First users' interactions with voice-controlled virtual assistants: A micro-longitudinal corpus study

Mathias Barthel
Institute for the German Language
Mannheim, Germany
barthel@ids-mannheim.de

Henrike Helmer
Institute for the German Language
Mannheim, Germany
helmer@ids-mannheim.de

Silke Reineke
Institute for the German Language
Mannheim, Germany
reineke@ids-mannheim.de

Abstract

We present a collection of (currently) about 5.500 commands directed to voice-controlled virtual assistants (VAs) by sixteen initial users of a VA system in their homes. The collection comprises recordings captured by the VA itself and with a conditional voice recorder (CVR) selectively capturing recordings including the VA-directed commands plus some surrounding context. Next to a description of the collection, we present initial findings on the patterns of use of the VA systems during the first weeks after installation, including usage timing, the development of usage frequency, distributions of sentence structures across commands, and (the development of) command success rates. We discuss the advantages and disadvantages of the applied collection-specific recording approach and describe potential research questions that can be investigated in the future, based on the collection, as well as the merit of combining quantitative corpus linguistic approaches with qualitative in-depth analyses of single cases.

1 Introduction

Human-computer interaction becomes increasingly more prevalent in industrialised societies. More recently, especially interactions with in-home intelligent virtual assistants (VAs) quickly grows in popularity and amount of use. While research on human interaction with technology moves more into the focus of the language sciences lately, it was established early with Suchman's (1987) seminal work on situated practices in the usage of "intelligent" machines (at that time a printer). Since then, ethnomethodological and conversation analytic (CA) research has addressed a variety of phenomena regarding the interaction between humans and AI-based technologies (for a comprehensive overview of studies see Mlynář et al. (in prep.)). CA-related studies (and studies interested in conversation analytic concepts), especially when investigating interaction with verbally controllable technology (voice-based virtual assistants, robots, chatbots etc.), have examined the organization of talk, like openings and closings (e.g., Pitsch et al. (2009)) and turn-taking in dyadic and multi-party interaction (Skantze, 2021), as well as on miscommunication and repair sequences (e.g., Krummheuer (2008); Pelikan and Broth (2016)).

Recent studies on interaction with VAs like *Amazon*'s *Echo Dot* or *Alphabet*'s *Google Home* have shown how VA systems are designed to help users diagnose and repair trouble (e.g. by rephrasing requests or asking clarifying questions (see Porcheron et al. (2018); Reeves et al. (2018)). Previous research also touched upon the question of how VAs are embedded in multiple ongoing activities in private settings (Porcheron et al., 2018), how reactions of VAs have effects on the progressivity in interaction (Fischer et al., 2019), how the integration of systems into everyday practices is connected to agency (Habscheid et al., 2023), and how a machine's 'participation' can be seen as situational and regulatory participation which becomes part of meaningful talk-in-interaction (Reeves and Porcheron, 2022).

While VAs are claimed to be designed to more and more resemble human interlocutors in their verbal behaviour, they still fall short of human-like interactional capacities in many tasks and on many occasions. Users however do not apply social rules 'mindlessly' onto VAs (Reeves and Porcheron, 2022). They adapt their talk in order to improve interaction with a VA (Pelikan and Broth, 2016), e.g. by altering prosody or rephrasing instructions (Porcheron et al., 2018), and they learn how to formulate probably successful commands (Reeves et al., 2018). Learning to efficiently deal with these weaknesses thus becomes a task of human users.

First users of a VA system hence need to learn the peculiarities of the system to be able to achieve successful goal-oriented interaction with the VA.

Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue, August, 16–17, 2023, Maribor.

Studying such adaptations to systems, CA-related research has hitherto mainly used single-case analyses only, typically focusing on specific moments of trouble. To systematically analyse and understand in what ways users adapt their use of VAs to the capabilities and limitations of the system, how they learn which strategies turn out to be successful, and which are the overarching longer-term patterns of use, we need to collect data of human-VA interaction over time and analyse them from a micro-longitudinal perspective. In our project, we aim at addressing this desideratum and adopt a mixed-methods approach that combines conversation analysis and interactional linguistics with quantitative analysis. Our overarching goal is a micro-longitudinal analysis of first users' adaption in interacting with the VA. The focus of this paper is on a quantitative overview of developments over time with regard to the timing dynamics of commands, their linguistic structures, and their success-rate.

The methods of data collection are described in section 2. Section 3 presents a description of the resulting collection of audio recordings, as well as a number of first exemplary findings. Finally, section 4 will offer a discussion of the achievements and downsides of the presented methods of data collection and processing, and give an outlook on future use cases for the collected data and the kinds of questions that can be investigated on their basis.

2 Methods

2.1 Participants

To be able to draw a picture of how humans use a VA in a natural setting, and in line with conversation analytic methodology, we recorded naturally occurring interactions of human users with a VA, focusing on recording human-VA interactions with ecological validity. We recorded first users' authentic interactions with VAs in their private living environments during their first weeks of using the VA. We searched for participants who had an a priori interest to get a VA system for their homes and asked them if they would be willing to participate in our study over a period of several weeks. We only included users who had no significant prior experience in using a VA system, so they are all novices in the field of VA communication. Participants got a small monetary compensation for their participation in the study and could keep the VA system after the end of the recording period. We obtained all participants' advanced informed written consent to use the recordings and VA log-files they provided for the purpose of the project. To date, we recorded six single participants or participating families with two to four members (mean age = 20 years, min = 3 years, max = 37 years) over a period of seven to ten weeks (mean = 66 days, min = 49 days, max = 72 days), starting from the first day of their usage of the VA.[1] This way, we were able to track potential changes over time in participants' usage behaviour and formulations of commands during the initial phase of interacting with their newly installed VA system.

