Sprachassistenzsysteme und ihre Interfaces

Eine medienlinguistische Analyse

Tim Hector

Abstract *Der vorliegende Beitrag untersucht aus einer medienlinguistischen Perspektive Interfaces von Sprachassistenzsystemen – und zwar sowohl die stimmbasierten Ein- und Ausgaben (Voice User Interfaces, VUIs) als auch darauf aufbauend deren Dokumentation in verknüpften Smartphone-Apps (Graphical User Interfaces, GUIs). Aus der kombinierten Interface-Analyse ergeben sich erstens Erkenntnisse über das Verhältnis zwischen den stimmbasierten Interfaces und dem Sprachgebrauch der Anwender*innen. Konkret wird ein sprachlich-sequenzielles Muster für Dialoge mit VUIs beschrieben und aufgezeigt, dass die Anwender*innen diese situativ in die soziale Praxis einbinden. Zweitens ergeben sich Einsichten über hintergründige Mechanismen zur Verdatung und Verarbeitung sprachlicher Praktiken. Die Dokumentation der Dialoge in visuellen Oberflächen erweist sich als fragmentarisch und zeigt auf, dass die Aufzeichnungen für eine maschinelle Weiterverarbeitung vorbereitet werden, die nur teilweise am sequenziellen Ablauf der Stimmein- und -ausgaben orientiert ist.*

Keywords *Voice User Interfaces; Interface Studies; Intelligente Persönliche Assistenten; Sprachassistenten; Mensch-Maschine-Interaktion; Machine Learning; Medienlinguistik*

1. Einleitung[1]

Menschen und Maschinen sprechen miteinander: Die Verbreitung von Sprachassistenzsystemen wie »Alexa« von Amazon oder Internet-Anwendungen wie ChatGPT ist dabei nur die jüngste Spitze einer längeren Entwicklung: Sprachdialogsysteme, Chatbots und Sprachassistenten weisen eine teils weit zurückreichende Mediengeschichte auf (vgl. etwa Baranovska/Höltgen 2018; Volmar 2019) und wurden historisch durch unterschiedliche technologische Verfahren ermöglicht (vgl. Bender/Koller 2020: 5192). Diese Anwendungen haben gemeinsam, dass sich ihre User Interfaces (Benutzer*innen-Schnittstellen) an Praktiken zur Organisation

1 Für hilfreiche Anregungen und Kommentare zu früheren Fassungen dieses Texts danke ich Benedikt Merkle, Niklas Strüver und Didem Leblebici sowie den anonymen Reviewer*innen.

zwischenmenschlicher Interaktionen orientieren. Die Maschinen eröffnen damit einen Raum für Interface-Dialoge, d.h. sprachbasierte Ein- und Ausgaben zur Steuerung der Maschine, wobei diese seitens der Maschine synthetisch generiert sind. Diese können in den genannten Anwendungen schrift- oder stimmbasiert erfolgen. Zugleich stehen diese Interface-Dialoge nicht für sich; vielmehr sind sie angebunden an unterschiedliche visuelle Bildschirm-Umgebungen (Sehflächen), die den Dialog selbst mit hervorbringen können, ihn evtl. beeinflussen, oder ihn dokumentieren und archivieren. Bei Chatbots können dies z.B. das Chatfenster, die visuelle Umgebung einer Webseite oder die Dokumentation des bisherigen Chatverlaufs sein. Bei Sprachassistenzsystemen, die im Zentrum des folgenden Beitrags stehen, stellt die verknüpfte Smartphone-Anwendung eine solche Umgebung dar, in der die Aktionen dokumentiert und das System gesteuert werden kann.

Der vorliegende Beitrag widmet sich den Interfaces stationärer Sprachassistenzsysteme. Die Systeme bestehen erstens aus einem Smart Speaker, einem Geräteverbund aus Lautsprecher, Mikrofon und Rechenmodul – u.a. zur Herstellung einer Internetverbindung und zur Erkennung des Aktivierungsworts, z.B. »Alexa«. Wird dies erkannt, wird der nachfolgende Input aufgezeichnet und zur Auswertung an einen Cloud-Server gesendet (vgl. Hoy 2018: 82). Die Auswertung erfolgt bei den untersuchten Systemen – anders als bei Anwendungen, die auf *Large Language Models (LLMs)* basieren – nicht rein probabilistisch, sondern funktionsbasiert (vgl. Hoy 2018: 83f.). Durch *Natural Language Understanding* (NLU) werden dabei in den aufgezeichneten und durch *Speech Recognition* transkribierten Äußerungen der Nutzer*innen zunächst einzelne Anwendungen (*Skills*) erkannt und der Input daraufhin nach anwendungsspezifischen Eingaben abgesucht (vgl. Bedford-Strohm 2017: 487f.). Beispielsweise wird, nachdem die Timer-Funktion erkannt wurde, die Spezifikation der Dauer des Timers erwartbar. Die anschließend produzierte Stimmausgabe (*Natural Language Generation*) ist entsprechend ebenfalls entlang bestimmter Parameter, der Funktionen und ihrer Spezifikationen, strukturiert. Mit den Geräten ist außerdem eine Smartphone-App verknüpft, die zur Einrichtung und Bedienung des Smart Speakers notwendig ist und weitere Funktionen ermöglicht. Der Beitrag untersucht Dialoge mit den Systemen erstens auf der Ebene der stimmbasierten Ein- und Ausgabe (*Voice User Interface, VUI*) und davon ausgehend zweitens auf der Ebene der Dokumentation dieser Dialoge in der App-Umgebung (*Graphical User Interface, GUI*).

Wie die knappe Beschreibung von Sprachassistenzsystemen schon andeutet, sind die verschiedenen maschinellen Prozesse, die Rechenoperationen und die daran beteiligten Entitäten, die für das Funktionieren von Alltagstechnologien notwendig sind, enorm komplex und nicht nur für die Anwender*innen zunehmend unverständlich und unzugänglich (vgl. Hadler/Haupt 2016a: 7). Diese Unzugänglichkeit entsteht, so Hadler/Haupt weiter, zum einen materiell durch das Verstecken von Kabeln und Drähten in materiellen Gehäusen, die selbst wiederum Schauplatz

von Zeichenprozessen sein können (vgl. Bartz et al. 2019).[2] Zum anderen entstehe durch die Vernetzung der Geräte untereinander und »Ubiquitous Computing« zunehmend eine Auflösung von Technologie in der Umgebung: »The technology is not only boxed in, but also dissolves into the environment« (Hadler/Haupt 2016a: 7). Es lässt sich ergänzen, dass auch softwareseitig Rechenoperationen und technische Prozesse durch Maskierungen in ›übersichtlichen‹ Eingabemasken für die Endnutzer*innen nicht (mehr) sichtbar sind. Sprachassistenzsysteme stehen dabei unter zwei Aspekten im Fokus – als ›Zentrale‹ für Smart-Home-Anwendungen, die die Steuerung vernetzter Geräte orchestriert (vgl. Strüver 2023a) und als stimmbasierte Technologie mit hintergründigen Verdatungs- und Auswertungsmechanismen.

Der Beitrag widmet sich insbesondere deshalb den User Interfaces: Dies verspricht, die praktische Funktionsweise zunehmend opaker Alltagstechnologien im Verhältnis zu den situierten alltäglichen Praktiken ihrer Anwender*innen einerseits und im Verhältnis zu digitalen Infrastrukturen andererseits aufschlüsseln zu können (vgl. Hadler/Haupt 2016a). Über die Interfaces besteht ein Zugriff auf das Verhältnis von Mensch und Maschine bzw. auf die Verhältnisse verschiedener maschineller Einheiten untereinander – und somit auf die Relationen von (alltäglichen) Praktiken und Infrastrukturen (vgl. Kaerlein 2020). Daher unternehme ich den Versuch einer Interface-Analyse für Sprachassistenten und folge insofern dem Appell Kaerleins (2020: 54): »Follow the interfaces«. In Anlehnung an die Losung der Akteur-Netzwerk-Theorie – »Follow the Actors« (Latour 2005: 12) – formuliert Kaerlein (2020: 54) mit diesem Credo den Vorschlag, »die Art und Weise von Vermittlungsprozessen zwischen den beteiligten menschlichen und nicht-menschlichen Entitäten« zu fokussieren. Dazu untersucht der Beitrag die Interfaces der Sprachassistenzsysteme detailliert auf einer empirischen Grundlage.

Interfaces sollen im Folgenden mit Hookway (2014: 59) als *Grenzfläche* in der Begegnung zweier verschiedenartiger Entitäten konzeptualisiert werden (siehe Abschnitt 2). Für VUIs sind dies ein sprechendes, menschliches Subjekt und eine cloudbasierte Recheneinheit, die als synthetische Stimme repräsentiert wird.[3] Die Grenzfläche im Fall der VUIs in Sprachassistenten ist als gesprochensprachlicher Dialog gestaltet – daher ist im Folgenden von *VUI-Dialogen* die Rede. Teilweise werden diese durch Lichtsignale auf der materiellen Oberfläche der Smart Speaker ergänzt. Interfaces konstituieren sich also in der Praxis erst im VUI-Dialog. Im Fall

2 Im Fall der Smart Speaker wird die Oberfläche etwa zur Platzierung von zusätzlichen Steuerungstasten und als Träger von Lichtzeichen zur Anzeige des Gerätestatus' funktionalisiert (siehe Abschnitt 2).

