Linguistic Frontiers • 7(1) • 2024
DOI: 10.2478/lf-2024-0010

Linguistic **Frontiers**

How a Child Learns to 'Talk' to a Smart Speaker: On the Emergence of Enlanguaged Practices

Original Study

Marie-Theres Fester-Seeger
Faculty of Cultural and Social Sciences, European University Viadrina, Frankfurt (Oder), Germany
Postdoctoral Fellow, Postdoc Network Brandenburg
e-mail: mt.fester.seeger@gmail.com
ORCID: 0000-0002-8635-359X

Received: November 3, 2023; Accepted: March 30, 2024

Abstract: In this paper, I am concerned with the socio-material practice of engaging with voice-enabled machines. Far from 'talking' to a smart speaker, a user must master the skill of composing a command while routinely engaging with the machine. While the practice relies on practical understanding and intelligibility, attention must be paid to the trans-situational aspects that enable the situated enactment of socio-material practices. By conceptualizing engagement with the smart speaker as an enlanguaged practice, I trace the ability to engage in a seemingly individualistic practice to a person's history of engagement in and with the world. Specifically, I consider how a pre-literate child relies on instances of recursive bodily coordination with her caregiver to learn how to engage with a smart speaker. Informed by the languaging perspective which treats language as multiscalar bodily verbal activity, I trace enlanguaging to the intricate interplay of dialogicality, temporality, and embodiment.

Key Words: languaging, enlanguaged practices, human-machine interaction, temporality, dialogicality, embodiment

1.0 INTRODUCTION

People have been talking to machines in their home environment for over a decade. Relying on human voice, smart speakers[1], also widely known as Alexa, Siri, or Google Home (Hoy 2018), changed how scholars approach human communication with machines. Marketed as personal assistants (Dickel, Schmidt-Jüngst 2020), voice assistants, such as Amazon's Alexa, Apple's Siri, or Google Home Nest, offer a variety of functions, from sending and reading text messages to asking basic questions, setting alarms and timers, playing music, telling jokes, controlling other devices in the household, or creating shopping lists (Hoy 2018). The mix of simple semiotic means, such as assigning a persona to the device, implementing a synthetic voice, and implementing personality traits, masks the complexity of a vast underlying system that grants functionality to a voice assistant (Crawford, Joler 2018; Natale 2020; Natale, Cooke 2021)[2]. This masking leads to a 'perceived autonomy' of machines (Gahrn-Andersen, Cowley 2021), significantly influencing how people perceive (Purington et al. 2017) and engage with voice assistants in their home environment. While some scholars understand human engagement with speech-enabled machines as communication (e.g., Beneteau et al. 2019; Gampe et al. 2023; Guzman 2019), others focus on how users integrate voice assistants in their everyday practices

1 In this paper, I use the terms 'smart speaker' and 'voice assistant' interchangeably to describe stationary voice-enabled user interfaces.

2 Crawford and Joler (2018) show how the working of voice assistants, such as Amazon's Alexa, depends on a complex interplay of "non-renewable materials, labor, and data." What appears to be a single action in human-machine engagement "requires a vast planetary network of materials, human past labor and data," which enables people to talk to machines.

(Porcheron et al. 2018). In addition to the latter, Hector (2023) examines the practices emerging around a voice assistant through an ethnomethodological lens. Given the smart speaker's unique feature of processing voice, the engagement with a voice user interface (VUI) allows for interaction with other screen-based devices (i.e. smart phones) and other people. The parallelism and overlapping of engaging with other devices and people give rise to distinct interactional practices, such as the installation of a smart speaker, overcoming silence or co-producing commands (Hector 2023; Porcheron et al. 2018). Unlike touch-screen-based interactions, interaction with voice user interfaces affords "accountability of action" (Porcheron et al. 2017, 212). Making one's actions with a smart speaker publicly available allows for distinct practices (ibid.). Porcheron and colleagues (2017), for example, identify the practice of mutually producing silence in multiparty interaction with a VUI when waiting for the completion of the device's audible output. In both cases, people orient to the bodily movements of the other. While Porcheron and colleagues' and Hector's work illuminate how people integrate voice-enabled technology in their everyday practices and highlight new emerging practices, with the focus on the sequential organization of talk the use of conversation analysis allows only for a situated approach to human interaction with technology. Concerned with the question of how "interaction with a VUI is achieved *within* talk-in-action" (Porcheron et al. 2018, 2), attention falls on how people *perform* requests to the device. Due to its technological constraints, requests must be directed to the device in a distinct manner: A request must always start with a distinct wake word to 'wake' the system (e.g., 'okay, Google' or 'Alexa') followed by, for example, an imperative (e.g., 'play music') or interrogative (e.g., 'What's the weather?') lexio-grammatical structure (Barthel et al. 2022). Far from conversing with the device (Porcheron et al. 2018; Due and Lüchow in press), engagement with voice-enabled devices is a distinct practice. While Due and Lüchow (in press) describe the practice as 'VUI-speak' not much attention falls on *how* people adapt to the practice of engaging with a smart speaker. While adults might quickly adapt a practical understanding (Schatzki 2002) of the sociomaterial practice of engaging with a smart speaker, a child has yet to gain such practical intelligibility (ibid.). That is, not only learn how to produce a command but also how to conceptually perceive the device as a speech-enabled machine (Gahrn-Andersen 2023).

Rather than assigning human-machine engagement to the situated exchange of the spoken word (and other modalities), this paper deals with *how* a child learns to engage in the specific practice of engaging with a smart speaker through their engagement with a caregiver. Consequently, the paper also explores how the skill of giving a command enables the child to engage in further practices around a smart speaker. Dealing with distinct technological constraints, a successful engagement with voice assistants relies on using a distinct wake word (e.g., 'okay, Google') followed by, for example, an imperative lexico-grammatical structure (e.g., play music). While adults might quickly adapt a practical understanding (Schatzki 2002) of how to engage with a voice assistant successfully, a child has yet to gain practical intelligibility (ibid.). Although the skill of composing a command depends on the use of 'words' and, from a practice theory view, might be deemed as 'discursive' (ibid.), I argue that this specific practice although it relies on denotative action can similarly treated as *non-discursive enlanguaged doings* (Gahrn-Andersen 2023). At the same time, the machine's functioning is based on a homogenous and denotative understanding of language, which treats words as stable and abstract entities; a child as a human living being engages in active verbal activity (cf. Cowley 2019). Building on the languaging perspective, which traces language, first, to people's bodily activity and, second, to symbols (Cowley 2009), the paper explores in an autoethnographic study how the skill of giving a command to a voice assistant depends on recursive coordinated moments between the child and a caregiver. What Gahrn-Andersen (2023) calls the *enlanguaged* must be traced to the interplay of dialogicality (i.e., the engagement with others), multiscalar temporality (i.e., integration of past events), and embodiment (i.e., the use of one's body). The past recursive engagement with others, where a child learns how to phrase a successful command and what 'words' to use, constrains a child's bodily activity with a voice assistant in such ways that it appears to be an individualistic practice of interacting with a machine. A perceived regularity and stability of practice, thus, emerges through close coordination with other people, things, and the environment. Languaging is an "entangled meshwork that links living, observing, and social action" (Cowley 2019, 1) and thus gives rise to distinct socio-material practices. The paper aims to investigate how the enlanguaged practice of engaging with a voice assistant emerges.

2.0 TECHNOLOGICAL CONSTRAINTS OF A VOICE ASSISTANT

Before turning to *how* a child learns to engage in the specific practice of engaging with a voice assistant, I briefly outline the technological constraints that determine a person's enlanguaged engagement with spoken dialogue systems. In order to mask the complex interplay of hardware and software choices that underlie the functioning of voice assistants, designers and programmers make use of distinct semiotic choices, such as the use of a human-like voice, to create the illusion of a person (Natale 2020; Natale 2023; Natale, Cooke 2021). Assigning agency to a device and can lead to establishing parasocial relationships (Gampe et al. 2023). However, treating speech-enabled machines as social actors (Clark, Fischer 2023) not only obscures complex underlying technological structures but also the intricate interplay of "non-renewable materials, labor, and data" (Crawford, Joler 2018).

2.1 Anatomy of a Voice Assistant

Giving a successful command to a voice assistant can be sketched out in terms of the five-step model of spoken dialogue systems, which includes different technological elements: 1) automatic speech recognition software (ASR), 2) natural language understanding (NLU), 3) dialogue management, 4) natural language generation, and 5) text-to-speech synthesis (Gunkel 2020; McTear et al. 2016). Once taking a closer look at these technological elements, it becomes apparent that unlike what companies' advertising suggests, a user essentially engages with *text* (i.e., digitalized written inscriptions) when talking to a machine (Terzopoulos, Satratzemi 2020). While phenomenologically, a verbal engagement with a voice user interface appears as 'talking' (Fischer 2011), the machine relies on the identification of distinct keywords in order to convert acoustic signals into electrical digital inscriptions to be processed further by natural language understanding techniques (McArthur 2020). Jokinen and McTear (2009) describe ASR as a "probabilistic pattern matching process" that converts a user's verbal input to convert it to digitalized written inscriptions "that represent the system's estimate of the user's utterance" (5). The system calculates "the most likely sequence of phonemes (individual speech sounds) from the occurrence of any one particular sound" (Gunkel 2020, 144). Importantly, the models need not to process each single word, rather detect words "related to a specific domain in which the ASR is intended to operate" (Gunkel 2020, 145). In order for a language model to, then, model "the probability of sequences of words" (Jokinen, McTear 2010, 5), it relies either on hand-coded "rules of a generative grammar" or on "calculating the likelihood of different word pairings from analyzing patterns in available data (books, online publications, newspaper, etc.).[3]. In their basis, automatic speech recognition systems are thus to be understood as probabilistic processes (McTear et al. 2016). Accordingly, spoken input (and what we analytically can distinguish as 'phonemes', 'words', or 'sentences') is matched against patterns in the trained data sets and thus "produces a hypothesized textual result" (Gunkel 2020, 145). In spoken dialogue systems the dynamic meets the static. While machines rely on repeatable static patterns, human living beings are the opposite. In relation to social robots, Fischer (2011) points out how human verbal interaction with a social robot or, in the case of this paper, a smart speaker, "can hardly be predicted" (31). The heterogenous character of language (Cowley 2019) not only allows for the unpredictable in terms of what has been said but also *how* it is said. The materiality of voice poses, therefore, challenges for the programmers and designers of smart speakers as acoustic sounds vary permanently in loudness, structure, and dynamics (McTear et al. 2016). Human language can never be understood in terms of absolute repetitions on which the machines, however, rely. What makes a smart speaker, therefore, so complex is their ability to process human language and, most importantly, human speech successfully in challenging environments (e.g., a multiparty household) (Mallidi et al. 2018). At its core, however, the machine works on predicting single tokens and words (Mahowald et al. 2023). It is important to note that the companies that produce the most known voice assistants do not give a detailed account of how they function. However, according to Gunkel (2020), each system can be described in its basis as extracting single linguistic elements such as verbs, adjectives, and nouns. Having made human speech processable for the machine, and hence, the dialogue system, the system does not need to deal with the entirety of spoken input but only operates on given command or question-answer structures. Dialogue management systems are linked with external information sources and produce messages to be sent to the user. As a result, today's voice assistants only function as they are connected to the web (Natale, Cooke 2021). Thus, the successful working of the machine depends on the underlying algorithms, which present information from the web to the user in distinct ways (Gillespie 2014). As voice interfaces, voice assistants can search the web, play music (through being connected to third-party providers, such as Spotify), read emails or answer phone calls. In its basis, dialogue management systems extract keywords to match a user's query with information retrieved from the web (ibid.).

