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

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

## 5.1 Projektvorstellung und Untersuchungsdesign

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

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

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

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


https://doi.org/10.1515/9783111574332-005

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

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

## 5.2 Datentyp I: Videoaufzeichnungen der Smart-Speaker-Ersteinrichtung

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

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

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

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

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

gend war auch eine ausführliche schriftliche Anleitung zur Installation der Aufnahmegeräte mit einer Abbildung exemplarischer Sitzkonstellationen sowie Formulare für die Einwilligung[147] und für die Aufwandsentschädigung. Darüber hinaus fand in sieben von acht Haushalten vor der Einrichtung (nach Eintreffen des Pakets) ein telefonisches oder videobasiertes Gespräch zwischen mindestens einem Haushaltsmitglied sowie den Erheber*innen statt, insbesondere zur technischen Instruktion im Hinblick auf die Aufnahmetechnik – allerdings nicht während der Inbetriebnahme selbst. Trotz des instruktiven Charakters dieser Bestandteile ist ethnomethodologisch interessant, wie die Teilnehmer*innen die Kameras platzieren und was sie selbst für relevante Ausschnitte halten.

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

## 5.3 Datentyp II: Audioaufzeichnungen der Smart-Speaker-Nutzung

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

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

**148** Für eine genauere Beschreibung siehe Kap. 3.3.

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

Es zeigt sich, dass für die Erhebung gesprächsanalytisch auswertbarer Daten eine andere Lösung gefunden werden musste. Bereits Porcheron et al. (2018) waren auf diese Herausforderung gestoßen und hatten zu diesem Zweck einen sogenannten „Conditional Voice Recorder“ (CVR) entwickelt. Dieser ahmt in seiner Funktionsweise in Teilen einen Smart Speaker nach: Er nimmt die akustischen Signale in der Umgebung, in der er platziert ist, auf und scannt sie auf das Aktivierungswort („Alexa“, „Hey/Okay Google“ bzw. „Hey Siri“). Außerdem speichert das Gerät temporär im Zwischenspeicher drei Minuten[149] Audiomaterial, das gelöscht und überschrieben wird, wenn kein Aktivierungswort fällt, aber gespeichert wird, sofern beim Scan ein Aktivierungswort erkannt wird. Sodann werden weitere drei Minuten aufgezeichnet und gespeichert, sodass sich Audioaufnahmen von ca. sechs Minuten ergeben. Wird ein weiteres Aktivierungswort während der Aufnahmezeit erkannt, verlängert sich die Aufnahmezeit entsprechend um weitere drei Minuten. Dazu ist der CVR aus verschiedenen Komponenten zusammengesetzt: Einem Raspberry Pi, Modell 3, einem damit verbundenen USB-Konferenzmikrofon, einem USB-Stick, auf dem die Aufnahmen gespeichert werden, einer SD-Karte, auf der das Betriebssystem gespeichert ist. Zusätzlich sind LED-Leuchten verbaut, die den Status des CVR anzeigen (alle Leuchten blinken: „Start“; grün leuchtet: „Aufnahme läuft, wird aber nicht gespeichert“; grün und rot leuchten: „Aufnahme läuft und wird gespeichert“; gelb leuchtet: Verarbeitungsfehler). Er sieht damit einem Smart Speaker – schon aufgrund seiner technischen Eigenschaften – nicht nur von seiner Funktionsweise her, sondern auch optisch (vgl. Abb. 16) im Hinblick auf Form und Größe ähnlich, wodurch er sich

**149** Die Länge des gespeicherten Audiomaterials ist variabel einstellbar.

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

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

Der von Porcheron et al. (2018) entwickelte und erfolgreich eingesetzte CVR war Grundlage für eine Implementierung und Weiterentwicklung des CVR im Zusammenhang des Forschungsprojekts. Nicht nur musste das Gerät entsprechend auf lokal zur Verfügung stehender Hardware auf Grundlage einer GitHub-Dokumentation[150] nachgebaut werden,[151] es ergab sich auch die Notwendigkeit einer Weiterentwicklung der bestehenden Software: Die vorliegende Programmierung – eine auf Python-Basis geschriebene und mit der als open source bereitgestellten Machine-Learning-Plattform TensorFlow[152] realisierte Hot-Word-Erkennung,[153] bei der eine Bibliothek von Trainingsdaten von SnowBoy zum Einsatz kam (vgl. Porcheron 2019) – war lediglich auf das Erkennen des Aktivierungsworts „Alexa“ ausgerichtet. Da jedoch alle drei der marktüblichen Hersteller berücksichtigt werden sollten, war diese Konfiguration nicht ausreichend. Die Weiterentwicklung wurde in Zusammenarbeit mit einer externen Firma realisiert,[154] die auch eine Dokumentation der vollzogenen Schritte veröffentlichte (siehe Kernel Concepts 2022). Eine umfangreiche Be-