2.2 Recording Methods

For data collection, a new VA system (*Amazon Alexa EchoDot*) was installed together with the participants in their home, either in the kitchen or in the living room. Additionally, a conditional voice recorder (CVR)[2] was placed in close proximity to the VA speaker for the recording period. The CVR is a device developed and previously used by Martin Porcheron (see Porcheron et al. (2018)) that captures audio snippets containing commands to the VA. The CVR-software uses a speech detection model[3] and is installed on a Raspberry Pi supplied with a conference microphone. We replicated the CVR and adapted it for our purposes.[4] Our version of the CVR continuously recorded 90-second stretches of audio, constantly overwriting these 90 seconds in a loop. Upon detecting the wake-word (*"Alexa"*), the CVR would save the last 90 seconds of recording and attach the following 90 seconds of recording to the file, creating 3-minute long audio snippets around each user command to the VA. This way, we were able to record the context in which users addressed the VA, the commands to the VA themselves, as well as the reactions by the VA plus potential follow-up context.[5] Whenever

[1] One additional household was excluded from data analysis in this study due do data scarcity, as the participants made use of the VA only in 11 days, producing only 81 commands.

[2] https://github.com/MixedRealityLab/conditional-voice-recorder

[3] The respective models were obtained by the Snowboy Hotword Detection Engine: https://github.com/Kitt-AI/snowboy

[4] Main changes were: We extended the recording time of the audio snippets from 120s to 180s, we wrote timestamps into the recording file names, we changed the LED-setup due to a mutable microphone and we added an RTC module that guarantees a power supply for the integrated system clock, so that we could disable wifi and bluetooth connections for privacy reasons.

[5] We only analyse the stretches of context that are relevant to the interaction with the VA.

they chose, participants could switch off the microphone attached to the CVR.

In addition to the CVR data, we also collected the audio recordings captured and stored by the VA system itself. The VA system saved audio recordings containing only the user commands, starting with the wake-word *"Alexa"*. Thus, these recordings are generally only a few seconds in length. On top of these VA audio recordings, the VA system kept a log of all user commands in a csv-file. These log files contain a transcription of each user command, generated by the VA's speech detection algorithm.[6]

Since both types of recordings have their advantages and drawbacks, both types of recordings were important for our purposes in order to achieve a collection of commands (and relevant context) that was as exhaustive as possible: CVR-recordings are based on a less well trained speech detection model than the one available to the VA. Hence, the CVR is prone to detection failures, occasionally missing to record actual user commands (i.e. false-negatives) (see section 3). Additionally, the CVR sometimes saves files based on false-positive detections of the wake-word. Due to the inferior speech detection model, false-positives and false-negatives are more common in the CVR-recordings than in the VA recordings. On the other hand, commands that did not trigger a verbal reaction by the VA are sometimes omitted from the list of VA recordings (and the respective csv-logfiles)[7]. Similarly, false-negatives also occur on the side of the VA, leading to no reaction in response to the wakeword. These false-negatives in turn can regularly be found within the CVR-recordings. Moreover, in comparison to the VA system, CVR-recordings contain context information leading up to the user command, the audio of the VA reaction to the command, and follow-up context including user reactions in third position following the VA reactions. Thus, CVR recordings are best suited for all studies that need to take into account the preceding context as well as the VAs response. As a complementary completion of the commands not recorded by the CVR, the VA recordings are however important for micro-longitudinal studies (e.g., on success- and failure-rates) that need to rely on a dataset as exhaustive as possible.

2.3 Data Pre-Processing

For the collection of human-VA interactions, the obtained recordings went through a number of pre-processing steps. After obtaining the CVR-recordings and the VA-recordings plus the VAs' log lists of commands that were issued by the participants during the recording period, we cleared the list of CVR-recordings from false positives by automatically matching the time stamps of the recordings with the time stamps of the logged commands in the VAs' log lists: Only CVR-recordings that contained at least one time stamp of a logged command were kept for further processing and inspection. As a next step, we manually checked and transcribed the remaining CVR-recordings that contained at least one logged command.[8] During this checking and transcription process, any additional commands that were contained in the CVR-recordings but not logged in the original list by the VA were also transcribed and added to the log list of issued commands. In a following step, all recorded commands were manually annotated for a number of factors, including whether the kind of command has been used before by the same user (form-based); a coarse category of what action was requested of the VA; what sentence type has been used for the command; what intonation contours have been used in the wake word and in the command proper; whether the command was successful in terms of the VA-output fitting to the command; and whether the output was followed by any additional comments on the side of the participant in third position. Moreover, we coded whether the original transcription by the speech detection algorithm of the VA was erroneous. In these cases, we corrected the transcription in question and kept a record of the original transcription of the command.[9]

[6] Copies of the VA log files and VA audio recordings were sent to us by the participants after the end of the respective recording periods. Before sending these data to us, participants had the chance to read the log file and listen to the recordings and decide to delete entries and recordings that they did not wish to share without any disadvantages or other consequences.

[7] Typically omitted commands include setting the volume or stopping a running playback of music. While these are omitted from VA recordings, they would still be present in the CVR-recordings.

[8] Iteratively developing and exhaustively implementing a coding scheme is a time-consuming process (Mundwiler et al., 2019; Stivers, 2015). At the point of submission, this checking and transcription has been completed for one participating family. See section 3.2 for more details.