3 Der Eindruck einer ›Einheit‹ wird nur auf der bedienbaren Oberfläche suggeriert. Die dahinterliegenden Prozesse erfordern hingegen die Beteiligung einer unüberschaubaren Anzahl weiterer Einheiten (vgl. Crawford/Joler 2018); siehe dazu Abschnitt 2.

der GUIs für Sprachassistenzsysteme – d.h. in der App-Umgebung – ist die Grenzfläche anders gestaltet, nämlich als Bildschirm-Sehfläche mit der Möglichkeit zur touchbasierten Nutzer*innen-Eingabe und der visuellen und teilweise auditiven Ausgabe. Ein konversationell gestalteter Austausch zwischen Nutzer*innen und Maschine ist hier nicht vorgesehen. Die kombinierte Analyse beider User Interfaces für Sprachassistenzsysteme soll die Praktiken zur Ausgestaltung der Grenzflächen sichtbar machen.

Als Analysematerial stehen dabei erstens Video- und Audio-Aufnahmen vom praktischen Umgang mit Smart Speakern zur Verfügung (siehe Abschnitt 3). Zweitens betrachte ich die Smartphone-Anwendung als Teil des Sprachassistenzsystems, mit einem Fokus auf darin hinterlegte Einträge von zuvor über den Smart Speaker durchgeführten Aktionen. Damit fokussieren die Analysen (Abschnitt 4) auf Prozesse, die typischerweise dem *front end* zugerechnet werden: (a) die Stimmein- und -ausgabe sowie entsprechende Lichtsignale, die Smart Speaker senden, als primär stimmbasiertes Interface, (b) die Smartphone-App als GUI mit eingebetteten Audio-Bestandteilen, und insbesondere als Rahmung für die Entstehung eines Nutzungsprotokolls sowie als Berührungspunkt von VUI und GUI.

2. Interfaces in Sprachassistenzsystemen

Interfaces erfahren in den letzten Jahren eine neue Aufmerksamkeit u.a. in medienwissenschaftlich ausgerichteten Software Studies, um digitale und vernetzte Technologien genauer zu untersuchen, ihren Charakter zu verstehen und einer Kritik zugänglich zu machen (vgl. u.a. Hadler/Haupt 2016a; Distelmeyer 2020; Kaerlein 2020; Ernst/Bächle 2020; Dieter 2022). Die Spezifik *stimmbasierter* Interfaces, wie sie in Sprachassistenzsystemen zum Einsatz kommen, liegt darin, dass ein sequenziell organisierter und zudem auf dem akustischen Kanal prozessierter Austausch zwischen Mensch und Maschine vollzogen wird – dieser Dialog gestaltet das aus, was bereits zuvor als »Grenzfläche« beschrieben wurde, d.h. die Ein- und Ausgaben von zwei verschiedenartigen Einheiten. Dieses Verständnis schließt an die etymologischen Ursprünge des Interface-Begriffs in den Naturwissenschaften an: Der Chemiker James Thomson bezeichnet damit die Grenzfläche, die beim Kontakt zweier unterschiedlicher Flüssigkeiten entsteht (vgl. Hookway 2014: 59f.).

Diese Konzeptionalisierung stellt auf die Dynamik in der Begegnung zweier unterschiedlich beschaffener ›Einheiten‹, Systeme oder Ordnungen ab. Damit konturieren sich Interfaces vielmehr erst dann, wenn die menschliche und die maschinelle Einheit oder »Ordnung« in der Praxis aufeinandertreffen (vgl. Hookway 2014:

45; Wirth 2016: 29).[4] Für VUIs ist dieses Aufeinandertreffen in Form menschlicher und maschinell-synthetischer gesprochensprachlicher Dialoge gestaltet. Ein VUI-Dialog konstituiert insofern genau diese Grenzfläche zwischen zwei ›Ordnungen‹ – der menschlichen und der maschinellen – und gestaltet sie aus. Die ›Einheit‹ des Maschinellen wird dabei v.a. durch die synthetische Stimme suggeriert, während tatsächlich eine Vielzahl von Komponenten am Zustandekommen einer gelungenen Operation mit einem Sprachassistenzsystem beteiligt sind (vgl. Crawford/Joler 2018). Die synthetischen Stimmen *repräsentieren* allerdings die dahinterliegenden, technischen Prozesse, wie Natale/Cooke (2021: 1009) konstatieren: »From a technical viewpoint, there isn't anything like one monolithic ›Alexa‹ or ›Siri‹. Rather than being individual entities, they are the integration of a wide range of different systems and algorithms«. Natale und Cooke ziehen Parallelen zu Metaphern, die in GUIs zum Einsatz kommen (z.B. ›Mülleimer‹ oder ›Ordner‹), die dazu vorgesehen sind, das Interface für die Nutzer*innen zu strukturieren (vgl. ebd.). Mit ihrem Namen, einer menschenähnlichen Stimme und der Simulation einer Persönlichkeit tragen die synthetischen Stimmen mit semiotischen Mitteln zur Steuerung der Dialoge zwischen Mensch und Maschine bei (vgl. ebd.: 1009ff.).[5]

Borbach (2019: 18f.) stellt mit Bezug auf akustische Interfaces fest, dass sich die medienwissenschaftliche Interface-Forschung überwiegend auf das Visuelle konzentriert hat, stellenweise unter Einbezug von Haptik. Interfaces hingegen, die sich über einen akustischen Kanal konstituieren, standen nicht im Fokus (siehe aber Borbach et al. i.E.). Dies ist insbesondere deswegen ein Desiderat, weil Interfaces als »Natural User Interfaces« geplant und konzipiert werden, die möglichst wenig auffallen, sich in das Gefüge von Praktiken einweben sollen und dabei den gesamten Körper durch zunehmende Multimodalität – u.a. Gesten oder eben gesprochene Sprache – ansprechen (vgl. Ernst/Bächle 2020: 416). Durch die Verwendung von gesprochener Sprache in Sprachassistenzsystemen ist diese Körperlichkeit besonders deutlich: Laute, die mit dem Stimmapparat des Menschen erzeugt werden, werden transkribiert und einer technischen Auswertung und Verdatung zugänglich gemacht.

VUI-Dialoge unterliegen dabei bestimmten Restriktionen, die sie von zwischenmenschlichen Interaktionen unterscheiden. Der Sprachgebrauch in VUI-Dialogen ist insofern für die maschinelle Verarbeitung vorbereitet und muss bestimmte Bedingungen erfüllen, damit die Begegnung zwischen Mensch und Maschine erfolgreich verlaufen kann: Analysen von VUI-Dialogen zeigen, dass

4 Zwar ist der Beitrag auf User Interfaces konzentriert, es muss jedoch betont werden, dass Schnittstellen und Grenzflächen auch jenseits davon entstehen – überall da, wo verschiedenartige Systeme aufeinandertreffen und ein Austausch vermittelt wird (vgl. Bratton 2016: 360); siehe auch die Typologie bei Cramer/Fuller (2008: 149).

5 Siehe dazu auch das erste Beispiel in Abschnitt 4.

Nutzer*innen sich nicht nur an gesprächsorganisatorischen Praktiken (wie z.B. sequenzieller Organisation oder Turn-Taking) orientieren, sondern auch an Annahmen über die Möglichkeiten und Grenzen der Technologien (vgl. Habscheid 2022; siehe auch Ernst 2017: 100). Konkret wirken dabei die Verarbeitungsfähigkeiten der Geräte stabilisierend auf die Ausbildung sprachlicher Praktiken als Interface-Praktiken: Es zeigt sich in Längsschnitt-Betrachtungen, dass Nutzer*innen im Verlaufe der ersten Wochen nach der Ersteinrichtung des Geräts in Abhängigkeit von der sprachlichen Gestaltung der Stimmeingabe erfahrungsbasiert höhere Erfolgsquoten bei der Nutzung von VUIs erzielen. Dabei ersetzen die Nutzer*innen z.B. bestimmte Satzarten durch andere und passen ihren Sprachgebrauch auf diese Weise an die Geräte an (vgl. Barthel/Helmer/Reineke 2023: 7). Hier ist die medienlinguistische Untersuchung der Interfaces wertvoll, um die Anatomie dieser Praktiken beschreiben zu können.

Auch die Sehfläche in der verknüpften Smartphone-App, einem GUI, stellt eine Grenzfläche her und repräsentiert vollzogene VUI-Dialoge in einer primär visuellen Darstellungsweise. Die Interface-Analyse betrachtet – ausgehend von VUI-Dialogen – beide Interfaces kombiniert und unternimmt den Versuch,

> »Verknüpfungen zwischen Praktiken und Infrastrukturen herzustellen, also nach den Umschlagspunkten und Vermittlungsschritten Ausschau zu halten, an denen aus Praxis Daten generiert werden bzw. Daten sich wiederum in Ketten von Praktiken übersetzen. […] Jede Übersetzung *zwischen* und *innerhalb von* Praktiken und Infrastrukturen lässt sich daraufhin befragen, nach welchen Regeln sie vollzogen wird, wie sie für die beteiligten Entitäten repräsentiert wird, und welche Interfaces in welcher Weise noch an diesem Prozess beteiligt sind.« (Kaerlein 2020: 54)

Mit der Konzentration auf Interfaces als Gegenstand der Analyse lassen sich also situierte Praktiken auf der einen mit Prozessen der infrastrukturellen Dimension auf der anderen Seite verbinden (vgl. Distelmeyer 2020: 62; Kaerlein 2020: 50). Mit dem Fokus auf Interfaces in der linguistischen Analyse kommt in den Blick, wie der Vermittlungsprozess zwischen den beteiligten ›Ordnungen‹ sprachlich gestaltet ist, wie die sprachlichen Äußerungen dokumentiert werden und somit, wie die Stellen beschaffen sind, an denen situierte, sprachliche Praktiken zu Interface-Praktiken werden und in der Folge verdatet und infrastrukturiert werden.

