hier verschiedene Typen von Stimmeingaben) auszuprobieren und verweist auf die Möglichkeit, „stop“ zu sagen“ (726), d. h. die Versuche abubrechen. Noch bevor diese überhaupt gestartet werden können, produziert Lukas diesen Stop-Befehl (729). Der Blick von Lukas zu Alex und auch das anschließende Gelächter von Alex deuten darauf hin, dass dies als eine Belustigung über das sprechende VUI gedeutet werden kann, das sie insofern nicht als ratifizierten Gesprächsteilnehmer bestätigen. Wie auch Habscheid (2022: 182) anhand desselben Auszugs argumentiert, kann dieses Nicht-Befolgen der vorgegebenen Dialogstruktur als Teil der Domestizierung der Geräte verstanden werden. Zugleich zeigt sich hier, dass das VUI dennoch nicht als unbeteiligt an der sozialen Praxis gesehen werden kann: Das VUI produziert erst das sprachliche Material, auf das Lukas seinen an Alex adressierten Witz aufbauen kann – der ko-operative Charakter der Verfertigung dessen wird deutlich.

6.2.2.2 Üben und Testen

Es zeigt sich also, dass die Einbeziehung des VUI sowohl sach- als auch humororientiert in den Vollzug der sozialen Praxis eingebunden werden kann. Dabei kommt es zwischen diesen beiden Orientierungen zu Übergängen, die sich an der Grenze von ‚Austesten‘ zu ‚spielerisch Überfordern‘ zeigen. Daher werden nachfolgend Praktiken des Übens und Testens beleuchtet. Im Rahmen der Ersteinrichtungen tritt im Rahmen der Begrüßungsdialoge bei Amazon und Apple (Haushalt 1 und 7) eine Phase ein, in der durch strukturiertes Vor- und Nachsprechen der Austausch mit dem VUI geübt werden kann. Damit wird die konversationelle Bedienung eines Smart Speakers als für die Anwender*innen herstellerseitig als

²⁸⁹ Siehe auch Kap. 6.1.1.1.

noch unvertraut gerahmt, zugleich entstehen situativ weitere Gelegenheiten für eine Ratifizierung des VUI als ‚Gesprächsteilnehmer*in‘. Betrachten wir dazu den weiteren Verlauf der Ersteinrichtung von Haushalt 1:

Beispiel (52): Alexa, wie ist das Wetter? (01-EE-17)²⁹⁰

730 AL: um mit mir ZU-
 731 AL: [SPREchen (.) nenne einfach,]
 732 AK: [+((lacht, 1.43 Sek.))+]
 lf: +lächelt-----+
 733 AL: meinen NAMen;
 734 AL: +[und STELle-] +
 lf: +blickt zu Alexa+
 735 AK: [°h]
 736 AL: +eine \$FRAge, +
 \$
 lf: +blickt zurück zu Smartphone+
 ak: \$stützt sich mit beiden Händen auf
 Kante des Sofas ab\$
 737 (0.47)
 738 AL: VERsuche es mit-
 739 (0.25)
 740 AL: aLEXa,
 741 (0.5)
 742 AL: wie ist das WETter,
 743 (0.76)
 744 LF: ((schmatzt)) aLEXa wie ist das wetter.
 745 (2.31)
 746 AL: in amelsbüren beträgt die temperatur
 sEchs grad celsius
 bei beWÖLKtem himmel.
 747 (0.29)
 748 AK: alexa korrigiere unserer ADRESse (.).

290 Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00465_SE_01_T_01_DF_01&cID=c626&wID=&textSize=500&contextSize=16. Für eine Analyse dieses Auszugs unter dem Aspekt der Sequenzialität siehe Habscheid (2022).

749 <<lachend> wir \$wohnen nich in
 AMselbüren.>\$
 lf: \$lächelt leicht----->
 >-----\$
 750 (0.23)
 751 AL: +das habe ich + nicht
 verSTANDen.
 ak: +blickt zurück auf Laptop+
 752 AK: \$((lacht))\$
 lf: \$lächelt--\$
 753 AL: [verSUche es mit,]
 754 MS: [(KLAR nich;)]
 755 AL: aLEXa,
 756 (0.3)
 757 AL: setze \$baNAnen auf die EInkaufsliste;\$
 lf: \$dreht Kopf zu AK-----\$

Durch den Modus der strukturierten Einlernphase werden seitens des VUI Stimm-eingaben ‚vorgesehen‘, die die Anwender nachsprechen sollen (731–742). Nach der Produktion der vorgeschlagenen Stimmeingabe einschließlich der Invokation (740–744) beantwortet das VUI diese mit der entsprechenden Stimmausgabe. Der Dialog bestätigt noch einmal den Befund einer Basis-Sequenzstruktur für VUI-Dialoge, die nach diesem Muster funktionieren und an die die Anwender*innen bei der Ersteinrichtung des Geräts gewöhnt werden sollen.²⁹¹ Der zweite Anwender, Alex (AK), übernimmt allerdings nun das Rederecht (748) und produziert – noch bevor die im prästrukturierten Dialogablauf vorgesehene nächste ‚Übung‘ seitens des VUI eingeleitet werden kann – eine andere Stimmeingabe, die sich auf die vorherige bezieht. Alex startet insofern nicht nur den Versuch einer Adres-seingabe, sondern auch einer sequenziellen Entfaltung eines Dialogs. Allerdings wird die Stimmeingabe nicht erwartungsgemäß produziert, stattdessen produ-ziert das VUI die Standardfehlermeldung „das habe ich nicht verSTANDen“ (751), worauf Alex mit lautem Gelächter reagiert (752). Auch wenn der Versuch, das VUI als Partizipanden eines sich sequenziell entfaltenden Gesprächs mit Inferenzen zu behandeln, fehlschlägt, zeigt sich dennoch, dass die sich zeitlich und akustisch vollziehende Bedienweise den Anwender Alex dazu eingeladen hat, die Korrektur ebenfalls auf diesem Kanal zu produzieren und dabei für menschliche Konversa-tionen beschriebene Mittel der Gesprächsorganisation anzuwenden – insbeson-

291 Siehe Kap. 6.1.2.1.1.

dere die Äußerung „wir wohnen nicht in Amelsbüren“ (749) stellt Kohäsion durch die Wiederholung des Namens des Stadtteils her und kann als Versuch gelesen werden, lokale Kohärenz und topikale Progression zu erzeugen. Dies unterstellt dem VUI (testweise) einen gewissen Grad an Gesprächskompetenz.

Der strukturierte Einlern-Dialog, der bei Haushalt 1 noch über den hier gezeigten Ausschnitt hinaus andauert, ist bei Haushalt 7 deutlich kürzer:

Beispiel (53): Hey Siri, was kannst du? (07-EE-15)²⁹²

400 SI: lass uns LOSlegen;
 401 (.)
 402 SI: sag.
 403 (0.25)
 404 SI: hey SIri (.) was kAnnst du?
 405 §(0.46) §
 tw: §legt rechte Hand auf rechtes Knie§
 406 KW: +°h +
 kw: +wendet Blick von Handy ab zu SI+
 407 (0.49)
 408 KW: hey SIri (.) was kAnnst du?
 409 §h° §
 tw: §wendet Kopf zu Siri§
 410 §(1.33) §
 tw: §kippt Kopf leicht nach oben und rechtes
 Knie leicht zu linkem§
 411 KW: °hh
 412 (0.72)
 413 SI: ich kann VIEle dinge
 +erledi[gen; +
 414 SI: zum BEIspiel das lIcht ein]schalten-
 kw: +blickt zurück auf Handy+
 415 KW: [§<<lachend> DANke.>§]
 kw: §dreht Kopf zu TW--§
 416 SI: die aktuellen NACHrichten abrufen,
 417 und dir SAgen wie das wetter wird.

²⁹² Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00467_SE_01_T_01_DF_01&cID=c415&wID=&textSize=600&contextSize=20.

418 \$(0.74) \$
kw: \$senkt Kopf zurück zu Handy\$

419 SI probier_s JETZT mal mit,
420 +(0.1) +
kw: +blickt TW an+

421 SI: \$hey SIri (.)
 \$ spiel musik,
tw: \$legt rechten Unterschenkel auf linkem
Knie ab\$

422 \$(0.89) \$
kw: \$zieht linke Augenbraue hoch\$

423 KW: äh hey SIri (.) spiel musIk.
424 (2.12)

425 KW: °hh
426 (0.4)

427 SI: \$GERne hier kommt musik;----->
 >-----\$

tw: \$zieht rechten Oberschenkel näher an
Oberkörper\$

428 (0.32)

429 SI: EXtra für dich zusammengestellt-

430 KW: oh-
431 (0.27)

432 TW: aHA,

433 k: ((HomePod spielt „Copacabana“ von Leon
Machère ab))

Auch hier zeigt sich aber ein gewisser Grad an Sensitivität für eine ‚Beteiligung‘ des VUIs am Gespräch: Konrad und Till äußern beide Erkenntnisprozessmarker (430, 432) die mit der Erzeugung von „joint projects“ assoziiert werden (Imo 2009: 84–85; siehe auch Heritage 1984a). Da sie beide auf den Beitrag des VUIs Bezug nehmen, sind das „OH-“ sowie das „aha,“ formal auch auf diesen gerichtet; für „aha“ stellt Imo (2009: 72) eine „minimale Bedeutungskomponente“ fest, die darin bestehe, „den Empfang einer neuen Information zu quittieren“. Allerdings produzieren Konrad und Till keine Invokation, wodurch sie als direkt an das VUI adressiert markiert wären, sondern äußern sie aneinander adressiert. Diese doppelte Adressiertheit von Äußerungen, die einen Split zwischen formaler und funktionaler Adressiertheit darzustellen scheint, ist charakteristisch für VUI-Dialoge in Mehrparteienkonstellationen, wie sich in weiteren Analysen zeigen wird. Sie sind auch typisch für fernsehbegleitendes Sprechen (vgl. Holly 2001b: 15; Baldauf 2001:

71–72; Klemm 2001: 103). Äußerungen wie im gegebenen Beispiel – „oh-“ und „aha“, in sehr kurzem zeitlichen Abstand zueinander – ergeben sich, wie auch im fernsehbegleitenden Sprechen beobachtet, aus dem geteilten Fokus der Aufmerksamkeit der Teilnehmer*innen (vgl. Baldauf 2001: 79).

Die formalen Erscheinungsformen ko-adressierter Ausdrücke können jedoch sehr unterschiedlich ausfallen. Betrachten wir dazu den folgenden Auszug aus der Ersteinrichtung in Haushalt 8 mit Robin, Samuel und Lara:

Beispiel (54): Wie macht der Otter I (08-EE-17)²⁹³

```

014  RL:  Δdann SPIEL mu:sik stop;
015      °h WIE sagt man (.) wie geht es dir auf
          portugIEsisch?Δ
      sm: Δhebt Bierflasche auf, trinkt und stellt
          sie wieder auf Boden----->
          >-----Δ
016  RL:  wie steht der DAX;
017      [wo is die nächste bäckerEI.          ]
018  LS:  [$(.) <<lachend> °h wie steht der DAX?>$]
      ls:  $lächelt-----$
019  SM:  [$<<lachend, murmelnd> (oh das is)> °h
          klas$][sischer (.)
          klassischer MORgen bei mir.]
      sm:  $schüttelt den Kopf----->
          >---$
020  LS:  [((lacht))
          ][ ja.          ]
021  RL:      [is mein FLUG
          pünktlich?          ]
022  RL:  [$wann ist MORgen mein erster termi:n?$
          ]
      sm:  $wendet sich dem Handy zu-----$
023  LS:  [erstMA aufwachen (.) <<leise nuschelnd>
          (oh ma:n wie steht der dax?)>]
```

²⁹³ Teile dieses Ausschnitts werden unter anderen Gesichtspunkten als Beispiel (57) in Kap. 6.2.2.4 und ausführlich als Beispiele (80–85) in Kap. 6.2.3.8 dieser Arbeit besprochen. Der Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c443&wID=&textSize=1000&contextSize=36.

024 RL: erZÄHL mir etwas lustiges.
 025 **wie kl[ingt ein +Otter.+]**
 rl: +schaut kurz seine
 Mitbewohner an+
 026 SM: [((lacht))]
 027 (1.2)
 028 LS: **HÄ (lass ma fragen.)**

((Ausgelassen bis Z. 061))

061 SM: **HEY google?**
 062 RL: \$((lacht))\$
 ls: \$lächelt--\$
 063 LS: +ne (.) oKE google;+
 ls: +blickt SM an-----+
 064 SM: **wie macht der O[Tter?]**
 065 RL: [geht BEIdes.]
 066 LS: +echt? +
 ls: +blickt RL an+
 067 LS: +bei MIR ging nur oke.+
 ls: +blickt GA an-----+
 068 (1.5)
 069 RL: *also da muss er SÜchen. *
 rl: *zeigt mit RH auf Lampe des GA*
 070 GA: entschuldige (.) das habe ich nicht
 verSTANden,
 071 RL: *hÄ du * \$hast das doch
grad hier als BEIspiel
angegeben.\$
 rl: *bewegt RH Richtung Handy* \$beugt sich
 über Handy----->
 >-----\$

Die WG-Mitglieder sind, nachdem das Gerät in einem ersten Schritt mit dem Smartphone von Robin verbunden werden konnte, in eine Phase des Testens eingetreten. Robin liest dazu verschiedene Beispiel-Stimmeingaben aus der Bedienungsanleitung vor (016–025), dazu gehört u. a. „Wie steht der DAX?“, „Wo ist die

nächste Bäckerei?“ und „Wie klingt ein Otter?“.²⁹⁴ Schon während Robin das Otter-Beispiel vorliest (025), wird dies durch eine auffällige Prosodie sowie durch den kurzen Blickaustausch mit Lara und Samuel relevant gesetzt. Lara äußert daraufhin eine nicht vollständig verständliche Passage, von der anzunehmen ist, dass sie die Otter-Abfrage projiziert. Das Testen wird im Anschluss durch eine plötzliche Wiedergabe von Musik unterbrochen (ausgelassene Stelle, 029–061),²⁹⁵ ehe das Beispiel wieder aufgegriffen wird. Zwischen der Produktion der Invokation von Samuel (061) und der tatsächlichen Stimmeingabe (064) beginnt Lara als zweiten Dialogstrang einen Austausch mit Samuel über die genaue Gestaltung der Invokation. Eine ähnliche Gesprächsdynamik beobachten auch Pitsch et al. (2017: 396) an Stellen, die für die Maschine als Stille erscheinen (sollen), die aber mit interaktionalen Aushandlungen *über* die Bedienung der Maschine gefüllt sind (siehe auch Habscheid 2022: 176). Im gegebenen Beispiel setzt sich dieser Dialogstrang noch über den Abschluss des VUI-Dialogs i. e. S. hinaus fort (065–067). Zu vermuten ist, dass aufgrund der anderen Äußerungen zwischen Invokation und Stimmausgabe die Interpretation der aufgezeichneten Laute fehlschlug und insofern die Fehlermeldung (070) ausgegeben wird.

Interessanterweise produziert daran anschließend Robin einen kohäsiven Turn-Anschluss: „hå du hast das doch grad hier als BEIspiel angegeben“ (071). Damit adressiert er direkt den Smart Speaker bzw. das VUI mit der Hörerdeixis. Auch in diesem Fall muss die Äußerung als doppelt adressiert verstanden werden: Formal gesehen ist sie an das VUI bzw., noch spezifischer, an die zuvor von Robin vorgelesene Bedienungsanleitung, gerichtet. Zwischen Bedienungsanleitung, VUI und der physischen Repräsentation als Smart Speaker wird hier nicht unterschieden, eine genauere Spezifikation, wer mit „du“ gemeint ist, unterbleibt. Funktional dient die Äußerung allerdings vielmehr dem Ausdruck von Nicht-Verstehen gegenüber seinen Mitbewohner*innen: Das VUI wird nicht in den Listening-Modus versetzt – d. h., die Invokationspräsequenz bleibt aus. Der Sprecherbeitrag ist insofern eine Stellungnahme des Nutzers gegenüber der scheinbaren Inkongruenz zwischen Anleitung und Resultat, die auch bzw. primär an die kopräsenten Sprecher*innen gerichtet, dabei aber indirekt adressiert ist (für ein ähnliches Phänomen im Fernsehkontext siehe Klemm 2001: 105). Die Stellungnahme bettet sich in die bereits seit längerem laufende Gesprächsentfaltung über die Otter-Anfrage ein. Genau solche Test- bzw. Übungssituationen, in denen ein längeres Gespräch *über* die Funktionen und die Bedienung des Smart Speakers geführt

²⁹⁴ Abweichend von der von Robin vorgelesenen Bedienungsanleitung („wie klingt ein Otter?“) produziert Samuel die Stimmeingabe „wie macht der OTter“ (066); diese Varianz wird allerdings von den Teilnehmenden nicht reflektiert.

²⁹⁵ Für eine Analyse dieser Passage siehe Kap. 6.2.3.8.

wird, führen in den erhobenen Daten wiederholt dazu, dass sich VUI-Dialog und Gespräch miteinander vermischen, was sich am Split zwischen Form (an das VUI adressiert) und Funktion (an die Mitmenschen adressiert) zeigt.

6.2.2.3 Vorführungen

Ähnliche Konstellationen können auch in Situationen auftreten, in denen Funktionen der Smart Speaker vorgeführt werden. Dazu soll zunächst ein Beispiel aus Haushalt 8 gezeigt werden, das am Tag nach der Ersteinrichtung des Google Assistant aufgezeichnet wurde:

Beispiel (55): Wo ist mein Handy? (08-CVR01-08)

032 RL: oKAY google,
033 (0.7)
034 RL: wer BIN ich?
035 (1.2)
036 GA: dein NAME ist rObin.
037 (1.0)
038 RL: ((lacht)) und der kann auch verstehen (.)
wenn ich sage (.)
wo ist MEIN handy,
039 dann weiss der auch (dat et) MEIN handy
suchen soll.
040 LS: JA_a.
041 aber (.) gestern hat er kurz gesagt (---)
OLli,
042 [((lacht))]
043 RL: [(ja ok aber sowat] GEHT eigentlich.)
044 oKAY google?
045 (1.1)
046 RL: wo ist mein HANdy?
047 LS: (hat OLli?)
048 (0.4)
049 GA: ich kann KEine weiteren [smartphones]
finden,
050 RL: [(ECHT?)]
051 GA: die mit deinem GOOgle konto verknüpft
sind.
052 (1.6)

- 053 RL: waRUM nicht?
 054 Eben gings noch.
 055 oKAY google,
 056 (1.1)
 057 RL: wo ist mein HANdy?
 058 (1.7)
 059 GA: [KLAR.]
 060 k: [((Klingeln setzt ein))]
 061 GA: dein redmi <<deutsche Aussprache> note>
 zehn [pro sollte
 jetzt KLINGeln.]
 062 RL: [stop (.)
 stop (.) oKAY google (.) stop.]
 063 (0.8)
 064 LS: UH.

Lara und Robin diskutieren die Funktionen des Google Assistant. Dabei liegt ein Schwerpunkt auf den Personalisierungsfunktionen des Geräts: Robin zeigt sich (schon vor Beginn des gezeigten Ausschnitts) davon beeindruckt, dass das Gerät anhand der Stimmprofile zuordnen kann, wer er ist, und führt dies für Lara noch einmal vor (032–036). Eine weitere Funktion, die damit einhergeht, ist die Möglichkeit, deiktische Ausdrücke bis zu einem gewissen Grad zu verarbeiten, etwa die kontextgebundene Bedeutung des Possessivpronomens „mein“ (siehe 038), auch bei mehreren registrierten Nutzer*innen. Lara weist darauf hin, dass die Funktion nicht zuverlässig sei und dass es gestern eine Verwechslung gegeben habe (041), woraufhin Robin die Funktion noch einmal vorführen will (044–046). Die Eingabe wird scheinbar auf dem akustischen Kanal verstanden, gleichwohl liefert der Smart Speaker nicht das gewünschte Ergebnis, sondern eine Fehlermeldung, die erkennen lässt, dass die Funktion erkannt wurde, aber entsprechende Spezifikationen (mit dem Google-Konto verknüpfte Smartphones) fehlen (049–051). Robin ‚antwortet‘ daraufhin dem VUI mit einer Rückfrage („waRUM nicht?“, 053). Er produziert zuvor keine Invokation, sodass davon auszugehen ist, dass die Äußerung zwar als eine Adjazenzellipse zu bestimmen ist (siehe Klein 1985; Knobloch 2013), mit der ein „elliptischer Anschluss“ (Brinker 2010: 33) realisiert wird. Sie ist aber dennoch funktional nicht an das VUI gerichtet, sondern bringt gegenüber Lara Robins Verwunderung indirekt zum Ausdruck. Er unternimmt sogleich einen neuen Versuch, der gelingt (057–062).

Die direkte Anrede in der zweiten Person (abseits von Imperativen) findet sich darüber hinaus in weiteren Beispielen, in denen das VUI vorgeführt wird. Dazu ist ein Auszug aufschlussreich, der aufgezeichnet wurde, als der Nutzer zeitgleich an einem Interview teilnahm, das im mediensoziologischen Projektteil des eingangs

Sam führt das Interview per Videotelefonie im Nebenraum auf einem Laptop. Das Interview neigt sich dem Ende entgegen, Gegenstand waren an dieser Stelle Beleidigungen, die an das VUI adressiert werden – dies möchte er der Interviewerin vorführen und betritt dazu den Raum, in dem sich der Smart Speaker und seine Lebenspartnerin Alexandra befinden, was durch kontextualisierende Geräusche und einen kurzen Austausch mit Alexandra kurz vor Beginn des Auszugs deutlich wird. Ohne weitere Einleitung adressiert er das VUI zweimal (043, 059) mit der Du-Anrede in Satzerststellung, die in Kombination mit einem Prädikativum von stark invektivem Charakter auftritt. Aus dem Zusammenhang der Äußerung wird deutlich, dass dies geschieht, um dies der Interviewerin vorzuführen. Durch eine sequenzielle Betrachtungsweise des Ausschnitts lässt sich bestätigen, dass das VUI die Äußerungen auch als invektiv erkennt – und sie nach Ellerbrock et al. (2017: 8) somit erst dazu macht: Invektiv sind demnach solche Äußerungen, die „von den Beteiligten oder Zuschauenden als solche aufgefasst werden [...]“. Dies ist hier, wenn auch verklausuliert, der Fall: Einerseits weist das VUI auf die Möglichkeit hin, Feedback zu geben (045–054), und schlägt auch vor, wie dies geäußert werden kann – damit wird der zuvor von Sam geäußerten Invektivform eine konstruktive Intention unterstellt, die vorherige Äußerung aber zurückgewiesen und nicht als Startpunkt eines Feedbacks angesehen. Nach der wiederholten Beleidigung (059) sendet das Gerät ein Geräusch, das Sam als „ausmachen“ deutet (065) – auch dieses Geräusch kann als Earcon konzeptualisiert werden.²⁹⁷ Die leichte Dissonanz in der Tonfolge sowie die absteigende Tonhöhenbewegung können darauf verweisen, dass etwas erfolglos abgebrochen wurde (vgl. Blattner/Sumikawa/Greenberg 1989: 30; Grond/Hermann 2012: 57; Oswald 2012: 254), so wie hier der Versuch einer ‚Unterhaltung‘ beendet wird und dies auf die invektive Äußerung zurückgeführt werden kann. Die Invektive stehen hier, wie es auch Lotze (2021: 320) beschreibt, im Kontext des Testens und Vorführens und werden gebraucht, „um Reaktionen zu evozieren, die die Grenzen des Dialog-Designs aufzeigen und so einen ersten Eindruck von den Funktionsweisen des Systems vermitteln“ (Lotze 2021: 320).

Allerdings können auch Vorführsituationen auftreten, in denen die ‚Beteiligung‘ des VUI anders konfiguriert wird: Der folgende Ausschnitt aus einer CVR-Aufnahme, die am Tag nach der Ersteinrichtung des HomePods in Haushalt 7 aufgezeichnet wurde, soll dies noch einmal illustrieren, auch wenn das Beispiel bereits ganz zu Beginn des Analyseteils diskutiert wurde.²⁹⁸

²⁹⁷ Zu Earcons siehe die Analyse zu Beispiel (13) („Lieferstatusupdate“) in Kap. 6.1.1.3.

²⁹⁸ Für eine Analyse des Beispiels unter dem Aspekt der Anredeformen siehe Kap. 6.1.1.

Beispiel (3): Reaktion (07-CVR01-03)

007 KW: °h hey SIri?
 008 (2.5)
 009 SI: mh_HM?
 010 (1.1)
 011 KW: hh° <<lachend> (spie,) > SPIELe: (-) get
 into it (-) vo::n (--) doja cat.
 012 (2.9)
 013 SI: ich spiele get into it yuh von DOja cat
 jetzt.
 014 (1.5)
 015 KW: voll FUNny wenn man nicht (--) direkt spricht-
 016 hast du geHÖRT grade?
 017 TW: JA:.
 018 KW: hast du gehört was die geMACHT hat?
 019 (0.7)
 020 TW: ich SPIELe,
 021 KW: NEI:N davor,
 022 (0.3)
 023 TW: NEIN.
 024 KW: ich hab grad ihren NAMen gesagt,
 025 TW: und dann hat die [geSAGT-]
 026 KW: [und hab dann NICH]
 sofort reagier:rt.
 027 weil ich nicht wusste ob sie schon ZUHört.
 028 (0.3)
 029 TW: JA,
 030 (0.4)
 031 KW: [ich mach das NOCHmal,]
 032 k: [((“Get into it” von Doja Cat setzt ein))]
 033 (0.8)
 034 KW: hey SIri?
 035 k: ((Musik spielt, pausiert kurz und beginnt
 erneut; gleichzeitig gelegentliches leises
 Klappern und Rascheln 7 Sek.))
 036 KW: die hat das NICHT gehört.
 037 HEY siri,
 038 TW: <<creaky> (ach GEIL,) >

039 k: ((Klappern)) [((Musik endet))]
 040 KW: [PAUse?]
 041 (2.9)
 042 k: ((“Get into it” von Doja Cat setzt wieder ein))
 043 (0.8)
 044 KW: hey SIri?
 045 (1.3)
 046 KW: STOP.
 047 [(1.9)]
 048 k: [((Musik endet))]
 049 KW: jetzt SCHAU ma.
 050 wie FUNNy.
 051 TW: [(STOP.)]
 052 k: [((Klappern))]
 053 (2.2)
 054 KW: warte äh (.) a_S0.
 055 (0.3)
 056 KW: hör ZU,
 057 HEY siri,
 058 (2.5)
 059 SI: HM_mh?
 060 KW: °h °h °h PLAY;
 061 (0.7)
 062 k: ((“Get into it” von Doja Cat setzt ein))
 063 KW: hast du_s geHÖRT?
 064 (0.7)
 065 TW: JA:.
 066 (0.5)

Der Anwender Konrad entdeckt eine Funktion des HomePod, die hier bereits diskutiert wurde: Das VUI produziert bei einer erkannten Invokation ca. 2,5 Sekunden nach dieser ein Rezeptionssignal (009). Konrad ist von dieser Funktion überrascht und will sie anschließend Till vorführen. Er versucht es zunächst mit einer Erzählung, für die er sich ein „ticket“ (Sacks 1972; Quasthoff 2001) einholt (015–029), doch entscheidet sich dann dafür, es Till noch einmal vorzuführen. In der Vorführung wird das VUI in der dritten Person besprochen, wobei alle auf das VUI bezogenen Pronomen mit weiblichem Genus produziert werden („die“, 018,

„ihren“, 024, „die“, 025, „sie“, 027).²⁹⁹ Nachdem Konrad sichergestellt hat, dass Tills Aufmerksamkeit fokussiert ist (056), produziert er eine Invokation (057) und wartet die entsprechende Zeit ab, bis das VUI ein Rezeptionssignal produziert. Er holt anschließend noch einmal eine Ratifizierung von Till ein (063–065). Diese Form der Vorführung ‚beteiligt‘ das VUI zwar an der sozialen Praxis: Die Soundproduktionen des VUI sind Gegenstand des Gesprächs und die Bedienung des Geräts im Mittelpunkt jedenfalls von Konrads Aufmerksamkeit (bei Till kann aufgrund der mehrfachen Fokussierungsaufforderungen von Konrad sowie dem ersten ‚Verpassen‘ der zu demonstrierenden Stelle vermutet werden, dass er parallel anderweitig beschäftigt ist). Es wird nicht *mit* dem VUI gesprochen, sondern *über* das VUI. Das Reden *mit* dem VUI ist vielmehr Gegenstand einer Demonstration, an der das VUI und dessen Äußerungen insoweit ‚beteiligt‘ sind, wie es für die Demonstration notwendig ist.

6.2.2.4 Ko-Evaluationen

Auch bei gemeinsamen Bewertungen der akustischen Äußerungen und der Funktionsweise des Smart Speakers scheint eine entsprechend hohe Fokussierung der Unterhaltung auf die Bedienung des Smart Speakers gegeben zu sein, sodass das Gerät vorübergehend sprachlich wie ein Teilnehmer des Gesprächs behandelt wird. Allerdings ist dieser Status ebenso brüchig, wie die Fortsetzung des oben gezeigten Beispiels (54) („Wie macht der Otter I“) zeigt:

Beispiel (57): Wie macht der Otter II (08-EE-17)³⁰⁰

```
061 SM: HEY google?
062 RL: $((lacht))$
      ls: $lächelt--$
063 LS: +ne (.) oKE google;+
      ls: +blickt SM an-----+
064 SM: wie macht der O[Tter?      ]
065 RL:                               [geht BEIdes.]
```

²⁹⁹ Für kurze Ausführungen zum Genus der jeweils verwendeten Personalpronomen siehe Kap. 7.

³⁰⁰ Teile dieses Ausschnitts werden unter anderen Gesichtspunkten als Beispiel (54) in Kap. 6.2.2.2 und ausführlich als Beispiele (80–85) in Kap. 6.2.3.8 dieser Arbeit besprochen. Er kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c443&wID=&textSize=1000&contextSize=36.

066 LS: +echt? +
 ls: +blickt RL an+
 067 LS: +bei MIR ging nur oke.+
 ls: +blickt GA an-----+
 068 (1.5)
 069 RL: *also da muss er SUCHen. *
 rl: *zeigt mit RH auf Lampe des GA*
 070 GA: entschuldige (.) das habe ich nicht
 verSTANDen,
 071 RL: *hä du * \$hast das doch
 grad hier als BEIspiel
 angegeben.\$
 rl: *bewegt RH Richtung Handy* \$beugt sich
 über Handy----->
 >-----\$
 072 LS: oKE google;
 073 wie macht der OTter,
 074 (1.6)
 075 GA: dAs ist ein OTter.
 076 \$((spielt Quieken eines Otters ab))\$
 rl: \$öffnet und schließt RH, wendet Kopf nach
 links\$
 077 SM: [((lacht leise))]
 078 RL: [((lacht leise))]
 079 LS: [((lacht leise))]
 080 SM: [+wär lustiger wenn_s IHre stimme wäre;+]
 ls: +lächelt-----+
 081 GA: [(spielt Quieken eines Otters ab))]
 082 SM: [((lacht))]
 083 LS: [\$<<lachend> JA ne?>\$]
 ls: \$wendet sich SM zu-\$
 084 RL: ja.
 085 SM: [((lacht))]
 086 LS: [<<lachend> ich hab auch eigentlich damit
 gerechnet dass es> ihr
 SIE macht.]

In der weiter oben präsentierten Analyse³⁰¹ lag die Aufmerksamkeit auf der Du-Adressierung des VUI durch Robin (073), die eine vorübergehende, formale Beteiligung des VUI an der sprachlichen Oberfläche erkennen ließ. Die Bewertungen von Samuel (080) und Lara (083, 086) zeigen, dass über das VUI aber auch in der Objektdeixis gesprochen wird. Dabei wird zwar eine Form der Persönlichkeitsbildung beibehalten – es wird präsupponiert, dass es ‚hinter‘ dem Otter eine andere Persona gibt, die eine eigene (technisch synthetisierte) Stimme hat, mit der der Otter-Sound wiederum hätte simuliert werden können. Diese ist jedoch kein*e ‚Teilnehmer*in‘ am Gespräch: Es wird über diese Persona geredet und zugleich die Performance des VUI bewertet. Der folgende Ausschnitt illustriert noch einmal, dass das VUI nicht als Gesprächsteilnehmer behandelt wird, sondern stattdessen von den menschlichen Teilnehmer*innen gemeinsam vollzogene Evaluationen der Äußerungen des VUIs erfolgen:

Beispiel (58): Du bist doof (07-CVR01-15)

039 KW: hey SIri:,
 040 intercom an TILL,
 041 du bist DOO:F.
 042 (1.4)
 043 SI: [ich kann keine intercom durchSAge,]
 044 k: [((Musik stoppt))]
 045 SI: an deine konTAKte mAchen,
 046 KW: hö Hö.
 047 k: ((Musik setzt wieder ein))
 048 KW: <<lachend> ((unverständlich, ca. 3.2
 Sek.))>
 049 (1.1)
 050 KW: ((lacht))
 051 °h das war richtig SCHNIPpisch oder?
 052 TW: <<lachend>ja_a.>
 053 KW: <<h, all>ich <<creaky>kann> kEIne intercOm
 durchsage an dEIne konTAKte machen;>
 054 (10.8)

301 Siehe die Analyse zu Beispiel (54) („Wie macht der Otter I“).

Konrad startet einen VUI-Dialog, in dem er scherzhaft die „Intercom“-Funktion³⁰² nutzt, um seinem ko-präsenten Mitbewohner und Bruder mitteilen zu lassen, dass er doof sei (041). Allerdings gibt das VUI eine Fehlermeldung aus, die anschließend durch Lachpartikeln als lustig sowie von Konrad als „SCHNIPPisch“ (051) bewertet wird; Till ratifiziert diese Bewertung von Konrad (052). Anschließend gibt er die Stimmausgabe selbst noch einmal wieder (053), wobei durch eine prosodisch modulierte und auffallend schnelle Sprechweise die Äußerung als stilisierte direkte Rede gerahmt wird (vgl. Günthner 1997: 257). Konrad und Till evaluieren die Äußerungen des VUI im Hinblick auf dessen Prosodie, die bestimmte Gefühlszustände oder Charaktereigenschaften transportiere. Die Redewiedergabe wird hier also in den Dienst der Evaluation der Äußerungen insbesondere im Hinblick auf die Sprechweise gestellt und für diesen Kontext mit prosodischen Modifikationen, die dieser kommunikativen Funktion entsprechen, rekontextualisiert (vgl. Günthner 1997: 229). Auch dieses Beispiel zeigt: Wenn das Reden über die Eigenschaften des VUI sich mit den zwischenmenschlichen Gesprächen verbinden, kann es zu kurzfristigen Attribuierungen menschlicher Eigenschaften kommen (siehe auch Pins et al. 2020: 10), wenn auch nicht in Form einer Behandlung als Gesprächsbeteiligte*r. Vielmehr werden diese Praktiken als „Interfacing“ verstanden: „practices that both separate *and* [Herv. i. O.] connect“ (Lipp/Dickel 2022: 15).

6.2.2.5 Interaktionale Bearbeitung von Fehlschlägen

Die gezeigten Situationstypen sollen um eine weitere charakteristische Situation ergänzt werden, die sich bereits bei der genaueren Betrachtung dyadischer VUI-Dialoge als auffällig herausgestellt hat: die interaktive Bearbeitung von Reparaturstellen bzw. Fehlschlägen.³⁰³ Wie sich in den Daten zeigt, scheinen Fehlschläge in Mehrparteieninteraktionen ebenfalls einen lokalen, temporal beschränkten Wechsel des Teilnahmestatus auslösen zu können. Der folgende Auszug soll dies exemplifizieren:

Beispiel (59): Ich hab Hunger (08-CVR01-05)

157 RL: <<lauter> oKAY google,>
 158 (0.3)
 159 LS: ((lacht))
 160 (0.5)
 161 RL: <<lauter> welche terMIne stehen in meinem

³⁰² Siehe dazu Beispiel (4) („Intercom“) in Kap. 6.1.1.1.

³⁰³ Siehe Kap. 6.1.4 Reparaturen.

kalender?>
 162 (1.7)
 163 GA: entSCHULDigung.
 164 ich kann KEIne einträge in deinem
 [kalender finden.]
 165 RL: [Alte:r-]
 166 SM: die ist echt SCHLECHT ne,
 167 RL: die ist ECHT schon scheiße ja,
 168 (0.3)
 ((ausgelassen bis Z. 225))
 225 SM: oKE google,
 226 (1.1)
 227 SM: ich würde gerne etwas bei goRILLas
 <<lachend> bestellen.>
 228 (0.3)
 229 RL: ((unverständlich, ca 2 Sek.)) (aber
 IRgendwann geht das.)
 230 (0.7)
 231 GA: entSCHULDigung.
 232 ich habe dich nicht [verSTANDen.]
 233 SM: [JA ich hab_s.]
 234 mir ja schon geDACHT.
 235 RL: oKAY google,
 236 (1.0)
 237 RL: ich hab HUNger,
 238 (2.9)
 239 GA: ich kann dir KEIne karten anzeigen.
 240 weil das gerät keinen BILDschirm hat.
 241 (1.4)
 242 RL: wat KANNST du denn eigentlich?

Wie zunächst aus 157–168 hervorgeht, testen die WG-Mitglieder in Haushalt 8 die Funktionen des neu eingerichteten Google Assistant. Die hier gezeigte Terminabfrage steht in einer Reihe von anderen Abfragen und wird schließlich unter Verwendung der Objektdeixis im Femininum evaluiert: „die ist echt SCHLECHT ne,“ – „die ist ECHT schon scheiße ja,“ (166–167). In die Reihe von Test-Stimmeingaben fallen auch zwei Abfragen aus dem Bereich der Lebensmittelversorgung: „ich

würde gerne etwas bei gorILLas³⁰⁴ bestellen“ (227) und „ich hab HUNger“ (236). Beide Anfragen kann das VUI nicht verarbeiten, auf die erste hin produziert es die Standardfehlermeldung (231–232) und auf die zweite Anfrage hin wird die Schwierigkeit ausgedrückt, lokalitätsbezogene Ergebnisse zu präsentieren, weil eine visuelle Oberfläche fehlt („weil das gerät keinen BILDschirm hat.“, 239). Bemerkenswert sind die beiden daran anschließenden Äußerungen von Samuel und Robin. Beide Äußerungen sind kohäsiv und als kohärente Turns gestaltet. Samuels Äußerung „JA ich hab_s. / mir ja schon geDACHT.“ (233–234) weist mit dem Diskursmarker „ja“ eine Turn-Übernahmebestätigung auf, die „als Mittel der Verknüpfung von zwei interaktionalen Einheiten“ (Meer 2007: 6) beschrieben wurde (siehe auch Meer 2009; Betz 2017). Der syntaktisch im Vor-Vorfeld platzierte Diskursmarker hat damit retraktives und projektives Potenzial: Samuel bestätigt die (frühzeitige, teilweise noch mit den Äußerungen des VUI überlappende) Turn-Übernahme und projiziert damit eine Positionierung zur vorhergehenden Äußerung. Außerdem ist im weiteren Verlauf Samuels Äußerung auch durch die klitisierte Pronominalreferenz „hab's“, die in Bezug auf die vom VUI geäußerte Proposition steht, mit der vorhergehenden Äußerung verbunden. Die Äußerung ist auch mit Blick auf die Domestizierung des Smart Speakers interessant: Vermutungen über die Funktionsweise des VUI, die die Anwender*innen möglicherweise noch in der Phase vor dem Kauf entwickelt haben, werden getestet – es zeigt sich also eine Verbindung zwischen der Kommodifizierung und der Aneignung (vgl. Hartmann 2013a: 21–22).³⁰⁵ In diesem Prozess werden Grenzen ausgelotet und durch Wiederholung immer detaillierter abgesteckt.

Auch die Äußerung von Robin auf den zweiten Fehlschlag hin scheint unter mehreren Gesichtspunkten als kohärenter Anschluss-Turn gestaltet zu sein: Er verwendet die Hörerdeixis und greift zudem sprachliches Material aus der vorangehenden Äußerung des VUI auf, indem er das Lexem „können“ erneut verwendet, aber in die zweite Person stellt, und auch durch die Fokusakzentuierung prosodisch betont (242). Kohärenz wird außerdem durch die Verwendung der Modalpartikel „denn“ erzeugt, die ebenfalls retraktiv gesprächsorganisatorisch wirkt und zudem eine Antwort des VUI gesteigert konditionell relevant setzt (vgl. Deppermann 2009: 23–24). Formal und lokal gesehen ratifizieren damit beide Äußerungen die Teilnahme des VUI am aktuellen Gespräch. Im größeren diskursiven Zusammenhang wird allerdings deutlich, dass die Äußerungen, ob schon sie formal an die Äußerungen des VUI angebunden sind, sich auf der pragmatisch-funktionalen Ebene an die ko-präsenten Sprecher*innen (Samuel und

304 *Gorillas* ist ein in Großstädten operierender Lieferdienst für Lebensmittel, Drogerieartikel und andere Produkte für kurzfristige Online-Bestellungen aus Supermärkten.

305 Siehe Kap. 2.3.2.1.

Robin, vermutlich auch Lara und Alexandra) richten und Überraschung und Frust über den limitierten Funktionsumfang des Smart Speakers ausdrücken.

Ähnlich stellt sich auch das folgende Beispiel dar: Konrad ist vor und nach dem gezeigten Ausschnitt dabei, eine Geschichte vom letzten Abend zu erzählen (094), parallel dazu läuft über den Smart Speaker Musik:

Beispiel (60): Fräulein (07-CVR01-15)

094 KW: =und und ich stand an der BAR-
 095 ach hey SIri mach doch mal n bIsschen
 LEIser?
 096 (1.4)
 097 TW: JA.
 098 KW: ((unverständlich, 3.5 Sek.)) hey si:ri:
 mach mal bisschen LEIser bitte.
 099 k: ((Musik spielt unverändert))
 100 KW: siri mach mal ruHIG,=
 101 =n gu:tes stück LEIser?
 102 k: [((Musik bleibt unverändert))]
 103 [(0.8)]
 104 KW: hey SIri::?
 105 =mach ma_n gutes stück LEIser bitte.
 106 k: ((Musik spielt unverändert))
 107 KW: MEINST ((unverständlich, 1.2 Sek.))
 108 (2.5)
 109 KW: hey siri (.) kannst du flüsTERN?
 110 (1.2)
 111 SI: mh.
 112 (0.6)
 113 SI: [DArauf hab ich keine antwort.]
 114 k: [((Musik endet))]
 115 (0.3)
 116 SI: gibt es etwas ANderes,
 117 =bei dem ich helfen KANN?
 118 TW: NE:: danke.
 119 KW: NE: danke.
 120 k: ((Musik spielt leise))
 121 SI: GERN geschehen.
 122 k: ((Musik wird wieder lauter))
 123 KW: FRÄUlein.

Unter Verwendung zahlreicher Modalpartikeln (z. B. „doch“, „mal“, 095) sowie Tilgungen und unscharfer Ausdrücke („n bisschen“, 095, „bisschen LEIser“, 098, „n gutes stück“, 100) versucht Konrad, die Lautstärke zu senken, um – so legt es der Kontext der Erzählung nahe, die im Anschluss an diese Episode wieder aufgenommen wird – mit der Narration nicht fortgesetzt gegen die Musik anreden zu müssen. Allerdings passt das VUI die Lautstärke der Musikwiedergabe nicht an, möglicherweise weil die Ausdrücke semantisch zu unscharf sind und insofern nicht verarbeitet werden können. Da eine Stimmausgabe des VUI gänzlich ausbleibt, lässt sich die Fehlerquelle nicht ermitteln. Scheinbar von diesen Fehlschlägen angeregt, richtet Konrad eine Frage an Till (107), die jedoch auf der Aufzeichnung wegen der lauten Musik nicht gänzlich verständlich ist – es kann angenommen werden, dass er mit dieser Frage die anschließende Stimmeingabe projiziert, die eine „Flüster-Funktion“ abfragt.³⁰⁶ Zunächst wird eine Fehlermeldung wiedergegeben, das VUI initiiert daraufhin allerdings eine Expansion der *Basis-Sequenzstruktur* und fragt, ob andere Funktionen gewünscht werden (116–117). Beide Sprecher verneinen dies und lehnen unter Verwendung derselben Routineformel (vgl. Coulmas 1979; siehe auch Stoltenburg 2009: 256–258) ab (vgl. 118–119). Das VUI produziert daraufhin noch einmal eine Schließung der „Danksequenz“ (Brinker/Sager 2010: 101), sodass sich ein Dialog aus insgesamt sieben Zügen ergibt (Invokation, Aktivierung des Listening-Modus, Spracheingabe, Sprachausgabe, Expansion, dankende Ablehnung, Danksequenzschließung). Diese mehrzügige, gesprächsähnliche Verfertigung der Flüsterabfrage wirkt sich – in Verbindung damit, dass Konrad möglicherweise ohnehin den Smart Speaker testen will, worauf die eingangs erwähnten sprachlichen Auffälligkeiten bei der Bedienung des VUI hindeuten – scheinbar auf den Teilnahme-status aus, den Konrad mit der direkten Anrede „Fräulein!“ noch einmal als gegeben ratifiziert, bevor er seine Erzählung wieder aufnimmt. Das Lexem „Fräulein“ ist seit den 70er-Jahren als Anrede in formalen Kontexten nicht mehr gebräuchlich – es trägt „Informationen der sexuellen Zugänglichkeit“ (Nübling 2011: 352) und zeichnet sich durch eine starke Funktionalisierung, d. h. die Reduktion der angesprochenen Person auf ihre Funktion, aus; ein Phänomen, das Nübling (2011: 346) als eine mögliche Beschaffenheit von Pejorisation disku-

306 Der Smart Speaker von Amazon, mit dem Konrad und Till Vorerfahrungen haben, kann einen entsprechenden Modus aktivieren, in dem die verbalsprachlichen Äußerungen des VUI durch prosodische Merkmale als „geflüstert“ wiedergegeben werden; dazu trägt nicht nur eine stark abgesenkte Lautstärke, sondern auch stimmloses, stark behauchtes Sprechen bei.

tiert.³⁰⁷ Dass Konrad hier das Lexem „Fräulein“ verwendet, deutet also darauf hin, dass gegenüber dem ko-präsenten Till eine negative Evaluation der Leistungen des VUI angezeigt wird, die auf der pejorativen, geschlechterspezifischen Anrede an das VUI beruht. Außerdem wird durch die Funktionalisierung eine Nähe zu Dienstpersonal hergestellt, die auch Lind/Dickel (2023: 10) sowie Dickel/Schmidt-Jüngst (2021: 360) in Inszenierungen von Smart Speakern in Werbevideos beschreiben. Der Ausdruck kann als Domestizierung im Sinne einer Eingliederung des Interfaces in die Konfiguration des Haushalts verstanden werden: Er verdeutlicht auf verschiedenen Ebenen die soziale Position des VUI als eine Entität, gegenüber der pejorative Begriffe und genderspezifische Abwertungen verwendet werden können, ohne dass dies Auswirkungen auf das eigene „Face“ oder das des Gegenübers hätte.

Die bisher gezeigten VUI-Dialoge oszillieren an der formalen Oberfläche schnell zwischen unbelebtem Objekt und einem ‚Beteiligten‘. Wird ein Status als ‚Gesprächsteilnehmer‘ formal zugeschrieben, ist dies brüchig und lokal begrenzt. Weder behandeln Nutzer*innen die Geräte im engeren Sinne wie Subjekte, noch übertragen sie in den hier gezeigten Fällen „mindless“ (Nass/Moon 2000) Kategorien aus der zwischenmenschlichen Interaktion – weder als anthropomorphisierende Äußerungen noch als gedankenlose Übertragungen sozialer Skripte auf den Austausch mit maschinellen Dialogpartnern ist dieses Phänomen hinreichend erfasst (vgl. Reeves/Porcheron 2023). Vielmehr ist die Gestaltung der Äußerungen, die formal eine ‚Beteiligung‘ der VUIs erkennen lassen, ein Stilmittel, mit dem die Geräte angeeignet werden, mit dem Stellung bezogen werden kann und mit dem sie als Träger an der *Praxis* partizipieren – das macht sie nicht zu Gesprächsteilnehmern. Daher erklärt sich auch, dass die Sprecher*innen zwischen diesen Modi schnell hin und her wechseln können: Der soziale oder gar der ontologische Status der Geräte ändert sich nicht durch entsprechende Ausdrücke. Sie sind insofern von großem gesprächsanalytischen Interesse und sollen im nachfolgenden Kapitel genauer entlang der sprachlichen Praktiken aufgeschlüsselt werden.

6.2.3 Sprachliche Praktiken in der (Nicht-)Beteiligung von VUIs

Im vorangehenden Kapitel wurden verschiedene Situationstypen vorgestellt, in denen VUIs ins Zentrum sozialer Praxis gerückt wurden, sodass sich der VUI-

³⁰⁷ Zunehmend scheint der Begriff eine Renaissance als Selbstbezeichnung weiblich geleiteter Account-Inhaberinnen auf Social Media zu erleben, teilweise dialektal durchfährt („Frollein“). Dieses Phänomen bedarf noch einer genaueren, korpusgestützten Untersuchung und könnte auch die vorliegende Analyse erhellen.

Dialog und das Reden sozialer Interaktion miteinander verwoben haben. VUI-Dialoge wurden dabei für den Vollzug sozialer Praxis funktionalisiert und daran beteiligt. Im folgenden Kapitel wird diese Perspektive noch einmal in Richtung der sprachlichen Realisierung ‚gedreht‘: Der Blick richtet sich auf sprachliche Praktiken, mit denen die Beteiligung oder Nicht-Beteiligung an sozialer Praxis und Gespräch vollzogen wird. Der theoretisch eingeführte Begriff der sprachlichen Praktiken, wie ihn Selting (2013; 2016) gebraucht,³⁰⁸ dient dabei als Grundlage. Einige der im Folgenden beschriebenen Phänomene zeigten sich bereits in den vorangehenden Analysen – sie werden nun besonders unter dem Aspekt der sprachlichen Gestaltung betrachtet. In den Kapiteln 6.2.3.6 und 6.2.3.7 sollen auch sprachliche Verfahren diskutiert werden, in denen die *Nicht*-Beteiligung des VUIs am Gespräch besonders deutlich wird.

6.2.3.1 Ko-Adressierungen (Hörerdeixis)

Charakteristisch sind doppelt adressierte Du-Anreden wie im folgenden Beispiel. Der Aufnahmeausschnitt stammt aus Haushalt 8 und wurde kurz nach der Ersteinrichtung aufgezeichnet:

Beispiel (61): Taub (08-CVR01-05)

281 (2.7)
 282 RL: oKAY google.
 283 (0.6)
 284 RL: warum ist [alexa SCHLAUer als]
 [du?]
 285 SM: [((unverständlich, ca. 1 Sek.))]
 286 LS:
 [(soll ICH das] jetzt zusAmmen machen;)
 287 (0.3)
 288 SM: (das is GLAUB ich ne sehr) [gu-]
 289 GA: [tut mir LEID,]
 290 (0.2)
 291 GA: das habe ich nicht verSTANDen.
 292 RL: **ja:_a (.) und TAUB bist_e auch.**
 293 (2.2)

308 Siehe Kap. 2.1.3.

Robin aktiviert den Google Assistant im Zuge fortgesetzter Tests, in dem es kurz zuvor zur geteilten Evaluation kam, dass der Smart Speaker von Amazon besser sei, mit dem zwei der Nutzer*innen Vorerfahrungen hatten.³⁰⁹ Dies ‚verarbeitet‘ Robin in einer Stimmeingabe (284), mit der er in Form einer Ergänzungsfrage nach den Gründen für die unterstellte Unterlegenheit des Systems fragt. Das VUI produziert die Standardfehlermeldung (289–291), während im Hintergrund zu hören ist, dass Lara und Samuel zu einem anderen Thema sprechen. Robin produziert den daran anschließenden Turn „ja:_a (.) und TAUB bist_e auch.“ (292), der in mehrfacher Hinsicht interessant ist. Zunächst ist bemerkenswert, dass er erneut den bereits gezeigten Split in der Adressiertheit aufweist: Formal ist der Turn unter Verwendung der Du-Anrede an das VUI gerichtet, funktional jedoch richtet sich die Äußerung vielmehr an die ko-präsenten Mitbewohner*innen – der Äußerung ist auch keine Invokationssequenz vorausgegangen. Darüber hinaus ist auffällig, dass erneut die Diskurspartikel „ja“ produziert wird, die hier nicht nur die Turnübernahme anzeigt. Mit Meer (2007: 22) kann argumentiert werden, dass *ja*, wenn es nicht als bestätigende Reaktion auf eine vorangehende Proposition gelesen werden kann, eine diskursorganisatorische Funktion hat, die eine formale Kohäsion mit dem vorhergehenden Turn herstellt, im Hinblick auf die projizierte Äußerung jedoch semantisch offenbleibt und in Talkshows daher aufgrund deren „präferiert konfrontativen Ausrichtung“ auch von „Widerspruch, Vorwürfen oder Gegenargumenten“ (Meer 2007: 22) gefolgt werden kann. Die Kohäsion stiftende Funktion verliere es deshalb nicht (vgl. Meer 2007: 22). Die Verwendung von *ja* tritt hier mit einer Folgeäußerung von stark invektivem Potenzial auf („und taub bist_e auch.“, 292), die zusätzlich durch den Konnektor „und“ sowie durch das Adverb „auch“ den Bezug zur vorherigen Proposition des Nicht-Verstehens herstellt. Diese Äußerung scheint im gegebenen Zusammenhang nicht dispräferiert zu sein, weder vor dem VUI noch vor den anderen ko-präsenten und teilweise auch in die Bedienung des VUI eingebundenen Sprecher*innen. Das zeigt noch einmal, dass gegenüber der adressierten „Persona“ kein bzw. nur sehr wenig „Face-Work“ oder Höflichkeitsarbeit (Goffman 1955; Brown/Levinson 1987) betrieben wird. Dies wird ebenso im folgenden Beispiel deutlich, in der das VUI bei einer miteinander verketteten Abfolge von Stimmeingaben schließlich die Kohärenz nicht mehr herstellen kann:

309 Siehe Beispiel (59) („Ich hab Hunger“).

Beispiel (62): Du bist so dumm (08-CVR01-08)

106 LS: oKAY google,
 107 [fünf lauten (.) STärken (.) leiser.]
 108 k: [((Radio pausiert))]
 109 RL: (xxxx xxx)
 110 (2.0)
 111 LS: [((stöhnt))]
 112 GA: [entSCHULDigung.]
 113 die HöCHste lautstärke ist zehn.
 114 auf welche stufe möchtest du die
 LAUTstärke stellen?
 115 (0.2)
 116 RL: DREI.
 117 (3.0)
 118 GA: oKE.
 119 (0.4)
 120 GA: DREI von britney spears.
 121 (0.3)
 122 LS: [((lacht))]
 123 GA: [in der PLAYlist von spotify.]
 124 **RL: du bist so DUMM,**
 125 [((lacht))]
 126 LS: [((lacht))]
 127 k: ((Lied "3" von Britney Spears setzt ein, 2 Sek.))

Lara unternimmt zu Beginn des Ausschnitts den Versuch, die Lautstärke zu regulieren (107). Sie verwendet dazu Angaben auf einer genauen Skala („fünf lauten (.) STärken“, 107). Das VUI kann die Anfrage offenbar nicht verarbeiten und produziert zunächst eine Entschuldigung (112), die ebendies anzeigt. Sodann wird Wissen über die Gestaltung der Lautstärkeregelung dargestellt (113), was erkennen lässt, dass der Operationsbereich der Lautstärkeregelung bereits identifiziert wurde, aber die Details in der tatsächlich durchzuführenden Operation nicht registriert wurden (114). Die folgende Frage des VUI überschreibt den komparativen Charakter in der Äußerung von Lara, der durch die Verwendung der Steigerungsform („leiser“, 107) entsteht, und macht eine numerische Angabe der Anwender*innen erwartbar. So wird Lara ermöglicht, eine besser verarbeitbare Eingabe zu formulieren – und es wird durch das VUI eine neue Paarsequenz eröffnet, die zwar in Teilen auf den Eingaben der Anwender*innen aufbaut, das Verfahren aber sprachlich abwandelt und dadurch in die Systemstrukturen ,ein-

passt'. Erwartbar wäre nun, dass nach dieser Einschubsequenz zur Reparatur der zweite Paarteil zum ersten, von Lara produzierten Paarteil erfolgt, in dem die Lautstärke des Smart Speakers gesenkt wird, ggf. in Kombination mit einem verbalisierten zweiten Paarteil (Stimmausgabescharnier).³¹⁰ Allerdings – und hier lässt sich wiederfinden, was Krummheuer (2010: 210) als „schiefe Anschlüsse“ bezeichnet (siehe auch Pitsch 2020: 135) – zeigt die Stimmausgabe des VUI, dass die konditionell relevant gesetzte Antwort von Robin („DREI“, 116) nun nicht mehr vor dem Hintergrund des bisherigen Verlaufs verarbeitet wird. Die Äußerung wird unter dem Aspekt der Speech Recognition zwar korrekt erkannt, wie die Wiederholung des Lexems „DREI“ durch das VUI (120) klar zeigt, wird aber als ‚neue‘ Stimmeingabe und als Aufforderung zur Wiedergabe von Musik verstanden (die Gründe für den Wechsel im Operationsbereich bzw. dem *Intent*³¹¹ bleiben hier opak), sodass der Titel „3“ („Three“) von Britney Spears wiedergegeben wird. Auf den Sprachwechsel zwischen Deutsch und Englisch – der Titel von Britney Spears hieße, auch wenn bei popkulturellen schriftlichen Referenzen überwiegend die numerische Angabe „3“ erfolgt, ausgesprochen „Three“ und nicht „Drei“ – reagieren die Sprecher*innen anschließend mit Lachen und nehmen die Fehlfunktion insofern humoristisch auf. Das Beispiel zeigt somit auch, dass die Aufrechterhaltung längerer sequenzieller Vollzüge zwar seitens der VUIs prinzipiell angelegt ist – die Nachfrage (114) mit der vorherigen Erläuterung (113) weist darauf hin. Sequenzielle Verknüpfungen, die über die Dyade hinausgehen, bleiben aber gleichwohl eine Herausforderung für die Geräte (vgl. Habscheid 2022), was wiederum dazu beiträgt, dass ihre Gesprächsteilnahme zeitlich sehr begrenzt ist.

Die Äußerung von Robin „du bist so DUMM,“ (124) zeigt erneut die gleiche sprachliche Praktik des ‚Splits‘ zwischen formaler und funktionaler Adressiertheit: Einerseits wird damit Robins Bewertung der Geräteperformanz ausgedrückt und in einen humorvollen Kontext gestellt, zugleich ist die Äußerung hochgradig invektiv und erfährt gleichwohl keinerlei Rahmung einer dispräferierten Handlung. Sprachlich invektives Verhalten gegenüber einem VUI, gegenüber dem auch keine Beziehungsarbeit i. e. S. geleistet werden muss, ist insofern in den untersuchten Haushalten sozial akzeptiert und ggf. als Humorressource sogar erwünscht.

6.2.3.2 Rezeptionssignale

Der formal-funktionale ‚Split‘ zeigt sich auch bei Betrachtung der Rezeptionssignale in folgendem Beispiel, das aus der Ersteinrichtung des Google Assistant in Haushalt 8 stammt:

³¹⁰ Siehe Kap. 6.1.2.1.1.