While Gunkel (2020) points out that queries that cannot be processed or fit a specific context are outfitted with pre-scripted replies, Natale (2020) views such pre-scripted responses as making up an essential aspect of a voice assistant's persona. In order to come across as, for example, funny or sassy, companies such as Apple and Amazon employ creative teams to craft scripted responses and jokes (Stroda 2020; Natale 2020). Viewing these as 'dramaturgical tricks,' Natale treats these assigned personal traits to the machine as tricks to deceive that these machines constantly harvest users' data to function properly. Natale (2020) notes that while the rise of deep learning has brought immense changes in AI, it has not touched upon all areas within the field of voice user interface design.[4] Consequently, "AI assistant developers can anticipate some of the most common queries and have writers come out with appropriate answers" (Natale 2020, 10), which deceives users into ascribing autonomy to a device. Once the dialogue management system finds the appropriate answer, it must be outputted as spoken text to the user. Through natural language generation algorithms, a response is then produced, which, through a text-to-speech function, is converted into a synthesized voice. These technological

3 On September 19, 2023, Deepgram's new speech-to-text model "Nova-2" was unveiled. Marketed as the most accurate and fastest ASR model, this model, according to the manufacturer, is "curated from nearly 6 million resources and incorporates an extensive library of high-quality human transcription" (Fox 2023, np).

4 On September 20, 2023, Amazon announced its plans to integrate generative AI in its voice assistant system Alexa (Rausch 2023).

Basic voice commands for all content providers

To do this:	Say "Hey Google," then:
Request a song	"Play [song name]" "Play [song name] by [artist name]" "Play [song name] from [album name]" "Play [song name] on [music service]" "Play songs like [song name]"
Request an artist	"Play [artist name]" "Play music by [artist name]" "Play [artist name] on [music service]" "Play songs like [artist name]"
Request an album	"Play [album name]" "Play [album name] by [artist name]" "Play [album name] by [artist name] on [music service]"
Play music based on genre, mood, or activity	"Play classical music" "Play happy music" "Play music for cooking" "Play [genre] on [music service]"
Play personalized suggested content from chosen service	"Play some music" "Play [genre] music on [music service]"
Shuffle	"Shuffle" "Shuffle [album]" "Shuffle some music" "Play [album] and shuffle" "Play [album] shuffled" "Play [album] on shuffle" You can also use an artist or playlist name instead of an album name.
Pause	"Pause" "Pause the music"
Resume	"Resume" "Continue playing"
Stop	"Stop" "Stop the music"
Play next song	"Next" "Skip" "Next song"
What's playing	"What's playing?" "What song is playing?" "What artist is playing?"
Control volume	"Louder" "Set volume to 40%"
Play music on your speakers, TV, or video device **Note:** You must use a Chromecast, Chromecast built-in TV, or Assistant built-in TV that is linked to Google Nest or Home speaker or display.	"Play music on my living room TV" "Play [genre] on my bedroom speakers"
Play music on a speaker group	"Play music on [speaker group name]" **Note**: Set up a speaker group in the Google Home app to enable this feature.

Figure 1: *Lists of basic voice commands for playing music on all Google Nest devices as listed on Google's support website (Google 2023).*

elements rely on deep learning algorithms that produce an audio waveform without relying on prerecorded samples but through articulatory synthesis (Gunkel 2020). The produced synthesized voice, then, makes up the main characteristic of a voice assistant. The deployed human-like voices (primarily female) work as anthropomorphic elements to foster what some call a parasocial relationship (Hoffman et al. 2021). This allows users to project a personality onto the assistant for an ongoing relationship.

2.2 Structure of a Command

So far, I have given rough sketch of the technical components which constitute a smart speaker. At their core, spoken dialogue systems depend on turn-to-turn structures (McTear 2009), such as simple question-response structures. For that acoustic signals need to be converted into string structures, for extracting the syntactic and semantic components of a user's utterances to which dialogue management systems create proper responses.

In order to be able to engage with the machine, the user must use a so-called wake word to 'wake up' the system. In the case of Amazon's Echo software, the wake word used and widely known is 'Alexa' for the Google assistant system, and their Google Home hardware is 'Okay, Google.' Only through the use of distinct lexical structures (Barthel et al. 2022) can the engagement with the device within the command response model be successful (Natale, Cooke 2021). Thus, the inputs a user has to give the device to work function as prompts. Natale and Cooke (2021) note that while a "computer interface that takes up the language of humans, they are also an interface that stimulates humans to take up the 'language' of computers, that is programming language" (1007). Developers of these systems provide users with clear documentation for how to engage with the device. For example, basic voice commands for playing music would be "Okay Google, play [song name]" or "Okay Google, play, [artist name]" (see Figure 1). The device, therefore, depends on the users to use imperatives and also interrogative ("What's the weather?") structures (Google 2023; Barthel et al. 2022).

Once a user manages to address the device successfully in terms of distinct lexicogrammatical structure, which resembles programming language as shown in Figure 1, they are able to engage in distinct practices with the device. Natale and Cooke (2021) conclude that "[a]n expert user of voice assistants will learn the commands that are most effective in order to have voice assistants operate as they wish […]" (1007). This documentation, however, works well for literate people. However, children who are not yet literate rely on recursive engagements with their caretakers to learn how to engage with these distinct voice user interfaces (see section 4). The structure of the device does not work beyond the realm of the command/respond model. Consequently, according to Due and Lüchow (in press), people engage in *VUI-Speak*. Recognizing this specific engagement with smart speaker as a participant practice, the authors highlight the five-part sequential structure that frames the human engagement with a smart speaker as follows: 1) human participants need to utter a wake-word in order to engage with the system, 2) the systems signals its readiness, for example, through lighting up briefly[5], 3) the human participant now gives their command adhering to a strict lexicogrammatical order (see Figure 1), to which 4) the system produces a response. 5) Depending on the anticipated output, the human participant either corrects their command and re-iterates the sequence or accepts the command and responds with silence. This action-based approach to conversation underlies the design of spoken dialogical systems (McTear et al. 2016). It thus uses structural units such as question-answer and offer-acceptance adjacency pairs (Schegloff, Sacks 1973).

Due to its technological underpinnings, the way humans engage with smart speakers, of course, contrasts significantly with human-to-human engagement. Thus, I argue it can be thought of as a unique practice that relies on the use of a wake word (see Due, Lüchow in press) and, as Natale and Cooke (2021) have put it, requires a person to adapt 'the language of a computer,' and to give distinct commands. The structure of smart speakers relies on a distinct perspective on language, emerging from Chomskyan approaches to assign a generative grammar to structure and from the sequential organization of talk (McTear et al. 2016). However, how human living beings *use* language differs significantly from the structuralist perspective implemented in a machine to make a machine work. In the end, machines work, in a simple sense, on the basis of pre-defined meanings, which leaves no room for any sort of creativity. Thus, human language must be made processable for a machine as it is broken into basic form-meaning distinctions (Bender, Koller 2020). Basic NLP principles are thus tokenization, stemming, and vectorization to process what widely is referred to as 'natural language.'

3.0 LANGUAGING, ENLANGUAGED PRACTICES, AND CONCEPTUAL ATTACHING

'Talking' to a machine is far from talking to another person. While a machine's working depends on concrete form-meaning distinctions, our verbal engagements with others go beyond the 'said.' Treating human engagement with a smart speaker as practice, my concern is, in particular, how a child learns to successfully engage with a smart speaker through recursive coordinative moments with a caregiver. The following highlights the heterogeneous character of languaging by emphasizing its dialogical, multiscalar, and embodied characteristics.

5 According to Stone (2021), when designing the first Amazon Echo, Jeff Bezos proposed to implement an LED light ring on top of the device that would light up in order to create a sense of social cues when talking to the device.

I argue that a person should be understood as a meshwork of past influences continuously determining their situated engagement with things and others. After outlining this view, key aspects of Schatzki's understanding of practices are highlighted and linked to Gahrn-Andersen's (2023) notion of enlanguaged practices and conceptual attaching.

3.1 Languaging

If one closely observes how humans talk with each other, one notices that human language extends the 'said' or denotative meaning (Gahrn-Andersen 2023). However, when language is treated as an abstract formal system, it can lead to a detachment of people or persons from language, thereby undermining its essence as a human social activity (see Cowley 2019). This reduction of language to a homogeneous "system of abstract forms and formal operations" overlooks the individuals who "produce and experience […] in concrete acts of language activity" (Thibault 2021, 4). These concrete acts of activity are intricately woven into "an entangled meshwork that links living, observing, and social action" (Cowley 2019, 1). Thibault (2021) highlights three tangible consequences of separating people from languaging, which should be a cause for concern:

> "1. The splitting of experiencing and observing self from the experienced and observed world;
> 2. The splitting of the experienced and observed world into the language system separated from its environment;
> 3. The splitting of the language system into its component parts, their principles of combination, their formal regularities, and so on." (Thibault 2021, 4).

Thibault emphasizes how languaging depends on people (to whom he refers as 'selves') from an observed and experienced world. This emphasis on observing emerges from Maturana's (1988) biological view on languaging, which greatly influenced the languaging perspective. Maturana argues that "[w]e human beings operate as observers, that is, we make distinctions in language." (26). Denying the widely accepted conceptualizations that people use talk to "denote and connote […] entities that exist independently from us" (ibid.). In fact, given that human reasoning is endowed with rationality, one is constantly exposed to one's own experiences. Thus, Maturana concludes that "any explanation or description of what we do is secondary to our experience of finding ourselves in the doing of what we do" (ibid.). While experience comes first, language comes second. In Maturana's famous words, "everything said is said by an observer to another observer that could be him- or herself, and the observer is a human being" (27). While not precisely mentioned by Maturana, observing is a highly embodied activity. Referring to the "chiasm between the various sense modalities, such that they continually couple or collaborate with one another" (Abram 1997, 128), Abram explains how the "synaesthesia between the human eyes and ears is especially concentrated in our relation to other animals" (129). Referring to indigenous hunting techniques, Abram holds that, especially in such practices, the interplay of eyes and ears fuse into a "hyperattentive organ." In his words, "We feel ourselves listening with our eyes and watching with our ears, ready to respond with our whole body to any change in the Other's behavior" (129). While the Other for Abram are animals or other aspects of an animistic surrounding, I argue that the ability to carefully and attentively observe and monitor the actions of an Other grounds languaging. As in newborn-caregiver interaction, Trevarthen (2011) observes that "infants, it appears, are born with motives and emotions for actions that sustain *human interactivity*" (121), which is brought forth through distinct coordinative ways of engaging with the movements of others, such as through distinct ways of imitating and mimicking the movements of others (Delafield-Butt, Trevarthen 2015; Trevarthen, Aitken 2001 Gahrn-Andersen, Cowley 2017). Through vocalizing or whole-body movements, infants "enter into a communicative and cooperative relationship" with other adults (Trevarthen 2011, 124). Thus, persons, whether adults or infants, must always be considered as interwoven in distinct social systems:

> "Not only is a fetus contingent, a part of a woman's body, but an adult, man or woman, is also Contingent, part of a larger whole, family or, community or ecosystem. We cannot afford to emphasize the individual too far, for no one -fetus, child, or adult – is independent of the actions and imaginations of others. Persons are human individuals shaped and succored by the reality of interdependence." (Bateson 1994, 63).

Thus, an individual cannot be conceived of, as it is often done in Western ideologies, as an individual self but must be accounted for as 'fluid' (Bateson 1994, 63) and a 'zone of entanglement' or a meshwork (Ingold 2008). While Bateson views a person as "held in a vessel of many strands, like the baskets closely woven by some Native American tribes" (ibid.), Becker (1999) points to Language/languaging to be understood "as something like a web" (233) that consists of vast and interwoven strands, which also can bear great holes and carry a sort of mysteriousness for a person.

