7 Schlussbetrachtung

Abschließend sollen in dieser dreiteiligen Schlussbetrachtung erstens die Ergebnisse noch einmal entlang der beiden Analyseteile zusammengefasst werden (Kap. 7.1). Darauf aufbauend folgt zweitens eine ausführlichere Diskussion der Ergebnisse, die sowohl konzeptionelle Fragen bearbeitet als auch methodologische Reflexionsarbeit leistet und daraus hervorgehende Anschlussfragen extrapoliert (Kap. 7.2). In einem Ausblick soll drittens von den Gegenständen der Arbeit ausgehend ein kurzer Blick auf rezente technologische Entwicklungen und Prognosen geworfen werden – und somit auf Phänomene, zu deren Erforschung erneut eine interdisziplinär arbeitende, empirische Sprachwissenschaft wichtige Beiträge leisten kann und sollte (Kap. 7.3).

7.1 Zusammenfassung der Ergebnisse

Die Arbeit fragte nach den sprachlichen Praktiken im Prozess der Domestizierung von Medien mit Voice-User-Interfaces und stellte dabei insbesondere auf die Einrichtung und Nutzung von Smart Speakern ab. Zentrale Erkenntnisinteressen der Arbeit waren, (1) welche sprachlichen Praktiken sich bei der Domestizierung stationärer Sprachassistenzsysteme mit VUIs zeigen, (2) inwieweit diese – zwischen Transformation und Emergenz – als Abwandlungen bereits untersuchter sprachlicher Praktiken beschrieben werden und inwieweit sie als emergente sprachliche Praktiken gelten können, sowie (3) wie Smart Speaker sprachlich zu Beteiligten an der sozialen Praxis werden. Dazu wurden im ersten Teil der Analyse ausschließlich dyadische Dialoge mit VUIs untersucht, um sprachliche Praktiken in der Bedienung von VUI-Geräten ausmachen zu können. Zu diesem Zweck wurde eine Kollektion von dyadischen VUI-Dialogen herangezogen.[351] Im zweiten Teil wurden dann auch Mehrparteienkonstellationen in den Blick genommen, um die Domestizierung und die Einbindung in die soziale Praxis *in situ* nachvollziehen zu können. Die Ergebnisse aus dem ersten Analyseteil konnten dabei nutzbar gemacht werden.

351 Wie bereits weiter oben reflektiert, kann für die ausgewählten Aufnahmen nicht sicher ausgeschlossen werden, dass neben VUI und Anwender*in noch andere Beteiligte während des VUI-Dialogs anwesend sind, was auch Auswirkungen auf die Gestaltung des VUI-Dialogs haben kann. Andere Beteiligte sind aber auf den ausgewählten Aufnahmen nicht zu hören und werden von den Sprecher*innen nicht relevant gesetzt.

https://doi.org/10.1515/9783111574332-007

7.1.1 Dyadische VUI-Dialoge

Die Analysen zu dyadischen VUI-Dialogen waren entlang von vier konversationsanalytischen Grundbegriffen strukturiert: Anredeformen, Sequenzialität, Turn-Taking und Reparaturen. Die Auswahl dieser Begriffe erfolgte aus den Daten heraus und auf Basis der von Schegloff (2006: 71) skizzierten „generic organizations of practice", die als praktische Lösungen überall dort aufzufinden sind, wo stabile Interaktionssituationen entstehen. Schegloff (2006: 71) entwickelt insgesamt sechs solcher Problembereiche, von denen auf Grundlage der Daten drei als besonders relevante Untersuchungsfelder ausgewählt wurden: die sequenzielle Organisation, das Turn-Taking sowie Reparaturen. Schegloff bezieht sich dabei auf zwischenmenschliche Interaktionen, für die jeder der genannten Problembereiche gesprächslinguistisch umfangreich beschrieben ist. Auch die sprachwissenschaftliche Literatur zum Austausch zwischen Menschen und Maschinen nimmt immer wieder auf diese Bereiche Bezug. Entsprechend eigneten sie sich nicht nur, um die Analyse sprachlicher Praktiken in VUI-Dialogen zu ordnen, sondern auch, um zu prüfen, wie die sprachlichen Praktiken sich als Abwandlungen bereits beschriebener sowie als emergente Praktiken darstellen.

Den Bereich der *Anredeformen* nennt Schegloff (2006) nicht als eine der „generic organizations". Dieses Analysefeld ergab sich aber als relevante Kategorie sowohl aus den Daten als auch aus der konversationsanalytischen Literatur heraus: Adressierungen, insbesondere in Form onymischer Anreden, traten als wiederkehrende kommunikative Aufgabe auf, die im Fall der VUI-Dialoge auch vor dem Hintergrund des Mithörens der Geräte in den privaten Wohnumgebungen von besonderer Relevanz ist (vgl. Waldecker/Volmar 2022): Um auf ein Aktivierungswort hin reagieren zu können, muss das Mithören dauerhaft erfolgen und das Gerät die empfangenen akustische Signale daraufhin prüfen, ob das Aktivierungswort genannt wird. Durch die Invokation, d. h. die Nennung und Erkennung eines gerätespezifischen Aktivierungsworts, verändert sich der Status des Geräts: Es verarbeitet dann alle folgenden akustischen Signale als sprachliche Eingaben. Dieser Modus muss nach jeder in sich geschlossenen sprachlichen Eingabe erneut hergestellt werden. Entsprechend lassen sich zur Invokation der Geräte sprachliche Praktiken der Adressierung finden, die zwar einerseits mit der Veränderung des Status einer Maschine vom Ruhe- in den Eingabemodus eine neue kommunikative Funktionalisierung aufweisen. Andererseits werden Invokationen praktisch als Summons-Answer-Sequenz verfertigt.

Die sprachlichen Praktiken, mit denen Summons-Answer-Sequenzen realisiert werden, haben konsequenterweise Ähnlichkeiten mit sprachlichen Praktiken aus zwischenmenschlichen Interaktionssituationen: Onymische Formen, ggf. in Verbindung mit einer Gesprächspartikel, sind als Summons auch in zwischen-

menschlichen Summons-Answer-Sequenzen beobachtbar, ebenso wie nonverbal produzierte Antworten (im Falle der VUIs das Aufleuchten entsprechender Elemente am Gerät, im Falle zwischenmenschlicher Interaktionen etwa eine körperliche Drehung oder ein Blick). Hierbei handelt es sich also um eine Abwandlung sprachlicher Praktiken aus einem bereits bekannten Repertoire. Als emergente sprachliche Praktik lässt sich die Verwendung der (allerdings ohnehin semantisch sehr weiten) Gesprächspartikel *okay* bei einer der beiden möglichen Invokationen für das Gerät von Google verstehen. Es lassen sich darüber hinaus nicht nur onymische Anreden im Rahmen von Summons-Answer-Sequenzen feststellen, sondern auch Anreden seitens des VUIs, die ebenfalls keine sprachliche Emergenz erkennen lassen. Vielmehr können die Verwendung onymischer Adressierungen und Du-Anreden durch die VUIs, in die im Ersteinrichtungsdialog auch die Nutzer*innen eingelernt werden, teilweise als sprachliche Verfahren verstanden werden, die Ausdruck eines auch seitens der Gerätehersteller vorangetriebenen Domestizierungsprozesses sind.

Auch im Hinblick auf die anderen drei untersuchten Bereiche sollen die Ergebnisse kurz zusammengefasst werden. Für das Untersuchungsfeld der *Sequenzialität* und der *Sequenzorganisation* lassen sich erstens Befunde bestätigen, die etwa von Krummheuer (2008; 2010), Gehle et al. (2015), Opfermann/Pitsch (2017), Pitsch (2020) bereits an anderen Geräten herausgearbeitet werden konnten: Der Austausch ist von einer Trägheit gekennzeichnet, die sich aus der Sequenzorganisation ergibt. Um damit verbundene sprachliche Praktiken genauer zu beleuchten, wurde das Teilkorpus der dyadischen Dialoge auf Muster im Ablauf von VUI-Dialogen hin untersucht. Als stabile Einheiten erweisen sich zwei Varianten eines Ablaufmusters, das als Basis-Sequenzstruktur für VUI-Dialoge beschrieben wurde, die nutzer*innenseitig initiiert werden:[352]

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

Diese Ablaufmuster können als emergente sprachliche Praktik verstanden werden, die nicht nur die Nutzer*innen immer wieder vollziehen, sondern an der sich – wie sich an den Dialogen zur Ersteinrichtung zeigen ließ – auch die Anbieter der Systeme orientieren. Dies hat einen zirkulären, selbstverstärkenden Ef-

352 Im gesamten Korpus (auch über die Kollektion der dyadischen VUI-Dialoge hinaus) ist – abgesehen von den Ersteinrichtungsdialogen, die einen Spezialfall darstellen, – ein einziger Fall eines VUI-seitig initiierten VUI-Dialogs dokumentiert. Dies könnte bei anderen Nutzer*innen auch anders gelagert sein, es wird jedoch aufgrund des geringen Vorkommens in meinen Daten auf die Formulierung einer Sequenzstruktur für diesen Typ VUI-Dialog verzichtet.

fekt: Im Design der Anwendungen wird mit solchen praktischen Verfestigungen gearbeitet, die sich durch die Gestaltung der Dialogsysteme wiederum weiter verstetigen können. Wie sich in den Daten zeigt, ist eine sequenzielle Kohärenz, die über dieses Ablaufmuster hinausreicht, nicht nachweisbar. Zwar ergaben sich an einigen Stellen Optionen für Einschübe (v. a. zur Durchführung von Reparaturen) und systemseitige Post-Expansionen, wobei für Letztere gezeigt werden konnte, dass sie Formen phatischer Kommunikation mit dem Zweck der Kund*innenbindung dienen können. Sequenzielle Verknüpfungen mehrerer Abläufe zur Kombination verschiedener, aufeinanderfolgender Stimmeingaben miteinander waren hingegen in dyadischen Konstellationen nicht nachweisbar. Für die sprachliche Gestaltung des Austauschs hat dies zur Folge, dass miteinander verknüpfte Anliegen, die mehrere Züge erfordern, in mehrere Teile zerfallen, die alle einzeln einer der beiden Varianten des Ablaufmusters entsprechen. VUIs können also Eingabe und Ausgabe als Paarsequenzen prozessieren; mit Sequenzialität, die jenseits von Sequenzen etwa durch konversationelle Implikaturen entsteht, sind die VUIs jedoch überfordert, wie auch der Abschnitt zu Reparaturen zeigen kann (siehe auch Habscheid 2022: 191). Entsprechend ist die Sequenzialität mit Krummheuer (2010: 229) als „aufgebrochen“ gut beschrieben: Sie ist auf die Durchführung einzelner Eingabe-Ausgabe-Prozesse hin ausgerichtet. Die Nutzer*innen passen sich dieser Präkonfiguration des Austauschs mit dem Interface einerseits an, andererseits entstehen gerade durch diese Eigenschaft der VUIs notwendige Reparaturen und ggf. Abbrüche; ferner testen Nutzer*innen – wie in den Mehrparteienkonstellationen noch deutlicher hervortritt – die Grenzen dieser Limitierung.

In der Gestaltung der einzelnen Turns bzw. *turn-constructional units* (TCUs) ist die Orientierung an der praktischen Verfertigung einer mit dem Sprechen vollzogenen Aktivität ebenfalls sichtbar, und zwar dann besonders deutlich, wenn die an das VUI gerichtete Stimmeingabe aus einem einzelnen freistehenden Lexem (z. B. „Licht“) oder kurzen, syntaktisch offenen Einheiten (z. B. „Timer löschen“) besteht. Nachweisbar sind darüber hinaus aber auch deontische Infinitive, Imperativ- und Fragesätze; nur außerhalb des Kontexts der dyadischen Nutzungssituationen zeigten sich sprachliche Praktiken außerhalb dieser Typologisierung.[353] Zusätzlich zeigt sich eine Tendenz zur Kürze in der Gestaltung der Stimmeingaben. Damit sind keine emergenten sprachlichen Praktiken beschrieben, sondern es zeigt sich vielmehr, dass sprachliche Praktiken in den Dienst anderer Praktiken gestellt werden, u. a. der Steuerung von Smart Home-Elementen oder Erinnerungsfunktionen. Die

353 Dies wiederum kann damit erklärt werden, dass im Rahmen der Ersteinrichtung und kurz danach das VUI zunächst ausgetestet, teilweise gezielt überfordert und an seine Grenzen geführt wird. Dabei kommt es zu einer höheren sprachlichen Variabilität als in der über die CVR-Aufnahmen dokumentierten routinisierten Nutzung.

sprachlichen Formen ähneln oder gleichen gar solchen, die auch in anderen Kontexten beobachtbar sind, in denen sprachliche Praktiken empraktischer Bestandteil von sozialen Praktiken sind, die kein Gespräch konstituieren (vgl. Goffman 1979: 14). Die knappe Gestaltung der Äußerungen muss nicht nur aus sprachökonomischen Gründen erfolgen, sondern es kann eine Kombination verschiedener Motive dahinterstecken; Baldauf (2002: 124) unterscheidet fünf: neben Sprachökonomie auch Rücksichtnahme auf andere Tätigkeiten, eine Sprecher*innenzentrierung in der Äußerung, gestalterische Absichten und beziehungsbezogene Motive. Im Hinblick auf die VUI-Dialoge wären als Motive für knappes Sprechen zu ergänzen: die Reduzierung der Fehleranfälligkeit (sowohl in der Artikulation der Spracheingabe als auch im Verständnis durch das VUI) und die Konzentration auf den durchzuführenden Prozess(schritt) sowie – verwandt mit dem von Baldauf beschriebenen Beziehungsmotiv – die Aufhebung der Notwendigkeit von „Face-Work“ (Goffman 1955) zwischen den Beteiligten.