3. Datengrundlage und Methode

Die nachfolgend analysierten Daten wurden im Rahmen des Projekts »Un/erbetene Beobachtung in Interaktion: Intelligente Persönliche Assistenten (IPA)« am Son-

derforschungsbereich Medien der Kooperation an der Universität Siegen[6] erhoben. Dazu wurde in insgesamt acht Haushalten – darunter Wohngemeinschaften, Paare und Single-Haushalte – unterschiedliches Material erfasst. Hierzu zählen drei für die folgende Auswertung relevante Datentypen: (1) Videoaufzeichnungen von der Inbetriebnahme und Ersteinrichtung eines Smart Speakers[7] in der Wohnumgebung, (2) Aufnahmen von der routinierten Nutzung in zwei verschiedenen Phasen von zwei bis vier Wochen und (3) Ausschnitte der in der Smartphone-App einsehbaren Protokolldaten.

Die *Videoaufzeichnungen* fertigten die Studienteilnehmer*innen selbst mit von den Forscher*innen bereitgestellten Kameras an. Die Forscher*innen waren während der Aufzeichnungen selbst nicht anwesend. Die *Audioaufzeichnungen* konnten mit Hilfe eines sogenannten »Conditional Voice Recorders« (CVR) erhoben werden.[8] Dieser zeichnet die Wohnumgebung fortlaufend akustisch auf, löscht aber auch das aufgezeichnete Audio-Material nach 180 Sekunden wieder, sodass stets drei Minuten im Zwischenspeicher verbleiben. Erkennt der Recorder durch seine eingebaute Spracherkennung eines der Aktivierungsworte der untersuchten VUIs (»Alexa«, »Hey/Okay Google« und »Hey Siri«), werden die drei Minuten gespeichert und drei weitere Minuten aufgezeichnet. Wird ein weiteres Aktivierungswort während der Aufzeichnungsphase erkannt, verlängert sich das Recording entsprechend. So ergeben sich Audio-Aufnahmen, in deren Verlauf das Aktivierungswort fällt, sodass die Anbahnung eines VUI-Dialogs und die anschließende kommunikative Bearbeitung analysierbar wird. Zusätzlich wurden die in den Smartphone-Apps hinterlegten *Protokolldaten* erhoben, d.h. Dokumentationen zuvor durchgeführter VUI-Dialoge. Alle drei Datentypen wurden nach dem gesprächsanalytischen Standard GAT 2 (vgl. Selting et al. 2009) bzw. dem multimodalen Transkriptionsstandard nach Mondada (2016) transkribiert und anschließend inventarisiert.[9] Insgesamt kann auf rund zwei Stunden Video-Material sowie rund 32 Stunden Audio-Material zurückgegriffen werden.

Die Audio-Aufzeichnungen werden mit einem gesprächsanalytischen Grundansatz untersucht. Damit ist gemeint, dass eine sequenzanalytische Vorgehensweise bei der Auswertung der Aufzeichnungen die Basis der Untersuchung bildet (vgl. Deppermann 2008: 53). Damit werden Muster in der Gestaltung von VUI-Dialogen erkennbar. Zugleich wird die Sequenzanalyse in den Dienst der medienlingu-

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

7 Dabei wurden Modelle von Amazon (Amazon Echo Dot, 3. und 4. Generation), Apple (HomePod Mini) und Google (GoogleHome Mini und Google Nest) eingesetzt.

8 Der CVR wurde erstmals von einer Forscher*innengruppe an der University of Nottingham entwickelt (vgl. Porcheron et al. 2018) und für das Forschungsprojekt in Siegen erweitert (vgl. Hector et al. 2022).

9 Für die Arbeit an den Transkripten danke ich Franziska Niersberger-Gueye und Sarah Diehl.

istischen Erforschung von sprachlichen Praktiken mit VUIs gestellt. Zusätzlich zu den analysierten Ausschnitten werden daher im Sinne einer ethnografischen Erweiterung gesprächsanalytischer Verfahren (vgl. Deppermann 2000) Hintergrundinformationen über die Sprecher*innen (z.B. deren Beziehung zueinander, deren Interaktionsgeschichte, deren Erfahrungen im Umgang mit Sprachassistenzsystemen u.a.) mit erhoben und teilweise in die Analysen einbezogen.[10] Im Falle der Untersuchung der Protokolldaten ist zwar immer noch die zur Verfügung stehende Aufzeichnung Ausgangspunkt der Analyse, allerdings wird diese hier in Beziehung zur grafischen App-Umgebung analysiert. Für Letzteres wird ausgehend von den sprachlichen Praktiken der Nutzer*innen deren weitere Dokumentation und Repräsentation in den Apps betrachtet. In diesem Sinne weist das methodische Vorgehen auch zu Teilen Merkmale der »Walkthrough«-Methode auf, in der systematisch dem Aktivitätsfluss einer Anwendung gefolgt und dies zugleich dokumentiert wird (vgl. Light et al. 2018: 882).[11] Die Auswahl der präsentierten Daten hat dabei in allen Fällen exemplifizierenden Charakter, die Analyse ist mithin qualitativ-explorativ zu verstehen.

4. Interface-Analyse

Die Analyse ist zweigeteilt und konzentriert sich im ersten Teil auf die VUI-Dialoge und ihre Entstehung. Dabei stehen die folgenden Leitfragen im Mittelpunkt: Wie sind VUI-Dialoge konversationell aufgebaut, d.h. wie nehmen die Nutzer*innen sprachlich Bezug auf ein maschinelles Gegenüber und wie sind die synthetisch generierten stimmlichen Äußerungen als Bezugnahmen auf die Eingaben gestaltet? Welche spezifischen sequenziellen Muster lassen sich dabei beschreiben? Lassen sich Verwebungen der Dialoge mit den sozialen Situationen beobachten, in denen sie ausgeführt werden? Der zweite Teil wendet sich der Dokumentation in der Smartphone-App zu: Wie sind »Aktionen« mit einem VUI, d.h. Stimmein- und -ausgaben, im GUI dokumentiert und repräsentiert? Welche Bestandteile eines VUI-Dialogs sind dabei inkludiert und exkludiert, welche Zusatzinformationen werden bereitgestellt? Welche Rückschlüsse erlaubt dies über die seitens der Dienstanbieter vorgenommene Datenverarbeitung und -verwertung?

10 Zu methodologischen Überlegungen für die Analyse von Smart Speakern in der Praxis siehe Hector (2022).

11 Die Methode ist allerdings v.a. für die Dokumentation und Kritik von GUIs, insbesondere Smartphone-Apps und deren Interfaces, etabliert. Da hier VUIs und die damit verbundenen sprachlichen Praktiken als Ausgangspunkt der Analyse gesetzt werden und zudem teilweise gesprächsanalytisch gearbeitet wird, ergibt sich eine etwas andere Vorgehensweise, die z.B. keine vollständige App-Dokumentation beinhaltet.

4.1 Sprachliche Praktiken in Voice User Interfaces

Um zunächst die seitens des VUI-Designs angelegten ›Optionen‹ genauer zu betrachten, lohnen die Eröffnungsdialoge, die bei der Ersteinrichtung eines Smart Speakers abgespielt werden, einer genaueren Betrachtung. Das nachfolgend präsentierte Beispiel (1) stammt aus einem Haushalt, in dem die 21-jährigen Zwillingsbrüder Till (TW) und Konrad (KW) als Wohngemeinschaft zusammenleben. In dem Ausschnitt nehmen sie den HomePod von Apple in Betrieb.

Beispiel (1): Einführungsdialog

```
389 SI:  hall[o ich] bin SIri;
390 TW:      [(h?) ]
391 p:   (0.2)
392 SI:  willkommen beim HOMEpod.
393  p:  (0.4)
394 SI:  du SIEHST es zwar nicht (.) aber ich wInke gerade,
395 KW:  ((lacht, ca. 2 Sek.))
396 SI:  wenn du meine aufmerksamkeit möchtest (.) sag EINfach-
397  p:  (0.2)
398 SI:  hey SIri-
399  p:  (0.9)
400 SI:  lass uns LOSlegen (.) sAg.
401  p:  (0.3)
402 SI:  hey sIri (.) was KANNST du?
403  p:  (0.5)
404 KW:  °h
405  p:  (0.5)
406 KW:  hey sIri (.) was KANNST du?
407  p:  (1.3)
408 KW:  °hh
409  p:  (0.7)
410 SI:  ich kann VIELE dinge erledi[gen(.) zum beispiel das lIcht
         einschalten (.) die aktu]ellen nAchrichten abrufen (.) und dir
         sagen wie das wEtter wird;
411 KW:                            [<<lachend> he he he DANke.>
                                  ]
412  p:  (0.7)
413 SI:  probier_s jetzt mal MIT (.) hey siri spiel musik,
414  p:  (0.9)
415 KW:  °h hey siri (.) spiel muSIK;
416  p:  (2.1)
417 KW:  °hh
418  p:  (0.4)
419 SI:  gerne hier kommt muSIK.
420  p:  (0.3)
421 SI:  [ ex]tra für dich zuSAMmenge[stellt;]
422 KW:  [HM.]
423 TW:                              [OH.    ]
424  p:  (0.3)
425 TW:  AHA,
426  p:  (3.2)
427 SI:  ((spielt „Copacabana“ von Leon Machère, ca. 10 Sek.))
```