**150** Siehe Porcheron (2019).

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

**152** Siehe https://www.tensorflow.org/.

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

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

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

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

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

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

**156** Siehe Kap. 5.2.

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

Vor dem Hintergrund der Debatte um die Natürlichkeit der Daten[157] und der von Gerwinski/Linz (2018: 108) geforderten Sensitivität diesbezüglich bei der Auswertung wurden bei der Erhebung mehrere Maßnahmen ergriffen, um die Beeinflussung der Aufnahmen durch die CVR-Technologie zu erkennen und einen Umgang damit entwickeln zu können. Erstens wurden die Weiterentwicklung und Anwendung der Technologie sowie die damit in Verbindung stehende Forschungspraxis selbst reflektiert (siehe Hector et al. 2022). Dabei wurden nicht nur die interaktiven Reflexionen der Aufgezeichneten, die sich in den Daten finden lassen, besprochen, sondern auch ausführlich das Vorgehen bei der Auswertung thematisiert und dabei die Praktiken herauskristallisiert, die bei der Forschung vollzogen wurden, um einen Umgang mit den Datentypen zu finden (vgl. Hector et al. 2022). Zweitens ist die Beobachtung durch einen zusätzlichen, von den Forscher*innen installierten ‚Mithörer' ein interessanter Untersuchungsgegenstand, auch in seiner Ähnlichkeit, aber auch Andersartigkeit zu Smart Speakern. Explizit mit den Charakteristika des CVR in Abgrenzung zum Smart Speaker setzen sich weitere Arbeiten auseinander, die das bereits mehrfach erwähnte Spannungsfeld zwischen Komfort und Beobachtung, in dem sich Smart Speaker bewegen, weiter beleuchten (Merkle/Hector 2025; Hector 2022) und gleichzeitig für die mitlaufende Erhebungstechnologie und den potenziellen Einfluss auf die in der vorliegenden Arbeit verwendeten Daten sensibilisieren. Drittens lassen sich für diese Arbeit sowohl ethnografische Erkenntnisse wie auch Einsichten aus den Interviews nutzbar machen, die mit den gleichen Haushalten geführt worden sind.

157 Siehe die Diskussion zum Ende von Kap. 4.3.

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

## 5.4 Datenaufbereitung

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

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

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

Teilweise erfolgten die Transkriptionen im Rahmen einer Kooperation mit dem Forschungs- und Lehr-Korpus Gesprochenes Deutsch (FOLK) am Leibniz-Institut für Deutsche Sprache, wo die Ersteinrichtungen aus vier von acht teilnehmenden Haushalten[160] archiviert werden konnten.[161] Diese können nach vorheriger Registrierung über die Datenbank für Gesprochenes Deutsch (DGD) im virtuellen Korpus „Sprachassistenten-Einrichtung“ zur wissenschaftlichen Nutzung abgerufen und einschließlich des Transkripts angesehen werden.[162] In Tab. 1 findet sich außerdem die Sprechereignis-ID für die im FOLK hinterlegten Ersteinrichtungen. Werden in den Analysen Ausschnitte aus diesen Daten besprochen, so wird stets ein Direktlink zum entsprechenden Ausschnitt angegeben.