Das Turn-Taking wird durch diese Art der Gestaltung von TCUs erleichtert, was reziprok auch zu einer Gestaltung der Turns in der entsprechenden Weise führt: Nutzer*innen orientieren sich an den notwendigen Parametern für eine Verarbeitung der Stimmeingaben, wodurch die Erfolgsquote (auch in der Rederechtsverteilung) steigt und die Nutzer*innen wiederum eher dazu tendieren, die erfolgreiche Turn-Gestaltung beizubehalten. Der Sprecher*innenwechsel fügt sich in die durch das Reden vollzogenen Handlungsschritte ein und scheint sich außerdem durch eine höhere Latenztoleranz seitens der Nutzer*innen auszuzeichnen, während die Latenztoleranz bzw. -flexibilität seitens der maschinellen Partizipanden geringer ist. Ob neben pragmatischen und segmentierenden Verfahren zur Anzeige von *transition-relevance places* (TRPs) auch andere *cues* produktiv sind (syntaktische Geschlossenheit, Intonationskontur o. Ä.), konnte im Rahmen der Untersuchung nicht gezeigt werden; die sprachlichen Praktiken der Nutzer*innen ließen hier ebenso wenig Rückschlüsse zu wie die Stimmausgaben der VUIs.

Eine andere sprachliche Praktik war in diesem Zusammenhang jedoch auffällig, worauf auch ihr Status in der Literatur hindeutet: Die 63 Aufnahmen umfassende Kollektion der dyadischen VUI-Dialoge dokumentiert insgesamt 133 VUI-Dialoge und darin acht Fälle von *barge-ins*, d. h. aktives Unterbrechen der Äußerungen des VUIs durch die Produktion einer neuen Invokation und Stimmeingabe. Die Notwendigkeit einer Aushandlung der Rederechtsverteilung, zu der es nach Schegloff (2000b) bei Überlappungen in zwischenmenschlichen Interaktionen kommt, ist hier suspendiert: Das VUI stoppt immer die Äußerungsproduktion bei Erkennung einer neuen Invokation und die neue Stimmeingabe überschreibt das vorherige kommunikative Projekt – ohne dass die Nutzer*innen dafür einen *account* produzieren. Diese sprachliche Praktik kann als Ausdruck einer Hierarchie zwischen Nutzer*innen

und VUIs verstanden werden, die auch in der Gestaltung der Stimmeingaben teilweise deutlich wird – etwa in der Eingabe „Stop“, die in allen untersuchten Haushalten regelmäßig auftritt und zur Beendigung der Äußerungsproduktion oder der aktuell durch das VUI abgespielten Musik sowie anderer laufender VUI-Aktivitäten dient und die widerspruchslos umgesetzt wird. Aushandlungen des Rederechts sind insofern teilweise zugunsten der Nutzer*innen aufgehoben (siehe auch Waldecker/Hector/Hoffmann 2024).

Im Kapitel zu *Reparaturen* konnte schließlich noch einmal die Trägheit des Austauschs genauer untersucht werden. Dazu wurden zunächst verschiedene Fälle unterschieden, sowohl danach, wer die reparaturbedürftige Stelle produziert (Nutzer*innen oder VUI), als auch danach, wer die Reparatur initiiert oder durchführt (angelehnt an konversationsanalytische Literatur, insbesondere Egbert 2009). Es zeigte sich, dass hyperlokal operierende Reparaturmechanismen aus der zwischenmenschlichen Interaktion, in denen z. B. einzelne Laute oder Lexeme in Sekunden ersetzt werden, nicht bzw. nur sehr eingeschränkt angewendet werden; eine Ausnahme stellten selbstinitiierte Selbstreparaturen der Nutzer*innen dar. Abgesehen davon wurden von den Nutzer*innen Reparaturen immer durch die Wiederholung einer gesamten Stimmeingabe realisiert, selbst dann, wenn nur ein einzelnes Lexem darin reparaturbedürftig war (dies wurde z. B. prosodisch oder syntaktisch als Reparandum markiert). Dies führt zu häufigen Wiederholungen des gesamten Basis-Sequenzablaufs, einschließlich der Invokation – eine Folge der maschinenseitig nicht bestehenden Kohärenz zwischen den einzelnen Sequenzen. Dieses Verfahren hat zudem zur Folge, dass stellenweise unklar bleibt, was als Reparandum zählt: die Stimmeingabe der Nutzer*innen oder die Stimmausgabe der VUIs. Dies ist ein Hinweis darauf, dass Reparaturen auch deswegen interessant sind, weil sie die sequenzielle Ordnung zwischen den beiden ‚Sprecher*innen‘ hinterfragen und die Ungleichheiten zwischen Nutzer*innen und VUI als ‚Beteiligte‘ an der Kommunikation deutlich werden lassen. Dazu trägt auch bei, dass die Reparaturen nicht immer erfolgreich sind: Es zeigen sich mehrere Fälle, in denen das kommunikative Projekt nach wiederholten Reparaturversuchen nutzer*innenseitig abgebrochen wird.

7.1.2 VUI-Dialoge in Mehrparteienkonstellationen

Der zweite Analyseteil widmete sich v. a. der Frage, ob und wie Smart Speaker sprachlich an der sozialen Praxis und an Gesprächen ‚beteiligt‘ werden. Entsprechend standen hier Mehrparteienkonstellationen im Fokus; ausgewertet wurden in diesem Teil Aufnahmen, auf denen mehrere Sprecher*innen zu hören waren. Für die Analysen wurde das Konzept der „materiellen Partizipanden des Tuns“

nach Hirschauer (2004; 2016) genutzt. Demzufolge können an der sozialen Praxis und der Durchführung einer Praktik menschliche wie nicht-menschliche „Partizipanden" mit unterschiedlichen Aktivitätsniveaus teilnehmen – sie sind Teil eines Praxisvollzugs. Dieser Vollzug gestaltet sich, wie schon in den Analysen im ersten Teil gezeigt werden konnte, in an der Oberfläche gesprächsartigen Sequenzen von akustischen Äußerungen, die die Anwender*innen und die VUIs produzieren. Mit diesem Umstand gehen die Anwender*innen von VUIs unterschiedlich um, wie sich in den Daten zeigte: Einerseits können die Anwender*innen nahtlos von ihrem *turn-by-turn talk* in eine an das VUI gerichtete Stimmeingabe übergehen (und zurück). Darüber hinaus konnte ein Fall beschrieben werden, in dem die sequenzielle Verfertigung des VUI-Dialogs auf der einen und die der zwischenmenschlichen Interaktion auf der anderen Seite sich kreuzen. Das spricht dafür, dass die sprachlichen Praktiken in VUI-Dialogen für die ko-präsenten Sprecher*innen als solche erkennbar sind, sodass sie nicht unbedingt einer Meta-Kommentierung bedürfen. Möglich sind aber auch die Produktion von prä-hoc- und post-hoc-Einbindungen, d. h. vorherigen oder anschließenden sprachlichen Verknüpfungen zwischen der laufenden sozialen Praxis und der Anwendung. Diese entstehen, wenn die Nutzung des VUIs stärker ins Zentrum rückt.

Insbesondere die Fälle, in denen das VUI im Zentrum der sozialen Praxis steht und auch sprachlich an diese angebunden wird, wurden einer genaueren Betrachtung unterzogen. In Mehrparteienkonstellationen zeigt sich noch einmal deutlich, dass die Geräte schon per Design nicht darauf ausgerichtet sind, die Eingaben von mehreren Nutzer*innen gleichzeitig zu verarbeiten. Entsprechend vollziehen sich auch hier dyadische VUI-Dialoge, die den beschriebenen Basis-Sequenzstrukturmustern folgen und zugleich in Mehrparteienkonstellationen eingebunden werden. Dabei wurden verschiedene soziale Praktiken identifiziert, in denen das VUI im Zentrum steht, mit entsprechenden Folgen für die sprachliche Gestaltung des Austauschs. Dazu zählen neben den Ersteinrichtungen des Smart Speakers auch das Üben bzw. Testen, Vorführungen der Smart Speaker für andere Personen (insbesondere Gäste), gemeinsame Bewertungen der VUI-Performanz und die interaktionale Bearbeitung von Fehlschlägen. In diesen Analysen war die Unterscheidung zwischen ‚Reden' und ‚Gespräch' (Goffman 1979: 6–7) wichtig: Für die soziale Praktik eines Gesprächs ist Reden zwar notwendig (wenn auch nicht fortlaufend und nicht notwendigerweise stimmlich), nicht jedes Reden konstituiert umgekehrt aber ein Gespräch, sondern es kann auch empraktischer Teil anderer Praktiken sein – mit entsprechenden Folgen für die sprachliche Gestaltung.

Nun zeigen sich aber gerade in den genannten Situationen Verknüpfungen in der – eben inkrementellen, situations -und materialgebundenen – sozialen Praxis: Zwischen dem gesprächsartigen Reden zur Bedienung eines VUI und der so-

zialen Praktik des Gesprächs bestehen Verbindungen. Das Sprechen *mit* dem VUI verband sich dann mit dem Gespräch *über* das VUI. Dies zeigt sich z. B. in der von den VUIs produzierten Eröffnung und Rahmung der Ersteinrichtung als gemeinsame Aktivität, die von den Anwender*innen gemeinsam mit dem VUI durchgeführt wird. Ebenso zeigt sich dies in den häufig daran anschließenden ersten Versuchen der Nutzung des Geräts, die mit den Ko-Beteiligten abgesprochen wird (vgl. Pitsch et al. 2017); das VUI wird hier in besonderer Weise zum ‚Beteiligten' an der sozialen Praxis. Sprachlich gestalten die Nutzer*innen dies u. a. durch direkt an das VUI gerichtete Adressierungen, Rezeptionssignale der Nutzer*innen, kohärente Anschlussturns an Äußerungen des VUIs und andere Formen, die auch in Vorführungssituationen und Bewertungen auftreten. Besonders auffällig ist die Produktion solcher Formen auch in Situationen, in denen die Nutzung des VUIs nicht gelingt. Zwar gelten weiterhin die Erkenntnisse, die aus den dyadischen Konstellationen für Reparaturen in VUI-Dialogen gezeigt wurden. Hinzu kommt aber eine deutlich beobachtbare Tendenz, unpassende oder falsche Äußerungen des VUI in Mehrparteienkonstellationen klar als solche zu kennzeichnen und dabei ebenfalls auf sprachliche Formen zuzugreifen, die die Geräte teilweise in laufende Gespräche zwischen den menschlichen Beteiligten eingliedern und sie so zu Partizipanden werden lassen.

Schon bei der situationsspezifischen Betrachtung wird deutlich, was durch die daran anknüpfende Untersuchung der genauen sprachlichen Praktiken zur Herstellung dieser Form der Partizipation noch einmal bestätigt wird: Der Beteiligungsstatus, der hier entsteht, ist nicht von Dauer, vielmehr ist er brüchig. Bei der Bedienung von VUIs werden zwar eine Reihe sprachlicher Praktiken eingesetzt, mit denen ein gewisser Grad an ‚Beteiligung' entsteht. Dieser wird jedoch auch sprachlich nicht konsequent umgesetzt: Kontinuierlich wechselnde Genus-Verwendungen, uneingeleitete Wechsel von Hörer- zu Objektdeixis, überlappende Äußerungen (auch die bereits erwähnten *barge-ins*) und auch objektifizierende Zeigegesten verdeutlichen, dass der zugeschriebene Status nur an der Oberfläche besteht und in den untersuchten Daten auch keine weiteren Sozialfolgen zeitigt.