Applying these metaphors of weaving baskets or the intricate ways of a web, one needs to ask what these strands hold together. What parts of past movements and situations are being interwoven and retained? This question of temporality and/or multiscalarity is one of the most crucial aspects of the languaging perspective (Cowley, Steffensen 2015; Cowley, Madsen 2014; Gahrn-Andersen 2019; Enfield 2014). Within this perspective, different approaches to temporality exist. An ecological-enactive view on multiscalar temporality (Loaiza et al. 2020; Steffensen, Pedersen 2014), for example, highlights "the structuring effect of histories of interaction" (Loaiza et al. 2020, 18) and refers these to, among others "mapping of temporal ranges, organising frames […]" and constitutional constraints (ibid.). Thus, Loaiza and colleagues (2020) describe

multiscalar temporality in terms of its "lived richness and depth of field of the present moment populated by non-local events and what appears to be absent in the here-and-now." (17). Their account allows for the expanded here-and-now and its constraints and underlying temporal scales (which can range from slower sociocultural temporal scales such as temporal ranges that constitute a dialogical system (cf. Steffensen, Pedersen 2014). While it is crucial to account for this underlying multiscalarity that 'amalgamates' in a given moment (Madsen 2017), others have pointed to a narrower view on *how* a person draws in distinct ways on the absent and brings about diachronic dispositions (Cowley, Fester-Seeger 2023). In such a person-oriented view, persons are treated as observers, and attention must be given to how people actively draw on past or non-local events in their ongoing bodily activity, which is playing out in real-time. Thus, one asks how people as human cognitive agents perform in a for-them familiar world and how they bring aspects of such a wolrd about (Fester-Seeger 2024a). With emphasis on a person's systemic embedding, one is left to ask how these systems and, thus, distinct histories come to emerge *for* a person. Here, one is well-advised to look at Maturana's (1988) notion of coordination and recursivity.

While Maturana does not specifically mention the role of the human body in languaging, others who advocate for the languaging perspective have. Raimondi (2019) views Maturana's notion of coordination as something that can be "achieved when the individual's action is oriented and constrained by the actions of the other" (22). He illustrates this with the following example:

> "To clarify the generative power of recursive coordination, it is best to see an example of how it functions. Let us consider a coordination such as the passing of toys between an infant and his/her caregiver. This is a "flat," non-recursive coordination. We can expect that playing this game, the adult will add vocalization to his gestures and movements. In other words, the set of operational components of this coordination include adult vocalization that we (but not the infant) recognize as utterances such as "teddy bear", "give it to me" etc. in line with what we said before, mean that we can understand the game of passing the teddy-bear as a configuration of consensually coordinated operations. However, things get more interesting when, over time, the routinization of this game eventually engenders the rise of more complex activities; say, when the adult asks, "Where is your teddy bear?" or "Bring me your teddy bear" and the infant becomes capable of pointing to or seeking it." (Raimondi 2019, 22).

Raimondi integrates the role of embodiment in this process of passing a teddy and the toddler learning to distinguish the teddy as an object, coordination here is understood in terms of 'consensually coordinated operations,' and how past events of coordination get integrated and bring forth "new forms of joint activity." (Raimondi 2019, 22). Eventually, certain vocalisations, such as 'pass the teddy,' become integrated with practices and become part of language. Raimondi shows how a sense of language and concepts emerges through recursive coordination with others.

Once one places human activity at the center of languaging (whose wordings are constrained by, among other things, lexis, usage, pragmatics, and syntax) one can give due attention to how people experience language, how they draw on what is not 'there,' and how they manage lived situations (see Cowley, Fester-Seeger 2023). Languaging thus goes beyond the word as it is spatiotemporally distributed, ecological, embodied, and highly dialogical. Attached to how people *do* language, the perspective enables one to investigate how people rely on human activity that includes verbal aspects. Once taking away the focus on the word, languaging enables one to trace how people bring about aspects of what it is absent through verbal activity. Bringing forth past circumstances in the form of evoking the absent determines how, for example, one reads messages, conceives of concepts, and motivates their storytelling in order to create understanding or knowing (Fester-Seeger 2024b). These aspects of the absent are traced in the interplay of bodily dynamics. Given the importance of embodied coordination, one can pursue what permeates one's bodily actions. What we generally describe in terms of 'words' are often better understood as 'repeatables' (van den Herik 2022; Love 1990). In languaging, we perceive similar patterns of activity as the 'same.' People attune to both bodily and verbal patterns over time (Cowley 2011) by using recursive coordination. As a result, they continuously incorporate aspects of the slower temporal scales (Raimondi 2019; Gahrn-Andersen 2019.

3.2 Enlanguaged Practices

Imagine observing grandmasters at a game of lighting chess. At the given moment, we can only see two people moving pieces without relying on talking to each other. Each person focused on the chessboard and the pieces in front of them. What appears to be a purely mentalist activity and a non-linguistic socio-material practice is, in fact, highly enlanguaged (Gahrn-Andersen 2023a): it depends on prior instances of bodily coordinative and verbal activity. While practices can be understood as routinized stable activities that enable one to manage unknown situations, they rely on a history of people's engagement with aspects of their environment. The idea of repetition and stability leads to neglecting the role of human activities in practices (Schäfer 2013). Once human activities move into the foreground, the general conception of practices as stable entities can no longer hold up (Barnes 2001, 30). While Rouse (2007), for example, differentiates between treating practices as "ephemeral doings" (such as wordings in languaging (See Cowley 2011)) and as "stable long-term patterns of activity" (639), Schatzki (2002) defines practices as "a temporally evolving, open-ended set of doings and sayings," whereby a sense of stability emerges through "practical understanding, rules and

teleoaffective structure6, and general understandings" (87). Schatzki (2002) differentiates, too, between 'dispersed' and 'integrative' practices. The former relates to practices that "center around a single type of action," such as "describing, questioning, reporting, and examining," and thus can be applied in various social contexts. For Schatzki, this kind of practice is 'rule-free' due to the characteristics of its use in various contexts. Integrative practices, in turn, hint at the complexity of integrating "multiple actions, projects, ends, and emotions" (88). Practical understanding relates to the procedural enacting of a specific practice and, thus, needs to be understood as "components of practice" (Welch, Warde 2017, 187), which grounds regularity. Schatzki's understanding of "knowing how to X, knowing how to identify X-ings, and knowing how to prompt as well as respond to X-ings" (Schatzki 2002, 77) constitutes a practice. Thus, while practices "exhibit regularity," they enable one to "embrace the irregular" (73). However, a practical understanding of how to do things in a specific routinized way can only emerge as one tries to manage the irregular. Hence, rather than establishing a dichotomy between the ephemeral and the idea of stable long-term patterns, the focus should turn to, one, how such seemingly stable patterns of activity permeate ephemeral doings and, two, how stability emerges through recursive instances of bodily dialogical coordination. Although Schatzki (2001) traces practices to "bodily doings and sayings" (72), contrary to the languaging perspective, he separates the two. While Schatzki describes doings as direct, perceivable bodily actions such as waving, running, or throwing, sayings are reduced to the linguistic or denotative. In other words, to what people 'say.' A game of chess in which no words are involved might, therefore, only be understood in terms of distinct bodily actions but not of any linguistic activity. This separation of the discursive and non-discursive leaves aside how practices emerge (Gahrn-Andersen 2023a). Seemingly non-linguistic practices, however, rely on prior instances of verbal activity. Practices are, therefore, trans-situational7 (Linell 2009). From an observer perspective, one can easily identify practices in terms of their procedural and situational enacting (Welch, Warde 2017, 187) and an individual's understanding of practice, which Schatzki identifies as practical intelligibility. In his view, "[i]t is always to an individual that a specific practice makes sense" (Schatzki 2002, 88). Schatzki fails to explore further how such an apparent individual capacity emerges. Situated and individualistic approaches to practices leave aside how people gain practical understanding and intelligibility. Much more attention must, therefore, fall on not practice as a stable entity but on people as human living beings who engage in practice-based activities. In line with Barnes' (2001) critique of the individual and social divide in practice theory, "human beings cannot be understood as independent calculative individuals; they stand revealed in their practices as profoundly, mutually susceptible social agents" (34). Once attention falls on human beings as fluid selves (see section 3), who are susceptible to the actions of others, practices must be treated as spatially and temporally distributed (Gahrn-Andersen 2023b). Turning to the grandmaster's game of lightening chess, finding an answer in Wittgenstein's language game of naming objects (2009, 21e), Gahrn-Andersen highlights how teaching a person to name objects arises out of distinct coordinative moments of pointing to while also naming the object. In the specific case of chess, one points to a wooden piece and identifies it as 'This is a king,' the language game simultaneously involves gaining practical understanding and intelligibility. Through active bodily engagement of directing a person's attention to aspects of their immediate environment, people learn to identify objects regarding their "practice-related relevance to other objects or doings" (Gahrn-Andersen 2023b, 16). Enlanguaged practices are not only trans-situational; that is, they rely on one's experience of having learned how to play chess, but also trans-practical; they allow for further practices. Thus, a novice learns that a wooden piece can be called a king and how and in which situations the king can be moved in a chess game. A person not only gains a practical understanding of how to move a piece but also practical intelligibility and an understanding of chess. In simpler words, attention must be given to the 'words' which precede the non-linguistic aspects of otherwise linguistic practices.

3.3 Conceptual Attaching

To successfully play a game of chess or, in the case of this paper, to engage with a voice assistant, a person relies on concept-infused perception. A chess player must be able to identify wooden pieces as chess pieces, while a child learns to "take up the 'language' of a computer" (Natale, Cooke 2021, 1007). Rather than following a mentalist approach to concepts where concepts are treated as innate, intracranial, and, thus, underlying language and perception (e.g., Potter 2018), Gahrn-Andersen (2021, 2023b) treats concept-infused perception as *activity*. Defining conceptual attaching as "the basic process whereby a cognizer conceptually identifies a thing in their perceptual horizon" (2023b, 2), the notion emerges out of radical embodied approaches to cognition and language (see section 3.1). In this view, human perception is not

6 Schatzki (2002) defines teleoaffective structures as "a range of normativized and hierarchically ordered ends, projects, and tasks, to varying degrees allied with normativized emotions and even moods" (80). Schatzki, however, does not refer to normativity in terms of 'acceptability' but oughtness, that is, a range of tasks or elements necessary to carry out a specific practice.