[9] The number of VA speech detection errors varied between participants, see section 3.1 for details.

3 Results

As described in section 2, our collection consists of two kinds of recordings: short audio files captured by the VA, containing just the user commands, and 3-minute recordings captured by the CVR, containing the preceding context of a command, the command itself plus the VA's reaction, as well as follow-up context after the exchange. We will first present the results of our analyses of the obtained VA-recordings in section 3.1, followed by a presentation of the results of initial analyses of a subset of the obtained CVR-recordings in section 3.2.

3.1 Analyses of obtained VA-recordings

For our analysis, we included audio-recorded human-VA interactions in six households for the first 49 to 72 days after the VA had been installed by the users. In total, we obtained 5502 commands that were recorded and logged by the VAs. On average, commands were 4.23 words long, including the wake word (SD = 2.49).[10] The intensity of usage and thus the number of commands recorded and logged by the VA varied considerably between participants (see Table 1): While participating household 5, for instance, only issued 165 commands that were logged by the VA, making use of the VA in 44 out of the recorded 72 days (61%), household 6 produced 2186 logged commands, using the VA in 55 out of the 67 days in the recording period (82%). Listening to all VA-recorded commands and comparing them to the VA-logged transcriptions, we found that of all commands logged by the VA system, the transcript of the command was erroneous in 8%. Proportions of mis-detections of speech input were found to vary across participating households: 1: 10.1%, 2: 9.1%, 3: 9.5%, 4: 3.0%, 5: 6.8%, 6: 9.9%.

Human-VA interactions were found to commonly happen in clusters of commands[11]. This means that, across all logged commands, the probability of a command being issued is highest right after a previous command and drops considerably after a few seconds, with 25% of commands being issued within the first 10 seconds after a previous command, 50% of commands being issued within 22 seconds, and 75% being issued within 182 seconds.This general pattern holds across all recorded households (Figure 1).

Household (*N* members)	*N* days of recording	*N* days of use	*N* logged commands
1 (2)	49	29	313
2 (4)	70	44	1033
3 (1)	69	40	429
4 (3)	68	61	1377
5 (2)	72	44	162
6 (4)	67	55	2186
total:	**395**	**273**	**5502**

Table 1: Recording details by participating households. *N* members specifies the number of regular users of the VA, *N* days of recording specifies the length of the recording periods, *N* days of use specifies the number of days containing commands to the VA, *N* logged commands specifies the number of commands logged by the VA during the recording period.

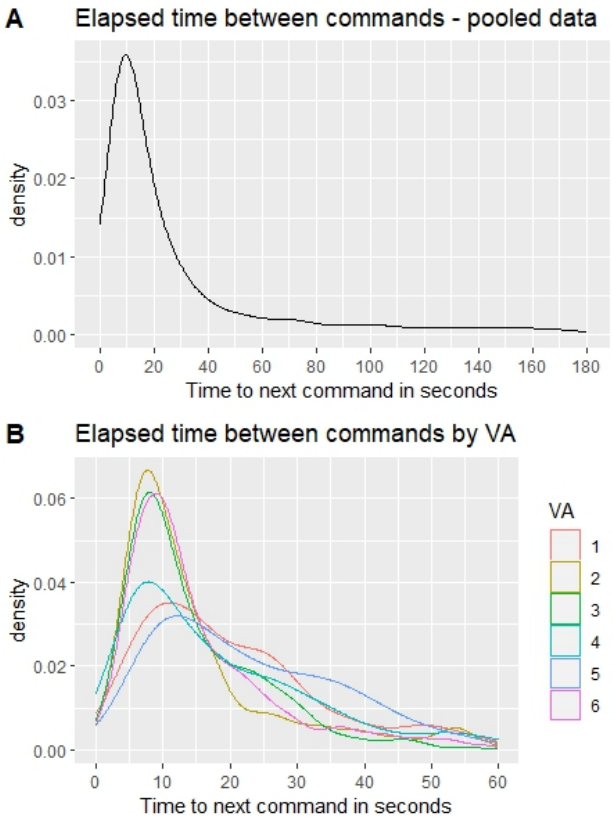

Figure 1: Density plots illustrating probabilities over time for a next command after a previous command. Top panel A shows data pooled by all users. Bottom panel B shows data by VA. In all VA users, the probability for a new command peaks between 7 and 13 seconds after the previous command. *N* = 5502.

[10]The average lengths of commands differed only slightly between households, with the smallest household mean being 3.95 words and the largest being 4.74. SDs for the different households were all between 2.39 and 2.65. VA responses varied much more in length, with a grand mean response length of 12.15 words (SD = 11.63).

[11]With 'commands', we mean the VA being addressed in an utterance by the user starting with the wakeword, mostly containing a request to the VA.

Another observation that holds across all recorded households is that the frequency of commands declines during the recording period. To quantify this observation, we built a general linear mixed effects regression model using the *R* package *lme4* (Bates et al., 2015; R Core Team, 2023), using a Poisson distribution to model the number of commands by the consecutive days of use during the recording period, with random intercepts and random slopes for day of use by household. The model output showed a significant effect of day of use ($\beta = -0.039, SE = 0.007, z = -5.453, p < .001$) and an intercept estimate of 3.438. Note that the link function is logarithmic, meaning that the modeled grand-average number of commands per day at the beginning of the recording period is 31 commands, with the number of commands decreasing by factor 1.04 on each consecutive day (Figure 2). While this factor (as well as the intercept) varied between households, it was found to be smaller than 1 for all households, meaning that the number of commands per day tended to decrease during the recording period for all households ($\beta_1 = -0.058, \beta_2 = -0.055, \beta_3 = -0.049, \beta_4 = -0.010, \beta_5 = -0.033, \beta_6 = -0.031$).