7.2 Diskussion der Ergebnisse

Im Folgenden sollen die oben zusammengefassten Befunde der Arbeit konzeptionell diskutiert werden – zunächst vor dem Hintergrund der eingeführten und in den Analysen verwendeten Konzepte, aber auch mit Blick auf Implikationen und Konsequenzen, die sich ggf. aus den Befunden auch für andere Konzepte ergeben können. Außerdem wird die Arbeit kurz methodologisch reflektiert und es wer-

den mögliche Folgefragen und Anschlussuntersuchungen diskutiert, die sich aus den Befunden der vorliegenden Arbeit und ihrer Diskussion ergeben haben. Theoretische Ausganspunkte der Arbeit waren neben dem Konzept der „Ko-Operation“ (Goodwin 2018) jeweils Begriffe von sozialen Praktiken, sprachlichen Praktiken und Medienpraktiken. Praktiken wurden dabei in ihrem reflexiven Spannungsverhältnis von übersituativer Einheiten- und Regelbildung auf der einen und situativen Aufführungen in der Praxis auf der anderen Seite verstanden (vgl. Schatzki 1996; Selting 2016; Schüttpelz/Meyer 2017). Mit dieser Perspektive konnte nachvollziehbar gemacht werden, welche sprachlichen Praktiken VUIs mit hervorbringen, wie VUIs situativ an der Praxis beteiligt werden und wie sie sich dadurch in soziale Praktiken einschreiben – und mithin domestiziert werden. Darauf soll nun im Einzelnen eingegangen werden.

7.2.1 Sprachliche Praktiken – Medienpraktiken im Interface

Das praxeologische Grundgerüst hat sich als stabile Ausgangslage erwiesen, um die ko-operative Verfertigung der Medien in den Blick zu nehmen und so die Eingliederung von VUIs in den Sprachgebrauch zu untersuchen. Es konnte beschrieben werden, wie sich in VUI-Dialogen sowohl neue sprachliche Praktiken ausbilden als auch sprachliche Praktiken aus dem Repertoire der Anwender*innen für die Anwendung im Kontext der VUI-Nutzung herangezogen wurden. Es konnte außerdem gezeigt werden, dass Nutzer*innen implizites Wissen[354] über Gesprächsorganisation und damit gekoppelte sprachliche Praktiken mit dem (zum Teil neu erworbenen) Wissen über die Möglichkeiten und Grenzen der Interfaces verbinden. Es wurde beobachtbar gemacht, wie Nutzer*innen einerseits bekannte Strategien der Gesprächsorganisation zum Gebrauch im VUI-Dialog ‚verlängern‘ – z. B. in der Gestaltung der Redezugorganisation oder bei Adressierungen –, andererseits dabei aber auch ihre Vorstellungen über die Systeme berücksichtigen, was z. B. im Rahmen von Reparaturen deutlich sichtbar wird (siehe auch das Beispiel bei Merkle/Hector 2025). Sprachliche Praktiken *leiten* (vgl. Distelmeyer 2020) insofern von kognitiven bzw. kognitionslinguistischen Prozessen und den sozialen Gegebenheiten in privaten Haushalten über die VUIs[355] hin zur cloudbasierten Datenspeicherung und -verwertung der

354 Implizites Wissen wird hier in der Verwendung von Ernst (2017) mit Bezügen zu Polanyi (1985[1966]), Collins (2010) und Loenhoff (2015) gebraucht. Als Ausgangsdefinition muss dabei genügen, was Ernst (2017: 99) zusammenfasst: „‚Implizit‘ ist dieses Wissen, weil es nicht bruchlos ‚expliziert werden kann‘“; siehe auch Kap. 2.1.4.

355 Involviert in diesen Prozess sind neben einem VUI eine unüberschaubare Anzahl weiterer Interfaces, die Verbindungen zwischen Soft- und Hardware herstellen (vgl. Cramer/Fuller 2008:

Systemanbieter, um dort u. a. als Basis der Dienstanbieter für kontinuierliches Interface-Design zu dienen. Von dort wiederum fließen darauf basierende Updates der VUIs – z. B. in der Speech Recognition und der Stimmsynthese – zu den Nutzer*innen zurück und können wiederum Modifikationen der sprachlichen Praktiken evozieren. Mit der Betrachtung dieser reflexiv verlaufenden Aneinanderreihung von Abläufen werden sprachliche Praktiken verständlich als Medienpraktiken. Sie sind der nutzer*innenseitig beobachtbare Bestandteil von „Interfacing" bei VUIs und Arbeit an den Bedingungen für Kooperation. Mit anderen Worten: Sprachliche Praktiken *sind* Medienpraktiken. Die Praktiken bekommen durch ihren Gebrauch in VUI-Dialogen also vor dem Hintergrund eines relationalen Interface-Begriffs[356] einen besonderen Status als reflexive, praktisch verfertigte Verbindung technischer Prozesse und Artefakte mit sozialen und kognitiven Gegebenheiten: „In Interfaces finden Übergänge zwischen situierter Kognition und dem impliziten Regelwissen von Praktiken statt" (Ernst 2017: 101): Die Nutzer*innen, so ließ sich anhand der Praktiken (wenn auch implizit) beobachten, verbinden im VUI-Dialog ihr implizites (sprachliches, gesprächsorganisatorisches) Regelwissen mit den situativ durch ein VUI gestellten kommunikativen Aufgaben. Dabei war keineswegs das Anliegen, auf Grundlage der hier vorgelegten Untersuchung trennscharf zu entscheiden, wo implizites Wissen anfängt und wo „situierte Kognition" (siehe Ataizi 2012) aufhört – ein Zugriff auf kognitive Prozesse war generell nicht Anliegen der Arbeit. Vielmehr war die Beschaffenheit der sprachlichen Praktiken von Interesse, die aber von diesen Prozessen an der Oberfläche zeugen.

7.2.2 Sprachliche Praktiken in VUI-Dialogen als Sprachregister?

Damit ist ein anderer Aspekt des sprachlichen Austauschs zwischen Menschen und Maschinen angesprochen: Die seit den 1980er-Jahren mit dem Aufkommen von „natural language interfaces" (Fischer 2006: 1) immer wieder gestellte Frage nach dem empirischen Gehalt eines eigenen *sprachlichen Registers* oder einer eigenen *Varietät* für den Austausch mit Computern (etwa Zoeppritz 1985; Krause/Hitzenberger 1992). Als Varietät versteht Fischer (2006: 25) mit Crystal (2001: 6–7) allgemein ein System sprachlicher Ausdrücke, deren Gebrauch durch verschiedene Einflussfaktoren bestimmt wird. Als Register wird genauer gesagt eine sprachliche Variation gefasst, die spezifische Eigenschaften aufweist, die nicht

149), aber aufgrund des sprachwissenschaftlichen Erkenntnisinteresses dieser Arbeit nicht näher untersucht worden sind.

356 Siehe ausführlich Kap. 2.1.4.

etwa durch geografische oder soziokulturelle Gegebenheiten bestimmt werden, sondern durch die Umstände im Vollzug des Sprechens (vgl. Coseriu 1988 [1980]: 25). Register sind mithin eine *gebrauchs*spezifische Varietät (vgl. Dittmar 2004: 217–218)[357] – Gegenstand ist also die Frage, ob der Gebrauch von Sprache im Austausch mit Computern hinreichend spezifische sprachstrukturelle oder funktionale Folgen zeitigt, um als eigenes Register – „Computer Talk" nach Zoeppritz (1985) – gelten zu können.

Krause (1992b: 16) geht spezifisch von der Entstehung eines *vereinfachten* Registers (*simplified register*) aus, ein Begriff, der von Ferguson (1975; 1982) vertreten und z. B. zur Beschreibung von *Foreigner Talk* oder *Baby Talk* herangezogen wurde (vgl. auch Krause 1992b: 4). Parallelen bestehen auch zu Eingaben bei der Benutzung von Datenbanken bzw. Suchmaschinen, die sprachlich als Schlüsselwort-Suche realisiert werden und u. a. an Booleschen Operatoren orientiert sind.[358] Allerdings ist dies für die meisten Systeme nicht mehr notwendig; für Fragehandlungen ist ein beginnender Wandel von Schlüsselwort-Suchen zu ‚natürlichsprachigen' konversationellen Fragen beobachtbar, auch wenn die Schlüsselwort-Sucheingaben noch zu überwiegen scheinen (vgl. White/Richardson/Yih 2015).[359] Die Formulierung von Suchanfragen im Allgemeinen weist außerdem insgesamt eine hohe sprachliche Variabilität auf, deren Entstehung und Einflussfaktoren noch nicht umfassend verstanden sind (vgl. Alaofi et al. 2022).

Durch die hier untersuchte *mündliche* Gestaltung von VUIs und den hohen Grad an Dialogizität im Interface könnte eine neue Spezifik auftreten, die bisher nicht ausführlich im Blickfeld der Forschung lag – eine der wenigen Untersuchungen, die zu dieser Frage beiträgt, kommt von Opfermann et al. (2017), und zeigt, dass Nutzer*innen sich in Dialogen mit einem Embodied Conversational Agent nicht durchgängig von einem *simplified register* Gebrauch machen, sondern auch Praktiken aus zwischenmenschlichen Interaktionen anwenden. Dies kann die vorliegende Untersuchung grundsätzlich bestätigen: Zwar sind auf der

357 Die begriffliche Debatte u. a. in der Soziolinguistik um Varietät, Register und Stil mit unterschiedlich nuancierten Vorschlägen soll hier nicht näher dargestellt werden (siehe etwa Auer 2012), es bleibt bei den Begriffsbestimmungen aus den bisherigen Arbeiten zu „Computer Talk" (Zoeppritz 1985).

358 Als Boolesche Operatoren für Suchmaschinen werden u. a. „UND", „NICHT" und „ODER" genutzt, mit denen einzelne Suchworte miteinander in Beziehung gesetzt werden. Diese kamen bereits in den 1970er-Jahren bei Interfaces für Bibliothekskataloge zum Einsatz und werden bis heute mehr oder weniger explizit bei der Gestaltung von Suchanfragen verwendet (vgl. Kerssens 2017: 222).

359 Dabei ist zu beachten, dass die Ergebnisse auf Suchanfragen beruhen, die zwischen November 2011 und Januar 2013 durchgeführt wurden. Eine Fortsetzung des von White/Richardson/Yih (2015) beobachteten Trends zur Formulierung konversationeller Fragen scheint also wahrscheinlich.

einen Seite durchaus wiederkehrende Charakteristika in der Gestaltung von Stimmeingaben beobachtbar. In der Analyse der vorliegenden Arbeit konnten insbesondere Ein-Wort-Eingaben (Nomen oder Adjektive), syntaktisch nicht geschlossene bzw. elliptische Phrasen (u. a. deonitische Infinitive), Imperativsätze sowie Ergänzungs- und Entscheidungsfragen beobachtet werden.[360] Bilden diese Praktiken jedoch zusammengenommen ein sprachliches Register bzw. eine sprachliche Varietät mit spezifischen Merkmalen? Diese Auflistung sprachlicher Praktiken ist auf der anderen Seite durchaus recht breit, zudem könnten auch weitere, in den hier ausgewerteten Daten nicht beobachtbare Praktiken auftreten. Darüber hinaus sind sie nicht unbedingt spezifisch für die Anwendung in VUI-Dialogen oder in Mensch-Computer-Konstellationen überhaupt. Was sich als Charakteristikum verallgemeinernd festhalten lässt, ist eine Tendenz zum „knappen Sprechen" (vgl. Baldauf 2002) in Verbindung mit Sprechen, das als „empraktisch" (Bühler 1965[1934]: 155–156) bekannt ist und insofern einen hohen Grad an Handlungsorientierung aufweist bzw. selbst die Handlung vollzieht und „Face-Work" (Goffman 1955) (temporär) suspendiert. Dies wurde linguistisch aber auch für andere Kontexte untersucht, z. B. für medizinische Operationen (siehe Mondada 2014b) oder Fahrschulstunden (vgl. Deppermann 2018a). Hier lässt sich insofern auch eine Ähnlichkeit zu instruktivem Sprechen feststellen, das in beiden von Mondada und Deppermann untersuchten Kontexten auftritt. Sie eint auch, dass ein höherer Aufwand zur Vermeidung von Missverständnissen betrieben wird. Dies ist auch in VUI-Dialogen beobachtbar: Die Tendenz zur Kürze könnte auch darin begründet liegen, dass der Austausch mit dem VUI als fehleranfällig und im Hinblick auf Reparaturen als umständlich beschrieben werden kann.