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

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

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

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

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

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

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

## 5.5 Untersuchte Haushalte und Auswertung

Insgesamt wurden in acht Haushalten Daten aus den zuvor beschriebenen Situationstypen erhoben. Dabei sind insgesamt sechs Videos von Ersteinrichtungen sowie insgesamt 226 CVR-Aufzeichnungen entstanden. Die Ersteinrichtungen sind im Durchschnitt rund 19 Minuten lang und umfassen eine Gesamtdauer von einer Stunde und rund 53 Minuten. Das CVR-Audiomaterial umfasst insgesamt 30 Stunden und rund 58 Minuten. Eine Übersicht über die erhobenen Daten in allen Haushalten liefert Tab. 1.[164]

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

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

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

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

| Pseudonym | Nr. | Dauer Ersteinrichtung (hh:mm:ss) | Anzahl CVR-Aufnahmen Phase I | Dauer CVR-Aufnahmen Phase I (hh:mm:ss) | Anzahl CVR-Aufnahmen Phase II | Dauer CVR-Aufnahmen Phase II (hh:mm:ss) | Gerät | FOLK Sprechereignis-ID |
|---|---|---|---|---|---|---|---|---|
| Faßbender | 01 | 00:24:21 | 11 | 01:11:31 | 8 | 01:11:01 | Amazon Echo Dot, 3. Gen. | FOLK_E_00465_SE_01 |
| Soellner | 02 | 00:19:04 | – | – | 3 | 00:29:51 | Amazon Echo Dot, 4. Gen. | FOLK_E_00466_SE_01 |
| Würz | 03 | 00:23:07 | 20 | 04:20:03 | 2 | 00:04:41 | Google Home Mini | |
| Ruhlange | 04 | – | – | – | 21 | 02:17:13 | Amazon Echo Dot, 3. Gen. | |
| Riker | 05 | – | – | – | 52 | 08:14:03 | Apple HomePod | |
| Reschke | 06 | 00:03:58 | – | – | 6 | 00:36:24 | Apple HomePod | |
| Waldes | 07 | 00:21:27 | 52 | 06:44:50 | 1 | 00:06:03 | Apple HomePod Mini | FOLK_E_00467_SE_01 |
| Matthäi | 08 | 00:21:44 | 9 | 01:02:46 | 41 | 04:40:01 | Google Home Nest | FOLK_E_00484_SE_01 |
| **SUMME** | **Σ** | **01:53:39** | **92** | **13:19:10** | **134** | **17:39:17** | | |

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

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

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

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

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

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

**Tab. 2:** Übersicht Fokushaushalte und Erhebungszeiträume.

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

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

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

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

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

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

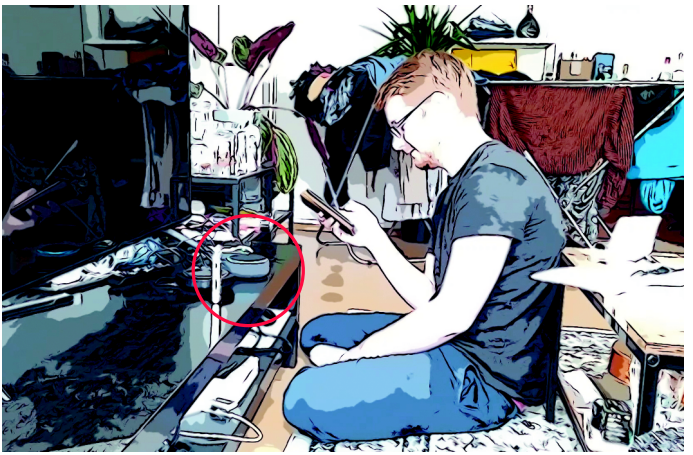

**Abb. 17:** Lukas (hier im Bild) und Alex (nicht im Bild) richten den Smart Speaker ein (Perspektive 1).

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

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

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

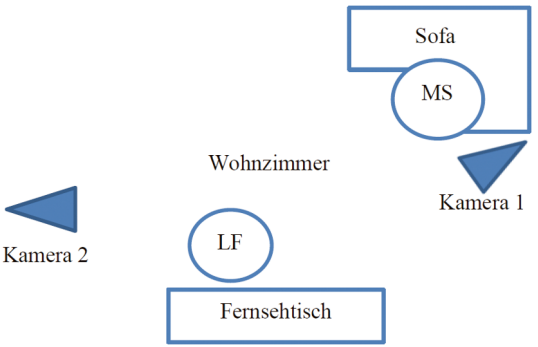


**Abb. 19:** Positionierung von Teilnehmenden und Kameras im Raum während der Ersteinrichtung in Haushalt 1, Darstellung von oben; Bild: Leibniz-Institut für Deutsche Sprache Mannheim (mit frdl. Genehmigung).