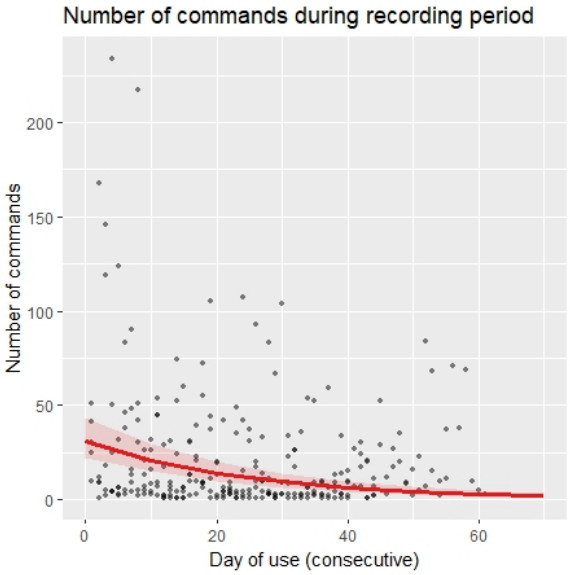

Figure 2: Development of frequencies of commands for consecutive days of use across all households. Dots represent the number of issued commands for each day of usage by any one household. Days without any commands are ignored. The red line represents the fit of a general linear mixed effects regression model (formula = $N_{commands} \sim dayOfUse + (1 + dayOfUse \,|\, household), family = poisson(link = log)$, see main text for details). The red ribbon represents 68% confidence intervals.

3.2 Initial analyses of a subset of CVR-recordings

As described in section 2.2, our collection consists of two kinds of audio recordings: short recordings of the commands made by the VA and 3-minute recordings containing the commands made by the additionally installed CVR. Analyses of the CVR-recordings are time consuming and still ongoing. Nevertheless, we exhaustively listened to all CVR-recordings of one of the participating households (household 1) that remained after excluding false positives as described in section 2.3. In this section, we present the analyses of the subset of the collection containing data of this example household, serving as a test case for the obtained recordings.

In addition to the 313 commands logged and recorded by this household's VA, we identified another 155 commands in the CVR-recordings that were not originally logged or recorded by the VA, and added these to the list of identified commands, leading to a total number of 468 identified commands. Note that while the VA did not log about a third of the issued commands, this does not mean that the VA was generally unresponsive to these commands. While the VA did indeed not respond to 50 of the total of the 468 identified commands (10.7%), the remaining commands triggered a response in the VA. Most of the originally unlogged commands were either commands to stop the ongoing output of the VA, or to adjust the output volume. These kinds of commands did not trigger a verbal response by the VA, but were generally complied to by directly stopping the current output or adjusting the output volume accordingly. On the flipside, 86 commands (18.4%) that were logged and recorded by the VA were not recorded by the CVR, in most cases probably because the wake word had not been detected, leading to false negatives.

We were interested in the distribution of success rates over different types of sentences (Figure 3). We thus coded all commands regarding their sentence type based on their syntactic structure. Of the 468 identified commands, 15 have a declarative sentence structure (3.2%, e.g., *"Alexa, ah das ist zu schwer"* ("Alexa, ah that's too difficult")), 135 have an imperative sentence structure (28.8%, e.g., *"Alexa, spiel mein Hörbuch"* ("Alexa, play my audiobook")), 105 have an interrogative sentence structure (22.4%, e.g., *"Alexa, wie wird das Wetter heute"* ("Alexa, how is the weather today")), 200 have an elliptical sentence structure (42.7%, e.g.

"Alexa, lauter" ("Alexa, louder")), and a single case has a deontic infinitive structure (*"Alexa, Werbung überspringen"* ("Alexa, skip ads")).[12] Another 12 commands have been aborted and not completely uttered (2.6%), mostly consisting of the wake word only.

We annotated all 468 commands regarding their outcome success. If the triggered VA response or output was relevant to (the surface structure of) the uttered command, the command was coded as 'success'. If, on the other hand, the VA response or output did not fit the command, it was coded as 'failure'. 298 commands successfully triggered relevant VA reactions, while 149 commands did not trigger the requested response or output and were thus coded as failures (Figure 3).[13] The proportion of failures was found to decline with increasing numbers of commands in a given sentence structure: In the most frequent category, imperative commands, only 26.5% of commands failed; in elliptical commands without a verb form, 29.6% failed; in commands with interrogative sentence structure, 40.7% failed; in the greatly rarer commands with declarative structure, 90.9% failed; and the single case with an infinitive verb form also failed. In a generalized linear model built with the *R* package *lme4*, the number of commands observed per sentence structure as a linear and a quadratic predictor for command success both turn out to be significant ($\beta_{linear} = 0.048, SE = 0.018, z = 2.587, p < .01$; $\beta_{quadratic} = -0.014, SE = 0.001, z = -1.956, p = .050$). This means that the higher the number of total commands used with a given sentence structure, the higher the proportion of successful commands in that structure (Figure 4).[14]

Interestingly, failing commands were found to be produced in clusters, with the probability of a command to fail being greatest right after a failing command, with no or not more than one successful command in between (Figure 5).