Wie schon Zoeppritz feststellt, lässt sich also keine Ausbildung spezifischer sprachlicher Praktiken zur Bedienung von Maschinen feststellen, vielmehr sind die Formen „mostly borrowed from styles already available for interpersonal communication" (Zoeppritz 1985: 20). Fischer (2006: 51) schreibt dazu: „If CT [Computer Talk, T.H.] is a register or sublanguage, it should display some homogeneity just on the basis of the fact that the speech is directed towards an automatic speech processing system". In ihrer eigenen empirischen Untersuchung auf Basis mehrerer Korpora kann sie dies nicht bestätigen, weder im Sinne eines strukturellen Registers – wie es von Krause (1992a: 157–158) bzw. Krause/Hitzenberger (1992) konstatiert wurde – noch im Hinblick auf eine funktionale Varietät, wie

360 Diese zeigten sich insbesondere in dyadischen Konstellationen (siehe dazu Kap. 6.1.3.3), die Mehrparteienkonstellationen weisen demgegenüber eine etwas höhere Variabilität auf, entstammen aber auch spezifischen Situationen, z. B. Ersteinrichtungen, Vorführungen und Tests, was die Aussagekraft für die routinierte Nutzung schmälert.

von Johnstone et al. (1994) vorgeschlagen.[361] Letztere argumentieren v. a. mit der funktional bedingten Reduktion von „Grounding“ (Clark/Schaefer 1989: 262) und Höflichkeitsarbeit (siehe Brown/Levinson 1987). Fischer (2006: 26–70) zeigt, dass diese Merkmale nicht als charakteristisch für ein sprachliches Register oder eine funktional ausgebildete Varietät gelten können: Zu groß sind die Unterschiede zwischen den Sprecher*innen sowie zwischen den verschiedenen Funktionen der untersuchten sprachlichen Formen, um für ein sprachstrukturelles Register zu argumentieren, zu wenig charakteristisch die untersuchten sprachlichen Praktiken, um für die Ausbildung einer funktionalen Varietät zu sprechen (vgl. Fischer 2006: 74). Dies zeigt sich auch in den vorliegenden Analysen: Das herausgearbeitete, weiter oben benannte Set an Formen zur TCU-Gestaltung für Stimmeingaben scheint nicht homogen zu sein – was sich schon auf Basis der hier gezeigten unterschiedlichen Fälle zeigen lässt. Die Suspendierung von „Face-Work“ (Goffman 1955) ist auch für andere empraktische Kontexte beobachtet worden. Zudem sind – jedenfalls an der sprachlichen Oberfläche – Formen von Face-Work beobachtbar. Gleichwohl ist Höflichkeit, wie auch von Fischer (2006: 70) angemerkt, eine vielschichtige Kategorie. Das in Mehrparteiensituationen beobachtete Auftreten von Invektiven und die Einbettungsformen in laufende zwischenmenschliche Interaktionssituationen deuten darauf hin, dass eine funktionale Differenzierung im Hinblick auf den sozialen Status der Geräte durchaus erfolgt.

Die verbindenden Eigenschaften von „Computer Talk“ sind, so argumentiert Fischer (2006) weiter, vielmehr *Strategien* im Umgang mit den Geräten, die auf die grundlegend anders ablaufende Sprachverarbeitung der VUIs (im Unterschied zum Menschen) und damit einhergehende Limitationen ebenso Rücksicht nehmen wie auf unterschiedliche Verarbeitungsweisen der situativen Gegebenheiten und unterschiedliches Weltwissen. Nutzer*innen produzieren dabei einen hohen Grad an Alignment im Hinblick auf Adressierungen, Lexik und Syntax (vgl. Fischer 2006: 107), den auch Lotze (2016: 287) am Beispiel von Chatbots feststellt.[362] Diese Sichtweise auf sprachliche Strategien als „Computer Talk“ kann anhand der Befunde zu sprachlichen Praktiken im Umgang mit VUIs bestätigt und teilweise

361 Dieser Unterschied beruht auf verschiedenen Ideen davon, welche konstitutiven Eigenschaften für ein Sprachregister gegeben sein müssen: „From the standpoint of language structure, registers differ in the type of repertoire involved, e.g., lexemes, prosody, sentence collocations, [...]; from standpoint of function, distinct registers are associated with social practices of every kind [...].“ (Agha 2004: 24).

362 Lotze (2016) arbeitet dabei mit einem sehr weit ausdifferenzierten Alignment-Begriff, der die Befunde von Fischer teilweise als „reaktives Alignment“ bestätigt; diese aufschlussreiche Diskussion soll hier nicht wiedergegeben werden.

ergänzt werden: Strategien der Nutzer*innen, die in spezifische sprachliche Praktiken münden, zeigen sich u. a.

- in der Basis-Sequenzstruktur, an der sich die Nutzer*innen im VUI-Dialog orientieren,
- an den unterschiedlichen Verfahren der sequenziellen Einbettung der VUI-Dialoge in zwischenmenschliche Interaktionssituationen,
- an den *barge-ins*, die Nutzer*innen einsetzen, um die Äußerungen des VUI zu unterbrechen und insofern keine Orientierung an Kategorien der Höflichkeit zeigen,
- an den unterschiedlichen Reparaturstrategien der Nutzer*innen, sowie auch
- am Einsatz von Invektiven zur interaktiven Bearbeitung von Fehlschlägen.

Diese Formen sind nicht distinkt im Sinne eines strukturellen oder funktionalen Registers. Sie sind aber als Set sprachlicher Strategien spezifisch und zeigen sich als emergente Zusammenhänge von Form und Funktion, die sich aus der situativen Praxis heraus zu verfestigen scheinen. Hinweise darauf geben die untersuchten Einlern-Dialoge, die etwa die Basis-Sequenzstruktur illustrieren und zu *barge-ins* einladen („Du kannst jederzeit ‚Stop' sagen") – sprachliche Praktiken der Nutzer*innen gehen also auch auf Phasen des Einlernens zurück.[363] Dies ist in den vorliegenden Daten ein wiederholt auftretendes Phänomen. Unterschiedliche Reaktionen der Sprachassistenzsysteme auf Invektive zeigen, dass dies seitens der Hersteller antizipiert und entsprechend im VUI-Design darauf reagiert wurde.

Letztlich muss im Hinblick auf diese Frage auch konstatiert werden, dass die vorliegende Arbeit mit ihrem qualitativ-explorativen Zuschnitt keine quantitativ gestützte Aussage über den Status der Registrierung spezifischer Formen treffen kann. Um die Routinisierung und die über Sprecher*innengruppen hinweg bestehende Etablierung als sprachliche Varietät beurteilen zu können, müsste ein deutlich größeres Korpus erhoben und durch Kodierungen ausgewertet werden, wie etwa in der Untersuchung von Barthel/Helmer/Reineke (2023) – darauf wird noch einmal eingegangen. Die genannten Strategien wären mögliche Ansatzpunkte, um hier weitere Untersuchungen anzustellen. So kann auch einer der sprachlichen Praktiken immer innewohnenden Übersituationalität als Voraussetzung für ihre Verfestigung und Routinisierung auf einer stabileren empirischen Grundlage nachgegangen werden.

363 Eine andere, nicht ausgewertete Quelle sind etwa E-Mail-Newsletter von Amazon an die Nutzer*innen, in denen neue Funktionen vorgestellt und länger existierende Funktionen in Erinnerung gerufen werden. Wie die Stimmeingaben dazu zu formulieren sind, wird in diesen Newslettern ebenfalls mitgeteilt.

7.2.3 Zur ‚Beteiligung' von VUIs an Praxis und Gespräch

Ein weiteres Ziel der Arbeit war, zu untersuchen, wie VUI-Dialoge und die soziale Praxis miteinander verwoben sind. Dazu wurde das Konzept der materiellen Partizipanden des Tuns nach Hirschauer (2004; 2016) nutzbar gemacht, der von einer verteilten Handlungsträgerschaft in der Praxis ausgeht, in der Aktivitätsniveaus diverser Partizipanden unterschiedlich aufgeteilt sein können. So konnten VUIs einerseits als Beteiligte an der sozialen Praxis betrachtet werden, zugleich wurde ihr Status in Gesprächen untersuchbar. Wie oben ausgeführt, kann dabei ein Auseinanderfallen von Form und Funktion konstatiert werden: Es zeigten sich sprachliche Praktiken zur Herstellung von formalen Gesprächsbeteiligungen eines VUIs auf der einen Seite, die aber auf der anderen Seite von brüchigem, temporärem Charakter waren und an vielen Stellen andere Funktionen als die tatsächliche Unterhaltung mit einem VUI erfüllten. Dabei ist der Austausch ähnlich zu Mensch-Tier-Interaktionen – sie sind auch oder überwiegend ein mehrfachadressiertes *display* anderer Praktiken (vgl. Tannen 2004; Torres Cajo/Bahlo 2016). Hierbei ergibt sich ein Kontinuum von mehr oder weniger stark ausgeprägten Mehrfachadressierungen: Wird zum Beispiel das Aktivierungswort („Alexa" o. a.) weggelassen, ist anzunehmen, dass die Nutzer*innen *primär* die ko-präsenten menschlichen Beteiligten adressieren und weniger das VUI selbst. Eine sprachliche Interaktion zwischen den adressierten Entitäten und den Sprecher*innen i. e. S ist dabei eine für alle Beteiligten offensichtliche Fiktion, die sich in die laufende soziale Praxis einfügt und innerhalb dieser andere Funktionen erfüllt. Die Beobachtungen zeigen, dass das Reden über den Smart Speaker die Integration solcher Formen begünstigt.

Mit der bereits mehrfach zitierten und diskutierten praxistheoretischen Konzeption von Hirschauer (2004; 2016) und der darin angelegten Vollzugsperspektive konnte aufgedeckt werden, dass VUIs als „materielle Partizipanden verteilten Handelns" (Hirschauer 2016: 51) gelten können. Ihre dialogischen Äußerungen werden als Ressourcen verwendet, um ko-operativ neue Koaktivitäten hervorzubringen. VUIs und ihre Äußerungen legen dabei die Ausführung bestimmter Praktiken nahe. Dies geschieht teilweise sehr explizit – z. B. wenn mit von den VUIs produzierten Adhortativen und Imperativen in den Ersteinrichtungsdialogen der Vollzug einzelner Praktiken ‚eingeübt' wird –, teilweise aber auch außerhalb des von den VUI-Designer*innen vorgesehenen Spektrums – etwa wenn die Nutzer*innen formal an das VUI gerichtete Dialogbeiträge produzieren, um die Äußerungen und Anschlussäußerungen als *display* von Frust zu verwenden oder wenn VUI-Äußerungen als Lachgegenstand funktionalisiert werden. Dies belegt noch einmal die These von Goodwin (2018: 444), dass der Vollzug von Praktiken von einem semiotischen Opportunismus gekennzeichnet ist – „with the ability to incorporate voraciously whatever

local materials might be used to construct the action required at just this moment". Sie werden damit also in Gespräche einbezogen und sind Teil des praktischen Vollzugs eines Gesprächs. Das macht sie – bzw. die Sprecher*innen machen sie – aber nicht zu Gesprächsteilnehmer*innen im engeren Sinn. Das Auseinanderfallen von formaler und funktionaler Ebene des Sprachlichen und die hohe Geschwindigkeit, mit der die zugeschriebenen Statuskategorien wechseln, die wiederum konzeptionell undurchlässig sind, bestätigen, dass die Kategorien aus Goffmans *Footing*-Konzept[364] im Hinblick auf ihren tatsächlichen Vollzug bei nicht-menschlichen Beteiligten an ihre Grenzen stoßen. Die Ergebnisse zeigen deutlich auf, dass es sich um nichtmenschliche Partizipanden handelt, die zwar mit Hirschauer an der Praxis partizipieren können, aber darum noch nicht kontinuierlich an einem Gespräch teilnehmen.

7.2.4 Domestizierung im *display* sprachlicher Praktiken

Die sprachlichen Praktiken mit VUIs sollten *im Prozess ihrer Domestizierung* untersucht werden. Dazu wurden zum einen Konzepte aus der Medienaneignungsforschung vorgestellt, diese wurden zum anderen mit dem Domestizierungsansatz, seinen Grundlagen und aktuellen Anwendungen und Kritik vermittelt. Als Teil der Aneignung bzw. Domestizierung sind also insbesondere Situationen der Ersteinrichtung, des Einlernens, Testens und Übens relevant, in denen sich zeigt, wie Nutzer*innen den Gebrauch anfangs ungewohnter Interfaces zur Gewohnheit werden lassen. Ertragreich zur Bearbeitung dieser Fragestellung waren insbesondere die Mehrparteienkonstellationen, in denen die VUIs im Zentrum der sozialen Praxis standen. In diesen Situationen war der Gebrauch von VUIs eingebettet in – den VUI-Dialogen vor- und nachgelagerte – zwischenmenschliche Interaktionen. In dyadischen Konstellationen waren solche expliziten Reflexionen nur im Ausnahmefall beobachtbar, sodass sie im Sinne eines konversationsanalytischen Vorgehens keine *displays* einer Bewertung enthielten – lediglich in zwei Aufzeichnungen aus demselben Haushalt äußert eine Nutzerin in dyadischen Konstellationen explizite Bewertungen. Abgesehen davon deuten in dyadischen Konstellationen lediglich nonverbale Elemente – z. B. vernehmliches Seufzen – auf Reflexionsprozesse hin. Während also die dyadischen Konstellationen eher Hinweise auf die Ausbildung sprachlicher Praktiken zur regelmäßigen Nutzung der VUIs geben, sind die Mehrparteienkonstellationen besser geeignet, um das Einlernen, Testen, Üben und die damit einhergehenden Reflexionen der Nutzer*innen

364 Siehe dazu Kap. 2.2.4.

einzufangen. Daher wird im Folgenden darauf der Fokus liegen, wenn es um die Domestizierungsprozesse geht.