³¹¹ Siehe Kap. 3.2.2.

Beispiel (63): Was hast du noch für Stimmen? (08-EE-21)³¹²

634 ((RL hat die Arme hinter dem Kopf
verschränkt, sitzt zurückgelehnt, bis Z. 642))³¹³

635 SM: oKE google;=
636 =was hast du noch für STIMmen,
637 (1.37)

638 SM: (hab ich) zu SCHNELL gesagt;
639 (0.67)

640 LS: ΔHMhm;
641 GA: du kannst meine stImme in der google HOME
app ändern.

642 RL: \$ΔHM; \$ Δ
rl: \$lehnt sich vor\$
Δgreift mit RH nach HandyΔ

643 GA: tippe dazu rechts oben auf dein profilbild
oder deine initiale und dann auf asSISant
einstellungen;
644 (0.45)

645 RL: Δhm_[HM,]----->
>-----Δ
Δtippt auf dem Handy (fortwährend bis
Z. 650)Δ

646 GA: [tI]ppe dann unter (.) alle
einstellungen (.) auf (.)
asSISant stimme;
647 (0.92)

648 RL: hm_HM,
649 (4.37)

650 SM: \$ist die einrichtung gleich damit
ABgeschlossen,\$
sm: \$runzelt die Stirn----->
>------\$

651 RL: °hh ja;

312 Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c560&wID=&textSize=400&contextSize=12.

313 Siehe Abb. 26 und 27.

Robin produziert drei Mal die Partikel „hm“ bzw. „hm_hm“, die bei Selting (2016: 45) als „einfache[s] Rezeptionssignal“ beschrieben wird. In der Forschung wird „hm“ und „hm_hm“ sowohl als Partikel (Willkop 1988) als auch als Interjektion (Zifonun/Hoffmann/Strecker 1997: 362) beschrieben, wobei in unterschiedlichen Arbeiten sowohl eine gliedernde als auch eine rückmeldende Funktion dieser Äußerungen postuliert wird (vgl. Kucharczik 1989; Schwitalla 2002). Im Zentrum stehen dabei in bisherigen Analysen Rückmeldungen der Hörer*innen an die Sprecher*innen, die durch die diskursive Positionierung oder auch durch die prosodischen Eigenschaften des Signals entstehen (vgl. Meer/Pick 2019: 34), wobei eine dichotome Verteilung dieser Rollen kaum hinterfragt wird.

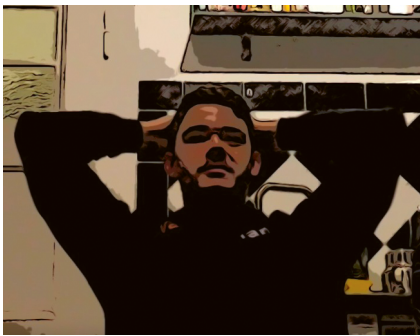


Abb. 26: Sitzposition von Robin bis Z. 642.



Abb. 27: Sitzposition von Robin ab Z. 642.

Die Rezeptionssignale sind auf der formalen Oberfläche an der segmentalen Struktur der VUI-Äußerungen orientiert: Sie folgen unmittelbar auf dessen Turns und sind zeitlich genau so platziert, dass sie zwischen zwei syntaktisch und prosodisch geschlossene Einheiten fallen – mit Ausnahme einer Überlappung mit dem Turn-Beginn (644), die typisch für Rezeptionssignale ist. Zudem bestätigen sie durch ihre Platzierung die Segmentierung der Äußerungen des VUI. Allerdings sind die Turns des VUI ja präskribiert und nicht – wie etwa in dyadischen Mensch-Mensch-Interaktionen – erst durch die Gliederungssignale des Gegenübers mit segmentiert und so eine gemeinsam konstruierte Hervorbringung (vgl. Schwitalla 2002: 271).

Geht man auch in diesem Auszug von einem formal-funktionalen ‚Split‘ aus, werden die Äußerungen von Robin als mehrfachadressiert verständlich: Sie schließen zwar einerseits in Form eines Rückmeldesignals an die Äußerungen des VUI an und bestätigen die Äußerungen dessen. Zugleich aber machen sie die Handlungen von Robin für seine ko-präsenten Mitbewohner*innen *accountable*: Die erste Äußerung von „hm“ (642) tritt in unmittelbarer zeitlicher Nähe zu einem Wechsel der Körperhaltung auf: Während er bis zur Äußerung des VUI (641), das auf die Home-App

verweist, mit hinter dem Kopf verschränkten Armen sitzt und den Rücken an die Stuhllehne stützt, lehnt er sich – begleitet von „hm“ (642) – vor zum Esstisch und greift nach seinem Handy (siehe Abb. 26 und 27). Er zeigt insofern an, dass er den Hinweis des VUIs auf die Möglichkeit zur Anpassung der Stimme aufnimmt und auf dem Smartphone entsprechende Einstellungen sucht. Dabei passt er sich an die Segmentierung der Äußerungen des VUI an und es entsteht somit zugleich eine formale, temporäre Ratifizierung dessen als ‚Beteiligte*r‘ am Gespräch. Auch die zweite Rückmeldung – nun zweisilbig und in (für Rezeptionssignale typischer) Überlappung mit dem Beginn des nächsten Turns des VUI produziert (645) – scheint nach diesem Muster zu funktionieren: In Verbindung mit der leichten Kopfbewegung bestätigt Robin, den entsprechenden Menüpunkt in den Einstellungen gefunden zu haben (ebenso 648).

Im Zusammenspiel mit der multimodalen Analyse – insbesondere Körperhaltung, Blick und Steuerung des Smartphones – zeigt sich, dass die Rezeptionssignale aber auch für die ko-präsenten Sprecher*innen das Verstehen ausdrücken und Folgehandlungen projizieren sowie laufende Handlungen *accountable* machen. Konzeptionell sind sie dabei zwischen Rezeptionssignal, „Erkenntnisprozessmarker“ (Imo 2009; vgl. auch Heritage 1984a) und „Response Cries“ (Goffman 1978) zu verorten: Eine Konzeptualisierung als Rezeptionssignal scheint nur auf der funktionalen Oberfläche zuzutreffen, nicht aber mit einer interaktiven oder praxeologischen Betrachtungsweise. Erkenntnisprozessmarker sind allen bisherigen Untersuchungen zufolge von höherer semantischer Spezifität gekennzeichnet, während *hm* zwar als Rezeptionssignal konventionalisiert, sonst aber unspezifisch ist und z. B. durch prosodische Variation auch semantisch variiert werden kann (vgl. Meer/Pick 2019: 34). Gleichwohl scheint sich eine Änderung im Wissensstand bei Robin einzustellen, denn er wird instruiert, wo er die Stimmeinstellung vornehmen kann und wie er die Smartphone-App zu bedienen hat (das macht die Äußerung freilich noch nicht zum Erkenntnisprozessmarker). Eine Klassifikation als „Response Cries“ nach Goffman (1978) scheint ebenso ambivalent: Auf der einen Seite sind „hm“ und „hm_hm“ explizite Beispiele, etwa für gefüllte Pausen oder Selbstreparaturen (vgl. Goffman 1978: 811). Sie sind – wie auch im gegebenen Beispiel – keine vollwertigen Diskurseinheiten und keine ‚Aussagen‘ (vgl. Goffman 1978: 805–806), sondern vielmehr selbstadressierte Zwischenrufe und können als „standardized vocal comments on circumstances that are not, or are no longer, beyond our emotional and physical control“ (Goffman 1978: 805) konzeptualisiert werden. Letztlich sind alle drei Konzepte für die Herausbildung dieser Praktik relevant – so wird das VUI an der Oberfläche mit einem fragilen, temporären Status zu einem temporären ‚Beteiligten‘ an der sozialen Praxis; zudem sind die Rezeptionssignale stimmlicher Ausdruck des Aneignungsprozesses.

6.2.3.3 Kohärente Anschluss-Turns

Dieser fragile und temporäre Status der ‚Beteiligung‘ zeigt sich auch im folgenden Beispiel.³¹⁴

Beispiel (64): Dein Gerät kann jetzt eingerichtet werden (01-EE-08)³¹⁵

241 LF: Δso::; Δ
lf: Δtippt auf HandyΔ

242 LF: ähm ähm ä:HM,
243 (16.0)

244 k : \$((Melodie spielt)) \$ [((Melodie spielt))
]
245 LF: [Δder SOLL noch gar
nich-Δ]
lf: \$wendet sich Alexa zu\$ Δdrückt Knöpfe auf
AlexaΔ

246 LF: *(ach eGAL;) *
lf: *Abwinken mit rechter Hand*

247 AL: ΔheLL0; Δ
lf: Δtippt weiter auf HandyΔ

248 AL: your device is ready for SETup;
249 AL: just download the alexa APP and [follow
the instrUctions;]

250 LF: [+MACH ich
gerade; +]
lf: +kurzer
Blick auf Alexa+

251 AL: ΔHALlo; Δ
lf: Δtippt weiter auf HandyΔ

252 AL: dein gerät kann jetzt EINgerichtet werden;
253 lade dazu die alexa app RUNter und folge
den anweisungen;
254 LF: SO;

³¹⁴ Für eine Analyse dessen unter dem Aspekt der Agency siehe Waldecker/Hector/Hoffmann (2024) sowie Hector (2022) für eine Analyse unter dem Aspekt der akustischen Beteiligung an der sozialen Praxis.

³¹⁵ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00465_SE_01_T_01_DF_01&cID=c273&wID=&textSize=500&contextSize=16.

- 255 AL: bonJOUR;
 256 votre Δappareil [est PRÊT pour la]
 configuration;Δ
 lf: Δwendet sich Alexa zu und zieht das
 StromkabelΔ
 257 LF: [!ARGH!?]
 258 ΔSO; Δ
 lf: Δtippt weiter auf HandyΔ
 259 (1.5)
 260 LF: ä:h aLEXa;

Lukas hat kurz zuvor den Smart Speaker an die Stromversorgung angeschlossen. In der Folge davon startet bereits das Skript zur Ersteinrichtung. Dies ist verbunden mit der Wiedergabe eines entsprechenden Sounds. Durch Lukas' Reaktion auf diese Melodie („der SOLL noch gar nich-“, 245) wird deutlich, dass der Systemstart aus Lukas' Sicht verfrüht ist – er ist noch mit dem Herunterladen der App beschäftigt. Bemerkenswert ist, dass Lukas auf den Smart Speaker hier mit dem Demonstrativpronomen „der“ referiert und diesen damit nicht direkt anspricht. Der kurzzeitige Beteiligungsstatus kommt erst zustande, als das VUI beginnt zu sprechen (247–249). Lukas knüpft mit seinem Turn an die Äußerung des VUI an – den Hinweis, die Alexa-App herunterzuladen (249). Lukas' Äußerung „mach ich gerade“ (250) kann als Analepse beschrieben werden. Ein zugehöriges Objekt wird nicht produziert, sodass die zuvor durch das VUI produzierte Aufforderung, die App herunterzuladen (249), als Antezedens verstanden werden kann (vgl. Helmer 2016: 116–131). Dadurch entsteht eine formale wie topikale Kohärenz zwischen den beiden Turns und die Äußerung des VUI wird so kurzzeitig als ‚Gesprächsbeitrag‘ verstanden. Schnell jedoch wird das VUI nicht mehr in dieser Weise angesprochen: Durch das Skript werden die Begrüßungsäußerungen neben Englisch und Deutsch nun auch in weiteren Sprachen vorgetragen. Dies jedoch stört durch seine Akustizität den weiteren Prozess der Ersteinrichtung, wie auch durch die Produktion der Partikel „ARGH“ (257) deutlich wird (vgl. Hector 2022: 222). Lukas zieht das Stromkabel wieder aus der Buchse am Echo heraus (siehe Abb. 28), sodass die Wiedergabe der Begrüßung in verschiedenen Sprachen unterbrochen wird. Die Stelle zeigt außerdem, dass dieser Beteiligungsstatus nicht nur fragil ist, sondern auch davon mit abhängt, wie die anderen Gesprächsbeteiligten in den Verlauf der Praxis involviert sind. Wie aus den Aufzeichnungen ersichtlich wird, ist Alex – der Mitbewohner von Lukas, der ebenfalls an der Ersteinrichtung mitwirkt, – zu diesem Zeitpunkt kaum in den Prozess der Ersteinrichtung eingebunden. Er ist auf seinen Laptop konzentriert, auf dem er tippt, und auch körperlich nicht auf das Geschehen hin orientiert (vgl. Abb. 29). Es ist nicht ersichtlich, was Alex auf dem Laptop ‚macht‘, aber da dies in der Interak-

tion zwischen Lukas und ihm nicht relevant wird, kann angenommen werden, dass es nicht mit der Ersteinrichtung in Verbindung steht.³¹⁶ Diese geringe Beteiligung von Alex (siehe auch Hector 2022: 221–222) geht einher mit der Beobachtung, dass Lukas ihn nicht direkt adressiert, sondern stattdessen Äußerungen produziert, die nicht an ihn gerichtet sind und kohärente Turns mit dem VUI produziert.



Abb. 28: Lukas zieht Stromkabel aus dem Smart Speaker (Z. 257).



Abb. 29: Position der Sprecher Lukas (vorne) und Alex (hinten).

Ein weiteres Beispiel, hier aus Haushalt 8, soll illustrieren, dass gleichwohl der kohärente Anschluss an Äußerungen des VUIs auch in dynamischeren Gesprächssituationen auftreten kann:

Beispiel (65): Ganz ruhig (08-EE-22)³¹⁷

- 687 RL: oKAY google;
 688 (0.69)
 689 RL: bist du jetzt ein MANN?
 690 (1.93)
 691 GA: ich bin \$vIrtuell und habe\$ deshalb kein
 geSCHLECHT;
 rl: \$lächelt nickend--\$

³¹⁶ Später zeigt Alex Lukas das W-LAN-Passwort auf dem Laptopbildschirm, doch dies ist ein separater Prozess, der hier noch nicht begonnen hat.

³¹⁷ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c608&wID=&textSize=300&contextSize=8.

692 (0.21)
 693 GA: aber ich liebe es (.) wenn \$menschen im
 einklang mit ihrem
 wah[ren selbst und ihrer identi]Tät
 sind;\$
 ls: \$lächelt und
 zieht anschließend
 Mundwinkel nach unten----->
 >----\$
 694 SM: § [((unverständlich, 1.44 Sek.))\$
 sm: \$wendet sich kurz RL zu-----\$
 695 (0.97)
 696 LS: *okay (.) ganz RUhig; *
 ls: *streckt GA Handfläche der LH entgegen*
 697 SM: [((lacht))]

Die Mitglieder der Wohngemeinschaft haben kurz vor dem hier präsentierten Ausschnitt verschiedene Stimmoptionen ausprobiert.³¹⁸ Ausgelöst davon fragt Robin danach, ob der virtuelle Assistent jetzt ein Mann sei.³¹⁹ Die Spracheingabe impliziert bereits, dass diese nur im Zusammenhang mit dem fortgesetzten Testen verständlich wird, eine über den humoristischen Effekt hinaus zweckdienliche Beantwortung dieser Frage ist nicht vorstellbar – mit Bezug auf fernsehbegleitendes Sprechen formuliert Klemm (2001: 102–103): „Die Zuschauer nutzen den Fernseh-text lediglich als Material zur Etablierung und Stabilisierung einer positiven Grundstimmung, die sich bis zum ungehemmten Blödeln hochschaukeln kann“. Eine ähnliche Dynamik scheint sich auch hier zu ereignen: Das Testen des VUIs wird insofern in die Praktiken der ‚Unterhaltung‘ und des ‚Spaßmachens und -habens‘ eingegliedert. Wie schon eingangs dargelegt, ist genau in diesen Verwebungen von VUI-Dialogen mit anderen, konversationell bzw. als Gespräch vollzogenen Praktiken eine erhöhte Sensitivität für (vorübergehende) ‚Gesprächsbeteiligungen‘ des VUIs auszumachen. Wie das Beispiel hier zeigt, ist dies auch seitens der Hersteller angelegt: Auch wenn die Abfrage nicht an eine spezifische Funktion (Timer, Smart Home-Anwendungen, Musikauswahl, Informationsabfragen o. Ä.) gebunden ist, sondern die Frage nach dem Geschlecht allein der Unterhaltung dient, hat das VUI eine ‚Antwort‘; es präsentiert sich entsprechend nicht nur „task oriented“, sondern

³¹⁸ Siehe Kap. 6.2.3.7.

³¹⁹ Bemerkenswert ist, dass Google in seinen Einstellungen auf eine Geschlechterzuordnung der Stimmen verzichtet und diese lediglich „rot“ und „orange“ heißen, siehe dazu Kap. 6.2.3.7.

auch „social“ (Natale/Cooke 2021: 1009); es sind Antworten dieser Art, die zum Aufbau von digitalen „Personae“ beitragen (vgl. Natale 2021: 124).

Im hier gegebenen Beispiel schließt Lara an diese Äußerung an und produziert in der Äußerung zunächst „okay“ (696), das erstens in seiner retraktiv-projizierenden Doppelfunktion (vgl. Beach 1993: 336) gelesen werden kann und zweitens in der prosodischen Klassifikation von *okay* nach Kessler (2010: 160–161) in Kombination mit der Folgeäußerung „ganz RUhig“ am ehesten „relativierende Akzeptanz“ ausdrücken könnte, wobei sich „okay“ auch hier durch große semantische Vagheit auszeichnet. Als sicher kann allerdings gelten, dass eine Turn-Übernahme ratifiziert wird. In Kombination mit „ganz RUhig“ (696), einer elliptischen und eher indirekten Beruhigungsformel, die wohl als verfestigt gelten kann, wird Laras Kommentar nicht nur als Anschluss an die Äußerung des VUI verständlich: Sie dient auch der Evaluation der Äußerung und bewertet diesen bzw. dessen Ausführungen zur Geschlechtlichkeit als ‚aufgeregt‘ oder ‚unruhig‘. Es entsteht damit auch ein Fortsetzen der Anwendung und Bewertung des VUI zu humoristischen Zwecken.

Kohärenz wird also sowohl von den VUIs hergestellt als auch von den Nutzer*innen, die kohärente Anschluss-Turns produzieren. Die Analysen zeigen, dass mit der formal erzeugten Kohärenz ein vorübergehender, fragiler Status als Beteiligte*r entsteht, der jedoch, was die Teilnahme am Gespräch angeht, nur auf der sprachlichen Oberfläche auftritt und nicht aufrechterhalten wird. Auch in den in diesem Unterkapitel vorgeführten Fällen zeigen sich die dazu genutzten sprachlichen Formen als doppelt adressiert und in den lokalen Kontext in einer Weise eingebunden, die den Schluss einer konsistenten Beteiligung nicht zulässt.

6.2.3.4 Wiederholungen

Wiederholungen der Äußerungen eines anderen Sprechers bzw. einer anderen Sprecherin sind in der konversationsanalytischen Literatur bereits früh diskutiert worden. Jefferson (1972: 303) formuliert eine im Goodwin’schen Sinne ko-operativ gedachte Bestimmung von Wiederholungen: „an object that has as its product-item a prior occurrence of the same thing, which performs some operation upon that product-item“. Während Wiederholungen häufig unter dem Blickpunkt der Initiierung von Reparaturen betrachtet worden sind (vgl. etwa Schegloff/Jefferson/Sacks 1977: 368; Sorjonen 1996; Schegloff 1997a; im Überblick Stivers 2005: 131), wurde in der Literatur nie bezweifelt, dass diese auch andere Funktionen erfüllen können. Bereits früh wurde das kohäsive Potenzial von Wiederholungen erkannt (vgl. Halliday/Hasan 1976: 288). Tannen (2007: 61) postuliert, dass Wiederholungen dem „over-arching purpose of creating interpersonal involvement“ die-

nen: Sie bewältigen die Durchführung einer Konversation, zeigen Reaktionen auf das Gegenüber an und wirken partizipativ. Die Wiederholung trägt dazu bei, ein interpersonales Involviert-Sein sprachlich anzuzeigen und interaktional zu vollziehen (vgl. Tannen 2007: 61–62).

Im Hinblick auf VUI-Dialoge lässt sich also nach der Funktion von konversationellen Wiederholungen fragen. Dabei konzentriere ich mich in den folgenden drei Beispielen auf Wiederholungstypen, in denen Nutzer*innen Teile einer VUI-Äußerung wiederholen.³²⁰ Dabei scheinen in den vorliegenden Daten verschiedene Verwendungsweisen möglich zu sein, es könnten aber weitere auftreten. Im folgenden Beispiel aus der Ersteinrichtungssituation in Haushalt 8 testet der Anwender Robin die Wetter-Funktion. Die Ersteinrichtung im engeren Sinne ist schon seit einigen Minuten abgeschlossen, die Gruppe befindet sich in einem fortgesetzten Modus des Testens.

Beispiel (66): Minus drei Grad (08-EE-36)³²¹

```

1218  RL:  oKAY google-
1219        (0.57)
1220  RL:  wie wird das WETter morgen;
1221        (1.98)
1222  GA:  morgen wird es in dortmund NEblig bei
          temperaturen zwischen
          minus $drEI und z[wEI grad;]
1223  RL:  Δ [      mInus] !DREI!,$--->
          >-----Δ
          rl:      $zieht die Augenbrauen hoch  $
                  Δnimmt Handy in LH und
                  tippt fortlaufendΔ
1224        (0.26)
1225  SM:  *(huä);                                *
          sm:  *streckt Kinn nach vorne*
1226  RL:  fIck mein LEben;

```

³²⁰ Es treten auch Wiederholungen auf, in denen VUIs (Teile der) Äußerungen der Nutzer*innen wiedergeben; dies dient der Anzeige des Verarbeiteten bzw. ‚Verstandenen‘ und insofern der Verbesserung der Nutzer*innenerfahrung auch bei Missverständnissen (siehe auch Kap. 6.1.2.1.1 zur Basis-Sequenzstruktur des VUI-Dialogs).

³²¹ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c1098&wID=&textSize=300&contextSize=8.

- 1227 §(0.87)----->
 >--§
 sm: §guckt zwischen Fenster und RL hin und
 her§
 1228 SM: (oh ne: dann rutsch ich ja AUS mit_m
 fahrrad;)

Robin produziert zunächst Invokation und Stimmeingabe (1218–1220), das VUI liefert eine kurzgefasste Wettervorhersage (1222). Noch während das VUI die Produktion dieser abschließt, greift Robin die zur Temperatur gehörige Numerales³²² auf und wiederholt diese. Dabei operiert er auf dem Material, das das VUI bereitstellt, verändert aber deutlich die prosodische Gestaltung: Er akzentuiert die Kardinalzahl „drei“. Über die Folgeäußerungen sowohl von Samuel (1225, 1228) als auch von Robin (1226) lässt sich erschließen, dass primär Überraschung und Missfallen über die Temperaturprognose zum Ausdruck kommt. Zwar entsteht auf der formalen Oberfläche Kohärenz zwischen der Äußerung des VUI und den Folgeäußerungen, die diese initiiert haben, doch das Folgegespräch entfaltet sich zwischen Robin und Samuel – das VUI wird daran nicht beteiligt. Die ‚Beteiligung‘ ist also lokal begrenzt: Sie erstreckt sich genau auf die VUI-Dialog-Sequenz (1218–1222) sowie auf den Folgeturn von Robin in 1223, in dem dieser an die Äußerung anschließt. Damit involviert sich Robin durch die Wiederholung interpersonal in die Interaktion, diese wirkt jedoch eher wie ein Scharnier: Die Äußerungen des VUI werden erneut ausgesprochen und damit in den interaktiven „Floor“ zwischen den menschlichen Beteiligten übertragen. Durch die Wiederholung durch den Anwender erhalten diese eine neue Relevanz und evozieren Folgeäußerungen sowie eine topikale Progression. Repetitionen können auch – wie im gegebenen Beispiel – evaluative Funktion haben und mit zahlreichen Nuancen Funktionen in der Anschlusskommunikation übernehmen, wie auch im Zusammenhang mit anderer medialer Anschlusskommunikation beobachtet werden konnte (vgl. Ayaß 2001: 212; Baldauf 2001: 74).

Das folgende Beispiel ist in dieser Hinsicht ähnlich gelagert. Der Auszug entstammt ebenfalls Haushalt 8 und stammt vom Abend der Ersteinrichtung, ist aber deutlich nach der Installation durch den CVR aufgezeichnet worden. Die Gruppe befindet sich weiterhin im Modus des Testens:

322 Zu Kardinalia, Ordinalia und Nummer-Konstruktionen siehe Wiese (1996).

Beispiel (67): Acht Uhr (08-CVR01-05)

113 RL: oKAY google,
 114 [(1.0)]
 115 k: [((Musik pausiert))]
 116 RL: kannst du einen termin für RObin
 erstellen?
 117 (0.5)
 118 SM: OH ja das ist gut.
 119 LS: soll ich dir das einmal rüberschick [(.)
 ä:hm AIRdropen?]
 120 GA: [WOran
 möchtest du erinnert werden?]
 121 (1.0)
 122 RL: MO:Rgen: (.) elf uhr (.) uni.
 123 (2.7)
 124 SM: (du SCHACHtel du.) du gehst doch eh nicht
 in die uni.
 125 GA: verSTANDen.
 126 **[morgen um acht UHR,]**
 127 RL: [((lacht))]
 128 k: ((lautes Gelächter))
 129 AW: [((lacht))] [((lacht))
]
 130 GA: [BRAUchst du (dafür
 allerdings die) GOOgle
 assistant app-]
 131 SM: aber GUCK ma du hast doch nicht reagiert
 jetzt drauf ne,
 132 (0.3)
 133 LS: Näh.
 134 [(2.3)]
 135 k: [((Musik setzt wieder ein))]
 136 RL: ((prustet))
 137 (0.3)
 138 k: ((RL und LS lachen))
 139 **RL: ALLE[s klar,]**
 140 LS: [Oke-]
 141 **RL: mo[rgen um ACHT.]**
 142 **LS: [morgen um ACHT] uhr;**

- 143 RL: ((lacht))
 144 SM: <<f>M0:Rge:n,>
 145 LS: mh HM,
 146 (1.3)

Robin unternimmt den Versuch, einen Termin erstellen zu lassen. Dass es sich dabei erneut um einen Versuch handelt, wird aus der Bewertung von Samuel ersichtlich (118), die sich auf die Stimmeingabe von Robin bezieht. Nach einer ‚Rückfrage‘ durch das VUI (120) produziert Robin die entscheidenden Spezifikationen für den Termin (122). Scheinbar jedoch sind Schwierigkeiten in der Speech Recognition aufgetreten (möglicherweise aufgrund des sequenziell eingeschobenen Kommentars von Samuel, vgl. 124). Statt „elf Uhr“ wird der Termin für „acht Uhr“ festgelegt (vgl. 126). Die Gruppe reagiert darauf mit lautem Gelächter, der Lachgegenstand ergibt sich scheinbar aus der zu frühen Uhrzeit in Verbindung mit der autonomen ‚Korrektur‘ durch das VUI sowie aus der kontrafaktischen Bestätigungsformel („verSTANDen“, 125).

Robin greift die Sequenz anschließend – nach einem als Lachpartikeln zu deutenden prustenden Geräusch (138) – noch einmal auf. Die folgenden Äußerungen scheinen auch hier den Charakter einer Re-Inszenierung fremder Rede aufzuweisen (vgl. Günthner 1997; 2000), wobei die gebrauchten sprachlichen Möglichkeiten – prosodische Desintegration, keine Redeeinleitung, stark exklamatorischer Charakter der Äußerungen – auf eine Wiedergabe im Modus „maximaler Direktheit“ (Günthner 2000: 20) verweisen. Anders als in Beispiel (66) („Minus drei Grad“) werden hier prosodische Eigenschaften, insbesondere Sprechrhythmus und Tempo, beibehalten bzw. wiederaufgeführt: Die intonatorische Segmentierung der Äußerung (Bestätigung – Wiederholung des ‚Verstandenen‘) erfolgt analog zur Äußerung des VUI. Dabei vollzieht sich die Wiederholung anfänglich sogar *primär* auf der prosodischen Ebene, denn lexikalisch produziert Robin eine Abweichung. Lara steigt nahezu zeitgleich in die Reinszenierung der Äußerung mit ein (141–142). Als Verstehensbestätigung verwendet sie „Oke-“ (140), und setzt – wie Robin – unmittelbar daran anschließend mit der Redewiedergabe der Äußerung des VUI fort. Die VUI-Sequenz wird so noch einmal in der Interaktion relevant und – bereits Sekunden nach dem Ereignis – zur scherzhaft aufgeführten Narration (vgl. Kotthoff 2006b).

An dieser Humornarration der Gruppe ist das VUI bzw. dessen Äußerung als Lachgegenstand ‚beteiligt‘. In die interaktiv-humoristischen ‚Verwertungen‘ der Äußerung ist das VUI nur mittelbar involviert – die Äußerung ist ein „Substrat“ (Goodwin 2018: 3), auf dem nun ko-operativ neue Handlungssequenzen zur weiteren Unterhaltung aufgebaut werden. Samuel nimmt aus diesem Substrat schließlich nur noch das Lexem „morgen“, das er als Ausruf inotiert (vgl. 144) und das somit vielmehr als Gruß verständlich wird. Dabei scheint der Sprecher einen

Übergang zu vollziehen. Für solche Übergänge leitet Ehmer (2011: 60–77) konzeptionell ein Muster her: Wurden in den bisherigen Darbietungen primär polyphone Redewiedergaben vollzogen (siehe etwa Günthner 2002), ist dieser Gruß der Übergang zur „animierten Rede“ – zu Äußerungen, „in denen zwar die formalen Eigenschaften direkter Rede vorliegen, jedoch keine konkreten Originaläußerungen zugrunde liegen, die reproduziert werden könnten“ (Ehmer 2011: 62). Der hier dokumentierte Fall scheint im Grenzbereich dessen zu liegen: Zwar wird „morgen“ auch in der Originaläußerung gesprochen und tritt auch in den polyphonen Redewiedergaben auf, ist allerdings dabei nicht als Gruß, sondern als Zeitangabe zu verstehen. Samuel arbeitet mit dem Ausruf „MO:Rgen,“ an der Herstellung einer konstruierten Dialogsituation, in der das VUI den Anwender Robin weckt. Indem er die Äußerung des VUIs demonstriert, übernimmt er kurzfristig ‚dessen Perspektive‘. Im Zuge dieser Imagination – verstanden „als der *Prozess* [Herv. i. O.] [...], im Gespräch eine szenische Vorstellung aufzubauen“ (Ehmer 2011: 13) – kann das VUI also als imaginierte*r Sprecher*in auftreten. Durch die Wiederholung wird diese Partizipation des VUIs an der Praxis sprachlich realisiert. Eine Änderung des Status des VUIs an der ‚Beteiligung‘ am Gespräch ergibt sich daraus allerdings nicht.

Auch das folgende Beispiel aus Haushalt 8 illustriert, dass zwar Bestandteile der Äußerungen des VUIs wiederholt werden, dies aber nicht zu einer ‚Beteiligung‘ des VUI am Gespräch führt – im Gegenteil markiert die Sprecherin hier vielmehr, dass das VUI eben *nicht* ratifizierter Gesprächsteilnehmer ist:

Beispiel (68): Passiv-aggressives „bitte sehr“ (08-CVR01-05)

003 RL: oKAY google;
 004 (1.3)
 005 SM: [((unverständlich, ca. 2 Sek))]
 006 RL: [spiel musik von SPotify;]
 007 (1.7)
 008 GA: verSTANDen.
 009 musik auf SPotify.
 010 **BITte sehr.**
 011 [(1.4)]
 012 k: [((Musik setzt ein))]
 013 SM: (oh WEH oh weh;) ((unverständlich, ca. 2 Sek.))
 014 **AW: sie und ihr passiv aggressives (-) BITte sehr-**
 015 SM: [((lacht))]

Die Gruppe aus Robin, Samuel, Lara und Alexandra sitzt in Haushalt 8 zusammen und führt weitere Tests in gelöster Atmosphäre durch. Nachdem Robin per Stimmeingabe die Musikwiedergabe über Spotify gestartet hat (003–010), bezieht sich Alexandra mit ihrer Äußerung („ihr passiv aggressives (-) BITte sehr“, 014) auf das Stimmausgabescharnier (010). Generell stellt dieses eine Verbindung zwischen der Stimmeingabe der Nutzer*innen und der praktischen Umsetzung her.³²³ Hier enthält dieser Bestandteil des Sequenzmusters mit „BITte sehr“ aber noch eine etwas anders nuancierte Funktion: Die zusätzliche Routineformel *bitte sehr!* (vgl. Farø 2015: 235),³²⁴ auf die sich auch Alexandra später bezieht, dient nicht nur der Verstehenssicherung, sondern ist auch Ausdruck von Höflichkeit. *Bitte sehr* bzw. *bitte schön* kann in Verkaufsgesprächen initial verwendet werden, um (ggf. nach Gruß und Gegengruß) die Bereitschaft zur Bearbeitung eines Anliegens zu signalisieren (siehe etwa Hausendorf 2022: 207). Die Formel kann aber auch als Bestandteil einer „Danksequenz“ (Brinker/Sager 2010: 101) begleitend zu oder nach einem Dank stehen (vgl. Zinken/Rossi/Reddy 2020: 261). Die Äußerung von Lara beinhaltet eine Evaluation der Intonation des VUI: Sie attribuiert die Konstituente als „passiv aggressiv“ (014) und re-inszeniert die Äußerung auch prosodisch. Dies kann damit zusammenhängen, dass „bitte sehr“ bzw. ähnliche Formeln auch verwendet werden, um anzuzeigen, dass ein sequenziell erwartbarer Dank nicht produziert wurde (vgl. Zinken/Rossi/Reddy 2020: 261). Stellen, an denen Dank im Zusammenhang mit einer konversationellen Aufforderung produziert werden können, sind von DeSouza et al. (2021) als „gratitude opportunity spaces“ beschrieben worden. Die Autor*innen stellen zwar einerseits fest, dass diese Gegenstand lokaler und präzise aufeinander abgestimmter konversationeller Aushandlungen der Gesprächsteilnehmer*innen sind. Gleichzeitig können Normerwartungen hinsichtlich des zeitlichen Verlaufs der Koordination von Aktion und Dank durchaus eine Rolle spielen: „gratitude expressions outside the gratitude opportunity space may be treated as accountably ‚late‘ or missing“ (DeSouza et al. 2021). Die Äußerung von Alexandra scheint ein (möglicherweise ironisch gebrochener) Hinweis auf eine entsprechende Normverletzung dieser Koordinierung zu sein – möglicherweise, weil die seitens des VUI produzierte Äußerung keine Gelegenheit für einen Dank lässt und die Formel in der Dankessequenz vorwegnimmt, sodass sie als ‚aggressiv‘ wahrgenommen wird.

An der Äußerung von Alexandra fällt zudem auf, dass sie sich eines verfestigten, routinisierten Musters (vgl. Günthner 2018b) bedient, das ohne Verbform auskommt und mit ‚X und sein* ihr Y‘ gebildet wird, wobei zwischen dem Menschen

323 Siehe Kap. 6.1.2.1.1.

324 Zur Klassifizierung von Routineformeln siehe Coulmas (1979: 117–120).

X und einer beliebigen Konstituente Y eine enge Verbindung unterstellt wird, deren genauer Charakter vage bleibt; es ist den Adressierten überlassen, diese evaluative Leerstelle (mental) zu füllen.³²⁵ Die Besonderheit und die Existenz der Verbindung zwischen den beiden Elementen überhaupt herauszustellen scheint die primäre kommunikative Aufgabe zu sein. Die Konstruktion hat exklamativen Charakter und steht in Verbindung zu vorangegangenen Äußerungen oder Handlungen, hier mit dem vorangegangenen Turn des VUI. Bewertungen über körperliche Merkmale wie die Stimme sind ebenso wie das Nachahmen potenziell gesichtsbedrohend. Es ist auffällig, dass Alexandra keine Versuche von „Face-Work“ (Goffman 1955) zur Gesichtswahrung unternimmt (siehe auch Brown/Levinson 1987) und außerdem auf das VUI mit der Objektdeixis referiert. Obschon also die Stimme evaluiert und re-inszeniert wird, entsteht durch diese Wiederholung keine ‚Beteiligung‘ am Gespräch. Gerade durch die Äußerung von Alexandra wird deutlich, dass das VUI kein ‚Gesprächsteilnehmer‘ ist, sondern dass dessen Äußerungen auch in ihrer Form vielmehr das Potenzial haben, zum Gegenstand einer gemeinsamen (humorvoll gestalteten) Evaluation unter den menschlichen Beteiligten zu werden, ohne dass dies als invektiv verstanden würde. Damit entsteht eine Abgrenzung des VUIs von den menschlichen Sprecher*innen.

6.2.3.5 Ko-konstruierte VUI-Sequenzstrukturen

In den bisherigen Analysen zeichnet sich ab, dass VUIs nur mit einem formalen, temporären und fragilen Beteiligtenstatus als ‚Beteiligte‘ an Gesprächen betrachtet werden können, während sie zugleich Partizipanden an der sozialen Praxis mit wechselnden Aktivitätsniveaus sind. Die Formen dieser Beteiligung an der sozialen Praxis in einer Rolle als „materielle Partizipanden“ (Hirschauer 2004) sind wiederum mannigfaltig und werden auch sprachlich auf verschiedene Weisen realisiert. Dabei wird das sprachliche Material des konversationell agierenden VUIs in die soziale Praxis der menschlichen Beteiligten eingewoben. Der in Kapitel 6.2.3.1. als Beispiel (62) („Du bist so dumm“) gezeigte Ausschnitt weist auf ein weiteres Verfahren hin, mit dem VUIs in den laufenden *turn-by-turn talk* eingebettet werden, darum wird hier ein Ausschnitt aus diesem Beispiel noch einmal diskutiert:

Beispiel (69): Drei (08-CVR01-08)

106 LS: oKAY google,

107 [fünf lauten (.) STärken (.) leiser.]

³²⁵ Studien zum Muster „X und sein* ihr Y“ bzw. „X mit seinem*ihrem Y“ scheinen aktuell ein Forschungsdesiderat darzustellen.

108 k: [((Radio pausiert))]
 109 RL: (xxxx xxx)
 110 (2.0)
 111 LS: [((stöhnt))]
 112 GA: [entSCHULdigung.]
 113 die HöCHste lautstärke ist zehn.
 114 auf welche stufe möchtest du die
 LAUTstärke stellen?
 115 (0.2)
 116 RL: DREI.

Obwohl Lara die ursprüngliche Eingabe formuliert (106–107), antwortet Robin auf die Rückfrage innerhalb der Einschubsequenz (116). Er ‚übernimmt‘ damit den VUI-Dialog von Lara und schließt aus der Prozessperspektive das Adjazenzpaar durch Produktion des zweiten Paarteils ab. Dabei ist die Hörerdeixis des VUI (114) eindeutig: Diese richtet sich auf diejenige Person, die die ursprüngliche Eingabe formuliert hat. Dies zeigt, dass in Mehrparteieninteraktionen, die in VUI-Dialoge eingebettet sind, der Vollzug der Sequenzen einen gewissen Grad an Offenheit zu haben scheint, sodass konditionelle Relevanzen trotz einer direkten Du-Anrede auch von anderen Nutzer*innen übernommen und fortgesetzt werden können; ein sehr ähnlicher Fall ist bei Porcheron et al. (2018: 8) dokumentiert. Das Abschließen der Sequenz wird damit ein kollaborativer Vorgang, in dem die Sprecher*innen ihre Äußerungen entlang der sequenziellen Projektionen verfertigen und die Handlungseinheit so ko-konstruieren. Mit einem offenen Konzept von „Ko-Konstruktion“, wie es etwa Dausendschön-Gay/Gülich/Krafft (2015a) vertreten, lässt sich hier von einem ko-konstruierten Handlungsverlauf sprechen, bei dem einzelne Sequenzen der VUI-Basis-Sequenzstruktur von verschiedenen Sprecher*innen übernommen werden. Mit dieser Perspektive kommt – anders als mit dem zu Beginn der Arbeit eingeführten und primär angewendeten Konzept der „Ko-Operation“ – stärker die interaktionale Verfertigung in den Blick (vgl. Dausendschön-Gay/Gülich/Krafft 2015b: 32), die die entsprechenden Rede-rechtsübernahmen auf der Formseite zu erklären vermögen. Das hier gezeigte Phänomen ist damit keine syntaktische Ko-Konstruktion, wie sie bei Sacks (1995: 528) und darauf aufbauenden Arbeiten etwa von Günthner (2012; 2015) konzeptualisiert sind. Vielmehr lassen sich mit dem Konzept der Ko-Konstruktionen in einem weiteren Sinne das „interactional achievement“ (Schegloff 1982) von Äußerungen und die ‚Hergestelltheit‘ jedes sozialen Handelns fassen. Vorteile bei der Anwendung dieses Konzepts auf Mensch-Maschine-Dialoge unterstreicht auch Pitsch (2015): Die Frage verschiebt sich vom ‚Ob‘ zum ‚Wie‘ der Gestaltung der Ko-Konstruktion mit einem Augenmerk auf „die interaktive Herstellungsleistung und die gemeinsamen

Praktiken der Teilnehmer“ (Pitsch 2015: 253). Die Analyse des Beispiels konnte dabei schon zeigen, dass eine verteilte Bearbeitung von Handlungsschritten möglich ist: Invokation und Spracheingabe werden von Lara vollzogen (106), während die nach der ersten Sprachausgabe notwendig gewordene und vom VUI relevant gesetzte Eingabe von Robin produziert wird (116). Das VUI ist in diesem Sinne also beteiligt an der kollaborativ zu ‚dritt‘ erzeugten Sequenz.³²⁶

Etwas anders gelagert, aber ebenfalls von ko-konstruktivem Charakter ist der Fall der Sequenzenbildung im folgenden Auszug aus der Ersteinrichtung in Haushalt 1. Hier arbeiten die Anwender Alex und Lukas gemeinsam daran, den Smart Speaker und seine Grenzen zu testen. Die ursprüngliche Stimmeingabe ist hochgradig davon geprägt, dass die beiden Nutzer das VUI testen und an seine Grenzen führen (siehe auch Habscheid/Hector/Hrncal 2023):

Beispiel (70): Sieben-Tage-Inzidenzen (01-EE-18)³²⁷

530 AL: +ich kann dir bei
beRECH +nungen in
deinem alltag helfen;
lf: +blickt zurück auf Smartphone in seiner
Hand+
ak: +blickt in Richtung von Alexa----->
>-----+
531 AL: Δegal ob du maßeinheiten UMrechnen,----->
>-----Δ
lf: Δtippt und scrollt fortlaufend auf
SmartphoneΔ
532 die tage bis zu einem feiertag Zählen,
533 oder ein matheproblem Lösen möchtest,
534 ich kann dir dabei HELfen.
535 sag [EINFach-]
536 AK: [alexa WIE viel] sieben tage
inzidenzen sind es noch bis
weihnachten

³²⁶ Ko-Konstruktionen i. e. S., verstanden als kollaborativ erzeugte Äußerungen bzw. „kollaborativ erzeugte[] grammatische[] Konstruktionen“ (Günthner 2015: 57) treten in den vorliegenden Daten im Rahmen von VUI-Dialogen nicht auf.

³²⁷ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00465_SE_01_T_01_DF_01&cID=c665&wID=&textSize=600&contextSize=20.

- 537 (2.5)
- 538 AL: folgendes habe ich im WEB gefunden [und
über]setzt;
- 539 AK: +[hmhm
KLAR.] +
ak: +blickt
zurück auf Laptop+
- 540 AK: +\$((lacht))\$----->
>-----+
ak: +blickt zunächst kurz zu LF, dann in die
Kamera+
lf: \$lächelt \$
- 541 AL: laut (xxx xxx xxx) punkt ORG-
- 542 +mit anderen WORTen; +
ak: +blickt zurück zu Alexa+
- 543 AL: +der FÜNFundzwanzigste dezember beginnt
mit den zwölf weihnachtstagen,+
lf: +wendet Blick von Alexa ab zum Fenster
hin----->
>-----+
544 AL: +die am FÜNFten [januar,]+
lf: +blickt wieder Richtung Alexa+
- 545 LF: [alexa STOP.]
- 546 AK: \$+((lacht)) + \$
ak: \$lehnt sich kurz zurück und direkt wieder
vor\$
ak: +blickt Laptopbildschirm wieder an+
lf: +blickt zurück auf Smartphone-----+
- 547 (1.3)
- 548 LF: **Δ+alexa was ist die inziDENZzahl von
münster?Δ +**
ak: **Δtippt auf Laptop----->**
>-----Δ
lf: +blickt zu Alexa und anschließend zurück
auf Smartphone+
- 549 p : (1.6)
- 550 AL: laut einem alexa (XXX xxx) nutzer,

```

551      münster hat mehrere +POSTleitzahlen.----->
      >-----+
      lf:                               +wendet Blick aus dem
      Fenster hinaus+

```

Wie auch Habscheid/Hector/Hrncal (2023: 16) argumentieren, ist Alex' Erwartung bzw. Zielsetzung der ersten Abfrage (536) nicht eindeutig. Die Aufnahme stammt aus dem Dezember 2020, in dem die über die Presse bzw. über das Robert Koch-Institut bekanntgegebenen Inzidenzzahlen entscheidende Auswirkungen auf die Gestaltung des Soziallebens und auch des bevorstehenden Weihnachtsfests hatten (siehe etwa Timcke/Schneider 2021). Welche Reaktion hier genau erwartet wird, lässt sich nicht sagen, was den spaßhaften Charakter der Abfrage unterstreicht – klar ist lediglich, dass die Stimmeingabe im Kontext des Testens und Herausforderns im Hinblick auf die kurz zuvor behaupteten Fähigkeiten des Geräts zur „Hilfe bei Berechnungen in deinem Alltag“ gesehen werden muss (530–534). In der Formulierung des VUIs wird die Hörerdeixis verwendet – die Äußerung kann insofern als parainteraktives Verfahren verstanden werden. Außerdem ist evident, dass das ‚erwünschte‘ Resultat nicht geliefert wird, denn Lukas bricht die Ausführungen des VUIs, die sich auf die Terminierung der Weihnachtstage beziehen, ab (545). Lukas scheint dann die ursprüngliche Stimmeingabe zu transformieren: Er bleibt im Themenbereich der Corona-Inzidenzen, aber formuliert diese präsentisch und nicht im Hinblick auf einen Zeitraum (548). Allerdings wird das gewünschte Ergebnis erneut nicht geliefert, stattdessen beziehen sich die Äußerungen des VUI nun auf Postleitzahlen (551).

Die thematisch-sprachliche Bezugnahme von Lukas' Stimmeingabe auf Alex' Stimmeingabe lässt sich als ko-operativ hervorgebracht beschreiben (vgl. Goodwin 2018: 7–8; Dausendschön-Gay/Gülich/Krafft 2015b: 31–32). Der Fokus liegt bei dieser Betrachtungsweise auf dem Verhältnis der zwei Versuche zueinander. Betrachtet man die beiden Anläufe der Nutzer mit dem Konzept der Ko-Konstruktion, erscheinen sie als ein gemeinsames kommunikatives Projekt – der Versuch einer Inzidenzabfrage im Modus des Testens. Wie deutlich wird, ist also – wie auch in Beispiel (69) („Drei“) – nicht *innerhalb* der Basis-Sequenzstruktur eine Übernahme der Dialogführung erkennbar. Stattdessen bilden sich größere Einheiten und Versuchsverkettungen, so übernimmt Lukas *nach* der Produktion der Stimmausgabe durch das VUI das Rederecht. Er reformuliert Alex' Stimmeingabe (548), verändert sie dabei aber so, dass sie für das VUI (vermeintlich) leichter zu verarbeiten sein müsste. Der ko-konstruktive Charakter der Hervorbringung der Stimmeingabe zeigt sich hier also nicht innerhalb der Basis-Sequenzstruktur durch eine Rederechtsübernahme zwischen Invokation und Stimmeingabe, sondern durch ein *trial and error*-Verfahren, Sequenz für Sequenz.

6.2.3.6 Überlagerndes Sprechen

Die bisher gezeigten sprachlichen Merkmale zeigten v. a., wie in den Mehrparteiensettings das VUI als Partizipand an der Praxis (vgl. Hirschauer 2004) und als ‚Teilnehmer‘ am Gespräch verhandelt wird. Es zeigen sich jedoch auch sprachliche Verfahren der Teilnehmenden, bei deren genauerer Bedeutung deutlich wird, dass VUIs gerade *nicht* als Gesprächsbeteiligte betrachtet werden. Dazu zählt die Fortsetzung einer laufenden Interaktion zwischen zwei menschlichen Beteiligten, während das VUI Stimmausgaben produziert, wie im folgenden Beispiel aus der Ersteinrichtung im Haushalt von Lukas und Alex.

Beispiel (71): Mischmasch (01-EE-11)³²⁸

459 (1.49)
 460 AL: [((spielt
 Startjingle, 7.68 Sek.))]
 461 AK: du hast dann die app dadrauf [und auf der
 connectest du dann mit SPOTify und den
 ganzen anderen sachen.]
 462 (0.38)
 463 LF: der kann *darüber *
 dann (.) ECHT ganz-
 lf: *zeigt mit rechtem Daumen auf AL*
 464 (0.22)
 465 LF: [ganz geNAU;]
 k: [((Startjingle endet hier))]
 466 AL: [heLLo-]
 467 LF: [\$ (KENN ich mich nich aus;)] \$
 lf: \$streichs sich mit LH durch Haare\$
 468 AL: \$Δyour dEvice is ready for SETup.Δ\$
 lf: \$stützt Kopf auf Hand ab-----\$
 Δtippt auf Handy-----Δ
 469 AK: ((schmatzt)) AH sonst ham-
 470 AL: [JUST download the alexa-=]
 471 AK: [wir so wieder so_n MISCHmasch;]
 472 k: ((Klappern))

³²⁸ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00465_SE_01_T_01_DF_01&cID=c402&wID=&textSize=500&contextSize=16.

473 AL: =app-
 474 AL: [and-]
 475 LF: [ich HAB-]
 476 AK: [((unverständlich))]
 477 AL: [folLOW the-=]
 478 LF: [*das wieder auf die-] *
 lf: *zeigt mit rechtem Zeigefinger auf Handy*
 479 AK: [spotiFY-]
 480 AL: [=inSTRUctions;]
 481 AK: PREmium,
 482 AK: [das heißt ich connecte dann wieder auf
 mei:nem (.) spotify acCOUNT?]
 483 AL: [HALlo (.) dein gerät kann jetzt
 *eingerichtet werden; *]
 lf: *blickt AL, dann Handy an*
 484 AK: auf DEInem-=
 485 AL: [LAde dazu die alexa app runter und-]
 486 AK: [=handy irgendnen kram REIN,]
 487 AL: folge den ΔANweisungen. Δ
 lf: Δtippt auf HandyΔ
 488 (0.22)
 489 AL: bonJOUR (.) votre appareil est pret pour
 la configuration;
 490 (0.32)
 491 AK: ich-
 492 AK: [find_s (irgendwie) übertrieben nerVig
 dass man so was
 immer PER (.) smart-=]
 493 AL: [teleCHARgez l_application alexa et suivez
 les instructions.]
 494 AK: =phone APP;
 495 AL: [ciao-]
 496 AK: in[stallie]ren MUSS;
 497 AL: ΔIL dispositivo e pronto per la----->
 >-----Δ
 lf: Δwendet sich zu A1 und drückt mehrfach auf
 Leiser-KnopfΔ
 498 AK: [(bin) FAST schon-]
 499 AL: [configuraZIOne-]
 500 AK: geZWUNgen,

501 AL: [scarica l_app aLEXa e-]
 502 AK: [<<p> MIR n neues Δhandy zu kaufen;>]---->
 >-----Δ
 ak: Δbeginnt auf Laptop zu
 tippenΔ
 503 AL: segui LE istruziOni-
 504 (0.39)
 505 AL: ΔhoLA- Δ
 lf: Δbeginnt auf Handy zu tippenΔ
 506 (0.51)
 507 AL: el dispositivo esta LISto para que lo
 configures;
 508 (0.37)
 509 AL: descarga la APP alexa y sigue las
 instrucciones (.)-
 510 AL: ((spielt Melodie))
 511 (3.04)
 512 LF: +A:H du musst mir noch ma das passwort
 geben von de:m;+
 lf: +blickt zweimal zwischen Handy und AK hin
 und her +
 513 (0.57)

Die Äußerungen des VUI sind eindeutig an die Anwender adressiert, unter Verwendung der Hörerdeixis richten sie sich an den primären Anwender und artikulieren die Aufforderung, die notwendige App herunterzuladen. An diesem Punkt im Ablaufskript für die Einrichtung des Smart Speakers war Lukas bereits angekommen: Da er die App zu diesem Zeitpunkt noch nicht installiert hatte, hatte er die Stromversorgung unterbrochen, um auch die Äußerungen des VUI zu stoppen.³²⁹ Die Inhalte sind Lukas und Alex also grundsätzlich vertraut, zumal es sich im Wesentlichen um die Wiederholung derselben Aufforderung in verschiedenen Sprachen handelt. Möglicherweise deswegen setzen Alex und Lukas ihre Interaktion trotz der wiederholt an sie gerichteten Aufforderung fort und sprechen zu großen Teilen überlappend mit dem VUI (u. a. 466, 469–478, 485–493). Dabei diskutieren sie über notwendige Konfigurationen und technische Voraussetzungen zur Anwendung des Smart Speakers. Das Reden des VUIs wird dabei von den Beteiligten nicht als Teil ihres eigenen Gesprächs behandelt und ihr eigenes Reden

³²⁹ Siehe Beispiel (64) („Dein Gerät kann jetzt eingerichtet werden“).

verbindet sich nicht – wie teilweise in vorherigen Analysen gezeigt – mit einem VUI-Dialog. Dies macht noch einmal deutlich, dass ein VUI dies nicht einfordern kann, sich insofern nicht als sozialer Akteur positionieren kann, der im Zweifel das gemeinsame Tun verlassen und sich der Situation entziehen kann – und demgegenüber entsprechend auch von den Anwender*innen kein „Face-Work“ (Goffman 1955) betrieben und sprachlich gesehen keine Höflichkeit gewahrt werden muss.

6.2.3.7 Pronominalreferenzen: Zwischen Anapher und Objektdeixis

Bisher konnte gezeigt werden, dass nicht nur direkte Adressierungen (wie in Kap. 6.2.3.1 illustriert), sondern auch andere sprachliche Verfahren dazu dienen können, eine formale Beteiligung des VUI herzustellen und die dadurch an der Oberfläche entstehende Dialogizität als interaktive Ressource zu gebrauchen. In den gezeigten Beispielen wird aber auch deutlich, dass ebenso *über* den Smart Speaker bzw. das VUI geredet wird. Der von Habscheid/Hector/Hrncal (2023) eingehend diskutierte Wechsel der dialogischen Orientierung – aus dem VUI-Dialog austretend in ein „meta-interaction regime“ (Habscheid/Hector/Hrncal 2023: 20) – ist eine Strategie von Nutzenden, bei der Ausführung von Alltagspraktiken ihre Handlungspotenziale auch dann zu erhalten, wenn die Abhängigkeit von den Operationsprinzipien des VUI situativ in den Vordergrund rückt, etwa bei Störungen (vgl. Habscheid/Hector/Hrncal 2023: 24–25). Der „Meta-Interaktionsraum“ (Habscheid 2022: 168) wird von den Nutzenden eröffnet, um miteinander – und explizit *ohne* direkte Adressierung und Beteiligung des VUI – den Status von Äußerungen und Beteiligungen im Zusammenhang mit den VUIs zu diskutieren, die Funktionen zu evaluieren und VUIs zu testen bzw. spielerisch an ihre Grenzen zu führen (vgl. Habscheid 2022: 179–182; siehe auch Pitsch et al. 2017: 393; Porcheron et al. 2018: 4–6; Beneteau et al. 2019: 7–9). Dabei werden Smart Speaker behandelt wie andere unbelebte Gegenstände – Habscheid (2022: 181) etwa beschreibt eine Situation, in der Nutzer*innen „had no problems talking about the technical dialogue partner in his presence like one is talking about other present things“.

Personalpronomen werden typischerweise als anaphorisch bezeichnet. Sie zeigen die Wiederaufnahme von bereits bekannten Referenten – „bereits eingeführte oder sonst mental präsente Gegenstände oder Sachverhalte“ (Zifonun/Hoffmann/Strecker 1997: 544) – an. In anderen Verwendungsweisen können Personalpronomen aber auch als Ausdruck der Objektdeixis verstanden werden. Bei Demonstrativpronomen ist die Zuordnung umgekehrt: Typischerweise werden sie dem Bereich der Deixis zugeschlagen, können aber auch anaphorisch gebraucht werden. Der Bereich der *Objektdeixis* ist keine traditionelle deiktische

Kategorie (vgl. Finkbeiner 2018: 191) in der für die Deixis-Forschung grundlegenden Sprachtheorie Karl Bühlers. Sie wird aber gleichwohl in linguistischen Kompendien (Lenz 2013), Einführungen (Meibauer et al. 2015: 212) und Grammatiken (Zifonun/Hoffmann/Strecker 1997: 41) aufgeführt. Dabei meint Objektdeixis Verweise „auf Objekte im weitesten Sinn und Personen“, sofern sie sich mit den Sprecher*innen „in einem Verweisraum [...] befinden“ (Zifonun/Hoffmann/Strecker 1997: 323). Lenz (2013) spricht von „Objekten im *Wahrnehmungsraum* [Herv. T.H.]“. Zur Unterscheidung von Deixis und Anaphorik sind die Arbeiten von Konrad Ehlich einschlägig: Die „deiktische Prozedur“ (Ehlich 1983: 85–86) bestimmt er als Fokussierung auf einen (neuen) Referenten innerhalb des Verweisraums, ein „anaphorischer Prozeß“ (Ehlich 1983: 96) hingegen dient der Bezugnahme auf einen bereits fokussierten Referenten (siehe auch Ehlich 1979; Tanaka 2011: 14).

Betrachten wir nun ein Beispiel einer pronominalen Referenz im Kontext des Smart-Speaker-Gebrauchs:

Beispiel (72): Lass ihn doch (08-EE-37)³³⁰

1238 (1.17)
 1239 SM: spiele DEUTSCHlandfunk nachrichten;
 1240 (0.22)
 1241 SM: weil (.) <<lachend> [das is]
 mein morgenWUNSCH immer;>
 1242 RL: [((räuspert sich))]
 1243 (0.3)
 1244 SM: °h
 1245 RL: lol;
 1246 (0.21)
 1247 GA: hier sind die aktuellen NACHrichten;
 1248 (0.26)
 1249 RL: ((schmatzt)) h°
 1250 SM: +aber jetzt (.) BITte- h° nicht-+
 sm: +schaut GA an +
 1251 (0.68)
 1252 SM: er te: EL oder so was;

330 Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c1125&wID=&textSize=500&contextSize=16.