7 Emphazising how sociocultural resources permeate the situated (i.e., specific encounters, specific participants, and time and place), Linell (2009) brings attention to how "interaction and practices are located on different time scales" (52). While situated interactions are tied to specific spatio/temporal domains, sociocultural practices extend over longer timescales. Thus, "participants in situated interactions contribute over time to sustaining changing the more long-term, situation-transcending practices" (ibid.).

achieved through mental representations (Noë, 2004) but is "an accomplishment stemming from our continuous explorations of the environment" (Gahrn-Andersen 2023b, 3). With the emphasis on human embodiment and skills of action, conceptual attaching relies on the "enactment of as-structures" (Gahrn-Andersen 2023b, 3). Building on Heidegger's notion of as-structures, Gahrn-Andersen (2023b) points out how perceiving things *as* something "cannot be taken in isolation from their socio-practical context" (3). In Heidegger's (2010) words (cited by Gahrn-Andersen), "we simply see and take things as they are: board, bench, house, policeman. However, this taking is always a taking within the context of dealing with something, and therefore is always a taking-as […] the as-character does not become explicit in the act" (122). Rather than perceiving things as they 'really are' (however that could look like, as Gahrn-Andersen remarks in an objectivist sense), concept-infused perception must be grounded in socio-practical doings. In other words, how people perceive things depends on their practical doings with the perceived things. For conceptual attaching to take place, human living beings must first be enculturated and possess "linguistic skill and know-how" and must "be actively oriented toward an object, a piece of equipment, person, a body-part" (Gahrn-Andersen 2021 7). Importantly, perception does not precede conceptual identification, rather both processes are intertwined. In taking things 'as' something, e.g., a wooden piece as a king, people bring forth histories of active engagement in the world with others. Gahrn-Andersen finds an example of conceptual attaching in Wittgenstein's language game, ' It could also be *this*'(Gahrn-Andersen, 2023). In this game, children perceive a chest as a house and must point out different aspects of it and ascribe to its house-like features. Wittgenstein refers to this, pointing it out as lightning. Children do so by uttering, 'Now, it's a…'. Once non-house related aspects are uttered or praxis logic of pointing out an aspect together with the utterance is not followed, the game terminates. Acts of conceptual attaching show that specific aspects of the chest are evoked as parts of a house and, therefore, constitute this game. Gahrn-Andersen points out that the example shows that the utterance 'Now, it's a house!' goes beyond semantics as the children must also deploy a specific practical understanding to play the game. Once they understand the game, children can engage in further practical activities in pretend play (e.g., playing house). This example shows that conceptual attaching arises through people's active engagement with aspects of their immediate environment. For Gahrn-Andersen, the children rely on conceptual know-how (that is, having an idea of a house as they rely on their own past experience of active engagement with and in a house), which becomes enacted through as-structures. The language game shows how conceptual attaching relies on situatedness, social consensus, and "a clear agent-to-world directionality" (Gahrn-Andersen 2021, 2). Gahrn-Andersen (2021, 2023b) construes in detail how conceptual attaching constitutes socio-practical activity. However, one is left to ask how people come to take certain things *as* something. This, so I argue, needs to be traced to instances of bodily dialogical recursive coordination.

4.0 "YOU ALWAYS HAVE TO SAY, 'OKAY, GOOGLE": A CASE STUDY ON HOW A CHILD LEARNS TO ENGAGE WITH A SMART SPEAKER

Much like Wittgenstein's "Now it's a house!" language game, successful engagement with a smart speaker depends on addressing the device through commands (e.g., "Okay Google, Play a song"). Through the instructions provided by the company when the device is installed, literate people can easily adopt the distinctive lexical-grammatical structure to construct a command. A non-literate child, however, has not yet acquired the practical understanding and intelligibility that would allow them to routinely engage with a smart speaker. Before a child is able to conceptualize a smart speaker as a voice user interface through that it can assess the Internet, a child is dependent on a caretaker's ability to conceptually identify the artifact as a smart speaker. The practice of engaging with a device depends on both denotative linguistic sequences and on what is not *said*: knowing how long to wait for the device's output and gaining an understanding of the machine's algorithm (Due and Lüchow in press). In a Schatzkian sense, the practice involves specific doings and sayings. I trace Gahrn-Andersen's (2023a) understanding of the *enlanguaged* to distinct past instances of recursive dialogical coordination. Although Gahrn-Andersen (2023a) focuses on how non-discursive practices (i.e., where no words are involved) bind people's past linguistic engagements with others, linguistic engagement with voice assistants can be treated in similar ways. Although speech is clearly involved in the process, a person needs to adapt to 'the language of the computer' (Natale, Cooke 2021) and, therefore, to form an understanding of how to react to the device's output. As computer linguistics relies on a classical view of language, where language is not traced to human activity but a linguistic system consisting of fixed codes (Love 2004), a user needs to engage with pre-defined rules and patterns of a language system. However, as outlined in section 3.1, how people *do* language goes beyond the 'word' (understood in terms of a stable entity) as languaging is spatiotemporally distributed, ecological, embodied and highly dialogical. Thus, when addressing a smart speaker, a user needs to adapt to formalized linguistic expressions. Even though a user uses voice to operate the device, the smart speaker relies on *text* – on static entities and abstraction of language (as outlined in section 2). People, however, engage in *languaging*, or human bodily activity in which the verbal 'plays a *part*' (Cowley 2019). Much more happens beyond the *word*. Languaging is, therefore, activity, in that it is sensorimotor perception-action that can, at once, be perceived in terms of 'wordings' (Cowley 2011). The concern, therefore, is not with 'words' as formal, abstract, repeatable entities but vocal gestures. There is no such thing as 'real' repetition

(Cowley, Nash 2013) because each action is preceded by other actions and with that moving in time and space. In linguistic coordination, people perceive 'wordings' as nonce (unrepeatable) events that occur during real-time interaction. In a present autoethnographic case study (Poulos 2021), I therefore investigate instances of bodily dialogical coordination where the intricate interplay of bodily micro- and pico-dynamics (e.g., the role of gaze, body posture and facial expression) play a central role (Thibault 2011). While practices appear as routinized and stable, they emerge out of the irregular – out of ephemeral actions. Through the coordination with others, such ephemeral actions, or pre-predicative perceptual acts (Gahrn-Andersen 2023b), gain on perceived regularity. Using multimedia event analysis (King, Thibault 2016), I explore in detail how a mother guides the actions of her child in such a way that the child can conceptually perceive of a command as a command to a non-human entity. The ethnographic case study builds on key points of cognitive ethnography which "combines traditional long-term participant observation with the micro-analysis of specific occurrences of events and practices" (Alač, Hutchins 2004, 632). In focusing on long-term observations, analytical focus falls on how participants integrate material artefacts, past events, socio-cultural constraints and aspects of their immediate environment in their process of gaining practical understanding and intelligibility. In this case study, I focus on not how a person comes to integrate a voice assistant as a participant in everyday practical activities (cf. Hector 2023), but on how a child learns how to engage with a smart speaker through commands and thus gains a specific praxis logic. The child integrates the ephemeral doings of her mother into her actions and thus builds a conceptual understanding of giving a command to a voice assistant. Thus, the child must perform as an active observer (Maturana 1988) who is sensitive to the movements of others in her immediate physical environment. Moreover, such views contribute to how so-called non-local resources (Steffensen 2013) emerge through recursive coordination and enable enlanguaged practices. In the analysis, therefore, I dissect important aspects of embodiment, dialogicality, and multiscalarity. Learning how to address the device *as* a non-human entity becomes essential for the child to implement other practices around the smart speaker (e.g., using the smart speaker for listening to music while cooking or brushing teeth).

We now turn to a German-speaking child, whom I named for the purposes of this paper Hannah. I observe how Hannah learns from her mother how to engage with a smart speaker, which involves the correct composing of a command. The study was conducted from March 2022 to March 2023. At the time of recording and introducing the smart speaker to the home environment, the child was three years and eight months old. Being introduced to the smart speaker, the child had no conceptualization of a smart speaker as a non-human entity and what it can do. Hannah thus relies heavily on her engagement with her caregivers. In what follows, I focus on two distinct instances from the third day of recording Hannah's engagements with the voice assistant. The first part points to coordinative moments between the mother and Hannah, and the second part to how Hannah unsuccessfully tries to turn the device off through verbal commands.

4.1 Learning how to give a command: coordinative moments between Hannah and her mother

The setting up of the device takes place using different material artifacts: the phone application "Google Home," (Hector 2023) through which the smart speaker is connected to the phone, and the printed instruction manual that came with the physical smart speaker. The installation process consists of the mother's coordinating activities through the phone application, where clear instructions must be followed. As the instructions in the phone application are given through written digital inscriptions, the mother, as a literate person, could easily set up the device. At the same time, the mother follows the manufacturer's instructions through digital inscriptions; the child, who cannot yet read or write, has to follow the caregiver's movements. Three days after setting up the device, the mother places the smart speaker in the living room, where the child sits at the dining table and draws. The mother places her phone in front of the table to film the child's engagement with the smart speaker. It is also important to note that at that time the smart speaker did not have a fixed location in the home and therefore, especially in the first few weeks after the purchase of the smart speaker, the location within the home environment changed quite frequently together with the movements of the person within the home ecology. The video[8] under study is 38.37 minutes long, from which I focus on three instances to investigate the coordinative moments between mother and child that contribute to the process of gaining a practical understanding of how to give a command to a smart speaker. As described in section 2, the child needs to enact a specific lexicogrammatical and sequentially organised structure in order to successfully engage with the device (e.g., Barthel et al. 2022).

Transcript 1: ## 00:44 –01:42
(Hannah drawing in livingroom, Pos. 12-30)
1 G: Das Mikrofon ist wieder eingeschaltet
The microphone is turned on

2 M: ah guck mal jetzt isse wach
Ah, see, she's awake now

3 M: Was möchteste der sagen, Hannah? Dem Google Gerät?
What would you like to tell her, Hannah? The Google device?

8 The video derives from a wider autoethnographic case study conducted by the author. The video is available for viewing at the following link: https://my.hidrive.com/lnk/WpcZ4F33L.

4 H: hm: Google (.) mach mein Lieblingslied an
Hm: Google (.) turn on my favorite song

5 M: musst du anfangs sagen ‚okay (.) Google'
You have to say in the beginning 'okay (.) Google'

6 H: okay (.) Google

7 M: HHH

Du musst immer anfangen zu sagen Okay Google
You always have to say 'Okay Google'

8 H: Okay Google (.)

9 M: und jetzt den
And now the

10 H: Hm. Was hat denn der Junge Goo:gle gesagt? Der Junge hat so lange geschlafen.
Hm. What did the boy Google say? The boy slept for so long.

11 M: Google hat so lange geschlafenhhh?
Google has slept so long hhh?

12 M: du kannst sagen hm: OKAY Google (.) spi:el (.) das Dino Lied
You could say hm: „OKAY Google (.) pla:y (.) the Dino song.

13 G: alles klar (.) das Dino-Lied von Simone Sommerland, Karsten Glück und die Kitafrösche (.) Hier ist es auf Spotify
all right (.) the Dino-Song from von Simone Sommerland, Karsten Glück und die Kitafrösche on Spotify.

In this instance, the mother has plugged in the smart speaker, which immediately leads to the following output from the device: "The microphone is turned on again".[9] The audible output does not yet evoke any bodily response from the child. She continues to draw, concentrating on her piece of paper (see Figure 2).

The mother responds to the audible output of the smart speaker by attributing the following observation when she says, "Ah, see, she's awake now" (2). Through her observation, the mother depicts the mechanical device as a social agent (Clark, Fischer 2023) as she assigns a persona to the device by using the third person pronoun 'she'. The mother's utterance prompts the child to move her gaze from her drawing to the smart speaker. Having drawn the child's attention to the device, the mother guides the child's engagement by saying, "What would you like to tell her, Hanna? The Google device?" (3). In the first part of the utterance, the mother again depicts the device as a social agent, while correcting herself in the second part of the utterance where she corrects herself by referring to the device as a non-human entity. There are two distinct instances of conceptual attaching (Gahrn-Andersen 2021, 2023): the mother takes the device *as* a depiction of a social agent, and, in terms of correcting herself, as a machine. The mother's intention might be to make sure that her daughter conceptually perceives the machine as a non-human entity. In response, Hannah turns to the phone and then back to the piece of paper as she says the following: "Ehm, Google, play my favourite song" (4). The structure of her utterance shows that the child has already gained some understanding of how to interact with the smart speaker, as she uses an imperative sentence structure ("turn on my favourite song") (Barthel et al.

9 The device offers the function of switching off the microphone.

Figure 2: *Hannah focused on drawing during the voice assistant's audible output, "The microphone is switched on." The mother standing to Hannah's right, and the Smart Speaker is located on the table in front of her.*

2022). However, there are two aspects missing from any "practical understanding" (Schatzki 2002) in the context of interacting with spoken dialogue systems: 1) conceptualising the use of a wake word, and 2) dealing with another practice of dealing with algorithms. Hannah does not know that the music function is connected to the audio streaming provider Spotify. She lacks a conceptualization of a smart speaker as a web interface (Natale, Cooke 2021).