In diesen Situationen tritt mit VUIs ein ungewohntes Interface in die häusliche Umgebung, das durch seine Sprech- und Dialogfähigkeit eine zuvor gesicherte Differenzierung zwischen Mensch und Maschine potenziell irritiert. Turkle (1995: 102) beschreibt, dass ein solcher Effekt verwischender Grenzen auch bei anderen Interfaces entstehen kann; für Smart Speaker, d. h. VUIs mit ihren ‚eigenen' Lautsprechern und insofern einer eigenen Materialität, ist dieser Effekt noch weitgehend unerforscht, Hinweise auf eine Verwischung ontologischer Grenzen konnten allerdings bereits gefunden werden (vgl. Guzman 2020; Weidmüller 2022). Dieser Effekt scheint in Mehrparteiensituationen noch stärker ausgeprägt zu sein (vgl. Etzrodt 2022). Dieser unsichere Status bedroht folglich auch die „ontological security" (Giddens 2003[1984]: 50) der Haushaltsmitglieder. Davon ausgehend könnte die bei Dizdar et al. (2021) ausbuchstabierte Perspektive der „Humandifferenzierung" helfen, um etwa das gezielte Heraus- und Überfordern der Geräte ebenso wie invektive Anschlusskommunikation zu verstehen. Das Konzept der Humandifferenzierung kann hier nicht umfassend eingeführt und beleuchtet werden, es soll aber kurz als möglicher, weiterführender Ansatz zur Domestizierungsforschung diskutiert werden. Anwendbar wäre dieser Gedanke z. B. auf analysierte Auszüge aus Ersteinrichtungen, in denen Nutzer*innen sprachlich deutlich betont nicht den Aufforderungen der VUIs folgen, sondern davon abweichen, sie umdeuten und die Geräte – teils als humorvoll gerahmte Ko-Aktivität – an ihre Grenzen bringen. Andere Situationen wären mehrfach funktionalisierte Invektive, wie sie v. a. in den Analysen scheiternder VUI-Dialoge herausgearbeitet wurden.

Dickel/Schmidt-Jüngst (2021: 343) argumentieren, dass Sprachassistenzsysteme in der Werbung vermarktet werden als „Entitäten, die (mehr oder weniger spezifische) soziale Positionen einnehmen sollen". Das bereits früh im Aneignungsprozess vollzogene Aufzeigen der Grenzen dieser Humanfiktion seitens der Nutzer*innen könnte also als gemeinschaftliche Vergewisserung der eigenen sozialen Position verstanden werden – und insofern als Praktiken der „ontologische[n] Außendifferenzierung" (Dizdar et al. 2021: 8), denn Humandifferenzierung vollzieht sich, so Dizdar et al. (2021: 11), u. a. sprachlich und praktisch. Dickel/Schmidt-Jüngst (2021: 342) sprechen von einer

> Konstitution der Außengrenzen des Humanen, also für die Arten und Weisen, in denen Nicht-Menschliches von Menschlichem unterschieden oder gerade nicht unterschieden wird. Denn durch solche Praktiken der Grenzziehung werden beide Seiten nicht nur zueinander ins Verhältnis gesetzt, sondern zugleich ihre jeweiligen Identitäten bestimmt. (Dickel/Schmidt-Jüngst 2021: 342)

Sprachliche Praktiken, mit denen die Grenzen der Interfaces aufgezeigt werden – gezieltes Überfordern, invektive Anschlussturns u. a. – können insofern als interaktive Arbeit an dieser Außengrenze verstanden werden. Die Nutzer*innen testen die Interfaces entlang der bereits weiter oben eingeführten Bruchlinie zwischen VUI-Dialog und zwischenmenschlicher Interaktion und zeigen sich selbst und übrigen Beteiligten durch entsprechende Praktiken auf, dass es sich bei VUI-Dialogen eben *nicht* um Interaktionen handelt, weil sich VUIs nicht als stabile Individuen erweisen. Die Nutzer*innen führen geradezu vor, dass die Parameter der Interaktion nach Goffman – geteilter Aufmerksamkeitsfokus, geteilte Wahrnehmung, wechselseitige Wahrnehmung dieser Wahrnehmung, koordinierte Ausrichtung von Handlungen am Gegenüber (vgl. Goffman 1983: 3) – nicht erfüllt sind. Insbesondere die fehlende Reflexivität der Geräte, d. h. eine wechselseitige Wahrnehmungswahrnehmung, fällt auf, wenn Äußerungen zwar an das VUI adressiert sind, diese aber zuvor gar nicht durch eine Invokationssequenz in einen entsprechenden Status versetzt wurden. Die fehlende Fähigkeit zur inkrementellen Handlungskoordination wird in Reparaturdialogen deutlich.

Zugespitzt könnte man sagen: Scheiternde VUI-Dialoge und deren interaktive Bearbeitung mindern das Irritationspotenzial von VUIs. Dabei tritt zur Außendifferenzierung auch eine Form der Binnendifferenzierung hinzu, die sich insbesondere in der auf VUI-Dialoge folgenden Anschlusskommunikation zeigt: Die Möglichkeit, Invektive an diesen maschinellen Kommunikationspartner zu richten, d. h. dem VUI kein zu wahrendes „Face“[365] zuschreiben zu müssen, bestellt eine soziale Hierarchie, derer sich hier vergewissert wird. Insoweit ko-präsente Mitglieder diese Äußerungen z. B. durch Lachen ratifizieren, schließen sie sich der durch den VUI-Dialog ausgedrückten sozialen Stellung an und bestätigen diese, auch für sich selbst. Sprachliche Praktiken diesen Typs werden somit verständlich als *display* von Domestizierung, als Bestandteil der Eingewöhnung, dem Zuschreiben einer Stellung des Geräts im Haushalt und somit der Vergewisserung der Grenzen des Maschinellen zur Sicherung der eigenen, humanen Überlegenheit. Sprachliche Praktiken, mit denen Nutzer*innen die sozialen Rahmen und Regulierungen in der Praxis abstecken, sind insofern Arbeit an der Erhaltung der „ontological security“ (Giddens 2003[1984]: 88).

Vor diesem Hintergrund hat es sich als produktiv erwiesen, in der Arbeit einen Begriff – den der *Interaktion* – zwischenmenschlichen Akteur*innen vorzubehalten. Dies war zunächst eine analytische Trennung; es ist aber – wie oben

365 Hier findet erneut der „Face“-Begriff von Goffman (1955) sowie in seiner sprachwissenschaftlichen Auseinandersetzung im Sinne einer Höflichkeitstheorie nach Brown/Levinson (1987) Anwendung.

ausgeführt – beobachtbar, dass die Teilnehmenden in ihren Äußerungen auch selbst Grenzen zwischen zwischenmenschlicher Interaktion und VUI-Dialog vollziehen, obschon einzelne Sequenzen zunächst das Gegenteil annehmen lassen könnten. Dies bestätigt grundsätzlich die Sichtweise von Suchman (1987; 1990; 2007; 2021), dass es sich um eine ressourcenlimitierte Form der Kommunikation handelt – VUI-Design wäre in diesem Sinne auch „less a matter of simulating human communication, than engeneering alternatives to talk's situated properties" (Suchman 2021: 77). Suchman (2007: 270) argumentiert, dass es sich um einen „dissymmetrischen" Austausch handelt: Während die Maschine im Hinblick auf die Wahrnehmung der Situation über die sensorisch erfassten Nutzer*innen-Eingaben hinaus limitiert ist, steht den Nutzer*innen mit dem Smart Speaker ein Artefakt als Austauschpartner gegenüber, das durch Design und darin integrierte Annahmen über gelingende Dialoggestaltung verfertigt wurde. Der extrem weit gefasste Begriff des VUI-*Dialogs* hat sich demgegenüber als tauglich erwiesen, um damit offen den an der Oberfläche konversationell vollzogenen Operationen begegnen zu können – die sprachlichen Praktiken der Nutzer*innen waren Teil der Dialoge und somit der Interfaces selbst. Ein solch weites Verständnis des Dialogbegriffs ist v. a. für Phänomene interessant, die konzeptionell noch nicht stabil gefasst werden konnten und deren konzeptioneller Status gerade Teil der Untersuchung ist. Ob er sich mittelfristig als tauglich erweist zur Beschreibung der sprachlichen Ein- und Ausgaben im VUI-Dialog, d. h. im „Interfacing" zwischen Mensch und Maschine, ist eine andere Frage.

Im Vergleich von dyadischen VUI-Dialogen und VUI-Dialogen in Mehrparteienkonstellationen ist auffällig, dass die linguistische Variabilität in Mehrparteienkonstellationen höher ist und die Tendenz zur Erzeugung formaler Kohärenz durch verschiedene sprachliche Verfahren nur in Mehrparteienkonstellationen besteht. Demgegenüber scheint in dyadischen VUI-Dialogen ein Fokus der Nutzer*innen auf ausgewählten Funktionen zu liegen, während der Großteil der Möglichkeiten mit VUIs nach der Testphase unausgeschöpft bleibt (vgl. Ammari et al. 2019; Pins/Alizadeh 2021). Situationen mit höherer sprachlicher Variabilität treten nach der Ersteinrichtung und den ersten Tagen danach erst wieder auf, wenn in Mehrparteienkonstellationen ein erneutes Testen oder Vorführen zum Zentrum der Praxis wird. Die sprachlichen Praktiken sind hier also als Evidenzen des Domestizierungsprozesses zu verstehen. Auch hierfür wäre eine quantitative Analyse zur Erhärtung dieses Befunds wünschenswert – Barthel/Helmer/Reineke (2023) zeigen hierfür vielversprechende Ansätze, die auf einem deutlich größeren Korpus und entsprechenden Kodierungen basieren und die beobachtete Tendenz bestätigen

können.[366] Sinnvoll wäre auch die Berücksichtigung haushaltsspezifischer Unterschiede in einem größeren und diverseren Korpus – so zeigt etwa Schneider (2021: 337) im Rahmen einer Interview-Studie, dass Nutzer*innen in Extremfällen „ein emotionales, positiv besetztes Verhältnis“ zu Smart Speakern aufbauen können, einschließlich ihrer Materialität.

Teil der Domestizierung und der Veralltäglichung von VUIs ist insofern, so kann zusammengefasst werden, einerseits das Einlernen, das Herausfordern und Testen von Grenzen – diese gehen mit einer hohen Variabilität in Bezug auf die sprachlichen Formen einher und zeigen die erwähnten Irritationspotenziale im Hinblick auf den ontologischen und partizipatorischen Status der Geräte. Andererseits deuten die Befunde darauf hin, dass sich die „Everydayification“ (Ayaß 2012: 3) von VUIs gerade dadurch auszeichnet, dass die sprachliche Variabilität in VUI-Dialogen nicht zunimmt und gesprächsähnlicher wird, sondern sinkt und die Technizität deutlicher hervortreten lässt – was allerdings durch methodologisch anders gelagerte Studien überprüft werden müsste (etwa Barthel/Helmer/Reineke 2023).

7.2.5 Smart Speaker im bewohnten Raum

Die linguistische Erforschung von Interaktion interessiert sich zunehmend für deren Multimodalität. Die immense Relevanz visueller, körperlicher und räumlicher Aspekte traten durch neue, technisch weiterentwickelte Analysemöglichkeiten deutlicher hervor als in früheren konversationsanalytischen Arbeiten.[367] Der Raum wurde dabei, so Schmitt/Hausendorf (2016: 13–16), zunächst mit dem Konzept des *Interaktionsraums* als „interaktive Ressource“ konzeptualisiert, in Fallstudien genauer untersucht (Hausendorf/Mondada/Schmitt 2012) und im Zusammenhang mit körperlichen Aspekten weiter exploriert (Schmitt 2013). In der Folge dieser genaueren Untersuchungen kam die „Ressourcenhaftigkeit des Raumes“ auch davon losgelöst in den Blick, die als den Interaktionen vorgelagert verstanden wurde (vgl. Schmitt/Hausendorf 2016: 16). Neben dem Interaktionsraum führen Hausendorf/Schmitt (2013; 2016a) daher die Konzepte der *Interaktionsarchite-*

366 Dabei sei – mit freundlicher Genehmigung der Vortragenden – auch auf den unveröffentlichten Vortrag von Mathias Barthel, Henrike Helmer und Silke Reineke bei der 17. Jahrestagung der International Pragmatics Association in Brüssel im Panel „(A)typical users of technology in social interaction“ (Chairs: Florence Oloff und Henrike Helmer) verwiesen. In diesem wurde die Frage der Routinisierung anhand eines größeren und kodierten Datensatzes in Kombination mit Einzelfallanalysen genauer beleuchtet, was die Produktivität von Mixed-Methods-Ansätzen (vorgeschlagen etwa von Kendrick 2017) bestätigt.