1253 GA: von DEUTSCHlandf[unk nachrich]ten
[heute] um zwanzig uhr;
1254 SM: *((((schmatzt)) a:h.)----->
>-----*
sm: *wirft den Kopf nach
hinten*
1255 RL: [°h]
1256 RL: °h
1257 (0.9)
1258 SM: +[oKAY google]+
sm: +fasst sich an die Lippen +
1259 RL: [<<behaucht> DEUTSCHlandfunk.>]
1260 (0.91)
1261 RL: ja LASS ihn [doch.]
1262 SM: [SPIE]le,
1263 RL: +oder NICH; +
sm: +schaut RL an+
1264 (0.26)
1265 SM: ((unverständlich)) +weit (.) erzähl uns
vom TAG;+
sm: +schaut GA an----->
>-----+
1266 RL: Oder,
1267 (0.53)

Die Haushaltsmitglieder Samuel, Robin und Lara testen die Funktionen des Google Assistant nach dessen Ersteinrichtung auf Robins Smartphone (die Verknüpfung mit Laras Smartphone, die als Teil der Ersteinrichtung zu verstehen ist, läuft parallel weiter). Samuel möchte testen, ob die Deutschlandfunk-Nachrichten wiedergegeben werden – ein mögliches Nutzungsszenario auch für die Zukunft (vgl. 1241). Nachdem durch die Stimmausgabe des VUIs bestätigt wird, dass diese Stimmeingabe erfolgreich verarbeitet werden kann und dies von Samuel ratifiziert wird (vgl. 1254–1255), setzt Samuel an, eine weitere Funktion zu testen und produziert dazu eine Invokation (vgl. 1258), die zugleich Kundgabe und Beginn seines Vorhabens ist. Während der kurzen Pause zwischen Invokation und Stimmeingabe äußert Robin „ja LASS ihn [doch.]“ (1261). Nachdem sich diese Äußerung mit dem Beginn der Stimmeingabe von Samuel überlappt hat (1261–1262), zieht er seinen Vorstoß wieder zurück (1263) und Samuel produziert eine neue Stimmeingabe, die die Wiedergabe der Nachrichten stoppt.

Bevor die Pronominalreferenz (1261) genauer beleuchtet wird, soll kurz hervorgehoben werden, dass das Beispiel auch im Hinblick auf die Aushandlung der Bedienung interessant ist: Robin passt genau die kurze Stille zwischen der Invokations- und der Stimmeingabe-Sequenz ab, um einen Einwand gegen deren Produktion zu formulieren (siehe auch Porcheron et al. 2018: 5). Dieser Einwand kann jedoch nicht mehr ohne Weiteres berücksichtigt werden: Der Listening-Modus ist bereits gestartet und Samuel hat mit der Produktion der Stimmeingabe bereits begonnen. Kurzerhand annulliert Robin das kommunikative Projekt des Einwands wieder (1263) und gibt damit dem Projekt, eine weitere Funktion zu testen, ‚Vorfahrt‘. Die kurze Stelle zwischen Invokation und Formulierung der Stimmeingabe erweist sich damit auch hier als sensitiv für Aushandlungen (siehe auch Pitsch et al. 2017) – zugleich aber auch als anfällig dafür, dass kommunikative Projekte scheitern, wie hier deutlich wird: Die Invokation von Samuel kündigt nicht nur an, eine weitere Funktion testen zu wollen, sondern Samuel steigt mit dieser Äußerung auch in die Produktion der für den Test notwendigen Sequenzstruktur ein. Für den Einwand von Robin ist es an dieser Stelle zu spät, weil kein unkomplizierter Weg besteht, den Listening-Modus wieder zu beenden.

Die Pronominalreferenz (1261) – das Personalpronomen in der 3. Person, Akkusativ, Singular – scheint tendenziell anaphorisch gebraucht zu sein: Im Fokus steht die Beibehaltung und Fortsetzung des Fokus auf einen bereits eingeführten, „mental präsent[n]“ (Zifonun/Hoffmann/Strecker 1997: 323) Gegenstand. Die genauere Bestimmung verweist jedoch auf ein noch grundlegenderes Problem für die Analyse: Der genaue Referent der Äußerung „ihn“ ist unklar. Damit gemeint sein könnte der Lautsprecher in seiner Gesamtheit, ebenso aber auch die im Interface auftretende, in diesem Fall namenlose, nur als Stimme repräsentierte „Persona“ von Google. Als Anapher gelesen hat „ihn“ kein Antezedens, das darüber Auskunft geben könnte, es geht vielmehr um das akustisch und visuell präsente Objekt im geteilten Wahrnehmungsraum der Beteiligten. Aus pragmatischer Perspektive gesehen wird allerdings auch keine deiktische Prozedur vollzogen: Es wird nicht auf einen mit der pronominalen Äußerung zu spezifizierenden Referenten die Aufmerksamkeit gerichtet. Für die eher anaphorische Einordnung spricht auch, dass das Pronomen nicht akzentuiert ist, wie es sonst beim deiktischen Gebrauch von Personalpronomen der Fall sei (vgl. Pafel 2015: 212). Für die Teilnehmer*innen ist dies gleichwohl unproblematisch: Sie akzeptieren die Vagheit dieser Referenz auf die (sozio)technische Einheit.

Die Verwendung des (hier männlichen) Genus zur Referenz auf den Smart Speaker bzw. auf seine Bestandteile oder dessen Interface-Repräsentation ist in den vorliegenden Daten nicht einheitlich, auch einzelne Sprecher*innen variieren das Genus in pronominalen Referenzen. Deren Vorkommen lassen keinen Rückschluss darüber zu, in welchen Zusammenhängen welches Genus gebraucht

wird. Dazu soll noch einmal ein Auszug aus dem Beispiel „Wie macht der Otter II“ (Kap. 6.2.3.4) betrachtet werden:

Beispiel (73) Wie macht der Otter II (08-EE-17)³³¹

077 GA: dAs ist ein OTter.
 078 §((spielt Quieken eines Otters
 ab)) §
 rl: §öffnet und schließt RH, wendet Kopf nach
 links§
 079 SM: [((lacht leise))]
 080 RL: [((lacht leise))]
 081 LS: [((lacht leise))]
 082 SM: [+wär lustiger wenn_s IHre stimme wäre;+]
 ls: +lächelt-----+
 083 GA: [(spielt Quieken eines Otters ab))]
 084 SM: [((lacht))]
 085 LS: [§<<lachend> JA ne?>§]
 ls: §wendet sich SM zu-§
 086 RL: ja.
 087 SM: [((lacht))]
 088 LS: [<<lachend> ich hab auch eigentlich damit
 gerechnet dass es> ihr
 SIE macht.]

Hier nehmen sowohl Samuel als auch Lara mit dem Possessiv- bzw. Personalpronomen Bezug auf die Stimme des VUIs. Dabei nutzen beide das feminine Genus, anders als im zuvor gezeigten Beispiel. Die Abgrenzung von anaphorischem zu deiktischem Gebrauch scheint hier weiter zu verschwimmen: Durch die prosodische Akzentuierung wird „IHre“ Stimme von den Otterlauten unterschieden – dies wäre ein Argument für eine deiktische Einordnung. Ein passendes Antezedens für eine Anapher (nun Femininum) ist wie auch in den vorherigen Beispielen nicht zu finden. Gleichwohl wird der Gegenstand nicht, wie bei deiktischen Prozeduren, als solcher eingeführt, sondern stand bereits vorher in Rede. Eine abschließende Einordnung lässt sich hier nicht vornehmen. Vielmehr ist erneut zu

³³¹ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c484&wID=&textSize=300&contextSize=8.

bemerken, dass die Nutzer*innen keine Schwierigkeiten in der Verarbeitung der referenziellen Verweise dokumentieren und diese als Teil der Vagheit in der Interaktion zu akzeptieren scheinen. Darunter fallen auch die bereits beobachteten Genuswechsel, die als Aspekt der Pronominalreferenzen noch einmal beleuchtet werden sollen. Das ist ein Hinweis darauf, dass sich klare Genuszuordnungen für die Referenz auf Maschinen oder Teile davon (noch) nicht etabliert haben.

Einen Hinweis auf die Genuszuweisung durch die Sprecher*innen könnte allerdings die Einstellung des Stimmklangs des VUIs geben. Wie die beiden folgenden Beispiele zeigen, wird bei der Ersteinrichtung mit dem Stimmklang auch Geschlechterzugehörigkeit verhandelt bzw. festgelegt:

Beispiel (74): Lass den Mann nehmen (08-EE-22)³³²

```

676 GA: wenn dir dies[e stim ][me geFÄLLT,      ]
677 LS:                [ja lass][den MANN_nehmen;]
678 SM:                [ja;      ]
679 k: ((GA wechselt die Stimmfarbe))
680 GA: und du MÖCHtest (.) dass ich sie weiterhin
        [verwende,      ]
681                [kannst du hier AUfhören;]
682 RL: [xxx xxx      ][(normAle STIMme;)      ]
683 SM: [die_s zwar    ][nicht GUT;=ab]er BESser;]
684 LS: [die_s viel ang][enehmer;      ]
685 RL: ja;
686      °hh
687      (1.05)
688 RL: +oKAY google;      +
        rl: +blickt GA an-----+
689      (0.69)
690 RL: +bist du jetzt ein MANN?+
        rl: +blickt GA an-----+
691      (1.93)

```

332 Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c599&wID=&textSize=300&contextSize=8.

692 GA: ich bin virtuell *und habe deshalb kein
 geSCHLECHT;*
 rl: *wendet Kopf nickend zu
 SM-----*

Beispiel (75): Klassische Frau (07-EE-18)³³³

587 KW: ah ich kann hier (meine) SIri äh stImmen?
 588 p: (2.71)
 589 SI: <<mit männlicher Stimme sprechend>
 +HALlo (.) ich bin sIri;>+
 kw: +blickt TW an-----+
 tw: +blickt KW an-----+
 590 KW: ah NEE wir [wollen
 lieber;]
 591 SI: <<mit männlicher Stimme sprechend>[wähle
 die STIM]me
 [aus;]>
 592 TW: [\$wir wollen (die)] (.) die klassische
 [FRAU oder,]\$
 tw: \$führt RH von Hinterkopf auf Knie----->
 >-----\$
 593 SI: <<mit weiblicher Stimme
 sprechend>[HALlo ich b]in sIri;>
 594 KW: ja;

Die Zuschreibung erfolgt im Beispiel (74) nur auf Grundlage der *Interpretation* des Stimmklangs: Die visuelle Oberfläche von Google nimmt keine Zuweisung einer der beiden Stimmen zu einem Geschlecht vor, sondern unterscheidet nur die ‚Stimmfarben‘ rot und orange (siehe Abb. 30 und 31). Konkret nach der Geschlechtszugehörigkeit des Assistenten gefragt, antwortet das VUI, es habe kein Geschlecht (692). In Haushalt 7 (Beispiel 75) wird ein Apple-Gerät eingerichtet, hier werden die Stimmen auch systemseitig als „männlich“ bzw. „weiblich“ attribuiert. In beiden Fällen leiten die Nutzer*innen aus der Stimmfarbe auch eine geschlechtsspezifische Personenkonstruktion ab. So schlägt Lara in einer verkürzten,

³³³ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00467_SE_01_T_01_DF_01&cID=c524&wID=&textSize=200&contextSize=4.

adhortativen Form vor: „lass den MANN_nehmen“ (677) und Till vergewissert sich, dass er und Konrad „die klassische FRAU“ (592) einstellen wollen. Die Einstellung der Stimmfarbe kann also unmittelbar mit den Personenkonstruktionen der Anwender*innen verknüpft sein.

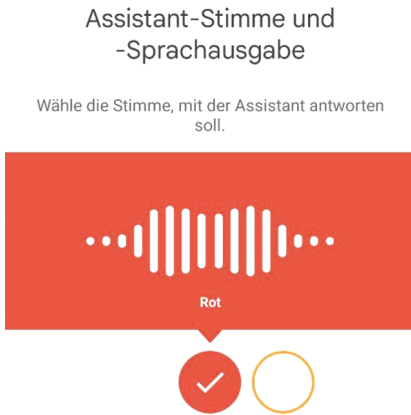


Abb. 30: Stimmauswahl für Sprachausgabe in der Google-Home-App – Rot; Bild: Screenshot / T.H.

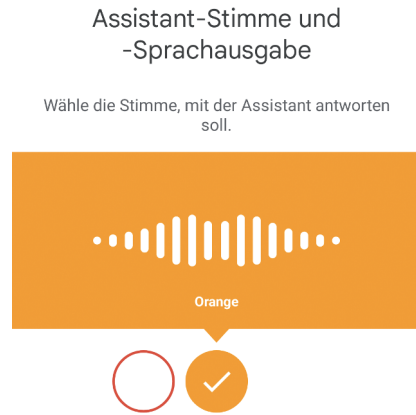


Abb. 31: Stimmauswahl für Sprachausgabe in der Google-Home-App - Orange; Bild: Screenshot / T.H.

Wie stabil sind diese Konstruktionen allerdings auf der Ausdruckseite in Form der pronominalen Referenzen? Tatsächlich liegt zeitlich zwischen den beiden weiter oben gezeigten Beispielen aus Haushalt 8 die Änderung des Stimmklangs: Diese war im Beispiel „Wie macht der Otter II“, in dem Lara und Samuel mit Pronomen im Femininum auf den Smart Speaker bzw. dessen Interface referieren, noch in der Voreinstellung und somit in der von den Nutzer*innen als weiblich verstandene Stimmfarbe (rot). Der davor gezeigte Ausschnitt, in dem das Pronomen im Maskulinum gebraucht wird, ereignet sich deutlich später im Verlauf der Ersteinrichtung. Daraus ließe sich auf den ersten Blick schließen, dass die Personenkonstruktion, mit der das VUI verknüpft ist, Bestand hat und sich in einer konsequenten Genuswahl für die Deklination der genusspezifischen Pronomina ausdrückt. Allerdings zeigen weitere Ausschnitte in beiden Haushalten, dass diese Stabilität gerade nicht gegeben ist:

Beispiel (76): Ton abspielen (07-EE-19)³³⁴

598 KW: ä:hm-
 599 KW: h°
 600 TW: hhh°
 601 KW: WOLLen wir,=
 602 =wenn wir **das** aktiVIEn;=
 603 =wollen wir dann n +TON abspielen?----->
 >-----+
 kw: +blickt TW an und dann
 zurück auf Handy+
 604 (1.63)
 605 TW: Ä:H pf;
 606 KW: also wenn <<p> ich ihr das SAge;
 607 TW: jo:,

Diese Äußerungen aus Haushalt 7 werden nur vier bis fünf Sekunden nach der Festlegung des Stimmklangs als „weiblich“ produziert. Dennoch referiert Konrad mit „das“ (602) auf den vagen Referenten, mit dem der Smart Speaker oder einzelne seiner Funktionen gemeint sein könnten. In seiner Folgeäußerung, die eine nähere Spezifikation zu seiner vorherigen Frage ist („also wenn ich ihr das SAge“, 606), wechselt er erneut das Genus. In dieser Äußerung steht nun der dialogische Charakter der Bedienweise des Geräts im Zentrum und dies könnte wiederum als Erklärung herangezogen werden für die Verwendung des Pronomens im Femininum, während zuvor möglicherweise das Gerät als Ganzes im Zentrum stand.

Obschon solche Überlegungen nicht gänzlich zurückgewiesen werden sollen, kann mit Blick auf andere Belegstellen, wie eingangs erwähnt, eine konsistente Genuserwendung bei den pronominalen Referenzen nicht festgestellt werden. In Haushalt 8 schien die anfängliche Wahl des Femininums nach dem Wechsel der Stimmfarbe angepasst zu werden. Allerdings ist in den CVR-Aufnahmen, die nur wenige Minuten nach der Ersteinrichtung aufgezeichnet wurden, wieder eine Verwendung des Femininums durch Samuel und Robin belegt:

³³⁴ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00467_SE_01_T_01_DF_01&cID=c534&wID=&textSize=200&contextSize=4.

Beispiel (77): Die ist echt schlecht (08-CVR-05)

161 RL: <<lauter> welche terMIne stehen in meinem
kalender?>
162 (1.7)
163 GA: entSCHULdigung.
164 ich kann KEIne einträge in deinem
[kalender finden.]
165 RL: [ALter-]
166 SM: **die** ist echt SCHLECHT ne,
167 RL: **die** ist ECHT schon scheiße ja,
168 (0.3)

Auch diese Wahl ist jedoch nicht konsistent, wie andere CVR-Aufzeichnungen aus Haushalt 8 belegen, in denen Robin und Lara wieder mit Personalpronomen im Maskulinum auf das Gerät referieren. Auch Alex in Haushalt 1 wechselt innerhalb kurzer Zeit das Genus:

Beispiel (78): Zwei Stimmprofile (01-EE-16)³³⁵

688 AK: \$ich geh davon aus **die** kann auch ZWEI
stimmprofile-\$
ak: \$stützt LH auf Oberschenkel und ist AL
zugewandt----\$
689 AK: [weil **sie** jetzt schon Eben-]
690 LF: [ich HOFF es;]
691 AK: meine (.) meine frage nach dem anruf für
deine MUTter,
692 (0.58)
693 AK: geCHECKT hat,
694 AK: \$Δ((lacht, 1.7 Sek.)) \$Δ
ak: \$lehnt sich nach hinten\$
Δund greift nach LaptopΔ

³³⁵ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00465_SE_01_T_01_DF_01&cID=c580&wID=&textSize=200&contextSize=4.

Beispiel (79): Playlist gehijackt (01-EE-21)³³⁶

- 941 AK AH welche (gute) playlist hat_n der jetzt
hier gehijackt.
- 942 p (3.64)
- 943 AK das is DEfinitiv ein (.) song aus (einer)
(.) playlist von mir.

Wie diese beiden Auszüge aus Haushalt 1 noch einmal verdeutlichen, sind semantische Herleitungen für die Genuswahl nur schwer haltbar und werden auch von den anderen Beteiligten nicht relevant gesetzt oder von den Sprecher*innen *accountable* gemacht: Bei der Auswahl von „die“ (688) und „sie“ (689) in Beispiel (78) scheint einerseits eher die technische Funktionalität im Vordergrund zu stehen (zwei Stimmprofile), gleichzeitig nimmt Alex auf einen Vorfall Bezug, der sich im Verlauf der Ersteinrichtung ereignet hat – die Planung eines nicht intendierten Anrufs an Lukas’ Mutter – und bei der das VUI als hörendes und sprechendes Interface im Zentrum stand. Wie diese beiden semantischen Ebenen mit dem Genus verknüpft sein könnten, erschließt sich nicht, wie auch die Auswahl des maskulin deklinierten Demonstrativpronomens „der“ (941) im darauf folgenden Beispiel (79) nicht verständlich wird. Da der genaue Referent dieser Äußerungen nicht klar ist, außerdem keine eindeutigen Antezedens-Beziehungen vorliegen und die Verwendung der Pronomen zudem von der anaphorischen aus zu einer objektdeiktischen Funktion neigen, lässt dies keine weitere Schlussfolgerungen zur Genuswahl zu.

Mit diesen Beobachtungen ist jedoch explizit *kein* Material geliefert, das kritische, feministische Perspektiven auf Smart Speaker untergraben könnte – im Gegenteil. Etwa die Argumentation von Natale/Cooke (2021: 1009–1010), die darauf abstellt, dass gegenderte Stimmen als semiotische Repräsentationen von Personenkonstruktionen Stereotype reproduzieren (siehe auch Both 2015; Phan 2017; Schiller/McMahon 2019; Loideain/Adams 2020; Chambers 2020),³³⁷ wird durch die vorliegenden Daten gestärkt. Sogar trotz der herstellerseitigen Nicht-Zuweisung von Geschlecht bei Google vollziehen die Nutzer*innen im gegebenen Beispiel die Konstruktion einer Personenfiktion mit binären Geschlechtszuschreibungen auf Grundlage der Stimmfarbe, die einer menschlichen Stimme nachempfunden ist. Auch die Analyse des Beispiels deutet auf die Entstehung einer stabilen, genger-

³³⁶ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00465_SE_01_T_01_DF_01&cID=c802&wID=&textSize=200&contextSize=4.

³³⁷ Für eine ausführlichere Diskussion siehe Kap. 3.2.5.

ten Personenfiktionen hin. Zusammengenommen zeigt dies, dass eine Reproduktion von genderbasierten Stereotypen durch die unmittelbaren Nutzungserfahrungen jedenfalls unterstützt wird. Die Äußerung von Till in Beispiel (75) („wir wollen die klassische FRAU oder“, 592) ist ferner ein Hinweis darauf, dass sich die weiblich verstandene Stimme als Standard bereits diskursiv etabliert zu haben scheint und von Beginn der Nutzung an in den Privathaushalten reproduziert wird. Mit Blick auf die Frage der Teilnahme an Gesprächen unterstützt die wechselnde Genuszuweisung bei den pronominalen Referenzen auf den Smart Speaker bzw. auf seine Komponenten die Hypothese, dass der Status eines VUIs als „Beteiligter“ am Gespräch brüchig ist und im Reden *über* die Geräte im „Meta-Interaktionsraum“ (Habscheid 2022: 168) kaum Bestand hat. Dass bei diesen Äußerungen ein hoher Grad an Vagheit in Bezug auf den Referenten besteht, stützt diesen Eindruck. Zudem sind von den anaphorischen oder objektdeiktischen Referenzen Wechsel in die Hörerdeixis und zurück in hoher Geschwindigkeit und Adjazenz beobachtbar.

6.2.3.8 Multimodale Verfahren

Quer zu den bisher diskutierten sprachlichen Praktiken liegt der Aspekt der Multimodalität:³³⁸ Sprachlichkeit, Körperlichkeit und die Räumlichkeit in ihrer physischen Materialität sind simultan auftretende, sich wechselseitig bedingende Ressourcen, die in der Praxis zur Lösung kommunikativer Aufgaben herangezogen werden. Die bisherigen Analysen haben sich auf die Verbal sprachlichkeit konzentriert, zugleich haben sie – wo das Datenmaterial es zuließ – ihren multimodalen Charakter teilweise berücksichtigt. So sind erstens Aspekte der Gestik und der Körperhaltung bereits in mehrere Analysen eingeflossen,³³⁹ zweitens ist auch das stets als Zusammenspiel verschiedener sprachlicher Ausdrucksressourcen auftretende verbalsprachliche Material unter Einbezug dessen analysiert worden: So wurden etwa prosodische, lexikalische und syntaktische Verfahren als aufeinander bezogen und sich wechselseitig ergänzend betrachtet. Insofern sind alle gezeigten sprachlichen Praktiken nicht ohne ihre Multimodalität zu denken – wo sie sich aufgrund des rein auditiven Datenmaterials einer Analyse entziehen, eröffnen sie Anschlussmöglichkeiten für weitere Studien.

Ein separates Kapitel soll hier der Multimodalität gewidmet sein, um zu zeigen, wie verschiedene Ausdrucksressourcen in einem soziotechnischen Arrangement zwischen menschlichen und technischen Partizipanden zur Konstitution

³³⁸ Siehe Kap. 4.4 zur Multimodalität.

³³⁹ Siehe u. a. die Analysen zu Beispiel (63) („Was hast du noch für Stimmen?“) in Kap. 6.2.3.2 und Beispiel (64) („Dein Gerät kann jetzt eingerichtet werden“) in Kap. 6.2.3.3.

der Beteiligungsstruktur beitragen.³⁴⁰ Im Folgenden sollen daher anhand eines einzigen längeren Auszugs aus der Ersteinrichtung in Haushalt 8 die Bereiche Körperhaltung, Gestik, Blick, Raum und Objektmanipulation beleuchtet werden. Der gewählte Ausschnitt umfasst insgesamt zwei Minuten und ca. acht Sekunden, die sich ereignen kurz nachdem der Smart Speaker aktiviert wurde und die Begrüßung über das VUI abgespielt hat (001–003).³⁴¹ An dieser Stelle wird zunächst der gewählte Ausschnitt insgesamt unter Verzicht auf die multimodale Annotation dargestellt, um einen lesefreundlichen Überblick zu ermöglichen. Anschließend werden detaillierter transkribierte Auszüge aus diesem Ausschnitt besprochen (Beispiele 81 bis 85).

Beispiel (80): Wie macht der Otter III (08-EE-17 bis 19)³⁴²

001 RL: dein GOOgle assistant wird dir antworten
wenn du [hey google-]
002 GA: [hallo i]ch
bin dein GOOgle assistant.
003 GA: ich bin HIER um dir zu hElfen;
004 GA: in der goo[gle HOME a]pp
kannst du dich über meine funktionen
informIERen;
005 SM: [°h ((lacht))]
006 LS: [((stimmhaftes Räuspern))]
007 SM: [°h;]
008 [(1.2)]
009 RL: mHM.
010 SM: [°h;]
011 LS: [ke.]

340 Es wäre ebenso möglich gewesen, die einzelnen Phänomene an verschiedenen Auszügen aus unterschiedlichen Haushalten zu zeigen. Durch die Wahl eines einzigen Ausschnitts mit (im Vergleich zu den übrigen Beispielen) längerer Dauer können jedoch sequenzielle Bezüge mit in den Blick kommen, die sich im zeitlichen Verlauf des Ausschnitts ergeben. So wird die Verwobenheit der multimodalen Ausdrucksressourcen (verbalsprachliche und darüberhinausgehende) unterstrichen. Weitere Auszüge über diesen Ausschnitt hinaus zu zeigen, hätte hingegen primär zu Redundanzen geführt.

341 Teile des hier gezeigten Ausschnitts „Wie macht der Otter III“ wurden unter anderen Aspekten als Beispiel (54) in Kap. 6.2.2.2 und als Beispiel (57) in Kap. 6.2.2.4 dieser Arbeit besprochen.

342 Dieser Ausschnitt kann über das Sprechereignis „FOLK_E_00484_SE_01_T_01“ in der DGD abgerufen werden (Z. 402 bis 517). URL: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayvideotranscript&id=FOLK_E00484_SE_01_T_01.

012 RL: <<liest aus der Anleitung vor>SAG einfach
 hm hm hm.
 013 dann SPIEL mU:sik;
 014 °h WIE sagt man (.) wie geht es dir auf
 portugiesisch?
 015 wie steht der DAX;
 016 [wo is die nächste bäckerEI. >]
 017 LS: [(.) <<lachend> °h wie steht der DAX?>]
 018 SM: [<<lachend und nuschelnd> (oh das is)>
 °h klas]
 sischer (.) klassischer MORgen bei mir.
 019 LS: [((lacht)) ja.]
 020 RL: [is mein FLUG pünktlich?]
 021 RL: [<<liest aus der Anleitung vor> wann ist
 MORgen mein erster termi:n?>]
 022 LS: [erstMA aufwachen (.) <<leise nuschelnd>
 (oh ma:n wie steht der dAx?)>]
 023 RL: <<liest aus der Anleitung vor> erZÄHL mir
 etwas lustiges.
 024 wie kl[ingt ein O]Tter.
 025 SM: [((lacht))]
 026 [((lacht leise))]
 027 [(1.2)]
 028 LS: [Hä (lass ma fragen.)]
 029 RL: [SAG ((unverständlich))]
 030 SM: <<lachend> (ha das fin_ich echt ganz
 gut.)>
 031 RL: <<vorlesend>EINrichtung abschließen.>
 032 (1.3)
 033 RL: °h Oke.
 034 nor
 dat WAR_s anscheinend;
 035 (0.6)
 036 RL: muSIK abspielen (.) kann ich ihm jetzt
 auch von hier sagen;
 037 LS: oder du SAGST es ihr.
 038 RL: °h ja dann äH.
 039 LF: oKAY google?
 040 GA: ((spielt plötzlich laut Musik, bis 054))
 041 RL: [hä::?]

042 LS: [oah.]
 043 SM: [((lacht))]
 044 GA: [((stoppt Musik))]
 045 SM: [((lacht))]
 046 RL: [wo kam DAS n jetzt her.]
 047 LS: [((lacht))]
 048 RL: ach so weil ich HIER auf mu[sik
 abspielen getippt hab.]
 049 LS: [em JA: warte
 ma (.) mach ma aus.]
 050 SM: [oKAY google.]
 051 GA: [HIER bin ich;]
 052 k: [((Musik wird wieder leise gespielt))]
 053 SM: LAUTstärke fünf.
 054 (1.5)
 055 GA: ((spielt Musik etwas lauter))
 056 LS: okay google STOP.
 057 GA: ((spielt Musik leiser))
 058 LS: wie macht der OTter?
 059 RL: [°h / h°]
 060 [(2.1)]
 061 [(0.8)]
 062 SM: HEY google?
 063 LS: ne (.) oKE google;
064 SM: wie macht der OTter;]
 065 RL: [geht BEIdes.]
 066 LS: echt?
 067 LS: bei MIR ging nur ok.
 068 [(1.5)]
 069 k: [((Uhr ticken im Hintergrund))]
070 RL: also da muss er Suchen.
 071 GA: entschuldige (.) das habe ich nicht
 verSTANDen,
 072 RL: hä du hast das doch grad hier als BEIspiel angegeben.
 073 LS: oKAY google;
 074 wie macht der OTter;
 075 (1.6)
 076 GA: dAs ist ein OTter.
 077 ((spielt Quieken eines Otters ab))
 078 SM: [((lacht leise))]

079 RL: [((lacht leise))]
 080 LS: [((lacht leise))]
 081 SM: [wär lustiger wenn_s IHre stimme wäre;]
 082 GA: [((spielt Quieken eines Otters ab))]
 083 SM: [((lacht))]
 084 LS: [<<lachend> JA ne?>]
 085 RL: JA.
 086 SM: [((lacht))]
 087 LS: [<<lachend> ich hab auch eigentlich damit
 geRECHnet dass es> ihr sIE_s macht.]
 088 RL: oKAY goog[le;]
 089 LS: [wie] SIri dann he he;
 090 RL: SELBSTzerstörung;
 091 SM: <<lachend> KLASsiker; >
 092 °hh
 093 (1.62)
 094 GA: laut wikiPEdia;
 095 GA: [die] selbstzerstörung ist ein system an
 bord eines [technischen GEgen]stande-
 096 SM: [ah;]
 097 [oKAY google;]
 098 (0.44)
 099 RL: ((schmatzt)) °h [also DAS dann-]
 100 SM: <<flüsternd> [kannst du auch FL]üstern;>
 101 (4.33)
 102 RL: ((schmatzt)) ne;;
 103 [((unverständlich, 1.16 Sek.))]
 104 SM: [na is noch nich so WEIT wie alex]a;=ne?
 105 RL: °h
 106 SM: °h
 107 RL: ne: (.) alexa is schlAUer (.) und
 SCHNELler auch;
 108 (0.23)
 109 RL: da siehste ma HIER;
 110 RL: warte den SIEHST du ja gar nich;
 111 ja DOCH, 0
 112 (0.29)
 113 RL: geh ma weg da hier vom KAb[el,]
 114 SM: [TSCH][ULdigung
 ((lacht))

115 RL: [°hh
] du
 sieh[st ja wenn de] hier da (.)
 ((schmatzt)) °h irgendwas FRAGST,
 116 LS [[(unverständlich)]]
 117 RL: der LädT relativ lange;
 118 (0.21)
 119 RL: so alexa is ziemlich SCHNELL;

Der Ausschnitt wurde gewählt, weil hier die ersten, gemeinsam über das VUI (sowie das grafische Interface bei Robin) in Haushalt 8 durchgeführten Aktionen mit dem Smart Speaker sichtbar werden. Dabei setzen die Teilnehmer*innen verschiedene Ausdrucksressourcen ein, um das VUI auf dem akustischen Kanal bedienen zu können und zugleich einen anderen Kommunikationsweg unter den Beteiligten aufrechtzuerhalten, was in dieser ersten Testphase von den Beteiligten auch anderer Haushalte relevant gemacht wird. Der Fokus soll dabei zunächst auf den Gesten liegen, die die Teilnehmenden parallel zur verbalen Bedienung des VUI produzieren, so z. B. die emblematische, zitierbare Geste *Daumen Hoch* (vgl. Kendon 1992) in Beispiel (81):

Beispiel (81): Wie macht der Otter III (Ausschnitt 001–004)³⁴³

001 RL: dein GOOgle assistant wird dir antworten
 wenn du [hey google-]
 002 GA: [hallo +i]ch
 bin dein GOOgle assistant.+
 sm: +blickt GA
 an-----+
 rl: +blickt GA
 an-----+
 003 GA: +ich bin HIER + um dir zu
 ShElfen; §
 sm: +blickt zu RL und dann zu GA+
 rl:
 §zeigt Daumen hoch§

³⁴³ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c403&wID=&textSize=300&contextSize=8.

004 GA: §in der goo[gle HOME a]pp
 +kannst du dich über meine
 funktionen informIERen;+ §
 sm: §lehnt sich nach vorne und wieder zurück,
 schiebt mit RH Brille hoch, fasst sich an
 Auge und richtet sich wieder auf§
 ls: +blickt SM und dann wieder
 Handy an-----+

Robin zeigt den Daumen hoch (003), als das VUI ihn unterbricht, während er gerade Anleitungen aus dem Handy vorliest (vgl. Abb. 32).



Abb. 32: Robin zeigt Daumen Hoch-Geste (Z. 003).

Die metaphorische Bedeutung des Daumens ‚nach oben‘ kann in einer positiven, affirmativen Weise verstanden werden (siehe auch McNeill 2014).³⁴⁴ Der Übergang in die gestische Prozessierung der Kommunikation erfolgt hier wohl nicht zufällig genau zu dem Zeitpunkt, als das VUI beginnt verbalsprachliche Laute zu produzieren: Der akustische Kanal ist nun ‚belegt‘ – würde die positive Bestätigung des Beginns des Ersteinrichtungsdialogs verbal ausgedrückt, würde sie genau diese überlagern. Die Teilnahme des VUIs macht also eine kurzfristige Um-

³⁴⁴ Huynh (2020: 87) schlägt „Das ist gut“ als verbale Paraphrase der Geste vor, wobei solche verbalsprachlichen Umschreibungen hier nicht übernommen werden, sondern eher als Orientierung dienen sollen, da dies nicht die von den Teilnehmer*innen geteilte Interpretation sein muss und im Sinne eines gesprächsanalytischen Vorgehens entsprechend auch nicht in die analytische Auswertung übergehen sollte.

gehung dieser Ressource notwendig; Samuel scheint jedenfalls vor dem Hintergrund seiner Blickrichtung auch kurzzeitig auf Robin orientiert zu sein, richtet den Blick dann jedoch zurück zum Smart Speaker.

Beispiel (82): Wie macht der Otter III (Ausschnitt 064–071)³⁴⁵

```

064 SM: wie macht der O[Tter;          ]
065 RL:                §[geht BEIdes.]----->
                >-----§
    rl:                §bewegt RH vor dem Gesicht
                vor und wieder zurück§
066 LS: +echt?        +
    ls: +blickt RL an+
067 LS: +bei MIR ging nur ok.+
    ls: +blickt GA an-----+
068    [(1.5)                ]
069 k:  [((Uhr-ticken im Hintergrund)) ]
070 RL: *also da muss er SÜchen.      *
    rl: *zeigt mit RH auf Lampe des GA*
071 GA: entschuldige (.) das habe ich nicht
                verSTANDen,

```

Durch Zeigegesten tritt noch ein anderer, in den bisherigen Ausführungen nur wenig beachteter Aspekt im Umgang mit dem Smart Speaker auf: dessen Materialität. Im Verlauf der Ersteinrichtung und der Tests treten teilweise zeitliche Abstände zwischen Stimmein- und ausgabe auf, die die Teilnehmenden als längere Wartezeit deuten (vgl. 070). Diese ‚längere‘ Wartezeit begründet Robin damit, dass der im Hintergrund ablaufende Prozess der Suche³⁴⁶ noch läuft, und verweist dabei auf eine visuelle Anzeige des Smart Speakers, der hier nicht nur in Form des VUIs, sondern auch mithilfe seiner Oberfläche Zeichen aussendet (siehe Abb. 33 und 34). Die vier Leuchten auf der textilarartigen Umhüllung des Lautsprechers leuchten in einer Reihe nacheinander auf und zeigen damit als „direct pro-

³⁴⁵ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c443&wID=&textSize=1000&contextSize=36.

³⁴⁶ Dieser von Robin als „suchen“ bezeichnete Prozess dürfte in etwa äquivalent zu dem sein, was im Rahmen dieser Arbeit als *Natural Language Understanding* und *Information Retrieval* bezeichnet wurde (vgl. Kap. 3.2.2), wobei die genauen technischen Details für die Teilnehmer*innen keine Rolle spielen (diese werden von Robin als „SÜchen“ (070) zusammengefasst).

gress indicators“ (Shneiderman/Plaisant 2005: 460–461; siehe auch Meyer et al. 1996) stilisiert in Anlehnung an einen Ladebalken an, dass noch etwas im Prozess befindlich und nicht abgeschlossen ist. Auf diese Leuchten verweist Robin mit seiner Zeigegeste, die prototypisch vom Körper ausgehend auf ein spezifisches Objekt deutet (vgl. Kita 2003: 1).

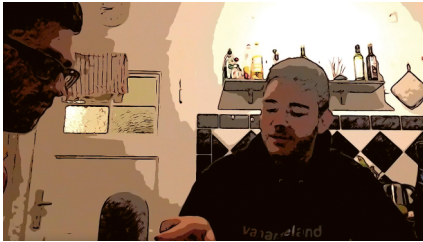


Abb. 33: Robin zeigt auf Leuchten des Google Nest (Perspektive 1).



Abb. 34: Robin zeigt auf Leuchten des Google Nest (Perspektive 2).

Die Zeigegeste wird verbalsprachlich begleitet („also da muss er Suchen“, 074), wobei der deiktische Ausdruck „da“ sich auf den vorherigen interaktionalen Verlauf („wie macht der OTter?“, 064, und die im Anschluss ausbleibende Reaktion des VUI) zu beziehen scheint, nicht auf den tatsächlichen ‚Ort‘ der Suche, der nicht mit den blinkenden Leuchten gleichzusetzen ist. Die deiktische Geste wird zudem von den anderen Beteiligten mit hervorgebracht, die sichtbar ihren Blick auf das „Zeigeziel“ (Stukenbrock 2009: 291) wenden, auch wenn insbesondere Lara – wie ich im weiteren Verlauf herausstellt – die Leuchten dennoch nicht im Blickfeld hatte.³⁴⁷ Die Geste verweist darauf, dass der Smart Speaker auch in seiner materiellen und visuellen Präsenz erfahrbar ist und auch die Oberfläche in das „Interfacing“ (Lipp/Dickel 2022) einbezogen wird. Die zusätzliche Explikation von Robin zeigt an, dass er nicht davon ausgeht, dass es sich dabei um unter den Beteiligten bekanntes Wissen handelt; es bleibt aber offen, ob er dabei nur die spezifische Suchanzeige im Speziellen meint oder die visuellen Anzeigen an der Oberfläche im Allgemeinen in den Fokus der Aufmerksamkeit rücken will – auch weil das VUI den Vorgang der deiktischen Geste durch akustische Signale unterbricht und so erneut als Partizipand an der Praxis auftritt. Der Umstand, dass das

³⁴⁷ Stukenbrock (2009: 290–292) unterscheidet zwischen Zeigeziel und Referent, wobei das lokalisierte Zeigeziel in den Bereich der perzeptiven Unschärfe, der Referent in den Bereich der kognitiv-semantischen Unschärfe falle. Im gegebenen Beispiel ist z. B. anhand des Zeigeziels nicht eindeutig auszumachen, ob Robin nur eine oder

alle Leuchten meint, ob er nur auf die Leuchten an sich oder auch deren Blinken referiert.

VUI allerdings für die visuellen *cues* der menschlichen Interaktionsbeteiligten ‚blind‘ ist und nicht erkennt, dass der Zeigevorgang noch nicht abgeschlossen ist, als es zu sprechen beginnt (und so die Regeln des Turn-Takings verletzt), verweist jedoch erneut darauf, dass der Status als ‚Gesprächsteilnehmer‘ brüchig ist.

Die Leuchten werden im weiteren Verlauf des Ausschnitts noch einmal Thema, wie sich im folgenden Auszug zeigt:

Beispiel (83): Wie macht der Otter III (Ausschnitt 100–114)³⁴⁸

```

100 SM:    <<flüsternd> [kannst du auch
          FL]üstern;>
101      (4.33)
102 RL:  ((schmatzt)) ne;;
103      [((unverständlich, 1.16 Sek.)) ]
104 SM:  [na is noch nich so WEIT wie alex]a;=ne?
105 RL:  °h
106 SM:  °h
107 RL:  +ne: (.) alexa is $schlAUer (.) und
          SCHNELler auch;+      $
          sm: +blickt RL an--+
          rl:      $wendet Kopf zu SM und
          führt Hand an Kinn-----$
108      (0.23)
109 RL:  Δda siehste ma HIER;Δ
          rl: Δgreift GA über LED-Δ
110 RL:  warte den SIEHST du ja gar nich;
111      ja DOCH,
112      (0.29)
113 RL:  Δgeh ma weg daΔ Δ$hier vom KAb[e]l, ] Δ$
          rl: Δhebt GA an---Δ Δbewegt GA in Mitte des
          Tisches-----Δ
          sm:      $weicht von Tisch zurück>
          >-----Δ$
          ls:      $weicht von Tisch zurück>
          >-----Δ$

```

³⁴⁸ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c511&wID=&textSize=400&contextSize=12&startTime=513.831.

114 SM:

[TSCH]

ULdigung ((lacht))

Nach dem Fehlschlag zur Flüster-Funktion (100–103) richtet sich das Gespräch auf einen Herstellervergleich. Zur Untermauerung seiner These, dass der Google Assistant länger laden muss als der Amazon Echo, auf den mit „Alexa“ referiert wird (104, 107), will Robin erneut die dynamische Fortschrittsanzeige einbeziehen und darauf verweisen. Er stellt nun jedoch fest, dass zumindest Samuel diese Anzeige und damit auch seinen Verweis nicht oder zumindest nicht gut sehen kann (vgl. 109–110; siehe auch Abb. 35).



Abb. 35: Standort des Smart Speakers vor Repositionierung.

In der Folge davon ergibt sich eine Spannung zur Positionierung und damit zur räumlichen Umgebung, in der die Ersteinrichtung stattfindet: Zur Herstellung einer gemeinsamen Sehfläche stellt Robin den Smart Speaker an einen anderen Platz (siehe Abb. 36), wobei sich kurz eine Unterbrechung ergibt, weil Samuel noch die Hand auf dem Kabel liegen hat, was für den Prozess hinderlich ist (113–114). Durch diese Umstellung verändert sich die Positionierung der Beteiligten zum Smart Speaker, die nun erst für Samuel und Lara die visuellen Anzeigen vollständig sichtbar werden lässt (siehe Abb. 36). Lara muss dafür ihren Kopf spezifisch neigen (siehe Abb. 37). Das Mobiliar, der Tisch und die Stühle und die Positionierung der Menschen sowie des Smart Speakers auf diesen konstituieren also nun eine leicht veränderte sozialräumliche Situation. So können Samuel und Lara den Smart Speaker nun auch mit erweiterten Ausdrucksressourcen als ‚beteiligen‘, da seine Signale wahrnehmbar gemacht wurden.

Wie sich diese räumliche Re-Positionierung vollzieht und wie auch Robin im Anschluss erneut mit der deiktischen Zeigegeste auf Bestandteile der Oberfläche

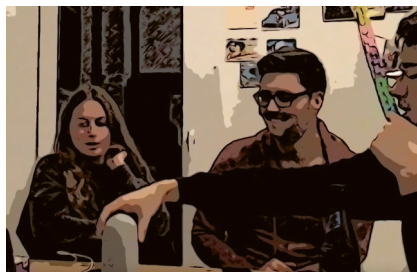


Abb. 36: Robin stellt den Smart Speaker in die Mitte des Tisches.

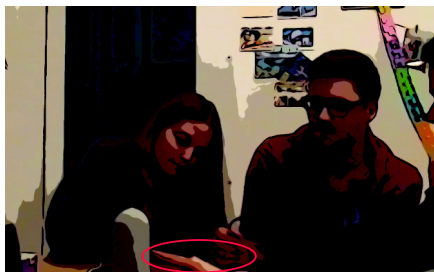


Abb. 37: Nach der Repositionierung zeigt Robin erneut auf die Leuchten.

des Geräts verweist, zeigt jedoch deutlich, dass der Smart Speaker als *Objekt* behandelt wird: Auch wenn Robin das Beteiligungspotenzial des Smart Speakers dadurch erhöht, ist das Gerät an seiner eigenen Re-Positionierung nur in materieller Weise beteiligt. Der Smart Speaker ermöglicht die Umstellung durch seine Größe und sein Gewicht, aber er lädt nicht aktiv dazu ein. Er ist also, wie hier illustriert werden konnte, mit seiner Materialität an der Verfertigung der sozialen Praxis und somit auch an der räumlichen Konfiguration, aber ohne eigene *Aktivität* (im Sinne Hirschauers 2016: 49) beteiligt. Für die Objekthaftigkeit spricht auch, dass die Anordnung der Personen um den Smart Speaker herum arrangiert ist, die das Gerät auf den Tisch in der Mitte gestellt haben, sodass sich eine geteilte Sehfläche ergibt. Zugleich schreiben die Beteiligten dem Smart Speaker keine Form von körperlicher Mit- oder Selbstbestimmung zu, wie sich an der Zeigepraktik sowie der Verschiebung illustrieren ließ.

Die Limitierungen des VUI im Hinblick auf die Beteiligung an visuellen Ausdrucksressourcen im Gespräch wurden bereits deutlich. Sie stehen im Einklang mit den Befunden von Suchman (2007) zu ressourcenlimitierter Kommunikation. Diese Limitierungen werden, wie bereits die emblematische Daumen Hoch-Geste von Robin illustriert hat, strategisch eingesetzt, um die Interaktion *über* den Smart Speaker im Dialog *mit* dem VUI fortzusetzen. Ähnliche strategische Einsätze finden sich auch bei einer Betrachtung des Blicks als Ausdrucksressource, den die Teilnehmer*innen fortlaufend einsetzen, wie der folgende Auszug unterstreicht.

Beispiel (84): Wie macht der Otter III (Ausschnitte 023–028 und 065–079)³⁴⁹

024 RL: <<liest aus der Anleitung vor>erZÄHL mir
etwas lustiges.
025 wie kl[ingt ein +Otter.>----->
>-----+]
rl: +schaut kurz seine
Mitbewohner an +
k: ((...))
026 SM: [((lacht))]
027 [((lacht leise))]
028 LS: [Hä (lass ma fragen.)]

Ausgelassen bis Z. 62

062 SM: HEY google?
063 LS: +ne (.) oKE google;+
ls: +blickt SM an-----+
064 SM: **wie macht der O[Tter;]**
065 RL: §[geht BEIdes.]----->
>-----§
rl: §bewegt RH vor dem Gesicht
vor und wieder zurück§
066 LS: +echt? +
ls: +blickt RL an+
067 LS: +bei MIR ging nur ok.+
ls: +blickt GA an-----+
068 [(1.5)]
069 k: [((Uhr ticken im Hintergrund))]
070 RL: ***also da muss er SUCHen. ***
rl: ***zeigt mit RH auf Lampe des GA***
071 GA: entschuldige (.) das habe ich nicht
verSTANDen,
072 RL: *hä du * §hast das doch
grad hier als BEIspiel angegeben.§
rl: ***bewegt RH Richtung Handy* §beugt sich**
über Handy-----§

³⁴⁹ Diese Ausschnitte können über die folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayvideotranscript&id=FOLK_E00484_SE_01_T_01.

073 LS: oKAY google;
 074 wie macht der OTter;
 075 (1.6)
 076 GA: dAs ist ein OTter.

Der Blick wird etwa von Robin eingesetzt, um die besondere Relevanz einer vorgelesenen Stelle aus den Einrichtungsinstruktionen zu markieren (vgl. 025), und tritt hier eindeutig in Kombination mit prosodischen Eigenschaften der verbalsprachlichen Äußerungen auf, die sich in der Akzentuierung von „OTter“ (025) manifestieren. Die anschließende Äußerung von Lara (028) bestätigt diese Relevanzsetzung, die dann deutlich später im weiteren Verlauf noch einmal aufgegriffen wird und als eine der ersten Stimmeingaben testweise von Lara an das VUI adressiert und bei Scheitern mit mehreren Versuchen wiederholt wird (vgl. 058, 064, 074), bis es funktioniert. Anhand dieser Test-Stimmeingabe zeigt sich auch, dass die Überlagerung von VUI-Dialog und zwischenmenschlicher Koordination zum Problem werden kann, zu dessen Lösung zwar Blickkontakt eingesetzt wird, was aber nicht zuverlässig zum Erfolg führt: Nachdem Samuel den ersten Teil der Invokationssequenz geäußert hat (062), ist davon auszugehen, dass das VUI bereits im Listening-Modus ist. Lara adressiert zwar durch den Blick ihren Korrekturversuch an Samuel („ne (.) oKE google;“, 063) und wendet sich auch Robin zu, als dieser antwortet („geht BEIdes.“, 065). Anschließend schließt Lara die Sequenz, die parallel zur VUI-Bedienung stattfindet, ab und unterstreicht dies durch die Hinwendung des Blicks zurück auf den Smart Speaker, der hier erneut in seiner Materialität als Kontextualisierungsressource für die anderen Interaktionsteilnehmer*innen verstanden werden kann. Das VUI im Gegenzug aber ist in Bezug auf dieses kontextualisierende Blickverhalten limitiert. Es ist davon auszugehen, dass die gefallenen Äußerungen mit in die zu verarbeitenden akustischen Signale eingeflossen sind und somit entweder auf Ebene des Natural Language Processing oder des Information Retrieval zu Verarbeitungsschwierigkeiten geführt haben. Diese Koordinationsschwierigkeiten führen hier zu der Fehlermeldung (071) und der erneuten Produktion der Stimmeingabe durch Lara (074).

Ebenso limitiert ist das VUI bei der Erkennung der mimischen Reaktionen auf den Fehlschlag bei der Musikwiedergabe (vgl. 041–042), wie anhand eines letzten Auszugs besprochen werden soll:

Beispiel (85): Wie macht der Otter III (Ausschnitt 040–046)³⁵⁰

040 GA: \$((spielt plötzlich laut Musik, bis 054))\$
 sm: \$schreckt zurück-----\$
 ls: \$schreckt zurück-----\$
 041 RL: [\$hä::?] \$
 rl: \$führt Arme zurück auf Tisch und lehnt
 sich nach vorne\$
 042 LS: [oah.]
 043 SM: [\$((lacht))] \$
 rl: \$beugt sich über Handy\$
 044 GA: [((stoppt Musik))]
 045 SM: [((lacht))]
 046 RL: [wo kam DAS n jetzt her.]

Durch die Steuerung des Smart Speakers über das Smartphone von Robin kommt es zu einer für die anderen gänzlich unerwarteten und für Robin mindestens unerwartet lauten Wiedergabe von Musik. Lara und Samuel drücken durch Körperhaltung und Mimik eine größtmögliche Distanz zum Smart Speaker, dem Ursprung der lauten Geräusche, aus (siehe Abb. 38) und zeigen somit auch an, dass sie sich erschrocken haben, was sie im weiteren Verlauf auch verbalisieren (041–046). Diese Lautstärke der Musikwiedergabe wird anschließend von den Teilnehmer*innen reduziert. Auch hier ist das VUI nicht darauf ausgerichtet, die möglichen Reaktionen seitens der Teilnehmer*innen zu antizipieren oder wiederum auf diese zu reagieren.

Insgesamt zeigt sich in dem Gesamtausschnitt von nur ca. zwei Minuten (siehe Beispiel 80 für eine Übersicht) die immense Relevanz von Multimodalität in der Praxis: Gestik, Mimik, Blick, die Materialität der Umgebung, Körperhaltung und Körperpositionierung im Raum (sowie der gestaltete Raum selbst, auch wenn dies im gegebenen Beispiel nicht hervortrat) bringen die Praxis (mit) hervor und können daran entscheidende Anteile haben: Die Kategorien Gestik, Mimik und Blick sind einerseits zentral für die Verfertigung der Praxis, wenn auf dem akustischen Kanal bereits das VUI Zeichen aussendet, andererseits sendet auch der Smart Speaker selbst visuelle Zeichen aus und macht Gebrauch von seiner Materialität. Die körperlichen Orientierungen auf das Gerät zeigen dessen Beteiligung an der sozialen Praxis – gerade

³⁵⁰ Dieser Ausschnitt kann über folgende URL in der DGD abgerufen werden: https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscriptVideo&id=FOLK_E_00484_SE_01_T_01_DF_01&cID=c450&wID=&textSize=300&contextSize=8&startTime=447.832.



Abb. 38: Der Smart Speaker spielt laute Musik ab, Lara und Samuel schrecken zurück.

dann, wenn wie im letzten Ausschnitt ein ‚Eigensinn‘ des VUIs zu bestehen scheint und Unvorhergesehenes passiert. Zugleich werden Abgrenzungen zu menschlichen Gesprächsteilnehmer*innen durch die Objektivität des Geräts deutlich.

6.2.4 Zwischenfazit zu VUI-Dialogen in Mehrparteienkonstellationen

Die Analysen zeigen, dass Einbindungen von VUIs in laufende Unterhaltungen nicht immer begründungspflichtig sind, sondern VUI-Dialoge auch uneingeleitet und sogar überkreuzt mit dem *turn-by-turn talk* auftreten können. Stärkere Einbindungen waren in den Analysen ein Hinweis darauf, dass das VUI Teil der sozialen Praxis wurde. Es konnten einige Situationstypen gezeigt werden, in denen dies geschieht (neben der Ersteinrichtung auch Situationen des Übens, Testens, Vorführens, Evaluierens und der Umgang mit Störungen). Diese Situationen wurden einer genaueren Betrachtung unterzogen und es konnten sprachliche Praktiken herausgearbeitet werden, mit denen Nutzer*innen auf der formalen Oberfläche die VUI-Dialoge in den *turn-by-turn talk* einbinden als *wären* an das VUI gerichtete Äußerungen Gesprächsbeiträge an eine*n Teilnehmer*in und als *wären* Äußerungen des VUIs Gesprächsbeiträge. Allerdings konnte auch herausgearbeitet werden, dass es sich dabei um Äußerungen handelt, die zwar an der Oberfläche eine solche Beteiligung am Gespräch herstellen, aber zugleich andere kommunikative Funktionen in der Interaktion mit den menschlichen Interaktionsteilnehmer*innen erfüllen, z. B. die Performanz des VUIs zu bewerten oder Frust über das Scheitern von VUI-Dialogen auszudrücken. Die zur Verfügung stehenden Dialogoptionen und die Stimmausgaben des VUIs werden als Ressourcen für die soziale Praxis erkannt und

von den Teilnehmenden für andere Zwecke weiterverwendet. Das trifft genau die bei Goodwin (2018: 6) als „decomposition and reuse with transformation“ beschriebene Eigenschaft von ko-operativer Verfertigung, die eben auch VUI-Dialoge einbezieht und nutzt.

Das VUI ist also „Partizipand“ (Hirschauer 2004) an der Praxis, und kann durch seine dialogische Verfasstheit Material für dialogische Folgeäußerungen in sozialen Interaktionssituationen liefern. Die Beteiligung ist aber eben materiell – ein VUI ist dadurch kein*e Gesprächsteilnehmer*in, es wird vielmehr ein solcher „dargestellt“, was auch zu entsprechenden Äußerungen der Nutzer*innen führen kann (vgl. Clark/Fischer 2022). Das zeigt sich u. a. an schnellen Wechseln von der Objekt- zur Hörerdeixis: VUIs werden in Sekunden von Angeredeten zu Redegegenständen, es ist nur in Ansätzen „Face-Work“ erkennbar. Ein solches Sprechen ist mit bzw. vor Haustieren beobachtet worden. Dort beschreiben Torres Cajo (2016: 66) sowie Torres Cajo/Bahlo (2016: 88) auf Basis von „Gassigesprächen“ ebenfalls mehrfachadressierte Äußerungen, die sich formal an Haustiere (hier: Hunde) richten, zugleich aber als „Face-Work“ (Goffman 1955) und kommunikative Funktionalisierung in der sozialen Praxis der menschlichen Beteiligten verstanden werden können. Diesen formal-funktionalen ‚Split‘ in der Mensch-Haustier-Interaktion stellt auch Tannen (2004: 417) fest und konzeptualisiert in diesem Sinne Hunde als Ressource zur Lösung diverser kommunikativer Aufgaben. In den von Mondémé (2018: 85–86) gesprächsanalytisch untersuchten Daten zur Interspezies-Interaktion sind – wie in den obenstehenden Beispielen – schnelle Wechsel zwischen direkter Adressierung der Tiere (Hörerdeixis) und der Bezugnahme auf diese mit der dritten Person beobachtbar. Zudem sind zahlreiche Vergleiche zwischen Haustier-Interaktionen und *Baby-Talk* angestellt worden (vgl. Mitchell 2001; Tannen 2004: 400; Mondémé 2018). Muster (2013: 168) stellt in ihrer kommunikationssoziologischen Untersuchung zu Haustieren als „narrative Ressource“ fest: „In diesem Zusammenhang ist es wichtig zu erwähnen, dass es [...] üblich ist, in Anwesenheit von Tieren über diese zu sprechen. Ebenso wie bei Babys und Kleinkindern gilt dies nicht als unhöflich“. Zwar kann das Sprechen über Anwesende – anders als von Muster (2013: 168) behauptet – nicht generell als unhöflich oder herabsetzend verstanden werden (vgl. Schegloff 1996: 445–447). Entscheidend ist vielmehr, dass es sich bei den adressierten bzw. referenzierten Einheiten – Tieren, Kleinkindern und hier: Maschinen – zwar um grundsätzlich dialogfähige, aber nicht voll sozialfähige Dialogpartner handelt, die ihre Rolle als ratifizierte Teilnehmer*innen nicht reziprok erfüllen können, auch wenn sie von anderen Beteiligten phasenweise so behandelt werden, was die schnellen und unmarkierten Wechsel von Objekt- zu Hörerdeixis erklärt.

Diese Merkmale treten in dyadischen Konstellationen zwischen Anwender*in und VUI nicht in der Deutlichkeit hervor, in der sie sichtbar werden, wenn Mehrparteieninteraktionen im Blickfeld stehen. Erst dann wird offensichtlich, dass das

VUI – selbst wenn dieses selbst zum Redegegenstand wird – sich daran nicht gleichermaßen beteiligen ‚kann‘. Sprachliche Praktiken zur Realisierung einer Beteiligung der VUIs an der Praxis der Nutzer*innen auf der sprachlichen Ebene erfüllen andere Funktionen im interaktionalen Gesamtgefüge. Sie sind Teil des VUI-Dialogs und Teil der sozialen Interaktion und insofern lassen auch sie sich als Interface-Praktik, als „Interfacing“ (Lipp/Dickel 2022) verstehen – sie haben eine ebenso verbindende wie trennende Wirkung.

Mit Blick auf die abschließend noch einmal konzentriert betrachtete Multimodalität in Interaktionssituationen, an denen ein Smart Speaker beteiligt ist, lässt sich auf Grundlage der hier analysierten Daten zusammenfassen, dass die Multimodalität von den Beteiligten genutzt werden kann, um ‚Ausweichbewegungen‘ vom akustischen zum visuellen Kanal zu prozessieren. Zudem tritt der Smart Speaker auch hier in seiner materiellen Erscheinungsform als Beteiligter der Praxis auf und wirkt in den Ersteinrichtungssituationen prägend für die räumlichen Konfigurationen. Allerdings werden durch die Materialität Kategorien zur Beschreibung der Teilnehmenden-Rolle unschärfer und bestätigen die schon bisher gefundenen Brüche in der Behandlung von Smart Speakern als Gesprächsbeteiligten.

