Learning how to talk: Co-producing action with and around voice agents

Stuart Reeves
Joel E. Fischer
Martin Porcheron
firstname.surname@nottingham.ac.uk
School of Computer Science, University of Nottingham

Rein Sikveland
R.O.Sikveland@lboro.ac.uk
Department of Social Sciences, Loughborough University

ABSTRACT

The domestication of voice interfaces, made accessible in consumer devices such as the Apple HomePod, Google Home or the Amazon Echo, has led to everyday talk becoming intertwined with—as well as acting as—device input. Whether intending to interact with voice interfaces or not, conversationalists must learn 'how to talk' *to* and *around* them as a matter of this domestication work. Taking an ethnomethodological conversation analysis approach, this paper interrogates some of the ways in which conversationalists deploy a variety of methods so as to manage and *design input* in line with the strictures of voice interface capabilities and collaboratively accomplish—co-produce—actions with *and around* such devices.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**.

KEYWORDS

Voice interfaces, ethnomethodology, conversation analysis

1 INTRODUCTION

While our prior work has investigated how voice interfaces, through the concerted effort of their users, come to be embedded into the social life of the home and its moral order [7–9], this paper addresses some of the ways in which conversationalists co-produce and work around device-relevant talk so as to do what we are calling *input design*. In our paper we examine such input design in the context of co-production.

MuC'19 Workshops, Hamburg, Deutschland

https://doi.org/10.18420/muc2019-ws-654

2 INPUT DESIGN: AN EXAMPLE

To ground co-production we first present a transcribed example[1] of some input design practices, drawn from a corpus of recordings of in-home interactions with the Amazon Echo—a 'smart speaker' that enables access to Amazon's Alexa voice-driven 'virtual assistant' service. The Echo lets its users perform various tasks via spoken interactions, such as setting a timer, creating a shopping list, or playing music, all of which tend to be initiated as a compound [wake word + directive/question] format (the 'wake word' could also be seen acting as a kind of summons). For example here is Susan initiating and the Amazon Echo (Alexa) responding:

```
SUS  Alexa (1.1) pl(h)ease del(h)ete shopping list
ALE  you can remove an item (0.2) or clear your list in the
     Alexa app.
```

Susan (parent) is joined by the rest of her family at the dinner table including Carl (parent), Liam and Emma (both children under 10). The family has been having trouble trying to manage a shopping list captured previously by the device. Liam is beginning to jokingly 'discipline' the device for 'misbehaving'. Although Liam and the family are clearly orienting to this as a humorous moment with the Echo as a prop in the following next fragment, Liam's formulations and reformulations are instructive for understanding input design practices in the context of conversation.

```
01 LIA  Alexa? (1.7) you are- (.) going to have a time out.
02      (0.3) now:, (1.1) err=sit on the: [n- (.) in (the)]
03 ALE                                    [no      timer  ]
04      is set.
05 SUS  Oh [(0.6) uh-oh
06 ALL      [((group laughter (2.5) ))]
07 LIA  sit in the naughty corner: (0.3)
08      f' [ten ] minutes (.) Alexa,
09 CAR     [can-]
10 LIA  ALEXA (0.2)
11      sit in the [naughty corter] for ten [minutes    ]
12 CAR             [ huh huh uh   ]
13 EMM                                     [huh (.) huh]
14      huh co(h)rn(h)er hhh .hhhhh heh
15 EMM  [heh eh     ]
16 ALE  [ten minutes] (0.5) starting now
```

[1] We employ Jeffersonian transcription.

Liam produces three distinct formulations of his directives to Alexa (lines 1-2/7-8/10-11). The first attempt—which contains a range of pauses and truncations or 'disfluencies'—includes two turn-completion units (TCU); approximately: "you are going to have a time out now" and "sit on the". Partway through the second (interrupted) TCU, Alexa responds to the content of the first (the "time" keyword) with "no timer is set" (lines 3-4). The second attempt by Liam then commences, albeit without the wake word placed at the start. While the rest of the utterance is more fluent, this time Liam *appends* "Alexa" (line 8) This second attempt can also be seen as a continuation of the first in that it repeats Liam's previous disrupted turn. This then blends into a third attempt with a louder, more pronounced utterance of the wake word (line 10) and a reiteration of the now fully formulated directive to "sit in the naughty corner for ten minutes".