367 Im Detail siehe Kap. 4.4.

ktur und der *Sozialtopografie* ein. Damit werden – als Interaktionsarchitektur – basale „Implikationen des Raumes hinsichtlich seiner interaktiven Nutzung“ (Hausendorf/Schmitt 2013: 17) einer Analyse zugänglich, außerdem auch die kulturell aufgeladenen, aus gesellschaftlichen Konventionen und dem Wissen um diese entspringenden Interpretationen dieser Gegebenheiten – die Sozialtopografie. Hausendorf/Schmitt (2013: 3) unterscheiden außerdem zwischen dem „gebauten“, „gestalteten“ und „ausgestatteten“ Raum als heuristische Unterscheidung für Architektur.[368] Der Schwerpunkt solcher Untersuchungen liegt bislang auf Interaktionen in institutionellen Zusammenhängen.

In der Praxis von Interaktionssituationen zur Einrichtung und Nutzung der Smart Speaker sind körperlich-räumlich-architektonische Kategorien unter mindestens drei Gesichtspunkten relevant. Erstens kann in den Ersteinrichtungen lokal die körperliche Positionierung zueinander, zum Smart Speaker als im (praktischen, nicht unbedingt räumlichen) Mittelpunkt stehendes Gerät und zu anderen interaktionsarchitektonischen Elementen als wesentlicher Träger der Praxis betrachtet werden. Dieser Aspekt hebt auf den „Raum als interaktive Ressource“ (Hausendorf/Mondada/Schmitt 2012) ab: die Nutzung des Raums zur Herstellung der Einrichtungssituation, um die Teilnehmendenkonfiguration auszuhandeln und das Zusammenspiel des Geräts mit den Teilnehmenden und der Umgebung näher zu bestimmen. Einige Aspekte dessen konnten in den vorangehenden Kapiteln behandelt und fortlaufend in den Analysen einbezogen werden,[369] sie könnten aber noch weiter vertieft werden.

Zweitens lassen sich Einpassungen der Smart Speaker als Ausstattungsgegenstände in den gestalteten (Wohn-)Raum beobachten, damit einhergehende An- und Herausforderungen beschreiben und ein Spannungsverhältnis zwischen gestaltetem und ausgestattem Raum ausmachen (siehe auch Hausendorf/Schmitt 2016b). Diese Perspektive scheint vielversprechend, um einen Beitrag zur Frage der Domestizierung der Geräte leisten zu können und den Domestizierungsansatz zu erweitern (vgl. Miggelbrink 2018: 87–88): Smart Speaker werden, sofern sich deren regelmäßige Nutzung in den Alltagspraktiken der Nutzer*innen etabliert, von einem Novum im Haushalt zu einem Teil der Ausstattung und damit der

368 Als „gebaut“ werden dabei nur unter größtem Aufwand veränderbare Elemente verstanden (etwa Wände, Dach, Boden, Raumaufteilung innerhalb der Wohnung), als „gestaltet“ innenarchitektonsiche Einrichtungen (Möbel, Lampen, Teppiche usw.) und als „ausgestattet“ leicht verrück- und verschiebbare Gegenstände (Dekorationsobjekte, Geschirr, technische Gegenstände usw.).

369 Dies ergibt sich auch aus der gesetzten Methodologie der Arbeit (siehe Kap. 4.4). Für Analysen, in denen multimodale Verfahren auch unter Einbezug des Interaktionsraums eine Rolle spielen siehe insbesondere Kap. 6.2.2.1 zu Ersteinrichtungen des Smart Speakers und 6.2.3.8 zu multimodalen Verfahren.

Interaktionsarchitektur;[370] sie treten ein in eine wechselseitige Beziehung mit dem Mobiliar. Dieses ‚Infrastruktur-Werden', das reziproke Verhältnis zwischen bestehenden räumlichen Ausstattungen und Gestaltungen und dem neuen Gerät, das erst noch zur Ausstattung wird, könnte anhand von Videoaufnahmen von Smart Speaker-Anwendungen nachvollzogen werden: Welche Rolle spielen Sicht- und Hörbarkeit oder andere interaktionsarchitektonische Kriterien bei der Platzierung (vgl. Hausendorf/Schmitt 2013: 8)? (Wie) werden diese Kriterien expliziert und verhandelt? Anhand der vorliegenden Daten konnte dieses Themenfeld ausschließlich auf Grundlage der Videos von Situationen der Inbetriebnahme bearbeitet werden. Dabei konnte herausgestellt werden, dass Nutzer*innen u. a. den Blick als Ausdrucksressource verwenden, während das VUI Äußerungen produziert und somit den akustischen Kanal ‚besetzt'. Außerdem zeigte sich, dass Smart Speaker auch über den akustischen Kanal hinaus in ihrer Materialität interaktional relevant sind – sie senden über die Oberfläche Signale aus, die wiederum von den Beteiligten in der Interaktion relevant gesetzt werden können. Sie können das Herstellen einer gemeinsamen Sehfläche aller Beteiligten mit dem Smart Speaker im Zentrum erforderlich machen. Allerdings konnte dieser Themenbereich nicht für die weiteren reziproken Prozesse zwischen Smart Speaker und Wohnumgebung untersucht werden, weil für die dauerhafte Nutzung der Geräte keine Videodaten vorliegen. Hier könnten sich Folgeuntersuchungen auf wohnräumliche Arrangements konzentrieren.

Drittens werden in der Konfiguration der Smart Speaker Fragen der Geräte- in Verbindung mit der Raumnutzung *herstellerseitig* relevant gesetzt, v. a. in den Ersteinrichtungsdialogen. Dabei steht der Charakter der Smart Speaker als Gemeinschaftsmedium in einem Spannungsverhältnis mit der individual angelegten Nutzung von User-Accounts für Kommunikations- und Unterhaltungsdienstleistungen: E-Mails, Kalender oder Musikstreaming sind personalisiert und möglicherweise zu privat für die geteilte Wohnumgebung. Wie bereits an anderer Stelle in dieser Arbeit gezeigt, ist auch die Bedienbarkeit über die Stimme ein Spezifikum dieser Geräte, das die Anwendung (im Gegensatz zum Smartphone) sicht- bzw. hörbar für andere Anwesende macht. Dieses Spannungsverhältnis ist nicht neu, es wurde etwa für Computer bereits beschrieben und Nutzer*innenstrategien untersucht, in denen unterschiedliche Raumnutzungskonzepte offengelegt werden konnten (vgl. Röser/Peil 2014). Auch die Untersuchung dieses Themenkomplexes konnte im Rahmen der vorliegenden Arbeit nicht nur aufgrund der Datenlage, sondern auch mit Blick auf

370 Letzteres wird noch plastischer, wenn über den Smart Speaker auch Smart Home-Anwendungen wie Licht, Klingel und Wärmeregulierung gesteuert und die Smart Speaker etwa von Amazon aktiv als Mittel zur Infrastrukturierung von Smart Home-Anwendungen positioniert werden (vgl. Strüver 2023a).

die Methodik nicht geleistet werden: Eine solche Untersuchung müsste sich anderen ethnografischen Methoden zuwenden und ggf. auch Interviews einbeziehen, die konversationsanalytisch angereichert werden könnten.

7.2.6 Methodologische Reflexion und Anschlussmöglichkeiten

Um eine von Meiler/Siefkes (2023: 321) zurecht geforderte Reflexion der Methodenwahl zu leisten, werden die obigen Ausführungen noch einmal methodologisch kondensiert; darauf aufbauend werden Anschlussmöglichkeiten für zukünftige Studien aufgezeigt. Die vorliegende Arbeit hat sich einer praxeologisch fundierten Erkundung und dem Aufdecken von Phänomenen und dem Verstehen dieser in ihrem situationalen Zusammenhang verschrieben – konkret sollten sprachliche Praktiken in VUI-Dialoge sichtbar gemacht werden, um ihren Bezug zu gesprächsorganisatorischen Prinzipien zu verstehen und die Verflochtenheit von sprachlichen Praktiken in VUI-Dialogen und sozialer Praxis zu beleuchten. Dabei stand eine ethnomethodologisch-konversationsanalytisch informierte Methodologie ethnografischer Gesprächsanalyse im Zentrum, die durch Sequenzanalyse und videobasierte Interaktionsanalyse einzelner Ausschnitte sprachliche Praktiken auffindbar und in ihrem Zusammenhang beschreibbar machen sollte. Die Puristik der Konversationsanalyse[371] wurde dabei zugunsten einer praxeologischen Betrachtung aufgegeben – zum Erkenntnisinteresse passte vielmehr eine ganzheitliche Betrachtung des Datenmaterials und die Zuhilfenahme ethnografischen Wissens außerhalb der reinen Ausschnitte. Zwar wurde dies im Rahmen der Analyse dyadischer VUI-Dialoge etwas eingegrenzt, gleichwohl wurde der Untersuchungsgegenstand von den reinen Aufzeichnungen ausgehend auch um Wissen über die soziale Konfiguration der untersuchten Haushalte ergänzt. Dabei wurde das Primat der Untersuchung von Sprache in ihrem ‚natürlichen Umfeld', d. h. zu tatsächlichen Nutzungssituationen und möglichst wenig von Aufnahmesituation und -kontext beeinflussten Sprechgelegenheiten berücksichtigt. Für die Beforschung eines neuen und sprachwissenschaftlich wenig erforschten Gegen-

371 Eine entscheidende Rolle dabei spielt die Bedeutung von Ethnografie bei der Auswertung der Daten sowie die Praktiken der Datengewinnung (vgl. Bergmann 2001: 921; Bergmann 2007b: 58): Während die Ethnomethodologie ein ganzes Spektrum ethnografischer Methoden zulässt (z. B. auch teilnehmende Beobachtung) und dabei auch Hintergrundinformationen über die soziale Situation, den Kontext und über die Teilnehmer*innen der Interaktion einbezieht, blendet die Konversationsanalyse gerade diese bewusst aus und konzentriert sich nur auf die Phänomene, die in den Datenkollektionen (d. h. Transkriptionen von Audioaufnahmen) zu finden sind (vgl. Day/Wagner 2008: 44).

stands schien dies geboten, um nicht von vornherein Faktoren aus dem sozialen Kontext auszuschließen, die im Vorfeld nicht bekannt sein konnten (vgl. Kendrick 2017: 3).

Als Herausforderung erwies sich bei der Datenerhebung, was als ‚Widerspenstigkeit alltäglicher Praxis' bezeichnet werden könnte: die Schwierigkeit, den Alltag als solchen zu erkunden und einer Analyse zugänglich zu machen, mithin eines der Ausgangsprobleme des ethnomethodologischen Forschungsprogramms (vgl. Garfinkel 1967: 31–32). Auch im Rahmen von am Alltag interessierten Studien aus der Medienforschung wurde dieses Problem regelmäßig konstatiert und mit unterschiedlichen Vorgehensweisen umkreist (vgl. hierfür und für das Folgende Ayaß 2012: 8–12). Um das häusliche Alltagsleben der Nutzer*innen von Medientechnologien zugänglich zu machen, haben Studien zur Domestizierung von Medientechnologien primär mit ethnografischen Methoden gearbeitet (vgl. Hartmann 2013a: 53), deren typisches Repertoire jedoch für eine linguistische Auswertung nicht hinreichend gewesen wäre. Studien, die an einer feingranularen Beschreibung von Sprach- und Medienpraktiken interessiert sind, geraten immer wieder in ein methodologisches Dilemma: Medienpraktiken können etwa im Rahmen von qualitativen Interviews nur schwer elizitiert werden, weil sie den Nutzer*innen selbst nicht bewusst sind; außerdem können sie selbst mit beobachtenden Methoden (z. B. längeren teilnehmenden Beobachtungen) nur schwer entdeckt werden, weil sie von solcher Feingranularität und Flüchtigkeit sind, dass sie auch den Analytiker*innen ohne technische Hilfsmittel der Aufzeichnung verborgen bleiben (vgl. Ayaß 2012: 12–13). Die Aufzeichnung wiederum beeinflusst potenziell den Verlauf des Geschehens und insbesondere die sprachlichen Praktiken, allerdings bleibt die genaue Art des Einflusses teilweise verborgen. In der vorliegenden Untersuchung kann in der Instruktion zur gemeinsamen Durchführung der Inbetriebnahme der Smart Speaker eine Beeinflussung gesehen werden, von der die Daten in Form von deutlich produzierten situativen Rahmungen Spuren tragen. Andererseits sind die erzeugten Mehrparteienkonstellationen sehr aufschlussreich für die ‚Beteiligung' von VUIs an der sozialen Praxis gewesen. Das Ideal ‚natürlich entstehender Daten' ist nie vollständig zu erreichen, sondern als Kontinuum mit verschiedenen Parametern zu betrachten, die neben dem Aspekt der ‚spontan' entstehenden Teilnehmendenkonstellation auch Fragen der zeitlichen, lokalen und qualitativen Arrangiertheit betreffen (vgl. Kendrick 2017: 4). Die Frage, welche Gespräche dieses Ur-Kriterium gesprächsanalytischer Forschung hinreichend erfüllen, kann nicht abschließend beantwortet werden (für eine ausführliche Diskussion siehe Gerwinski/Linz 2018: 107–112). Hier bleibt für jede Studie erneut eine Abwägung zu treffen und entsprechende Entscheidungen sind durch eine gründliche Reflexion abzusichern.