7 Schlussbetrachtung

Abschließend sollen in dieser dreiteiligen Schlussbetrachtung erstens die Ergebnisse noch einmal entlang der beiden Analyseteile zusammengefasst werden (Kap. 7.1). Darauf aufbauend folgt zweitens eine ausführlichere Diskussion der Ergebnisse, die sowohl konzeptionelle Fragen bearbeitet als auch methodologische Reflexionsarbeit leistet und daraus hervorgehende Anschlussfragen extrapoliert (Kap. 7.2). In einem Ausblick soll drittens von den Gegenständen der Arbeit ausgehend ein kurzer Blick auf rezente technologische Entwicklungen und Prognosen geworfen werden – und somit auf Phänomene, zu deren Erforschung erneut eine interdisziplinär arbeitende, empirische Sprachwissenschaft wichtige Beiträge leisten kann und sollte (Kap. 7.3).

7.1 Zusammenfassung der Ergebnisse

Die Arbeit fragte nach den sprachlichen Praktiken im Prozess der Domestizierung von Medien mit Voice-User-Interfaces und stellte dabei insbesondere auf die Einrichtung und Nutzung von Smart Speakern ab. Zentrale Erkenntnisinteressen der Arbeit waren, (1) welche sprachlichen Praktiken sich bei der Domestizierung stationärer Sprachassistenzsysteme mit VUIs zeigen, (2) inwieweit diese – zwischen Transformation und Emergenz – als Abwandlungen bereits untersuchter sprachlicher Praktiken beschrieben werden und inwieweit sie als emergente sprachliche Praktiken gelten können, sowie (3) wie Smart Speaker sprachlich zu Beteiligten an der sozialen Praxis werden. Dazu wurden im ersten Teil der Analyse ausschließlich dyadische Dialoge mit VUIs untersucht, um sprachliche Praktiken in der Bedienung von VUI-Geräten ausmachen zu können. Zu diesem Zweck wurde eine Kollektion von dyadischen VUI-Dialogen herangezogen.³⁵¹ Im zweiten Teil wurden dann auch Mehrparteienkonstellationen in den Blick genommen, um die Domestizierung und die Einbindung in die soziale Praxis *in situ* nachvollziehen zu können. Die Ergebnisse aus dem ersten Analyseteil konnten dabei nutzbar gemacht werden.

³⁵¹ Wie bereits weiter oben reflektiert, kann für die ausgewählten Aufnahmen nicht sicher ausgeschlossen werden, dass neben VUI und Anwender*in noch andere Beteiligte während des VUI-Dialogs anwesend sind, was auch Auswirkungen auf die Gestaltung des VUI-Dialogs haben kann. Andere Beteiligte sind aber auf den ausgewählten Aufnahmen nicht zu hören und werden von den Sprecher*innen nicht relevant gesetzt.

7.1.1 Dyadische VUI-Dialoge

Die Analysen zu dyadischen VUI-Dialogen waren entlang von vier konversationsanalytischen Grundbegriffen strukturiert: Anredeformen, Sequenzialität, Turn-Taking und Reparaturen. Die Auswahl dieser Begriffe erfolgte aus den Daten heraus und auf Basis der von Schegloff (2006: 71) skizzierten „generic organizations of practice“, die als praktische Lösungen überall dort aufzufinden sind, wo stabile Interaktionssituationen entstehen. Schegloff (2006: 71) entwickelt insgesamt sechs solcher Problembereiche, von denen auf Grundlage der Daten drei als besonders relevante Untersuchungsfelder ausgewählt wurden: die sequenzielle Organisation, das Turn-Taking sowie Reparaturen. Schegloff bezieht sich dabei auf zwischenmenschliche Interaktionen, für die jeder der genannten Problembereiche gesprächslinguistisch umfangreich beschrieben ist. Auch die sprachwissenschaftliche Literatur zum Austausch zwischen Menschen und Maschinen nimmt immer wieder auf diese Bereiche Bezug. Entsprechend eigneten sie sich nicht nur, um die Analyse sprachlicher Praktiken in VUI-Dialogen zu ordnen, sondern auch, um zu prüfen, wie die sprachlichen Praktiken sich als Abwandlungen bereits beschriebener sowie als emergente Praktiken darstellen.

Den Bereich der *Anredeformen* nennt Schegloff (2006) nicht als eine der „generic organizations“. Dieses Analysefeld ergab sich aber als relevante Kategorie sowohl aus den Daten als auch aus der konversationsanalytischen Literatur heraus: Adressierungen, insbesondere in Form onymischer Anreden, traten als wiederkehrende kommunikative Aufgabe auf, die im Fall der VUI-Dialoge auch vor dem Hintergrund des Mithörens der Geräte in den privaten Wohnumgebungen von besonderer Relevanz ist (vgl. Waldecker/Volmar 2022): Um auf ein Aktivierungswort hin reagieren zu können, muss das Mithören dauerhaft erfolgen und das Gerät die empfangenen akustische Signale daraufhin prüfen, ob das Aktivierungswort genannt wird. Durch die Invokation, d. h. die Nennung und Erkennung eines gerätespezifischen Aktivierungsworts, verändert sich der Status des Geräts: Es verarbeitet dann alle folgenden akustischen Signale als sprachliche Eingaben. Dieser Modus muss nach jeder in sich geschlossenen sprachlichen Eingabe erneut hergestellt werden. Entsprechend lassen sich zur Invokation der Geräte sprachliche Praktiken der Adressierung finden, die zwar einerseits mit der Veränderung des Status einer Maschine vom Ruhe- in den Eingabemodus eine neue kommunikative Funktionalisierung aufweisen. Andererseits werden Invokationen praktisch als Summons-Answer-Sequenz verfertigt.

Die sprachlichen Praktiken, mit denen Summons-Answer-Sequenzen realisiert werden, haben konsequenterweise Ähnlichkeiten mit sprachlichen Praktiken aus zwischenmenschlichen Interaktionssituationen: Onymische Formen, ggf. in Verbindung mit einer Gesprächspartikel, sind als Summons auch in zwischen-

menschlichen Summons-Answer-Sequenzen beobachtbar, ebenso wie nonverbal produzierte Antworten (im Falle der VUIs das Aufleuchten entsprechender Elemente am Gerät, im Falle zwischenmenschlicher Interaktionen etwa eine körperliche Drehung oder ein Blick). Hierbei handelt es sich also um eine Abwandlung sprachlicher Praktiken aus einem bereits bekannten Repertoire. Als emergente sprachliche Praktik lässt sich die Verwendung der (allerdings ohnehin semantisch sehr weiten) Gesprächspartikel *okay* bei einer der beiden möglichen Invokationen für das Gerät von Google verstehen. Es lassen sich darüber hinaus nicht nur onymische Anreden im Rahmen von Summons-Answer-Sequenzen feststellen, sondern auch Anreden seitens des VUIs, die ebenfalls keine sprachliche Emergenz erkennen lassen. Vielmehr können die Verwendung onymischer Adressierungen und Du-Anreden durch die VUIs, in die im Ersteinrichtungsdialog auch die Nutzer*innen eingelernt werden, teilweise als sprachliche Verfahren verstanden werden, die Ausdruck eines auch seitens der Gerätehersteller vorangetriebenen Domestizierungsprozesses sind.

Auch im Hinblick auf die anderen drei untersuchten Bereiche sollen die Ergebnisse kurz zusammengefasst werden. Für das Untersuchungsfeld der *Sequenzialität* und der *Sequenzorganisation* lassen sich erstens Befunde bestätigen, die etwa von Krummheuer (2008; 2010), Gehle et al. (2015), Opfermann/Pitsch (2017), Pitsch (2020) bereits an anderen Geräten herausgearbeitet werden konnten: Der Austausch ist von einer Trägheit gekennzeichnet, die sich aus der Sequenzorganisation ergibt. Um damit verbundene sprachliche Praktiken genauer zu beleuchten, wurde das Teilkorpus der dyadischen Dialoge auf Muster im Ablauf von VUI-Dialogen hin untersucht. Als stabile Einheiten erweisen sich zwei Varianten eines Ablaufmusters, das als Basis-Sequenzstruktur für VUI-Dialoge beschrieben wurde, die nutzer*innenseitig initiiert werden:³⁵²

- (A) Invokation – (Listening-Modus –) Stimmeingabe – (Stimmausgabe/Scharnier –) praktische Umsetzung.
- (B) Invokation – (Listening-Modus –) Stimmeingabe – Stimmausgabe.

Diese Ablaufmuster können als emergente sprachliche Praktik verstanden werden, die nicht nur die Nutzer*innen immer wieder vollziehen, sondern an der sich – wie sich an den Dialogen zur Ersteinrichtung zeigen ließ – auch die Anbieter der Systeme orientieren. Dies hat einen zirkulären, selbstverstärkenden Ef-

³⁵² Im gesamten Korpus (auch über die Kollektion der dyadischen VUI-Dialoge hinaus) ist – abgesehen von den Ersteinrichtungsdialogen, die einen Spezialfall darstellen, – ein einziger Fall eines VUI-seitig initiierten VUI-Dialogs dokumentiert. Dies könnte bei anderen Nutzer*innen auch anders gelagert sein, es wird jedoch aufgrund des geringen Vorkommens in meinen Daten auf die Formulierung einer Sequenzstruktur für diesen Typ VUI-Dialog verzichtet.

fekt: Im Design der Anwendungen wird mit solchen praktischen Verfestigungen gearbeitet, die sich durch die Gestaltung der Dialogsysteme wiederum weiter verstetigen können. Wie sich in den Daten zeigt, ist eine sequenzielle Kohärenz, die über dieses Ablaufmuster hinausreicht, nicht nachweisbar. Zwar ergaben sich an einigen Stellen Optionen für Einschübe (v. a. zur Durchführung von Reparaturen) und systemseitige Post-Expansionen, wobei für Letztere gezeigt werden konnte, dass sie Formen phatischer Kommunikation mit dem Zweck der Kund*innenbindung dienen können. Sequenzielle Verknüpfungen mehrerer Abläufe zur Kombination verschiedener, aufeinanderfolgender Stimmeingaben miteinander waren hingegen in dyadischen Konstellationen nicht nachweisbar. Für die sprachliche Gestaltung des Austauschs hat dies zur Folge, dass miteinander verknüpfte Anliegen, die mehrere Züge erfordern, in mehrere Teile zerfallen, die alle einzeln einer der beiden Varianten des Ablaufmusters entsprechen. VUIs können also Eingabe und Ausgabe als Paarsequenzen prozessieren; mit Sequenzialität, die jenseits von Sequenzen etwa durch konversationelle Implikaturen entsteht, sind die VUIs jedoch überfordert, wie auch der Abschnitt zu Reparaturen zeigen kann (siehe auch Habscheid 2022: 191). Entsprechend ist die Sequenzialität mit Krummheuer (2010: 229) als „aufgebrochen“ gut beschrieben: Sie ist auf die Durchführung einzelner Eingabe-Ausgabe-Prozesse hin ausgerichtet. Die Nutzer*innen passen sich dieser Präkonfiguration des Austauschs mit dem Interface einerseits an, andererseits entstehen gerade durch diese Eigenschaft der VUIs notwendige Reparaturen und ggf. Abbrüche; ferner testen Nutzer*innen – wie in den Mehrparteienkonstellationen noch deutlicher hervortritt – die Grenzen dieser Limitierung.

In der Gestaltung der einzelnen Turns bzw. *turn-constructive units* (TCUs) ist die Orientierung an der praktischen Verfertigung einer mit dem Sprechen vollzogenen Aktivität ebenfalls sichtbar, und zwar dann besonders deutlich, wenn die an das VUI gerichtete Stimmeingabe aus einem einzelnen freistehenden Lexem (z. B. „Licht“) oder kurzen, syntaktisch offenen Einheiten (z. B. „Timer löschen“) besteht. Nachweisbar sind darüber hinaus aber auch deontische Infinitive, Imperativ- und Fragesätze; nur außerhalb des Kontexts der dyadischen Nutzungssituationen zeigten sich sprachliche Praktiken außerhalb dieser Typologisierung.³⁵³ Zusätzlich zeigt sich eine Tendenz zur Kürze in der Gestaltung der Stimmeingaben. Damit sind keine emergenten sprachlichen Praktiken beschrieben, sondern es zeigt sich vielmehr, dass sprachliche Praktiken in den Dienst anderer Praktiken gestellt werden, u. a. der Steuerung von Smart Home-Elementen oder Erinnerungsfunktionen. Die

³⁵³ Dies wiederum kann damit erklärt werden, dass im Rahmen der Ersteinrichtung und kurz danach das VUI zunächst ausgetestet, teilweise gezielt überfordert und an seine Grenzen geführt wird. Dabei kommt es zu einer höheren sprachlichen Variabilität als in der über die CVR-Aufnahmen dokumentierten routinisierten Nutzung.

sprachlichen Formen ähneln oder gleichen gar solchen, die auch in anderen Kontexten beobachtbar sind, in denen sprachliche Praktiken empirischer Bestandteil von sozialen Praktiken sind, die kein Gespräch konstituieren (vgl. Goffman 1979: 14). Die knappe Gestaltung der Äußerungen muss nicht nur aus sprachökonomischen Gründen erfolgen, sondern es kann eine Kombination verschiedener Motive dahinterstecken; Baldauf (2002: 124) unterscheidet fünf: neben Sprachökonomie auch Rücksichtnahme auf andere Tätigkeiten, eine Sprecher*innenzentrierung in der Äußerung, gestalterische Absichten und beziehungsbezogene Motive. Im Hinblick auf die VUI-Dialoge wären als Motive für knappes Sprechen zu ergänzen: die Reduzierung der Fehleranfälligkeit (sowohl in der Artikulation der Spracheingabe als auch im Verständnis durch das VUI) und die Konzentration auf den durchzuführenden Prozess(schritt) sowie – verwandt mit dem von Baldauf beschriebenen Beziehungsmotiv – die Aufhebung der Notwendigkeit von „Face-Work“ (Goffman 1955) zwischen den Beteiligten.

Das Turn-Taking wird durch diese Art der Gestaltung von TCUs erleichtert, was reziprok auch zu einer Gestaltung der Turns in der entsprechenden Weise führt: Nutzer*innen orientieren sich an den notwendigen Parametern für eine Verarbeitung der Stimmeingaben, wodurch die Erfolgsquote (auch in der Rederechtsverteilung) steigt und die Nutzer*innen wiederum eher dazu tendieren, die erfolgreiche Turn-Gestaltung beizubehalten. Der Sprecher*innenwechsel fügt sich in die durch das Reden vollzogenen Handlungsschritte ein und scheint sich außerdem durch eine höhere Latenztoleranz seitens der Nutzer*innen auszuzeichnen, während die Latenztoleranz bzw. -flexibilität seitens der maschinellen Partizipanten geringer ist. Ob neben pragmatischen und segmentierenden Verfahren zur Anzeige von *transition-relevance places* (TRPs) auch andere *cues* produktiv sind (syntaktische Geschlossenheit, Intonationskontur o. Ä.), konnte im Rahmen der Untersuchung nicht gezeigt werden; die sprachlichen Praktiken der Nutzer*innen ließen hier ebenso wenig Rückschlüsse zu wie die Stimmausgaben der VUIs.

Eine andere sprachliche Praktik war in diesem Zusammenhang jedoch auffällig, worauf auch ihr Status in der Literatur hindeutet: Die 63 Aufnahmen umfassende Kollektion der dyadischen VUI-Dialoge dokumentiert insgesamt 133 VUI-Dialoge und darin acht Fälle von *barge-ins*, d. h. aktives Unterbrechen der Äußerungen des VUIs durch die Produktion einer neuen Invokation und Stimmeingabe. Die Notwendigkeit einer Aushandlung der Rederechtsverteilung, zu der es nach Schegloff (2000b) bei Überlappungen in zwischenmenschlichen Interaktionen kommt, ist hier suspendiert: Das VUI stoppt immer die Äußerungsproduktion bei Erkennung einer neuen Invokation und die neue Stimmeingabe überschreibt das vorherige kommunikative Projekt – ohne dass die Nutzer*innen dafür einen *account* produzieren. Diese sprachliche Praktik kann als Ausdruck einer Hierarchie zwischen Nutzer*innen

und VUIs verstanden werden, die auch in der Gestaltung der Stimmeingaben teilweise deutlich wird – etwa in der Eingabe „Stop“, die in allen untersuchten Haushalten regelmäßig auftritt und zur Beendigung der Äußerungsproduktion oder der aktuell durch das VUI abgespielten Musik sowie anderer laufender VUI-Aktivitäten dient und die widerspruchslös umgesetzt wird. Aushandlungen des Rederechts sind insofern teilweise zugunsten der Nutzer*innen aufgehoben (siehe auch Waldecker/Hector/Hoffmann 2024).

Im Kapitel zu *Reparaturen* konnte schließlich noch einmal die Trägheit des Austauschs genauer untersucht werden. Dazu wurden zunächst verschiedene Fälle unterschieden, sowohl danach, wer die reparaturbedürftige Stelle produziert (Nutzer*innen oder VUI), als auch danach, wer die Reparatur initiiert oder durchführt (angelehnt an konversationsanalytische Literatur, insbesondere Egbert 2009). Es zeigte sich, dass hyperlokal operierende Reparaturmechanismen aus der zwischenmenschlichen Interaktion, in denen z. B. einzelne Laute oder Lexeme in Sekunden ersetzt werden, nicht bzw. nur sehr eingeschränkt angewendet werden; eine Ausnahme stellten selbstinitiierte Selbstreparaturen der Nutzer*innen dar. Abgesehen davon wurden von den Nutzer*innen Reparaturen immer durch die Wiederholung einer gesamten Stimmeingabe realisiert, selbst dann, wenn nur ein einzelnes Lexem darin reparaturbedürftig war (dies wurde z. B. prosodisch oder syntaktisch als Reparandum markiert). Dies führt zu häufigen Wiederholungen des gesamten Basis-Sequenzablaufs, einschließlich der Invokation – eine Folge der maschinenseitig nicht bestehenden Kohärenz zwischen den einzelnen Sequenzen. Dieses Verfahren hat zudem zur Folge, dass stellenweise unklar bleibt, was als Reparandum zählt: die Stimmeingabe der Nutzer*innen oder die Stimmausgabe der VUIs. Dies ist ein Hinweis darauf, dass Reparaturen auch deswegen interessant sind, weil sie die sequenzielle Ordnung zwischen den beiden ‚Sprecher*innen‘ hinterfragen und die Ungleichheiten zwischen Nutzer*innen und VUI als ‚Beteiligte‘ an der Kommunikation deutlich werden lassen. Dazu trägt auch bei, dass die Reparaturen nicht immer erfolgreich sind: Es zeigen sich mehrere Fälle, in denen das kommunikative Projekt nach wiederholten Reparaturversuchen nutzer*innenseitig abgebrochen wird.

7.1.2 VUI-Dialoge in Mehrparteienkonstellationen

Der zweite Analyseteil widmete sich v. a. der Frage, ob und wie Smart Speaker sprachlich an der sozialen Praxis und an Gesprächen ‚beteiligt‘ werden. Entsprechend standen hier Mehrparteienkonstellationen im Fokus; ausgewertet wurden in diesem Teil Aufnahmen, auf denen mehrere Sprecher*innen zu hören waren. Für die Analysen wurde das Konzept der „materiellen Partizipanden des Tuns“

nach Hirschauer (2004; 2016) genutzt. Demzufolge können an der sozialen Praxis und der Durchführung einer Praktik menschliche wie nicht-menschliche „Partizipanden“ mit unterschiedlichen Aktivitätsniveaus teilnehmen – sie sind Teil eines Praxisvollzugs. Dieser Vollzug gestaltet sich, wie schon in den Analysen im ersten Teil gezeigt werden konnte, in an der Oberfläche gesprächsartigen Sequenzen von akustischen Äußerungen, die die Anwender*innen und die VUIs produzieren. Mit diesem Umstand gehen die Anwender*innen von VUIs unterschiedlich um, wie sich in den Daten zeigte: Einerseits können die Anwender*innen nahtlos von ihrem *turn-by-turn talk* in eine an das VUI gerichtete Stimmeingabe übergehen (und zurück). Darüber hinaus konnte ein Fall beschrieben werden, in dem die sequenzielle Verfertigung des VUI-Dialogs auf der einen und die der zwischenmenschlichen Interaktion auf der anderen Seite sich kreuzen. Das spricht dafür, dass die sprachlichen Praktiken in VUI-Dialogen für die ko-präsenten Sprecher*innen als solche erkennbar sind, sodass sie nicht unbedingt einer Meta-Kommentierung bedürfen. Möglich sind aber auch die Produktion von prä-hoc- und post-hoc-Einbindungen, d. h. vorherigen oder anschließenden sprachlichen Verknüpfungen zwischen der laufenden sozialen Praxis und der Anwendung. Diese entstehen, wenn die Nutzung des VUIs stärker ins Zentrum rückt.

Insbesondere die Fälle, in denen das VUI im Zentrum der sozialen Praxis steht und auch sprachlich an diese angebunden wird, wurden einer genaueren Betrachtung unterzogen. In Mehrparteienkonstellationen zeigt sich noch einmal deutlich, dass die Geräte schon per Design nicht darauf ausgerichtet sind, die Eingaben von mehreren Nutzer*innen gleichzeitig zu verarbeiten. Entsprechend vollziehen sich auch hier dyadische VUI-Dialoge, die den beschriebenen Basis-Sequenzstrukturmustern folgen und zugleich in Mehrparteienkonstellationen eingebunden werden. Dabei wurden verschiedene soziale Praktiken identifiziert, in denen das VUI im Zentrum steht, mit entsprechenden Folgen für die sprachliche Gestaltung des Austauschs. Dazu zählen neben den Ersteinrichtungen des Smart Speakers auch das Üben bzw. Testen, Vorführungen der Smart Speaker für andere Personen (insbesondere Gäste), gemeinsame Bewertungen der VUI-Performanz und die interaktionale Bearbeitung von Fehlschlägen. In diesen Analysen war die Unterscheidung zwischen ‚Reden‘ und ‚Gespräch‘ (Goffman 1979: 6–7) wichtig: Für die soziale Praktik eines Gesprächs ist Reden zwar notwendig (wenn auch nicht fortlaufend und nicht notwendigerweise stimmlich), nicht jedes Reden konstituiert umgekehrt aber ein Gespräch, sondern es kann auch empirischer Teil anderer Praktiken sein – mit entsprechenden Folgen für die sprachliche Gestaltung.

Nun zeigen sich aber gerade in den genannten Situationen Verknüpfungen in der – eben inkrementellen, situations- und materialgebundenen – sozialen Praxis: Zwischen dem gesprächsartigen Reden zur Bedienung eines VUI und der so-

zialen Praktik des Gesprächs bestehen Verbindungen. Das Sprechen *mit* dem VUI verband sich dann mit dem Gespräch *über* das VUI. Dies zeigt sich z. B. in der von den VUIs produzierten Eröffnung und Rahmung der Ersteinrichtung als gemeinsame Aktivität, die von den Anwender*innen gemeinsam mit dem VUI durchgeführt wird. Ebenso zeigt sich dies in den häufig daran anschließenden ersten Versuchen der Nutzung des Geräts, die mit den Ko-Beteiligten abgesprochen wird (vgl. Pitsch et al. 2017); das VUI wird hier in besonderer Weise zum ‚Beteiligten‘ an der sozialen Praxis. Sprachlich gestalten die Nutzer*innen dies u. a. durch direkt an das VUI gerichtete Adressierungen, Rezeptionssignale der Nutzer*innen, kohärente Anschlussturns an Äußerungen des VUIs und andere Formen, die auch in Vorführungssituationen und Bewertungen auftreten. Besonders auffällig ist die Produktion solcher Formen auch in Situationen, in denen die Nutzung des VUIs nicht gelingt. Zwar gelten weiterhin die Erkenntnisse, die aus den dyadischen Konstellationen für Reparaturen in VUI-Dialogen gezeigt wurden. Hinzu kommt aber eine deutlich beobachtbare Tendenz, unpassende oder falsche Äußerungen des VUI in Mehrparteienkonstellationen klar als solche zu kennzeichnen und dabei ebenfalls auf sprachliche Formen zuzugreifen, die die Geräte teilweise in laufende Gespräche zwischen den menschlichen Beteiligten eingliedern und sie so zu Partizipanden werden lassen.

Schon bei der situationsspezifischen Betrachtung wird deutlich, was durch die daran anknüpfende Untersuchung der genauen sprachlichen Praktiken zur Herstellung dieser Form der Partizipation noch einmal bestätigt wird: Der Beteiligungsstatus, der hier entsteht, ist nicht von Dauer, vielmehr ist er brüchig. Bei der Bedienung von VUIs werden zwar eine Reihe sprachlicher Praktiken eingesetzt, mit denen ein gewisser Grad an ‚Beteiligung‘ entsteht. Dieser wird jedoch auch sprachlich nicht konsequent umgesetzt: Kontinuierlich wechselnde Genus-Verwendungen, uneingeleitete Wechsel von Hörer- zu Objektdeixis, überlappende Äußerungen (auch die bereits erwähnten *barge-ins*) und auch objektifizierende Zeigegesten verdeutlichen, dass der zugeschriebene Status nur an der Oberfläche besteht und in den untersuchten Daten auch keine weiteren Sozialfolgen zeitigt.

7.2 Diskussion der Ergebnisse

Im Folgenden sollen die oben zusammengefassten Befunde der Arbeit konzeptionell diskutiert werden – zunächst vor dem Hintergrund der eingeführten und in den Analysen verwendeten Konzepte, aber auch mit Blick auf Implikationen und Konsequenzen, die sich ggf. aus den Befunden auch für andere Konzepte ergeben können. Außerdem wird die Arbeit kurz methodologisch reflektiert und es wer-

den mögliche Folgefragen und Anschlussuntersuchungen diskutiert, die sich aus den Befunden der vorliegenden Arbeit und ihrer Diskussion ergeben haben. Theoretische Ausgangspunkte der Arbeit waren neben dem Konzept der „Ko-Operation“ (Goodwin 2018) jeweils Begriffe von sozialen Praktiken, sprachlichen Praktiken und Medienpraktiken. Praktiken wurden dabei in ihrem reflexiven Spannungsverhältnis von übersituativer Einheiten- und Regelbildung auf der einen und situativen Aufführungen in der Praxis auf der anderen Seite verstanden (vgl. Schatzki 1996; Selting 2016; Schüttpelz/Meyer 2017). Mit dieser Perspektive konnte nachvollziehbar gemacht werden, welche sprachlichen Praktiken VUIs mit hervorbringen, wie VUIs situativ an der Praxis beteiligt werden und wie sie sich dadurch in soziale Praktiken einschreiben – und mithin domestiziert werden. Darauf soll nun im Einzelnen eingegangen werden.

7.2.1 Sprachliche Praktiken – Medienpraktiken im Interface

Das praxeologische Grundgerüst hat sich als stabile Ausgangslage erwiesen, um die ko-operative Verfertigung der Medien in den Blick zu nehmen und so die Eingliederung von VUIs in den Sprachgebrauch zu untersuchen. Es konnte beschrieben werden, wie sich in VUI-Dialogen sowohl neue sprachliche Praktiken ausbilden als auch sprachliche Praktiken aus dem Repertoire der Anwender*innen für die Anwendung im Kontext der VUI-Nutzung herangezogen wurden. Es konnte außerdem gezeigt werden, dass Nutzer*innen implizites Wissen³⁵⁴ über Gesprächsorganisation und damit gekoppelte sprachliche Praktiken mit dem (zum Teil neu erworbenen) Wissen über die Möglichkeiten und Grenzen der Interfaces verbinden. Es wurde beobachtbar gemacht, wie Nutzer*innen einerseits bekannte Strategien der Gesprächsorganisation zum Gebrauch im VUI-Dialog ‚verlängern‘ – z. B. in der Gestaltung der Redezugorganisation oder bei Adressierungen –, andererseits dabei aber auch ihre Vorstellungen über die Systeme berücksichtigen, was z. B. im Rahmen von Reparaturen deutlich sichtbar wird (siehe auch das Beispiel bei Merkle/Hector 2025). Sprachliche Praktiken *leiten* (vgl. Distelmeyer 2020) insofern von kognitiven bzw. kognitionslinguistischen Prozessen und den sozialen Gegebenheiten in privaten Haushalten über die VUIs³⁵⁵ hin zur cloudbasierten Datenspeicherung und -verwertung der

³⁵⁴ Implizites Wissen wird hier in der Verwendung von Ernst (2017) mit Bezügen zu Polanyi (1985[1966]), Collins (2010) und Loenhoff (2015) gebraucht. Als Ausgangsdefinition muss dabei genügen, was Ernst (2017: 99) zusammenfasst: „Implizit‘ ist dieses Wissen, weil es nicht bruchlos ‚expliziert werden kann‘; siehe auch Kap. 2.1.4.

³⁵⁵ Involviert in diesen Prozess sind neben einem VUI eine unüberschaubare Anzahl weiterer Interfaces, die Verbindungen zwischen Soft- und Hardware herstellen (vgl. Cramer/Fuller 2008:

Systemanbieter, um dort u. a. als Basis der Dienstleister für kontinuierliches Interface-Design zu dienen. Von dort wiederum fließen darauf basierende Updates der VUIs – z. B. in der Speech Recognition und der Stimmsynthese – zu den Nutzer*innen zurück und können wiederum Modifikationen der sprachlichen Praktiken evozieren. Mit der Betrachtung dieser reflexiv verlaufenden Aneinanderreihung von Abläufen werden sprachliche Praktiken verständlich als Medienpraktiken. Sie sind der nutzer*innenseitig beobachtbare Bestandteil von „Interfacing“ bei VUIs und Arbeit an den Bedingungen für Kooperation. Mit anderen Worten: Sprachliche Praktiken *sind* Medienpraktiken. Die Praktiken bekommen durch ihren Gebrauch in VUI-Dialogen also vor dem Hintergrund eines relationalen Interface-Begriffs³⁵⁶ einen besonderen Status als reflexive, praktisch verfertigte Verbindung technischer Prozesse und Artefakte mit sozialen und kognitiven Gegebenheiten: „In Interfaces finden Übergänge zwischen situierter Kognition und dem impliziten Regelwissen von Praktiken statt“ (Ernst 2017: 101): Die Nutzer*innen, so ließ sich anhand der Praktiken (wenn auch implizit) beobachten, verbinden im VUI-Dialog ihr implizites (sprachliches, gesprächsorganisatorisches) Regelwissen mit den situativ durch ein VUI gestellten kommunikativen Aufgaben. Dabei war keineswegs das Anliegen, auf Grundlage der hier vorgelegten Untersuchung trennscharf zu entscheiden, wo implizites Wissen anfängt und wo „sitierte Kognition“ (siehe Ataizi 2012) aufhört – ein Zugriff auf kognitive Prozesse war generell nicht Anliegen der Arbeit. Vielmehr war die Beschaffenheit der sprachlichen Praktiken von Interesse, die aber von diesen Prozessen an der Oberfläche zeugen.

7.2.2 Sprachliche Praktiken in VUI-Dialogen als Sprachregister?

Damit ist ein anderer Aspekt des sprachlichen Austauschs zwischen Menschen und Maschinen angesprochen: Die seit den 1980er-Jahren mit dem Aufkommen von „natural language interfaces“ (Fischer 2006: 1) immer wieder gestellte Frage nach dem empirischen Gehalt eines eigenen *sprachlichen Registers* oder einer eigenen *Varietät* für den Austausch mit Computern (etwa Zoeppritz 1985; Krause/Hitzenberger 1992). Als Varietät versteht Fischer (2006: 25) mit Crystal (2001: 6–7) allgemein ein System sprachlicher Ausdrücke, deren Gebrauch durch verschiedene Einflussfaktoren bestimmt wird. Als Register wird genauer gesagt eine sprachliche Variation gefasst, die spezifische Eigenschaften aufweist, die nicht

149), aber aufgrund des sprachwissenschaftlichen Erkenntnisinteresses dieser Arbeit nicht näher untersucht worden sind.

356 Siehe ausführlich Kap. 2.1.4.

etwa durch geografische oder soziokulturelle Gegebenheiten bestimmt werden, sondern durch die Umstände im Vollzug des Sprechens (vgl. Coseriu 1988 [1980]: 25). Register sind mithin eine *gebrauchsspezifische* Varietät (vgl. Dittmar 2004: 217–218)³⁵⁷ – Gegenstand ist also die Frage, ob der Gebrauch von Sprache im Austausch mit Computern hinreichend spezifische sprachstrukturelle oder funktionale Folgen zeitigt, um als eigenes Register – „Computer Talk“ nach Zoeppritz (1985) – gelten zu können.

Krause (1992b: 16) geht spezifisch von der Entstehung eines *vereinfachten* Registers (*simplified register*) aus, ein Begriff, der von Ferguson (1975; 1982) vertreten und z. B. zur Beschreibung von *Foreigner Talk* oder *Baby Talk* herangezogen wurde (vgl. auch Krause 1992b: 4). Parallelen bestehen auch zu Eingaben bei der Benutzung von Datenbanken bzw. Suchmaschinen, die sprachlich als Schlüsselwort-Suche realisiert werden und u. a. an Booleschen Operatoren orientiert sind.³⁵⁸ Allerdings ist dies für die meisten Systeme nicht mehr notwendig; für Fragehandlungen ist ein beginnender Wandel von Schlüsselwort-Suchen zu ‚natürlichsprachigen‘ konversationellen Fragen beobachtbar, auch wenn die Schlüsselwort-Sucheingaben noch zu überwiegen scheinen (vgl. White/Richardson/Yih 2015).³⁵⁹ Die Formulierung von Suchanfragen im Allgemeinen weist außerdem insgesamt eine hohe sprachliche Variabilität auf, deren Entstehung und Einflussfaktoren noch nicht umfassend verstanden sind (vgl. Alaofi et al. 2022).

Durch die hier untersuchte *mündliche* Gestaltung von VUIs und den hohen Grad an Dialogizität im Interface könnte eine neue Spezifik auftreten, die bisher nicht ausführlich im Blickfeld der Forschung lag – eine der wenigen Untersuchungen, die zu dieser Frage beiträgt, kommt von Opfermann et al. (2017), und zeigt, dass Nutzer*innen sich in Dialogen mit einem Embodied Conversational Agent nicht durchgängig von einem *simplified register* Gebrauch machen, sondern auch Praktiken aus zwischenmenschlichen Interaktionen anwenden. Dies kann die vorliegende Untersuchung grundsätzlich bestätigen: Zwar sind auf der

357 Die begriffliche Debatte u. a. in der Soziolinguistik um Varietät, Register und Stil mit unterschiedlich nuancierten Vorschlägen soll hier nicht näher dargestellt werden (siehe etwa Auer 2012), es bleibt bei den Begriffsbestimmungen aus den bisherigen Arbeiten zu „Computer Talk“ (Zoeppritz 1985).

358 Als Boolesche Operatoren für Suchmaschinen werden u. a. „UND“, „NICHT“ und „ODER“ genutzt, mit denen einzelne Suchworte miteinander in Beziehung gesetzt werden. Diese kamen bereits in den 1970er-Jahren bei Interfaces für Bibliothekskataloge zum Einsatz und werden bis heute mehr oder weniger explizit bei der Gestaltung von Suchanfragen verwendet (vgl. Kerssens 2017: 222).

359 Dabei ist zu beachten, dass die Ergebnisse auf Suchanfragen beruhen, die zwischen November 2011 und Januar 2013 durchgeführt wurden. Eine Fortsetzung des von White/Richardson/Yih (2015) beobachteten Trends zur Formulierung konversationeller Fragen scheint also wahrscheinlich.

einen Seite durchaus wiederkehrende Charakteristika in der Gestaltung von Stimmeingaben beobachtbar. In der Analyse der vorliegenden Arbeit konnten insbesondere Ein-Wort-Eingaben (Nomen oder Adjektive), syntaktisch nicht geschlossene bzw. elliptische Phrasen (u. a. deonitische Infinitive), Imperativsätze sowie Ergänzungs- und Entscheidungsfragen beobachtet werden.³⁶⁰ Bilden diese Praktiken jedoch zusammengekommen ein sprachliches Register bzw. eine sprachliche Varietät mit spezifischen Merkmalen? Diese Auflistung sprachlicher Praktiken ist auf der anderen Seite durchaus recht breit, zudem könnten auch weitere, in den hier ausgewerteten Daten nicht beobachtbare Praktiken auftreten. Darüber hinaus sind sie nicht unbedingt spezifisch für die Anwendung in VUI-Dialogen oder in Mensch-Computer-Konstellationen überhaupt. Was sich als Charakteristikum verallgemeinernd festhalten lässt, ist eine Tendenz zum „knappen Sprechen“ (vgl. Baldauf 2002) in Verbindung mit Sprechen, das als „empraktisch“ (Bühler 1965[1934]: 155–156) bekannt ist und insofern einen hohen Grad an Handlungsorientierung aufweist bzw. selbst die Handlung vollzieht und „Face-Work“ (Goffman 1955) (temporär) suspendiert. Dies wurde linguistisch aber auch für andere Kontexte untersucht, z. B. für medizinische Operationen (siehe Mondada 2014b) oder Fahrschulstunden (vgl. Deppermann 2018a). Hier lässt sich insofern auch eine Ähnlichkeit zu instruktivem Sprechen feststellen, das in beiden von Mondada und Deppermann untersuchten Kontexten auftritt. Sie eint auch, dass ein höherer Aufwand zur Vermeidung von Missverständnissen betrieben wird. Dies ist auch in VUI-Dialogen beobachtbar: Die Tendenz zur Kürze könnte auch darin begründet liegen, dass der Austausch mit dem VUI als fehleranfällig und im Hinblick auf Reparaturen als umständlich beschrieben werden kann.

Wie schon Zoeppritz feststellt, lässt sich also keine Ausbildung spezifischer sprachlicher Praktiken zur Bedienung von Maschinen feststellen, vielmehr sind die Formen „mostly borrowed from styles already available for interpersonal communication“ (Zoeppritz 1985: 20). Fischer (2006: 51) schreibt dazu: „If CT [Computer Talk, T.H.] is a register or sublanguage, it should display some homogeneity just on the basis of the fact that the speech is directed towards an automatic speech processing system“. In ihrer eigenen empirischen Untersuchung auf Basis mehrerer Korpora kann sie dies nicht bestätigen, weder im Sinne eines strukturellen Registers – wie es von Krause (1992a: 157–158) bzw. Krause/Hitzenberger (1992) konstatiert wurde – noch im Hinblick auf eine funktionale Varietät, wie

360 Diese zeigten sich insbesondere in dyadischen Konstellationen (siehe dazu Kap. 6.1.3.3), die Mehrparteienkonstellationen weisen demgegenüber eine etwas höhere Variabilität auf, entstammen aber auch spezifischen Situationen, z. B. Ersteinrichtungen, Vorführungen und Tests, was die Aussagekraft für die routinierte Nutzung schmälert.

von Johnstone et al. (1994) vorgeschlagen.³⁶¹ Letztere argumentieren v. a. mit der funktional bedingten Reduktion von „Grounding“ (Clark/Schaefer 1989: 262) und Höflichkeitsarbeit (siehe Brown/Levinson 1987). Fischer (2006: 26–70) zeigt, dass diese Merkmale nicht als charakteristisch für ein sprachliches Register oder eine funktional ausgebildete Varietät gelten können: Zu groß sind die Unterschiede zwischen den Sprecher*innen sowie zwischen den verschiedenen Funktionen der untersuchten sprachlichen Formen, um für ein sprachstrukturelles Register zu argumentieren, zu wenig charakteristisch die untersuchten sprachlichen Praktiken, um für die Ausbildung einer funktionalen Varietät zu sprechen (vgl. Fischer 2006: 74). Dies zeigt sich auch in den vorliegenden Analysen: Das herausgearbeitete, weiter oben benannte Set an Formen zur TCU-Gestaltung für Stimmeingaben scheint nicht homogen zu sein – was sich schon auf Basis der hier gezeigten unterschiedlichen Fälle zeigen lässt. Die Suspendierung von „Face-Work“ (Goffman 1955) ist auch für andere empirische Kontexte beobachtet worden. Zudem sind – jedenfalls an der sprachlichen Oberfläche – Formen von Face-Work beobachtbar. Gleichwohl ist Höflichkeit, wie auch von Fischer (2006: 70) angemerkt, eine vielschichtige Kategorie. Das in Mehrparteiensituationen beobachtete Auftreten von Invektiven und die Einbettungsformen in laufende zwischenmenschliche Interaktionssituationen deuten darauf hin, dass eine funktionale Differenzierung im Hinblick auf den sozialen Status der Geräte durchaus erfolgt.

Die verbindenden Eigenschaften von „Computer Talk“ sind, so argumentiert Fischer (2006) weiter, vielmehr *Strategien* im Umgang mit den Geräten, die auf die grundlegend anders ablaufende Sprachverarbeitung der VUIs (im Unterschied zum Menschen) und damit einhergehende Limitationen ebenso Rücksicht nehmen wie auf unterschiedliche Verarbeitungsweisen der situativen Gegebenheiten und unterschiedliches Weltwissen. Nutzer*innen produzieren dabei einen hohen Grad an Alignment im Hinblick auf Adressierungen, Lexik und Syntax (vgl. Fischer 2006: 107), den auch Lotze (2016: 287) am Beispiel von Chatbots feststellt.³⁶² Diese Sichtweise auf sprachliche Strategien als „Computer Talk“ kann anhand der Befunde zu sprachlichen Praktiken im Umgang mit VUIs bestätigt und teilweise

361 Dieser Unterschied beruht auf verschiedenen Ideen davon, welche konstitutiven Eigenschaften für ein Sprachregister gegeben sein müssen: „From the standpoint of language structure, registers differ in the type of repertoire involved, e.g., lexemes, prosody, sentence collocations, [...]; from standpoint of function, distinct registers are associated with social practices of every kind [...].“ (Agha 2004: 24).

362 Lotze (2016) arbeitet dabei mit einem sehr weit ausdifferenzierten Alignment-Begriff, der die Befunde von Fischer teilweise als „reaktives Alignment“ bestätigt; diese aufschlussreiche Diskussion soll hier nicht wiedergegeben werden.

ergänzt werden: Strategien der Nutzer*innen, die in spezifische sprachliche Praktiken münden, zeigen sich u. a.

- in der Basis-Sequenzstruktur, an der sich die Nutzer*innen im VUI-Dialog orientieren,
- an den unterschiedlichen Verfahren der sequenziellen Einbettung der VUI-Dialoge in zwischenmenschliche Interaktionssituationen,
- an den *barge-ins*, die Nutzer*innen einsetzen, um die Äußerungen des VUI zu unterbrechen und insofern keine Orientierung an Kategorien der Höflichkeit zeigen,
- an den unterschiedlichen Reparaturstrategien der Nutzer*innen, sowie auch
- am Einsatz von Invektiven zur interaktiven Bearbeitung von Fehlschlägen.

Diese Formen sind nicht distinkt im Sinne eines strukturellen oder funktionalen Registers. Sie sind aber als Set sprachlicher Strategien spezifisch und zeigen sich als emergente Zusammenhänge von Form und Funktion, die sich aus der situativen Praxis heraus zu verfestigen scheinen. Hinweise darauf geben die untersuchten Einlern-Dialoge, die etwa die Basis-Sequenzstruktur illustrieren und zu *barge-ins* einladen („Du kannst jederzeit ‚Stop‘ sagen“) – sprachliche Praktiken der Nutzer*innen gehen also auch auf Phasen des Einlernens zurück.³⁶³ Dies ist in den vorliegenden Daten ein wiederholt auftretendes Phänomen. Unterschiedliche Reaktionen der Sprachassistenzsysteme auf Invektive zeigen, dass dies seitens der Hersteller antizipiert und entsprechend im VUI-Design darauf reagiert wurde.

Letztlich muss im Hinblick auf diese Frage auch konstatiert werden, dass die vorliegende Arbeit mit ihrem qualitativ-explorativen Zuschnitt keine quantitativ gestützte Aussage über den Status der Registrierung spezifischer Formen treffen kann. Um die Routinisierung und die über Sprecher*innengruppen hinweg bestehende Etablierung als sprachliche Varietät beurteilen zu können, müsste ein deutlich größeres Korpus erhoben und durch Kodierungen ausgewertet werden, wie etwa in der Untersuchung von Barthel/Helmer/Reineke (2023) – darauf wird noch einmal eingegangen. Die genannten Strategien wären mögliche Ansatzpunkte, um hier weitere Untersuchungen anzustellen. So kann auch einer der sprachlichen Praktiken immer innewohnenden Übersituationalität als Voraussetzung für ihre Verfestigung und Routinisierung auf einer stabileren empirischen Grundlage nachgegangen werden.

363 Eine andere, nicht ausgewertete Quelle sind etwa E-Mail-Newsletter von Amazon an die Nutzer*innen, in denen neue Funktionen vorgestellt und länger existierende Funktionen in Erinnerung gerufen werden. Wie die Stimmangaben dazu zu formulieren sind, wird in diesen Newslettern ebenfalls mitgeteilt.

7.2.3 Zur ‚Beteiligung‘ von VUIs an Praxis und Gespräch

Ein weiteres Ziel der Arbeit war, zu untersuchen, wie VUI-Dialoge und die soziale Praxis miteinander verwoben sind. Dazu wurde das Konzept der materiellen Partizipanden des Tuns nach Hirschauer (2004; 2016) nutzbar gemacht, der von einer verteilten Handlungsträgerschaft in der Praxis ausgeht, in der Aktivitätsniveaus diverser Partizipanden unterschiedlich aufgeteilt sein können. So konnten VUIs einerseits als Beteiligte an der sozialen Praxis betrachtet werden, zugleich wurde ihr Status in Gesprächen untersuchbar. Wie oben ausgeführt, kann dabei ein Auseinanderfallen von Form und Funktion konstatiert werden: Es zeigten sich sprachliche Praktiken zur Herstellung von formalen Gesprächsbeteiligungen eines VUIs auf der einen Seite, die aber auf der anderen Seite von brüchigem, temporärem Charakter waren und an vielen Stellen andere Funktionen als die tatsächliche Unterhaltung mit einem VUI erfüllten. Dabei ist der Austausch ähnlich zu Mensch-Tier-Interaktionen – sie sind auch oder überwiegend ein mehrfach-adressiertes *display* anderer Praktiken (vgl. Tannen 2004; Torres Cajo/Bahlo 2016). Hierbei ergibt sich ein Kontinuum von mehr oder weniger stark ausgeprägten Mehrfachadressierungen: Wird zum Beispiel das Aktivierungswort („Alexa“ o. a.) weggelassen, ist anzunehmen, dass die Nutzer*innen *primär* die ko-präsenten menschlichen Beteiligten adressieren und weniger das VUI selbst. Eine sprachliche Interaktion zwischen den adressierten Entitäten und den Sprecher*innen i. e. S ist dabei eine für alle Beteiligten offensichtliche Fiktion, die sich in die laufende soziale Praxis einfügt und innerhalb dieser andere Funktionen erfüllt. Die Beobachtungen zeigen, dass das Reden über den Smart Speaker die Integration solcher Formen begünstigt.

Mit der bereits mehrfach zitierten und diskutierten praxistheoretischen Konzeption von Hirschauer (2004; 2016) und der darin angelegten Vollzugsperspektive konnte aufgedeckt werden, dass VUIs als „materielle Partizipanden verteilten Handelns“ (Hirschauer 2016: 51) gelten können. Ihre dialogischen Äußerungen werden als Ressourcen verwendet, um ko-operativ neue Koaktivitäten hervorzu-bringen. VUIs und ihre Äußerungen legen dabei die Ausführung bestimmter Praktiken nahe. Dies geschieht teilweise sehr explizit – z. B. wenn mit von den VUIs produzierten Adhortativen und Imperativen in den Ersteinrichtungsdialogen der Vollzug einzelner Praktiken ‚eingeübt‘ wird –, teilweise aber auch außerhalb des von den VUI-Designer*innen vorgesehenen Spektrums – etwa wenn die Nutzer*innen formal an das VUI gerichtete Dialogbeiträge produzieren, um die Äußerungen und Anschlus-säußerungen als *display* von Frust zu verwenden oder wenn VUI-Äußerungen als Lachgegenstand funktionalisiert werden. Dies belegt noch einmal die These von Goodwin (2018: 444), dass der Vollzug von Praktiken von einem semiotischen Opportunismus gekennzeichnet ist – „with the ability to incorporate voraciously whatever

local materials might be used to construct the action required at just this moment“. Sie werden damit also in Gespräche einbezogen und sind Teil des praktischen Vollzugs eines Gesprächs. Das macht sie – bzw. die Sprecher*innen machen sie – aber nicht zu Gesprächsteilnehmer*innen im engeren Sinn. Das Auseinanderfallen von formaler und funktionaler Ebene des Sprachlichen und die hohe Geschwindigkeit, mit der die zugeschriebenen Statuskategorien wechseln, die wiederum konzeptionell undurchlässig sind, bestätigen, dass die Kategorien aus Goffmans *Footing*-Konzept³⁶⁴ im Hinblick auf ihren tatsächlichen Vollzug bei nicht-menschlichen Beteiligten an ihre Grenzen stoßen. Die Ergebnisse zeigen deutlich auf, dass es sich um nicht-menschliche Partizipanden handelt, die zwar mit Hirschauer an der Praxis partizipieren können, aber darum noch nicht kontinuierlich an einem Gespräch teilnehmen.

7.2.4 Domestizierung im *display* sprachlicher Praktiken

Die sprachlichen Praktiken mit VUIs sollten *im Prozess ihrer Domestizierung* untersucht werden. Dazu wurden zum einen Konzepte aus der Medienaneignungsforschung vorgestellt, diese wurden zum anderen mit dem Domestizierungsansatz, seinen Grundlagen und aktuellen Anwendungen und Kritik vermittelt. Als Teil der Aneignung bzw. Domestizierung sind also insbesondere Situationen der Ersteinrichtung, des Einlernens, Testens und Übens relevant, in denen sich zeigt, wie Nutzer*innen den Gebrauch anfangs ungewohnter Interfaces zur Gewohnheit werden lassen. Ertragreich zur Bearbeitung dieser Fragestellung waren insbesondere die Mehrparteienkonstellationen, in denen die VUIs im Zentrum der sozialen Praxis standen. In diesen Situationen war der Gebrauch von VUIs eingebettet in – den VUI-Dialogen vor- und nachgelagerte – zwischenmenschliche Interaktionen. In dyadischen Konstellationen waren solche expliziten Reflexionen nur im Ausnahmefall beobachtbar, sodass sie im Sinne eines konversationsanalytischen Vorgehens keine *displays* einer Bewertung enthielten – lediglich in zwei Aufzeichnungen aus demselben Haushalt äußert eine Nutzerin in dyadischen Konstellationen explizite Bewertungen. Abgesehen davon deuten in dyadischen Konstellationen lediglich nonverbale Elemente – z. B. vernehmliches Seufzen – auf Reflexionsprozesse hin. Während also die dyadischen Konstellationen eher Hinweise auf die Ausbildung sprachlicher Praktiken zur regelmäßigen Nutzung der VUIs geben, sind die Mehrparteienkonstellationen besser geeignet, um das Einlernen, Testen, Üben und die damit einhergehenden Reflexionen der Nutzer*innen

³⁶⁴ Siehe dazu Kap. 2.2.4.

einzufangen. Daher wird im Folgenden darauf der Fokus liegen, wenn es um die Domestizierungsprozesse geht.

In diesen Situationen tritt mit VUIs ein ungewohntes Interface in die häusliche Umgebung, das durch seine Sprech- und Dialogfähigkeit eine zuvor gesicherte Differenzierung zwischen Mensch und Maschine potenziell irritiert. Turkle (1995: 102) beschreibt, dass ein solcher Effekt verwischender Grenzen auch bei anderen Interfaces entstehen kann; für Smart Speaker, d. h. VUIs mit ihren ‚eigenen‘ Lautsprechern und insofern einer eigenen Materialität, ist dieser Effekt noch weitgehend unerforscht, Hinweise auf eine Verwischung ontologischer Grenzen konnten allerdings bereits gefunden werden (vgl. Guzman 2020; Weidmüller 2022). Dieser Effekt scheint in Mehrparteiensituationen noch stärker ausgeprägt zu sein (vgl. Etzrodt 2022). Dieser unsichere Status bedroht folglich auch die „ontological security“ (Giddens 2003[1984]: 50) der Haushaltsmitglieder. Davon ausgehend könnte die bei Dizdar et al. (2021) ausbuchstabierte Perspektive der „Humandifferenzierung“ helfen, um etwa das gezielte Heraus- und Überfordern der Geräte ebenso wie invektive Anschlusskommunikation zu verstehen. Das Konzept der Humandifferenzierung kann hier nicht umfassend eingeführt und beleuchtet werden, es soll aber kurz als möglicher, weiterführender Ansatz zur Domestizierungsforschung diskutiert werden. Anwendbar wäre dieser Gedanke z. B. auf analysierte Auszüge aus Ersteinrichtungen, in denen Nutzer*innen sprachlich deutlich betont nicht den Aufforderungen der VUIs folgen, sondern davon abweichen, sie umdeuten und die Geräte – teils als humorvoll gerahmte Ko-Aktivität – an ihre Grenzen bringen. Andere Situationen wären mehrfach funktionalisierte Invektive, wie sie v. a. in den Analysen scheiternder VUI-Dialoge herausgearbeitet wurden.

Dickel/Schmidt-Jüngst (2021: 343) argumentieren, dass Sprachassistentensysteme in der Werbung vermarktet werden als „Entitäten, die (mehr oder weniger spezifische) soziale Positionen einnehmen sollen“. Das bereits früh im Aneignungsprozess vollzogene Aufzeigen der Grenzen dieser Humanfiktion seitens der Nutzer*innen könnte also als gemeinschaftliche Vergewisserung der eigenen sozialen Position verstanden werden – und insofern als Praktiken der „ontologische[n] Außendifferenzierung“ (Dizdar et al. 2021: 8), denn Humandifferenzierung vollzieht sich, so Dizdar et al. (2021: 11), u. a. sprachlich und praktisch. Dickel/Schmidt-Jüngst (2021: 342) sprechen von einer

Konstitution der Außengrenzen des Humanen, also für die Arten und Weisen, in denen Nicht-Menschliches von Menschlichem unterschieden oder gerade nicht unterschieden wird. Denn durch solche Praktiken der Grenzziehung werden beide Seiten nicht nur zueinander ins Verhältnis gesetzt, sondern zugleich ihre jeweiligen Identitäten bestimmt. (Dickel/Schmidt-Jüngst 2021: 342)

Sprachliche Praktiken, mit denen die Grenzen der Interfaces aufgezeigt werden – gezieltes Überfordern, invektive Anschlussturns u. a. – können insofern als interaktive Arbeit an dieser Außengrenze verstanden werden. Die Nutzer*innen testen die Interfaces entlang der bereits weiter oben eingeführten Bruchlinie zwischen VUI-Dialog und zwischenmenschlicher Interaktion und zeigen sich selbst und übrigen Beteiligten durch entsprechende Praktiken auf, dass es sich bei VUI-Dialogen eben *nicht* um Interaktionen handelt, weil sich VUIs nicht als stabile Individuen erweisen. Die Nutzer*innen führen geradezu vor, dass die Parameter der Interaktion nach Goffman – geteilter Aufmerksamkeitsfokus, geteilte Wahrnehmung, wechselseitige Wahrnehmung dieser Wahrnehmung, koordinierte Ausrichtung von Handlungen am Gegenüber (vgl. Goffman 1983: 3) – nicht erfüllt sind. Insbesondere die fehlende Reflexivität der Geräte, d. h. eine wechselseitige Wahrnehmungswahrnehmung, fällt auf, wenn Äußerungen zwar an das VUI adressiert sind, diese aber zuvor gar nicht durch eine Invokationssequenz in einen entsprechenden Status versetzt wurden. Die fehlende Fähigkeit zur inkrementellen Handlungskoordination wird in Reparaturdialogen deutlich.

Zugespißt könnte man sagen: Scheiternde VUI-Dialoge und deren interaktive Bearbeitung mindern das Irritationspotenzial von VUIs. Dabei tritt zur Außendifferenzierung auch eine Form der Binnendifferenzierung hinzu, die sich insbesondere in der auf VUI-Dialoge folgenden Anschlusskommunikation zeigt: Die Möglichkeit, Invektive an diesen maschinellen Kommunikationspartner zu richten, d. h. dem VUI kein zu wahrendes „Face“³⁶⁵ zuschreiben zu müssen, bestellt eine soziale Hierarchie, derer sich hier vergewissert wird. Insoweit ko-präsente Mitglieder diese Äußerungen z. B. durch Lachen ratifizieren, schließen sie sich der durch den VUI-Dialog ausgedrückten sozialen Stellung an und bestätigen diese, auch für sich selbst. Sprachliche Praktiken diesen Typs werden somit verständlich als *display* von Domestizierung, als Bestandteil der Eingewöhnung, dem Zuschreiben einer Stellung des Geräts im Haushalt und somit der Vergewisserung der Grenzen des Maschinellen zur Sicherung der eigenen, humanen Überlegenheit. Sprachliche Praktiken, mit denen Nutzer*innen die sozialen Rahmen und Regulierungen in der Praxis abstecken, sind insofern Arbeit an der Erhaltung der „ontological security“ (Giddens 2003[1984]: 88).

Vor diesem Hintergrund hat es sich als produktiv erwiesen, in der Arbeit einen Begriff – den der *Interaktion* – zwischenmenschlichen Akteur*innen vorzubehalten. Dies war zunächst eine analytische Trennung; es ist aber – wie oben

³⁶⁵ Hier findet erneut der „Face“-Begriff von Goffman (1955) sowie in seiner sprachwissenschaftlichen Auseinandersetzung im Sinne einer Höflichkeitstheorie nach Brown/Levinson (1987) Anwendung.

ausgeführt – beobachtbar, dass die Teilnehmenden in ihren Äußerungen auch selbst Grenzen zwischen zwischenmenschlicher Interaktion und VUI-Dialog vollziehen, obschon einzelne Sequenzen zunächst das Gegenteil annehmen lassen könnten. Dies bestätigt grundsätzlich die Sichtweise von Suchman (1987; 1990; 2007; 2021), dass es sich um eine ressourcenlimitierte Form der Kommunikation handelt – VUI-Design wäre in diesem Sinne auch „less a matter of simulating human communication, than engineering alternatives to talk’s situated properties“ (Suchman 2021: 77). Suchman (2007: 270) argumentiert, dass es sich um einen „dissymmetrischen“ Austausch handelt: Während die Maschine im Hinblick auf die Wahrnehmung der Situation über die sensorisch erfassten Nutzer*innen-Eingaben hinaus limitiert ist, steht den Nutzer*innen mit dem Smart Speaker ein Artefakt als Austauschpartner gegenüber, das durch Design und darin integrierte Annahmen über gelingende Dialoggestaltung verfertigt wurde. Der extrem weit gefasste Begriff des *VUI-Dialogs* hat sich demgegenüber als tauglich erwiesen, um damit offen den an der Oberfläche konversationell vollzogenen Operationen begegnen zu können – die sprachlichen Praktiken der Nutzer*innen waren Teil der Dialoge und somit der Interfaces selbst. Ein solch weites Verständnis des Dialogbegriffs ist v. a. für Phänomene interessant, die konzeptionell noch nicht stabil gefasst werden konnten und deren konzeptioneller Status gerade Teil der Untersuchung ist. Ob er sich mittelfristig als tauglich erweist zur Beschreibung der sprachlichen Ein- und Ausgaben im VUI-Dialog, d. h. im „Interfacing“ zwischen Mensch und Maschine, ist eine andere Frage.