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

**168** Siehe Kap. 6.2.

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

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

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

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

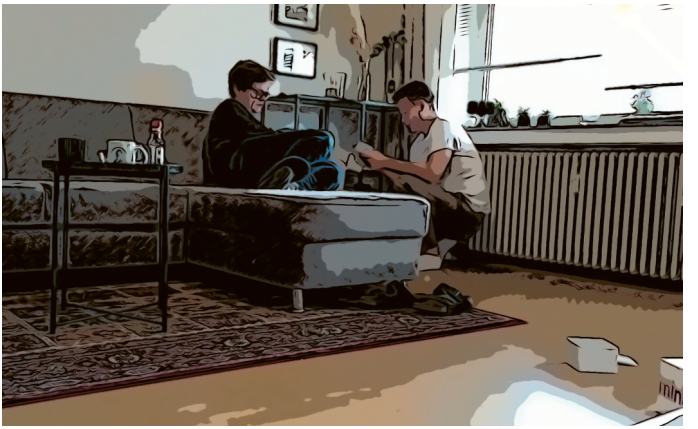

**Abb. 21:** Konrad (rechts) und Till (links) richten den Smart Speaker ein (Perspektive 2).

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

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

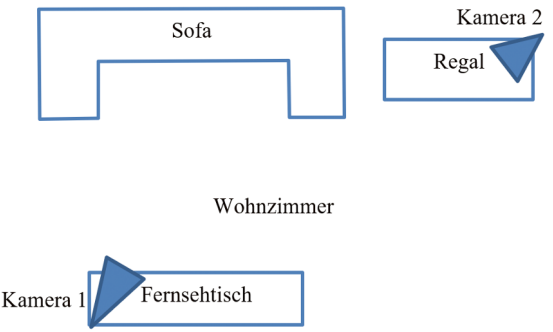


**Abb. 22:** Positionierung der Kameras im Raum während der Ersteinrichtung in Haushalt 7, Darstellung von oben; Bild: Leibniz-Institut für Deutsche Sprache, Mannheim (mit frdl. Genehmigung).

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

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

Die Wohngemeinschaft von Samuel Matthäi (SM), Lara Schiffer (LS) und Robin Lorentz (RL) aus Dortmund nutzt den Smart Speaker Google Nest. Samuel ist 30 Jahre alt und studiert mit dem Abschlussziel Lehramt; er arbeitet seit mehreren Jahren als Vertretungslehrkraft im Inklusionsbereich. Robin ist 23 Jahre alt, er studiert im Master ebenfalls mit dem Abschlussziel Lehramt. Lara ist 24 Jahre alt und studiert im Master. Sie beschreiben das Verhältnis zueinander als vertrauensvoll und freundschaftlich. Samuel und Robin haben regelmäßig ihre Lebenspartnerinnen zu Besuch, die teilweise auf den ausgewerteten Aufnahmen zu hören sind; die Lebenspartnerin von Robin, Alexandra Wormsberge (AW), ist auch während der Ersteinrichtung anwesend, aber kaum von der Aufnahme erfasst; sie beteiligt sich auch nicht an der Ersteinrichtung des Google Nest, die am gemeinsamen Küchentisch sitzend durchgeführt wird (siehe Abb. 25).
Die Einrichtung findet vollständig in der abgebildeten Konstellation statt (siehe Abb. 23–24), es wird währenddessen Bier getrunken und die Stimmung ist ausgelassen. Wie auch auf den Aufnahmen zu hören ist, die während der Ersteinrichtung gemacht werden, ist auch hier das Sozialleben noch von den pandemiebedingten Kontaktbeschränkungen im Winter 2021/22 geprägt, sodass – so geben die Haushaltsmitglieder an – selten andere Besucher*innen empfangen werden. Der Smart Speaker wird in der Küche platziert und steht nach Angaben von Samuel lange Zeit auf dem Küchentisch, wo er auch eingerichtet wurde, obwohl diese Positionierung