Für eine längerfristige, situationsungebundene Beobachtung stellt sich zudem die Problematik, dass (auch der rein häusliche) Alltag nicht ‚vollständig' aufgezeichnet werden kann und ein minimal-invasives Vorgehen nicht nur zur Vermeidung von Spuren des Beobachterparadoxons, sondern auch aus forschungsethischen Gründen erforderlich macht. Eine videobasierte Erfassung des gesamten Alltags der Nutzer*innen über einen Zeitraum von mehreren Wochen hätte auch die Analytiker*innen vor große Herausforderungen bei der Bewältigung des Umfangs an gesammelten Daten gestellt und wäre unter forschungsethischen Aspekten mit der Erhebung von weitaus zu viel redundantem Material verbunden gewesen. So wurde die beschriebene, audiobasierte CVR-Technologie eingesetzt, um eine alltagsbasierte Nutzung von VUIs überhaupt zugänglich zu machen. Dabei ergibt sich bei der Auswertung die Problematik, dass der Alltag nicht immer in ‚Situationen' verläuft und die Sprecher*innen keineswegs immer verbalsprachlich anzeigen, was im Praxisvollzug gerade getan oder unterlassen wird und wer daran beteiligt ist. Die CVR-Aufzeichnungen stellen Bruchstücke aus dem Alltag der Nutzer*innen dar und von den Teilnehmenden produzierte Situationsrahmungen, wie sie bei den Ersteinrichtungen beobachtbar und analysierbar waren, treten zu großen Teilen nicht auf. Dieser Umstand lässt, gerade ohne auf Videomaterial zurückgreifen zu können, Interpretationen zur sozialen Konstellation durchaus unsicher werden. Die Analysen wurden insofern, wo sie sich auf die CVR-Daten beziehen, vorsichtig formuliert und präsentiert.

Für eine Untersuchung mit diesen Charakteristika wurde (typischerweise) eine verhältnismäßig kleine Datenbasis gewählt, um bei den Analysen tiefer ins Detail gehen zu können, wie es ein sequenzanalytisches Vorgehen erforderlich macht (vgl. Deppermann 2008: 55). Die Bildung von Kollektionen in einem iterativen Wechselspiel aus Gegenstandskonstitution, Sampling und Gegenstandsanalyse ist ein in der Gesprächsanalyse etabliertes Vorgehen (vgl. Deppermann 2008: 95), das auch hier angewendet wurde, um das erhobene Material auszuwerten. Basis dieser Methode ist ein sequenzanalytisches Vorgehen, das keine isolierten Äußerungen betrachtet, sondern Ausschnitte aus der Praxis.

Die Gliederung des ersten Teils entlang der „generic organisations of practice" (Schegloff 2006: 71) ist bereits ein Ergebnis sequenzanalytischer Beschreibungen in früheren Stadien der Untersuchung. Wie aus den obigen Ausführungen hervorgeht, ergeben sich aus diesen Analysen Anschlussfragen – insbesondere im Bereich der Routinisierung und Registrierung, dem Zusammenhang von sprachlichen Praktiken und Domestizierung sowie im Hinblick auf die Multimodalität. Anschlussfähig an die vorliegenden Befunde wären insofern ausgehend von der Methodologie Erweiterungen mit folgenden Linien:

1. Ein Ausbau der Untersuchung mit einer stärker **quantitativen Ausrichtung**; dazu wären sowohl ein größeres Korpus, eine systematische Kodierung wie auch eine in Teilen statistische Auswertung notwendig. Damit könnte der Verlauf der Domestizierung und ihr Niederschlag in sprachlichen Praktiken mit größerer empirischer Sicherheit eingefangen werden. Solche Ansätze sind durchaus mit konversationsanalytischen Grundprinzipien vereinbar (vgl. Pitsch et al. 2009b; Stivers 2015) und lassen Rückschlüsse auf „die Robustheit eines Phänomens“ (Pitsch 2023: 128) zu. Barthel/Helmer/Reineke (2023) zeigen bereits Vorteile eines solchen Vorgehens auf und berichten über erste Ergebnisse; die Autor*innen erproben dabei auch Mixed-Methods-Ansätze (s. o.), bei denen einzelne Sequenzanalysen mit korpusbasierten Auswertungen kombiniert werden. Solche Studien könnten auch zu weiteren Erkenntnissen im Hinblick auf die Registrierung kommen. Die vorliegenden Ergebnisse können für die Konzeptionierung und Durchführung einer solchen Studie wichtige Ansatzpunkte sein.
2. Eine stärkere Fokussierung auf die **Multimodalität** von VUI-Dialogen könnte die Rolle von Blick, Gestik, Mimik und körperlicher Orientierung zur Anzeige von ‚Beteiligung‘ stärker beleuchten und prüfen, ob diese außerhalb der hier erhobenen Ersteinrichtungsdialoge ebenfalls eingesetzt werden und wie sie funktionalisiert sind. Die hier vorgelegten Befunde lassen verschiedene Funktionen multimodaler Ausdrücke erkennen.
3. Um die Interaktionsbedingungen – z. B. bei der genaueren Betrachtung multimodaler Verfahren im Umgang mit VUIs – konstant zu halten, wären hier – wie etwa bei Etzrodt (2022) erprobt – im Sinne eines Methodenpluralismus in der empirischen Sprachwissenschaft (vgl. Kendrick 2017) auch **Laborstudien** möglich, was nicht zugleich eine quantitativ-hypothesengeleitete Auswertung bedingen muss; vielmehr kann gleichwohl ein qualitativ-entdeckendes Vorgehen angewendet werden (vgl. Pitsch 2023: 123). Auch hierfür liefern die Erträge aus der vorliegenden Arbeit Anknüpfungspunkte, die eine Orientierung für den Aufbau semiexperimenteller Settings geben könnten.
4. Sich auf die Multimodalität zu fokussieren könnte allerdings auch im Gegenteil bedeuten, sich einer (video-basierten) **Analyse des bewohnten Raums** zuzuwenden, in den VUIs eingebunden werden (siehe etwa Kesselheim 2016; Hausendorf/Schmitt 2016b). Diesem Aspekt wurde in der vorliegenden Arbeit nur in Ansätzen nachgegangen, dabei weisen Passagen aus den Ersteinrichtungen darauf hin, dass Mobiliar und Smart Speaker in eine Beziehung zueinander treten, die sprachlich mit-konstituiert und reflektiert wird (s. o.). Die Materialität von Smart Speakern kommt dabei noch stärker zum Vorschein – etwa wenn die Nutzer Konrad und Till den zunächst auf dem Fußboden ausgepackten HomePod gemeinsam an seinen späteren Standort tragen. Metho-

dologisch dürfte dann – anders als im zuvor ausgeführten Punkt – keine Abkehr von der Erfassung der Wohnumgebung tatsächlicher Nutzer*innen erfolgen, sondern müssten vielmehr Methoden angewendet werden, die die Verhältnisse zwischen Raum und Interaktion sichtbar machen können.

Über die generierten Erkenntnisse hinaus liegt also der Mehrwert einer Studie von qualitativ-explorativem Charakter, wie sie hier durchgeführt und präsentiert wurde, darin, dass sie aufzeigen kann, wo sich eine weitere Stratifizierung der Erhebung, aber auch tiefere Detaillierung in der Auswertung zukünftiger Untersuchungen anbieten könnte. Die Erkenntnisse stehen somit nicht nur für sich, sondern erledigen in besonderer Weise auch ‚Vorarbeiten' für Anschlussuntersuchungen.

7.3 Ausblick

Die bis hierher ausgeführten Analysen und Schlussfolgerungen beziehen sich teilweise auf Technologien, die bereits zum Zeitpunkt der Drucklegung dieser Arbeit von neueren Technologien überholt wurden: So hat etwa Amazon einen Nachfolger für den EchoDot (EchoPop) und Geräte mit Display auf den Markt gebracht (z. B. EchoShow). Andere auf dem Markt verfügbare Smart Home-Geräte haben sich in Qualität und Quantität rasant weiterentwickelt. Wie bereits in der Einleitung erwähnt, wurde außerdem zwischenzeitlich die Wirtschaftlichkeit von Smart Speakern für die Systemanbieter in Zweifel gezogen (vgl. Kim 2022), was wiederum unterschiedliche Prognosen für die Zukunft von Smart Speakern hervorbrachte. Die vorliegende Arbeit konnte nichtsdestoweniger erhellen, wie ungewohnte Interfaces zur Gewöhnung werden und welche zentrale Rolle die sprachlichen Praktiken bei der Einbindung von VUIs in die soziale Praxis spielen. Sie konnte zeigen, wie Nutzer*innen ihr gesprächsorganisatorisches Wissen in den Dienst von „Interfacing" (Lipp/Dickel 2022) stellen, sie konnte Konsequenzen spezifischer Eigenschaften von VUIs herausarbeiten und Situationen ihrer versprachlichten Domestizierung beschreiben. Dies alles sind wichtige Schritte, um die Veralltäglichung von Technologien verständlich zu machen – und ob diese dabei in absoluten Werten ‚neu' sind, ist für das medienlinguistische Erkenntnisinteresse nachrangig, wenn sie zum Untersuchungszeitpunkt ‚neu' in die untersuchten Haushalte einziehen. Darüber hinaus kann die Beleuchtung von VUI-Dialogen im Sinne Pitschs (2015; 2023) auch als Forschungsinstrument für die Gesprächsforschung bzw. die Interaktionale Linguistik eingesetzt werden – Erkenntnisse aus der Nutzung von VUIs lassen auch Rückschlüsse über die Verwendung von Sprache in ihrem interaktionalen Gebrauch zu.

Diesen Wert der Untersuchung muss man sich vor Augen halten, wenn der Blick nun auf zukünftige Entwicklungen gerichtet werden soll, die zu großen Teilen bereits Fahrt aufgenommen haben. So ist anzunehmen, dass sich durch die bereits angekündigte Einbindung von Large Language Models (LLMs) wie *ChatGPT* in VUIs (vgl. Malik 2023) die Sprachsynthese und Dialogizität erneut signifikant verändern werden. Die Verknüpfung von LLMs mit externen Wissensdatenbanken ist ebenfalls bereits in der Entwicklung (siehe etwa Thoppilan et al. 2022), obschon sich diese Entwicklungen aktuell auf schriftbasierte Systeme beziehen (vgl. Albrecht 2023: 46) und die Systeme nicht ohne Weiteres kompatibel zu sein scheinen (vgl. Strüver 2025). Werden diese technologischen Systeme allerdings miteinander kombiniert, ist mit einer neuen Qualität von VUIs zu rechnen. Diese Entwicklungen könnten auch zu einer erneuten Wende oder jedenfalls einer Veränderung in Bezug auf die Bewertung der Berichte führen, der zufolge Smart Speaker nicht wirtschaftlich sind – erste Diskussionen und Spekulationen diesbezüglich sind in entsprechenden Magazinen und Foren bereits im Gang (vgl. etwa Poti 2023). Technologisch deutlich avanciertere Stimmsyntheseverfahren, die bereits technisch machbar sind und etwa die eigene Stimme „klonen“ können (vgl. Grünewald 2023), werden auch die Erfahrungen der Nutzer*innen verändern und die ‚Maschinenhaftigkeit‘ der „Personae“ der bisher verfügbaren VUIs reduzieren. Die 2018 bei der Google I/O vorgestellte Anwendung Google Duplex hatte bereits Charakteristika menschlicher Sprache-in-Interaktion integriert, z. B. deutlich ungleichmäßigere Intonation und Sprechtempo und weitere prosodische Merkmale sowie die Produktion von Zögerungspartikeln und Rezeptionssignalen (vgl. Leviathan/Matias).[372] Zwar hat Google ein solches Dialog-System nicht in der Breite verfügbar gemacht, es ist allerdings anzunehmen, dass die neueren technologischen Entwicklungen im Bereich der LLM-basierten Sprachsynthese auch auf die Gestaltung von Stimmeingaben der VUIs einen Einfluss haben werden. Hier entstehen entsprechend erneut Aufgaben für eine gesprächsanalytisch arbeitende Medienlinguistik, die flexibel für die technologischen Entwicklungen bleiben muss. Eine Linguistik, die mit den Eigenschaften gesprochener Sprache und mit der Erfassung von ‚Dialogen‘ im Alltag ebenso umzugehen weiß wie mit medienwissenschaftlichen Konzepten und Debatten, ist hier gefragt, um die Einbindung KI-basierter Anwendungen in unterschiedlichste Bereiche des privaten und institutionellen Alltags erforschen zu können.