Im Vergleich von dyadischen VUI-Dialogen und VUI-Dialogen in Mehrparteienkonstellationen ist auffällig, dass die linguistische Variabilität in Mehrparteienkonstellationen höher ist und die Tendenz zur Erzeugung formaler Kohärenz durch verschiedene sprachliche Verfahren nur in Mehrparteienkonstellationen besteht. Demgegenüber scheint in dyadischen VUI-Dialogen ein Fokus der Nutzer*innen auf ausgewählten Funktionen zu liegen, während der Großteil der Möglichkeiten mit VUIs nach der Testphase unausgeschöpft bleibt (vgl. Ammari et al. 2019; Pins/Alizadeh 2021). Situationen mit höherer sprachlicher Variabilität treten nach der Ersteinrichtung und den ersten Tagen danach erst wieder auf, wenn in Mehrparteienkonstellationen ein erneutes Testen oder Vorführen zum Zentrum der Praxis wird. Die sprachlichen Praktiken sind hier also als Evidenzen des Domestizierungsprozesses zu verstehen. Auch hierfür wäre eine quantitative Analyse zur Erhärtung dieses Befunds wünschenswert – Barthel/Helmer/Reineke (2023) zeigen hierfür vielversprechende Ansätze, die auf einem deutlich größeren Korpus und entsprechenden Kodierungen basieren und die beobachtete Tendenz bestätigen

können.³⁶⁶ Sinnvoll wäre auch die Berücksichtigung haushaltsspezifischer Unterschiede in einem größeren und diverseren Korpus – so zeigt etwa Schneider (2021: 337) im Rahmen einer Interview-Studie, dass Nutzer*innen in Extremfällen „ein emotionales, positiv besetztes Verhältnis“ zu Smart Speakern aufbauen können, einschließlich ihrer Materialität.

Teil der Domestizierung und der Veralltäglichen von VUIs ist insofern, so kann zusammengefasst werden, einerseits das Einlernen, das Herausfordern und Testen von Grenzen – diese gehen mit einer hohen Variabilität in Bezug auf die sprachlichen Formen einher und zeigen die erwähnten Irritationspotenziale im Hinblick auf den ontologischen und partizipatorischen Status der Geräte. Andererseits deuten die Befunde darauf hin, dass sich die „Everydayification“ (Ayaß 2012: 3) von VUIs gerade dadurch auszeichnet, dass die sprachliche Variabilität in VUI-Dialogen nicht zunimmt und gesprächsähnlicher wird, sondern sinkt und die Technizität deutlicher hervortreten lässt – was allerdings durch methodologisch anders gelagerte Studien überprüft werden müsste (etwa Barthel/Helmer/Reineke 2023).

7.2.5 Smart Speaker im bewohnten Raum

Die linguistische Erforschung von Interaktion interessiert sich zunehmend für deren Multimodalität. Die immense Relevanz visueller, körperlicher und räumlicher Aspekte traten durch neue, technisch weiterentwickelte Analysemöglichkeiten deutlicher hervor als in früheren konversationsanalytischen Arbeiten.³⁶⁷ Der Raum wurde dabei, so Schmitt/Hausendorf (2016: 13–16), zunächst mit dem Konzept des *Interaktionsraums* als „interaktive Ressource“ konzeptualisiert, in Fallstudien genauer untersucht (Hausendorf/Mondada/Schmitt 2012) und im Zusammenhang mit körperlichen Aspekten weiter exploriert (Schmitt 2013). In der Folge dieser genaueren Untersuchungen kam die „Ressourcenhaftigkeit des Raumes“ auch davon losgelöst in den Blick, die als den Interaktionen vorgelagert verstanden wurde (vgl. Schmitt/Hausendorf 2016: 16). Neben dem Interaktionsraum führen Hausendorf/Schmitt (2013; 2016a) daher die Konzepte der *Interaktionsarchite-*

³⁶⁶ Dabei sei – mit freundlicher Genehmigung der Vortragenden – auch auf den unveröffentlichten Vortrag von Mathias Barthel, Henrike Helmer und Silke Reineke bei der 17. Jahrestagung der International Pragmatics Association in Brüssel im Panel „(A)typical users of technology in social interaction“ (Chairs: Florence Oloff und Henrike Helmer) verwiesen. In diesem wurde die Frage der Routinisierung anhand eines größeren und kodierten Datensatzes in Kombination mit Einzelfallanalysen genauer beleuchtet, was die Produktivität von Mixed-Methods-Ansätzen (vorgeschlagen etwa von Kendrick 2017) bestätigt.

³⁶⁷ Im Detail siehe Kap. 4.4.

ktur und der *Sozialtopografie* ein. Damit werden – als Interaktionsarchitektur – basale „Implikationen des Raumes hinsichtlich seiner interaktiven Nutzung“ (Hausendorf/Schmitt 2013: 17) einer Analyse zugänglich, außerdem auch die kulturell aufgeladenen, aus gesellschaftlichen Konventionen und dem Wissen um diese entspringenden Interpretationen dieser Gegebenheiten – die Sozialtopografie. Hausendorf/Schmitt (2013: 3) unterscheiden außerdem zwischen dem „gebauten“, „gestalteten“ und „ausgestatteten“ Raum als heuristische Unterscheidung für Architektur.³⁶⁸ Der Schwerpunkt solcher Untersuchungen liegt bislang auf Interaktionen in institutionellen Zusammenhängen.

In der Praxis von Interaktionssituationen zur Einrichtung und Nutzung der Smart Speaker sind körperlich-räumlich-architektonische Kategorien unter mindestens drei Gesichtspunkten relevant. Erstens kann in den Ersteinrichtungen lokal die körperliche Positionierung zueinander, zum Smart Speaker als im (praktischen, nicht unbedingt räumlichen) Mittelpunkt stehendes Gerät und zu anderen interaktionsarchitektonischen Elementen als wesentlicher Träger der Praxis betrachtet werden. Dieser Aspekt hebt auf den „Raum als interaktive Ressource“ (Hausendorf/Mondada/Schmitt 2012) ab: die Nutzung des Raums zur Herstellung der Einrichtungssituation, um die Teilnehmendenkonfiguration auszuhandeln und das Zusammenspiel des Geräts mit den Teilnehmenden und der Umgebung näher zu bestimmen. Einige Aspekte dessen konnten in den vorangehenden Kapiteln behandelt und fortlaufend in den Analysen einbezogen werden,³⁶⁹ sie könnten aber noch weiter vertieft werden.

Zweitens lassen sich Einpassungen der Smart Speaker als Ausstattungsgegenstände in den gestalteten (Wohn-)Raum beobachten, damit einhergehende An- und Herausforderungen beschreiben und ein Spannungsverhältnis zwischen gestaltetem und ausgestatteten Raum ausmachen (siehe auch Hausendorf/Schmitt 2016b). Diese Perspektive scheint vielversprechend, um einen Beitrag zur Frage der Domestizierung der Geräte leisten zu können und den Domestizierungsansatz zu erweitern (vgl. Miggelbrink 2018: 87–88): Smart Speaker werden, sofern sich deren regelmäßige Nutzung in den Alltagspraktiken der Nutzer*innen etabliert, von einem Novum im Haushalt zu einem Teil der Ausstattung und damit der

³⁶⁸ Als „gebaut“ werden dabei nur unter größtem Aufwand veränderbare Elemente verstanden (etwa Wände, Dach, Boden, Raumaufteilung innerhalb der Wohnung), als „gestaltet“ innenarchitektonische Einrichtungen (Möbel, Lampen, Teppiche usw.) und als „ausgestattet“ leicht verrück- und verschiebbare Gegenstände (Dekorationsobjekte, Geschirr, technische Gegenstände usw.).

³⁶⁹ Dies ergibt sich auch aus der gesetzten Methodologie der Arbeit (siehe Kap. 4.4). Für Analysen, in denen multimodale Verfahren auch unter Einbezug des Interaktionsraums eine Rolle spielen siehe insbesondere Kap. 6.2.2.1 zu Ersteinrichtungen des Smart Speakers und 6.2.3.8 zu multimodalen Verfahren.

Interaktionsarchitektur;³⁷⁰ sie treten ein in eine wechselseitige Beziehung mit dem Mobiliar. Dieses ‚Infrastruktur-Werden‘, das reziproke Verhältnis zwischen bestehenden räumlichen Ausstattungen und Gestaltungen und dem neuen Gerät, das erst noch zur Ausstattung wird, könnte anhand von Videoaufnahmen von Smart Speaker-Anwendungen nachvollzogen werden: Welche Rolle spielen Sicht- und Hörbarkeit oder andere interaktionsarchitektonische Kriterien bei der Platzierung (vgl. Hausendorf/Schmitt 2013: 8)? (Wie) werden diese Kriterien expliziert und verhandelt? Anhand der vorliegenden Daten konnte dieses Themenfeld ausschließlich auf Grundlage der Videos von Situationen der Inbetriebnahme bearbeitet werden. Dabei konnte herausgestellt werden, dass Nutzer*innen u. a. den Blick als Ausdrucksressource verwenden, während das VUI Äußerungen produziert und somit den akustischen Kanal ‚besetzt‘. Außerdem zeigte sich, dass Smart Speaker auch über den akustischen Kanal hinaus in ihrer Materialität interaktionsrelevant sind – sie senden über die Oberfläche Signale aus, die wiederum von den Beteiligten in der Interaktion relevant gesetzt werden können. Sie können das Herstellen einer gemeinsamen Sehfläche aller Beteiligten mit dem Smart Speaker im Zentrum erforderlich machen. Allerdings konnte dieser Themenbereich nicht für die weiteren reziproken Prozesse zwischen Smart Speaker und Wohnumgebung untersucht werden, weil für die dauerhafte Nutzung der Geräte keine Videodaten vorliegen. Hier könnten sich Folgeuntersuchungen auf wohnräumliche Arrangements konzentrieren.

Drittens werden in der Konfiguration der Smart Speaker Fragen der Geräte- in Verbindung mit der Raumnutzung *herstellerseitig* relevant gesetzt, v. a. in den Ersteinrichtungsdialogen. Dabei steht der Charakter der Smart Speaker als Gemeinschaftsmedium in einem Spannungsverhältnis mit der individual angelegten Nutzung von User-Accounts für Kommunikations- und Unterhaltungsdienstleistungen: E-Mails, Kalender oder Musikstreaming sind personalisiert und möglicherweise zu privat für die geteilte Wohnumgebung. Wie bereits an anderer Stelle in dieser Arbeit gezeigt, ist auch die Bedienbarkeit über die Stimme ein Spezifikum dieser Geräte, das die Anwendung (im Gegensatz zum Smartphone) sicht- bzw. hörbar für andere Anwesende macht. Dieses Spannungsverhältnis ist nicht neu, es wurde etwa für Computer bereits beschrieben und Nutzer*innenstrategien untersucht, in denen unterschiedliche Raumnutzungskonzepte offengelegt werden konnten (vgl. Röser/Peil 2014). Auch die Untersuchung dieses Themenkomplexes konnte im Rahmen der vorliegenden Arbeit nicht nur aufgrund der Datenlage, sondern auch mit Blick auf

³⁷⁰ Letzteres wird noch plastischer, wenn über den Smart Speaker auch Smart Home-Anwendungen wie Licht, Klingel und Wärmeregulierung gesteuert und die Smart Speaker etwa von Amazon aktiv als Mittel zur Infrastrukturierung von Smart Home-Anwendungen positioniert werden (vgl. Strüver 2023a).

die Methodik nicht geleistet werden: Eine solche Untersuchung müsste sich anderen ethnografischen Methoden zuwenden und ggf. auch Interviews einbeziehen, die konversationsanalytisch angereichert werden könnten.

7.2.6 Methodologische Reflexion und Anschlussmöglichkeiten

Um eine von Meiler/Siefkes (2023: 321) zurecht geforderte Reflexion der Methodenwahl zu leisten, werden die obigen Ausführungen noch einmal methodologisch kondensiert; darauf aufbauend werden Anschlussmöglichkeiten für zukünftige Studien aufgezeigt. Die vorliegende Arbeit hat sich einer praxeologisch fundierten Erkundung und dem Aufdecken von Phänomenen und dem Verstehen dieser in ihrem situationalen Zusammenhang verschrieben – konkret sollten sprachliche Praktiken in VUI-Dialoge sichtbar gemacht werden, um ihren Bezug zu gesprächsorganisatorischen Prinzipien zu verstehen und die Verflochtenheit von sprachlichen Praktiken in VUI-Dialogen und sozialer Praxis zu beleuchten. Dabei stand eine ethnomethodologisch-konversationsanalytisch informierte Methodologie ethnografischer Gesprächsanalyse im Zentrum, die durch Sequenzanalyse und videobasierte Interaktionsanalyse einzelner Ausschnitte sprachliche Praktiken auffindbar und in ihrem Zusammenhang beschreibbar machen sollte. Die Puristik der Konversationsanalyse³⁷¹ wurde dabei zugunsten einer praxeologischen Betrachtung aufgegeben – zum Erkenntnisinteresse passte vielmehr eine ganzheitliche Betrachtung des Datenmaterials und die Zuhilfenahme ethnografischen Wissens außerhalb der reinen Ausschnitte. Zwar wurde dies im Rahmen der Analyse dyadischer VUI-Dialoge etwas eingegrenzt, gleichwohl wurde der Untersuchungsgegenstand von den reinen Aufzeichnungen ausgehend auch um Wissen über die soziale Konfiguration der untersuchten Haushalte ergänzt. Dabei wurde das Primat der Untersuchung von Sprache in ihrem ‚natürlichen Umfeld‘, d. h. zu tatsächlichen Nutzungssituationen und möglichst wenig von Aufnahmesituation und -kontext beeinflussten Sprechgelegenheiten berücksichtigt. Für die Beforschung eines neuen und sprachwissenschaftlich wenig erforschten Gegen-

³⁷¹ Eine entscheidende Rolle dabei spielt die Bedeutung von Ethnografie bei der Auswertung der Daten sowie die Praktiken der Datengewinnung (vgl. Bergmann 2001: 921; Bergmann 2007b: 58): Während die Ethnomethodologie ein ganzes Spektrum ethnografischer Methoden zulässt (z. B. auch teilnehmende Beobachtung) und dabei auch Hintergrundinformationen über die soziale Situation, den Kontext und über die Teilnehmer*innen der Interaktion einbezieht, blendet die Konversationsanalyse gerade diese bewusst aus und konzentriert sich nur auf die Phänomene, die in den Datenkollektionen (d. h. Transkriptionen von Audioaufnahmen) zu finden sind (vgl. Day/Wagner 2008: 44).

stands schien dies geboten, um nicht von vornherein Faktoren aus dem sozialen Kontext auszuschließen, die im Vorfeld nicht bekannt sein konnten (vgl. Kendrick 2017: 3).

Als Herausforderung erwies sich bei der Datenerhebung, was als ‚Widerpenstigkeit alltäglicher Praxis‘ bezeichnet werden könnte: die Schwierigkeit, den Alltag als solchen zu erkunden und einer Analyse zugänglich zu machen, mithin eines der Ausgangsprobleme des ethnomethodologischen Forschungsprogramms (vgl. Garfinkel 1967: 31–32). Auch im Rahmen von am Alltag interessierten Studien aus der Medienforschung wurde dieses Problem regelmäßig konstatiert und mit unterschiedlichen Vorgehensweisen umkreist (vgl. hierfür und für das Folgende Ayaß 2012: 8–12). Um das häusliche Alltagsleben der Nutzer*innen von Medientechnologien zugänglich zu machen, haben Studien zur Domestizierung von Medientechnologien primär mit ethnografischen Methoden gearbeitet (vgl. Hartmann 2013a: 53), deren typisches Repertoire jedoch für eine linguistische Auswertung nicht hinreichend gewesen wäre. Studien, die an einer feingranularen Beschreibung von Sprach- und Medienpraktiken interessiert sind, geraten immer wieder in ein methodologisches Dilemma: Medienpraktiken können etwa im Rahmen von qualitativen Interviews nur schwer elizitiert werden, weil sie den Nutzer*innen selbst nicht bewusst sind; außerdem können sie selbst mit beobachtenden Methoden (z. B. längeren teilnehmenden Beobachtungen) nur schwer entdeckt werden, weil sie von solcher Feingranularität und Flüchtigkeit sind, dass sie auch den Analytiker*innen ohne technische Hilfsmittel der Aufzeichnung verborgen bleiben (vgl. Ayaß 2012: 12–13). Die Aufzeichnung wiederum beeinflusst potenziell den Verlauf des Geschehens und insbesondere die sprachlichen Praktiken, allerdings bleibt die genaue Art des Einflusses teilweise verborgen. In der vorliegenden Untersuchung kann in der Instruktion zur gemeinsamen Durchführung der Inbetriebnahme der Smart Speaker eine Beeinflussung gesehen werden, von der die Daten in Form von deutlich produzierten situativen Rahmungen Spuren tragen. Andererseits sind die erzeugten Mehrparteienkonstellationen sehr aufschlussreich für die ‚Beteiligung‘ von VUIs an der sozialen Praxis gewesen. Das Ideal ‚natürlich entstehender Daten‘ ist nie vollständig zu erreichen, sondern als Kontinuum mit verschiedenen Parametern zu betrachten, die neben dem Aspekt der ‚spontan‘ entstehenden Teilnehmendenkonstellation auch Fragen der zeitlichen, lokalen und qualitativen Arrangiertheit betreffen (vgl. Kendrick 2017: 4). Die Frage, welche Gespräche dieses Ur-Kriterium gesprächsanalytischer Forschung hinreichend erfüllen, kann nicht abschließend beantwortet werden (für eine ausführliche Diskussion siehe Gerwinski/Linz 2018: 107–112). Hier bleibt für jede Studie erneut eine Abwägung zu treffen und entsprechende Entscheidungen sind durch eine gründliche Reflexion abzusichern.

Für eine längerfristige, situationsungebundene Beobachtung stellt sich zudem die Problematik, dass (auch der rein häusliche) Alltag nicht ‚vollständig‘ aufgezeichnet werden kann und ein minimal-invasives Vorgehen nicht nur zur Vermeidung von Spuren des Beobachterparadoxons, sondern auch aus forschungsethischen Gründen erforderlich macht. Eine videobasierte Erfassung des gesamten Alltags der Nutzer*innen über einen Zeitraum von mehreren Wochen hätte auch die Analytiker*innen vor große Herausforderungen bei der Bewältigung des Umfangs an gesammelten Daten gestellt und wäre unter forschungsethischen Aspekten mit der Erhebung von weitaus zu viel redundantem Material verbunden gewesen. So wurde die beschriebene, audiobasierte CVR-Technologie eingesetzt, um eine alltagsbasierte Nutzung von VUIs überhaupt zugänglich zu machen. Dabei ergibt sich bei der Auswertung die Problematik, dass der Alltag nicht immer in ‚Situationen‘ verläuft und die Sprecher*innen keineswegs immer verbalsprachlich anzeigen, was im Praxisvollzug gerade getan oder unterlassen wird und wer daran beteiligt ist. Die CVR-Aufzeichnungen stellen Bruchstücke aus dem Alltag der Nutzer*innen dar und von den Teilnehmenden produzierte Situationsrahmen, wie sie bei den Ersteinrichtungen beobachtbar und analysierbar waren, treten zu großen Teilen nicht auf. Dieser Umstand lässt, gerade ohne auf Videomaterial zurückgreifen zu können, Interpretationen zur sozialen Konstellation durchaus unsicher werden. Die Analysen wurden insofern, wo sie sich auf die CVR-Daten beziehen, vorsichtig formuliert und präsentiert.

Für eine Untersuchung mit diesen Charakteristika wurde (typischerweise) eine verhältnismäßig kleine Datenbasis gewählt, um bei den Analysen tiefer ins Detail gehen zu können, wie es ein sequenzanalytisches Vorgehen erforderlich macht (vgl. Deppermann 2008: 55). Die Bildung von Kollektionen in einem iterativen Wechselspiel aus Gegenstandskonstitution, Sampling und Gegenstandsanalyse ist ein in der Gesprächsanalyse etabliertes Vorgehen (vgl. Deppermann 2008: 95), das auch hier angewendet wurde, um das erhobene Material auszuwerten. Basis dieser Methode ist ein sequenzanalytisches Vorgehen, das keine isolierten Äußerungen betrachtet, sondern Ausschnitte aus der Praxis.

Die Gliederung des ersten Teils entlang der „generic organisations of practice“ (Schegloff 2006: 71) ist bereits ein Ergebnis sequenzanalytischer Beschreibungen in früheren Stadien der Untersuchung. Wie aus den obigen Ausführungen hervorgeht, ergeben sich aus diesen Analysen Anschlussfragen – insbesondere im Bereich der Routinisierung und Registrierung, dem Zusammenhang von sprachlichen Praktiken und Domestizierung sowie im Hinblick auf die Multimodalität. Anschlussfähig an die vorliegenden Befunde wären insofern ausgehend von der Methodologie Erweiterungen mit folgenden Linien:

1. Ein Ausbau der Untersuchung mit einer stärker **quantitativen Ausrichtung**; dazu wären sowohl ein größeres Korpus, eine systematische Kodierung wie auch eine in Teilen statistische Auswertung notwendig. Damit könnte der Verlauf der Domestizierung und ihr Niederschlag in sprachlichen Praktiken mit größerer empirischer Sicherheit eingefangen werden. Solche Ansätze sind durchaus mit konversationsanalytischen Grundprinzipien vereinbar (vgl. Pitsch et al. 2009b; Stivers 2015) und lassen Rückschlüsse auf „die Robustheit eines Phänomens“ (Pitsch 2023: 128) zu. Barthel/Helmer/Reineke (2023) zeigen bereits Vorteile eines solchen Vorgehens auf und berichten über erste Ergebnisse; die Autor*innen erproben dabei auch Mixed-Methods-Ansätze (s. o.), bei denen einzelne Sequenzanalysen mit korpusbasierten Auswertungen kombiniert werden. Solche Studien könnten auch zu weiteren Erkenntnissen im Hinblick auf die Registrierung kommen. Die vorliegenden Ergebnisse können für die Konzeptionierung und Durchführung einer solchen Studie wichtige Ansatzpunkte sein.
2. Eine stärkere Fokussierung auf die **Multimodalität** von VUI-Dialogen könnte die Rolle von Blick, Gestik, Mimik und körperlicher Orientierung zur Anzeige von ‚Beteiligung‘ stärker beleuchten und prüfen, ob diese außerhalb der hier erhobenen Ersteinrichtungsdialoge ebenfalls eingesetzt werden und wie sie funktionalisiert sind. Die hier vorgelegten Befunde lassen verschiedene Funktionen multimodaler Ausdrücke erkennen.
3. Um die Interaktionsbedingungen – z. B. bei der genaueren Betrachtung multimodaler Verfahren im Umgang mit VUIs – konstant zu halten, wären hier – wie etwa bei Etzrodt (2022) erprobt – im Sinne eines Methodenpluralismus in der empirischen Sprachwissenschaft (vgl. Kendrick 2017) auch **Laborstudien** möglich, was nicht zugleich eine quantitativ-hypothesengeleitete Auswertung bedingen muss; vielmehr kann gleichwohl ein qualitativ-entdeckendes Vorgehen angewendet werden (vgl. Pitsch 2023: 123). Auch hierfür liefern die Erträge aus der vorliegenden Arbeit Anknüpfungspunkte, die eine Orientierung für den Aufbau semiexperimenteller Settings geben könnten.
4. Sich auf die Multimodalität zu fokussieren könnte allerdings auch im Gegenteil bedeuten, sich einer (video-basierten) **Analyse des bewohnten Raums** zuzuwenden, in den VUIs eingebunden werden (siehe etwa Kesselheim 2016; Hausendorf/Schmitt 2016b). Diesem Aspekt wurde in der vorliegenden Arbeit nur in Ansätzen nachgegangen, dabei weisen Passagen aus den Ersteinrichtungen darauf hin, dass Mobiliar und Smart Speaker in eine Beziehung zueinander treten, die sprachlich mit-konstituiert und reflektiert wird (s. o.). Die Materialität von Smart Speakern kommt dabei noch stärker zum Vorschein – etwa wenn die Nutzer Konrad und Till den zunächst auf dem Fußboden ausgepackten HomePod gemeinsam an seinen späteren Standort tragen. Metho-

dologisch dürfte dann – anders als im zuvor ausgeführten Punkt – keine Abkehr von der Erfassung der Wohnumgebung tatsächlicher Nutzer*innen erfolgen, sondern müssten vielmehr Methoden angewendet werden, die die Verhältnisse zwischen Raum und Interaktion sichtbar machen können.

Über die generierten Erkenntnisse hinaus liegt also der Mehrwert einer Studie von qualitativ-explorativem Charakter, wie sie hier durchgeführt und präsentiert wurde, darin, dass sie aufzeigen kann, wo sich eine weitere Stratifizierung der Erhebung, aber auch tiefere Detaillierung in der Auswertung zukünftiger Untersuchungen anbieten könnte. Die Erkenntnisse stehen somit nicht nur für sich, sondern erledigen in besonderer Weise auch ‚Vorarbeiten‘ für Anschlussuntersuchungen.

7.3 Ausblick

Die bis hierher ausgeführten Analysen und Schlussfolgerungen beziehen sich teilweise auf Technologien, die bereits zum Zeitpunkt der Drucklegung dieser Arbeit von neueren Technologien überholt wurden: So hat etwa Amazon einen Nachfolger für den EchoDot (EchoPop) und Geräte mit Display auf den Markt gebracht (z. B. EchoShow). Andere auf dem Markt verfügbare Smart Home-Geräte haben sich in Qualität und Quantität rasant weiterentwickelt. Wie bereits in der Einleitung erwähnt, wurde außerdem zwischenzeitlich die Wirtschaftlichkeit von Smart Speakern für die Systemanbieter in Zweifel gezogen (vgl. Kim 2022), was wiederum unterschiedliche Prognosen für die Zukunft von Smart Speakern hervorbrachte. Die vorliegende Arbeit konnte nichtsdestoweniger erhellen, wie ungewohnte Interfaces zur Gewöhnung werden und welche zentrale Rolle die sprachlichen Praktiken bei der Einbindung von VUIs in die soziale Praxis spielen. Sie konnte zeigen, wie Nutzer*innen ihr gesprächsorganisatorisches Wissen in den Dienst von „Interfacing“ (Lipp/Dickel 2022) stellen, sie konnte Konsequenzen spezifischer Eigenschaften von VUIs herausarbeiten und Situationen ihrer versprachlichten Domestizierung beschreiben. Dies alles sind wichtige Schritte, um die Veralltäglichung von Technologien verständlich zu machen – und ob diese dabei in absoluten Werten ‚neu‘ sind, ist für das medienlinguistische Erkenntnisinteresse nachrangig, wenn sie zum Untersuchungszeitpunkt ‚neu‘ in die untersuchten Haushalte einziehen. Darüber hinaus kann die Beleuchtung von VUI-Dialogen im Sinne Pitschs (2015; 2023) auch als Forschungsinstrument für die Gesprächsforschung bzw. die Interaktionale Linguistik eingesetzt werden – Erkenntnisse aus der Nutzung von VUIs lassen auch Rückschlüsse über die Verwendung von Sprache in ihrem interaktionalen Gebrauch zu.

Diesen Wert der Untersuchung muss man sich vor Augen halten, wenn der Blick nun auf zukünftige Entwicklungen gerichtet werden soll, die zu großen Teilen bereits Fahrt aufgenommen haben. So ist anzunehmen, dass sich durch die bereits angekündigte Einbindung von Large Language Models (LLMs) wie *ChatGPT* in VUIs (vgl. Malik 2023) die Sprachsynthese und Dialogizität erneut signifikant verändern werden. Die Verknüpfung von LLMs mit externen Wissensdatenbanken ist ebenfalls bereits in der Entwicklung (siehe etwa Thoppilan et al. 2022), obschon sich diese Entwicklungen aktuell auf schriftbasierte Systeme beziehen (vgl. Albrecht 2023: 46) und die Systeme nicht ohne Weiteres kompatibel zu sein scheinen (vgl. Strüver 2025). Werden diese technologischen Systeme allerdings miteinander kombiniert, ist mit einer neuen Qualität von VUIs zu rechnen. Diese Entwicklungen könnten auch zu einer erneuten Wende oder jedenfalls einer Veränderung in Bezug auf die Bewertung der Berichte führen, der zufolge Smart Speaker nicht wirtschaftlich sind – erste Diskussionen und Spekulationen diesbezüglich sind in entsprechenden Magazinen und Foren bereits im Gang (vgl. etwa Poti 2023). Technologisch deutlich avanciertere Stimmssyntheseverfahren, die bereits technisch machbar sind und etwa die eigene Stimme „klonen“ können (vgl. Grünewald 2023), werden auch die Erfahrungen der Nutzer*innen verändern und die ‚Maschinenhaftigkeit‘ der ‚Personae‘ der bisher verfügbaren VUIs reduzieren. Die 2018 bei der Google I/O vorgestellte Anwendung Google Duplex hatte bereits Charakteristika menschlicher Sprache-in-Interaktion integriert, z. B. deutlich ungleichmäßigere Intonation und Sprechtempo und weitere prosodische Merkmale sowie die Produktion von Zögerungspartikeln und Rezeptionssignalen (vgl. Leviathan/Matias).³⁷² Zwar hat Google ein solches Dialog-System nicht in der Breite verfügbar gemacht, es ist allerdings anzunehmen, dass die neueren technologischen Entwicklungen im Bereich der LLM-basierten Sprachsynthese auch auf die Gestaltung von Stimmeingaben der VUIs einen Einfluss haben werden. Hier entstehen entsprechend erneut Aufgaben für eine gesprächsanalytisch arbeitende Medienlinguistik, die flexibel für die technologischen Entwicklungen bleiben muss. Eine Linguistik, die mit den Eigenschaften gesprochener Sprache und mit der Erfassung von ‚Dialogen‘ im Alltag ebenso umzugehen weiß wie mit medienwissenschaftlichen Konzepten und Debatten, ist hier gefragt, um die Einbindung KI-basierter Anwendungen in unterschiedlichste Bereiche des privaten und institutionellen Alltags erforschen zu können.

Eine weitere Entwicklung spricht dafür, dass die bisherigen Anbieter Smart Speaker nicht umgehend einstellen, aber möglicherweise modifizieren werden.

372 Die Demonstration ist auf YouTube unter <https://www.youtube.com/watch?v=D5VN56jQMWM> zu finden (zuletzt abgerufen am 22.05.2025).

Smart Speaker werden, so argumentiert Strüver (2023a; 2023b), für die Nutzer*innen zur wesentlichen Plattform, um Geräte in Smart Home-Environments einzurichten, zu konfigurieren und zu verwalten. Viel deutet demzufolge darauf hin, dass die Dienstanbieter von Sprachassistenzsystemen wie Amazon ein verändertes Ziel mit der Vermarktung von Smart Speakern verfolgen: die Etablierung dieser Geräte als zentrale Schnittstelle für Smart Home-Anwendungen sowie als Schnittpunkte für Nutzer*innen und Entwickler*innen. Dabei scheint für die Hersteller nicht entscheidend zu sein, dass das Interface stimmbasiert ist, denn bereits seit einigen Jahren sind Smart Speaker, die neben einem VUI auch ein Graphical User Interface (GUI) anbieten und beide Funktionen kombinieren, marktgängig und verbreiten sich weiter – zentral ist das Zusammenlaufen der verschiedenen Smart Home-Geräte in einer einzigen Applikation als Teil eines Ökosystems. Das „Interfacing“ (Lipp/Dickel 2022) erfolgt also über verbale, visuelle und taktile Praktiken, die linguistisch in ihrem Zusammenspiel untersucht werden müssen. Dabei sind auch die multisensorischen Praktiken im Hinblick auf die gesteuerten Smart Home-Anwendungen einzubeziehen: Neben Praktiken des Sprechens, Hörens und Berührens können auch Praktiken des Fühlens oder Schmeckens relevant sein, z. B. bei smarten Thermostaten oder ‚Smart Cooking‘ (siehe Graf 2023) – dies macht eine multisensorische Analyse der Praxis im Sinne Mondadas (2021) und die iterative Anpassung des Methodenspektrums zum Einbezug technischen ‚Sensings‘ erforderlich (siehe auch Scholz 2022; Hector et al. 2025). So schlagen Albert/Hamann (2021: 5) bereits die Verwendung von Blick oder anderen multimodalen Alternativen zur Aktivierung von VUIs vor – anstatt der Verwendung eines Aktivierungsworts zur Invokation der Interfaces. VUIs finden zudem auch in Kontexten außerhalb der Wohnumgebung Verwendung, z. B. in Autos, aber auch in öffentlichen Institutionen oder in der Industrie. Insofern scheint es lohnenswert, Interface-Praktiken auch außerhalb der privaten Wohnumgebung zu betrachten. Insbesondere die Einbindung von VUI-Dialogen in andere soziale Settings könnte die Befunde aus der vorliegenden Arbeit noch erweitern.

Öffentliche Diskurse über die zunehmende Leistungsfähigkeit von KI-Anwendungen sind zu Recht als „Hype“ bezeichnet worden (vgl. Bender/Koller 2020: 5185; Albrecht 2023: 17). Diese Diskursentwicklung wurde wiederholt von verschiedenen Seiten scharf kritisiert. So bezeichnet etwa der renommierte Medienwissenschaftler Alexander Galloway ‚Künstliche Intelligenz‘ auf seiner Webseite als „total scam“ (Galloway 2021; siehe auch Groß/Jordan 2023: 11). Die Kritik zielt im Kern darauf ab, dass sowohl die irreführende Verwendung von anthropozentrischen Begriffen wie ‚Verstehen‘ und ‚Lernen‘ (vgl. Bender/Koller 2020) als auch die wiederholte Betonung der Opazität und Unergründlichkeit der technischen Verfahren die tatsächlichen Herausforderungen verschleiern, die durch KI-basierte Anwendungen gesellschaftlich entstehen (vgl. Groß/Jordan 2023: 11). Diskurslinguistische

Analysen können die Dynamik und Folgen solcher Diskurse aufdecken (vgl. Kalwa 2025); durch medienlinguistische Untersuchungen im Alltagskontext der Nutzer*innen lassen sich diskursiv erzeugte und möglicherweise überhöhte Vorstellungen von KI-basierten Smart Home-Anwendungen zudem immer wieder einer empirischen Prüfung unterziehen. Damit können sprachwissenschaftliche Untersuchungen auch einen Ausgangspunkt sowie eine empirische Grundlage für eine kritische Betrachtung sozialer Herausforderungen durch KI-basierte Anwendungen schaffen.

Als gegenwärtige und zukünftige gesellschaftliche Herausforderungen, die sich spezifisch durch VUIs als KI-basierte Anwendungen ergeben, lassen sich unterschiedliche Punkte benennen, zu deren Bearbeitung die Angewandte Sprachwissenschaft – zusätzlich zu den oben genannten Aspekten – Beiträge leisten kann und sollte. Dazu gehören weiterhin virulente Fragen um Datenschutz, Privatsphäre und Sicherheit. Wie bereits mehrfach im Verlauf der Arbeit adressiert, entsteht durch die Verwertung der aufgezeichneten Sprachdaten ein Spannungsfeld zwischen der hintergründigen, für die Nutzer*innen überwiegend opaken Verwertung von Daten auf der einen und der Subjektautonomie der Nutzer*innen auf der anderen Seite; in diesem Spannungsfeld neigen Nutzer*innen zu einem „Privacy Cynicism“, der auf der Unausweichlichkeit der Nutzung bei gleichzeitiger Limitierung der eigenen Handlungsoptionen beruht (vgl. Ranzini/Lutz/Hoffmann 2023). Dieses Spannungsfeld betrifft die alltägliche Anwendung digitaler Technologien allgemein (vgl. Hoffmann 2023), aber Smart Speaker im Besonderen, eben weil sie als dauerhaft ‚mithörende‘ Geräte in privaten Wohnumgebungen platziert werden und Aufgezeichnetes cloudbasiert aus- und weiterverwerten. Dadurch werden die Grenzen zwischen privaten und öffentlichen Räumen herausgefordert (vgl. Lutz/Newlands 2021). Diese Prozesse stellen die Handlungsmacht der Nutzer*innen nicht nur im medienpraktischen Umgang mit VUIs (vgl. Habscheid/Hector/Hrncal 2023) infrage, sondern auch im Gesamtzusammenhang und dem steigenden Einfluss der Betreiberfirmen. Die Sprachwissenschaft kann hier z. B. durch sequenzanalytische Betrachtungen der Nutzung (vgl. Waldecker/Hector/Hoffmann 2024) und der davon systemseitig angefertigten Protokolle (vgl. Habscheid et al. 2021) sowie durch diskursanalytische Untersuchungen (vgl. Lind/Dickel 2023) interdisziplinär integrierte Beiträge leisten, die Grundlage für eine weiterführende Kritik sein können.

Die Trainingsmethoden und die Datensätze, auf denen die KI-basierten Berechnungen zur Sprachsynthese erfolgen, sind bei den derzeit gängigen Systemen in vielerlei Hinsicht undurchschaubar – insbesondere das etablierteste, von OpenAI betriebene System *ChatGPT* ist im Hinblick auf Code-Offenheit, die Zugänglichkeit der Trainingsdaten für die LLMs und andere Aspekte intransparent, was wiederum eine Reihe von Risiken mit sich bringt (vgl. Liesenfeld/Lopez/Din-

gemanse 2023). Sie sind damit anfälliger für die Wiedergabe vorurteilsbehafteter Inhalte, u. a. für die Reproduktion von Gender-Stereotypen und Gender-Binarität (vgl. Dev et al. 2021; Savoldi et al. 2021), aber auch für religiöse, rassistische und klassistische Biases (vgl. Apprich et al. 2018), die sich aus den Daten heraus ergeben und keiner ethikbasierten Kontrolle unterzogen wurden. Verbinden sich diese Formen von Verzerrungen mit dem Bias, der durch die stimmbasierte Steuerung von VUIs entsteht und die Optionen für Nutzer*innen auf opake Weise restringiert (vgl. Natale/Cooke 2021), könnten diese Effekte sich wechselseitig verstärken. Darüber hinaus liegt der Fokus in der Entwicklung von VUIs derzeit eindeutig auf den Standardvarietäten großer Sprachgemeinschaften. Sprachen kleinerer Sprachgemeinschaften und auch größere Varietäten (z. B. Schweizerdeutsch) werden durch die VUIs entweder nicht hinreichend erkannt oder synthetisiert oder beides. Damit begünstigen VUIs marginalisierende Effekte auf Sprecher*innen von Dialekten, Minderheitensprachen, L2-Sprecher*innen (vgl. Markl 2022) oder Nutzer*innen in mehrsprachigen Haushalten (vgl. Leblebici 2025) sowie mit sprachbezogenen Einschränkungen (vgl. Albert/Hamann/Stokoe 2023).

Die Intransparenz von KI führt nicht nur zu möglichen Verzerrungen im Hinblick auf die Inhalte, sondern begünstigt auch die Verschleierung ausbeuterischer Arbeitspraktiken (vgl. Liesenfeld/Lopez/Dingemanse 2023: 5). Während einerseits über den Einfluss von KI-basierten Systemen auf die Arbeitswelt im Sinne eines veränderten Arbeitens *mit* KI diskutiert wird (z. B. Kirsch-Kreinsen/Karacic 2019), stabilisiert sich andererseits gerade für VUIs ein prekärer Arbeitssektor *für* KI. Diese Arbeit, die häufig die Datenschutzproblematik noch weiter verstärkt (vgl. Fuest 2019), ist zu nicht unerheblichen Teilen gesellschaftlich ‚unsichtbar‘ – sie wird von nicht-privilegierten Bevölkerungsgruppen ausgeübt, die als ‚Clickworkers‘ einen entscheidenden Anteil am Funktionieren der entsprechenden Dienste haben (vgl. Crawford/Joler 2018: XVIII; siehe auch Kaerlein 2020: 53), weil ihre Arbeit für die Generierung und das Training der Modelle sowie den reibungslosen Ablauf der KI-basierten Anwendungen notwendig ist. Zugleich findet sie unter prekären Bedingungen statt und ist psychologisch belastend (vgl. Perrigo 2023). Ferner werden auch die Nutzer*innen zu Mitwirkenden an „Datenarbeit“ im „Capture-Kapitalismus“ (Heilmann 2015). Die erfassten Daten werden abgeschöpft und zu kommerziellen Zwecken der Systemanbieter verwertet (siehe auch Turow 2021), was allerdings gegenüber den zuerst genannten Herausforderungen vielmehr als konsequente Folge von Technologie im Dienst kapitalistischer Ordnungssysteme erscheint, wie auch Heilmann (2015: 47) herausstellt. Insgesamt manifestieren, so lässt sich konstatieren, die mit diesen Prozessen einhergehenden Verlagerungen von Arbeitslasten globale Machtverhältnisse, die den Einfluss großer Tech-Konzerne stabilisieren. Dies berührt auch Nachhaltigkeitsaspekte und Ressourcenverbrauch (vgl. Crawford/Joler 2018). Die Angewandte Sprachwissenschaft ist in diesen Bereichen, die hier nur angerissen

werden sollten, als empirische Grundlagendisziplin gefragt, etwa um Paradigmen und praktische Verfertigung des VUI-Designs sichtbar zu machen (vgl. Khemani/Reeves 2022) und Kritik untermauern zu können.

Es ist offensichtlich, dass diesen Herausforderungen nur interdisziplinär begegnet werden kann – sowohl empirisch als auch theoretisch sind fachübergreifende Kooperationen notwendig, auch um disziplinäre Erkenntnisinteressen befriedigen zu können (siehe auch Hepp/Loosen 2023). Die Linguistik sollte dabei den Anspruch haben, gesellschaftlich relevante Ergebnisse zur Einbindung von KI-Anwendungen im Alltag zu liefern. Die vergleichsweise hohe methodologische Festigung der Linguistik in unterschiedlichen Bereichen (etwa der Gesprächs-, Diskurs- und Medienlinguistik) und das fundierte Empirieverständnis sind Ressourcen des Fachs, die unbedingt auch in den Dienst kultur- und medienwissenschaftlicher, mediensoziologischer und anderer Erkenntnisinteressen gestellt werden sollten. So kann die Linguistik neben einer fundierenden auch eine integrierende Funktion erfüllen, was angesichts der gleichzeitig stattfindenden Ausdifferenzierung von Subdisziplinen in den Geistes- und Sozialwissenschaften dringend geboten zu sein scheint.

8 Literatur

- Ägel, Vilmos und Mathilde Hennig (Hrsg.) (2006): Zugänge zur Grammatik der gesprochenen Sprache. Berlin u. a.: De Gruyter.
- Agha, Asif (2004): Registers of Language. In: Duranti, Alessandro (Hrsg.): A companion to linguistic anthropology. Oxford: Blackwell, 23–45.
- Agre, Philip (1997): Toward a critical technical practice: Lessons learned in trying to reform AI. In: Bowker, Geoffrey et al. (Hrsg.): Social Science, Technical Systems, and Cooperative Work. Beyond the Great Divide. Mahwah: Erlbaum, 167–195.
- Agre, Philip und David Chapman (1987): What are plans for? In: Maes, Pattie (Hrsg.): Designing autonomous agents: Theory and practice from biology to engineering and back. Cambridge: MIT-Press, 17–34.
- Aiello, Luca et al. (2012): People Are Strange When You're a Stranger: Impact and Influence of Bots on Social Networks. In: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media. Palo Alto: The AAAI Press, 10–17.
- Alač, Morana et al. (2020): Talking to a Toaster. How Everyday Interactions with Digital Voice Assistants Resist a Return to the Individual. In: Evental Aesthetics 9 (1), 3–53.
- Alaofi, Marwah et al. (2022): Where Do Queries Come From? In: Amigo, Enrique et al. (Hrsg.): Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2850–2862.
- Albert, Saul und Magnus Hamann (2021): Putting wake words to bed. We speak wake words with systematically varied prosody, but CUIs don't listen. In: CUI 2021 – 3rd Conference on Conversational User Interfaces. New York: ACM Press, 1–5.
- Albert, Saul, Magnus Hamann und Elizabeth Stokoe (2023): Conversational User Interfaces in Smart Homecare Interactions: A Conversation Analytic Case Study. In: Lee, Minha et al. (Hrsg.): Proceedings of the 5th International Conference on Conversational User Interfaces. New York: ACM Press, 1–12.
- Albrecht, Steffen (2023): ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen. Berlin: Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag.
- Albrecht, Uwe (2020): Digitale Welt für Einsteiger: Alexa und Amazon Echo. Berlin: Stiftung Warentest.
- Alm, Maria (2007): Also darüber lässt sich ja streiten! Die Analyse von ‚also‘ in der Diskussion zu Diskurs- und Modalpartikeln. Stockholm: Almqvist & Wiksell.
- Amadeo, Ron (2022): Amazon Alexa is a „colossal failure“, on pace to lose \$10 billion this year. Layoffs reportedly hit the Alexa team hard as the company's biggest money loser, 21.11.2022. URL: <https://arstechnica.com/gadgets/2022/11/amazon-alexa-is-a-colossal-failure-on-pace-to-lose-10-billion-this-year/> (21.09.2023).
- Amann, Klaus und Stefan Hirschauer (1997): Die Befremdung der eigenen Kultur. Ein Programm. In: Hirschauer, Stefan und Klaus Amann (Hrsg.): Die Befremdung der eigenen Kultur. Zur ethnographischen Herausforderung soziologischer Empirie. Frankfurt a. M.: Suhrkamp, 7–52.
- Amazon (2021): Häufig gestellte Fragen zu Alexa und Alexa-Geräten. URL: <https://www.amazon.de/gp/help/customer/display.html?nodeId=201602230> (21.09.2023).
- Amazon (2023): Alexa Design Guide. URL: <https://developer.amazon.com/de-DE/alexa/alexa-haus> (21.09.2023).

- Amazon Science (2023): Research Area Conversational AI / Natural-language processing. URL: <https://www.amazon.science/research-areas/conversational-ai-natural-language-processing> (04.09.2023).
- Amendt, Gerhard (1995): *Du oder Sie. 1945 – 1968 – 1995*. Bremen: Ikaru.
- Ammari, Tawfiq et al. (2019): Music, Search, and IoT. How People (Really) Use Voice Assistants. In: *ACM Transactions on Computer-Human Interaction* 26 (3), 1–28.
- Androutsopoulos, Jannis und Florian Busch (Hrsg.) (2020): *Register des Graphischen. Variation, Interaktion und Reflexion in der digitalen Schriftlichkeit*. Berlin u. a.: De Gruyter.
- Apple (2023): HomePod oder HomePod mini als Intercom-Sprechanlage verwenden. URL: <https://support.apple.com/de-de/HT206149> (21.09.2023).
- Apprigh, Clemens et al. (2018): *Pattern Discrimination*. Minneapolis/London: University of Minnesota Press.
- Apthorpe, Noah et al. (2017), *Spying on the Smart Home: Privacy Attacks and Defenses on Encrypted IoT Traffic*. URL: <http://arxiv.org/pdf/1708.05044v1> (07.03.2025).
- Arminen, Ilkka (2005): *Institutional Interaction. Studies of Talk at Walk*. Ashgate: Hants.
- Ashworth, Boone (2022): Alexa Might Not Get Much Smarter Than It Is Right Now. In: *Wired*, 12.11.2022. URL: <https://www.wired.com/story/amazon-alexa-reportedly-losing-money/> (21.03.2023).
- Ataizi, Murat (2012): *Situated Cognition*. In: Seel, Norbert (Hrsg.): *Encyclopedia of the Sciences of Learning*. Boston: Springer, 3082–3084.
- Auer, Peter (1986): Kontextualisierung. In: *Studium Linguistik* 19, 22–47.
- Auer, Peter (1997): Formen und Funktionen der Vor-Vorfeldbesetzung im gesprochenen Deutsch. In: Schlobinski, Peter (Hrsg.): *Syntax des gesprochenen Deutsch*. Wiesbaden: VS Verlag für Sozialwissenschaften, 55–91.
- Auer, Peter (2000): On line-Syntax – Oder: was es bedeuten könnte, die Zeitlichkeit der mündlichen Sprache ernst zu nehmen. In: *Sprache und Literatur* 31 (1), 43–56.
- Auer, Peter (2012): Sprachliche Heterogenität im Deutschen. *Linguistik zwischen Variation, Varietäten und Stil*. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 42 (166), 7–28.
- Auer, Peter (Hrsg.) (2013): *Sprachliche Interaktion. Eine Einführung anhand von 22 Klassikern*. Berlin u. a.: De Gruyter.
- Auer, Peter (2016): *Dialogus in dialogum*. In: *Zeitschrift für germanistische Linguistik* 44 (3), 357–368.
- Auer, Peter (2017): Epilogue. Imperatives – The language of immediate action. In: Sorjonen, Marja-Leena, Liisa Raevaara und Elizabeth Couper-Kuhlen (Hrsg.): *Imperative Turns at Talk: The Design of Directives in Action*. Amsterdam: John Benjamins, 411–423.
- Auer, Peter (2018): Gaze, addressee selection and turn-taking in three-party interaction. In: Brône, Geert (Hrsg.): *Eye-tracking in interaction. Studies on the role of eye gaze in dialogue*. Amsterdam: John Benjamins, 197–232.
- Auer, Peter (2020): Die Struktur von Redebeiträgen und die Organisation des Sprecherwechsels. In: Birkner, Karin et al. (Hrsg.): *Einführung in die Konversationsanalyse*. Berlin/Boston: De Gruyter, 106–235.
- Auer, Peter und Susanne Günthner (2003): Die Entstehung von Diskursmarkern im Deutschen – ein Fall von Grammatikalisierung? In: *Interaction and Linguistic Structures* 38.
- Aune, Margrethe (2007): Energy comes home. In: *Energy Policy* 35 (11), 5457–5465.
- Ayaß, Ruth (1993): Auf der Suche nach dem verlorenen Zuschauer. In: Holly, Werner und Ulrich Püschel (Hrsg.): *Medienrezeption als Aneignung. Methoden und Perspektiven qualitativer Medienforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften, 27–41.

- Ayaß, Ruth (2001): Werbespots. In: Holly, Werner, Ulrich Püschel und Jörg Bergmann (Hrsg.): *Der sprechende Zuschauer. Wie wir uns Fernsehen kommunikativ aneignen*. Wiesbaden: Westdeutscher Verlag, 201–225.
- Ayaß, Ruth (2012): Introduction. Media appropriation and everyday life. In: Ayaß Ruth und Cornelia Gerhardt (Hrsg.): *The Appropriation of Media in Everyday Life*. Amsterdam: John Benjamins, 1–16.
- Ayaß, Ruth (2016): Medienethnografie. In: Auerbeck-Lietz, Stefanie und Michael Meyen (Hrsg.): *Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft*. Wiesbaden: Springer VS, 335–346.
- Bachtin, Michail (1981): *The Dialogic Imagination. Four Essays*. Edited by Michael Holquist. Austin: University of Texas Press.
- Bachtin, Michail (1986): The Problem of Speech Genres. In: Bachtin, Michail (Hrsg.): *Speech Genres and Other Late Essays*. Austin: University of Texas Press, 60–102.
- Bakardjieva, Maria (2005): *Internet society: The Internet in everyday life*. London: Sage.
- Bakardjieva, Maria (2006): Domestication running wild. From the moral economy of the household to the mores of a culture. In: Berker Thomas et al. (Hrsg.): *Domestication of Media and Technology*. Maidenhead: Open University Press, 62–79.
- Baldauf, Heike (2001): Strukturen und Formen des fernsehbegleitenden Sprechens. In: Holly, Werner, Ulrich Püschel und Jörg Bergmann (Hrsg.): *Der sprechende Zuschauer. Wie wir uns Fernsehen kommunikativ aneignen*. Wiesbaden: Westdeutscher Verlag, 61–82.
- Baldauf, Heike (2002): *Knappes Sprechen*. Tübingen: Niemeyer.
- Bangerter, Adrian, Herbert Clark und Anna Katz (2004): Navigating Joint Projects in Telephone Conversations. In: *Discourse Processes* 37 (1), 1–23.
- Baranovska, Marianna und Stefan Höltgen (Hrsg.) (2018a): *Hello, I'm Eliza. Fünfzig Jahre Gespräche mit Computern*. Bochum: projektverlag.
- Baranovska, Marianna und Stefan Höltgen (2018b): Vorwort: TELL ME YOUR PROBLEM. In: Baranovska, Marianna und Stefan Höltgen (Hrsg.): *Hello, I'm Eliza. Fünfzig Jahre Gespräche mit Computern*. Bochum: projektverlag, 9–20.
- Barthel, Mathias, Henrike Helmer und Silke Reineke (2023): First users' interactions with voice-controlled virtual assistants: A micro-longitudinal corpus study. In: *SemDial Proceedings (Pre-Publication)*, 2023. URL: <https://mezzanine.um.si/en/conference/semDial-2023-marilogue/#proceedings> (21.03.2025).
- Barth-Weingarten, Dagmar (2008): Interactional Linguistics. In: Antos, Gerd und Eija Ventola (Hrsg.): *Handbook of Interpersonal Communication*. Berlin u. a.: De Gruyter, 77–106.
- Bassett, Caroline (2007): *The arc and the machine. Narrative and new media*. Manchester: Manchester University Press.
- Bauer, Angelika (2020): Reparaturen. In: Birkner, Karin et al. (Hrsg.): *Einführung in die Konversationsanalyse*. Berlin/Boston: De Gruyter, 331–414.
- Bauer, Nathalie (2016): Onymische Anredeformen in computervermittelter Kommunikation – zur vokativen Verwendung von Rufnamen in WhatsApp-Interaktionen. In: *StudSpIn – Studentische Arbeitspapierreihe „Sprache und Interaktion“* 26.
- Baum, Patrick (2018): Die Tücke im System. „Starke“ KI im Science-Fiction-Film. In: Baranovska, Marianna und Stefan Höltgen (Hrsg.): *Hello, I'm Eliza. Fünfzig Jahre Gespräche mit Computern*. Bochum: projektverlag, 199–210.
- Baumann, Melanie (2023): Cheat-Code für den Alexa Super Modus. Amazon Alexa Cheat Code anwenden – so geht's. In: *Home & Smart*, 23.01.2023. URL: <https://www.homeandsmart.de/alexa-cheat-code> (21.03.2025).

- Bausinger, Hermann (1984): Media, technology and daily life. In: *Media, Culture and Society* 6, 343–351.
- Bayer, Klaus (1979): Die Anredepronomen Du und Sie – Thesen zu einem semantischen Konflikt im Hochschulbereich. In: *Deutsche Sprache* 7, 212–219.
- Beach, Wayne (1993): Transitional regularities for ‚casual‘ „Okay“ usages. In: *Journal of Pragmatics* 19 (4), 325–352.
- Bedford-Strohm, Jonas (2017): Voice First? Eine Analyse des Potentials von intelligenten Sprachassistenten am Beispiel Amazon Alexa. In: *Communicatio Socialis* 4, 485–494.
- Behagel, Otto (1927): *Von deutscher Sprache. Aufsätze, Vorträge und Plaudereien*. Lahr: Schauenburg.
- Bender, Emily und Alexander Koller (2020): Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In: Jurafsky, Dan et al. (Hrsg.): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 5185–5198.
- Beneteau, Erin et al. (2019): Communication Breakdowns Between Families and Alexa. In: Brewster, Stephen et al. (Hrsg.): *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 1–13.
- Beneteau, Erin et al. (2020a): Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics. In: Bernhaupt, Regina et al. (Hrsg.): *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 1–13.
- Beneteau, Erin et al. (2020b): Assumptions Checked: How Families Learn About and Use the Echo Dot. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4 (1), 1–23.
- Bentley, Frank et al. (2018): Understanding the Long-Term Use of Smart Speaker Assistants. In: *Proceedings of the 2018 ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. New York: ACM Press, 1–24.
- Bergmann, Jörg (1985): Flüchtigkeit und methodische Fixierung sozialer Wirklichkeit: Aufzeichnungen als Daten der interpretativen Soziologie. In: Bonß, Wolfgang und Heinz Hartmann (Hrsg.): *Entzauberte Wissenschaft: Zur Relativität und Geltung soziologischer Forschung. Sonderband 3 der Zeitschrift Soziale Welt*. Göttingen: Schwarz, 299–320.
- Bergmann, Jörg (2001): Das Konzept der Konversationsanalyse. In: Brinker, Klaus et al. (Hrsg.): *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung*. 2. Halbband. Berlin u. a.: De Gruyter, 919–927.
- Bergmann, Jörg (2007a): Ethnomethodologie. In: Flick, Uwe, Ernst Kardorff und Ines Steinke (Hrsg.): *Qualitative Forschung. Ein Handbuch*. Reinbek: Rowohlt, 118–135.
- Bergmann, Jörg (2007b): Harold Garfinkel und Harvey Sacks. In: Flick, Uwe, Ernst Kardorff und Ines Steinke (Hrsg.): *Qualitative Forschung. Ein Handbuch*. Reinbek: Rowohlt, 51–62.
- Bergmann, Jörg (2011): Nachruf Harold Garfinkel (1917–2011). In: *Zeitschrift für Soziologie* 40 (4), 227–232.
- Bergmann, Jörg und Christian Meyer (Hrsg.) (2021a): *Ethnomethodologie reloaded. Neue Werkinterpretationen und Theoriebeiträge zu Harold Garfinkels Programm*. Bielefeld: transcript.
- Bergmann, Jörg und Christina Meyer (2021b): Reflexivity, Indexicality, Accountability. Zur theoretisch-programmatischen Grundlegung der Ethnomethodologie. In: Bergmann, Jörg und Christian Meyer (Hrsg.): *Ethnomethodologie reloaded. Neue Werkinterpretationen und Theoriebeiträge zu Harold Garfinkels Programm*. Bielefeld: transcript, 37–56.
- Bergmann, Jörg und Bernd Ulmer (1993): Medienrekonstruktion als kommunikative Gattungen. In: Holly, Werner und Ulrich Püschel (Hrsg.): *Medienrezeption als Aneignung. Methoden und*

- Perspektiven qualitativer Medienforschung. Wiesbaden: VS Verlag für Sozialwissenschaften, 81–102.
- Berker, Thomas et al. (2006): Introduction. In: Berker Thomas et al. (Hrsg.): *Domestication of Media and Technology*. Maidenhead: Open University Press, 1–17.
- Besch, Werner (1996): *Duzen, Siezen, Titulieren. Zur Anrede im Deutschen heute und gestern*. Göttingen: Vandenhoeck & Ruprecht.
- Besch, Werner (2004): Anredeformen des Deutschen im geschichtlichen Wandel. In: Besch, Werner et al. (Hrsg.): *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*. 3. Teilband. Berlin/New York: De Gruyter.
- Betz, Emma (2017): Diskursmarker aus konversationsanalytischer Sicht: Prosodisch integriertes *ja* am Beginn von responsiven Turns. In: Blühdorn, Hardarik et al. (Hrsg.): *Diskursmarker im Deutschen. Reflexionen und Analysen*. Göttingen: Verlag für Gesprächsforschung, 183–205.
- Betz, Emma und Arnulf Deppermann (2021): OKAY in responding and claiming understanding. In: Betz, Emma et al. (Hrsg.): *OKAY across Languages. Toward a comparative approach to its use in talk-in-interaction*. Amsterdam: John Benjamins, 55–92.
- Betz, Emma et al. (2021) (Hrsg.): *OKAY across Languages. Toward a comparative approach to its use in talk-in-interaction*. Amsterdam: John Benjamins.
- Bezemer, Jeff et al. (2011): „Scissors, Please“: The Practical Accomplishment of Surgical Work in the Operating Theater. In: *Symbolic Interaction* 34 (3), 398–414.
- Birkner, Karin (2001): *Bewerbungsgespräche mit Ost- und Westdeutschen. Eine kommunikative Gattung in Zeiten gesellschaftlichen Wandels*. Berlin u. a.: De Gruyter.
- Birkner, Karin (2020a): Grundlegendes. In: Birkner, Karin et al. (Hrsg.): *Einführung in die Konversationsanalyse*. Berlin/Boston: De Gruyter, 3–31.
- Birkner, Karin (2020b): Sequenzstruktur. In: Birkner, Karin et al. (Hrsg.): *Einführung in die Konversationsanalyse*. Berlin/Boston: De Gruyter, 236–330.
- Birkner, Karin et al. (Hrsg.) (2020): *Einführung in die Konversationsanalyse*. Berlin/Boston: De Gruyter.
- Bitkom (2022), *Die Zukunft der Consumer Technology*. Berlin: URL: https://www.bitkom.org/sites/main/files/2022-08/220823_CT_Studie_2022.pdf (07.03.2025).
- Blattner, Meera, Denise Sumikawa und Robert Greenberg (1989): Earcons and Icons: Their Structure and Common Design Principles. In: *Human-Computer Interaction* 4 (1), 11–44.
- Bleakley, Anna et al. (2021): „Hey Guguru“: Exploring Non-English Linguistic Barriers for Wake Word Use. In: *Joint Proceedings of the ACM IUI 2021 Workshops*, April 13–17, 2021, College Station, USA, 1–5.
- Böhm, Markus (2020): Hoch, hoch, runter, runter ruhe in Frieden. In: *DER SPIEGEL*, 27.02.2020. URL: <https://www.spiegel.de/netzwelt/games/tod-des-konami-code-erfinders-hoch-hoch-runter-runter-ruhe-in-frieden-a-d408df68-f79d-48ef-99e9-f44ad7fe9aa1> (21.09.2023).
- Bohus, Dan und Eric Horvitz (2009): Models for Multiparty Engagement in Open-World Dialog. In: *Proceedings of SIGDAL 2009*, 225–234.
- Bohus, Dan und Eric Horvitz (2010), Microsoft Research Technical Report MSR-TR 2010-115. 1-10: URL: <https://www.microsoft.com/en-us/research/publication/computational-models-for-multiparty-turn-taking/> (07.03.2025).
- Borbach, Christoph (2018): Sprachspiele|Stimmsynthesen. Zur nachrichtentechnischen Genese des auditiven Pendants von ELIZA. In: Baranovska, Marianna und Stefan Hölzgen (Hrsg.): *Hello, I'm Eliza. Fünfzig Jahre Gespräche mit Computern*. Bochum: projektverlag, 177–198.
- Borbach, Christoph (2019): Navigating (through) Sound. Auditory Interfaces in Maritime Navigation Practice, 1900–1930. In: *Interface Critique Journal* 2, 17–33.