Der Ausschnitt zeigt die ersten verbalsprachlichen Äußerungen des eingerichteten Smart Speakers. Die synthetisierte Stimme, die die maschinelle Seite des Interfaces darstellt, präsentiert sich – durch ihre onymische, d.h. namentliche, Selbstvorstellung (Z. 389) und durch den Verweis auf einen unsichtbaren menschlichen Körper (Z. 394) – als menschenähnliche Einheit, die adressiert werden kann. Damit wird die Metaphorik eingeführt, die in VUIs zur Strukturierung des Interfaces eingesetzt wird (vgl. Natale/Cooke 2021: 1009; siehe auch Abschnitt 2). Diese Strukturierung des stimmbasierten Interfaces und das Einlernen in diese Metapher setzt sich im weiteren Verlauf des Ausschnitts weiter fort: Die synthetische Stimme äu-

ßert eine mögliche Stimmeingabe und fordert die Nutzer auf, dies nachzumachen (Z. 400ff.), was Konrad auch entsprechend realisiert (Z. 406). Diese ›Übung‹ wird ein weiteres Mal wiederholt (Z. 413ff.). Dabei wird ein bestimmter sequenzieller Ablauf geprobt: eine Adressierung durch eine Interjektion und das folgende Onym »Siri«, gefolgt zunächst von einer kurzen Pause, in der das Gerät den Listening-Modus aktivieren kann und die nachfolgenden akustischen Signale verarbeitet, und anschließend von der Stimmeingabe, auf die wiederum eine synthetisierte Stimmausgabe erfolgt. Dieser Sequenzablauf lässt sich wie folgt darstellen (vgl. Hector i.V.):

Invokation [Interjektion + Onym] – Aktivierung des Listening-Modus – Stimmeingabe – Stimmausgabe

Bei der zweiten ›Übung‹ im oben dargestellten Beispiel folgt auf die Stimmausgabe noch die praktische Umsetzung der in der Stimmeingabe angeforderten Tätigkeit. Die Stimmausgabe erscheint insofern eher als dialogisches Scharnier zwischen der Stimmeingabe und der praktischen Umsetzung (z.B. Wiedergabe von Musik):

Invokation [Interjektion + Onym] – Aktivierung des Listening-Modus – Stimmeingabe – Stimmausgabe-Scharnier – Praktische Umsetzung

Dieser Sequenzablauf scheint auf den ersten Blick selbstverständlich zu sein, ist aber nicht trivial, denn dabei verlängern sich Praktiken der zwischenmenschlichen Gesprächsorganisation ins VUI. So ist der Ablauf der Eröffnung eines Austauschs – für den VUI-Dialog als Invokation und anschließende Aktivierung des Listening-Modus beschrieben – für zwischenmenschliche Interaktionen umfangreich als *Summons-Answer-Sequenz* diskutiert worden (vgl. Schegloff 1968). Dabei ist noch eine Parallele auffällig: Der Apple HomePod signalisiert durch ein visuelles Signal an seiner Oberfläche, dass der Listening-Modus aktiviert wird, also die nachfolgenden akustischen Signale als Stimmeingaben verarbeitet werden. In der von Schegloff (vgl. ebd.) beschriebenen Summons-Answer-Sequenz für zwischenmenschliche Interaktionen können ebenfalls visuelle Signale eine Reaktion auf einen ›Summons‹, also eine erste Adressierung sein – z.B. ein Blick oder eine veränderte Körperhaltung. Die Flexibilität im Hinblick auf die sprachliche Realisierung der Invokation in VUI-Dialogen ist allerdings eingeschränkt: Diese muss im VUI-Dialog immer mit der Interjektion »Hey« sowie dem Onym »Siri« realisiert werden, die Möglichkeit eines sprachlich anders gestalteten Summons besteht (anders als in zwischenmenschlichen Summons-Answer-Sequenzen, die hierfür eine Vielzahl verschiedener verbaler und nonverbaler Realisierungsmöglichkeiten ausgebildet haben) aufgrund der technischen Gegebenheiten nicht.

Das VUI erfordert für sein erfolgreiches Zustandekommen als Dialog außerdem eine strikte Abfolge von Redezügen, die hier erprobt und vorgemacht wird. Die Redezugorganisation, die in der konversationsanalytischen Forschung bekannt ist als *Turn-Taking* (vgl. Sacks/Schegloff/Jefferson 1974), wird in zwischenmenschlichen In-

teraktionen von kompetenten Sprecher*innen mühelos und »mit schlafwandlerischer Sicherheit und Präzision« (Auer 2020: 106) vollzogen – auch dann, wenn z.B. auf dem konversationellen *floor* eine Auseinandersetzung darüber stattfindet, welche*r der Sprecher*innen als nächstes am Zug ist. Das Rederecht in VUI-Dialogen ist aber nicht wie in Alltagsgesprächen Gegenstand einer fortlaufenden Aushandlung, sondern ist vielmehr – ähnlich wie bei institutionellen Zuweisungen (vgl. Clayman 2012) – strikt reguliert: Erst nach der Invokation und der Aktivierung des Listening-Modus kann nutzer*innenseitig gesprochen werden, und zwar bis zum Beginn der Verarbeitung, daraufhin folgt die Stimmausgabe. Wird von dieser Abfolge abgewichen, droht der VUI-Dialog zu scheitern und das konstituierte Interface folglich zusammenzubrechen. Das Zustandekommen des Interfaces ist insofern an spezifische und fragile konversationelle Bedingungen geknüpft, die sich aus dessen akustisch-sequenziellem Charakter ergeben. Diese bauen auf gesprächsorganisatorischen Praktiken auf, die jedoch durch den Gebrauch in einem VUI neu konfiguriert werden: Der rigide sequenzielle Ablauf einschließlich der Notwendigkeit einer Invokation vor jeder Stimmein- und -ausgabe sowie die Restriktionen etwa in der Lexemwahl, im syntaktischen Aufbau (vgl. Barthel/Helmer/Reineke 2023) und im Turn-Taking, bilden spezifische Interface-Bedingungen, die VUI-Dialoge maßgeblich von zwischenmenschlichen Interaktionen unterscheiden. Das Interface konstituiert sich also genau in dem entstehenden Korridor, dem Zusammenspiel zwischen den technologischen Vorgaben, Möglichkeiten und Grenzen und dem gesprächsorganisatorischen Wissen der Nutzer*innen.

Wie sich in Beispiel (1) ebenfalls zeigt, sind VUI-Dialoge an die sozialen Situationen angebunden, in denen sie entstehen. Hinweise darauf sind etwa das Gelächter (Z. 395) und das sequenzielle Schließen der Sequenz etwa durch Responsive und Erkenntnisprozessmarker (u.a. Z. 411, 422f.), die zwar einerseits auf der sprachlichen Oberfläche einen fortgesetzten VUI-Dialog andeuten, die allerdings bei genauerer Betrachtung auch der Blickverhältnisse und der Körperhaltung[12] vielmehr als an die ko-präsenten menschlichen Interaktionsteilnehmer*innen gerichtet verständlich werden. Teile der Äußerungen in VUI-Dialogen sind insofern nur über die situative, soziale Einbindung, die lokalen Bedingungen ihrer Produktion, erklärbar. Wie das folgende Beispiel zeigt, können Sprecher*innen einen ganzen VUI-Dialog funktionalisieren, um gänzlich andere kommunikative Aufgaben in einer sozialen Situation zu erfüllen:

12 Eine detaillierte multimodale Betrachtung führt hier zu weit; siehe dazu etwa Hector (2022) oder Habscheid et al. (2023: 17f.).

Beispiel (2a): Pudding

```
068  AS:  <<nuschelnd> hab so lust auf NACHtisch.>
069       MAN.
070   p:  (2.2)
071  SR:  ((unverständlich, ca. 3 Sek.))
072       jo KOMM.
073       MAchen;
074       MAchen.
075  AS:  (--) nja is ja AUCH kacke.
076  SR:  ne ich MACH dir jetzt n pudding.
077       [<<f>PUDdi:ng;>    ]
078  AS:  [ich weiß gar nicht] ob wir überhaupt pudding HAben
          oder so was.
079  SR:  aLEXa?
080   p:  (0.7)
081  SR:  spiel PUDding;
082   p:  (2.0)
083  AL:  pudding von sheef von SPOtify;
084  SR:  ((lacht))
085  AL:  ((spielt Musik, „Pudding“ von Sheef, bis Z.100))
```