Observing Hannah's utterance, the mother responds with slight laughter which causes Hannah to smile about her given command. The mother uses this opportunity to teach the child how to address the smart speaker. Although the child is still focused on her drawing, the mother engages with her. What follows is a coordinative moment revolving around the idea of using a wake word in order to engage with the smart speaker successfully. The mother tells the child that she has to say 'okay (.) Google" before giving a command. The child imitates her mom's utterance (5-6). A closer look shows that Hannah's audible imitation differs slightly from her mom's (see Figure 2).

The mother's utterance starts with an upward pitch movement on the syllable 'o,' whereas Hannah's pitch moves downwards. The pitch of both differs in that the mother moves from around 258 Hz to 295 Hz while the child moves from 377 Hz to 326 Hz. Although there is a pitch discrepancy, Hannah's vocal movement seems to approaching her mother's pitch. The wording 'Google' differs in vocal movement between the two. While the mother keeps a steady pitch movement, Hannah's vocal activity can be divided by the two syllables of the wording 'google' in two different sections. That is, 'goo' and 'gle.' Her pitch moves from 242 to 222 Hz on the first syllable and slightly decreases from 255 Hz to 227 Hz on the second syllable. In contrast, her mother's utterance remains almost steady at around 220 Hz. This comparison of the mother's and Hannah's attuning of vocal dynamics shows how teaching to give a command extends the mere repetition of saying the 'same' (Cowley, Nash 2013). As argued by Gahrn-Andersen (2023), a dichotomy between 'sayings' and 'bodily doings' becomes untenable from a languaging perspective. Bodily dynamics underlie and bring about the said (Cowley 2014). In this instance, the child tests and adapts her vocalizing strategy to her mother to engage with the smart speaker successfully and not the smart speaker directly (Gampe et al. 2023). This coordinative moment between mother and Hannah grounds the enlanguaged practices of engaging with a smart speaker *for* the child.

As she 'repeated' *okay, google*, the child paused and kept looking at her paper. After roughly a second, the mother continues guiding Hannah in her actions, as she says, "and (.) now (.) the." Hannah remains focused as she begins to move her upper body, possibly preparing to give a command to the device. The device responds with the message 'All good,' prompting Hannah to smile slightly and turn her attention to the phone recording her (Figure 4).

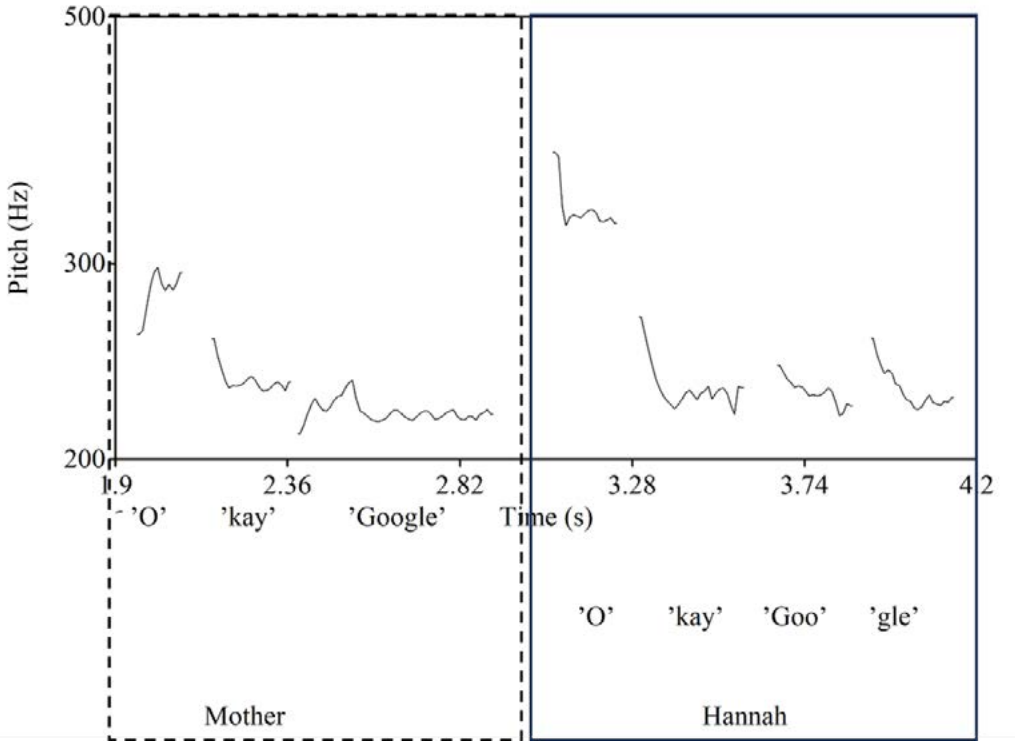

***Figure 3**: Pitch contour of mother's (dotted rectangle) and Hannah's utterance (solid rectangle) 'Okay Google.' Generated with the speech analysis software Praat (Boersma, Weenink 2023)*

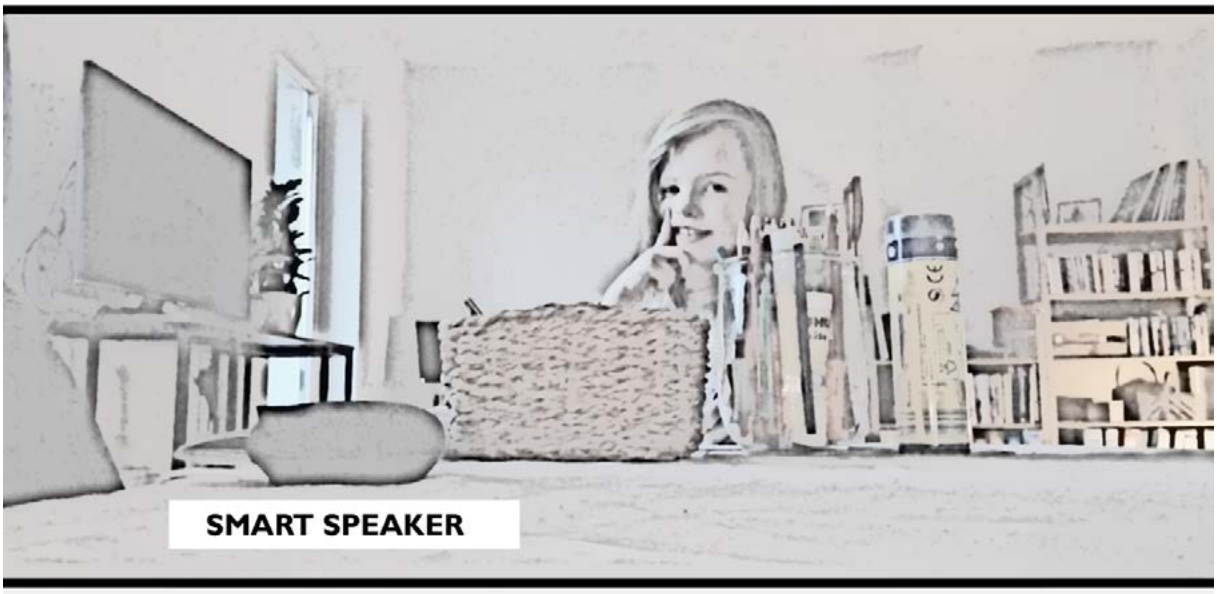

***Figure 4**: Hannah's reaction to the voice assistant's output "all good" after engaging with her mother to construct a command.*

By teaching Hannah to use the wake word 'okay Google,' the mother points to distinct aspects that constitute a smart speaker. In a Wittgensteinian sense, she makes particular aspects light up (Gahrn-Andersen 2023). As Hannah adapts her vocalizing to her mothers, her utterances become conceptually constrained. That is, Hannah begins to learn to treat the utterance 'Okay Google' as a constitutive action (or 'way with wordings') within the sociomaterial practice of engaging with a smart speaker.

In response to the device's output, Hannah pauses and asks, "What did the boy (.) what did Google say?" (10). While speaking, Hannah alternates her gaze between her drawing and the phone. She then continues as she turns to her mother and says, "The boy (.) goo= google slept for so long," after she finishes her utterance, she looks directly at her mother with a smile (Figure 4). Hannah's observation prompts her mother to laugh and to repeat her daughter's utterance. This instance shows that Hannah has not yet fully conceptualized the device as a smart speaker manufactured by the company Google. Rather she treats it as a social agent for she has not yet gained any understanding of the concept 'smart speaker'. Although Hannah appears to correct herself as she says, 'Google,' she does not refer to the company. Rather in her understanding, it might be the name of the 'boy.' Hannah's conceptualization of the device as a boy and assigning it agency ('it slept for so long') does not match the mother's observation. This clearly shows how each party acts from a different standpoint of experiences (cf. Maturana 1988).

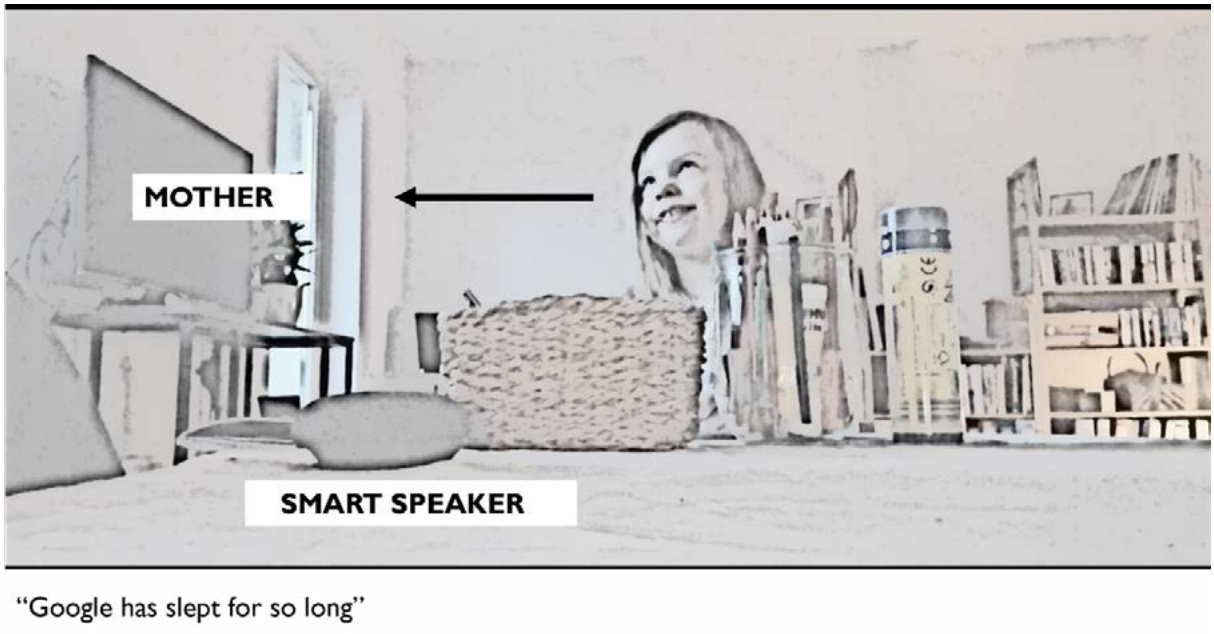

***Figure 5:** Hannah coordinating with their mother about the smart speaker's output.*

In this instance, the mother demonstrates a teaching moment by giving a clear command to the device. She begins by addressing Hannah, saying "You could say," before turning to the device and giving the command in a clear and distinct manner: "Okay, google, play the dino song." As the mother gives the command, Hannah continues to focus on her drawing. Even though Hannah does not look directly at her mother, she is sensitive to her mother's actions and engages with her in subtle ways. She reacts to her mother's utterance in two sequences: 1) after the mother has uttered "okay Google" and 2) "play the dino song." While not fully mimicking her mother's utterances, Hannah subtly mumbles the syllable 'oh' after 'Okay, Google' and the syllable 'play' after 'play the dino song.' Hannah closely observes her mother's actions and selectively chooses to which instance to give importance. Hannah has not yet gained practical understanding and intelligibility which is necessary for engaging successfully with a smart speaker. However, through mimicry and close observations of her mother's actions, Hannah learns how to verbalise the command as a constitutive action in a sociomaterial practice. After this interaction, Hannah was left alone for the next 20 minutes, during which she drew and listened to music.