Eine weitere Entwicklung spricht dafür, dass die bisherigen Anbieter Smart Speaker nicht umgehend einstellen, aber möglicherweise modifizieren werden.

372 Die Demonstration ist auf YouTube unter https://www.youtube.com/watch?v=D5VN56jQMWM zu finden (zuletzt abgerufen am 22.05.2025).

Smart Speaker werden, so argumentiert Strüver (2023a; 2023b), für die Nutzer*innen zur wesentlichen Plattform, um Geräte in Smart Home-Environments einzurichten, zu konfigurieren und zu verwalten. Viel deutet demzufolge darauf hin, dass die Dienstanbieter von Sprachassistenzsystemen wie Amazon ein verändertes Ziel mit der Vermarktung von Smart Speakern verfolgen: die Etablierung dieser Geräte als zentrale Schnittstelle für Smart Home-Anwendungen sowie als Schnittpunkte für Nutzer*innen und Entwickler*innen. Dabei scheint für die Hersteller nicht entscheidend zu sein, dass das Interface stimmbasiert ist, denn bereits seit einigen Jahren sind Smart Speaker, die neben einem VUI auch ein Graphical User Interface (GUI) anbieten und beide Funktionen kombinieren, marktgängig und verbreiten sich weiter – zentral ist das Zusammenlaufen der verschiedenen Smart Home-Geräte in einer einzigen Applikation als Teil eines Ökosystems. Das „Interfacing" (Lipp/Dickel 2022) erfolgt also über verbale, visuelle und taktile Praktiken, die linguistisch in ihrem Zusammenspiel untersucht werden müssen. Dabei sind auch die multisensorischen Praktiken im Hinblick auf die gesteuerten Smart Home-Anwendungen einzubeziehen: Neben Praktiken des Sprechens, Hörens und Berührens können auch Praktiken des Fühlens oder Schmeckens relevant sein, z. B. bei smarten Thermostaten oder ‚Smart Cooking' (siehe Graf 2023) – dies macht eine multisensorische Analyse der Praxis im Sinne Mondadas (2021) und die iterative Anpassung des Methodenspektrums zum Einbezug technischen ‚Sensings' erforderlich (siehe auch Scholz 2022; Hector et al. 2025). So schlagen Albert/Hamann (2021: 5) bereits die Verwendung von Blick oder anderen multimodalen Alternativen zur Aktivierung von VUIs vor – anstatt der Verwendung eines Aktivierungsworts zur Invokation der Interfaces. VUIs finden zudem auch in Kontexten außerhalb der Wohnumgebung Verwendung, z. B. in Autos, aber auch in öffentlichen Institutionen oder in der Industrie. Insofern scheint es lohnenswert, Interface-Praktiken auch außerhalb der privaten Wohnumgebung zu betrachten. Insbesondere die Einbindung von VUI-Dialogen in andere soziale Settings könnte die Befunde aus der vorliegenden Arbeit noch erweitern.

Öffentliche Diskurse über die zunehmende Leistungsfähigkeit von KI-Anwendungen sind zu Recht als „Hype" bezeichnet worden (vgl. Bender/Koller 2020: 5185; Albrecht 2023: 17). Diese Diskursentwicklung wurde wiederholt von verschiedenen Seiten scharf kritisiert. So bezeichnet etwa der renommierte Medienwissenschaftler Alexander Galloway ‚Künstliche Intelligenz' auf seiner Webseite als „total scam" (Galloway 2021; siehe auch Groß/Jordan 2023: 11). Die Kritik zielt im Kern darauf ab, dass sowohl die irreführende Verwendung von anthropozentrischen Begriffen wie ‚Verstehen' und ‚Lernen' (vgl. Bender/Koller 2020) als auch die wiederholte Betonung der Opazität und Unergründlichkeit der technischen Verfahren die tatsächlichen Herausforderungen verschleierten, die durch KI-basierte Anwendungen gesellschaftlich entstehen (vgl. Groß/Jordan 2023: 11). Diskurslinguistische

Analysen können die Dynamik und Folgen solcher Diskurse aufdecken (vgl. Kalwa 2025); durch medienlinguistische Untersuchungen im Alltagskontext der Nutzer*innen lassen sich diskursiv erzeugte und möglicherweise überhöhte Vorstellungen von KI-basierten Smart Home-Anwendungen zudem immer wieder einer empirischen Prüfung unterziehen. Damit können sprachwissenschaftliche Untersuchungen auch einen Ausgangspunkt sowie eine empirische Grundlage für eine kritische Betrachtung sozialer Herausforderungen durch KI-basierte Anwendungen schaffen.

Als gegenwärtige und zukünftige gesellschaftliche Herausforderungen, die sich spezifisch durch VUIs als KI-basierte Anwendungen ergeben, lassen sich unterschiedliche Punkte benennen, zu deren Bearbeitung die Angewandte Sprachwissenschaft – zusätzlich zu den oben genannten Aspekten – Beiträge leisten kann und sollte. Dazu gehören weiterhin virulente Fragen um Datenschutz, Privatsphäre und Sicherheit. Wie bereits mehrfach im Verlauf der Arbeit adressiert, entsteht durch die Verwertung der aufgezeichneten Sprachdaten ein Spannungsfeld zwischen der hintergründigen, für die Nutzer*innen überwiegend opaken Verwertung von Daten auf der einen und der Subjektautonomie der Nutzer*innen auf der anderen Seite; in diesem Spannungsfeld neigen Nutzer*innen zu einem „Privacy Cynicism“, der auf der Unausweichlichkeit der Nutzung bei gleichzeitiger Limitierung der eigenen Handlungsoptionen beruht (vgl. Ranzini/Lutz/Hoffmann 2023). Dieses Spannungsfeld betrifft die alltägliche Anwendung digitaler Technologien allgemein (vgl. Hoffmann 2023), aber Smart Speaker im Besonderen, eben weil sie als dauerhaft ‚mithörende‘ Geräte in privaten Wohnumgebungen platziert werden und Aufgezeichnetes cloudbasiert aus- und weiterverwerten. Dadurch werden die Grenzen zwischen privaten und öffentlichen Räumen herausgefordert (vgl. Lutz/Newlands 2021). Diese Prozesse stellen die Handlungsmacht der Nutzer*innen nicht nur im medienpraktischen Umgang mit VUIs (vgl. Habscheid/Hector/Hrncal 2023) infrage, sondern auch im Gesamtzusammenhang und dem steigenden Einfluss der Betreiberfirmen. Die Sprachwissenschaft kann hier z. B. durch sequenzanalytische Betrachtungen der Nutzung (vgl. Waldecker/Hector/Hoffmann 2024) und der davon systemseitig angefertigten Protokolle (vgl. Habscheid et al. 2021) sowie durch diskursanalytische Untersuchungen (vgl. Lind/Dickel 2023) interdisziplinär integrierte Beiträge leisten, die Grundlage für eine weiterführende Kritik sein können.

Die Trainingsmethoden und die Datensätze, auf denen die KI-basierten Berechnungen zur Sprachsynthese erfolgen, sind bei den derzeit gängigen Systemen in vielerlei Hinsicht undurchschaubar – insbesondere das etablierteste, von OpenAI betriebene System *ChatGPT* ist im Hinblick auf Code-Offenheit, die Zugänglichkeit der Trainingsdaten für die LLMs und andere Aspekte intransparent, was wiederum eine Reihe von Risiken mit sich bringt (vgl. Liesenfeld/Lopez/Din-

gemanse 2023). Sie sind damit anfälliger für die Wiedergabe vorurteilsbehafteter Inhalte, u. a. für die Reproduktion von Gender-Stereotypen und Gender-Binarität (vgl. Dev et al. 2021; Savoldi et al. 2021), aber auch für religiöse, rassistische und klassistische Biases (vgl. Apprich et al. 2018), die sich aus den Daten heraus ergeben und keiner ethikbasierten Kontrolle unterzogen wurden. Verbinden sich diese Formen von Verzerrungen mit dem Bias, der durch die stimmbasierte Steuerung von VUIs entsteht und die Optionen für Nutzer*innen auf opake Weise restringiert (vgl. Natale/Cooke 2021), könnten diese Effekte sich wechselseitig verstärken. Darüber hinaus liegt der Fokus in der Entwicklung von VUIs derzeit eindeutig auf den Standardvarietäten großer Sprachgemeinschaften. Sprachen kleinerer Sprachgemeinschaften und auch größere Varietäten (z. B. Schweizerdeutsch) werden durch die VUIs entweder nicht hinreichend erkannt oder synthetisiert oder beides. Damit begünstigen VUIs marginalisierende Effekte auf Sprecher*innen von Dialekten, Minderheitensprachen, L2-Sprecher*innen (vgl. Markl 2022) oder Nutzer*innen in mehrsprachigen Haushalten (vgl. Leblebici 2025) sowie mit sprachbezogenen Einschränkungen (vgl. Albert/Hamann/Stokoe 2023).

Die Intransparenz von KI führt nicht nur zu möglichen Verzerrungen im Hinblick auf die Inhalte, sondern begünstigt auch die Verschleierung ausbeuterischer Arbeitspraktiken (vgl. Liesenfeld/Lopez/Dingemanse 2023: 5). Während einerseits über den Einfluss von KI-basierten Systemen auf die Arbeitswelt im Sinne eines veränderten Arbeitens *mit* KI diskutiert wird (z. B. Kirsch-Kreinsen/Karacic 2019), stabilisiert sich andererseits gerade für VUIs ein prekärer Arbeitssektor *für* KI. Diese Arbeit, die häufig die Datenschutzproblematik noch weiter verstärkt (vgl. Fuest 2019), ist zu nicht unerheblichen Teilen gesellschaftlich ‚unsichtbar' – sie wird von nicht-privilegierten Bevölkerungsgruppen ausgeübt, die als ‚Clickworkers' einen entscheidenden Anteil am Funktionieren der entsprechenden Dienste haben (vgl. Crawford/Joler 2018: XVIII; siehe auch Kaerlein 2020: 53), weil ihre Arbeit für die Generierung und das Training der Modelle sowie den reibungslosen Ablauf der KI-basierten Anwendungen notwendig ist. Zugleich findet sie unter prekären Bedingungen statt und ist psychologisch belastend (vgl. Perrigo 2023). Ferner werden auch die Nutzer*innen zu Mitwirkenden an „Datenarbeit" im „Capture-Kapitalismus" (Heilmann 2015). Die erfassten Daten werden abgeschöpft und zu kommerziellen Zwecken der Systemanbieter verwertet (siehe auch Turow 2021), was allerdings gegenüber den zuerst genannten Herausforderungen vielmehr als konsequente Folge von Technologie im Dienst kapitalistischer Ordnungssysteme erscheint, wie auch Heilmann (2015: 47) herausstellt. Insgesamt manifestieren, so lässt sich konstatieren, die mit diesen Prozessen einhergehenden Verlagerungen von Arbeitslasten globale Machtverhältnisse, die den Einfluss großer Tech-Konzerne stabilisieren. Dies berührt auch Nachhaltigkeitsaspekte und Ressourcenverbrauch (vgl. Crawford/Joler 2018). Die Angewandte Sprachwissenschaft ist in diesen Bereichen, die hier nur angerissen

werden sollten, als empirische Grundlagendisziplin gefragt, etwa um Paradigmen und praktische Verfertigung des VUI-Designs sichtbar zu machen (vgl. Khemani/Reeves 2022) und Kritik untermauern zu können.

Es ist offensichtlich, dass diesen Herausforderungen nur interdisziplinär begegnet werden kann – sowohl empirisch als auch theoretisch sind fachübergreifende Kooperationen notwendig, auch um disziplinäre Erkenntnisinteressen befriedigen zu können (siehe auch Hepp/Loosen 2023). Die Linguistik sollte dabei den Anspruch haben, gesellschaftlich relevante Ergebnisse zur Einbindung von KI-Anwendungen im Alltag zu liefern. Die vergleichsweise hohe methodologische Festigung der Linguistik in unterschiedlichen Bereichen (etwa der Gesprächs-, Diskurs- und Medienlinguistik) und das fundierte Empirieverständnis sind Ressourcen des Fachs, die unbedingt auch in den Dienst kultur- und medienwissenschaftlicher, mediensoziologischer und anderer Erkenntnisinteressen gestellt werden sollten. So kann die Linguistik neben einer fundierenden auch eine integrierende Funktion erfüllen, was angesichts der gleichzeitig stattfindenden Ausdifferenzierung von Subdisziplinen in den Geistes- und Sozialwissenschaften dringend geboten zu sein scheint.