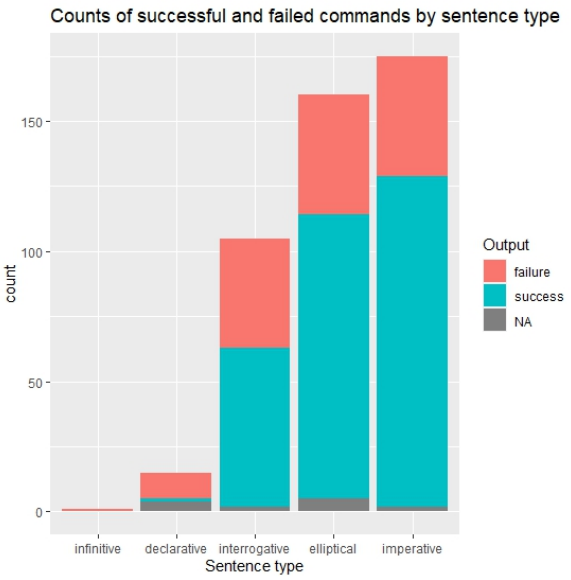

Figure 3: Number of successful and failed commands by sentence type in analysed subset. $N = 454$.

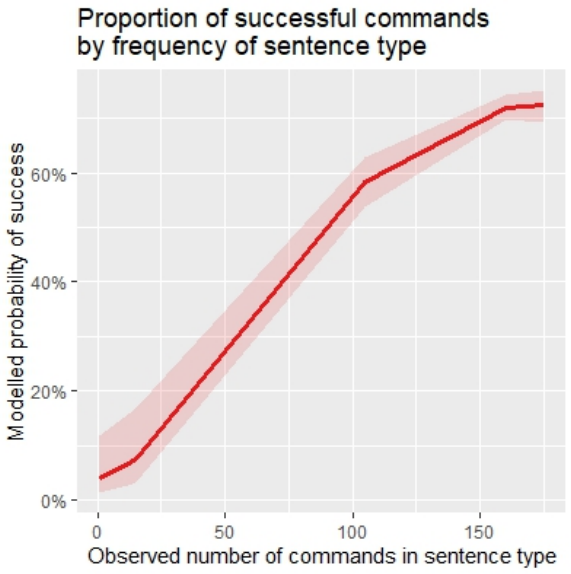

Figure 4: Modeled probability of success of a command as predicted by the observed frequency of the command's sentence structure. More frequent sentence types show higher proportions of successful commands (see main text for details). Formula = $success \sim N_{structure} + I(N_{structure}{}^2), family = binomial(link = logit)$.

In order to test whether users successfully adapt their commands to the VA system over time, adjusting the input so as to increase the success rate, we analysed how the frequency of failing commands changes over time as the users get more experienced with the VA system. While failed com-

[12] "The vast majority of commands are single sentences. Rare instances of multi-sentence commands are generally not successful, apparently mainly because the VA does not log more than the first main sentence of the command. Example (translated): *User:* alexa play macklemore and like this <<singing> this is the moment>. *VA:* this is macklemore on spotify. (plays some other song by same artist instead of specific song). In this example the VA did not log the complete command and responded only to the first part of the command (logged command: *alexa play macklemore and like this*).

[13] 7 aborted commands and 14 uncategorisable commands were not coded.

[14] Note that the relatively frequent stopping-commands ("*Alexa stop*") were coded as 'imperative' here. If they were coded as 'elliptical', the general pattern of this result would not change.

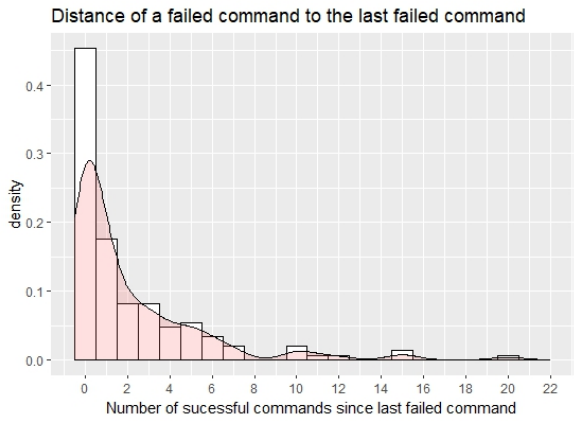

Figure 5: Frequency distribution of the number of successful commands between two failing commands. As the most frequent case, a failing command follows directly after a previous failing command ($N = 67$), with the second most frequent case being a single intermittent successful command ($N = 26$). $N_{total} = 149$.

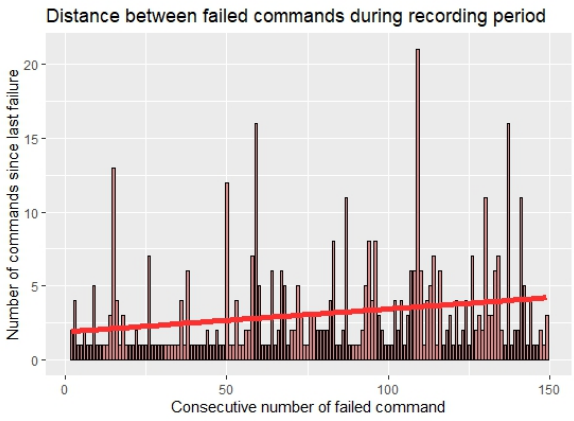

Figure 6: Development over time of the number of commands since the last failing command in analysed subset. For each failed command on the x-axis, bars show how many commands ago the last failure was located in the usage history. If the number of commands since last failure is shown to equal 1, this failed command followed directly upon a previous failed command; if the number is shown to be equal to 2, one successful command has been issued after the previous failed command, and so on. The regression line in red shows that the frequency of failing commands significantly decreased during the recording period, with a slope of 0.02 (formula = $distanceToLastFailure \sim position_{failedCommand}$; $p = .013$).

mands were very common initially, with about every second command failing to trigger an intended response or output, the success rate in the analysis subset of the collection approximately doubled during the recording period of 49 days (out of which the VA was used on 29 days). Hence, at the end of the recording period, only about one in four commands failed to elicit a desired response or output, which makes for an average increase in success rate of 0.66% per failed command (Figure 6).