Der Auszug beinhaltet einen VUI-Dialog, der zur Wiedergabe des Musiktitels »Pudding« von Sheef führt (Z. 079–083). Der VUI-Dialog selbst ist dabei sprachlich gestaltet wie in Beispiel (1): Auf die Invokation (Z. 079) folgt die Stimmeingabe (Z. 081). Daraufhin produziert die synthetische Stimme zunächst eine Stimmausgabe mit Scharnierfunktion und anschließend wird der angeforderte Titel wiedergegeben. Der Blick auf das Vorgeschehen zu diesem VUI-Dialog macht diesen allerdings in seinem Kontext als situierte Humoraktivität verständlich: Andrea (AR), die Ehefrau von Sam (SR), äußert im Anschluss an eine weiter zurückliegende Interaktion, dass sie sich Nachtisch wünscht (Z. 068). Nachdem Sam mehrfach eine entsprechende Aktivität ankündigt (Z. 072–077) und Andrea entsprechende Skepsis äußert (Z. 078), realisiert er daraufhin den VUI-Dialog zur Wiedergabe des entsprechenden Lied-Titels. Damit bricht er humorvoll mit seiner ursprünglichen Ankündigung, tatsächlich einen Pudding zu machen. Für sich genommen stellt das Ziel des VUI-Dialogs also die Wiedergabe des entsprechenden Liedtitels dar, im Kontext der Nachtisch-Diskussion ist er aber vielmehr Mittel zum Zweck der Gestaltung eines humorvollen Beitrags. Die spezifischen Bedingungen des Interfaces führen also zu einem gewissen Grad an Vorhersehbarkeit und Vor*führ*barkeit: Durch die Regelhaftigkeit eines VUI-Dialogs und dessen Plan-Basiertheit können sie situiert als Humoraktivität eingesetzt werden (letzteres gilt freilich auch, wenn Plan-Brüche entstehen und Reparaturen notwendig werden oder der VUI-Dialog scheitert, siehe etwa Hector/Hrncal 2024). Durch seine Prozessierung auf dem akustischen Kanal ist der VUI-Dialog auch für Andrea hör- und nachvollziehbar. Somit wird die Aktivität des Abspielens des Liedes (nicht die Liedwiedergabe selbst) zum Humor-Gegenstand. Dies zeigt, dass VUI-Dialoge mit ihrer lokalen, sozialen Umgebung verwoben sind und darin z.B. Humorfunktionen übernehmen können. In dieser Eingebundenheit können sie für die Nutzer*innen auf einer zweiten Ebene anders Sinn erzeugen als lediglich als Dialog zur Bedienung eines Geräts.

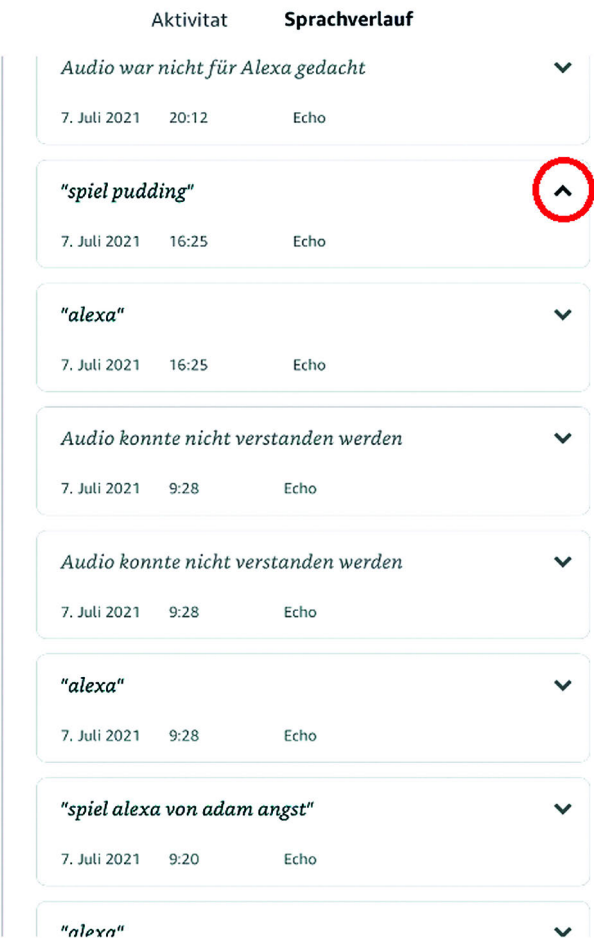
4.2 Repräsentationen sprachlicher Praktiken in grafischen Interfaces

Folgt man den Interfaces digitaler Assistenzsysteme weiter, führt von den VUIs der nächste Schritt zu den Smartphone-Anwendungen, die in einer – je nach System unterschiedlich ausgestalteten – komplementären Beziehung zu Smart Speakern stehen. Im Folgenden steht die visuelle Darstellung des Aktivitätenprotokolls eines Smart Speakers von Amazon im Mittelpunkt.[13] Dieser ›Sprachverlauf‹ arbeitet mit der visuellen Repräsentation gesprochensprachlicher Eingaben in Form von Transkripten und eröffnet zudem die Möglichkeit einer akustischen Wiedergabe des Aufgezeichneten (vgl. Habscheid et al. 2021). Zugleich ist diese Darstellung ein Hinweis darauf, welche Bestandteile der sprachlichen Äußerungen erfasst und verarbeitet wurden. Die Repräsentation eines Ereignisses entsteht hier auf einer Sehfläche, in die die Audio-Aufzeichnung des erfassten Ausschnitts eingebettet ist. Diese Repräsentation ist ein ›Zerrspiegel‹ der VUI-Dialoge als GUI, das die Begegnungen von Mensch und Maschine noch einmal neu kontextualisiert, unter anderen Bedingungen wiederholbar werden lässt und somit eine neue Begegnung schafft, die über visuelle Repräsentation und Eingaben über den Touch-Bildschirm (oder über den Computerbildschirm und die -maus bzw. -tastatur) vermittelt wird. Die Grenzfläche zwischen Technologie und Mensch wird hier also einer Befragung aus einem zweiten Blickwinkel zugänglich und zeichnet ein Bild der VUI-Dialoge aus der maschinellen Perspektive.

Die Aufzeichnung im Sprachverlauf ist standardmäßig aktiviert, kann jedoch auch unterbrochen oder dauerhaft deaktiviert werden. In dem entsprechenden Menü finden sich die VUI-Dialoge »als verschriftete isolierte Einzelelemente ohne die Situation, in der sie realisiert werden« (Habscheid et al. 2021: 40). Betrachten wir als Beispiel, wie hier die »Aktivität« aus dem Beispiel (2a) (»Pudding«) weiter oben dokumentiert ist:

13 Für eine detailliertere Beschreibung der App-Oberfläche siehe Habscheid et al. (2021: 18–21).

Beispiel (2b): Pudding (Aktivitätenprotokoll)



Der Eintrag für die Aktivität ist dabei eingebettet in eine Reihe anderer Einträge im Sprachverlauf. Dabei ist festzustellen, dass die Invokation als eigene Aktivität darunter dokumentiert ist und insofern die Ganzheitlichkeit des VUI-Dialogs in dieser Darstellung aufgebrochen wurde. Nicht nur der situative Kontext, sondern auch der Zusammenhang der eigentlichen sprachlichen Praktik – der gesprächsähnliche Sequenzverlauf – werden in der visuellen Darstellung aufgelöst. Auch Reparaturversuche der Nutzer*innen, die sich über mehr als einen VUI-Dialog erstrecken – z.B. die wiederholte und ggf. prosodisch, syntaktisch oder lexisch angepasste Formulierung einer Stimmeingabe (vgl. Hector/Hrncal 2024) – sind nicht in ihrem Zusammenhang dokumentiert. In der Übersichtsdarstellung ist insofern nicht nur

der sequenzielle Verlauf eines einzelnen VUI-Dialogs nicht zusammenhängend dokumentiert, sondern auch Beziehungen zwischen einzelnen VUI-Dialogen können nicht abgebildet werden, obwohl etwa bei scheiternden und anschließend reparierten Anfragen mehrere Einzel-Dialoge kurz hintereinander auftreten, die sprachlich aufeinander aufbauen (siehe etwa das Beispiel bei Merkle/Hector i.E.).

Die App-Nutzer*innen können sich die Details einzelner Aufzeichnungen durch eine Berührung der Pfeilsymbole am rechten Rand der Sehfläche anzeigen lassen (im Beispiel (2b) ist bereits der Eintrag für »Pudding« ausgewählt und hervorgehoben). Der Eintrag erscheint daraufhin als ein größerer Eintrag zwischen den eingeklappten Einträgen, wie in Beispiel (2c) zu sehen:

Beispiel (2c): Pudding (Aktivitätenprotokoll)

Dieses Menü enthält neben der Wiedergabefunktion eine Transkription des Audios sowie eine ausschließlich *schriftliche* Dokumentation der Reaktion des Sprachassistenten. Außerdem findet sich ein roter Button zur Löschung des Audios. In einem weiteren Abschnitt finden sich Datum, Uhrzeit und das genutzte Gerät. Wiederum darunter besteht die Möglichkeit eines »Feedbacks« über den Sprachverlauf. Der Eintrag ist insofern nicht mehr eingebettet in die soziale Situation der Humoraktivität, die weiter oben beschrieben wurde. Vielmehr werden nun Fragmente des VUI-Dialogs visuell repräsentiert. Auch anschließende Kommentierungen wie im folgenden Beispiel werden dabei getilgt:

Beispiel (3a): Tell me why

```
026   k:  ((Klappergeräusche, ca. 9.8 Sek.))
027  AS:  <<singend> tell me WHY,
028       AIN_T (nothing),
029       mh_mh_mh_MH_mh;
030       ELL me why,>
031  SR:  aLEXa?
032       SPIEL,
033       BACKstreet boys tell me why.
034   p:  (2.0)
035  AL:  hier ist i want it that WAY,
036       auf amazon MUsic.
037  SR:  ah (.) (so HEIßt das),
038  AS:  JA.
039  AL:  ((spielt Musik, „I want it that way", Backstreet
          Boys, bis Z. 046))
```