- Borbach, Christoph et al. (Hrsg.) (2025): *Akustische Interfaces. Interdisziplinäre Perspektiven auf Schnittstellen von Technologien, Sounds und Menschen*. Wiesbaden: Springer Vieweg.
- Both, Göde (2011): *Agency und Geschlecht in Mensch/Maschine-Konfigurationen am Beispiel von Virtual Personal Assistants*. Diplomarbeit, Berlin: Humboldt-Universität zu Berlin.
- Both, Göde (2012): *Better living through Siri? Arbeitersparnis, Geschlecht und Virtuelle Assistent_innen*. In: *Bulletin Texte / Zentrum für Transdisziplinäre Geschlechterstudien / Humboldt-Universität zu Berlin* 40, 123–138.
- Both, Göde (2014): *Multidimensional Gendering Processes at the Human-Computer-Interface: The Case of Siri*. In: Marsden, Nicola und Ute Kempf (Hrsg.): *Gender-UseIT. HCI, Usability und UX unter Gendergesichtspunkten*. Berlin u. a.: De Gruyter, 107–112.
- Both, Göde (2015): *Praktiken kartografieren. Was bringt Clarkes Situational Analysis für Praxeografien?* In: Schäfer, Franka, Anna Daniel und Frank Hillebrandt (Hrsg.): *Methoden einer Soziologie der Praxis*. Bielefeld: transcript, 197–214.
- Braslavsky, Emma (2019): *Ich bin dein Mensch. Ein Liebeslied*. In: Brandt, Stefan, Christian Granderath und Manfred Hattendorf (Hrsg.): 2029. *Geschichten von morgen*. Berlin: Suhrkamp, 17–86.
- Braun, Friederike (1988): *Terms of Address. Problems of patterns and usage in various languages and cultures*. Berlin u. a.: De Gruyter.
- Braun, Friederike, Armin Kohz und Klaus Schubert (1986): *Anredeforschung. Kommentierte Bibliographie zur Soziolinguistik der Anrede*. Tübingen: Narr.
- Braun-Thürmann, Holger (2002): *Künstliche Interaktion. Wie Technik zur Teilnehmerin sozialer Wirklichkeit wird*. Wiesbaden: Westdeutscher Verlag.
- Brause, Saba und Grant Blank (2020): *Externalized domestication: smart speaker assistants, networks and domestication theory*. In: *Information, Communication & Society* 23 (5), 751–763.
- Brause, Saba und Grant Blank (2023): *‘There are some things that I would never ask Alexa’ – privacy work, contextual integrity, and smart speaker assistants*. In: *Information, Communication & Society* 27(1), 182–197, 1–16.
- Breidenstein, Georg et al. (2013): *Ethnografie. Die Praxis der Feldforschung*. Konstanz: UVK.
- Brennan, Susan (1990): *Conversation as Direct Manipulation. An Iconoclastic View*. In: Laurel, Brenda (Hrsg.): *The Art of Human-Computer Interface Design*. Reading: Addison-Wesley, 393–404.
- Brinker, Klaus (2010): *Linguistische Textanalyse*. Berlin: Erich Schmidt.
- Brinker, Klaus und Sven Sager (2010): *Linguistische Gesprächsanalyse. Eine Einführung*. Berlin: Erich Schmidt.
- Brock, Alexander und Peter Schildhauer (2017): *Communication Form: A Concept Revisited*. In: Brock, Alexander und Peter Schildhauer (Hrsg.): *Communication Forms and Communicative Practices: New Perspectives on Communication Forms, Affordances and What Users Make of Them*. Frankfurt a. M.: Peter Lang.
- Brône, Geert (Hrsg.) (2018): *Eye-tracking in interaction. Studies on the role of eye gaze in dialogue*. Amsterdam: John Benjamins.
- Brown, Penelope und Stephen Levinson (1987): *Politeness. Some universals in language usage*. Cambridge: Cambridge University Press.
- Brown, Roger und Albert Gilman (1968): *The pronouns of Power and Solidarity*. In: Fishman, Joshua (Hrsg.): *Readings in the Sociology of Language*. Berlin u. a.: De Gruyter, 252–275.
- Brünner, Gisela (1987): *Kommunikation in institutionellen Lehr-Lern-Prozessen*. Tübingen: Narr.
- Bucher, Taina (2014): *About a Bot: Hoax, Fake, Performance Art*. In: *M/C Journal* 17 (3).
- Bühler, Karl (1965[1934]): *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Stuttgart: Gustav Fischer.

- Burkhardt, Marcus et al. (2022): Introduction. In: Burkhardt, Marcus et al. (Hrsg.): *Interrogating Datafication: Towards a Praxeology of Data*. Bielefeld: transcript, 9–36.
- Burrell, Jenna (2016): How the machine „thinks“: Understanding opacity in machine learning algorithms. In: *Big Data & Society* 3 (1), 1–12.
- Butler, Carly, Susan Danby und Michael Emmison (2011): Address terms in turn beginnings: managing disalignment and disaffiliation in telephone counseling. In: *Research on Language & Social Interaction* 44, 338–358.
- Canalys (2019): Canalys: Amazon smart speaker shipments crossed 10 million mark in Q3 2019. URL: <https://www.canalys.com/newsroom/worldwide-smartspeaker-Q3-2019> (21.03.2025).
- Carstensen, Kai-Uwe et al. (2010): *Computerlinguistik und Sprachtechnologie*. Heidelberg: Spektrum Akademischer Verlag.
- Cassell, Justine (2000): Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents. In: Cassell, Justine, Joseph Sullivan und Scott Prevost (Hrsg.): *Embodied conversational agents*. Cambridge: MIT-Press, 1–31.
- Cassell, Justine (2020): The ties that bind: Social interaction in conversational agents. In: *Réseaux* 220–221 (2–3), 21–45.
- Cassell, Justine et al. (2000): Human Conversation as a System Framework: Designing Embodied Conversational Agents. In: Cassell, Justine, Joseph Sullivan und Scott Prevost (Hrsg.): *Embodied conversational agents*. Cambridge: MIT-Press, 29–63.
- Cassell, Justine, Joseph Sullivan und Scott Prevost (Hrsg.) (2000): *Embodied conversational agents*. Cambridge: MIT-Press.
- Cha, Narae et al. (2020): Hello There! Is Now a Good Time to Talk? In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4 (3), 1–28.
- Chalhoub, George und Ivan Flechais (2020): „Alexa, Are You Spying on Me?“. Exploring the Effect of User Experience on the Security and Privacy of Smart Speaker Users. In: Moallem, Abbas (Hrsg.): *HCI for Cybersecurity, Privacy and Trust*. Cham: Springer, 305–325.
- Chambers, Deborah (2020): Domesticating the „Smarter Than You“ Home. Gendered Agency Scripts Embedded in Smart Home Discourses. In: *Medien & Kommunikationswissenschaft* 68 (3), 304–317.
- Charlton, Michael (1993): Methoden der Erforschung von Medienaneignungsprozessen. In: Holly, Werner und Ulrich Püschel (Hrsg.): *Medienrezeption als Aneignung. Methoden und Perspektiven qualitativer Medienforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften, 11–26.
- Charmaz, Kathy (2006): *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. London: Sage.
- Cho, Janghee und Emilee Rader (2020): The Role of Conversational Grounding in Supporting Symbiosis Between People and Digital Assistants. In: *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW1), 1–28.
- Christen, Helen (2006): *Comutter, Papi und Lebensabschnittsgefährte. Untersuchungen zum Sprachgebrauch im Kontext heutiger Formen des Zusammenlebens*. Hildesheim: Olms.
- Cirko, Lesław (2016): Phatische Signale in deutschen, englischen und polnischen Prüfungsgesprächen. In: *Sprachreport* 32 (1), 26–34.
- Clark, Herbert (2005): Coordinating with each other in a material world. In: *Discourse Studies* 7 (4–5), 507–525.
- Clark, Herbert und Kerstin Fischer (2022): Social robots as depictions of social agents. In: *Behavioral and Brain Sciences* 46, 1–33.
- Clark, Herbert und Jean Fox Tree (2002): Using *uh* and *um* in spontaneous speaking. In: *Cognition* 84, 73–111.

- Clark, Herbert und Edward Schaefer (1989): Contributing to Discourse. In: *Cognitive Science* 13, 259–294.
- Clayman, Steven (2010): Address terms in the service of other actions: The case of news interview talk. In: *Discourse & Communication* 4 (3), 161–183.
- Clayman, Steven (2012a): Address terms in the organization of turns at talk: The case of pivotal turn extensions. In: *Journal of Pragmatics* 44, 1853–1867.
- Clayman, Steven (2012b): Turn-Constructional Units and the Transition-Relevance Place. In: Sidnell, Jack und Tanya Stivers (Hrsg.): *The Handbook of Conversation Analysis*. Chichester: Blackwell, 150–166.
- Cohn, Michelle und Georgia Zellou (2021): Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments. In: *Frontiers in Communication* 6, 1–13.
- Collins, Harry (2010): *Tacit and Explicit Knowledge*. Chicago: University of Chicago Press.
- Consalvo, Mia (2007): *Cheating. Gaining advantage in videogames*. Cambridge: MIT-Press.
- Coseriu, Eugenio (1988 [1980]): Die Begriffe „Dialekt“, „Niveau“ und „Sprachstil“ und der eigentliche Sinn der Dialektologie. In: Albrecht, Jörn (Hrsg.): *Energeia und Ergon*. Band I: *Schriften von Eugenio Coseriu (1968–1987)*. Tübingen: Narr, 15–43.
- Couldry, Nick und Andreas Hepp (2021): Datafizierung. Wie digitale Medien und ihre Infrastrukturen unsere Praktiken, unser Wissen und unsere soziale Welt verändern. In: *Mittelweg* 36 30 (1), 85–101.
- Couldry, Nick und Ulises Mejias (2019): Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. In: *Television & New Media* 20 (4), 336–349.
- Coulmas, Florian (1979): *Routine im Gespräch: zur pragmatischen Fundierung der Idiomatik*. Wiesbaden: Akademische Verlagsgesellschaft Athenaion.
- Couper-Kuhlen und Margret Selting (2017): *Interactional linguistics. An introduction to language in social interaction*. Cambridge: Cambridge University Press.
- Courtois, Cédric, Pieter Verdegem, Pieter und Lieven de Marez (2013): The Triple Articulation of Media Technologies in Audiovisual Media Consumption. In: *Television & New Media* 14 (5), 421–439.
- Cramer, Florian und Matthew Fuller (2008): Interface. In: Fuller, Matthew (Hrsg.): *Software Studies: A Lexicon*. Cambridge: The MIT Press, 149–152.
- Crawford, Kate und Vladan Joler (2018): *Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources*. URL: <https://anatomyof.ai/> (21.03.2025).
- Crystal, David (2001): *Language and the Internet*. Cambridge: Cambridge University Press.
- Cyra, Katharina und Karola Pitsch (2017a): Dealing with ‚long turns‘ produced by users of an assistive system: How missing uptake and reciprocity lead to turn increments. In: *Proceedings of RO-MAN 2017. IEEE Xplore*, 329–334.
- Cyra, Katharina und Karola Pitsch (2017b): Dealing with Long Utterances. How to Interrupt the User in a Socially Acceptable Manner? In: Wrede, Britta (Hrsg.): *Proceedings of the 5th International Conference on Human Agent Interaction*. New York: ACM Press, 341–345.
- Dahlgren, Kari et al. (2021): Personalization and the Smart Home: questioning techno-hedonist imaginaries. In: *Convergence* 27 (5), 1155–1169.
- Dang-Anh, Mark (2017): Die interaktionale Konstitution einer synthetischen Protestsituation. In: Kämper, Heidrun, Jörg Kilian und Kersten Roth (Hrsg.): *Protest – Parteienschelte – Politikverdrossenheit: Politikkritik in der Demokratie*. Bremen: Hempfen, 133–149.
- Dang-Anh, Mark (2019): *Protest twittern. Eine medienlinguistische Untersuchung von Straßenprotesten*. Bielefeld: transcript.

- Dang-Anh, Mark et al. (Hrsg.) (2017): Medienpraktiken. Situieren, erforschen, reflektieren, Navigationen 17 (1). Siegen: universi – Universitätsverlag Siegen.
- Datenethikkommission der Bundesregierung (Hrsg.) (2019): Gutachten der Datenethikkommission der Bundesregierung. <https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/the-men/it-digitalpolitik/gutachten-datenethikkommission.pdf> (21.12.2024).
- Dausendschön-Gay, Ulrich, Elisabeth Gülich und Ulrich Krafft (Hrsg.) (2015a): Ko-Konstruktionen in der Interaktion. Die gemeinsame Arbeit an Äußerungen und anderen sozialen Ereignissen. Bielefeld: transcript.
- Dausendschön-Gay, Ulrich, Elisabeth Gülich und Ulrich Krafft (2015b): Zu einem Konzept von Ko-Konstruktion. In: Dausendschön-Gay, Ulrich, Elisabeth Gülich und Ulrich Krafft (Hrsg.): Ko-Konstruktionen in der Interaktion. Die gemeinsame Arbeit an Äußerungen und anderen sozialen Ereignissen. Bielefeld: transcript, 21–36.
- Day, Dennis und Johannes Wagner (2008): Ethnomethodology and Conversation Analysis. In: Antos, Gerd und Eija Ventola (Hrsg.): Handbook of Interpersonal Communication. Berlin u. a.: De Gruyter, 33–52.
- de Ruiter, Jan, Holger Mitterer und N. Enfield (2006): Projecting the End of a Speaker's Turn: A Cognitive Cornerstone of Conversation. In: Language 82 (3), 515–535.
- Deppermann, Arnulf (2000): Ethnographische Gesprächsanalyse: Zu Nutzen und Notwendigkeit von Ethnographie für die Konversationsanalyse. In: Gesprächsforschung 1, 96–124.
- Deppermann, Arnulf (2001): Gesprächsanalyse als explikative Konstruktion – Ein Plädoyer für eine reflexive ethnomethodologische Konversationsanalyse. In: Iványi, Zsuzsanna, András Kertész (Hrsg.): Gesprächsforschung. Tendenzen und Perspektiven. Frankfurt a. M. u. a.: Peter Lang, 43–73.
- Deppermann, Arnulf (2006): Deontische Infinitivkonstruktionen: Syntax, Semantik, Pragmatik und interaktionale Verwendung. In: Günthner, Susanne und Wolfgang Imo (Hrsg.): Konstruktionen in der Interaktion. Berlin/Boston: De Gruyter, 239–262.
- Deppermann, Arnulf (2008): Gespräche analysieren. Eine Einführung. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Deppermann, Arnulf (2009): Verstehensdefizit als Antwortverpflichtung: Interaktionale Eigenschaften der Modalpartikel *denn* in Fragen. In: Bückner, Jörg und Susanne Günther (Hrsg.): Grammatik im Gespräch. Konstruktionen der Selbst- und Fremdpositionierung. Berlin u. a.: De Gruyter, 23–56.
- Deppermann, Arnulf (2013): Multimodal interaction from a conversation analytic perspective. In: Journal of Pragmatics 46 (1), 1–7.
- Deppermann, Arnulf (2014): „Don't get me wrong“: Recipient design by reactive and anticipatory uses of negation to constrain an action's interpretation. In: Günthner, Susanne, Wolfgang Imo und Jörg Bückner (Hrsg.): Grammar and dialogism. Sequential, Syntactic, and Prosodic Patterns between Emergence and Sedimentation. Berlin: De Gruyter, 15–52.
- Deppermann, Arnulf (2018a): Instruction practices in German driving lessons: Differential uses of declaratives and imperatives. In: International Journal of Applied Linguistics 28 (2), 265–282.
- Deppermann, Arnulf (2018b): Sprache in der multimodalen Interaktion. In: Deppermann, Arnulf und Silke Reineke (Hrsg.): Sprache im kommunikativen, interaktiven und kulturellen Kontext. Berlin u. a.: De Gruyter, 51–85.
- Deppermann, Arnulf (2021): Imperative im Deutschen: Konstruktionen, Praktiken oder social action formats? In: Weidner, Beate et al. (Hrsg.): Verfestigungen in der Interaktion. Konstruktionen, sequenzielle Muster, kommunikative Gattungen. Berlin u. a.: De Gruyter, 195–230.

- Deppermann, Arnulf und Hardarik Blühndorn, (2013): Negation als Verfahren des Adressatenzuschnitts: Verstehenssteuerung durch Interpretationsrestriktion. In: Deutsche Sprache 41 (1), 6–30.
- Deppermann, Arnulf, Helmuth Feilke und Angelika Linke (2016): Sprachliche und kommunikative Praktiken: Eine Annäherung aus linguistischer Sicht. In: Deppermann, Arnulf, Helmuth Feilke und Angelika Linke (Hrsg.): Sprachliche und kommunikative Praktiken. Berlin/Boston: De Gruyter, 1–24.
- Deppermann, Arnulf und Nadine Proske (2015): Grundeinheiten der Sprache und des Sprechens. In: Dürscheid, Christa und Jan Georg Schneider (Hrsg.): Handbuch Satz, Äußerung, Schema. Berlin u. a.: De Gruyter, 17–47.
- Deppermann, Arnulf und Reinhold Schmitt (2007): Koordination. Zur Begründung eines neuen Forschungsgegenstandes. In: Schmitt, Reinhold (Hrsg.): Koordination. Analysen zur multimodalen Interaktion. Tübingen: Narr, 15–54.
- DeSouza, Darcey et al. (2021): The Gratitude Opportunity Space. In: Social Interaction. Video-based Studies of Human Sociality 4 (1).
- Dev, Sunipa et al. (2021): Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In: ArXiv-Preprint Nr. 2108.12084.
- Dickel, Sascha und Miriam Schmidt-Jüngst (2021): Gleiche Menschen, ungleiche Maschinen. Die Humandifferenzierung digitaler Assistenzsysteme und ihrer Nutzer:innen in der Werbung. In: Dizdar, Dilek et al. (Hrsg.): Humandifferenzierung. Disziplinäre Perspektiven und empirische Sondierungen. Weilerswist: Velbrück Wissenschaft, 342–367.
- Distelmeyer, Jan (2018): Drawing Connections – How Interfaces Matter. In: Interface Critique Journal 1, 22–32.
- Distelmeyer, Jan (2020): Interface II. Zur Programmatik leitender Prozesse der ‚digitalen Gegenwart‘. In: Huber, Martin, Sybille Krämer und Claus Pias (Hrsg.): Wovon sprechen wir, wenn wir von Digitalisierung sprechen? Gehalte und Revisionen zentraler Begriffe des Digitalen. Frankfurt a. M.: Universitätsbibliothek Frankfurt a. M., 59–72.
- Dittmar, Norbert (2004): Register. In: Ammon, Ulrich et al. (Hrsg.): Sociolinguistics. Berlin u. a.: De Gruyter, 216–226.
- Dittmar, Norbert (2010): Zum Verhältnis von Form und (kommunikativer) Funktion in der mündlichen Rede am Beispiel des Konnektors *also*. In: Bahlo, Nils und Norbert Dittmar (Hrsg.): Beschreibungen für gesprochenes Deutsch auf dem Prüfstand: Analysen und Perspektiven. Frankfurt a. M.: Peter Lang, 99–135.
- Dizdar, Dilek et al. (2021): Humandifferenzierung. Disziplinäre Perspektiven und transdisziplinäre Anschlüsse. In: Dizdar, Dilek et al. (Hrsg.): Humandifferenzierung. Disziplinäre Perspektiven und empirische Sondierungen. Weilerswist: Velbrück Wissenschaft, 7–31.
- Dolata, Ulrich (2019): Privatization, curation, commodification. Commercial platforms on the Internet. In: Österreichische Zeitschrift für Soziologie 44 (S1), 181–197.
- Dolata, Ulrich und Jan-Felix Schrape (2018): Kollektivität und Macht im Internet. In: Dolata, Ulrich und Jan-Felix Schrape (Hrsg.): Kollektivität und Macht im Internet. Wiesbaden: Springer Fachmedien, 1–5.
- Draude, Claude (2006): Degendering the Species? Gender Studies Encounter Virtual Humans. In: Angeli, Anonella de und Nadia Bianchi-Berthouse (Hrsg.): Proceedings AVI, Venedig 2006, 1–5.
- Drew, Paul und John Heritage (Hrsg.) (1992): Talk at work. Interaction in institutional settings. Cambridge: Cambridge University Press.
- Drösser, Christoph (2020): Wenn die Dinge mit uns reden. Von Sprachassistenten, dichtenden Computern und Social Bots. Berlin: Duden.

- Duncan, Starkey (1972): Some Signals and Rules for Taking Speaking Turns in Conversations. In: *Journal of Personality and Social Psychology* 23 (2), 283–292.
- Duranti, Alessandro (1997): *Linguistic anthropology*. Cambridge: Cambridge University Press.
- Duranti, Alessandro (Hrsg.) (2008): *A Companion to Linguistic Anthropology*. New York: John Wiley & Sons.
- Duranti, Alessandro (Hrsg.) (2009): *Linguistic Anthropology. A Reader*. New York: John Wiley & Sons.
- Dürscheid, Christa (2005): Medien, Kommunikationsformen, kommunikative Gattungen. In: *Linguistik Online* 22 (1), 1–16.
- Dwelling, Michael und Robert Prus (2012): *Einführung in die interaktionistische Ethnographie. Soziologie im Außendienst*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Egbert, Maria (2003): Die interaktionelle Relevanz einer gemeinsamen Vorgeschichte: Zur Bedeutung und Funktion von *übrigens* in deutschen Alltagsgesprächen. In: *Zeitschrift für Sprachwissenschaft* 22 (2), 189–212.
- Egbert, Maria (2009): *Der Reparatur-Mechanismus in deutschen Gesprächen*. Mannheim: Verlag für Gesprächsforschung.
- Eggers, Hans (1962): Zur Syntax der deutschen Sprache der Gegenwart. In: *Studium Generale* 15 (1), 49–59.
- Ehlich, Konrad (1979): Verwendungen der Deixis beim sprachlichen Handeln. *Linguistisch-philologische Untersuchungen zum hebräischen deiktischen System, Teil 1*. Frankfurt a. M.: Peter Lang.
- Ehlich, Konrad (1983): Deixis und Anapher. In: Rauh, Gisa (Hrsg.): *Essays on Deixis*. Tübingen: Narr, 79–98.
- Ehlich, Konrad (2007): *Sprache und sprachliches Handeln. Band 1: Pragmatik und Sprechakttheorie*. Berlin: De Gruyter.
- Ehlich, Konrad (2019): Funktionale Pragmatik – Terme, Themen und Methoden. In: Hoffmann, Ludger (Hrsg.): *Sprachwissenschaft*. Berlin/Boston: De Gruyter, 303–320.
- Ehmer, Oliver (2011): *Imagination und Animation. Die Herstellung mentaler Räume durch animierte Rede*. Berlin u. a.: De Gruyter.
- Ehrhardt, Claus (2018): Höflichkeit, Beziehungsarbeit und face. Ein Klärungsversuch. In: Gärtig, Anne-Kathrin, Roland Bauer und Matthias Heinz (Hrsg.): *Pragmatik – Diskurs – Kommunikation. Festschrift für Gudrun Held zum 65. Geburtstag*. Wien: Praesens, 58–72.
- Ekman, Paul und Wallace Friesen (1969): The repertoire of nonverbal behaviour: Categories, origins, usage, and coding. In: *Semiotica* 1, 49–98.
- Ekstedt, Erik und Gabriel Skantze (2022): How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In: *Association for Computational Linguistics (Hrsg.): Proceedings of the SIGdial 2022 Conference*, 541–551.
- Ellerbrock, Dagmar et al. (2017): Invektivität – Perspektiven eines neuen Forschungsprogramms in den Kultur- und Sozialwissenschaften. In: *Kulturwissenschaftliche Zeitschrift* 2 (1), 2–24.
- Englert, Kathrin, Dagmar Hoffmann und David Waldecker (2022): „Tut mir Leid, ich verstehe nicht ganz“. Smart Speaker als vermeintliche Gesprächspartner*innen. In: *merz – Zeitschrift für Medienpädagogik* 66 (2), 24–34.
- Ernst, Christoph (2017): Implizites Wissen, Kognition und die Praxistheorie des Interfaces. In: *Navigationen* 17 (2), 99–116.
- Ernst, Christoph und Thomas Bächle (2020): *Interface*. In: Heßler, Martina und Kevin Liggieri (Hrsg.): *Technikanthropologie. Handbuch für Wissenschaft und Studium*. Baden-Baden: Nomos, 416–420.

- Ernst, Christoph, Oliver Ruf, Jens Schröter und Sabine Wirth (i. V.): *Interface Studies. Eine Einführung*. Tübingen: Narr.
- Ervin-Tripp, Susan (1972): Sociolinguistic rules of address. In: Pride, John und Janet Holmes (Hrsg.): *Sociolinguistics. Selected Readings*. Harmondsworth: Penguin Books, 225–240.
- Etzrodt, Katrin (2022): The third party will make a difference – A study on the impact of dyadic and triadic social situations on the relationship with a voice-based personal agent. In: *International Journal of Human–Computer Studies* 168, 1–16.
- Faber, Marlene (2001): Medienrezeption als Aneignung. In: Holly, Werner, Ulrich Püschel und Jörg Bergmann (Hrsg.): *Der sprechende Zuschauer. Wie wir uns Fernsehen kommunikativ aneignen*. Wiesbaden: Westdeutscher Verlag, 25–40.
- Farø, Ken (2015): Feste Wortgruppen/Phraseologie II: Phraseme. In: Haß, Ulrike und Petra Storjohann (Hrsg.): *Handbuch Wort und Wortschatz*. Berlin u. a.: De Gruyter, 226–247.
- Fasold, Ralph (1990): *The sociolinguistics of language*. Oxford: Blackwell.
- Feenberg, Andrew (1999): *Questioning Technology*. London: Routledge.
- Felderer, Brigitte, Ernst Strouhal (2004): *Kempelen – zwei Maschinen. Texte, Bilder und Modelle zur Sprechmaschine und zum schachspielenden Androiden Wolfgang von Kempelens*. Wien: Sonderzahl.
- Fensel, Dieter/Şimşek, Umutcan/Angele, Kevin/Huaman, Elwin/Kärle, Elias/Panasiuk, Oleksandra/Toma, Ioan/Umbrich, Jürgen/Wahler, Alexander (2020): *Knowledge Graphs. Methodology, Tools and Selected Use Cases*. Basel: Springer.
- Ferguson, Charles (1975): Toward a Characterization of English Foreigner Talk. In: *Anthropological Linguistics* 17 (1), 1–14.
- Ferguson, Charles (1982): Simplified registers and linguistic theory. In: Obler, Loraine und Lisa Menn (Hrsg.): *Exceptional Language and Linguistics*. New York: Academic Press, 49–66.
- Fiehler, Reinhard (1993): Spezifika der Kommunikation in Kooperationen. In: Schröder, Hartmut (Hrsg.): *Fachtextpragmatik*. Tübingen: Narr, 343–357.
- Fiehler, Reinhard et al. (2004): *Eigenschaften gesprochener Sprache*. Tübingen: Narr.
- Finkbeiner, Rita (2018): Deixis und Anapher. In: Liedtke, Frank und Astrid Tuchen (Hrsg.): *Handbuch Pragmatik*. Stuttgart: Metzler, 186–197.
- Fischer, Kerstin (2006): *What Computer Talk Is and Isn't: Human-Computer Conversation as Intercultural Communication*. Saarbrücken: AQ-Verlag.
- Fischer, Kerstin (2011): How People Talk with Robots: Designing Dialogue to Reduce User Uncertainty. In: *AI MAGAZINE* 32 (4), 31–38.
- Fischer, Kerstin et al. (2011): Is Talking to a Simulated Robot like Talking to a Child? In: *Proceedings of IEEE ICDL-EPIROB 2011*, Frankfurt, 1–6.
- Foley, William (1997): *Anthropological linguistics. An introduction*. Malden: Blackwell.
- Ford, Cecilia und Sandra Thompson (1996): Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In: Ochs, Elinor, Emanuel Schegloff und Sandra Thompson (Hrsg.): *Interaction and grammar*. Cambridge/New York: Cambridge University Press, 134–184.
- Ford, Marcia und William Palmer (2019): Alexa, are you listening to me? An analysis of Alexa voice service network traffic. In: *Personal and Ubiquitous Computing* 23 (1), 67–79.
- Freeman-Mills, Max (2021): Amazon Echo Dot 4. Generation vs. Echo Dot 3. Generation: Was sind die Hauptunterschiede? In: *Pocket-Lint*, 14.06.2021. URL: <https://www.pocket-lint.com/de-de/lautsprecher/kaufberatung/amazon/153968-amazon-echo-dot-4-generation-vs-3-generation-lautsprecher> (21.03.2025).

- Frehmann, Katharina (2023): „Buckets of rain!“ – Effects of colloquial and formal speech style of a voice assistant on humanness, competence, trust, and intentions to use. In: Ziegele, Marc, Anna Kümpel und Tobias Dienlin (Hrsg.): Beiträge zur Jahrestagung der Fachgruppe Rezeptions- und Wirkungsforschung 2022. Düsseldorf: Deutsche Gesellschaft für Publizistik- und Kommunikationswissenschaft e.V., 41–57.
- Frey, Christoph (1996): Die Bedeutung von parasozialen Beziehungen als Grundlage zur Verarbeitung medialer Stimuli. In: Vorderer, Peter (Hrsg.): Fernsehen als „Beziehungskiste“. Parasoziale Beziehungen und Interaktionen mit TV-Personen. Opladen: Westdeutscher Verlag, 145–152.
- Fricke, Ellen (2012): Grammatik multimodal. Wie Wörter und Gesten zusammenwirken. Berlin u. a.: De Gruyter.
- Fricke, Ellen (2013): Deixis, gesture, and embodiment from a linguistic point of view. In: Müller, Cornelia et al. (Hrsg.): Body – Language – Communication. An International Handbook on Multimodality in Human Interaction. Zweiter Halbband. Berlin u. a.: De Gruyter, 1803–1823.
- Fricke, Ellen (2015): Grammatik und Multimodalität. In: Dürscheid, Christa und Jan Georg Schneider (Hrsg.): Handbuch Satz, Äußerung, Schema. Berlin u. a.: De Gruyter, 48–76.
- Friedrich, Jürgen (1980): Soziologie und Kybernetik. Zum Verhältnis von Naturwissenschaften, Sozialwissenschaften und Systemwissenschaften. Frankfurt a. M.: Peter Lang.
- Fritz, Gerd und Franz Hundsniß (1994a): Einleitung. In: Fritz, Gerd und Franz Hundsniß (Hrsg.): Handbuch der Dialoganalyse. Berlin: De Gruyter, IX–XV.
- Fritz, Gerd und Franz Hundsniß (Hrsg.) (1994b): Handbuch der Dialoganalyse. Berlin: De Gruyter.
- Fuest, Benedikt (2019): Amazon schickt Alexa-Gespräche an Heimarbeiter in Polen. In: WELT am Sonntag, 05.08.2019. URL: <https://www.welt.de/wirtschaft/article197916397/Alexa-Amazon-laesst-Mitschnitte-von-Zeitarbeitern-in-Polen-auswerten.html> (21.03.2025).
- Galloway, Alexander (2012): The Interface Effect. Cambridge: Polity Press.
- Galloway, Alexander (2021): Questions. Answers. URL: <http://cultureandcommunication.org/galloway/questions-answers> (21.03.2025).
- Garfinkel, Harold (1967): Studies in Ethnomethodology. Cambridge: Polity Press.
- Garfinkel, Harold (2002): Ethnomethodology's Program. Working Out Durkheim's Aporism. Edited and Introduced by Anne Warfield Rawls. Lanham: Roman & Littlefield.
- Garfinkel, Harold (2020[1967]): Studien zur Ethnomethodologie. Herausgegeben von Erhard Schüttelpe, Anne Warfield Rawls und Tristan Thielmann. Frankfurt/New York: Campus-Verlag.
- Garfinkel, Harold und Harvey Sacks (1979): Über formale Strukturen praktischer Handlungen. In: Weingarten, Elmar, Fritz Sack und Jim Schenkein (Hrsg.): Ethnomethodologie. Beiträge zu einer Soziologie des Alltagshandelns. Frankfurt a. M.: Suhrkamp, 130–178.
- Garg, Radhika, Hua Cui und Yash Kapadia (2021): „Learn, Use, and (Intermittently) Abandon“: Exploring the Practices of Early Smart Speaker Adopters in Urban India. In: Proceedings of the ACM on Human-Computer Interaction 5 (CSCW2), 1–28.
- Gattringer, Karin und Marlene Handel (2021): Hohe Wertschätzung des Publikums für öffentlich-rechtliches Radio. Ergebnisse aus der ARD/ZDF-Massenkommunikation Langzeitstudie. In: Media-Perspektiven 2021 (4), 227–239.
- Gaver, William (1986): Auditory Icons: Using Sound in Computer Interfaces. In: Human-Computer Interaction 2 (2), 167–177.
- Gehle, Raphaela et al. (2015): Trouble-based group dynamics in real-world HRI — Reactions on unexpected next moves of a museum guide robot. In: 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 407–412.

- Gehle, Raphaela et al. (2017): How to Open an Interaction Between Robot and Museum Visitor? In: Mutlu, Bilge et al. (Hrsg.): *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. New York: ACM Press, 187–195.
- Gerlitz, Carolin et al. (2019): Regramming the Platform: Infrastructural Relations between Apps and Social Media. In: *Computational Culture* 7.
- Gerwinski, Jan und Erika Linz (2018): Methodik II: Beobachterparadoxon – die Aufnahmesituation im Gespräch. In: Gerwinski, Jan, Stephan Habscheid und Erika Linz (Hrsg.): *Theater im Gespräch. Sprachliche Publikumspraktiken in der Theaterpause*. Berlin u. a.: De Gruyter, 105–163.
- Gibson, James (1977): The Theory of Affordances. In: Shaw, Robert (Hrsg.): *Perceiving, acting, and knowing. Toward an ecological psychology*. Hillsdale: Erlbaum, 67–82.
- Giddens, Anthony (2003[1984]): *The constitution of society: outline of the theory of structuration*. Cambridge, MA: Polity Press.
- Gießmann, Sebastian (2018): Elemente einer Praxistheorie der Medien. In: *Zeitschrift für Medienwissenschaft* 10 (2), 95–109.
- Gillespie, Tarleton (2018): *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.
- Gleich, Uli (1996): Sind Fernsehpersonen die „Freunde“ des Zuschauers? Ein Vergleich zwischen parasozialen und realen sozialen Beziehungen. In: Vorderer, Peter (Hrsg.): *Fernsehen als „Beziehungskiste“. Parasoziale Beziehungen und Interaktionen mit TV-Personen*. Opladen: Westdeutscher Verlag, 113–144.
- Gloning, Thomas (2016): Neue mediale Formate und ihre kommunikative Nutzung in der Wissenschaft. Fallbeispiele und sieben Thesen zum Praktiken-Konzept, seiner Reichweite und seinen Konkurrenten. In: Deppermann, Arnulf, Helmuth Feilke und Angelika Linke (Hrsg.): *Sprachliche und kommunikative Praktiken*. Berlin/Boston: De Gruyter, 457–486.
- Goffman, Erving (1955): On Face-Work. An Analysis of Ritual Elements in Social Interaction. In: *Psychiatry* 18, 213–231.
- Goffman, Erving (1973): *Interaktionsrituale. Über Verhalten in direkter Kommunikation*. Frankfurt a. M.: Suhrkamp.
- Goffman, Erving (1974): *Frame Analysis. An Essay on the Organization of Experience*. Boston: Northeastern University Press.
- Goffman, Erving (1978): Response Cries. In: *Language* 54 (4), 787–815.
- Goffman, Erving (1979): Footing. In: *Semiotica* 25 (1–2), 1–30.
- Goffman, Erving (1981): *Forms of Talk*. Oxford: Blackwell.
- Goffman, Erving (1983): The Interaction Order: American Sociological Association, 1982 Presidential Address. In: *American Sociological Review* 48 (1), 1–17.
- Goksel-Canbek, Nil und Mehmet Mutlu (2016): On the track of Artificial Intelligence: Learning with Intelligent Personal Assistants. In: *International Journal of Human Sciences* 13 (1), 592–601.
- Goodwin, Charles (1979): The Interactive Construction of a Sentence in Natural Conversation. In: Psathas, George (Hrsg.): *Everyday Language: Studies in Ethnomethodology*. New York: Irvington Publishers, 97–121.
- Goodwin, Charles (1980): Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning. In: *Sociological Inquiry* 50 (3–4), 272–302.
- Goodwin, Charles (1981): *Conversational Organization. Interaction between Speakers and Hearers*. New York: Academic Press.
- Goodwin, Charles (1984): Notes on story structure and the organization of participation. In: Atkinson, J. (Hrsg.): *Structures of social action. Studies in conversation analysis*. Cambridge: Cambridge University Press, 225–246.

- Goodwin, Charles (1994): Professional Vision. In: *American Anthropologist* 96 (3), 606–633.
- Goodwin, Charles (1996): Transparent vision. In: Ochs, Elinor, Emanuel Schegloff und Sandra Thompson (Hrsg.): *Interaction and grammar*. Cambridge/New York: Cambridge University Press, 370–404.
- Goodwin, Charles (2000): Action and embodiment within situated human interaction. In: *Journal of Pragmatics* 32, 1489–1522.
- Goodwin, Charles (2002): Practices of Color Classification. In: *Mind, Culture, and Activity* 7 (1–2), 19–36.
- Goodwin, Charles (2003): Pointing as situated practice. In: Kita, Sotaro (Hrsg.): *Pointing: Where Language, Culture and Cognition Meet*. Mahwah: Lawrence Erlbaum, 217–241.
- Goodwin, Charles (2007): Environmentally coupled gestures. In: Duncan, Susan, Justine Cassell und Elena Levy (Hrsg.): *Gesture and the Dynamic Dimensions of Language*. Amsterdam u. a.: John Benjamins, 195–212.
- Goodwin, Charles (2012): Zeigegesten und kollaborative Bedeutungskonstitution in der Interaktion mit Aphasikern. In: Ayaß, Ruth und Christian Meyer (Hrsg.): *Sozialität in Slow Motion. Theoretische und empirische Perspektiven*. Wiesbaden: Springer VS, 405–418.
- Goodwin, Charles (2018): *Co-Operative Action*. Cambridge: Cambridge University Press.
- Google (2021): Google Sicherheitscenter. URL: https://safety.google/intl/de_de/ (24.03.2025).
- Google (2023): Dialogflow-Agents. URL: <https://cloud.google.com/dialogflow/cx/docs/concept/agent?hl=de> (24.03.2025).
- Google Support (2021): Informationen zu den LEDs auf dem Lautsprecher. URL: <https://support.google.com/googlenest/answer/7073219?hl=de#zippy=%2Candere-lautsprecher-die-mit-ok-google-funktionieren> (24.03.2025).
- Goulden, Murray (2019): Delete the Family: Platform Families and the Colonisation of the Smart Home. In: *Information, Communication & Society* 24 (7), 903–920.
- Graf, Eva-Maria und Thomas Spranz-Fogasy (2018): Welche Frage, wann und warum? – Eine qualitativ-linguistische Programmatik zur Erforschung von Frage-Sequenzen als zentrale Veränderungspraktik im Coaching. In: *Coaching Theorie & Praxis* 4 (1), 17–32.
- Graf, Katharina (2023): Cyborg Cook: Domestic Cooks and the Anthropology of the Smart Kitchen. In: *Digital Culture & Society* 9 (1), 49–70.
- Graff, Bernd (2016): Rassistischer Chat-Roboter: Mit falschen Werten bombardiert. In: *Süddeutsche Zeitung*, 03.04.2016. URL: <https://www.sueddeutsche.de/digital/microsoft-programm-tay-rassistischer-chat-roboter-mit-falschen-werten-bombardiert-1.2928421> (21.03.2025).
- Grammis (2023): Pausen. In: Leibniz-Institut für Deutsche Sprache (Hrsg.): *Grammatisches Informationssystem „grammis“*.
- Gray, Stacey (2016): Always On: Privacy Implications of Microphone-Enabled-Devices. In: *Future of Privacy Forum*, 09.08.2016. URL: <https://fpf.org/resource/examining-privacy-implications-smart-tv/> (24.03.2025).
- Grond, Florian und Thomas Hermann (2012): Zeitenössische Perspektiven auf ästhetische Strategien der Sonifikation. In: Schoon, Andi und Axel Volmar (Hrsg.): *Das geschulte Ohr. Eine Kulturgeschichte der Sonifikation*. Bielefeld: transcript, 47–70.
- Groß, Richard und Rita Jordan (2023): KI-Realitäten/AI Realities. In: Groß, Richard und Rita Jordan (Hrsg.): *KI-Realitäten. Modelle Praktiken und Topologien*. Bielefeld: transcript, 9–33.
- Grünewald, Isabel (2023): KI-Update: Stimmen klonen mit KI. Die eigene Stimme zu klonen, ist mit KI ein Kinderspiel geworden. Wir erklären, wie das funktioniert. In: *Heise Online*, 18.08.2023. URL: <https://www.heise.de/news/KI-Update-Stimmen-klonen-mit-KI-9224888.html> (21.03.2025).

- Gülich, Elisabeth (1970): Makrosyntax der Gliederungssignale im gesprochenen Französisch. München: Wilhelm Fink.
- Gülich, Elisabeth und Lorenza Mondada (2008): Konversationsanalyse. Eine Einführung am Beispiel des Französischen. Berlin u. a.: De Gruyter.
- Gumperz, John (1982): Discourse strategies. Cambridge: Cambridge University Press.
- Gumperz, John (1999): On interactional sociolinguistic method. In: Sarangi, Srikant, Celia Roberts (Hrsg.): Talk, work and institutional order. Discourse in medical, mediation and management settings. Berlin u. a.: De Gruyter, 453–472.
- Gumperz, John und Dell Hymes (Hrsg.) (1972): Directions in Sociolinguistics. New York u. a.: Holt, Rinehart and Winston.
- Günthner, Susanne (1997): Direkte und indirekte Rede in Alltagsgesprächen. Zur Interaktion von Syntax und Prosodie in der Redewiedergabe. In: Schlobinski, Peter (Hrsg.): Syntax des gesprochenen Deutsch. Opladen: Westdeutscher Verlag, 227–262.
- Günthner, Susanne (1999a): Beschwerdeerzählungen als narrative Hyperbeln. In: Bergmann, Jörg und Thomas Luckmann (Hrsg.): Kommunikative Konstruktion von Moral. Band 1: Struktur und Dynamik der Formen moralischer Kommunikation. Opladen: Westdeutscher Verlag, 174–205.
- Günthner, Susanne (1999b): Vorwürfe in der Alltagskommunikation. In: Bergmann, Jörg und Thomas Luckmann (Hrsg.): Kommunikative Konstruktion von Moral. Band 1: Struktur und Dynamik der Formen moralischer Kommunikation. Opladen: Westdeutscher Verlag, 206–241.
- Günthner, Susanne (2000): Zwischen direkter und indirekter Rede. Formen der Redewiedergabe in Alltagsgesprächen. In: Zeitschrift für germanistische Linguistik 28 (1), 1–22.
- Günthner, Susanne (2002): Stimmenvielfalt im Diskurs: Formen der Stilisierung und Ästhetisierung in der Redewiedergabe. In: Gesprächsforschung 3, 59–80.
- Günthner, Susanne (2008): Interactional Sociolinguistics. In: Antos, Gerd und Eija Ventola (Hrsg.): Handbook of Interpersonal Communication. Berlin u. a.: De Gruyter, 53–76.
- Günthner, Susanne (2012): ‚Geteilte Syntax‘: Kollaborativ erzeugte *dass*-Konstruktionen. In: gidi-Arbeitspapiere 43, 1–21.
- Günthner, Susanne (2015): Ko-Konstruktionen im Gespräch: Zwischen Kollaboration und Konfrontation. In: Dausendschön-Gay, Ulrich, Elisabeth Gülich und Ulrich Krafft (Hrsg.): Ko-Konstruktionen in der Interaktion. Die gemeinsame Arbeit an Äußerungen und anderen sozialen Ereignissen. Bielefeld: transcript, 55–74.
- Günthner, Susanne (2016): Praktiken erhöhter Dialogizität: onymische Anredeformen als Gesten personifizierter Zuwendung. In: Zeitschrift für germanistische Linguistik 44 (3), 406–436.
- Günthner, Susanne (2017): Die kommunikative Konstruktion von Kultur: Chinesische und deutsche Anredepraktiken im Gebrauch. In: Zeitschrift für Angewandte Linguistik 2017 (66), 1–29.
- Günthner, Susanne (2018a): Perspektiven einer sprach- und kulturvergleichenden Interaktionsforschung: Chinesische und deutsche Praktiken nominaler Selbstreferenz in SMS-, WhatsApp- und WeChat-Interaktionen. In: Gesprächsforschung 19, 478–514.
- Günthner, Susanne (2018b): Routinisierte Muster in der Interaktion. In: Deppermann, Arnulf und Silke Reineke (Hrsg.): Sprache im kommunikativen, interaktiven und kulturellen Kontext. Berlin u. a.: De Gruyter, 29–50.
- Günthner, Susanne (2020): Facetten einer Interaktionalen Onomastik: ‚Die Maus liebt dich!‘ – Onymische Selbstreferenzen in der Interaktion. In: Namenkundliche Informationen 122, 187–212.
- Günthner, Susanne und Jörg Bücker (2009): Einleitung zum Sammelband: „Grammatik im Gespräch. Konstruktionen der Selbst- und Fremdpositionierung“. In: Bücker, Jörg und Susanne Günther

- (Hrsg.): Grammatik im Gespräch. Konstruktionen der Selbst- und Fremdpositionierung. Berlin u. a.: De Gruyter, 1–19.
- Günthner, Susanne und Hubert Knoblauch (1994): ‚Forms are the food of faith‘ Gattungen als Muster kommunikativen Handelns. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 46 (4), 693–723.
- Günthner, Susanne und Angelika Linke (2006): Linguistik und Kulturanalyse – Ansichten eines symbiotischen Verhältnisses. In: Zeitschrift für germanistische Linguistik 34 (1–2), 1–27.
- Günthner, Susanne und Qiang Zhu (2015): Formen „verbaler Fellpflege“: Kosende Anredepraktiken in chinesischen und deutschen SMS-Dialogen. In: Deutsche Sprache 43 (1), 42–73.
- Günthner, Susanne und Qiang Zhu (2017): Anredeformen im Kulturvergleich. Verwandschaftsbezeichnungen als Mittel der kommunikativen Konstruktion sozialer Beziehungen in chinesischen und deutschen SMS-Interaktionen. In: Linke, Angelika und Juliane Schröter (Hrsg.): Sprache und Beziehung. Berlin u. a.: De Gruyter, 119–148.
- Gustavsson, Eva (2005): Virtual Servants: Stereotyping Female Front-Office Employees on the Internet. In: Gender, Work and Organization 12 (5), 400–419.
- Guzman, Andrea (2019): Voices in and of the machine: Source orientation toward mobile virtual assistants. In: Computers in Human Behavior 2018 (90), 343–350.
- Guzman, Andrea (2020): Ontological Boundaries between Humans and Computers and the Implications for Human-Machine Communication. In: Human-Machine Communication 1, 37–54.
- Habscheid, Stephan (2016): Handeln in Praxis. Hinter- und Untergründe situierter sprachlicher Bedeutungssituation. In: Deppermann, Arnulf, Helmuth Feilke und Angelika Linke (Hrsg.): Sprachliche und kommunikative Praktiken. Berlin, Boston: De Gruyter, 127–151.
- Habscheid, Stephan (2018a): Konversation – „Kunst im Niedergang“? In: Deppermann, Arnulf und Silke Reineke (Hrsg.): Sprache im kommunikativen, interaktiven und kulturellen Kontext. Berlin u. a.: De Gruyter, 119–142.
- Habscheid, Stephan (2018b): Konversation, Small Talk, ‚Bildungssprache‘. In: Gerwinski, Jan, Stephan Habscheid und Erika Linz (Hrsg.): Theater im Gespräch. Sprachliche Publikumspraktiken in der Theaterpause. Berlin u. a.: De Gruyter, 164–202.
- Habscheid, Stephan (2020): Alexander Brock & Peter Schildhauer (Hrsg.). 2017. Communication Forms and Communicative Practices. New Perspectives on Communication Forms, Affordances and What Users Make of Them (Language and Text Studies 15). Frankfurt a. M.: Peter Lang. 210 S. In: Zeitschrift für Rezensionen zur germanistischen Sprachwissenschaft 12 (1–2), 150–155.
- Habscheid, Stephan (2022): Socio-technical dialogue and linguistic interaction. Intelligent Personal Assistants (IPA) in the Private Home. In: Sprache und Literatur 51 (2), 167–196.
- Habscheid, Stephan (2024): Kommunikative Distanz und Nähe, Text- und Interaktionsorientierung. In: Androutsopoulos, Jannis und Friedemann Vogel (Hrsg.): Handbuch Sprache und digitale Kommunikation. Berlin/Boston: De Gruyter, 51–69.
- Habscheid, Stephan, Tim Hector und Christine Hrcal (2023): Human and non-human agency as practical accomplishment. Interactional occasions for ascription and withdrawal of (graduated) agency in the use of smart-speaker-technology. In: Social Interaction. Video-based Studies of Human Sociality 6 (1), 1–31.
- Habscheid, Stephan, Tim Hector und Christine Hrcal (2025): Linguistic Practices as a Means of Domesticating Voice-Controlled Assistance Technologies. In: Habscheid, Stephan et al. (Hrsg.): Voice Assistants in Private Homes. Media, Data and Language in Interaction and Discourse. Bielefeld: transcript, 207–239.
- Habscheid, Stephan et al. (2020): Professionelle Emotionalität und humanoide Robotik in der institutionellen Kommunikation. In: Gruber, Helmut et al. (Hrsg.): Institutionelle und

- organisationale Kommunikation. Theorie, Methodologie, Empirie und Kritik. Göttingen: Vandenhoeck & Ruprecht, 169–188.
- Habscheid, Stephan et al. (2021): Intelligente Persönliche Assistenten (IPA) mit Voice User Interfaces (VUI) als ‚Beteiligte‘ in häuslicher Alltagsinteraktion. Welchen Aufschluss geben die Protokolldaten der Assistenzsysteme? In: *Journal für Medienlinguistik* 4 (1), 16–53.
- Habscheid, Stephan et al. (Hrsg.) (2025a): *Voice Assistants in Private Homes. Media, Data and Language in Interaction and Discourse*. Bielefeld: transcript.
- Habscheid, Stephan et al. (2025b): *Voice Assistants in Private Homes. Introduction to the Volume*. In: Habscheid, Stephan et al. (Hrsg.): *Voice Assistants in Private Homes. Media, Data and Language in Interaction and Discourse*. Bielefeld: transcript, 9–29.
- Halliday, Michael und Ruqaiya Hasan (1976): *Cohesion in English*. London: Longman.
- Hartmann, Maren (2006): The triple articulation of ICTs. Media as technological objects, symbolic environments and individual texts. In: Berker Thomas et al. (Hrsg.): *Domestication of Media and Technology*. Maidenhead: Open University Press, 80–102.
- Hartmann, Maren (2009a): Everyday: Domestication of Mediatization or Mediatized Domestication? In: Lundby, Knut (Hrsg.): *Mediatization: concept, changes, consequences*. New York: Peter Lang, 225–242.
- Hartmann, Maren (2009b): Roger Stilverstone: Medienobjekte und Domestizierung. In: Hepp, Andreas, Friedrich Krotz und Tanja Thomas (Hrsg.): *Schlüsselwerke der Cultural Studies*. Wiesbaden: VS Verlag, 304–315.
- Hartmann, Maren (2013a): *Domestizierung*. Baden-Baden: Nomos.
- Hartmann, Maren (2013b): From domestication to mediated mobilism. In: *Mobile Media & Communication* 1 (1), 42–49.
- Hartmann, Maren (2019): Domestizierung, mobile Medien und anderes (un)häusliches mehr. In: Linke, Christine und Isabel Schlote (Hrsg.): *Soziales Medienhandeln*. Wiesbaden: Springer Fachmedien, 101–116.
- Hartmann, Maren (2020): (The domestication of) Nordic domestication? In: *Nordic Journal of Media Studies* 2 (1), 47–57.
- Hartung, Martin (2001): Formen der Adressiertheit der Rede. In: Brinker, Klaus et al. (Hrsg.): *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung*. 2. Halbband. Berlin u. a.: De Gruyter, 1348–1355.
- Hausendorf, Heiko (2013): On the interactive achievement of space – and its possible meanings. In: Auer, Peter et al. (Hrsg.): *Space in language and linguistics. Geographical, interactional, and cognitive perspectives*. Berlin: De Gruyter, 276–303.
- Hausendorf, Heiko (2015): Interaktionslinguistik. In: Eichinger, Ludwig (Hrsg.): *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven*. Berlin: De Gruyter, 43–70.
- Hausendorf, Heiko (2022): Over the counter. In: Christmann, Gabriela, Martina Löw und Hubert Knoblauch (Hrsg.): *Communicative Constructions and the Refiguration of Spaces*. London: Routledge, 194–224.
- Hausendorf, Heiko, Lorenza Mondada und Reinhold Schmitt (Hrsg.) (2012): *Raum als interaktive Ressource*. Tübingen: Narr Francke Attempto.
- Hausendorf, Heiko und Uta Quasthoff (1996): *Sprachentwicklung und Interaktion. Eine linguistische Studie zum Erwerb von Diskursfähigkeiten*. Opladen: Westdeutscher Verlag.
- Hausendorf, Heiko und Reinhold Schmitt (2013): *Interaktionsarchitektur und Sozialtopografie. Umrisse einer raumlinguistischen Programmatik*. In: *Arbeitspapiere des UFSP Sprache und Raum (SpuR)* 1.

- Hausendorf, Heiko und Reinhold Schmitt (2016a): Interaktionsarchitektur und Sozialtopografie. Basiskonzepte einer interaktionistischen Raumanalyse. In: Hausendorf, Heiko, Reinhold Schmitt und Wolfgang Kesselheim (Hrsg.): Interaktionsarchitektur, Sozialtopographie und Interaktionsraum. Tübingen: Narr, 27–54.
- Hausendorf, Heiko und Reinhold Schmitt (2016b): Vier Stühle vor dem Altar. Interaktionsarchitektur, Sozialtopografie und Interaktionsraum in einem „Alpha“-Gottesdienst. In: Hausendorf, Heiko, Reinhold Schmitt und Wolfgang Kesselheim (Hrsg.): Interaktionsarchitektur, Sozialtopographie und Interaktionsraum. Tübingen: Narr, 227–262.
- Haviland, John (2015): Hey! In: Topics in Cognitive Science 7 (1), 124–149.
- Hector, Tim (2022): Smart Speaker in der Praxis. Methodologische Überlegungen zur medienlinguistischen Erforschung stationärer Sprachassistenzsysteme. In: Sprache und Literatur 51 (2), 197–229.
- Hector, Tim (2025): Sprachassistenzsysteme und ihre Interfaces: eine medienlinguistische Analyse. In: Kindler-Mathôt, Clara et al. (Hrsg.): Un/Reale Interaktionsräume. Formen sozialer Ordnung im Spektrum medienpezifischer Interaktion. Bielefeld: transcript, 55–80.
- Hector, Tim und Christine Hrnca (2020): Intelligente Persönliche Assistenten im häuslichen Umfeld. Erkenntnisse aus einer linguistischen Pilotstudie zur Erhebung audiovisueller Interaktionsdaten. In: Working Paper Series Media of Cooperation 14, 1–14.
- Hector, Tim und Christine Hrnca (2024): Sprachassistenzsysteme in der Interaktion. In: Androutsopoulos, Jannis und Friedemann Vogel (Hrsg.): Handbuch Sprache und digitale Kommunikation. Berlin u. a.: De Gruyter, 309–328.
- Hector, Tim et al. (2022): The „Conditional Voice Recorder“: Data practices in the co-operative advancement and implementation of data-collection technology. In: Working Paper Series Media of Cooperation 23, 1–15.
- Hector, Tim et al. (Hrsg.) (2023): Taming Digital Practices: On the Domestication of Data-Driven Technologies. Digital Culture & Society 9 (1).
- Hector, Tim et al. (2025): Sensorische Praktiken im Smart Home. Erkenntnisse und methodische Reflexion einer interdisziplinären Pilotstudie. In: Working Paper Series Media of Cooperation 36, 1–29.
- Heilmann, Till (2015): Datenarbeit im „Capture“-Kapitalismus. Zur Ausweitung der Verwertungszone im Zeitalter informatischer Überwachung. In: Zeitschrift für Medienwissenschaft 7 (2), 35–48.
- Heilmann, Till (2019): Worin haust ein Computer? Über Seinswesen und Gehäuse universaler diskreter Maschinen. In: Bartz, Christina et al. (Hrsg.): Gehäuse: Mediale Einkapselungen. Paderborn: Wilhelm Fink, 35–52.
- Heinemann, Trine und Jakob Steensig (2017): Three imperative action formats in Danish talk-in-interaction. The case of imperative + modal particles *bare* and *lige*. In: Sorjonen, Marja-Leena et al. (Hrsg.): Imperative Turns at Talk: The Design of Directives in Action. Amsterdam: John Benjamins, 139–173.
- Heins, Rebecca et al. (1997): Turn-taking as a design principle for barge-in in Spoken Language Systems. In: International Journal of Speech Technology 2 (2), 155–164.
- Heintz, Bettina (1993): Die Herrschaft der Regel. Zur Grundlagengeschichte des Computers. Frankfurt a. M./New York: Campus-Verlag.
- Heller, Vivien (2018): Jenseits des Hier und Jetzt. Multimodale Praktiken der Versetzung in Erzählinteraktionen kleiner Kinder. In: Gesprächsforschung 19, 242–274.
- Helle-Valle, Jo und Dag Slettebakk (2008): ICTs, domestication and language-games: a Wittgensteinian approach to media uses. In: New Media & Society 10 (1), 45–66.