4.2 Hannah's engagement with the smart speaker: How the diachronic impacts the synchronic

For about 12 minutes, Hannah drew and listened to her favourite songs on the device. Suddenly, the output changed to French, which was unfamiliar to Hannah. She noticed the change in her soundscape and decided to turn off the device by giving a command.

In the following section, Hannah uses the device independently. While focused on her drawing, Hannah notices that the device's output has switched to an unfamiliar language. Consequently, she decides to turn off the device. As she selects new crayons, she says 'Stop' (Figure 6).

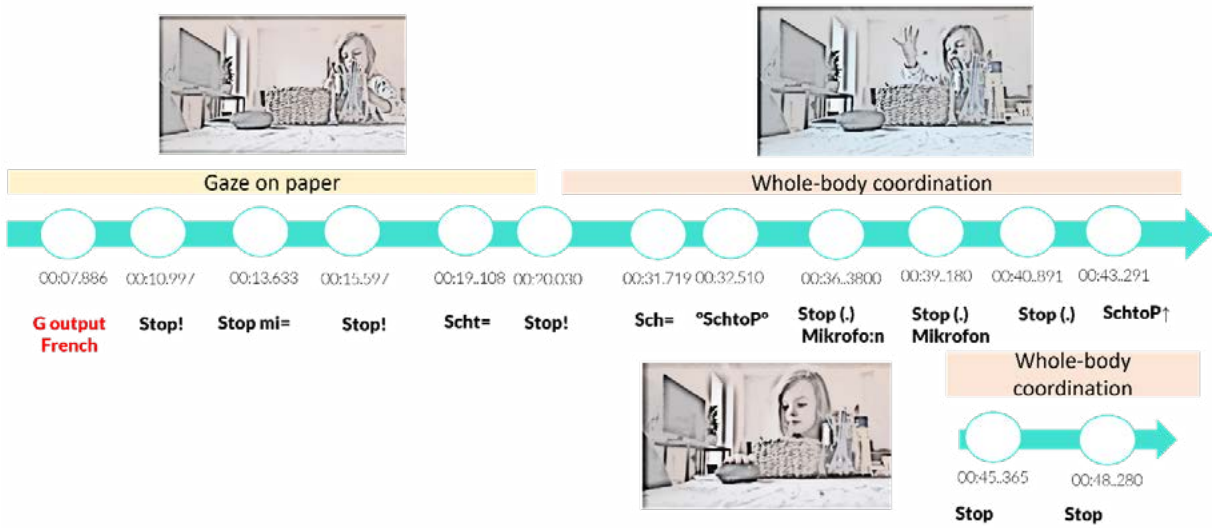

***Figure 6**: Interaction trajectory of Hannah's direct engagement with the voice assistant.*

***Figure 7:** Hannah engages verbally with the device while directing her attention elsewhere*

***Figure 8**: Hannah directed at voice assistant and enacted distinct speech-cum-gesture for 'stop'*

As Hannah says, 'Stop!', she does not look at the device. Instead, she chooses a different colored crayon. As the device does not turn off, Hannah attempts to turn it off again using her voice. After approximately 1800ms, while looking at the paper in front of her, she says, 'Stop mic=.' She clearly intends to address the voice assistant as a 'microphone.' She integrates a past event that occurred roughly 13 minutes prior, when her mother plugged in the device and the system sent the message 'the microphone is switched on again.'

Hannah's subsequent actions were thus affected by the device's output. Although Hannah was not focused on the device, she was sensitive to the device's output. About one second later, Hannah once again utters 'stop' while keeping her gaze forward. No noticeable changes occur with the device. Approximately four seconds later, Hannah's body posture changes more noticeably. In Figure 8, she directs her gaze towards the device and raises her right hand with her palm facing it.

Hannah enacts a distinct speech-cum-gesture to embody to make someone or something stop. As she engages this distinct movement, she brings about past events from her wider and narrower autobiographical history. This movement is rooted in her past experiences, particularly in kindergarten, where she learned to communicate her boundaries to signal her boundaries using this specific speech-cum-gesture, i.e., turning a flat palm to another person while uttering the wording 'stop'. The human-like voice coming from the device creates an illusion of a person (Natale 2020), which might trigger an

***Figure 9**: Hannah enacts speech-cum-gesture 'stop' during installation of smart speaker.*

experiential transition for Hannah. Perceiving the device as something with personality traits, Hannah acts from the background of her engagement with other people. This is shown when Hannah connects the wording 'stop' with a distinct hand gesture.

During the installation phase two days prior[10], the mother coordinates with the device through the 'Google Home' smartphone application and follows the instructions provided. She is instructed on how to give a command using the wake word 'Okay Google,' and can test the function by asking the device to play music. The mother directed Hannah's attention to the device, saying, 'Look, the device plays a song for you.' Hannah approached the smart speaker, which her mother held in her left hand and her phone in her right. The mother engaged with Hannah, asking her to say 'stop'. Hannah immediately made the distinct speech and gesture for 'stop' (Figure 9).

Recognizing her mistake in not guiding Hannah correctly by telling her to use the wake word 'okay, Google,' the mother instructs Hannah to say, 'Okay Google, stop.' This is one of Hannah's first experiences with giving a command to a smart speaker. While the mother, as a literate person, was able to successfully coordinate with the smart speaker through the phone application, the child must rely on her mother's guidance. When Hannah looked at the smart speaker, she saw a perhaps weirdly shaped device. However, she did not see or take the smart speaker *as* something. Thus there was, as yet, no possibility for conceptual attaching (Gahrn-Andersen 2021). Hannah's pre-predicative perceptual acts could only take shape through her mother's languaging. As her mother holds the device and, at the same time, says the wake word, she elicits a change in Hannah's conceptual perception of the device, that is, for her to 'take' the device *as* a smart speaker. However, until she fully conceptualises the device as a spoken dialogical system, much more moments of bodily dialogical coordination need to shape her conceptual perception, which enable her to treat a command as a constitutive action in the sociomaterial practice of engaging with a smart speaker.

When attempting to turn off the smart speaker independently, the child relies on past coordinated moments with her mother. This can be traced to how Hannah uses only the one-word utterance 'stop,' thus displaying a sense of giving a command. However, she still needs to gain a complete understanding of what Natale and Cooke (2021) refer to as 'the language of a computer' and Due and Lüchow (in press) in more detail as VUI-Speak. Hannah integrates several past occasions of close engagement with her mother in her relatively near past (i.e., she refers to the device as 'microphone' and uses the one-word utterance 'stop') and events from her wider past (i.e., she executes her distinct speech-cum-gesture) in her iterative attempt to engage with the device. After her spoken utterance and her speech-cum-gesture appear not to be successful in turning off the music, Hannah shifts attention now entirely to the device as she leans forward and holds her hands over the device and utters at the same time 'stop.' However, Hannah observes no change in the state of the smart speaker. As a result, she repeatedly says 'stop' and physically interacts with the smart speaker, as she touches it (Figure 10). This clearly shows how Hannah has not yet gained a conceptual understanding of the device.

Despite her best efforts, no change happens. She used the 'stop'- utterance 15 times within around 35 seconds before she accepted her fate and decided to return to her drawing. When her mother enters the room a minute later, she notices the change in output and directs Hannah's attention to it. Once again, a cooperative moment between Hannah and her mother emerges as her mother verbally guides her to successfully form

10 The video is available for viewing at the following link: https://my.hidrive.com/lnk/GDqIc19Kw.

Figure 10: *Hannah engages directly with a smart speaker through touch.*

a command to play a specific song ("Okay Google, play Hakunah Matata"). In doing so, the mother points to the voice-enabled aspect of the smart speaker and thus helps Hannah to gain a practical understanding of how to use the specific utterance of a command.

5.0 HOW ENLANGUAGED PRACTICES EMERGE

Although the idea of 'talking' to a machine often emerges from orthodox views on language that assume a consistent and stable linguistic system, close observation of how people engage with machines shows how the heterogenous nature of language contradicts any notion of a homogenous linguistic system. While the working of a speech-enabled computer system relies on linguistic stability and regularity (as outlined in section 2), this is inapplicable to the actions of human living beings. The case study focused on a three- and eight-month-old child, demonstrating that interacting with smart speakers requires practical understanding and intelligibility (Schatzki 2002). Most importantly, the child has to learn how to address the smart speaker. Mastering such expertise becomes a prerequisite for enabling a practice with these devices.

While other studies focus on how people interact with voice assistants in various daily activities (e.g., Hector 2023; Porcheron et al. 2018), this study examines *how* a child learns to engage with a smart speaker successfully through recursive engagement with others. Importantly, while adults can easily conceptually identify the device as a smart speaker (due to an ability to read and exposure to the concept through, for example, advertisement), a child who has never encountered the device lacks conceptual understanding. Gahrn-Andersen (2023b) points out how concept-infused perception determines sociomaterial practices. In order to engage with a smart speaker, one needs to know how to address the device through commands and needs to have a general understanding of the machine as a web interface (Natale, Cooke 2021). Thus, this distinct sociomaterial practice depends on the practical understanding of executing conceptually constrained actions. That is, verbalizing and constructing a command. Practical understanding undergirds ideas of regularity and stability, which, according to Schatzki (2002), also needs to be traced to the entanglement of people and artifacts, and the following of rules. A user of a smart speaker must not only know how to procedurally perform a practice, but the practice must also make sense to the user, i.e., it must have practical intelligibility.

In the case presented, gaining a practical understanding of how to command a smart speaker derives from the dialogical coordinative and recursive bodily engagement between a mother and a child. The child has not only learned to 'say' the right thing, or form a particular imperative structure, but must also synthesise pasts to grasp how the device is not contextually implemented. When the child verbalized, "Play my favorite song," it showed some understanding of how to phrase a command verbally but did not yet understand how to engage with the algorithms 'in' the machine. The child lacks an understanding of conceiving a smart speaker as web interfaces (Natale, Cooke 2021). While the direct and situated engagement with the device could be described as what Schatzki (2002) calls dispersed practice, that is, as a single action, giving a command requires, simultaneously, to engage with algorithms. The child is confronted with the unknown and intangible. Although the mother understands that a specific streaming platform supplies the music to the device, the child lacks these experiences and, thus, an understanding of engaging with algorithms. The integrative practice of engaging with a smart speaker depends on the practical understanding of engaging with a distinct kind of material artifact. Specific semiotic means, such as a human-like synthesized voice and an ascribed persona, disguise the complexity of the voice assistant system (Crawford, Joler 2018; Natale 2020; Natale, Cooke 2021), which leads to the device's anthropomorphization (Dickel, Schmidt-Jüngst 2020).. The child is therefore inclined to perceive the device as a social actor (Clark, Fischer 2023), as shown in calling it a "boy". Thus, through the recursive dialogical engagement with her mother, the child not only learns how to give a command, but also gains a practical understanding of the device. Furthermore, through the mother's guidance the child learns to conceptually perceive the device as a smart speaker. Gahrn-Andersen (2023) highlights how concept-infused perception, or conceptual attaching, grounds sociomaterial practices. Importantly, both practical intelligibility and conceptual attaching emerge out of a history of engagement with others and artefacts. The child learns to 'take' the device as a voice-enabled non-human entity as she observes and reacts to the socio-practical actions of her mother.