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

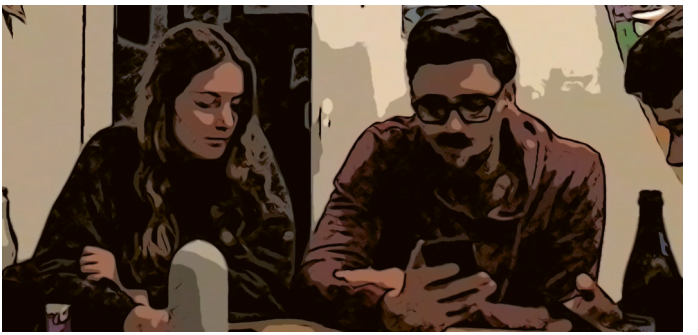

**Abb. 24:** Lara (links), Samuel (Mitte) und Robin (rechts, halb im Bild) richten den Smart Speaker ein (Perspektive 2).

nicht optimal sei, weil er beim Essen oder anderen am Küchentisch zu verrichtenden Aktivitäten gelegentlich störe. Es sei aber auch kein besserer Ort gefunden worden.
Die Beteiligung der drei Bewohner*innen an der Ersteinrichtung ist ausgeglichen, wobei Robin sich als primärer Anwender positioniert, indem er sein Smartphone als erstes mit dem Smart Speaker verknüpft. Samuels Smartphone kann ebenfalls kurz darauf verknüpft werden. Lara ist hingegen bei den im weiteren Verlauf der Aufnahme vollzogenen Tests des Smart Speakers noch abgelenkt, weil es sich als Herausforderung erweist, ihr Smartphone mit dem Smart Speaker zu verbinden (die Gründe dafür bleiben unklar).

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

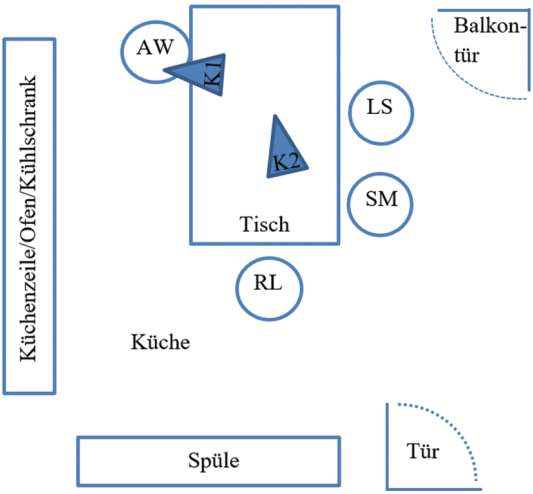


**Abb. 25:** Positionierung von Teilnehmenden und Kameras im Raum während der Ersteinrichtung in Haushalt 8, Darstellung von oben; Bild: Leibniz-Institut für Deutsche Sprache Mannheim (mit frdl. Genehmigung).

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

## 5.6 Transkriptionskonventionen

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

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

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

Sequenzielle Struktur/Verlaufsstruktur

| | |
|---|---|
| `[ ]` | Überlappungen und Simultansprechen |
| `[ ]` | |

Ein- und Ausatmen

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

Pausen

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

Sonstige segmentale Konventionen

| | |
|---|---|
| `und_äh` | Verschleifungen innerhalb von Einheiten |
| `äh öh äm` | Verzögerungssignale, sog. „gefüllte Pausen" |
| `:` | Dehnung, Längung, ca. 0.2–0.5 Sek. |
| `::` | Dehnung, Längung, ca. 0.5–0.8 Sek. |
| `:::` | Dehnung, Längung, ca. 0.8–1.0 Sek. |

Lachen und Weinen

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

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

Rezeptionssignale

| | |
|---|---|
| hm ja nein | einsilbige Signale |
| hm_hm ja_a | zweisilbige Signale |
| ʔhmʔhm | mit Glottalverschlüssen, meistens verneinend |

Sonstige Konventionen

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

Akzentuierung

| | |
|---|---|
| akZENT | Fokusakzent |
| ak!ZENT! | extra starker Akzent |

Tonhöhenbewegung am Ende von Intonationsphrasen

| | |
|---|---|
| ? | hoch steigend |
| , | mittel steigend |
| – | gleichbleibend |
| ; | mittel fallend |
| . | tief fallend |

Multimodale Transkriptionen von Gestik und Blick

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

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

Kleinbuchstaben versehen. Dabei werden im Einzelnen folgende Symbole verwendet:

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

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

**171** Siehe dazu Kap. 3.3.