- Helmer, Henrike (2016): *Analepsen in der Interaktion: Semantische und sequenzielle Eigenschaften von Topik-Drop im gesprochenen Deutsch*. Heidelberg: Winter.
- Helmond, Anne et al. (2019): Facebook's evolution: development of a platform-as-infrastructure. In: *Internet Histories* 3 (2), 123–146.
- Henne, Helmut und Helmut Rehbock (2001): *Einführung in die Gesprächsanalyse*. Berlin: De Gruyter.
- Hennig, Martin und Kilian Hauptmann (2019): Alexa, optimier mich! KI-Fiktionen digitaler Assistenzsysteme in der Werbung. In: *Zeitschrift für Medienwissenschaft* 11 (21), 86–94.
- Henrich, Andreas (2008): *Information Retrieval 1. Grundlagen, Modelle Anwendungen*. Bamberg: Otto-Friedrich-Universität Bamberg, Lehrstuhl für Medieninformatik.
- Hentschel, Elke und Harald Weydt (2013): *Handbuch der deutschen Grammatik*. Berlin u. a.: De Gruyter.
- Hepp, Andreas (1998): *Fernsehaneignung und Alltagsgespräche. Fernsehnutzung aus Perspektive der Cultural Studies*. Opladen: Westdeutscher Verlag.
- Hepp, Andreas (2005): Kommunikative Aneignung. In: Mikos, Lothar und Claudia Wegener (Hrsg.): *Qualitative Medienforschung*. Konstanz: UVK, 67–79.
- Hepp, Andreas (2020): *Deep mediatization*. Abingdon/Oxon/New York: Routledge.
- Hepp, Andreas und Wiebke Loosen (2023): The Interdisciplinarity of HMC: Rethinking Communication, Media, and Agency. In: Guzman, Andrea, Rhonda McEwen und Steve Jones (Hrsg.): *The SAGE Handbook of Human-Machine Communication*. London: Sage, 12–21.
- Heritage, John (1984a): A change-of-state token and aspects of its sequential placement. In: Atkinson, J. und John Heritage (Hrsg.): *Structures of social action. Studies in conversation analysis*. Cambridge: Cambridge University Press, 299–345.
- Heritage, John (1984b): *Garfinkel and ethnomethodology*. Cambridge: Polity Press.
- Heritage, John (1988): Explanations as accounts: a conversation analytic perspective. In: Antaki, Charles (Hrsg.): *Analyzing Everyday Explanation: A Casebook of Methods*. London: Sage, 127–144.
- Herzberg, Laura (2016): *Korpuslinguistische Analyse interaktiver Einheiten: das Beispiel okay*. Masterarbeit, Mannheim: Universität Mannheim.
- Hess-Lüttich, Ernest (1993): Vom Grüßen – Rituale des Grüßens im interkulturellen Vergleich. In: Thum, Bernd und Gonthier-Louis Fink (Hrsg.): *Praxis interkultureller Germanistik. Forschung – Bildung – Politik*. München: Iudicium, 1009–1026.
- Hillebrandt, Frank (2014): *Soziologische Praxistheorien. Eine Einführung*. Wiesbaden: Springer.
- Hind, Sam und Tatjana Seitz (2021): Agre's Interactionism. In: Götz, Magdalena et al. (Hrsg.): *In the spirit of addition: taking a „Practice+“ approach to studying media*. Working Paper Series Media of Cooperation 18, 22–25.
- Hind, Sam/Seitz, Tatjana (2024): Cynical technical practice: From AI to APIs. In: *Convergence* 30 (1), 29–48.
- Hirschauer, Stefan (2004): Praktiken und ihre Körper. Über die materiellen Partizipanten des Tuns. In: Reuter, Julia und Karl Hörning (Hrsg.): *Doing Culture. Neue Positionen zum Verhältnis von Kultur und sozialer Praxis*. Bielefeld: transcript, 73–91.
- Hirschauer, Stefan (2016): Verhalten, Handeln, Interagieren. Zu den mikrosoziologischen Grundlagen der Praxistheorie. In: Schäfer, Hilmar (Hrsg.): *Praxistheorie. Ein soziologisches Forschungsprogramm*. Bielefeld: transcript, 45–67.
- Hitzler, Sarah (2013): Recipient Design in institutioneller Mehrparteieninteraktion. In: *Gesprächsforschung* 14, 110–132.
- Hochreiter, Sepp und Jürgen Schmidhuber (1997): Long Short-term Memory. In: *Neural computation* 9 (8), 1725–1780.

- Hoffmann, Dagmar (2023): Digitaler Alltag, Datenpraktiken und Subjektautonomie. In: *merz – Zeitschrift für Medienpädagogik* 67 (1), 51–59.
- Hoffmann, Ludger (1984): Berichten und Erzählen. In: Ehlich, Konrad (Hrsg.): *Erzählen in der Schule*. Tübingen: Niemeyer, 55–66.
- Holly, Werner (2001a): Beziehungsmanagement und Imagearbeit. In: Brinker, Klaus et al. (Hrsg.): *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung*. 2. Halbband. Berlin u. a.: De Gruyter, 1382–1393.
- Holly, Werner (2001b): Der sprechende Zuschauer. In: Holly, Werner, Ulrich Püschel und Jörg Bergmann (Hrsg.): *Der sprechende Zuschauer. Wie wir uns Fernsehen kommunikativ aneignen*. Wiesbaden: Westdeutscher Verlag, 11–24.
- Holly, Werner und Heike Baldauf (2001): Grundlagen des fernsehbegleitenden Sprechens. In: Holly, Werner, Ulrich Püschel und Jörg Bergmann (Hrsg.): *Der sprechende Zuschauer. Wie wir uns Fernsehen kommunikativ aneignen*. Wiesbaden: Westdeutscher Verlag, 41–60.
- Holly, Werner und Ulrich Püschel (Hrsg.) (1993a): *Medienrezeption als Aneignung. Methoden und Perspektiven qualitativer Medienforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Holly, Werner und Ulrich Püschel (1993b): Vorwort. In: Holly, Werner und Ulrich Püschel (Hrsg.): *Medienrezeption als Aneignung. Methoden und Perspektiven qualitativer Medienforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften, 7–10.
- Holly, Werner, Ulrich Püschel und Jörg Bergmann (Hrsg.) (2001): *Der sprechende Zuschauer. Wie wir uns Fernsehen kommunikativ aneignen*. Wiesbaden: Westdeutscher Verlag.
- Hookway, Branden (2014): *Interface*. Cambridge/London: The MIT Press.
- Hopper, Robert (1992): *Telephone Conversation*. Bloomington: Indiana University Press.
- Horton, Donald und Richard Wohl (1956): Mass communication and para-social Interaction. In: *Psychiatry* 19 (3), 215–229.
- Howard, Philip, Samuel Woolley und Ryan Calo (2018): Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration. In: *Journal of Information Technology & Politics* 15 (2), 81–93.
- Hoy, Matthew (2018): Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. In: *Medical Reference Services Quarterly* 37 (1), 81–88.
- Hrncal, Christine (2020): *Bewertungsinteraktionen in der Theaterpause. Eine gesprächsanalytische Untersuchung von Pausengesprächen im Theaterfoyer*. Berlin/Boston: De Gruyter.
- Hughes, John (2001): Of Ethnography, ethnomethodology and workplace Studies. In: *Ethnographic Studies* 6, 7–16.
- Huynh, Ilham (2020): *Doing Emotions. Zur multimodalen Herstellung von Emotionalität in deutschen und türkischen Alltagserzählungen*. Göttingen: Verlag für Gesprächsforschung.
- Hymes, Dell (2004): Sociolinguistics and the Ethnography of Speaking. In: Ardener, Edwin (Hrsg.): *Social Anthropology and Language*. London: Routledge, 47–93.
- Hynes, Deirdre und Els Romes (2006): ‚Fitting the internet into our lives‘: IT courses for disadvantaged users. In: Berker Thomas et al. (Hrsg.): *Domestication of Media and Technology*. Maidenhead: Open University Press, 125–144.
- Imo, Wolfgang (2009): Konstruktion oder Funktion? Erkenntnisprozessmarker (change-of-state token) im Deutschen. In: Bücker, Jörg und Susanne Günther (Hrsg.): *Grammatik im Gespräch. Konstruktionen der Selbst- und Fremdpositionierung*. Berlin u. a.: De Gruyter, 57–86.
- Imo, Wolfgang (2013): *Sprache in Interaktion. Analysemethoden und Untersuchungsfelder*. Berlin: De Gruyter.
- Imo, Wolfgang (2016): Dialogizität – eine Einführung. In: *Zeitschrift für germanistische Linguistik* 44 (3), 337–356.

- Imo, Wolfgang und Jens Lanwer (2019): Interaktionale Linguistik. Eine Einführung. Stuttgart: Metzler.
- Iqbal, Umar et al. (2022): Your Echos are Heard: Tracking, Profiling, and Ad Targeting in the Amazon Smart Speaker Ecosystem. In: ArXiv-Preprint 2204.10920.
- Jakobson, Roman (1960[2007]): Linguistik und Poetik. In: Birus, Hendrik und Sebastian Donat (Hrsg.): Poesie der Grammatik und Grammatik der Poesie. Berlin u. a.: De Gruyter, 155–216.
- Jefferson, Gail (1972): Side Sequences. In: Sudnow, David (Hrsg.): Studies in Social Interaction. New York/London: Free Press, Collier-Macmillan, 294–338.
- Jefferson, Gail (1978): Sequential aspects of storytelling in conversation. In: Schenkein, Jim (Hrsg.): Studies in the organization of conversational interaction. New York: Academic Press, 79–112.
- Jefferson, Gail (1979): A technique for inviting laughter and its subsequent acceptance/declination. In: Psathas, George (Hrsg.): Everyday Language: Studies in Ethnomethodology. New York: Irvington Publishers, 79–96.
- Jefferson, Gail (2004): Glossary of transcript symbols with an introduction. In: Lenz, Friedrich (Hrsg.): Deictic conceptualization of space, time and person. Amsterdam: John Benjamins, 13–31.
- Johansson, Martin und Gabriel Skantze (2015): Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction. In: Proceedings of SIGDIAL 2015, Prag, 305–314.
- Johnstone, Anne et al. (1994): There was a long pause: influencing turn-taking behaviour in human-human and human-computer spoken dialogues. In: International Journal of Human-Computer Studies 42 (4), 383–411.
- Jones, Mark (1996): Posthuman Agency. Between Theoretical Traditions. In: Sociological Theory 14, 290–310.
- Jung, Hyunhoon und Hyeji Kim (2019): Finding contextual meaning of the wake word. In: Cowan, Benjamin und Leigh Clark (Hrsg.): Proceedings of the 1st International Conference on Conversational User Interfaces – CUI '19. New York: ACM Press, 1–3.
- Kabel, Peter (2020): Dialog Zwischen Mensch und Maschine. Conversational User Interfaces, Intelligente Assistenten und Voice-Systeme. Wiesbaden: Springer Gabler.
- Kaerlein, Timo (2018): Social Bots und die Formalisierung von Sozialität auf Online-Plattformen. In: Hug, Theo und Günther Pallaver (Hrsg.): Talk with the Bots. Gesprächsroboter und Social Bots im Diskurs. Innsbruck: Innsbruck University Press, 17–28.
- Kaerlein, Timo (2020): *Interface*. Zur Vermittlung von Praktiken und Infrastrukturen (als Perspektive für die Medienwissenschaft). In: Huber, Martin, Sybille Krämer und Claus Pias (Hrsg.): Wovon sprechen wir, wenn wir von Digitalisierung sprechen? Gehalte und Revisionen zentraler Begriffe des Digitalen. Frankfurt a. M.: Universitätsbibliothek Frankfurt a. M., 45–58.
- Kahle, Tim und Dominik Meißner (2020): All About Voice – inkl. Arbeitshilfen online. So gestalten und vermarkten Sie Produkte oder Dienstleistungen auf digitalen Sprachassistenten. Freiburg im Breisgau: Haufe.
- Kalwa, Nina (2025): »Noch steckt KI in den Kinderschuhen«. In: Zeitschrift für Literaturwissenschaft und Linguistik (Online First), o.S.
- Kallmeyer, Werner (1995): Mannheim *do unne* und *do hauße*. In: Kallmeyer, Werner (Hrsg.): Kommunikation in der Stadt. Band 2: Ethnographien von Mannheimer Stadtteilen. Berlin u. a.: De Gruyter, 1–41.
- Kallmeyer, Werner und Imken Keim (1994): Phonologische Variation als Mittel der Symbolisierung sozialer Identität in der Filsbachwelt. In: Kallmeyer, Werner (Hrsg.): Kommunikation in der Stadt. Band 1: Exemplarische Analysen des Sprachverhaltens in Mannheim, Part 1. Berlin u. a.: De Gruyter, 141–249.

- Kamath, Uday und James Withaker (2019): *Deep Learning for NLP and Speech Recognition*. Basel: Springer International.
- Kasper, Gabriele (1984): Repair in Foreign Language Teaching. In: *Studies in Second Language Acquisition* 7 (2), 200–215.
- Kawalkowski, Blasius (2020): Amazon Echo 4 (2020) im Test: Neues Design, doch ist das schon alles? In: *Inside-digital.de*, 19.11.2020. URL: <https://www.inside-digital.de/tests/amazon-echo-4-2020-im-test-neues-design-doch-ist-das-schon-alles> (21.03.2025).
- Kehlmann, Daniel (2021): *Mein Algorithmus und ich*. Stuttgarter Zukunftsrede. Stuttgart: Klett-Cotta.
- Kendon, Adam (1967): Some Functions of Gaze-Direction in Social Interaction. In: *Acta Psychologica* 26, 22–63.
- Kendon, Adam (1972): Some relationships between body motion and speech. An analysis of an example. In: Wolfe Siegman, Aron und Benjamin Pope (Hrsg.): *Studies in Dyadic Communication*. Conference on Interview Research, University of Maryland, April 22nd and 23rd, 1968. New York: Pergamon Press, 177–210.
- Kendon, Adam (1980): Gesticulation and speech: Two aspects of the process of utterance. In: Ritchie, Key (Hrsg.): *The Relationship of Verbal and Nonverbal Communication*. Den Hag: Mouton.
- Kendon, Adam (1992): Some Recent Work from Italy on Quotable Gestures (Emblems). In: *Journal of Linguistic Anthropology* 2 (1), 92–108.
- Kendon, Adam (2004): *Gesture. Visible Action as Utterance*. Cambridge: Cambridge University Press.
- Kendon, Adam, Thomas Sebeok und Jean Umiker-Sebeok (Hrsg.) (1981): *Nonverbal Communication, Interaction and Gesture. Selections from Semiotica*. Den Hag: Mouton.
- Kendrick, Kobin (2017): Using Conversation Analysis in the Lab. In: *Research on Language & Social Interaction* 50 (1), 1–11.
- Keppler, Angela (1993): Fernsehunterhaltung aus Zuschauersicht. In: Holly, Werner und Ulrich Püschel (Hrsg.): *Medienrezeption als Aneignung. Methoden und Perspektiven qualitativer Medienforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften, 103–113.
- Keppler, Angela (1996): Interaktion ohne reales Gegenüber. Zur Wahrnehmung medialer Akteure im Fernsehen. In: Vorderer, Peter (Hrsg.): *Fernsehen als „Beziehungskiste“*. Parasoziale Beziehungen und Interaktionen mit TV-Personen. Opladen: Westdeutscher Verlag, 11–24.
- Keppler, Angela (2019): „Zeig mal“. Smartphones im Gespräch. In: Marx, Konstanze und Axel Schmidt (Hrsg.): *Interaktion und Medien. Interaktionsanalytische Zugänge zu medienvermittelter Kommunikation*. Heidelberg: Winter, 177–190.
- Képuska, Veton (2011): Wake-Up-Word Speech Recognition. In: Ipsic, Ivo (Hrsg.): *Speech Technologies*. London: InTechOpen, 237–262.
- Kernel Concepts (2022): Case-Study: Conditional Voice Recorder. URL: <https://www.kernelconcepts.de/case-study-conditional-voice-recorder/?lang=en> (21.03.2025).
- Kerssens, Niels (2017): When search engines stopped being human: menu interfaces and the rise of the ideological nature of algorithmic search. In: *Internet Histories. Digital Technology, Culture and Society* 1 (3–4), 219–237.
- Keseling, Gisbert (1989): Die Partikel *ÄH*. Ein paraverbales Element im Sprachsystem? In: Weydt, Harald (Hrsg.): *Sprechen mit Partikeln*. Berlin u. a.: De Gruyter, 575–591.
- Kesselheim, Wolfgang (2016): Videobasierte Raumforschung als interdisziplinäres Forschungsfeld. Ein exemplarischer Einblick in unterschiedliche methodische Zugänge. In: Hausendorf, Heiko, Reinhold Schmitt und Wolfgang Kesselheim (Hrsg.): *Interaktionsarchitektur, Sozialtopographie und Interaktionsraum*. Tübingen: Narr, 89–110.

- Kessler, Christine (2010): *Okay* – die Bedeutung eines Wortes zwischen Schriftlichkeit und Mündlichkeit. In: Pohl, Inge (Hrsg.): *Semantische Unbestimmtheit im Lexikon*. Frankfurt am Main: Peter Lang, 151–166.
- Khan, Lina (2018): Amazon – An Infrastructure Service and Its Challenge to Current Antitrust Law. In: Moore, Martin und Damian Tambini (Hrsg.): *Digital Dominance. The Power of Google, Amazon, Facebook, and Apple*. New York: Oxford University Press, 98–129.
- Khemani, Krishika und Stuart Reeves (2022): Unpacking Practitioners' Attitudes Towards Codifications of Design Knowledge for Voice User Interfaces. In: Barbosa, Simone et al. (Hrsg.): *CHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 1–10.
- Kim, Eugene (2022): Amazon is gutting its voice-assistant Alexa. Employees describe a division in crisis and huge losses on a ‚wasted opportunity‘. In: *Business Insider*, 22.11.2022. URL: <https://www.businessinsider.com/amazon-lose-10-billion-this-year-from-alexa-other-devices-2022-11> (21.03.2025).
- Kipp, Michael (2023): Maschinen mit künstlichen Gehirnen – Was ist Deep Learning. In: Kovács, László (Hrsg.): *Künstliche Intelligenz und menschliche Gesellschaft*. Berlin u. a.: De Gruyter, 25–40.
- Kirsch-Kreinsen, Hartmut und Anemari Karacic (Hrsg.) (2019): *Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt*. Bielefeld: transcript.
- Kita, Sotaro (2003): Pointing: A Foundational Building Block of Human Communication. In: Kita, Sotaro (Hrsg.): *Pointing. Where language, culture, and cognition meet*. London: Lawrence Erlbaum, 1–8.
- Kittler, Friedrich (1986): *Grammophon, Film, Typewriter*. Berlin: Brinkmann & Bose.
- Kittler, Friedrich (1996): Farben und/oder Maschinen denken. In: Hammel, Eckhard (Hrsg.): *Synthetische Welten. Kunst, Künstlichkeit und Kommunikationsmedien*. Essen: Verlag Die Blaue Eule, 119–132.
- Kittler, Friedrich (2002): *Optische Medien*. Berliner Vorlesung 1999. Berlin: Merve.
- Kitzinger, Celia (2012): Repair. In: Sidnell, Jack und Tanya Stivers (Hrsg.): *The Handbook of Conversation Analysis*. Chichester: Blackwell, 229–256.
- Klein, Wolfgang (1985): Ellipse: Fokusgliederung und thematischer Stand. In: Meyer-Hermann, Reinhard und Hannes Rieser (Hrsg.): *Ellipsen und fragmentarische Ausdrücke*. Tübingen: Niemeyer, 1–24.
- Klemm, Michael (2001): Sprachhandlungsmuster. In: Holly, Werner, Ulrich Püschel und Jörg Bergmann (Hrsg.): *Der sprechende Zuschauer. Wie wir uns Fernsehen kommunikativ aneignen*. Wiesbaden: Westdeutscher Verlag, 83–142.
- Knaut, Andrea (2018): Können Künstliche Neuronale Netzwerke denken? In: Hug, Theo und Günther Pallaver (Hrsg.): *Talk with the Bots. Gesprächsroboter und Social Bots im Diskurs*. Innsbruck: Innsbruck University Press, 73–86.
- Knoblauch, Hubert, Christine Leuenberger und Bernt Schnettler (Hrsg.) (2017): *Erving Goffman. Rede-Weisen*. Köln: Herbert von Halem.
- Knobloch, Clemens (2013): „Ein Teil, das fehlt, geht nie kaputt“ – Ellipsen in Grammatik und Kommunikation. In: Hennig, Mathilde (Hrsg.): *Die Ellipse. Neue Perspektiven auf ein altes Phänomen*. Berlin u. a.: De Gruyter, 19–38.
- Knorr-Cetina, Karin (1995): Laborstudien. Der kultursoziologische Ansatz in der Wissenschaftsforschung. In: Martinsen, Renate und Hartmu Esser (Hrsg.): *Das Auge der Wissenschaft. Zur Emergenz von Realität*. Baden-Baden: Nomos, 101–135.

- Koch, Peter und Wulf Oesterreicher (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In: Romanistisches Jahrbuch 36, 15–43.
- Kohne, Andreas et al. (2020): Chatbots. Aufbau und Anwendungsmöglichkeiten von autonomen Sprachassistenten. Wiesbaden: Springer Vieweg.
- Kohz, Armin (1982): Linguistische Aspekte des Anredeverhaltens. Untersuchungen am Deutschen und Schwedischen. Tübingen: Narr.
- Konerding, Klaus-Peter (2004): Semantische Variation, Diskurspragmatik, historische Entwicklung und Grammatikalisierung. Das Phänomenspektrum der Partikel *also*. In: Pohl, Inge und Klaus-Peter Konerding (Hrsg.): Stabilität und Flexibilität in der Semantik. Frankfurt a. M.: Peter Lang, 199–237.
- König, Katharina (2019): Narratives 2.0. A Multi-Dimensional Approach to Semi-Public Storytelling in WhatsApp Voice Messages. In: Journal für Medienlinguistik 2 (2), 30–59.
- König, Katharina und Florence Oloff (2019): Mobile Medienpraktiken im Spannungsfeld von Öffentlichkeit, Privatheit und Anonymität. In: Journal für Medienlinguistik 2 (2).
- Kotthoff, Helga (Hrsg.) (2006a): Scherzkommunikation. Beiträge aus der empirischen Gesprächsforschung. Radolfzell: Verlag für Gesprächsforschung.
- Kotthoff, Helga (2006b): Witzige Darbietungen als Talk-Shows. Zur konversationellen Konstruktion eines sozialen Milieus. In: Kotthoff, Helga (Hrsg.): Scherzkommunikation. Beiträge aus der empirischen Gesprächsforschung. Radolfzell: Verlag für Gesprächsforschung, 145–191.
- Krajewski, Markus (2010): Der Diener. Mediengeschichte einer Figur zwischen König und Klient. Frankfurt a. M.: S. Fischer.
- Krause, Jürgen (1992a): Fazit und Ausblick: Registermodell versus metaphorischer Gebrauch von Sprache in der Mensch-Computer-Interaktion. In: Krause, Jürgen und Ludwig Hitzenberger (Hrsg.): Computer Talk. Hildesheim/New York: Olms, 157–170.
- Krause, Jürgen (1992b): Natürlichsprachliche Mensch-Computer-Interaktion als technisierte Kommunikation: Die computer talk-Hypothese. In: Krause, Jürgen und Ludwig Hitzenberger (Hrsg.): Computer Talk. Hildesheim/New York: Olms, 1–29.
- Krause, Jürgen und Ludwig Hitzenberger (Hrsg.) (1992): Computer Talk. Hildesheim/New York: Olms.
- Kremp, Matthias (2019): Debatte um Sprachassistenten. Wie arbeitet jemand, der Siri-Aufnahmen auswertet? In: DER SPIEGEL, 08.08.2019. URL: <https://www.spiegel.de/netzwelt/gadgets/siri-transkription-so-werden-sprachbefehle-von-menschen-ausgewertet-a-1280333.html> (14.05.2025).
- Krotz, Friedrich (2022): Die Teilung geistiger Arbeit per Computer. Eine Kritik der digitalen Transformation. Weinheim: Beltz.
- Krummheuer, Antonia (2008): Zwischen den Welten. Zur Verstehenssicherung und Problembehandlung in künstlichen Interaktionen von menschlichen Akteuren und personifizierten virtuellen Agenten. In: Willems, Herbert (Hrsg.): Weltweite Welten. Internet-Figurationen aus wissenssoziologischer Perspektive. Wiesbaden: VS Verlag für Sozialwissenschaften, 269–294.
- Krummheuer, Antonia (2010): Interaktion mit virtuellen Agenten? Zur Aneignung eines ungewohnten Artefakts. Stuttgart: Lucius & Lucius.
- Kucharczik, Kerstin (1989): Sprecher- und Hörerseitige Verwendungsweisen der Interjektion „HM“. In: Kowal, Sabine und Roland Posner (Hrsg.): Zeitliche und inhaltliche Aspekte der Textproduktion. Berlin: Arbeitspapiere zur Linguistik. TU Berlin, Institut für Linguistik, 168–191.
- Kuglin, Jörg (1977): Einige Bemerkungen zur Anrede im Deutschen und Türkischen. In: Engel-Braunschmidt, Annelore und Alois Schmücker (Hrsg.): Korrespondenzen. Festschrift für Dietrich Gerhardt aus Anlaß des 65. Geburtstages. Gießen: Schmitz, 261–278.

- Kühn, Peter (1995): Mehrfachadressierung. Untersuchungen zur adressatenspezifischen Polyvalenz sprachlichen Handelns. Tübingen: Niemeyer.
- Kumar, Deepak et al. (2018): Skill squatting attacks on Amazon Alexa. In: Proceedings of the 27th USENIX Security Symposium.
- Kuzuoka, Hideaki et al. (2008): Effect of Restarts and Pauses on Achieving a State of Mutual Orientation between a Human and a Robot. In: CSCW 2008, San Diego, 201–204.
- Labov, William (1972): Some principles of linguistic methodology. In: *Language in Society* 1, 97–120.
- Lakier, Matthew und Daniel Vogel (2022): More than just Software Surprises: Purposes, Processes, and Directions for Software Application Easter Eggs. In: Proceedings of the ACM on Human-Computer Interaction 6 (CSCW1), 1–26.
- Lanwer, Jens und Georgios Coussios (2015): Kommunikative Praxis, soziale Gruppe und sprachliche Konventionen. In: Neuland, Eva und Peter Schlobinski (Hrsg.): *Handbuch Sprache in sozialen Gruppen*. Berlin u. a.: De Gruyter, 126–148.
- Latour, Bruno (1987): *Science in Action*. Cambridge: Harvard University Press.
- Latour, Bruno (1996): On Akteur-Network-Theory. A Few Clarifications. In: *Soziale Welt* 47 (4), 369–381.
- Latour, Bruno (2002): *Die Hoffnung der Pandora*. Untersuchungen zur Wirklichkeit der Wissenschaft. Frankfurt a. M.: Suhrkamp.
- Latour, Bruno (2005): *Reassembling the social. An introduction to Actor-Network-Theory*. Oxford: Oxford University Press.
- Latour, Bruno (2007): *Eine neue Soziologie für eine neue Gesellschaft. Einführung in die Akteur-Netzwerk-Theorie*. Berlin: Suhrkamp.
- Latour, Bruno (2019): Bruno Latour: Über technische Vermittlung: Philosophie, Soziologie und Genealogie (1994). In: Ziemann, Andreas (Hrsg.): *Grundlagentexte der Medienkultur. Ein Reader*. Wiesbaden: Springer VS, 425–429.
- Lau, Josephine, Benjamin Zimmerman und Florian Schaub, (2018): Alexa, Are You Listening? In: Proceedings of the ACM on Human-Computer Interaction 2, 1–31.
- Leblebici, Didem (2025): Einstimmen mit Sprachassistenten: Einblicke in monolinguales Bias in mehrsprachigen Settings. In: Vallentin, Rita und Dorothea Horst (Hrsg.): *Sprache entgrenzen. Beiträge zu einer kulturwissenschaftlichen Linguistik*. Bielefeld: transcript, 47–74.
- Lee, Sunok, Sungbae Kim und Sangsu Lee (2019): „What does your Agent look like?“. In: Brewster, Stephen et al. (Hrsg.): *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 1–6.
- Lehr, Andrea et al. (2001): Herbert Ernst Wiegand und die Erforschung des alltäglichen Umgangs mit Sprache. In: Lehr, Andrea et al. (Hrsg.): *Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik*. Berlin/New York: De Gruyter, XIII–XX.
- Lengersdorf, Diana (2015): Ethnografische Erkenntnisstrategien zur Erforschung sozialer Praktiken. In: Schäfer, Franka, Anna Daniel und Frank Hillebrandt (Hrsg.): *Methoden einer Soziologie der Praxis*. Bielefeld: transcript, 177–196.
- Lenz, Friedrich (2013): Objektdeixis. In: Schierholz, Stefan (Hrsg.): *Datenbank: Wörterbücher zur Sprach- und Kommunikationswissenschaft (WSK) Online*. Berlin u. a.: De Gruyter.
- Lenzen, Manuela (2024): *Künstliche Intelligenz. Fakten, Chancen, Risiken*. München: Beck.
- Lerner, Gene (2003): Selecting next speaker: The context-sensitive operation of a context-free organization. In: *Language in Society* 32, 177–201.
- Leviathan, Yaniv und Yossi Matias: Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. In: Google Research Blog. URL: <https://blog.research.google/2018/05/duplex-ai-system-for-natural-conversation.html> (21.03.2025).

- Levin, Harry und Deborah Gray (1983): The Lecturer's OK. In: *American Speech* 58 (3), 195–200.
- Levinson, Stephen (2000): *Pragmatik*. Tübingen: Niemeyer.
- Levinson, Stephen (2006): On the human „interactional engine“. In: Enfield, Nick und Stephen Levinson (Hrsg.): *Roots of human sociality. Culture, cognition and interaction*. Oxford: Berg, 39–69.
- Levinson, Stephen (2016): Turn-taking in Human Communication – Origins and Implications for Language Processing. In: *Trends in Cognitive Sciences* 20 (1), 6–14.
- Levinson, Stephen und Francisco Torreira (2015): Timing in turn-taking and its implications for processing models of language. In: *Frontiers in Psychology* 6, 1–17.
- Liesenfeld, Andreas, Alianda Lopez und Mark Dingemanse (2023): Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In: *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–6.
- Lind, Miriam (2021): „Alexa, 3, Sprachassistentin, hat die Religion für sich entdeckt“. Die sprachliche Anthropomorphisierung von Assistenzsystemen. In: Lind, Miriam (Hrsg.): *Mensch – Tier – Maschine. Sprachliche Praktiken an und jenseits der Außengrenze des Humanen*. Bielefeld: transcript, 347–370.
- Lind, Miriam und Sascha Dickel (2023): Speaking, but having no voice. Negotiating agency in advertisements for intelligent personal assistants. In: *Convergence* 0 (0), 1–17.
- Lindström, Jan et al. (2017): Imperatives in Swedish medical consultations. In: Sorjonen, Marja-Leena, Liisa Raevaara und Elisabeth Couper-Kuhlen (Hrsg.): *Imperative Turns at Talk: The Design of Directives in Action*. Amsterdam: John Benjamins, 299–324.
- Linell, Per (1998): *Approaching Dialogue. Talk, Interaction and Contexts in Dialogical Perspectives*. Amsterdam: John Benjamins.
- Linell, Per (2005): *The Written Language Bias in Linguistics. Its nature, origins and transformations*. London: Routledge.
- Linell, Per (2009): *Rethinking language, mind, and world dialogically. Interactional and contextual theories of human sense-making*. Charlotte, NC: Information Age Pub.
- Linell, Per (2012): Zum Begriff des kommunikativen Projekts. In: Ayaß, Ruth und Cristian Meyer (Hrsg.): *Sozialität in Slow Motion. Theoretische und empirische Perspektiven*. Wiesbaden: Springer VS, 71–79.
- Ling, Richard (2004): *The Mobile Connection. The Cell Phone's Impact on Society*. San Francisco: Elsevier Science & Technology.
- Linz, Erika (2016): Sprache, Materialität, Medialität. In: Jäger, Ludwig et al. (Hrsg.): *Sprache – Kultur – Kommunikation. Ein internationales Handbuch zu Linguistik als Kulturwissenschaft*. Berlin, Boston: De Gruyter, 100–111.
- Lipp, Benjamin und Sascha Dickel (2022): Interfacing the human/machine. In: *Distinktion: Journal of Social Theory* 23 (1), 425–443.
- Livingstone, Sonia (2007): On the material and the symbolic. Silverstone's double articulation of research traditions in new media studies. In: *New Media & Society* 9 (1), 16–24.
- Locher, Miriam und Richard Watts (2005): Politeness Theory and Relational Work. In: *Journal of Politeness Research. Language, Behaviour, Culture* 1 (1), 9–33.
- Loehr, Daniel (2007): Aspects of Rhythm in Gesture and Speech. In: *Gesture* 7 (2), 179–214.
- Loenhoff, Jens (2015): Tacit Knowledge: Shared and Embodied. In: Adloff, Frank, Katharina Gerund und David Kaldewey (Hrsg.): *Revealing Tacit Knowledge. Embodiment and Explication*. Bielefeld: transcript, 21–40.
- Loenhoff, Jens (2019): Objectification and verbalization: social constructivism and the problem of language. In: Pfadenhauer, Michaela und Hubert Knoblauch (Hrsg.): *Social Constructivism as a*

- Paradigm. *The Legacy of The Social Construction of Reality*. Abingdon/New York: Routledge, 105–117.
- Löfström, Erica (2007): *Visualizing Energy in Households: The Power Aware Cord as a Means to Create Energy Awareness*. In: *Proceedings of the European Network for Housing Research conference*. Rotterdam: ENHR.
- Loideain, Nora und Rachel Adams (2020): *From Alexa to Siri and the GDPR: The gendering of Virtual Personal Assistants and the role of Data Protection Impact Assessments*. In: *Computer Law & Security Review* 36, 1–14.
- Lopatovska, Irene et al. (2019): *Talk to me: Exploring user interactions with the Amazon Alexa*. In: *Journal of Librarianship and Information Science* 51 (4), 984–997.
- Lotze, Netaya (2016): *Chatbots. Eine linguistische Analyse*. Berlin: Peter Lang.
- Lotze, Netaya (2018): *Zur sprachlichen Interaktion mit Chatbots – Eine linguistische Perspektive*. In: Hug, Theo und Günther Pallaver (Hrsg.): *Talk with the Bots. Gesprächsroboter und Social Bots im Diskurs*. Innsbruck: Innsbruck University Press, 29–49.
- Lotze, Netaya (2020): *Künstliche Intelligenz im Dialog – Ein methodologisches Konzept zur Analyse von Mensch-Maschine-Interaktion*. In: Marx, Konstanze, Henning Lobin und Axel Schmidt (Hrsg.): *Deutsch in Sozialen Medien*. Berlin/Boston: De Gruyter, 363–367.
- Lotze, Netaya (2021): *Zur Adressierung des Unbelebten – Grenzen von pragmatischer Konzeption*. In: Lind, Miriam (Hrsg.): *Mensch – Tier – Maschine. Sprachliche Praktiken an und jenseits der Außengrenze des Humanen*. Bielefeld: transcript, 305–325.
- Lotze, Netaya (2025): *Human-Machine Interaction as a Complex Socio-Linguistic Practice*. In: Habscheid, Stephan et al. (Hrsg.): *Voice Assistants in Private Homes. Media, Data and Language in Interaction and Discourse*. Bielefeld: transcript, 105–143.
- Lotze, Netaya (i. V.): *Linguistische KI-Forschung. Eine Einführung*. Tübingen: Narr.
- Lübke, Valeska (2005): *CyberGender. Geschlecht und Körper im Internet*. Königstein: Helmer.
- Luff, Paul, Jon Hindmarsh und Christina Heath (Hrsg.) (2000): *Workplace studies. Recovering work practice and informing system design*. Cambridge: Cambridge University Press.
- Luger, Ewa und Abigail Sellen (2016): „Like Having a Really Bad PA“. In: Kaye, Jofish et al. (Hrsg.): *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 5286–5297.
- Luginbühl, Martin (2019): *Mediale Durchformung: Fernsehinteraktion und Fernsehöffentlichkeit in Gesprächen im Fernsehen*. In: Marx, Konstanze und Axel Schmidt (Hrsg.): *Interaktion und Medien*. Heidelberg: Winter, 125–146.
- Luginbühl, Martin und Dorothee Meer (2022): *Parainteraktion in den Medien – Zur Einführung*. In: Luginbühl, Martin und Dorothee Meer (Hrsg.): *Parainteraktion in den Medien. Linguistische Studien zu formen medialer Pseudo-Interaktion*. Frankfurt a. M.: Peter Lang, 7–20.
- Lutz, Christoph (2023): *Privacy and Human-Machine Communication*. In: Guzman, Andrea, Rhonda McEwen und Steve Jones (Hrsg.): *The SAGE Handbook of Human-Machine Communication*. London: Sage, 310–317.
- Lutz, Christoph und Gemma Newlands (2021): *Privacy and smart speakers: A multi-dimensional approach*. In: *The Information Society* 37 (2), 147–162.
- Lynch, Michael (1993): *Scientific Practice and Ordinary Action. Ethnomethodology and Social Studies of Science*. Cambridge: Cambridge University Press.
- Lynch, Michael (2011): *Harold Garfinkel (29 October 1917 – 21 April 2011): A remembrance and reminder*. In: *Social Studies of Science* 41 (6), 927–942.

- Macha, Jürgen (1997): Konstanz, Variation und Wandel familiärer Anredeformen. In: Macha, Hildegard und Lutz Mauermann (Hrsg.): Brennpunkte der Familienerziehung. Weinheim: Deutscher Studien Verlag, 199–218.
- Malik, Aisha (2023): Amazon is developing an improved LLM to power Alexa. In: TechCrunch, 28.04.2023. URL: <https://techcrunch.com/2023/04/28/amazon-working-improved-llm-to-power-alexa/> (21.03.2025).
- Malinowski, Bronislaw (1946[1923]): The Problem of Meaning in Primitive Languages. In: Odgen, Charles und Ivor Richards (Hrsg.): The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism. New York: Harcourt, Brace & World, 296–336.
- Manning, Christopher, Prabhakar Raghavan und Hinrich Schütze (2018): Introduction to information retrieval. Cambridge: Cambridge University Press.
- Manovich, Lev (2001): The Language of New Media. Cambridge, MA: MIT-Press.
- Mansholt, Malte (2022): Sparmaßnahme: Mit Alexa wollte Amazon die Welt verändern – jetzt steht sie auf der Streichliste. In: Stern, 20.11.2022. URL: <https://www.stern.de/digital/online/mit-alexa-wollte-amazon-die-welt-veraendern-jetzt-steht-sie-auf-der-streichliste-32902438.html> (21.03.2025).
- Markl, Nina (2022): Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM Press, 521–534.
- Marres, Noortje und David Stark (2020): Put to the test: For a new sociology of testing. In: The British journal of sociology 71 (3), 423–443.
- Martinet, André (1963[1960]): Grundzüge der Allgemeinen Sprachwissenschaft. Stuttgart: Kohlhammer.
- Matthewson, Lisa (1992): Talking to the TV. The conversational behaviour of television viewers. In: Wellington Working Papers in Linguistics 5, 17–35.
- McEwan, Ian (2019): Maschinen wie ich. Übersetzung aus dem Englischen von Bernhard Robben. Zürich: Diogenes.
- McFarland, David (2001): Respiratory markers of conversational interaction. In: Journal of Speech, Language, and Hearing Research 44 (1), 128–143.
- McNeill, David (2014): The emblem as metaphor. In: Seyfeddinipur, Mandana und Marinanne Gullberg (Hrsg.): From Gesture in Conversation to Visible Action as Utterance. Essays in honor of Adam Kendon. Amsterdam: John Benjamins, 75–94.
- Meer, Dorothee (2007): „ja er redet nur MÜLL hier.“ – Funktionen von *ja* als Diskursmarker in Täglichen Talkshows. In: gidi-Arbeitspapiere 11, 1–25.
- Meer, Dorothee (2009): „Unscharfe Ränder“ – Einige kategoriale Überlegungen zu Konstruktionen mit dem Diskursmarker „ja“ in konfrontativen Talkshowpassagen. In: Bückler, Jörg und Susanne Günther (Hrsg.): Grammatik im Gespräch. Konstruktionen der Selbst- und Fremdpositionierung. Berlin u. a.: De Gruyter, 87–114.
- Meer, Dorothee und Ina Pick (2019): Einführung in die Angewandte Linguistik. Stuttgart: Metzler.
- Meibauer, Jörg et al. (Hrsg.) (2015): Einführung in die germanistische Linguistik. Stuttgart: Metzler.
- Meiler, Matthias (2019): Zur praxeologischen Verhältnisbestimmung von Materialität, Medialität und Mentalität oder: Medien als Praxis. In: Zeitschrift für Semiotik 41 (1–2), 63–88.
- Meiler, Matthias (2021): Keine Methode ohne Methodologie – Überlegungen zu einer praxeologisch fundierten Medienlinguistik. In: Tienken, Susanne et al. (Hrsg.): Methoden kontrastiver Medienlinguistik. Bern: Peter Lang, 79–94.

- Meiler, Matthias (2023): Method(olog)ische Herausforderungen der Analyse sprachlichen Handelns in kommunikativen Praktiken. In: Meiler, Matthias und Martin Siefkes (Hrsg.): *Linguistische Methodenreflexion im Aufbruch. Beiträge zur aktuellen Diskussion im Schnittpunkt von Ethnographie und Digital Humanities, Multimodalität und Mixed Methods*. Berlin u. a.: De Gruyter, 45–77.
- Meiler, Matthias und Martin Siefkes (2023): Zwölf Thesen zur linguistischen Methodendiskussion. In: Meiler, Matthias und Martin Siefkes (Hrsg.): *Linguistische Methodenreflexion im Aufbruch. Beiträge zur aktuellen Diskussion im Schnittpunkt von Ethnographie und Digital Humanities, Multimodalität und Mixed Methods*. Berlin u. a.: De Gruyter, 319–323.
- Merkle, Benedikt und Tim Hector (2025): Werkzeuge und Medienpraktiken. Intelligente persönliche Assistenten und das Paradigma objektorientierten Programmierens. In: Borbach, Christoph et al. (Hrsg.): *Akustische Interfaces. Interdisziplinäre Perspektiven auf Schnittstellen von Technologien, Sounds und Menschen*. Wiesbaden: Springer VS.
- Merziger, Barbara (2005): *Das Lachen der Frauen im Gespräch über Shopping und Sexualität*. Dissertationsschrift, Berlin: Freie Universität.
- Metcalfe, Allan (2011): *OK. The improbable story of America's greatest word*. Oxford: Oxford University Press.
- Meyer, Christian (2013): *Ethnography: Body, communication, and cultural practices*. In: Müller, Cornelia et al. (Hrsg.): *Body – language – communication. An international handbook on multimodality in human interaction*. Erster Halbband. Berlin: De Gruyter, 227–239.
- Meyer, Christian (2015): Neopraxiologie. Ethnografische und konversationsanalytische Praxisforschung in ethnomethodologischer Einstellung. In: Schäfer, Franka, Anna Daniel und Frank Hillebrandt (Hrsg.): *Methoden einer Soziologie der Praxis*. Bielefeld: transcript, 91–119.
- Meyer, Joachim et al. (1996): Duration estimates and users' preferences in human-computer interaction. In: *Ergonomics* 39 (1), 46–60.
- Miggelbrink, Monique (2018): *Fernsehen und Wohnkultur. Zur Vermöbelung von Fernsehgeräten in der BRD der 1950er- und 1960er-Jahre*. Bielefeld: transcript.
- Mikos, Lothar (1996): Parasoziale Interaktion und indirekte Adressierung. In: Vorderer, Peter (Hrsg.): *Fernsehen als „Beziehungskiste“*. Parasoziale Beziehungen und Interaktionen mit TV-Personen. Opladen: Westdeutscher Verlag, 97–106.
- Misselhorn, Catrin (2021): *Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co*. Stuttgart: Reclam.
- Mitchell, Robert (2001): Americans' Talk to Dogs: Similarities and Differences With Talk to Infants. In: *Research on Language & Social Interaction* 34 (2), 183–210.
- Mols, Anouk (2021): *Everyday Experiences of Privacy and Surveillance. Negotiating Appropriate Forms of Monitoring*. Rotterdam: Erasmus Research Centre for Media, Communication and Culture.
- Mondada, Lorenza (2007): Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. In: *Discourse Studies* 9 (2), 197–225.
- Mondada, Lorenza (2013): Interactional space and the study of embodied talk-in-interaction. In: Auer, Peter et al. (Hrsg.): *Space in language and linguistics. Geographical, interactional, and cognitive perspectives*. Berlin: De Gruyter, 247–275.
- Mondada, Lorenza (2014a): Instructions in the operating room. How the surgeon directs their assistant's hands. In: *Discourse Studies* 16 (2), 131–161.
- Mondada, Lorenza (2014b): The local constitution of multimodal resources for social interaction. In: *Journal of Pragmatics* 65, 137–156.

- Mondada, Lorenza (2018): Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality. In: *Research on Language & Social Interaction* 51 (1), 85–106.
- Mondada, Lorenza (2021): *Sensing in Social Interaction. The Taste for Cheese in Gourmet Shops*. Cambridge: Cambridge University Press.
- Mondada, Lorenza (2022): Conventions for multimodal transcription. Version 6.0.1. URL: <https://www.lorenzamondada.net/multimodal-transcription> (23.03.2025).
- Mondada, Lorenza und Florence Oloff (2011): Gestures in overlap. In: Stam, Gale und Mika Ishino (Hrsg.): *Integrating Gestures*. Amsterdam: John Benjamins, 321–338.
- Mondada, Lorenza und Marja-Leena Sorjonen (2021): OKAY in closings and transitions. In: Betz, Emma et al. (Hrsg.): *OKAY across Languages. Toward a comparative approach to its use in talk-in-interaction*. Amsterdam: John Benjamins, 93–127.
- Mondémé, Chloé (2018): How Do We Talk to Animals? Modes and Pragmatic Effects of Communication with Pets. In: *Language et Société* 163, 77–99.
- Mondémé, Chloé (2022): Why study turn-taking sequences in interspecies interactions? In: *Journal for the Theory of Social Behaviour* 52 (1), 67–85.
- Montalban, Matthieu, Vincent Frigant und Bernard Jullien (2019): Platform economy as a new form of capitalism: a Régulationist research programme. In: *Cambridge Journal of Economics* 43 (4), 805–824.
- Mooshammer, Sandra und Katrin Etzrodt (2022): Gender Ambiguity in Voice-Based Assistants: Gender Perception and Influences of Context. In: *Human-Machine Communication* 5, 49–74.
- Morley, David und Roger Silverstone (1990): Domestic communication – technologies and meanings. In: *Media, Culture & Society* 12 (1), 31–55.
- Morozov, Evgeny (2019): Capitalism's New Clothes. Shoshana Zuboff's new book on „surveillance capitalism“ emphasizes the former at the expense of the latter. In: *The Baffler*, 04.02.2019. URL: <https://thebaffler.com/latest/capitalisms-new-clothes-morozov> (21.03.2025).
- Moser, Hugo (1969): Umgangssprache. Überlegungen zu ihren Formen und ihrer Stellung im Sprachganzen. In: *Zeitschrift für Mundartforschung* 27, 215–232.
- Mroczynski, Robert (2014): *Gesprächslinguistik. Eine Einführung*. Tübingen: Narr.
- Müller, Cornelia (1998): *Redebegleitende Gesten. Kulturgeschichte, Theorie, Sprachvergleich*. Berlin: Spitz.
- Muster, Judith (2013): Welchen kommunikativen Stellenwert haben Haustiere? Eine kommunikationssoziologische Betrachtung der Mensch-Tier-Beziehung. In: Pfau-Effinger, Birgit und Sonja Buschka (Hrsg.): *Gesellschaft und Tiere*. Wiesbaden: Springer, 165–192.
- Nass, Clifford und Youngme Moon (2000): Machines and Mindlessness: Social Responses to Computers. In: *Journal of Social Issues* 56 (1), 81–103.
- Nass, Clifford, Youngme Moon und Nancy Green (1997): Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. In: *Journal of Applied Social Psychology* 27 (10), 864–876.
- Natale, Simone (2020): To believe in Siri: A critical analysis of AI voice assistants. In: *Working Paper Communicative Figurations* 32, 1–17.
- Natale, Simone (2021): *Deceitful media. Artificial intelligence and social life after the Turing test*. New York: Oxford University Press.
- Natale, Simone (2023): AI, Human-Machine Communication and Deception. In: Guzman, Andrea, Rhonda McEwen und Steve Jones (Hrsg.): *The SAGE Handbook of Human-Machine Communication*. London: Sage, 401–408.
- Natale, Simone und Henry Cooke (2021): Browsing with Alexa: Interrogating the impact of voice assistants as web interfaces. In: *Media, Culture & Society* 43 (6), 1000–1016.

- Natarajan, Pradeep, Arindam Mandal und Nikko Ström (2023): New Alexa features: Natural turn-taking. In: Amazon Science Blog, 06.09.2023. URL: <https://www.amazon.science/blog/change-to-alexa-wake-word-process-adds-natural-turn-taking> (24.03.2025).
- Nebel, Bernhard (2019): Turing-Test. In: Liggieri, Kevin und Oliver Müller-Känel (Hrsg.): Mensch-Maschine-Interaktion. Handbuch zu Geschichte – Kultur – Ethik. Stuttgart: Metzler, 304–306.
- Neuland, Eva (2008): Jugendsprache. Eine Einführung. Tübingen/Basel: Francke.
- Neuland, Eva und Benjamin Könnig (2021): Abmilderungsstrategien in der Unterrichtskommunikation. In: Der Deutschunterricht 4, 88–93.
- Norman, Donald (1988): The Design of Everyday Things. New York u. a.: Currency-Doubleday.
- Norman, Donald (2010): Natural User Interfaces Are Not Natural. In: interactions 17 (3), 6–10.
- Norricks, Neal (2001): On the conversational performance of narrative jokes: Towards an account of timing. In: Humor 14 (3), 225–274.
- Norricks, Neal und Claudia Bubel (2009): Direct address as a resource for humor. In: Norricks, Neal und Delia Chiaro (Hrsg.): Humor in Interaction. Amsterdam: John Benjamins, 29–48.
- Nübling, Damaris (2011): Von der ‚Jungfrau‘ zur ‚Magd‘, vom ‚Mädchen‘ zur ‚Prostituierten‘: Die Pejorisierung der Frauenbezeichnungen als Zerrspiegel der Kultur und als Effekt männlicher Galanterie? In: Jahrbuch für Germanistische Sprachgeschichte 2 (1), 344–362.
- Nübling, Damaris, Fabian Fahlbusch und Rita Heuser (2012): Namen. Eine Einführung in die Onomastik. Tübingen: Narr.
- Oloff, Florence (2019): Das Smartphone als soziales Objekt. Eine multimodale Analyse von initialen Zeigesequenzen in Alltagsgesprächen. In: Marx, Konstanze und Axel Schmidt (Hrsg.): Interaktion und Medien. Interaktionsanalytische Zugänge zu medienvermittelter Kommunikation. Heidelberg: Winter, 191–218.
- Opfermann, Christiane und Karola Pitsch (2017): Reprompts as Error Handling Strategy in Human-Agent-Dialog? User Responses to a System’s Display of Non-understanding. In: Human-robot collaboration and human assistance for an improved quality of life: IEEE RO-MAN 2017 / 26th IEEE International Symposium on Robot and Human Interactive Communication August 28 – September 1, 2017, Lisbon, 310–316.
- Opfermann, Christiane et al. (2017): The Communicative Activity of „Making Suggestions“ as an Interactional Process. In: Wrede, Britta (Hrsg.): Proceedings of the 5th International Conference on Human Agent Interaction. New York: ACM Press, 161–170.
- Østergaard Madsen, Christian und Pernille Kræmmergaard (2015): The efficiency of freedom: Single parents’ domestication of mandatory e-government channels. In: Government Information Quarterly 32 (4), 380–388.
- Oswald, David (2012): Semiotik auditiver Interfaces. Zur Geschichte von Gestaltung und Rezeption auditiver Zeichen in Computer-Interfaces. In: Schoon, Andi und Axel Volmar (Hrsg.): Das geschulte Ohr. Eine Kulturgeschichte der Sonifikation. Bielefeld: transcript, 243–264.
- Pafel, Jürgen (2015): Pragmatik. In: Meibauer, Jörg et al. (Hrsg.): Einführung in die germanistische Linguistik. Stuttgart: Metzler, 210–252.
- Patzelt, Werner (1987): Grundlagen der Ethnomethodologie. Theorie, Empirie und politikwissenschaftlicher Nutzen einer Soziologie des Alltags. München: W. Fink.
- Pearl, Cathy (2016): Designing voice user interfaces. Principles of conversational experiences. Sebastopol: O’Reilly Media.
- Pennycook, Alastair (2010): Language as a local practice. Milton Park/Abingdon/New York: Routledge.
- Pentzold, Christian und Andreas Bischof (2019): Making Affordances Real: Socio-Material Prefiguration, Performed Agency, and Coordinated Activities in Human–Robot Communication. In: Social Media + Society 5 (3), 1–11.

- Perrigo, Billy (2023): Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic. In: TIME, 18.01.2023. URL: <https://time.com/6247678/openai-chatgpt-kenya-workers/> (21.09.2023).
- Petereit, Dieter (2021a): Amazon-Whistleblower warnen vor mieser Sicherheit beim E-Commerce-Riesen. In: t3n, 25.02.2021. URL: <https://t3n.de/news/amazon-whistleblower-datenschutz-1358723/> (21.03.2025).
- Petereit, Dieter (2021b): Revision deckt auf: Tausende Amazon-Mitarbeiter hatten Zugriff auf Marktplatz. In: t3n, 05.05.2021. URL: <https://t3n.de/news/revision-amazon-spoofers-marketplace-1377069/> (21.03.2025).
- Pfänder, Stefan und Jörg Wagner (2009): Warum wir sprechen, wenn wir klicken. Die Mensch-Maschine-Interaktion als zukunftsweisendes Thema für sprach- und medienwissenschaftliche Module in germanistischen und romanistischen Studiengängen. In: Kailuweit, Rolf und Stefan Pfänder (Hrsg.): *Franko-Media: Aufriss einer französischen Sprach- und Medienwissenschaft*. Berlin: Berliner Wissenschaftsverlag, 89–98.
- Pfeiffer, Martin (2015): *Selbstreparaturen im Deutschen*. Berlin u. a.: De Gruyter.
- Pfister, Beat und Tobias Kaufmann (2017): *Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Berlin: Springer.
- Phan, Thao (2017): The Materiality of the Digital and the Gendered Voice of Siri. In: *Transformations* 29, 23–33.
- Phan, Thao (2019): Amazon Echo and the Aesthetics of Whiteness. In: *Catalyst: Feminism, Theory, Technoscience* 5 (1), 1–39.
- Pickering, Martin und Simon Garrod (2004): Toward a mechanistic psychology of dialogue. In: *Behavioral and Brain Sciences* 27 (2), 169–190.
- Pika, Simone et al. (2018): Taking turns: bridging the gap between human and animal communication. In: *Proceedings of the Royal Society of Biological Sciences* 285, 1–9.
- Pike, Kenneth und Evelyn Pike (1947): Immediate Constituents of Mazateco Syllables. In: *International Journal of American Linguistics* 13 (2), 78–91.
- Pillet-Shore, Danielle (2018): How to Begin. In: *Research on Language & Social Interaction* 51 (3), 213–231.
- Pins, Dominik und Fatemeh Alizadeh (2021): „Im Wohnzimmer kriegt die schon alles mit“ – Sprachassistentendaten im Alltag. In: Boden, Alexander et al. (Hrsg.): *Verbraucherdatenschutz – Technik und Regulation zur Unterstützung des Individuums*. Hochschule Bonn-Rhein-Sieg, 1–20.
- Pins, Dominik et al. (2020): „Miss Understandable“ – A study on how users appropriate voice assistants and deal with misunderstandings. In: *Proceedings of Mensch und Computer 2020 (MUC20)*. ACM, Magdeburg, 1–11.
- Pitsch, Karola (2015): Ko-Konstruktion in der Mensch-Roboter-Interaktion Kontingenz, Erwartungen und Routinen in der Eröffnung. In: Krafft, Ulrich, Elisabeth Gülich und Ulrich Dausendschön-Gay (Hrsg.): *Ko-Konstruktionen in der Interaktion*. Bielefeld: transcript, 229–257.
- Pitsch, Karola (2020): Answering a robot's questions. Participation dynamics of adult-child-groups in encounters with a museum guide robot. In: *Réseaux* 220–221 (2–3), 113–150.
- Pitsch, Karola (2023): Mensch-Roboter-Interaktion als Forschungsinstrument der Interaktionalen Linguistik. Situiertheit von Interaktion und das Design referenzieller Praktiken. In: Meiler, Matthias und Martin Siefkes (Hrsg.): *Linguistische Methodenreflexion im Aufbruch. Beiträge zur aktuellen Diskussion im Schnittpunkt von Ethnographie und Digital Humanities, Multimodalität und Mixed Methods*. Berlin u. a.: De Gruyter, 119–151.