Given the observation of an increasing success rate over time in combination with the distribution of failing commands across sentence types, we investigated the development of success rates by sentence type in more detail (Figure 7).

Elliptic commands, including standardized commands like setting the volume ('louder', 'softer'), are found to be constantly used over time, showing a high success rate already in the first week of use and even becoming more successful over time. Imperatives, including highly frequent standardized commands like 'stop', are continuously used over time as well. In contrast to ellipticals, however, they don't tend to become more successful over time. Declarative commands, which are mostly failing, are rather rare from the beginning and eventually fade out completely. In qualitative single-case analyses, we can see how declaratives that are not successful in the local sequential context are repaired and, more specifically, replaced by other formats (like interrogatives) for the same use case. E.g. *"Alexa, we'd like to play a game for five year olds"*

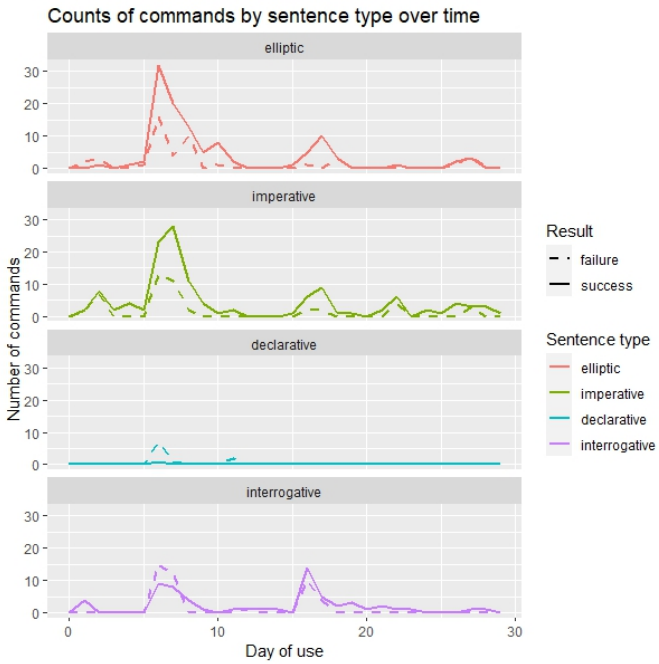

Figure 7: Numbers of failing and successful commands by sentence type during recording period. A single failing infinitive command on day 27 is not plotted here. 12 aborted commands, which are generally failing, are also not plotted here.

(see Transcript 1, lines 01-04 in the Appendix) is locally repaired by the interrogative *"Alexa, can we play a game with you"* (Transcript 1, line 10). The fact that alternative commands are successful (either directly or after several attempts, see below and Transcript 2 in the Appendix for the eventually successful request) and declarative commands fade out in the course of the recording period suggests that declarative sentence types are abandoned in favor of more successful command types. In interrogative commands, we find the success rate to increase over time. Investigating the data with more in-depth analyses shows that this is not the case because interrogatives are used with less trouble in general. Instead, unsuccessful variants of interrogatives are also locally repaired and replaced by types of interrogatives that turn out to be more successful. Transcripts 1 and 2 in the Appendix show that an unsuccessful *can we*-interrogative (*"Alexa can we play a game with you"*, Transcript 1, line 10) is replaced by a successful *wh*-interrogative (*"Alexa what games are there"*, Transcript 2, line 08). Overall, we find that (typically unsuccessful) *can you/we/I*-interrogatives fade out over time in favor of other, more successful, types of interrogatives. We take these first examples as evidence for experience-based, goal-oriented adaptations of users' behaviour in interaction with the VA that lead to a reduction in the proportion of failing commands over time.

4 Discussion and Prospect

In this first description of the new collection of first users' interactions with virtual assistants (VAs), we presented initial observations of patterns of use during the first weeks after installation of the VA. Comprising over 5.000 commands to the VA that were captured in six households with a total of sixteen members, the collection was found to be suitable for micro-longitudinal analyses of the development of patterns of interactions with the VA system. A CVR, selectively recording audio snippets only, has proven to be suitable for field recordings in private settings over a longer period of time. Continuous recording, as well as longitudinal video-recording, would be much more intrusive and less efficient in terms of capturing sequences of focal interest (i.e., sequences featuring interactions with the VA). Moreover, selective recording with a CVR proved to be a practical approach to meet relevant ethical questions, since recordings get limited to stretches of time that are directly relevant to the target research questions of the project. Three-minute stretches of recording have proven to be an apt compromise to grasp sufficient context without covering excess unrelated interaction. This approach also minimised the amount of recorded data, leading to computational efficiency during data curation, inspection, and annotation. Notwithstanding these advantages of this way of audio-only recording, they obviously come with the downside of some situational aspects remaining unanalysable to us: Without video recordings of the relevant sequences, most of the time we are unable to detect with certainty if or when and how users turn to another channel of input (like, e.g., their cellphone), or when they chose to control the VA by pressing a button (e.g. to adjust the output volume). Nonetheless, the applied recording methods strike a worthwhile balance between highly informative, goal-focused data and low invasiveness.[15]

