

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: [alexa 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: [alexa 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).

Wie bereits dargestellt, wird mit „übrigens“ eine „graduelle Themenentwicklung“ (Egbert 2003: 199) durch das VUI eingeleitet. Die Fortsetzung durch das VUI ist sequenziell interessant: Die Äußerung ist einerseits thematisch an die vorherige Timer-Sequenz angeschlossen, indem das VUI initiativ beginnt, einen Vorschlag oder eine Instruktion zu entfalten, was bei einer gewünschten Verlängerung der Timer-Zeit gesagt werden kann (016) – genauer kann dies nicht spezifiziert werden, weil der Nutzer die Äußerung des VUI abbricht (018). Zugleich ist die Äußerung handlungslogisch nur lose verbunden, weil die vorhergehende Äußerung mit Blick auf den Handlungsvollzug („Timer einstellen“) abgeschlossen war und die Expansion des VUI dafür nicht nötig gewesen wäre – der Abbruch und die nutzerseitige Reduzierung des Austauschs auf die Dyade lässt sich mit der Perspektive der Sequenzorganisation auch als nutzerseitige Unterbindung der Entfaltung einer Postexpansion lesen. Gleichwohl entsteht dadurch eine Postsequenz, bestehend aus Expansion (016 f.) und Abbruch (018). Das Sequenzmuster für eine Postexpansion sieht also wie folgt aus:

Invocation – (Listening-Modus →) Eingabe – Stimmausgabe/Scharnier – praktische Umsetzung – *Post-Expansion*.

Das gezeigte Beispiel (8) präsentiert einen Fall, in dem das VUI einen „Vorschlag“ unterbreitet – eine Expansion mit dem Ziel der weiteren Nutzung neuer, dem Anwender möglicherweise noch unbekannter Funktionen.²¹⁶ Eine etwas anders gelagerte Postexpansion zeigt sich im bereits diskutierten Beispiel (7) („Schneege-stöber“), auf das hier noch einmal eingegangen werden soll.

Beispiel (7): Schneege-stöber (01-CVR01-01)²¹⁷

```

071      [(16.7)                               ]
072 k:    [((vereinzelt Hintergrundgeräusche)) ]
073 LF:   aLEXa.
074      (1.2)
075 LF:   wie ist das wetter DRAUßen?
076      (1.2)
077 AL:   in amelsbüren beträgt die temperatur
          aktuell minus ZWEI grad celsius bei
          bewölktem himmel,
```

²¹⁶ Für die Analyse eines weiteren Beispiels mit „sequenzeröffnender Expansion“, in dem das VUI eine andere Funktion „vorschlägt“, siehe Habscheid et al. (2021: 38).

²¹⁷ Für eine Betrachtung desselben Beispiels unter dem Aspekt der Anredeformen siehe Kap. 6.11.2.

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 ‚reibungslosen‘ 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 marktgängigen 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 hilfsweise 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 Folgereparatur (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
gesucht;>]
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 Reparaturinitiierung 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 Reparaturinitiierung verstanden würde. Dann allerdings läge die Initiierung 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 Gesprächsbeteiligtenstatus 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 schOtti-sche 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 GRAdе,
 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). Zusammengefasst 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 Stimmeeingaben) 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 Musikkwiedergabe 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.

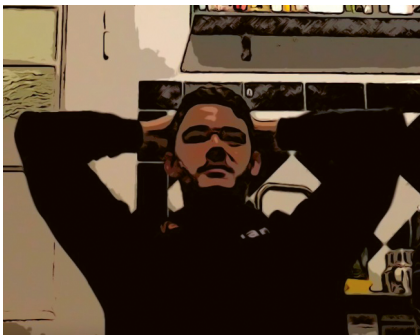


Abb. 26: Sitzposition von Robin bis Z. 642.



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 Numerale³²² 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 AIRdroppen?]
 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 Stimm Ausgaben 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.

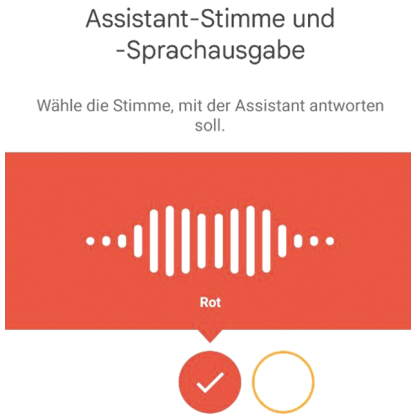


Abb. 30: Stimmauswahl für Sprachausgabe in der Google-Home-App – Rot; Bild: Screenshot / T.H.

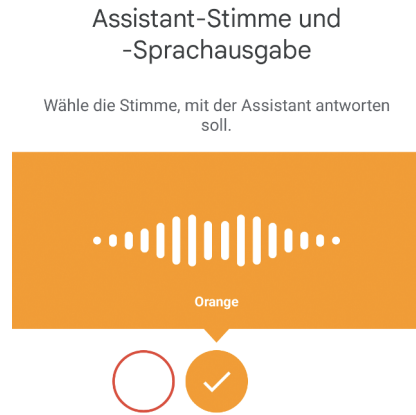


Abb. 31: Stimmauswahl für Sprachausgabe in der Google-Home-App - Orange; Bild: Screenshot / T.H.

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, gegender-

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 Verbalsprachlichkeit 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:  [((Uhrlicken 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 textilartigen 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).

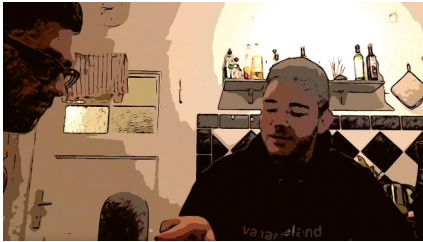


Abb. 33: Robin zeigt auf Leuchten des Google Nest (Perspektive 1).



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 des



Abb. 36: Robin stellt den Smart Speaker in die Mitte des Tisches.

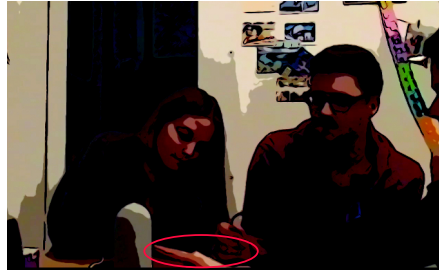


Abb. 37: Nach der Repositionierung zeigt Robin erneut auf die Leuchten.

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.