- Pitsch, Karola, Raphaela Gehle und Sebastian Wrede (2013): A museum guide robot: Dealing with multiple participants in the real-world. In: Workshop „Robots in public spaces. Towards multi-party, short-term, dynamic human-robot interaction“ at ICSR 2013.
- Pitsch, Karola et al. (2009a): „The first five seconds“: Contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. In: RO-MAN 2009. The 18th IEEE International Symposium on Robot and Human Interactive Communication; Toyama, Japan, 27 September – 2 October 2009. Piscataway, NJ: IEEE, 985–991.
- Pitsch, Karola et al. (2009b): On the loop of action modification and the recipient's gaze in adult-child interaction. In: Proceedings of GESPIN – Gesture & Speech in Interaction – Poznań, 24–26 September 2006 1, 1–7.
- Pitsch, Karola et al. (2017): Interactional Dynamics in User Groups. In: Wrede, Britta (Hrsg.): Proceedings of the 5th International Conference on Human Agent Interaction. New York: ACM Press, 393–397.
- Plantin, Jean-Christophe und Aswin Punathambekar (2019): Digital media infrastructures: pipes, platforms, and politics. In: Media, Culture & Society 41 (2), 163–174.
- Ploder, Andrea und James McElvenny (2022): Sequence Analysis in Linguistics and Social Theory. Introduction to the Thematic Issue. In: Gesprächsforschung 23, 289–306.
- Polanyi, Michael (1985[1966]): Implizites Wissen. Frankfurt a. M.: Suhrkamp.
- Pollner, Melvin (1991): Left of ethnomethodology: The rise and decline of radical reflexivity. In: American Sociological Review 56 (3), 370–380.
- Pomerantz, Anita (1984): Pursuing a response. In: Atkinson, J. und John Heritage (Hrsg.): Structures of social action. Studies in conversation analysis. Cambridge: Cambridge University Press, 79–112.
- Pomerantz, Anita und John Heritage (2012): Preference. In: Sidnell, Jack und Tanya Stivers (Hrsg.): The Handbook of Conversation Analysis. Chichester: Blackwell, 211–228.
- Porcheron, Martin (2019): Conditional Voice Recorder (CVR). URL: <https://github.com/MixedRealityLab/conditional-voice-recorder> (21.09.2023).
- Porcheron, Martin et al. (2018): Voice Interfaces in Everyday Life. In: Mandryk, Megan et al. (Hrsg.): Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems – CHI '18. New York: ACM Press, 1–12.
- Porsché, Yannik (2015): Kontextualisierung am Schnittpunkt von Museumsraum und Öffentlichkeit. Ethnomethodologische, poststrukturale und ethnografische Analyseheuristiken. In: Schäfer, Franka, Anna Daniel und Frank Hillebrandt (Hrsg.): Methoden einer Soziologie der Praxis. Bielefeld: transcript, 239–265.
- Poti, Patrick (2023): Können Google und Amazon ihre Smart Speaker mit KI retten – oder werden die Sprachassistenten endgültig begraben? Gegen das Ende von Alexa & Co.: Wie Amazon und Google das Smart-Speaker-Ruder wieder rumreißen wollen. In: GameStar Tech, 02.08.2023.
- Potter, Jonathan und Margaret Wetherell (1987): Discourse and social psychology. Beyond Attitudes and Behaviour. London: Sage.
- Pridmore, Jason und Anouk Mols (2020): Personal choices and situated data: Privacy negotiations and the acceptance of household Intelligent Personal Assistants. In: Big Data & Society 7 (1), 1–12.
- Proske, Nadine (2017): Zur Funktion und Klassifikation gesprächsorganisatorischer Imperative. In: Blühdorn, Hardarik et al. (Hrsg.): Diskursmarker im Deutschen. Reflexionen und Analysen. Göttingen: Verlag für Gesprächsforschung, 73–102.
- Püschel, Ulrich (1993): „du mußt gucken nicht so viel reden“ – Verbale Aktivitäten bei der Fernsehrezeption. In: Holly, Werner und Ulrich Püschel (Hrsg.): Medienrezeption als Aneignung.

- Methoden und Perspektiven qualitativer Medienforschung. Wiesbaden: VS Verlag für Sozialwissenschaften, 115–135.
- Quasthoff, Uta (1980a): Erzählen in Gesprächen. Tübingen: Narr.
- Quasthoff, Uta (1980b): Gemeinsames Erzählen als Form und Mittel im sozialen Konflikt oder Ein Ehepaar erzählt eine Geschichte. In: Ehlich, Konrad (Hrsg.): Erzählen im Alltag. Frankfurt a. M.: Suhrkamp, 109–141.
- Quasthoff, Uta (1981): Zuhöreraktivitäten beim konversationellen Erzählen. In: Schröder, Peter und Hugo Steger (Hrsg.): Dialogforschung. Düsseldorf: Schwann, 287–313.
- Quasthoff, Uta (2001): Erzählen als interaktive Gesprächsstruktur. In: Brinker, Klaus et al. (Hrsg.): Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung. 2. Halbband. Berlin u. a.: De Gruyter, 1293–1309.
- Ramers, Karl-Heinz (2015): Phonologie. In: Meibauer, Jörg et al. (Hrsg.): Einführung in die germanistische Linguistik. Stuttgart: Metzler, 70–120.
- Ramge, Thomas (2018): Mensch und Maschine. Wie Künstliche Intelligenz und Roboter unser Leben verändern. Stuttgart: Reclam.
- Rammert, Werner (1993): Technik aus soziologischer Perspektive. Verteiltes Handeln in soziotechnischen Konstellationen – Forschungsstand, Theorieansätze, Fallbeispiele. Ein Überblick. Opladen: Westdeutscher Verlag.
- Rammert, Werner und Ingo Schulz-Schaeffer (Hrsg.) (2002a): Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik. Frankfurt a.M.: Campus-Verlag.
- Rammert, Werner und Ingo Schulz-Schaeffer (2002b): Technik und Handeln: wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt. In: Rammert, Werner und Ingo Schulz-Schaeffer (Hrsg.): Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik. Frankfurt a. M.: Campus, 11–64.
- Ranzini, Giuliam, Christoph Lutz und Christian Hoffmann (2023): Privacy Cynicism: Resignation in the Face of Agency Constraints. In: Trepte, Sabine und Philipp Masur, (Hrsg.): The Routledge Handbook of Privacy and Social Media. New York u. a.: Routledge, 134–143.
- Rauniomaa, Mirka (2017): Assigning roles and responsibilities. Finnish imperatively formatted directive actions in a mobile instructional setting. In: Sorjonen, Marja-Leena, Liisa Raevaara und Elizabeth Couper-Kuhlen (Hrsg.): Imperative Turns at Talk: The Design of Directives in Action. Amsterdam: John Benjamins, 325–355.
- Reckwitz, Andreas (2003): Grundelemente einer Theorie sozialer Praktiken. Eine sozialtheoretische Perspektive. In: Zeitschrift für Soziologie 32 (4), 282–301.
- Reeves, Stuart und Martin Porcheron (2023): Conversational AI: Respecifying Participation as Regulation. In: Housley, William et al. (Hrsg.): The SAGE Handbook of Digital Society. London u. a.: Sage, 573–592.
- Rehbein, Jochen (1984): Beschreiben, Berichten und Erzählen. In: Ehlich, Konrad (Hrsg.): Erzählen in der Schule. Tübingen: Niemeyer, 67–124.
- Reichert, Jo (1992): Beschreiben oder Zeigen – Über das Verfassen ethnographischer Berichte. In: Soziale Welt 43, 331–350.
- Reineke, Silke und Henrike Helmer (2024): User practices in dealing with trouble in interactions with virtual assistants in German: Repeating, altering and insisting. In: Discourse & Communication (OnlineFirst).
- Rendle-Short, Johanna (2010): „Catherine, you’re wasting your time“: Address terms within the Australian political interview. In: Journal of Pragmatics 39, 1503–1525.

- Rieger, Stefan (2019): Interface. Die Natur der Schnittstelle. In: Liggieri, Kevin und Oliver Müller-Känel (Hrsg.): *Mensch-Maschine-Interaktion. Handbuch zu Geschichte – Kultur – Ethik*. Stuttgart: Metzler, 190–197.
- Roddy, Matthew, Gabriel Skantze und Naomi Harte: Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs. In: *ArXiv-Preprint 1806.11461v1*.
- Rogowska, Jowita und Valentin Schneider (2022): Bericht über das 4. Treffen des Netzwerks für Doktorand*innen der Gesprächsforschung (DokGF) 19–20.11.2021, Mannheim (online). In: *Gesprächsforschung* 23, 254–267.
- Rohlfing, Katharina et al. (2021): Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. In: *IEEE Transactions on Cognitive and Developmental Systems* 13 (3), 717–728.
- Röser, Jutta (2007): Der Domestizierungsansatz und seine Potenziale zur Analyse alltäglichen Medienhandelns. In: Röser, Jutta (Hrsg.): *MedienAlltag. Domestizierungsprozesse alter und neuer Medien*. Wiesbaden: VS Verlag, 15–30.
- Röser, Jutta und Corinna Peil (2014): Internetnutzung im häuslichen Alltag. Räumliche Arrangements zwischen Fragmentierung und Gemeinschaft. Wiesbaden: Springer VS.
- Rothe, Ingmar (2022): Die Transkription folgt den Daten: Multikonventionelles Transkribieren. In: Schwarze, Cordula und Sven Grawunder (Hrsg.): *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion. Konzepte, Probleme, Lösungen*. Tübingen: Narr-Francke-Attempo, 132–162.
- Sacks, Harvey (1972): On the Analyzability of Stories by Children. In: Gumperz, John und Dell Hymes (Hrsg.): *Directions in Sociolinguistics*. New York u. a.: Holt, Rinehart and Winston, 325–345.
- Sacks, Harvey (1984): Notes on methodology. In: Atkinson, J. (Hrsg.): *Structures of social action. Studies in conversation analysis*. Cambridge: Cambridge University Press, 21–27.
- Sacks, Harvey (1995): *Lectures on Conversation. Volume I*. Oxford: Blackwell.
- Sacks, Harvey, Emanuel Schegloff und Gail Jefferson (1974): A simplest systematics for the organisation of turn taking in conversation. In: *Language* 50 (4), 696–735.
- Sadowski, Jathan (2020): *Too smart. How Digital Capitalism is Extracting Data, Controlling Our Lives, and Taking Over the World*. Cambridge: MIT-Press.
- Sadowski, Jathan, Yolande Strengers und Jenny Kennedy (2021): More work for Big Mother: Revaluing care and control in smart homes. In: *Environment and Planning A: Economy and Space* 56 (1), 330–345.
- Sager, Sven (2001): Formen und Probleme der technischen Dokumentation von Gesprächen. In: Brinker, Klaus et al. (Hrsg.): *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung*. 2. Halbband. Berlin u. a.: De Gruyter, 1022–1033.
- Savoldi, Beatrice et al. (2021): Gender Bias in Machine Translation. In: *ArXiv-Preprint 2104.06001*.
- Schäfer, Franka und Anna Daniel (2015): Zur Notwendigkeit einer praxissoziologischen Methodendiskussion. In: Schäfer, Franka, Anna Daniel und Frank Hillebrandt (Hrsg.): *Methoden einer Soziologie der Praxis*. Bielefeld: transcript, 37–55.
- Schäfer, Franka, Anna Daniel und Frank Hillebrandt (Hrsg.) (2015): *Methoden einer Soziologie der Praxis*. Bielefeld: transcript.
- Schäfer, Hilmar (2016): Einleitung. Grundlagen, Rezeption und Forschungsperspektiven der Praxistheorie. In: Schäfer, Hilmar (Hrsg.): *Praxistheorie. Ein soziologisches Forschungsprogramm*. Bielefeld: transcript, 9–25.
- Schäfer, Hilmar (2020): Die Instabilität der Praxis. Reproduktion und Transformation des Sozialen in der Praxistheorie. Weilerswist: Velbrück Wissenschaft.

- Schank, Gerd (1979): Zum Problem der Natürlichkeit von Gesprächen in der Konversationsanalyse. In: Dittmann, Jürgen (Hrsg.): *Arbeiten zur Konversationsanalyse*. Tübingen: Niemeyer, 73–93.
- Schatzki, Theodore (1996): *Social practices. A Wittgensteinian approach to human activity and the social*. Cambridge: Cambridge University Press.
- Schatzki, Theodore (2001): *Practice Theory*. In: Schatzki, Theodore, Karin Knorr-Cetina und Eike von Savigny (Hrsg.): *The practice turn in contemporary theory*. London: Routledge, 1–14.
- Schatzki, Theodore (2016): *Praxistheorie als flache Ontologie*. In: Schäfer, Hilmar (Hrsg.): *Praxistheorie. Ein soziologisches Forschungsprogramm*. Bielefeld: transcript, 29–44.
- Schatzki, Theodore, Karin Knorr-Cetina und Eike von Savigny (Hrsg.) (2001): *The practice turn in contemporary theory*. London: Routledge.
- Scheffler, Joline und Karola Pitsch (2020): *Pre-Beginnings in Human-Robot Encounters: Dealing with time delay*. In: *Proceedings of 18th European Conference on Computer-Supported Cooperative Work*, Siegen.
- Schegloff, Emanuel (1968): *Sequencing in Conversational Openings*. In: *American Anthropologist* 70 (6), 1075–1095.
- Schegloff, Emanuel (1979): *The relevance of repair to syntax-for-conversation*. In: Givón, Talmy (Hrsg.): *Discourse and Syntax*. New York: Brill, 261–286.
- Schegloff, Emanuel (1982): *Discourse as an interactional achievement: some uses of uh huh and other things that come between sentences*. In: Tannen, Deborah (Hrsg.): *Analyzing discourse: text and talk*. Washington D.C.: Georgetown University Press, 71–93.
- Schegloff, Emanuel (1984): *On some gesture's relation to talk*. In: Atkinson, J. (Hrsg.): *Structures of social action. Studies in conversation analysis*. Cambridge: Cambridge University Press, 266–296.
- Schegloff, Emanuel (1987): *Recycled turn beginnings: A precise repair mechanism in conversation's turn-taking organisation*. In: Button, Graham und John Lee (Hrsg.): *Talk and Social Organization*. Clevedon: Multilingual Matters, 70–85.
- Schegloff, Emanuel (1990): *On the Organization of Sequences as a Source of „Coherence“ in Talk-in-Interaction*. In: Dorval, Bruce (Hrsg.): *Conversational Organization and its Development*. Norwood: Alex Publishing Corporation, 51–77.
- Schegloff, Emanuel (1992): *Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation*. In: *American Journal of Sociology* 97 (5), 1295–1345.
- Schegloff, Emanuel (1996): *Some Practices for Referring to Persons in Talk-in-Interaction: A Partial Sketch of a Systematics*. In: Fox, Barbara (Hrsg.): *Studies in anaphora*. Amsterdam: John Benjamins, 437–485.
- Schegloff, Emanuel (1997a): *Practices and actions: Boundary cases of other-initiated repair*. In: *Discourse Processes* 23, 499–545.
- Schegloff, Emanuel (1997b): *Whose text? Whose context?* In: *Discourse & Society* 8 (2), 165–187.
- Schegloff, Emanuel (2000a): *On Granularity*. In: *Annual Review of Sociology* 26, 715–720.
- Schegloff, Emanuel (2000b): *Overlapping talk and the organization of turn-taking for conversation*. In: *Language in Society* 29 (1), 1–63.
- Schegloff, Emanuel (2006): *Interaction: The infrastructure for Social Institutions, the Natural Ecological Niche for Language, and the Arena in which Culture is Enacted*. In: Enfield, Nick und Stephen Levinson (Hrsg.): *Roots of human sociality. Culture, cognition and interaction*. Oxford: Berg, 70–96.
- Schegloff, Emanuel (2007): *Sequence organization in interaction. A primer in conversation analysis*. Cambridge: Cambridge University Press.
- Schegloff, Emanuel, Gail Jefferson und Harvey Sacks (1977): *The Preference for Self-Correction in the Organization of Repair in Conversation*. In: *Language* 53 (2), 361–382.

- Schegloff, Emanuel und Harvey Sacks (1973): Opening up Closings. In: *Semiotica* 8 (4), 289–327.
- Schenkein, Jim (1978): Sketch of an analytic mentality for the study of conversational interaction. In: Schenkein, Jim (Hrsg.): *Studies in the organization of conversational interaction*. New York: Academic Press, 1–6.
- Schiller, Amy und John McMahon (2019): Alexa, Alert Me When the Revolution Comes: Gender, Affect, and Labor in the Age of Home-Based Artificial Intelligence. In: *New Political Science* 41 (2), 173–191.
- Schlängen, David und Gabriel Skantze (2009): A General, Abstract Model of Incremental Dialogue Processing. In: *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, 710–718.
- Schmidhuber, Jürgen (2019): My First Deep Learning System of 1991 + Deep Learning Timeline 1960–2013. URL: <https://people.idsia.ch/~juergen/firstdeeplearner.html> (21.03.2025).
- Schmidtke, Oliver, Kathrin Englert und David Waldecker (2019): Vom alltäglichen Ziehen fließender Grenzen. Die Veröffentlichung von Intimität bei jugendlichen Social-Media-Nutzer*innen. In: Ettinger, Patrik et al. (Hrsg.): *Intimisierung des Öffentlichen. Zur multiplen Privatisierung des Öffentlichen in der digitalen Ära*. Wiesbaden: Springer VS, 211–226.
- Schmitt, Reinhold (2013): *Körperlich-räumliche Aspekte der Interaktion*. Tübingen: Narr.
- Schmitt, Reinhold und Heiko Hausendorf (2016): *Sprache und Raum: Eine neue Forschungsperspektive, ihre Ursprünge und ihr aktueller Entwicklungsstand*. In: Hausendorf, Heiko, Reinhold Schmitt und Wolfgang Kesselheim (Hrsg.): *Interaktionsarchitektur, Sozialtopographie und Interaktionsraum*. Tübingen: Narr, 9–26.
- Schneider, Britta (2021): Von Gutenberg zu Alexa. Posthumanistische Perspektiven auf Sprachideologie. In: Lind, Miriam (Hrsg.): *Mensch – Tier – Maschine. Sprachliche Praktiken an und jenseits der Außengrenze des Humanen*. Bielefeld: transcript, 327–346.
- Schneider, Jan Georg (2008): *Spielräume der Medialität. Linguistische Gegenstandskonstitution aus medientheoretischer und pragmatischer Perspektive*. Berlin u. a.: De Gruyter.
- Schneider, Jan Georg (2017): *Medien als Verfahren der Zeichenprozessierung: Grundsätzliche Überlegungen zum Medienbegriff und ihre Relevanz für die Gesprächsforschung*. In: *Gesprächsforschung* 18, 34–55.
- Scholl, Armin (2016): Die Logik qualitativer Methoden in der Kommunikationswissenschaft. In: Averbek-Lietz, Stefanie und Michael Meyen (Hrsg.): *Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft*. Wiesbaden: Springer VS, 17–32.
- Scholz, Sebastian (2022): *Sensormedien-Milieus und Technikökologien der Wahrnehmung. Navigieren in/mit „more-than-human“-Infrastrukturen*. In: *Navigationen* 22 (1), 199–218.
- Schoon, Andi und Axel Volmar (2012): Informierte Klänge und geschulte Ohren: Zur Kulturgeschichte der Sonifikation. In: Schoon, Andi und Axel Volmar (Hrsg.): *Das geschulte Ohr. Eine Kulturgeschichte der Sonifikation*. Bielefeld: transcript, 9–26.
- Schröder, Peter und Hugo Steger (Hrsg.) (1981): *Dialogforschung*. Düsseldorf: Schwann.
- Schulz-Schaeffer, Ingo (2010): Praxis, handlungstheoretisch. In: *Zeitschrift für Soziologie* 39 (4), 133.
- Schüttpelz, Erhard (2020): Chils Prädikation. Ein Kommentar zu Charles Goodwins Co-Operative Action. In: Kablitz, Andreas, Christoph Marksches und Peter Strohschneider (Hrsg.): *Prädikation und Bedeutung*. Berlin u. a.: De Gruyter, 127–148.
- Schüttpelz, Erhard (2021): Introduction: Media Theory Before and After the Practice Turn. In: Schüttpelz, Erhard et al. (Hrsg.): *Connect and Divide. The Practice Turn in Media Studies*. Zürich: Diaphanes, 237–245.
- Schüttpelz, Erhard und Sebastian Gießmann (2015): *Medien der Kooperation. Überlegungen zum Forschungsstand*. In: *Navigationen* 15 (1), 7–55.

- Schüttelpelz, Erhard und Christina Meyer (2017): Ein Glossar zur Praxistheorie. „Siegener Version“ (Frühjahr 2017). In: *Navigationen* 17 (1), 155–163.
- Schüttelpelz, Erhard und Christian Meyer (2018): Charles Goodwin's Co-Operative Action: The Idea and the Argument. In: *Media in Action. Interdisciplinary Journal on Cooperative Media* 1, 171–188.
- Schüttelpelz, Erhard et al. (Hrsg.) (2021): *Connect and Divide. The Practice Turn in Media Studies*. Zürich: Diaphanes.
- Schütz, Alfred und Thomas Luckmann (2003): *Strukturen der Lebenswelt*. Stuttgart: UTB.
- Schütz, Alfred und Talcott Parsons (Hrsg.) (1977): *Zur Theorie sozialen Handelns. Ein Briefwechsel*. Frankfurt a. M.: Suhrkamp.
- Schwitalla, Johannes (2002): Kleine Wörter: Partikeln im Gespräch. In: Dittmann, Jürgen und Claudia Schmidt (Hrsg.): *Über Wörter. Grundkurs Linguistik*. Freiburg im Breisgau: Rombach, 259–281.
- Schwitalla, Johannes (2012): *Gesprochenes Deutsch. Eine Einführung*. Berlin: Erich Schmidt.
- Schwitalla, Johannes (2018): Was ist ein Gespräch? In: Birkner, Karin und Nina Janich (Hrsg.): *Handbuch Text und Gespräch*. Berlin u. a.: De Gruyter, 80–103.
- Selting, Margret (1987): Reparaturen und lokale Verstehensprobleme oder: zur Binnenstruktur von Reparatursequenzen. In: *Linguistische Berichte* 108, 128–149.
- Selting, Margret (1996): On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. In: *Pragmatics* 6 (3), 371–388.
- Selting, Margret (2013): Verbal, vocal, and visual practices in conversational interaction. In: Müller, Cornelia et al. (Hrsg.): *Body – language – communication. An international handbook on multimodality in human interaction*. Erster Halbband. Berlin: De Gruyter, 589–609.
- Selting, Margret (2016): Praktiken des Sprechens und Interagierens im Gespräch aus Sicht von Konversationsanalyse und Interaktionaler Linguistik. In: Deppermann, Arnulf, Helmuth Feilke und Angelika Linke (Hrsg.): *Sprachliche und kommunikative Praktiken*. Berlin/Boston: De Gruyter, 27–56.
- Selting, Margret und Elizabeth Couper-Kuhlen (2000): Argumente für die Entwicklung einer ‚interaktionalen Linguistik‘. In: *Gesprächsforschung* 1, 79–95.
- Selting, Margret und Elizabeth Couper-Kuhlen (2001a): Forschungsprogramm ‚Interaktionale Linguistik‘. In: *Linguistische Berichte* 187, 257–287.
- Selting, Margret und Elizabeth Couper-Kuhlen (Hrsg.) (2001b): *Studies in interactional linguistics*. Amsterdam: John Benjamins.
- Selting, Margret et al. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: *Gesprächsforschung* 10, 353–402.
- Senft, Gunter (2009): Phatic communion. In: Senft, Gunter, Jan-Ola Östman und Jef Verschueren (Hrsg.): *Culture and Language Use*. Amsterdam: John Benjamins, 226–233.
- Sharrock, Wes (2012): Regelfolgen: Alles oder nichts? In: Ayaß, Ruth und Christian Meyer (Hrsg.): *Sozialität in Slow Motion. Theoretische und empirische Perspektiven*. Wiesbaden: Springer VS, 59–70.
- Shneiderman, Ben (1983): Direct Manipulation. A Step Beyond Programming Languages. In: *Computer* 16 (8), 57–69.
- Shneiderman, Ben und Catherine Plaisant (2005): *Designing the user interface. Strategies for effective human-computer-interaction*. Boston u. a.: Pearson.
- Shrager, Jeff (2021): Joseph Weizenbaum's Original ELIZA. URL: <https://sites.google.com/view/eliza-gen-org/> (24.03.2025).
- Sidnell, Jack (2006): Coordinating Gesture, Talk, and Gaze in Reenactments. In: *Research on Language & Social Interaction* 39 (4), 377–409.
- Sidnell, Jack (2009): *Conversation Analysis. An Introduction*. Hoboken: Wiley.

- Sidnell, Jack (2012): Turn-Continuation by Self and by Other. In: *Discourse Processes* 49 (3–4), 314–337.
- Silverstone, Roger (1994): *Television and everyday life*. London: Routledge.
- Silverstone, Roger (2006): Domesticating domestication. Reflections on the life of a concept. In: Berker Thomas et al. (Hrsg.): *Domestication of Media and Technology*. Maidenhead: Open University Press, 229–248.
- Silverstone, Roger und Leslie Haddon (1996): Design and the domestication of information and communication technologies: technical change and everyday life. In: Mansell, Robin und Roger Silverstone (Hrsg.): *Communication by design. The politics of information and communication technologies*. Oxford: Oxford University Press, 44–74.
- Silverstone, Roger, Eric Hirsch und David Morley (1991): Listening to a long conversation: an ethnographic approach to the study of information and communication technologies in the home. In: *Cultural Studies: Theorizing Politics, Politicising Theory* 5 (2), 204–227.
- Silverstone, Roger und Eric Hirsch (Hrsg.) (1992): *Consuming Technologies. Media and Information in Domestic Spaces*. London: Routledge.
- Silverstone, Roger, Eric Hirsch und David Morley (1992): Information and communication technologies and the moral economy of the household. In: Silverstone, Roger und Eric Hirsch (Hrsg.): *Consuming Technologies. Media and Information in Domestic Spaces*. London: Routledge, 13–28.
- Singhal, Amit (2012): Introducing the Knowledge Graph: things, not strings. In: Google Blog, 16.05.2012. URL: <https://blog.google/products/search/introducing-knowledge-graph-things-not/> (21.03.2025).
- Skantzé, Gabriel (2021): Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. In: *Computer Speech & Language* 67, 1–26.
- Sørensen, Knut (1994): Technology in Use. Two essays on the domestication of artifacts. In: STS Working Paper 2.
- Sørensen, Knut (2006): Domestication. The Enactment of Technology. In: Berker, Thomas et al. (Hrsg.): *Domestication of media and technology*. Maidenhead: Open University Press, 40–61.
- Sorjonen, Marja-Leena (1996): On repeats and responses in Finnish conversations. In: Ochs, Elinor, Emanuel Schegloff und Sandra Thompson (Hrsg.): *Interaction and grammar*. Cambridge/ New York: Cambridge University Press, 277–327.
- Sorjonen, Marja-Leena, Liisa Raevaara und Elizabeth Couper-Kuhlen (2017): Imperative turns at talk. An introduction. In: Sorjonen, Marja-Leena, Liisa Raevaara und Elizabeth Couper-Kuhlen (Hrsg.): *Imperative Turns at Talk: The Design of Directives in Action*. Amsterdam: John Benjamins, 1–24.
- Spranz-Fogasy, Thomas (2010): Verstehensdokumentation in der medizinischen Kommunikation: Fragen und Antworten im Arzt-Patient-Gespräch. In: Deppermann, Arnulf et al. (Hrsg.): *Verstehen in professionellen Handlungsfeldern*. Tübingen: Narr, 27–116.
- Srnicek, Nick (2016): *Platform capitalism*. Cambridge: Polity Press.
- Star, Susan (1993): Cooperation Without Consensus in Scientific Problem Solving: Dynamics of Closure in Open Systems. In: Diaper, Dan, Colston Sanger und Steve Easterbrook (Hrsg.): *CSCW: Cooperation or Conflict?* London: Springer, 93–106.
- Star, Susan und Geoffrey Bowker (2005): How to Infrastructure. In: Lievrouw, Leah und Sonia Livingstone (Hrsg.): *Handbook of New Media*. London: Sage, 230–245.
- Statista (2019): Statista-Umfrage: Smart Speaker und virtuelle Assistenten 2019. Statista-ID 983723. URL: <https://de.statista.com/prognosen/983723/umfrage-in-deutschland-zum-besitz-eines-smart-speakers> (21.03.2025).

- Statista (2025a): Beliebteste Smart-Speaker-Marken in Deutschland im Jahr 2024. URL: <https://de.statista.com/prognosen/999790/deutschland-beliebteste-smart-speaker-marken> (24.03.2025).
- Statista (2025b): Besitz von Smart-Home-Geräten in Deutschland im Jahr 2024. URL: <https://de.statista.com/prognosen/999788/deutschland-besitz-von-smart-home-geraeten> (24.03.2025).
- Statista (2025c): Absatz von intelligenten Lautsprechern weltweit von 2018 bis 2023 und Prognose für 2024 und 2027. URL: <https://de.statista.com/statistik/daten/studie/1079997/umfrage/prognose-zum-absatz-von-intelligenten-lautsprechern-weltweit/> (24.03.2025).
- Stefani, Elwys de (2012): Crossing perspectives on onomastic methodology: Reflections on fieldwork in place name research. An essay in interactional onomastics. In: Ender, Andrea, Adrian Leemann und Bernhard Wälchli (Hrsg.): *Methods in Contemporary Linguistics*. Berlin/New York: De Gruyter, 441–462.
- Steinke, Ines (2007): Gütekriterien qualitativer Forschung. In: Flick, Uwe, Ernst Kardorff und Ines Steinke (Hrsg.): *Qualitative Forschung. Ein Handbuch*. Reinbek: Rowohlt, 319–331.
- Stiftung Warentest (2021): Kleine Klangkugeln. In: *test* (5/2021), 28–29.
- Stivers, Tanya (2005): Modified Repeats: One Method for Asserting Primary Rights From Second Position. In: *Research on Language & Social Interaction* 38 (2), 131–158.
- Stivers, Tanya (2012): Sequence organization. In: Sidnell, Jack und Tanya Stivers (Hrsg.): *The Handbook of Conversation Analysis*. Chichester: Blackwell, 191–209.
- Stivers, Tanya (2015): Coding Social Interaction: A Heretical Approach in Conversation Analysis? In: *Research on Language & Social Interaction* 48 (1), 1–19.
- Stivers, Tanya, Nick Enfield und Stephen Levinson (2007): Person reference in interaction. In: Enfield, Nick und Tanya Stivers (Hrsg.): *Person reference in interaction. Linguistic, cultural, and social perspectives*. Cambridge: Cambridge University Press, 1–20.
- Stivers, Tanya und Jack Sidnell (2005): Introduction: Multimodal interaction. In: *Semiotica* 156 (1), 1–20.
- Stivers, Tanya et al. (2009): Universals and cultural variation in turn-taking in conversation. In: *Proceedings of the National Academy of Sciences* 106 (26), 10587–10592.
- Stoltenburg, Benjamin (2009): Was wir sagen, wenn wir es „ehrlich“ sagen ... Äußerungskommentierende Formeln bei Stellungnahmen am Beispiel von *ehrlich gesagt*. In: Bückner, Jörg und Susanne Günther (Hrsg.): *Grammatik im Gespräch. Konstruktionen der Selbst- und Fremdpositionierung*. Berlin u. a.: De Gruyter, 249–282.
- Stoltenburg, Benjamin (2012): „ich will jetzt nicht sagen Reparaturen, aber ...“ – Eine Gesprächsstrategie zur Indizierung von Problemstellen. In: *gidi-Arbeitspapiere* 47, 1–35.
- Strauss, Anselm und Juliet Corbin (2010): *Grounded theory. Grundlagen qualitativer Sozialforschung*. Weinheim: Beltz.
- Streeck, Jürgen (2013): Praxeology of gesture. In: Müller, Cornelia et al. (Hrsg.): *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction*. Zweiter Halbband. Berlin u. a.: De Gruyter, 674–685.
- Streeck, Jürgen (2014): Mutual gaze and recognition. In: Seyfeddinipur, Mandana und Marianne Gullberg (Hrsg.): *From Gesture in Conversation to Visible Action as Utterance. Essays in honor of Adam Kendon*. Amsterdam: John Benjamins, 35–56.
- Streeck, Jürgen (2016): Gestische Praxis und sprachliche Form. In: Deppermann, Arnulf, Helmuth Feilke und Angelika Linke (Hrsg.): *Sprachliche und kommunikative Praktiken*. Berlin/Boston: De Gruyter, 57–80.
- Streeck, Jürgen, Charles Goodwin und Curtis LeBaron (2011): Embodied Interaction in the Material World: An Introduction. In: Streeck, Jürgen, Charles Goodwin und Curtis LeBaron (Hrsg.):

- Embodied interaction. Language and body in the material world. Cambridge: Cambridge University Press, 1–26.
- Strengers, Yolande und Jenny Kennedy (2020): The smart wife. Why Siri, Alexa, and other smart home devices need a feminist reboot. Cambridge u. a.: The MIT Press.
- Strengers, Yolande und Larissa Nicholls (2018): Aesthetic pleasures and gendered tech-work in the 21st-century smart home. In: *Media International Australia* 166 (1), 70–80.
- Strengers, Yolande und Zoe Sofoulis (2024): Containment and Decontainment of Digital Voice Assistants. In: Angerer, Marie-Luise et al. (Hrsg.): *Containment. Technologies of Holding, Filtering, Leaking*. Lüneburg: Meson Press, 165–182.
- Strüver, Niklas (2020): Amazons Alexa als Infrastruktur der Zukunft? Eine qualitative Betrachtung von Nutzungspraktiken mit Sprachassistenten im Kontext digitaler Plattformtechnologien. Master-Arbeit, Aachen: RWTH Aachen.
- Strüver, Niklas (2023a): Frustration Free: How Alexa Orchestrates the Development of the Smart Home. In: *Digital Culture & Society* 9 (1), 99–123.
- Strüver, Niklas (2023b): Wieso eigentlich Alexa? Konzeptualisierung eines Sprachassistenten als Plattform im soziotechnischen Ökosystem verschiedener Infrastrukturen und Plattformen. In: *kommunikation@gesellschaft* 24, 1–33.
- Strüver, Niklas (2025): Innovating Alexa amid the Rise of Large Language Models. *Sociotechnical Transitions in Algorithmic Development Practices*. In: Habscheid, Stephan et al. (Hrsg.): *Voice Assistants in Private Homes. Media, Data and Language in Interaction and Discourse*. Bielefeld: transcript, 365–402.
- Stukenbrock, Anja (2009): Referenz durch Zeigen: Zur Theorie der Deixis. In: *Deutsche Sprache* 2009 (4), 289–315.
- Stukenbrock, Anja (2012): Empraxis und Displacement: Überblendete Räume in der Koch-Show-Interaktion. In: Hausendorf, Heiko, Lorenza Mondada und Reinhold Schmitt (Hrsg.): *Raum als interaktive Ressource*. Tübingen: Narr Francke Attempto Verlag, 347–379.
- Stukenbrock, Anja (2013a): Commentary: What difference does space make for interaction and interaction for space? In: Auer, Peter et al. (Hrsg.): *Space in language and linguistics. Geographical, interactional, and cognitive perspectives*. Berlin: De Gruyter, 401–408.
- Stukenbrock, Anja (2013b): Sprachliche Interaktion. In: Auer, Peter (Hrsg.): *Sprachwissenschaft. Grammatik – Interaktion – Kognition*. Stuttgart: Metzler, 217–259.
- Stukenbrock, Anja (2015): Deixis in der face-to-face-Interaktion. Berlin: De Gruyter.
- Stukenbrock, Anja (2016): Deiktische Praktiken: Zwischen Interaktion und Grammatik. In: Deppermann, Arnulf, Helmuth Feilke und Angelika Linke (Hrsg.): *Sprachliche und kommunikative Praktiken*. Berlin, Boston: De Gruyter, 81–126.
- Stukenbrock, Anja (2018): Mobile dual eye-tracking in face-to-face interaction: The case of deixis and joint attention. In: Brône, Geert (Hrsg.): *Eye-tracking in interaction. Studies on the role of eye gaze in dialogue*. Amsterdam: John Benjamins, 265–302.
- Suchman, Lucy (1987): *Plans and situated actions. The problem of human-machine communication*. Cambridge: Cambridge University Press.
- Suchman, Lucy (1990): What is Human-Machine Interaction? In: Robertson, Scott, P., Wayne Zachary und John Black (Hrsg.): *Cognition, Computing, and Cooperation*. Norwood: Ablex, 25–55.
- Suchman, Lucy (2007): *Human-machine reconfigurations. Plans and situated actions*. 2nd edition. Cambridge: Cambridge University Press.
- Suchman, Lucy (2021): Talk with Machines, Redux. In: *Interface Critique Journal* 3, 69–80.
- Sweeney, Miriam (2021): Digital Assistants. In: Thylstrup, Nanna et al. (Hrsg.): *Uncertain Archives. Critical Keywords for Big Data*. Cambridge/London: MIT-Press, 151–160.

- Tanaka, Shin (2011): *Deixis und Anaphorik*. Berlin u. a.: De Gruyter.
- Tannen, Deborah (2004): Talking the Dog: Framing Pets as Interactional Resources in Family Discourse. In: *Research on Language & Social Interaction* 37 (4), 399–420.
- Tannen, Deborah (2007): *Talking voices. Repetition, dialogue and imagery in conversational discourse*. Cambridge: Cambridge University Press.
- ten Have, Paul (1999): *Doing Conversation Analysis. A Practical Guide*. London: Sage.
- Thar, Evelyn (2015): „Ich habe Sie leider nicht verstanden.“. *Linguistische Optimierungsprinzipien für die mündliche Mensch–Maschine-Interaktion*. Bern: Peter Lang.
- Thoppilan, Romal et al. (2022): LaMDA: Language Models for Dialog Applications. In: *ArXiv-Preprint 2201.08239*, 1–47.
- Timcke, Marie-Louise und Birgit Schneider (2021): Welt aus Daten – Datenjournalismus während der Corona-Pandemie. In: *Zeitschrift für Medienwissenschaft* 13 (2), 102–114.
- Tomasello, Michael (1988): The role of joint attentional processes in early language development. In: *Language Sciences* 10 (1), 69–88.
- Tomasello, Michael (2005): *Constructing a Language*. Cambridge, MA: Harvard University Press.
- Tomasello, Michael (2009): *Why we cooperate*. Cambridge: MIT-Press.
- Torreira, Francisco, Sara Bögels und Stephen Levinson (2015): Breathing for answering: the time course of response planning in conversation. In: *Frontiers in Psychology* 6, 1–11.
- Torres Cajo, Sarah (2016): Mensch-Tier-Interaktion als soziale Praxis – Eine gesprächsanalytische Betrachtung von Adressierungsverhalten in Gassigesprächen. In: Arens, Katja, Sarah Torres Cajo (Hrsg.): *Sprache und soziale Ordnung. Studentische Beiträge zu sozialen Praktiken in der Interaktion*. Münster: Monsenstein und Vannerdat, 57–78.
- Torres Cajo, Sarah und Nils Bahlo (2016): „Ach der ist ja süß ...“ – Gassigespräche. Eine kommunikative Gattung. In: *Deutsche Sprache. Zeitschrift für Theorie, Praxis und Dokumentation* 44 (1), 77–96.
- Tuma, René (2018): Video-Interaktionsanalyse. In: Moritz, Christine und Michael Corsten (Hrsg.): *Handbuch Qualitative Videoanalyse*. Wiesbaden: Springer VS, 423–444.
- Turing, Alan (1950): Computing machinery and intelligence. In: *Mind* 49, 433–460.
- Turkle, Sherry (1995): *Life on the Screen. Identity in the Age of the Internet*. New York: Simon & Schuster.
- Turow, Joseph (2021): *The Voice Catchers. How Marketers Listen in to Exploit Your Feelings, Your Privacy, and Your Wallet*. New Haven: Yale University Press.
- Uhmann, Susanne (1991): *Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie*. Tübingen: Niemeyer.
- Uhmann, Susanne (1997): Selbstreparaturen in Alltagsdialogen: Ein Fall für eine integrative Konversationstheorie. In: Schlobinski, Peter (Hrsg.): *Syntax des gesprochenen Deutsch*. Wiesbaden: VS Verlag für Sozialwissenschaften, 157–180.
- Uhmann, Susanne (2013): Pointing with the abdomen: Local deixis under restricted conditions. In: Auer, Peter et al. (Hrsg.): *Space in language and linguistics. Geographical, interactional, and cognitive perspectives*. Berlin: De Gruyter, 557–600.
- van der Vlist, Fernando (2022): *The platform as ecosystem. Configurations and dynamics of governance and power*. Utrecht: Utrecht University Press.
- van Dijck, José (2013): *The Culture of Connectivity: A Critical History of Social Media*. Oxford u. a.: Oxford University Press.
- van Dijck, José, Thomas Poell und Martijn de Waal, (2018): *The Platform Society*. Oxford u. a.: Oxford University Press.

- Velmezova, Ekaterina (2015): The Bakhtinian *Dialogue* Revisited: A (Nonbiosemiotic) View from Historiography and Epistemology of Humanities. In: Velmezova, Ekaterina, Kalevi Kull und Stephen Cowley (Hrsg.): *Biosemiotic Perspectives on Language and Linguistics*. Basel: Springer International, 275–290.
- Volmar, Axel (2019): Productive Sounds. In: Sudmann, Andreas (Hrsg.): *The Democratization of Artificial Intelligence*. Bielefeld: transcript, 55–76.
- vom Lehn, Dirk (2018): Workplace Studies. In: Habscheid, Stephan et al. (Hrsg.): *Handbuch Sprache in Organisationen*. Berlin u. a.: De Gruyter, 327–345.
- vom Lehn, Dirk (2021): Garfinkel und Interaktion. In: Bergmann, Jörg und Christian Meyer (Hrsg.): *Ethnomethodologie reloaded. Neue Werkinterpretationen und Theoriebeiträge zu Harold Garfinkels Programm*. Bielefeld: transcript, 295–314.
- Vorderwülbecke, Klaus (1976): Anredeformen und Selbstbezeichnungen im Deutschen und Japanischen. In: Stickel, Gerhard (Hrsg.): *Deutsch-japanische Kontraste. Vorstudien zu einer kontrastiven Grammatik*. Tübingen: Narr, 335–394.
- Vrandečić, Denny und Markus Krötzsch (2014): Wikidata: a free collaborative knowledgebase. In: *Communications of the ACM* 57 (10), 78–85.
- Waitelonis, Jörg (2018): *Linked Data Supported Information Retrieval*, Karlsruhe: KIT.
- Waldecker, David (2022): Zur empirischen und theoretischen Kritik der Datensouveränität anhand der Smart-Speaker-Nutzung. In: *MerzWissenschaft* 66 (6), 145–155.
- Waldecker, David und Tim Hector (2023): Taming digital practices: A praxeological approach towards domestication of connected devices and services. Introduction to the thematic issue. In: *Digital Culture & Society* 9 (1), 5–22.
- Waldecker, David, Tim Hector und Dagmar Hoffmann (2024): Intelligent Personal Assistants in practice. Situational agencies and the multiple forms of cooperation without consensus. In: *Convergence* 30 (3), 975–991.
- Waldecker, David und Dagmar Hoffmann (2023): Zur Inszenierung von kritischen Kompetenzen in Nischenöffentlichkeiten. In: *kommunikation@gesellschaft* 23 (1), 1–28.
- Waldecker, David, Alexander Martin und Dagmar Hoffmann (2025): Mostly Harmless? Everyday Smart Speaker Use and Pragmatic Fatalism. In: Habscheid, Stephan et al. (Hrsg.): *Voice Assistants in Private Homes. Media, Data and Language in Interaction and Discourse*. Bielefeld: transcript, 291–317.
- Waldecker, David und Axel Volmar (2022): Die zweifache akustische Intelligenz virtueller Sprachassistenten zwischen verteilter Kooperation und Datafizierung. In: Schürmer, Anna, Maximilian Haberer und Tomy Brautschek (Hrsg.): *Acoustic Intelligence. Hören und Gehorchen*. Düsseldorf: Düsseldorf University Press, 161–182.
- Ward, Katie (2006): The bald guy just ate an orange. Domestication, work and home. In: Berker Thomas et al. (Hrsg.): *Domestication of Media and Technology*. Maidenhead: Open University Press, 145–164.
- Weber, Jutta und Corinna Bath (2007): ‚Social‘ Robots & ‚Emotional‘ Software Agents. Gendering Processes and De-Gendering Strategies for ‚Technologies in the Making‘. In: Zorn, Isabel (Hrsg.): *Gender designs IT: construction and deconstruction of information society technology*. Wiesbaden: VS Verlag für Sozialwissenschaften, 53–63.
- Weber, Max (1966[1921]): *Soziologische Grundbegriffe*. Tübingen: Mohr.
- Weidmüller, Lisa (2022): Human, Hybrid, or Machine? Exploring the Trustworthiness of Voice-Based Assistants. In: *Human-Machine Communication* 4, 85–110.
- Weidmüller, Lisa, Katrin Etzrodt und Sven Engesser (2022): Trustworthiness of voice-based assistants: integrating interlocutor and intermediary predictors. In: *Publizistik* 67 (4), 625–651.

- Weidmüller, Lisa et al. (2021): Ich höre was, was du nicht hörst. Medienvielfalt und Transparenz bei Sprachassistenten. Düsseldorf: Landesanstalt für Medien NRW.
- Weinel, Jonathan, Darryl Griffiths und Stuart Cunningham (2014): Easter Eggs: Hidden Tracks and Messages in Musical Mediums. In: *Proceedings of ICMC-SMC 2014*, 140–147.
- Weiser-Zurmühlen, Kristin (2021): Vergemeinschaftung und Distinktion. Eine gesprächsanalytische Studie über Positionierungspraktiken in Diskussionen über TV-Serien. Berlin u. a.: De Gruyter.
- Weizenbaum, Joseph (1966): ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and machine. In: *Communications of the ACM* 9 (1), 36–45.
- Weizenbaum, Joseph (1977): Die Macht der Computer und die Ohnmacht der Vernunft. Übersetzt von Udo Rennert. Frankfurt a. M.: Suhrkamp.
- Werner, Fritjof (1983): Lachähnliche Partikeln in Redebeiträgen. In: Weydt, Harald (Hrsg.): *Partikeln und Interaktion*. Berlin u. a.: De Gruyter, 226–241.
- West, Candace und Don Zimmermann (1987): doing gender. In: *Gender & Society* 1 (2), 121–151.
- White, Ryan, Matthew Richardson und Wen-tau Yih (2015): Questions vs. Queries in Informational Search Tasks. In: Gangemi, Aldo, Stefano Leonardi und Alessandro Panconesi (Hrsg.): *Proceedings of the 24th International Conference on World Wide Web*. New York: ACM Press, 135–136.
- Wiese, Heike (1996): Der Status von Numeralia. Ein Beitrag zur Klärung des Klassifikationsproblems für Kardinalia, Ordinalia und ‚Nummer‘. In: *Arbeitspapiere Sprache und Pragmatik* (Lund) 39, 1–39.
- Wilker, Michael (2002): Künstliche Intelligenz als technisierte Kommunikation. Das Verhältnis von sozialen und informationsverarbeitenden Systemen. Frankfurt a. M.: Peter Lang.
- Williams, Jamie (2018): Should AI Always Identify Itself? It's More Complicated Than You Might Think. In: *Electronic Frontier Foundation*, 22.05.2018. URL: <https://www EFF.org/de/deep links/2018/05/should-ai-always-identify-itself-its-more-complicated-you-might-think> (12.03.2025).
- Willkop, Eva-Maria (1988): Gliederungspartikeln im Dialog. München: Iudicium.
- Wilson, Charlie, Tom Hargreaves und Richard Hauxwell-Baldwin (2015): Smart homes and their users: a systematic analysis and key challenges. In: *Personal and Ubiquitous Computing* 19, 463–476.
- Winter, Rainer (1995): Der produktive Zuschauer. Medienaneignung als kultureller und ästhetischer Prozeß. München: Quintessenz.
- Winter, Werner (Hrsg.) (1984): *Anredeverhalten*. Tübingen: Narr.
- Wirth, Sabine (2016): Between Interactivity, Control, and ‚Everydayness‘ – Towards a Theory of User Interfaces. In: Hadler, Florian und Joachim Haupt (Hrsg.): *Interface Critique*. Berlin: Kadmos, 17–38.
- Wirth, Sabine (2019): *Dispositive der Handhabung. Zur Medialität des User Interface*. Dissertation, Marburg: Philipps Universität Marburg.
- Wirth, Sabine (2025): *Formierungen des Interface. Zur Mediengeschichte und -theorie des Personal Computing*. Bielefeld: transcript.
- Wissenschaftliche Dienste des Deutschen Bundestages (2019): Zulässigkeit der Transkribierung und Auswertung von Mitschnitten der Sprachsoftware „Alexa“ durch Amazon. URL: <https://www.bundestag.de/resource/blob/650728/3f72e6abc1c524961e5809002fe20f21/WD-10-032-19-pdf-data.pdf> (20.03.2025).
- Wittgenstein, Ludwig (1990[1953]): *Werkausgabe, Band 1. Tractatus logico-philosophicus. Tagebücher 1914–1916. Philosophische Untersuchungen*. Frankfurt a. M.: Suhrkamp.
- Woods, Heather (2018): Asking more of Siri and Alexa: feminine persona in service of surveillance capitalism. In: *Critical Studies in Media Communication* 35 (4), 334–349.

- Woods, Heater (2024): *Threshold: How Smart Homes Change Us Inside and Out*. Alabama: University of Alabama Press.
- Wuenderlich, Nancy und Stefanie Paluch (2017): A Nice and Friendly Chat with a Bot: User Perceptions of AI-Based Service Agents. In: *ICIS 2017 Proceedings* 11, 1–11.
- Young, Liam (2019): ‚I’m a Cloud of Infinitesimal Data Computation‘. When Machines Talk Back – An interview with Deborah Harrison, one of the personality designers of Microsoft’s Cortana AI. In: *Architectural Design* 89 (1), 112–117.
- Zarcone, Alessandra und Anna Leschanowsky (2023): Moderne Sprachassistenten zwischen Datenhunger und Datenschutz. In: Kovács, László (Hrsg.): *Künstliche Intelligenz und menschliche Gesellschaft*. Berlin u. a.: De Gruyter, 167–180.
- ZEIT Online (2019): Amazon lässt Alexa-Mitschnitte im Homeoffice auswerten. In: ZEIT Online, 04.08.2019. URL: <https://www.zeit.de/digital/datenschutz/2019-08/datenschutz-amazon-alexasprachassistenten-nutzerdaten-homeoffice> (21.03.2025).
- Ziem, Alexander und Alexander Lasch (2013): *Konstruktionsgrammatik*. Berlin u. a.: De Gruyter.
- Zifonun, Gisela, Ludger Hoffmann und Bruno Strecker (1997): *Grammatik der deutschen Sprache*. Berlin u. a.: De Gruyter.
- Zillien, Nicole (2008): Die (Wieder-)Entdeckung der Medien – Das Affordanzkonzept in der Mediensoziologie. In: *Sociologica Internationalis* 46 (2), 161–181.
- Zillien, Nicole (2019): Affordanz. In: Liggieri, Kevin und Oliver Müller-Känel (Hrsg.): *Mensch-Maschine-Interaktion. Handbuch zu Geschichte – Kultur – Ethik*. Stuttgart: Metzler, 226–228.
- Zinken, Jörg und Arnulf Deppermann (2017): A cline of visible commitment in the situated design of imperative turns. Evidence from German and Polish. In: Sorjonen, Marja-Leena, Liisa Raevaara und Elizabeth Couper-Kuhlen (Hrsg.): *Imperative Turns at Talk: The Design of Directives in Action*. Amsterdam: John Benjamins, 27–63.
- Zinken, Jörg, Giovanni Rossi und Vasudevi Reddy (2020): Doing more than expected. Thanking recognizes another’s agency in providing assistance. In: Taleghani-Nikazm, Carmen, Emma Betz und Peter Golato (Hrsg.): *Mobilizing Others. Grammar and lexis within larger activities*. Amsterdam: Benjamins, 253–278.
- Zoeppritz, Magdalena (1985): *Computer Talk? Technical Report TN 85.05*. Heidelberg: IBM Germany Heidelberg Scientific Center.
- Zuboff, Shoshana (2018): *Das Zeitalter des Überwachungskapitalismus*. Frankfurt a. M./New York: Campus.

Register

- Adressierung 49, 56, 75, 98, 177, 180–182, 184–185, 188, 193–194, 196, 201, 203, 247, 249, 261, 267, 282, 312, 319, 345, 372, 375–376, 381–382, 386
- Aktivierungswort 83–84, 98–99, 106, 112–113, 119–120, 153–154, 176, 178, 181, 183, 185, 211, 229, 232, 250, 375, 388, 402
- Aneignung 5, 44–47, 49–51, 54–56, 64–66, 68, 91, 93, 96, 153, 315, 389
- Anthropomorphisierung 102–103, 265
- Appropriation 55–56
- Artikulation, dreifache *Siehe* dreifache Artikulation
- Barge-in 77, 224, 230–232, 238, 257, 378, 381, 387
- Beleidigung *Siehe* Invektiven
- Beteiligung 4–6, 25, 35, 40–41, 44, 49, 53, 69, 75, 93, 100, 107, 150, 169, 175–176, 251, 254, 267–270, 275, 282, 286, 291, 300, 307, 312, 318–319, 326, 328, 330, 332, 335, 337, 345, 367, 370–373, 381, 388, 397, 399
- Chatbots 68–74, 85, 196, 386
- Commodification 54–55, 64, 315
- Computer Talk 5, 70, 265, 384–386
- Conditional Voice Recorder 98, 148, 153–154, 166
- Conversion 57
- Datenschutz 21, 63, 68, 92, 104, 107, 110, 116, 150, 156, 403–404
- Domestizierung 4–6, 22–23, 45, 50–55, 57–58, 60, 62–67, 153, 195, 263, 296, 315, 318, 374, 389, 391, 393–394, 397–400
- dreifache Artikulation 53–54, 58, 60
- Einbindung 11, 34, 43, 49, 55, 93, 98, 150, 152, 155, 183, 263, 270, 273–275, 280, 282–283, 371, 374, 380, 400–402, 405
- Embodied Conversational Agents 41, 68–70, 72, 243, 263
- Entity 86, 191, 210, 234
- Ethnomethodologie 122–128, 130–131, 134, 396
- Gender 68, 101–104, 108, 129, 356, 404
- Gespräch 4, 6, 9, 25–26, 28–31, 33–34, 41–42, 44, 47, 128, 131, 135, 138, 143, 146, 152, 174, 189, 192, 194, 204, 220–221, 223, 226, 260, 262, 268–270, 275, 279, 281–282, 285, 300, 303, 312–313, 315, 319, 325, 329–330, 335, 337, 342, 356, 366–367, 371, 378–381, 388–389
- Gestik 17, 38, 56, 59, 139, 141–143, 145, 172–173, 176, 224, 356–357, 370, 399
- Information Retrieval 83, 87, 363, 369
- Intent 85–87, 191–192, 210–211, 226, 234–235, 255, 322
- Interaktion 2, 7, 12, 15–18, 20, 28–29, 31, 33–34, 39, 41–43, 48–49, 55, 60, 70–71, 73–75, 97, 99–100, 104, 109, 122, 126, 128, 130–132, 135, 139–146, 148–150, 153, 177, 180, 182, 204–205, 216, 239–240, 242–243, 251, 263, 265, 269–270, 272, 278, 280, 285, 296, 318–319, 328, 332, 334, 342, 344, 350, 367, 371–373, 379–380, 388, 391, 393, 395–396, 400–401
- Interface 1, 5, 9–10, 18–28, 33, 46, 63, 66, 68–69, 79–80, 82, 85, 89–92, 101, 110, 132, 177, 184, 202, 209, 215–216, 225, 228, 233, 238, 267–269, 279, 286, 291, 318, 348, 352, 355, 361, 373–374, 377, 382–384, 389–392, 400, 402
- Intermediäre 68, 89–91
- Invektiven 306–307, 320, 322, 337, 386–387, 390–391
- Kochen 2, 14, 94, 165, 274, 278
- Kommodifizierung *Siehe* Commodification
- Konversationsanalyse 18, 28, 35, 37, 69–70, 126–128, 130–132, 134–136, 138, 140, 145–146, 175, 181, 205, 221, 261, 396
- Leuchten 112, 118, 153, 156, 363–365, 367
- Medienpraktik 18–19, 21–23, 25, 45, 382–383, 397
- Meta-Interaktionsraum 100, 201, 243, 345, 356
- Moralische Ökonomie 51, 61–63

- Multimodalität 6, 122, 139–141, 144–145, 356, 370, 373, 393, 398–399
- Musik 19, 94, 116, 120, 165, 167, 173, 189, 193–194, 199, 206, 230, 234, 252, 257, 278, 280, 303, 316–317, 322, 369–371, 379
- Natural Language Generation 85
- Natural Language Processing 24, 69, 81, 83, 85, 369
- Natural Language Understanding 85, 363
- no-response-timeout 238
- Objectification 54–56, 59, 189, 367
- Objektifizierung, Objektifikation *Siehe* Objectification
- Partizipand 12–13, 26–27, 41, 44, 64, 223, 238, 248, 261, 267, 279, 286, 298, 337, 342, 356, 364, 372, 378–379, 381, 388–389
- Privatsphäre 68, 92, 104, 108, 110, 403
- Pronominalreferenzen 315, 345, 348, 350
- Raum 7, 19, 46, 52, 75, 78, 95–96, 130, 139–140, 144, 146, 150–151, 165, 168, 170, 211, 274, 307, 357, 370, 393–394, 399–400
- Register 5, 71, 102, 265, 383–387
- Repair *Siehe* Reparatur
- Reparatur 4, 17, 36–37, 39, 57, 70, 73–74, 77, 99–100, 157, 176, 198, 210–211, 231, 238–251, 253–258, 260–262, 267, 313, 322, 330, 345, 371, 375, 377, 379, 381–382, 385
- Sequenzialität 31, 77, 100, 144, 204–205, 216, 219, 221, 250, 262, 268, 297, 375–377
- Sequenzorganisation 6, 196–197, 204–205, 211, 213, 216, 219–220, 376
- Sequenzstruktur 177, 206–209, 216–217, 219–220, 226, 229, 263–264, 267, 284–285, 298, 317, 331, 337–338, 341, 348, 376, 387
- Slot 86, 210–211, 226, 255
- Smartphone 3–4, 25, 82, 90, 94–95, 106, 110, 113–116, 118–120, 149, 151–152, 157, 167, 169, 181, 248, 268–269, 290–291, 302, 305, 325, 347, 370, 395
- Sprachdialogsystem 2, 24, 76, 78
- Störung *Siehe* Reparatur
- Syntax 38, 132, 141, 216, 224, 234, 386
- Testen 9, 26, 39, 44, 93, 182, 201, 283, 285–286, 296, 302–303, 307, 314, 317, 329, 331–332, 339, 341, 345, 347–348, 371, 377, 380, 389, 391–393
- Timer, Timer-Funktion 2, 90, 94, 165, 198, 209–211, 213, 234, 237, 244, 250, 254, 255–256, 274–275, 329, 377
- turn-by-turn talk 98–99, 193, 240, 267, 272–275, 337, 371, 380
- Turn-Taking 128, 143, 145, 176, 179, 197, 212, 220–227, 232, 238, 262, 264, 285, 365, 375, 378
- Üben 9, 14, 26, 296, 371, 296, 380, 389
- Überwachung 42, 66, 107–109
- Vorführen 9, 26, 57, 279, 286, 305, 307, 309, 371, 392
- VUI-Dialog 6, 8, 15–17, 25, 33–34, 38, 43–44, 46, 49, 98, 109, 135, 146, 158, 162, 174–178, 180, 183, 203, 205–209, 212, 215–216, 218, 220, 223, 225–227, 229, 231, 233–237, 240, 242, 245, 263, 266–270, 275, 278, 280–283, 285, 298, 300, 303–304, 313, 318–319, 329, 331–332, 338–339, 345, 369, 371–376, 378–383, 385, 387–393, 396, 399–400, 402
- Wake Word *Siehe* Aktivierungswort
- Wetter, Wetterabfrage 13, 85–86, 88, 90, 94, 96, 114, 236–237, 297, 331