Der Auszug aus demselben Haushalt von Andrea und Sam zeigt, dass der VUI-Dialog als Reaktion darauf verstanden werden kann, dass Andrea ein Lied von den Backstreet Boys singt, woraufhin Sam dieses über den Smart Speaker wiedergeben lassen will. Dabei formuliert er die Stimmeingabe als Imperativ (Z. 032), gefolgt von Interpret und der Songzeile »tell me why«, die Andrea zuvor gesungen hatte (Z. 029f.). Die Stimmausgabe des Smart Speakers kündigt die Wiedergabe des Titels an, ›korrigiert‹ dabei aber im Stimmausgabe-Scharnier vor dem Abspielen des Lieds die Eingabe zum eigentlichen Titel des Stücks (»I want it that way«, Z. 035). Dies wird von Sam und Andrea kommentiert, bei Sam eingeführt durch einen Erkenntnisprozessmarker (vgl. Imo 2009), der die durch die synthetische Stimme durchgeführte Reparatur quittiert, und von Andrea bestätigt (Z. 037ff.). Das Singen von Andrea, das den VUI-Dialog initiiert, ebenso wie die Reparatursequenz und die Invokation sind nicht im Aktivitätenverlauf dokumentiert:

Beispiel (3b): Tell me why

Die beschriebenen visuellen Elemente für einen einzelnen Eintrag (Beispiele (2c) und (3b)) lassen sich noch genauer in ihrer Funktionalität analysieren. Wie Habscheid et al. (2021: 44) feststellen, kann die »angebotene Möglichkeit einer Löschung […] als Reaktion von Amazon auf die öffentliche Debatte bezüglich der Verwendung von personenbezogenen Daten durch Internetfirmen verstanden werden«. Die rote Hervorhebung und deutliche Platzierung des Buttons unterstützen auch visuell den

Eindruck einer möglichen Kontrolle durch die Nutzer*innen, obschon nicht deutlich wird, was eine Löschung des Audios tatsächlich bewirkt – wie eine Inspektion der App-Funktionen zeigt, bleiben der visuelle Eintrag und die zugehörigen Zusatz-Daten in der App erhalten und zum Zeitpunkt der Ansicht in der App hat eine gewisse Verarbeitung notwendigerweise schon stattgefunden.

Die Möglichkeit des nutzer*innenseitigen »Feedbacks« ist eine Funktion, mit der der Hersteller die Endkund*innen an der Arbeit am Trainingsdatenmaterial beteiligt, in das die aufgezeichneten Daten anschließend eingehen. Die Funktion ist ein Hinweis darauf, dass das aufgezeichnete Material, die Transkription und die ausgelöste Aktivität in ihrem Zusammenhang weiter für die Optimierung der Spracherkennungsverfahren verwertet werden (siehe dazu ausführlich Strüver 2023b: 15f.), und zwar sowohl durch automatisierte Verfahren wie auch durch menschliche Trainings- und Kontrollprozesse, die im Wesentlichen durch prekär beschäftigte Clickworker*innen durchgeführt werden (vgl. Waldecker/Volmar 2022: 167). Wenn dabei von »redaktionelle[r] Mikroarbeit« (ebd.) die Rede ist, lässt der Blick auf das Aktivitätenprotokoll im Abgleich mit der sozialen Situation verständlich werden, mit welcher Herausforderung die Arbeiter*innen hierbei im Dienst des KI-Trainings konfrontiert werden: Der Rekonstruktion des Verhältnisses fragmentarischer Ton-Aufzeichnungen zu deren Transkript und der dokumentierten »Aktion«.

Die verbal vollzogenen Interface-Praktiken werden also in Datenpunkte und somit in Infrastrukturen übersetzt. Die Ansicht der Aktivitäten in GUIs sind Dokumentation der Verdatung und Plattformisierung[14] des vollzogenen VUI-Dialogs (und somit der sozialen Praxis). An dieser Übersetzung sind zu diesem Zeitpunkt zwei Interfaces beteiligt: Das VUI als akustisch prozessierte Begegnung zwischen Mensch und Computer und das GUI als visuelles Abbild dieser Begegnung, genauer gesagt als Abbild eines Datenbank-Eintrags zur Dokumentation und weiteren Interaktion mit dieser Begegnung. Im Abgleich des Eintrags mit der längeren Audio-Aufzeichnung von der sozialen Praxis, aus der heraus er entstanden ist, werden diese Übersetzungen als »Umschlagspunkte« (Kaerlein 2020: 54) erkennbar, »an denen aus Praxis Daten generiert werden bzw. Daten sich wiederum in Ketten von Praktiken übersetzen« (ebd.). Die Übertragung lässt aus einem sich dynamisch in der Zeitlichkeit entfaltenden Geschehen einen geronnenen und sich primär in der Räumlichkeit ausdehnenden, wiederholbaren Datenpunkt werden, der mit zusätzlichen Informationen (einer Transkription, Datum und Uhrzeit der Anfrage, verwendetes Gerät, daran anschließende Reaktion des Systems) versehen und zur Weiterverarbeitung vorbereitet ist.

14 Zur Konzeption von Sprachassistenzsystemen als Plattform-Technologien siehe Goulden (2019) und Strüver (2023a: 104f.).

4.3 Repräsentationen des Nicht-Verstehens

Die genauere Betrachtung des GUIs zeigt außerdem die Fehleranfälligkeit und -verarbeitung der Systeme, wie das folgende Beispiel aus einem anderen Haushalt zeigt:

Beispiel (4): Licht an[15]

Dokumentiert ist hier ein Fall, in dem die Transkription des Audios nicht zur Tonspur passt, was auf Fehler im Speech-Recognition-Prozess hindeutet. Tatsächlich ist in dem Audio undeutlich zunächst der Name eines im Haushalt lebenden Bewohners, Alex, zu hören – wohl der Auslöser für die Fehlaktivierung – und fälschlicherweise als Invokation verarbeitet; anschließend ist eine nicht gut zu verstehende Folgeäußerung dokumentiert:

```
001  M1:  aLEX (hat_n bestimmten gefunden.)
```

Die daran anschließend produzierte und schriftlich hinterlegte Äußerung – »Tut mir leid, ich kann keine Gruppe oder kein Gerät mit dem Namen licht finden« – bezieht sich entsprechend auf den systemseitig verstandenen Text. Davon abweichend wird im folgenden Beispiel der Fehler systemseitig erkannt und klassifiziert:

15 In diesem und in den folgenden Beispielen wird auf die Gegenüberstellungen zwischen Protokolleinträgen und Aufzeichnungen der sozialen Situation verzichtet, da diese für die Illustration der nachfolgenden Argumente nicht wesentlich sind.

Beispiel (5): Audio war nicht für Alexa gedacht

Eine gesprächslinguistische Transkription des hinterlegten Audios zeigt an, dass in der Tat kein Versuch eines VUI-Dialogs erkennbar ist. Auch hier wird vermutlich durch den Namen des Bewohners Alex die Invokation ausgelöst:

```
001  M1:  aLEX (du hattest das ma in kunst.)
```

Allerdings wird die anschließende Äußerung nicht transkribiert. Zugleich kann konstatiert werden, dass die Aufzeichnung dennoch verarbeitet und gespeichert wurde. Auch zu diesem Audio besteht die Möglichkeit, es löschen zu lassen oder die Ausführung des »Befehls« zu bewerten. Die zwar einerseits »nicht für Alexa gedachte« und nicht transkribierte Aufnahme wird so dennoch Teil des Sprachverlaufs und einer hintergründigen Verarbeitung. Davon noch einmal zu differenzieren sind Einträge, zu denen eine leicht davon abweichende Klassifizierung hinterlegt ist:

Beispiel (6): Audio konnte nicht verstanden werden

Anstatt eines Transkripts des Audios findet sich hier der Vermerk »Audio konnte nicht verstanden werden«. Eine Transkription der Audio-Aufzeichnung ergibt eine Äußerung, die in einem persönlichen Gespräch oder Telefonat gefallen sein dürfte:

```
001  M2:  dachte mir so (LASS ma_s) training;
```

Für die Nutzer*innen besteht hier keine Transparenz darüber, warum die Klassifizierung hier anders ausgefallen ist als in Beispiel (5) – offensichtlich war auch dieses Audio »nicht für Alexa gedacht«. Ebenso unklar ist, warum überhaupt zu diesem Zeitpunkt das Recording aktiviert war.

Die Beispiele (5) und (6) belegen, dass Verarbeitungsfehler antizipiert werden und entsprechende Klassifizierungen dafür vorgesehen sind. Die Darstellung im GUI weist insofern darauf hin, dass – in der Begegnung mit der nächsten maschinellen Einheit – auch scheinbar fehlerhafte oder versehentlich entstandene Aufzeichnungen weiterverarbeitet werden können: Sie werden *als Fehler* dokumentiert, klassifiziert und ausgewertet. Aus dieser Perspektive betrachtet, handelt es sich bei den hier dokumentierten Einträgen insofern also nicht um Verarbeitungsfehler, sondern um einkalkulierte (und ethisch problematische) Vorgänge zur zusätzlichen Gewinnung von Audio-Daten. Dies verweist auf den uneindeutigen Zusammenhang zwischen dem Verhalten der Nutzer*innen und dessen Verdatung, den auch Paßmann/Gerlitz (2014) für Social-Media-Plattformen wie Twitter und Facebook besprechen.[16] Sie betonen, dass mit den aufgezeichneten Plattform-

16 Auch Sprachassistenzsysteme können als Plattform-Technologien betrachtet werden, siehe dazu Goulden (2019) und Strüver (2023a: 104f.).