Although Schatzki (2002) seems to distinguish between discursive and non-discursive actions, Gahrn-Andersen (2023a) points out how practices in which no overt language is used must be understood as enlanguaged: The ability to enact these non-discursive enlanguaged actions must be traced back to recursive moments of close bodily dialogical verbal engagement with other people (e.g. a teacher). Even though the practice of engaging with a smart speaker depends on the verbal, it too exemplifies non-discursive enlanguaged doings that presuppose specific lexico-grammatical structures and a certain understanding and conceptualization of the practice. The case study highlights distinct moments of coordinated engagement between the mother and the child, where the mother verbally guided the child to use the wake word in order to be able to engage with the device successfully. Within these close coordinative moments, the child closely observes the mother not only through gaze but is also audibly sensitive and sensible to any changes in their direct physical environment (Abram 1997). One way of reacting to the mother's verbal guidance was, among others, through mimicry. Rather than repeating the 'same,' mimicry allowed the child to receive a sense of engagement with the device *for* herself.

Within this process of mimicry and imitation, the child picks out distinct aspects of her experiences of engaging with her mother. Whether learning to pass a teddy (Raimondi 2019) or engaging with a smart speaker, the child integrates past events from close dialogical coordinative moments into future actions. As an active observer, the child (and her mother) is sensitive and sensible to changes in her immediate physical environment. This showed especially how the child integrated past events from her narrower and wider past. The child's solitary engagement with the device clearly depends on integrating diachronic aspects in synchronic activity. In her attempt to turn the device off, the child, one, refers to the device as 'microphone,' thus integrating past output coming from the device and, two, enacts a distinct speech-cum-gesture for 'stop,' which originates from her past engagements with other people from a distinct social system (i.e., kindergarten). Further, using the utterance 'stop' for pausing, the voice assistant's output could be traced to a past coordinative moment between mother and child during the installation process.

Enlanguaged practices depend on the interplay of dialogicality, multiscalar temporality, and embodiment: they presuppose a world where 'language' and languaging are distributed. Thus, what appears to be stable or regular emerges from temporally evolving recursive engagement of practices through the languaging guidance of others. Practical intelligibility can, therefore, not be assigned to be an individual phenomenon but emerges through people's past engagements with others. Human living beings should be understood as zones of entanglements (Ingold 2008) and as embedded in various distinct social systems. People, therefore, act against the background of their social embedding.

6. CONCLUDING REMARKS

Due to the technological structure of a voice assistant, human participants engage with speech-enabled machines through the specific practice of giving a command. What appears to be a verbal activity and said to simulate 'talking' to a machine is a concrete practice based on specific practical understanding and intelligibility (Schatzki 2002). While literate adults can quickly learn to adapt to this practice due to written instructions given by the developers and designers, for example, smartphone applications, pre-literate children rely on their close coordinative bodily engagements with a caregiver. In this paper, therefore, the focus did not fall on the ways a child 'talks' to a machine and adapts to the technological constraints (Gampe et al. 2023) but on *how* a child *learns* through recursive bodily dialogical coordinative moments with others to engage with these specific technological constraints. Given Gahrn-Andersen's (2023a) notion of enlanguaged practices, I show that successfully engaging in situated activity with smart speaker can presuppose the integration of past events in a child's wider and narrower autobiographical and sociocultural history. Using ethnography, the paper identifies distinct coordinative moments between mother and child that add to a child's understanding of engaging in a distinct practice. Theoretically and methodologically informed by the languaging perspective, the paper traces *enlanguaging* to the interplay of dialogicality, temporality, and embodiment. The child, therefore, needs to be understood as a human living being who is, one, sensible and sensitive to changes in their immediate environment and, two, acts from the basis of their embedding in distinct social systems. This embedding emerges through recursive moments of coordination- or 'doing things together' with others (Thibault 2011; Cowley 2019).

What appears regular and homogeneous depends on previous engagement with the irregular. Once attention is paid to the heterogeneous nature of human language activity, a better understanding of human engagement with artificial intelligence may emerge.

Acknowledgements

I would like to thank Rasmus Gahrn-Andersen for his insightful comments on the manuscript of this paper and for accepting this paper as part of the special issue. This research has been funded through the Postdoc Network Brandenburg.

REFERENCES

Abram, D., 1997/2017. The Spell of the Sensuous. New York: Vintage Books.

Alač, M., Hutchins, E., 2004. I see what you are saying: Action as cognition in fMRI brain mapping practice. *Journal of Cognition and Culture*, 4(3–4), 629–661, available at: <https://doi.org/10.1163/1568537042484977>.

Barnes, B., 2001. Practice as collective action. In Schatzki, T.R., Cetina, K.K., von Savigny, E. (Eds.), *The Practice Turn in Contemporary Theory*. London and New York: Routledge, pp. 10–-23.

Barthel, M., Helmer, H., & Reineke, S., 2022. First users' interactions with voice-controlled virtual assistants : A micro-longitudinal corpus study. In: *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*.

Bateson, M. C., 1994. *Peripheral Visions: Learning along the Way*. New York: HarperCollins

Becker, A., 1999. A short, familiar essay on person. *Language Sciences*, 21(3), 229–236, available at: <. https://doi.org/10.1016/s0388-0001(98)00025-4 >.

Bender, E. M., Koller, A. 2020, July. Climbing towards NLU: On meaning, form, and understanding in the age of data. In: *Proceedings of the 58th annual meeting*

of the association for computational linguistics, pp. 5185-5198.

Barthel, M., Helmer, H., Reineke, S., 2022. First users' interactions with voice-controlled virtual assistants : A micro-longitudinal corpus study. In: *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*.

Beneteau, E., Richards, O. K., Zhang, M. et al.,2019. Communication breakdowns between families and Alexa. In: *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–13, available at: <. https://doi.org/10.1145/3290605.3300473 >.

Boersma, P., Weenink, D., 2023. Praat: Doing phonetics by computer., University of Amsterdam, available at: < https://www.fon.hum.uva.nl/praat/ >.

Clark, H. H., Fischer, K., 2023. Social robots as depictions of social agents. *Behavioral and Brain Sciences*, 46(July), available at: < https://doi.org/10.1017/S0140525X22000668 >.

Cowley, S. J., 2009. Distributed language and dynamics. *Pragmatics & Cognition*, 17(3), 495–508, available at: < https://doi.org/10.1075/pc.17.3.01cow >.

Cowley, S. J., 2011. Taking a language stance. *Ecological Psychology*, 23(3), 185-209.

Cowley, S. J., 2014. Linguistic embodiment and verbal constraints: Human cognition and the scales of time. *Frontiers in Psychology*, 5(OCT), available at: < https://doi.org/10.3389/fpsyg.2014.01085 >.

Cowley, S. J., 2019. The Return of Languaging. *Chinese Semiotic Studies*, 15(4), 483–512.

Cowley, S. J., Fester-Seeger, M. T., 2023. Re-evoking absent people: what languaging implies for radical embodiment. *Linguistic Frontiers*, 6(2),64-–77, available at: < https://doi.org/10.2478/lf-2023-0012 >.

Cowley, S. J., Gahrn-Andersen, R., 2021. Drones, robots and perceived autonomy: implications for living human beings. *AI and Society*, 0123456789, 3–6, available at: < https://doi.org/10.1007/s00146-020-01133-5 >.

Cowley, S. J., Steffensen, S. V., 2015. Coordination in Language. *Interaction Studies*, 16(3), 474–494.

Cowley, S., Madsen, J. K., 2014. Time and temporality: Linguistic distribution in human life-games. *Cybernetics & Human Knowing*, 21(1–2), 172–185.

Cowley, S., Nash, L., 2013. Language, interactivity and solution probing: Repetition without repetition. *Adaptive Behavior*, 21(3), 187–198, available at: < https://doi.org/10.1177/1059712313482804 >.

Crawford, K., Joler, V. 2018. *Anatomy of an AI System: The Amazon Echo as an anatomical map of human labor, data and planetary resources*, available at: <https://anatomyof.ai/>.

Delafield-Butt, Jonathan T. and Colwyn Trevarthen. 2015. The Ontogenesis of Narrative: From Moving to Meaning. *Frontiers in Psychology*, 6(September),1–16.

Dickel, S., Schmidt-Jüngst, M. 2021. Gleiche Menschen, ungleiche Maschinen. Die Humandifferenzierung digitaler Assistenzsysteme und ihrer Nutzer:innen in der Werbung. In Dizdar, D.,Hirschauer, S., Paulmann, J., Schabacher, G., (Eds.), *Humandifferenzierung. Disziplinäre Perspektiven und empirische Sondierungen,* Weilerswist: Velbrück Wissenschaft, pp. 342–367.

Due, B. L., Lüchow, L., In press. VUI-Speak: There is Nothing Conversational about "Conversational User Interfaces". In Muhle, F., Bock, I. (Eds.), *Social Robots in Institutional Interaction*. Bielefeld: Bielefeld University Press.

Enfield, N., 2014. Causal dynamics of language. In Enfield, N., Kockelman, P., Sidnell, J. (Eds.), *The Cambridge Handbook of Linguistic Anthropology (Cambridge Handbooks in Language and Linguistics)*. Cambridge: Cambridge University Press, pp. 319-336.

Fischer, K. 2011. Interpersonal variation in understanding robots as social actors. In: *HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, May, 53–60, available at: <https://doi.org/10.1145/1957656.1957672>

Fester-Seeger, M. T., 2024a. Becoming a Knower: Fabricating Knowing Through Coaction. *Social Epistemology*, 38 (1), 49–-69.

Fester-Seeger, M.T. 2024b. Human presencing: an alternative perspective on human embodiment and its implications for technology. AI & Society, Online first, available at: <https://doi.org/10.1007/s00146-024-01874-7>

Fox, J., 2023. Introducing Nova-2: The Fastest, Most Accurate Speech-to-Text API, available at: < https://deepgram.com/learn/nova-2-speech-to-text-api. >.

Gahrn-Andersen, R., 2023a. On the constitutional relevance of non - discursive enlanguaged doings to sociomaterial practices. *Pragmatics and Society*, October, available at: < https://doi.org/https://doi.org/10.1075/ps.22037.gah >.

Gahrn-Andersen, R., 2023b. Enacting Practices:

Perception, Expertise and Enlanguaged Affordances. *Social Epistemology*, 00(00), 1–13, available at: < https://doi.org/10.1080/02691728.2023.2261397 >.

Gahrn-Andersen, R., 2021. Conceptual attaching in perception and practice-based behavior. *Lingua*, 249, 102960, available at: https://doi.org/10.1016/j.lingua.2020.102960

Gahrn-Andersen, R., 2019. Interactivity and Languaging. *Chinese Semiotic Studies*, 15(4), 653–674, available at: < https://doi.org/10.1515/css-2019-0033 >.

Gahrn-Andersen, R., Cowley, S. J., 2017. Phenomenology & sociality: How extended normative perturbations give rise to social agency. *Intellectica*, 67(1), 379–398, available at: < http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2017-44573-016&site=ehost-live%0Ahttp://cowley@sdu.dk%0Ahttp://rga@sdu.dk >.

Gahrn-Andersen, R., Cowley, S. J., 2021. Autonomous technologies in human ecologies : enlanguaged cognition , practices and technology. *AI & Society*, 0123456789, available at: < https://doi.org/10.1007/s00146-020-01117-5 >.

