1 Einleitung, Forschungsfragen, Aufbau der Arbeit

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

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

https://doi.org/10.1515/9783111574332-001

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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