Daten über die Aktivitäten der Nutzer*innen[17] kein vollständiges, umfänglich verständliches Bild eines individuellen Nutzungsverhaltens gezeichnet werden kann – ein Befund, der sich für VUI-Dialoge bestätigen lässt. Sie argumentieren weiter, dass insbesondere soziales Verhalten mit solchen Formen fragmentarischer Daten kaum erfassbar ist: »Es handelt sich zwar um vorstrukturierte Aktivitäten, doch diese sind zugleich unterbestimmt« (Paßmann/Gerlitz 2014: 2). Genau diese »Unterbestimmtheit oder Vagheit« (ebd.: 3) wird im Sprachverlauf sichtbar und ist kein Zufall: Die Gliederung nach einzelnen ›Aktivitäten‹ (nicht nach der tatsächlichen sequenziell organisierten Ausführung des VUI-Dialogs), die nur sehr kurzen und kaum zuzuordnenden Audio-Ausschnitte sowie die unterschiedlichen Klassifizierungen von Nicht-Verstehen legen den Schluss nahe, dass hier die Vorstrukturierung im Sinne einer weiteren Verarbeitung der Aufzeichnungen als Trainingsdaten für die Verbesserung der Dienste u.a. im Bereich der Speech Recognition sichtbar wird. Diese ist auf möglichst vergleichbare und daher kurze Aufzeichnungen als Daten angewiesen, um möglichst effizient Mustererkennung betreiben zu können (vgl. Pasquinelli 2019: 11) – nicht aber auf vollständige Bilder des Nutzungsverhaltens.

5. Fazit und Ausblick

Die Konzeptionalisierung von Interfaces als Grenzfläche nach Hookway (2014: 59) und die darauf aufbauende Analyse von VUI-Dialogen, mit denen diese Grenzfläche ausgestaltet wird, konnte verdeutlichen, dass bestimmte sprachliche Bedingungen erfüllt sein müssen, um in einem VUI-Dialog einen Austausch zwischen Nutzer*innen und Maschine zu etablieren. Beide Einheiten sind dazu auf einen rigiden Sequenzablauf orientiert. Die sprachlichen Praktiken folgen zwar im Grundsatz gesprächsorganisatorischen Prinzipien, unterliegen jedoch bestimmten Restriktionen. Sie sind mithin Interface-Praktiken, d.h. Äußerungen, deren sprachliche Gestaltung nicht nur an gesprächsorganisatorischen Prinzipien ausgerichtet ist, sondern auch Spuren des Gebrauchs im Interface trägt. VUI-Dialoge sind zudem durch die situativen, sozialen Kontexte geprägt, in denen sie stattfinden, und können dabei auch über die Bedienung der Maschine hinaus funktionalisiert werden.

Während die analysierten Audio-Aufzeichnungen also die soziale Situation und die Konstruktion des VUI-Dialogs verständlich werden lassen, macht der Aktivitätenverlauf als GUI die maschinelle Perspektive auf die Situation sichtbar: Es do-

17 An dieser Stelle sind die Aktivitäten der Nutzer*innen gemeint, die im Rahmen bestimmter, von den Betreiberfirmen angebotenen und vorstrukturierten Optionen stattfinden, z.B. »Like« und »Share« auf Social Media oder eben ein vordefinierter Ablauf für einen VUI-Dialog bei Sprachassistenzsystemen.

kumentiert, wie sprachliche Eingaben segmentiert und zur weiteren Verarbeitung vorbereitet wurden – und erweist sich dabei als blind für die Situiertheit der Nutzer*innen. Der Vergleich bestätigt insofern einerseits die ethnomethodologische Perspektive Suchmans (2007: 4f.): »human-machine communications take place at a very limited site of interchange; that is, through actions of the user that actually change the machine's state«. Andererseits zeigt sich, dass die cloudbasierte Auswertung und der Bedarf nach Audiodaten, z.B. zum Training der Spracherkennung, eine neue Dimension einbringt: Wie am Schluss gezeigt wurde, sind hier eben nicht nur diejenigen Aktionen dokumentiert, die unmittelbar eine Veränderung im Status der Maschine auslösen, sondern auch solche, die das gerade nicht getan haben. Dadurch wird eine zweite Motivation für die Aufzeichnung und Speicherung der Audios verdeutlicht, die über die reine Anwendung durch die Nutzer*innen hinausgeht: Das Sammeln von Trainingsdaten für die Verbesserung von Speech Recognition (vgl. auch Waldecker/Volmar 2022).

Sprachassistenten scheinen nicht in erhofftem Umfang rentabel zu sein – insbesondere steht die Vermutung im Raum, dass die aufgezeichneten Daten sich nicht hinreichend monetarisieren lassen (vgl. Amadeo 2022). Konversationelle Interfaces haben dennoch Konjunktur, denn anders als die hier untersuchten Interfaces basiert die jüngste Welle in diesem Bereich auf generativen Anwendungen Künstlicher Intelligenz auf Basis von LLMs, die vereinfacht gesagt mit statistischen Berechnungen die wahrscheinlichste Antwort auf einen Input generieren können und insofern nicht mehr auf die Verknüpfung mit einzelnen Funktionen angewiesen sind (vgl. Bender et al. 2021: 616). Eine Weiterentwicklung entsprechender Technologien auf akustische Prozessierung ist zeitnah zu erwarten (vgl. Knisella 2023). Medienhistorisch bedeutet dies für die Interfaces die Fortsetzung einer Linie, die Waldecker/Volmar (vgl. 2022: 168) herausarbeiten: Die Reduktion der Gelenktheit des Dialogs, die etwa bei Voice-Dial-Systemen oder in Call-Center-Anrufen durch eine akustische Menüführung noch deutlich stärker ist als bei Sprachassistenzsystemen oder auf LLMs basierenden Chatprogrammen. VUI-Dialoge laufen, wie sich auch an den Analysen zeigt, bisher sehr musterhaft und funktionsorientiert ab. Die Rigidität der VUIs könnte dann ebenfalls abnehmen – sie könnten im Hinblick auf ihre sequenzielle und sprachliche Gestaltung flexibler werden und mehr Raum für Vagheit lassen. Damit wären aber auch deutlich längere und eben gesprächsähnlichere Konversationen produzierbar, was nicht unbedingt nutzer*innenseitig gewünscht sein muss und die Fragilität von VUI-Dialogen erhöhen könnte.

Literatur

Amadeo, Ron (2022): Amazon Alexa is a »Colossal Failure«, on Pace to Lose $10 Billion this Year. Abrufbar unter: https://arstechnica.com/gadgets/2022/11/am

azon-alexa-is-a-colossal-failure-on-pace-to-lose-10-billion-this-year/ (Stand: 01.02.2024).

Auer, Peter (2020): »Die Struktur von Redebeiträgen und die Organisation des Sprecherwechsels«, in: Karin Birkner/Peter Auer/Angelika Bauer/Helga Kotthoff (Hg.), Einführung in die Konversationsanalyse, Berlin/Boston: de Gruyter, S. 106–235.

Baranovska, Marianna/Stefan Höltgen (Hg.) (2018): *Hello, I'm Eliza. Fünfzig Jahre. Gespräche mit Computern*, Bochum/Freiburg: projektverlag.

Barthel, Mathias/Henrike Helmer/Silke Reineke (2023): »First Users‹ Interactions with Voice-Controlled Virtual Assistants: A Micro-Longitudinal Corpus Study«, in: Proceedings of SemDial 2023. Abrufbar unter: https://mezzanine.um.si/en/conference/semdial-2023-marilogue/#proceedings (Stand: 02.05.2024).

Bartz, Christina/Timo Kaerlein/Monique Miggelbrink/Christoph Neubert (2019): »Zur Medialität von Gehäusen. Einleitung«, in: Christina Bartz/Timo Kaerlein/Monique Miggelbrink/Christoph Neubert (Hg.), Gehäuse: Mediale Einkapselungen, Paderborn: Wilhelm Fink, S. 9–32.

Bedford-Strohm, Jonas (2017): »Voice First? Eine Analyse des Potentials von intelligenten Sprachassistenten am Beispiel Amazon Alexa«, in: Communicatio Socialis 4, S. 485–494.

Bender, Emily/Timnit Gebru/Angelina McMillan-Major/Shmargaret Shmitchell (2021): »On the Dangers of Stochastic Parrots«, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York: ACM Press, S. 610–623.

Bender, Emily/Alexander Koller (2020): »Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data«, in: Dan Jurafsky/Joyce Chai/Natalie Schluter/Joel Tetreault (Hg.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg: Association for Computational Linguistics, S. 5185–5198.

Borbach, Christoph (2019): »Navigating (Through) Sound. Auditory Interfaces in Maritime Navigation Practice, 1900–1930«, in: Interface Critique Journal 2, S. 17–33.

Borbach, Christoph/Timo Kaerlein/Robert Stock/Sabine Wirth (Hg.) (i.E.): *Akustische Interfaces. Interdisziplinäre Perspektiven auf Schnittstellen von Technologien, Sounds und Menschen*, Wiesbaden: Springer Vieweg.

Bratton, Benjamin H. (2016): The Stack. On Software and Sovereignity, Cambridge: MIT Press.

Clayman, Steven (2012): »Turn-Cunstructional Units and the Transition-Relevance Place«, in: Jack Sidnell/Tanya Stivers (Hg.), The Handbook of Conversation Analysis, Chichester: Blackwell, S. 150–166.

Cramer, Florian/Matthew Fuller (2008): »Interface«, in: Matthew Fuller (Hg.), Software Studies: A Lexikon, Cambridge: MIT Press, S. 149–152.