Analysing the data captured by the VA, i.e., audio-recordings and lists of logged transcripts of just the commands (section 3.1), we found a clear and consistent pattern of users to interact with the VA in clusters of commands, meaning that the probability of a command being uttered is highest right after a command has been uttered and quickly declines within a few seconds. Additionally, we found failing commands to cluster as well, with frequently no or not more than one successful command between two failing commands that do not trigger the intended response or output in the VA. These two results seem likely to be related. If a user tries to achieve a certain goal and fails with an initial attempt, any follow-up pursuit of that goal might also fail, due to limitations of the VA. Similarly, a regularly observable pattern of commands that leads to one successful command between two failing commands is a successful stopping command after an initial failing command that triggered unintended output in the VA, followed by a second (possibly again failing) attempt to pursue the initial goal.[16]

While we found failing commands to be initially very frequent in the subset of our collection that

[15] Alongside this project, we collect non-longitudinal video data of VA users to be able to study the use of alternative types of input (like button presses, phone control, etc.) and embodied orientation and conduct.

[16] Example (translated): *VA*: (starts song). *User*: alexa that's the wrong one. *VA*: (beep). *User*: alexa stop. VA: (stops playback). *User*: alexa what other version is there? (no response by VA).

we analysed in more detail, the frequency of failing commands was found to decrease with increasing time of use (see section 3.2). While future analyses will still have to show whether this finding generalises across users, it might be related to the globally observed pattern that the number of uttered commands generally declines during the recorded first weeks of using the VA. The combination of these two findings offers at least two non-exclusive explanations: First, the number of commands might decline over time because users get to know and memorise the limitations in the VA's use cases and consequently try to use the system for less goals, hence uttering less commands. And second, with more experience, users learn to need less commands to achieve their interactional goals. This last scenario of increasing user efficiency might at least partly be caused by users reducing the variation of commands as they get used to the VA system, honing in on more standardised formats that become known to work. Speaking in favour of this possibility, we found that, at least in the subset of the collection that was analysed in more detail, the number of commands using sentence structures that lead to more failures (i.e., declaratives and interrogatives) is lower across the recording period than the number of sentence structures that lead to less failures (i.e., elliptical commands and interrogatives). Given the observation that the frequency of failures tends to decline over time, this structure-frequency effect might well be the result of a learning process that intensifies over time: As users repeatedly fail to achieve their intended goals with commands in a particular sentence structure, they might use the structure less frequently and learn to use other, more successful structures instead. First qualitative analyses of commands support this hypothesis: Local failures (like with declaratives) are found to lead to a local variation of formats in order to repair the trouble (e.g. substituting an unsuccessful declarative with an interrogative format). As a consequence, this can lead to a consistent usage of successful formats and strategies over time. We aim to validate this conceivable pattern, analysing a greater number of users in the course of the current project.

The learning effects contributing to the development and changes of usage patterns, including adaptions to characteristics of the VA system, are a central aspect of our intended future investigations that can be run on the presented collection of human-VA interactions. While we expect meaningful insights to be based on further quantitative analyses on a larger data basis of CVR-recordings, we also plan to adopt more in-depth qualitative analysis regarding the occasions and reasons for specific quantitative results (de Ruiter and Albert, 2017). For instance, we intend to identify possible 'crucial' moments, e.g. at the end of repair sequences, after which users learn how to successfully formulate a specific request, adapting their usage behaviour. Similarly, we plan to analyse in more detail which types of commands do more typically work and why, taking into account both the characteristics of the commands as well as the inherent limitations of the VA system that cannot be mitigated by adaptions in users' behaviour (Pelikan and Broth, 2016; Reeves et al., 2018). Moreover, as the CVR-recordings cover the context around commands, we will be able to investigate the sequential structure of user-VA interactions more thoroughly, analysing user comments in third-position after the VA's response to a command, as well as potential explicit ascriptions (of actions, intentions, etc.) to the VA, both addressed to co-present users and to the VA itself (see also Habscheid et al. (2023)). These investigations of longer sequences will also enable us to conduct a more in-depth analysis of repair sequences (see also Krummheuer (2010)) and their outcomes, as well as their development over time of usage. The results of these future analyses should shed light on the question on what levels users adapt to the VA, and how human-VA interactions change with an accumulating history of interacting with the respective VA.

On the basis of the presented new collection, we expect to generate fruitful insights into the dynamics of human-VA interaction. Due to its size and focus, the collection lends itself to mixed-methods approaches, with intended future investigations likely profiting from mutually informing insights from quantitative and qualitative analyses. While the former offer powerful tools to discover global usage patterns, the latter, especially conversation analytic qualitative single case analyses, offer apt methods to identify fine-grained aspects of sequential patterns and unveil additional information about occasions, reasons and routinization of users' behaviour and the practices they develop over time. A combination of both approaches will be necessary to draw an encompassing picture of change in practices of VA users over time.

References

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4.

de Ruiter, J. P. and Albert, S. (2017). An Appeal for a Methodological Fusion of Conversation Analysis and Experimental Psychology. *Research on Language and Social Interaction*, 50(1):90–107.

Fischer, J. E., Reeves, S., Porcheron, M., and Sikveland, R. O. (2019). Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–8, Dublin Ireland. ACM.

Habscheid, S., Hector, T., and Hrncal, C. (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. *Social Interaction. Video-Based Studies of Human Sociality*, 6(1):1–31.