Gampe, A., Zahner-Ritter, K., Müller, J. J. et al., 2023. How children speak with their voice assistant Sila depends on what they think about her. Computers in Human Behavior, 143(February), 107693, available at: < https://doi.org/10.1016/j.chb.2023.107693 >.

Giere, R. N., 2004. The problem of agency in scientific distributed cognitive systems. *Journal of Cognition and Culture*, 4(3–4), 759–774, available at: < https://doi.org/10.1163/1568537042484887 >.

Gillespie T. 2014. The relevance of algorithms. In Gillespie T., Boczkowski, P.J., Foot, K.A. (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: MIT Press, pp. 167-194.

Google, 2023. *Listen to music on Google Nest speakers and displays*, available at: < https://support.google.com/googlenest/answer/7030379?hl=en-AU >.

Gunkel, D. J., 2020. *An Introduction to Communication and Artificial Intelligence*. Cambridge, UK and Medford, MA: Polity Press.

Guzman, A. L., 2019. Voices in and of the machine: Source orientation toward mobile virtual assistants. *Computers in Human Behavior*, 90(January 2018), 343–350, available at: < https://doi.org/10.1016/j.chb.2018.08.009 >.

Guzman, A. L., 2018. What is human-machine communication, anyway? In Guzman, A. L. (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves*. New York, NY: Peter Lang, pp. 1 – 28.

Guzman, A. L., Lewis, S. C., 2020. Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media and Society*, 22(1), 70–86, available at: < https://doi.org/10.1177/1461444819858691 >.

Hector, T., 2023. Smart Speaker in der Praxis. Methodologische Überlegungen zur medienlinguistischen Erforschung stationärer Sprachassistenzsysteme. *Sprache und Literatur*, 51(2), 197–229, available at: < https://doi.org/10.30965/25890859-05002021 >.

Heidegger, M., 2010. *Being and Time*. J. Stambaugh (trans). Albany: State University of New York Press.

Hepp, A., Loosen, W., Dreyer, S. et al., 2023. ChatGPT, Lamda, and the hype around communicative ai: The automation of communication as a field of research in media and communication studies. *Human-Machine Communication*, 6, 41–63, available at: < https://doi.org/10.30658/hmc.6.4 >.

Hoffman, A., Owen, D., Calvert, S. L., 2021. Parent reports of children's parasocial relationships with conversational agents: Trusted voices in children's lives. *Human Behavior and Emerging Technologies*, 3(4), 606–617, available at: <https://doi.org/10.1002/ hbe2.271>.

Hoy, M. B., 2018. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. Medical Reference Services Quarterly, 37(1), 81–88, available at: < https://doi.org/10.1080/02763869.2018.1404391 >.

Ingold, T., 2008. Bindings against boundaries: Entanglements of life in an open world. *Environment and Planning A*, 40(8), 1796–1810, available at; <https://doi.org/10.1068/a40156>.

Jokinen, K., McTear, M., 2010. *Spoken dialogue systems. Synthesis lectures on human language technologies*. San Rafael, Cal.: Morgan and Claypool.

Linell, Per. 2009. *Rethinking language, mind, and world dialogically: Interactional and contextual theories of human sense-making*. Charlotte, NC: Information Age Publishing, Inc.

Loaiza, J. M., Trasmundi, S. B., Steffensen, S. V., 2020. Multiscalar Temporality in Human Behaviour: A Case Study of Constraint Interdependence in Psychotherapy. *Frontiers in Psychology*, 11(1685).

Love, N. 2004. Cognition and the language myth. *Language Sciences*, 26(6 SPEC. ISS.), 525–544, available at:

<https://doi.org/10.1016/j.langsci.2004.09.003>

Love, N., 1990. The Locus of Languages in a Redefined Linguistics*. In: Davis, H.G., Taylor, T.J. (Eds.), *Redefining Linguistics (RLE Linguistics A: General Linguistics) (1st ed.)*. London: Routledge, available at: https://doi.org/10.4324/9781315880273.

MacArthur, E., 2014. The iPhone Erfahrung: Siri, the auditory unconscious, and Walter Benjamin's Aura. In Weiss, D.M., Propen, A.D., Emmerson Reid, C. (Eds.), *Design, Mediation, and the Posthuman*. Lanham: Lexington Books, pp. 113–127.

Madsen, J. K., 2017. Time during time: Multi-scalar temporal cognition. In Cowley, S.J., Vallée-Tourangeau, F., (Eds.), *Cognition beyond the Brain: Computation, Interactivity and Human Artifice, Second Edition*, 155–174, available at: < https://doi.org/10.1007/978-3-319-49115-8_8 >.

Mahowald, K., Ivanova, A. A., Blank, I. A. et al. 2024. Dissociating language and thought in large language models: a cognitive perspective. *Trends in Cognitive Sciences*, available at: <. http://arxiv.org/abs/2301.06627 >.

Mallidi, S. H., Maas, R., Goehner, K. et al., 2018. Device-directed utterance detection. In: *Proceedings of the Annual Conference of the International Speech Communication Association*, INTERSPEECH, 2018-Sept, 1225–1228, available at: < https://doi.org/10.21437/Interspeech.2018-1531 >.

Maturana, H. R., 1988. Reality: The Search for Objectivity or the Quest for a Compelling Argument. *The Irish Journal of Psychology*, 9(1), 25–82, available at: < https://doi.org/10.1080/03033910.1988.10557705 >.

McTear, M., Callejas, Z., Griol, D., 2016. *The Conversational Interface: Talking to Smart Devices*. Basel, Switzerland: Springer Publishing Company.

Mühlhoff, R., 2020. Human-aided artificial intelligence: Or, how to run large computations in human brains? Toward a media sociology of machine learning. *New Media and Society*, 22(10), 1868–1884, available at: <. https://doi.org/10.1177/1461444819885334 >.

Natale, S., 2020. To believe in Siri: A critical analysis of AI voice assistants. *Communicative Figurations*, Working Paper, 32, available at: < www.kommunikative-figurationen.de >.

Natale, S., Cooke, H., 2021. Browsing with Alexa: Interrogating the impact of voice assistants as web interfaces. Media, Culture and Society, 43(6), 1000–1016, available at: < https://doi.org/10.1177/0163443720983295 >.

Natale, S., 2023. AI, Human-Machine Communication and Deception. In Guzman, A., McEwen, R., Jones, S. (Eds.), *The Sage Handbook of Human-Machine Communication*. London, UK: Sage, pp. 401-408.

Noë, A., 2004. *Action in perception*. Cambridge, MA: MIT Press.

Porcheron, M., Fischer, J. E., Reeves, S. et al., 2018. Voice interfaces in everyday life. In: *Conference on Human Factors in Computing Systems - Proceedings*, 2018-April, available at: < https://doi.org/10.1145/3173574.3174214 >.

Porcheron, M., Fischer, J. E., Sharples, S., 2017. "Do Animals Have Accents?". In: ***Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, pp.*** 207–219, available at: <. https://doi.org/10.1145/2998181.2998298 >.

Potter, M.C., 2018. The immediacy of conceptual processing. In De Almeida, R.G., Gleitman, L.R. (Eds.), *On Concepts, Modules, and Language: Cognitive Science at Its Core*. Oxford University Press, Oxford, pp. 239-248.

Poulos, C.N. 2021. *Essentials of Autoethnography.* Washington, DC: American Psychological Association.

Purington, A., Taft, J. G., Sannon, S. et al., 2017. "Alexa is my new BFF": Social roles, user satisfaction, and personification of the Amazon Echo. In: *Conference on Human Factors in Computing Systems - Proceedings*, Part F1276, pp. 2853–2859, available at: <. https://doi.org/10.1145/3027063.3053246 >.

Raimondi, V., 2019. The Role of Languaging in Human Evolution: An approach based on the theory of natural drift. *Chinese Semiotic Studies*, 15(4), 675–696, available at: <https://doi.org/10.1515/css-2019-0034>.

Rausch, D., 2023. *Previewing the future of Alexa*, available at: <https://www.aboutamazon.com/news/devices/amazon-alexa-generative-ai>.

Rouse, J., 2007. Practice Theory. In Turner, S. P., Risjord, M. W. (Eds.), *Handbook of Philosophy of Anthropology and Sociology*. Boston: Elsevier, pp. 639–- 682.

Schatzki, T. R., 2001. Introduction: practice theory. In Schatzki, T. R., Cetina, K. K., von Savigny, E. (Eds.), *The Practice Turn in Contemporary Theory*. London and New York: Routledge, pp. 10–-23.

Schatzki, T. R., 2002. *The site of the social: a philosophical account of the constitution of social life and change.* University Park, PA: Pennsylvania State University Press.

Schäfer, H., 2013. *Die Instabilität der Praxis: Reproduktion und Transformation des Sozialen in der Praxistheorie*. Weilerswist: Velbrück Wissenschaft.

Schegloff, E. A., Sacks, H. 1973. Opening up closings. *Semiotica*, 8, 289–327.

Steffensen, S.V. 2013. Human Interactivity: Problem-Solving, Solution-Probing and Verbal Patterns in the Wild. In Cowley, S.J., Vallée-Tourangeau, F. (Eds), *Cognition Beyond the Brain*. London: Springer, pp. 195–221.

Steffensen, S., Pedersen, S. B., 2014. Temporal Dynamics in Human Interaction. *Cybernetics & Human Knowing*, 21(1–2), 80–97, available at: <. http://www.ingentaconnect.com/content/imp/chk/2014/00000021/F0020001/art00007 >.

Stone, B., 2021. *Amazon Unbound: Jeff Bezos and the invention of a global empire*. London: Simon & Schuster.

Stroda, U., 2020. 'Siri, tell me a joke': Is there laughter in a transhuman future? In Hrynkow, C. (Ed.) *Spiritualities, ethics, and implications of human enhancement and artificial intelligence*. Wilmington, De.: Vernon Press, pp. 69–85.

Terzopoulos, G., Satratzemi, M., 2020. Voice Assistants and Smart Speakers in Everyday Life and in Education. *Informatics in Education*, 19(3), 473–490, available at: <. https://doi.org/10.15388/infedu.2020.21 >.

Thibault, P.J., 2020. *Distributed Languaging, Affective Dynamics, and the Human Ecology Volume I: The Sense-making Body (1st ed.)*. London: Routledge, available at: <https://doi.org/10.4324/9781351215589>.

Thibault, P., King, M., 2016. Interactivity, Values and the Microgenesis: A Distributed Cognition Perspective. In Chi-Hung,C., Ng,C.,Fox, R., Nakano,R. (Eds.), *Reforming Learning and Teaching in Asia-Pacific Universities: Influences of Globalised Processes in Japan, Hong Kong and Australia*. Singapore: Springer, pp. 173-211.

Trevarthen, C., 2011. What Is It like to Be a Person Who Knows Nothing? Defining the Actie Intersubjective Mind of a Newborn Human Being. *Infant and Child Development*, 20,119–35.

Trevarthen, C., Aitken, K.J., 2001. Infant Intersubjectivity: Research, Theory, and Clinical Applications. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 42(1), 3–48.

Waldecker, D., Hector, T., Hoffmann, D., 2023. Intelligent Personal Assistants in practice. Situational agencies and the multiple forms of cooperation without consensus. *Convergence*, 0(0), 1–17, available at: <. https://doi.org/10.1177/13548565231189584 >.

Welch, D., Warde, A., 2017. How should we understand 'general understandings'? In Hui, A., Schatzki, T., Shove, E. (eEds.), *The Nexus of practices: Connections, constellations, practitioners*. London and New York: Routledge, pp. 183 – 196.

Wittgenstein, L. 2009. *Philosophical Investigations*. Hoboken: Blackwell.

van den Herik, J. C., 2022. The reflexive roots of reference. *Language Sciences*, 89, 1–14, available at: <https://doi.org/10.1016/j.langsci.2021.101446>