Crawford, Kate/Vladan Joler (2018): »Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources«. Abrufbar unter: https://anatomyof.ai/ (Stand: 24.04.2024).

Deppermann, Arnulf (2000): »Ethnographische Gesprächsanalyse: Zu Nutzen und Notwendigkeit von Ethnographie für die Konversationsanalyse«, in: Gesprächsforschung 1, S. 96–124.

Deppermann, Arnulf (2008): Gespräche analysieren. Eine Einführung, Wiesbaden: VS Verlag für Sozialwissenschaften.

Dieter, Michael (2022): »Interface Critique at Large«, in: Convergence 0 (0), S. 1–17.

Distelmeyer, Jan (2020): »Interface II. Zur Programmatik leitender Prozesse der ›digitalen Gegenwart‹«, in: Huber/Krämer/Pias, Wovon sprechen wir, wenn wir von Digitalisierung sprechen? Gehalte und Revisionen zentraler Begriffe des Digitalen, S. 59–72.

Ernst, Christoph (2017): »Implizites Wissen, Kognition und die Praxistheorie des Interfaces«, in: Navigationen 17 (2), S. 99–116.

Ernst, Christoph/Thomas Bächle (2020): »Interface«, in: Martina Heßler/Kevin Liggieri (Hg.), Technikanthropologie. Handbuch für Wissenschaft und Studium, Baden-Baden: Nomos, S. 416–420.

Goulden, Murray (2019): »Delete the Family: Platform Families and the Colonisation of the Smart Home«, in: Information, Communication & Society 24 (7), S. 1–18.

Habscheid, Stephan (2022): »Socio-Technical Dialogue and Linguistic Interaction. Intelligent Personal Assistants (IPA) in the Private Home«, in: Sprache und Literatur 51 (2), S. 167–196.

Habscheid, Stephan/Tim Hector/Christine Hrncal (2023): »Human and Non-Human Agency as Practical Accomplishment. Interactional Occasions for Ascription and Withdrawal of (Graduated) Agency in the Use of Smart-Speaker-Technology«, in: Social Interaction. Video-Based Studies of Human Sociality 6 (1), 1–31.

Habscheid, Stephan/Tim Hector/Christine Hrncal/David Waldecker (2021): »Intelligente Persönliche Assistenten (IPA) mit Voice User Interfaces (VUI) als ›Beteiligte‹ in häuslicher Alltagsinteraktion. Welchen Aufschluss geben die Protokolldaten der Assistenzsysteme?«, in: Journal für Medienlinguistik 4 (1), S. 16–53.

Hadler, Florian/Joachim Haupt (2016a): Towards a Critique of Interfaces. In: Hadler/Haupt, Interface Critique, S. 7–13.

Hadler, Florian/Joachim Haupt (2016b) (Hg.) *Interface Critique*, Berlin: Kadmos.

Hector, Tim (i.V.): *Smart Speaker im Dialog. Sprachliche Praktiken mit Voice User Interfaces.* Berlin/Boston: de Gruyter.

Hector, Tim (2022): »Smart Speaker in der Praxis. Methodologische Überlegungen zur medienlinguistischen Erforschung stationärer Sprachassistenzsysteme«, in: Sprache und Literatur 51 (2), S. 197–229.

Hector, Tim/Christine Hrncal (2024): »Sprachassistenzsysteme in der Interaktion«, in: Jannis Androutsopoulos/Friedemann Vogel (Hg.), Handbuch Sprache und digitale Kommunikation, Berlin u.a.: de Gruyter, S. 309–328.

Hector, Tim/Franziska Niersberger-Gueye/Franziska Petri/Christine Hrncal (2022): »The ›Conditional Voice Recorder‹: Data Practices in the Co-Operative Advancement and Implementation of Data-Collection Technology«, in: Working Paper Series Media of Cooperation 23, S. 1–15.

Hookway, Branden (2014): Interface, Cambridge/London: The MIT Press.

Hoy, Matthew B. (2018): »Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants«, in: Medical Reference Services Quarterly 37 (1), S. 81–88.

Huber, Martin/Sybille Krämer/Claus Pias (Hg.) (2020): *Wovon sprechen wir, wenn wir von Digitalisierung sprechen? Gehalte und Revisionen zentraler Begriffe des Digitalen,* Frankfurt a.M.: Universitätsbibliothek Johann Christian Senckenberg.

Imo, Wolfgang (2009): »Konstruktion oder Funktion? Erkenntnisprozessmarker (Change-of-State Tokens) im Deutschen«, in: Jörg Bücker/Susanne Günther (Hg.), Grammatik im Gespräch. Konstruktionen der Selbst- und Fremdpositionierung, Berlin u.a.: de Gruyter, S. 57–86.

Kaerlein, Timo (2020): »Interface. Zur Vermittlung von Praktiken und Infrastrukturen (als Perspektive für die Medienwissenschaft)«, in: Huber/Krämer/Pias, Wovon sprechen wir, wenn wir von Digitalisierung sprechen? Gehalte und Revisionen zentraler Begriffe des Digitalen, S. 45–58.

Knisella, Bret (2023): Google Assistant and Alexa Are Both Getting Generative AI Makeovers. Welcome to the ChatGPT Era. Abrufbar unter: https://synthedia.substack.com/p/google-assistant-and-alexa-are-both (Stand: 01.02.2024).

Latour, Bruno (2005): Reassembling the Social. An Introduction to Actor-Network-Theory, Oxford: Oxford University Press.

Light, Ben/Jean Burgess/Stefanie Duguay (2018): »The Walkthrough Method: An Approach to the Study of Apps«, in: New Media and Society 20 (3), S. 881–900.

Merkle, Benedikt/Tim Hector (i.E.): » Werkzeuge und Medienpraktiken. Intelligente persönliche Assistenten und das Paradigma objektorientierten Programmierens«, in: Christoph Borbach/Timo Kaerlein/Robert Stock/Sabine Wirth (Hg.), Akustische Interfaces.

Mondada, Lorenza (2016): Conventions for Multimodal Transcription. Abrufbar unter: https://franzoesistik.philhist.unibas.ch/fileadmin/user_upload/franzoesistik/mondada_multimodal_conventions.pdf (Stand: 21.09.2023).

Natale, Simone/Henry Cooke (2021): »Browsing with Alexa: Interrogating the Impact of Voice Assistants as Web Interfaces«, in: Media, Culture & Society 43 (6), S. 1000–1016.

Pasquinelli, Matteo (2019): »How a Machine Learns and Fails«, in: spheres – Journal for Digital Cultures 5, S. 1–17.

Paßmann, Johannes/Carolin Gerlitz (2014): »›Good‹ Platform-Political Reasons for ›Bad‹ Platform Data. Zur sozio-technischen Geschichte der Plattformaktivitäten ›Fav‹, ›Retweet‹ und ›Like‹«, in: Mediale Kontrolle unter Beobachtung 3 (1), S. 1–40.

Porcheron, Martin/Joel Fischer/Stuart Reeves/Sarah Sharples (2018): »Voice Interfaces in Everyday Life«, in: Regan Mandryk/Mark Hancock/Mark Perry/Anna Cox (Hg.), Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems – CHI ›18, New York: ACM Press, S. 1–12.

Sacks, Harvey/Emanuel Schegloff/Gail Jefferson (1974): »A Simplest Systematics for the Organisation of Turn Taking in Conversation«, in: Language 50 (4), S. 696–735.

Schegloff, Emanuel (1968): »Sequencing in Conversational Openings«, in: American Anthropologist 70 (6), S. 1075–1095.

Selting, Margret/Peter Auer/Dagmar Barth-Weingarten/Jörg Bergmann/Pia Bergmann/Karin Birkner/Elizabeth Couper-Kuhlen/Arnulf Deppermann/Peter Gilles/Susanne Günthner/Martin Hartung/Friederike Kern/Christine Mertzlufft/Christian Meyer/Miriam Morek/Frank Oberzaucher/Jörg Peters/Uta Quasthoff/Wilfried Schütte/Anja Stukenbrock/Susanne Uhmann (2009): »Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)«, in: Gesprächsforschung 10, S. 353–402.

Strüver, Niklas (2023a): »Frustration Free: How Alexa Orchestrates the Development of the Smart Home«, in: Digital Culture & Society 9 (1), S. 99–123.

Strüver, Niklas (2023b): »Wieso eigentlich Alexa? Konzeptualisierung eines Sprachassistenten als Infrastruktur und Plattform im soziotechnischen Ökosystem Amazons«, in: kommunikation@gesellschaft 24, S. 1–33.

Suchman, Lucy (2007): Human-Machine Reconfigurations. Plans and Situated Actions. 2nd edition, Cambridge: Cambridge University Press.

Volmar, Axel (2019): »Productive Sounds«, in: Andreas Sudmann (Hg.), The Democratization of Artificial Intelligence, Bielefeld: transcript, S. 55–76.

Waldecker, David/Axel Volmar (2022): »Die zweifache akustische Intelligenz virtueller Sprachassistenten zwischen verteilter Kooperation und Datafizierung«, in: Anna Schürmer/Maximilian Haberer/Tomy Brautschek (Hg.), Acoustic Intellingence. Hören und Gehorchen, Düsseldorf: Düsseldorf University Press, S. 161–182.

Wirth, Sabine (2016): »Between Interactivity, Control, and ›Everydayness‹ – Towards a Theory of User Interfaces«, in: Florian Hadler/Joachim Haupt (Hg.), Interface Critique, Berlin: Kadmos, S. 17–38.