Krummheuer, A. L. (2008). Zwischen den Welten: Verstehenssicherung und Problembehandlung in künstlichen Interaktionen von menschlichen Akteuren und personifizierten virtuellen Agenten. In Willems, H., editor, *Weltweite Welten*, pages 269–294. VS Verlag für Sozialwissenschaften, Wiesbaden.

Krummheuer, A. L. (2010). *Interaktion mit virtuellen Agenten? Zur Aneignung eines ungewohnten Artefakts.* Lucius, Stuttgart.

Mlynář, J., de Rijk, L., Liesenfeld, A., Albert, S., Stommel, W., and Topinková, R. (in prep.). Interacting with AI: A scoping review of empirical ethnomethodological and conversation analytic studies.

Mundwiler, V., Kreuz, J., Müller-Feldmeth, D., Luginbühl, M., and Hauser, S. (2019). Quantitative und qualitative Zugänge in der Gesprächsforschung: Methodologische Betrachtungen am Beispiel einer Studie zu argumentativen Gruppendiskussionen. *Gesprächsforschung*, 20:323–383.

Pelikan, H. R. and Broth, M. (2016). Why That Nao?: How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4921–4932, San Jose California USA. ACM.

Pitsch, K., Kuzuoka, H., Suzuki, Y., Sussenbach, L., Luff, P., and Heath, C. (2009). "The first five seconds": Contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 985–991, Toyama, Japan. IEEE.

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Montreal QC Canada. ACM.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reeves, S. and Porcheron, M. (2022). Conversational AI: Respecifying participation as regulation. In *Handbook of Digital Society*, pages 573–592.

Reeves, S., Porcheron, M., and Fischer, J. (2018). 'This is not what we wanted': designing for conversation with voice interfaces. *Interactions*, 26(1):46–51.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., and Uhmann, S. (2011). A system for transcribing talk-in-interaction: GAT 2. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 12:1–51.

Skantze, G. (2021). Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

Stivers, T. (2015). Coding Social Interaction: A Heretical Approach in Conversation Analysis? *Research on Language and Social Interaction*, 48(1):1–19.

Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press, Cambridge.

Acknowledgments

We thank Svenja Filthaut for automating the detection of false-positive CVR-recordings.

Author contributions

MB: field recording, data annotation, formal analysis, data curation, writing - original draft, writing - review and editing, visualization, supervision
HH: conceptualization, methodology, writing - original draft, software adaptation CVR, hardware conceptualization and building CVRs (final version), data annotation, supervision, project administration
SR: conceptualization, methodology, writing - review and editing, supervision, software and hardware prototype CVR (first replication)

A Appendix

List of coding categories

- commandID
  - unique identifier of each command
- date
  - calender date of command
- time
  - clock time of command
- commandTranscription
  - text transcript of user command
- sentenceType
  - grammatical sentence type of command
    - declarative
    - elliptical
    - imperative
    - infinitive
    - interrogative
    - abortion
- responseTranscription
  - text transcript of VA reponse to user command
- dayOfRecording
  - day of recording, also counting days when no command was produced
- dayOfUse
  - day of use of the VA, not counting days when no command was produced
- VAtranscriptionCorrect
  - coding if the automatic speech recognition process transcribed the user command correctly
    - yes
    - no
- commandSuccess
  - coding whether the command triggered a fitting output to from the VA
    - yes
    - no
    - unclear

Transcript 1:[17]
CVR03-recording-220724162605

MO = Mother; CH = Child; AL = Alexa

01	MO	aLExa,
02		wir möchten ein SPIEL spielen?
		we'd like to play a game
03		(0.3)
04	MO	fü:r FÜNFjährige.
		for five year olds
05		(1.6)
06	AL	entschuldigung das weiß ich leider nicht
		sorry I do not know that unfortunately
07		(2.0)
08	MO	hö? ((lacht))
		huh? ((laughs))
09	CH	((kichert))
		((chuckles))
10	MO	°hh alExa können wir ein SPIEL mit dir spielen.
		Alexa can we play a game with you
11		(2.5)
12	MO	aLExa,
13		(1.0)
14	MO	°h können wir ein SPIEL mit dir spielen.
		can we play a game with you
15		(1.2)
16	AL	um musik aus deiner amazon musik bibliothek
		abzuspielen frage einfach nach dem song interpreten
		oder dem album das du gerne hören möchtest
		to play music from your amazon music library
		just ask for the song, artist or album
		you would like to listen to

About ten minutes after Transcript 1, mother and child try again to play a game with Alexa. In between, there was one successful request done by the child: After saying "Alexa what can we play", Alexa responds "Okay then let's choose a great game", offering a list of possible games. The mother tries to replicate this, first using an unsuccessful interrogative, and then eventually formulating a successful request (Transcript 2):

[17]Transcripts were created based on GAT2 transcription conventions (Selting et al., 2011).

Transcript 2:
CVR03-recording-220724163721

MO = Mother; CH = Child; AL = Alexa

01	MO	((lacht)) °h mach DU doch nochmal;=
		((laughs)) you do that again
02		=das hast du eben SUper gemacht.
		you have just done great
03		°h WAS hast du sie gefragt-=
		what did you ask her
04		=was für SPIEle gibt es. gell,
		what games are there, right
05		(0.9)
06	CH	JAha.
		yes
07	MO	ja FRAG se nochma;=
		well ask her again
08		=aLExa was für SPIEle gibt es.
		Alexa what games are there
09		(1.7)
10	AL	okay spiele, lass uns eins zum spielen finden
		okay games, let us find one to play