Liam engages in a series of repairs through formulating and reformulating his utterances [10], interactively exploring via self-repair [1] different possibilities with the device and at the same time displaying what is entailed in formulating adequately designed device input. This is all conducted in and through a shared joke (Liam's directives to the device are met with repeated, sequentially organised eruptions of laughter from the others).

3 CO-PRODUCING ACTION IN DEVICE INPUT DESIGN

For ethnomethodology and conversation analysis, the collaborative production of action is a pervasive feature of everyday social organisation [3]. Conversation analysis in particular has extensively documented how sentences in progress are coordinatedly produced by conversationalists, leading to a variety of talk phenomena such as choral co-production and other-completion of utterances (e.g., see [4–6]). Drawing from this work we focus on the ways in which voice-driven interfaces in the home present quite distinctively new *methodological* challenges for conversationalists' production *and* co-production practices, as above.

Co-production is shot through routine device actions. For instance, use of the wake word as initiator projects the next action (for example, a directive or question), and makes it available to others to also complete [4]. But unlike conversation, utterances directed towards voice interfaces are subject to a range of technological hurdles (speech-to-text transcription, lexical parsing, dialogue management, text-to-speech generation) that constrain voice input in various ways (and are largely unavailable from a users' point of view). Thus co-production practices must be adapted to fit in appropriate ways to the rigidities of these 'conversational' interfaces in order to support initiation, production and turn-by-turn interactional progressivity of the talk environment with / around the device [2].

Our paper unpacks co-production and input design practices as follows. Firstly we examine how actions with the device may be anticipated with **pre-initiations** that are formulated to project the possibility of further device-directed talk and in doing so prepare the interactional environment with others (e.g., to gather support for a particular use of the device, to create space for others to suspend their own utterances to support voice recognition, etc.). Secondly we explore the **joint production** of Alexa-directed utterances that are formulated as distinctive summons-like [wake word + directive / question] formulations. This includes a range of self- and other-repair practices [1, 10], as well as more 'competitive' types of co-production in which prosodic and other methods are employed to manage access to device input. Thirdly we investigate the utterances that are **designedly not input**, i.e., crafted by conversationalists so as to avoid accidental triggerings of the device. These kinds of utterances that take advantage of voice interface strictures to continue parallel conversations or offer support to others using the device.

REFERENCES

[1] Paul Drew, Traci Walker, and Richard Ogden. 2013. Self-repair and action construction. In *Conversational Repair and Human Understanding*, Jack Sidnell and Geoffrey Raymond Makoto Hayashi (Eds.). Cambridge University Press, Cambridge, U.K., Chapter 3, 71–94.

[2] Joel Fischer, Stuart Reeves, Martin Porcheron, and Rein Sikveland. 2019. Progressivity for Voice Interface Design. In *Conversational User Interfaces 2019*.

[3] Charles Goodwin. 1979. The Interactive Construction of a Sentence in Natural Conversation. In *Everyday Language: Studies in Ethnomethodology*, George Psathas (Ed.). Irvington Publishers, New York, 97–121.

[4] Gene H. Lerner. 1991. On the syntax of sentences-in-progress. *Language in Society* 20 (1991), 441–458. https://doi.org/DOI:https://doi.org/10.1017/S0047404500016572

[5] Gene H. Lerner. 2002. Turn-sharing: The choral co-production of talk-in-interaction. In *The language of turn and sequence*, Barbara A. Fox Celia E. Ford and Sandra A. Thompson (Eds.). Oxford University Press USA, New York, Chapter 9, 225–257.

[6] Gene H. Lerner. 2004. Collaborative turn sequences. John Benjamins, Amsterdam / Philadelphia, 225–256.

[7] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 ACM Conference on Human Factors in Computing Systems (CHI '18)*. ACM, ACM, New York, NY, USA. https://doi.org/10.1145/3173574.3174214

[8] Stuart Reeves and Martin Porcheron. 2018. Talking with Alexa. *The Psychologist* 31 (December 2018), 37–39.

[9] Stuart Reeves, Martin Porcheron, and Joel Fischer. 2018. 'This is Not What We Wanted': Designing for Conversation with Voice Interfaces. *Interactions* 26, 1 (Dec. 2018), 46–51. https://doi.org/10.1145/3296699

[10] Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53, 2 (1977), 361–382. https://doi.org/10.2307/413107