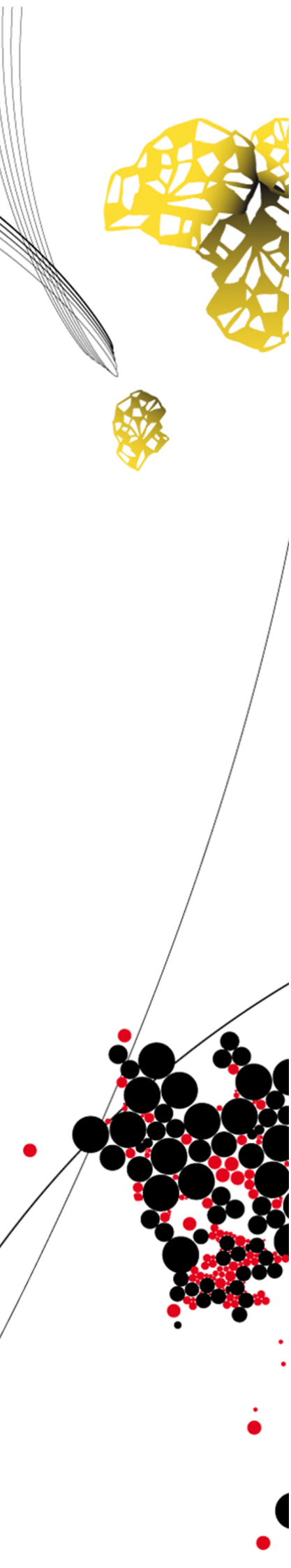

UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

Exploring the Role of Generated Backchannels in Human-CA Collaboration: Implications for Task Duration and User Perception

Roel Leenders

M.Sc. Thesis
August 2023

Supervisors:
Dr. Mariët Theune
Dr. Joyce Karreman
Dr. Myungho Lee

Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Preface

Writing a thesis proved to be a challenging, but exciting experience. Having dedicated most of my waking hours to this endeavour for the last months, I can wholeheartedly express my pride in the outcome and appreciate the academic growth it has fostered within me. Finishing this project, however, was only made possible due to the support of my supervisors, family, and friends. I'd like to especially thank Dr. Mariët Theune; Her optimism, captivating lectures, and insightful feedback were instrumental in shaping my academic growth during my Masters and throughout the duration of my thesis. Moreover, I'd like to extend this heartfelt gratitude to Dr. Joyce Karreman and Dr. Myungho Lee, whose feedback and insights proved to be invaluable.

Abstract

This thesis explores the role of backchannels (BCs) in the collaboration between humans and Conversational Agents (CAs). BCs, which are (non)verbal responses or cues that an interlocutor provides to indicate their attention and understanding in a conversation (e.g. ”uhu”, ”really?”), play a significant role in the establishment of a mutual understanding and common ground. Incorporating BCs in a human-CA collaborative context may make interactions feel more natural and human-like, which may enhance the overall collaborative experience.

First, in order to establish a thorough understanding of the field of human-CA collaboration, using a systematic literature review, we provide an overview of 1) current human-CA collaborative studies, 2) the respective collaborative models, and 3) the evaluation methods used. We conclude that, although there is an increase in popularity within human-CA collaborative research, it still remains largely unexplored. Furthermore, as most collaborative tasks are domain-specific, most studies define their own task-specific collaborative model, with no universal model or framework available. Finally, there's notable variation in the evaluation methods used across studies, mainly driven by the collaborative task's specific objectives. This causes most studies to evaluate the task performance or the user's perception of the collaboration within their specific collaborative domain (e.g. pair programming, creative games, discussions).

Subsequently, in order to conduct a user study in the context of human-CA collaboration using BCs, a BC model had to be implemented and evaluated. This model was realised using the Voice Activity Projection (VAP) model, which provides online, continuous, BC predictions using the voice activation of both the CA and the human. An initial user study, evaluating the perceived naturalness of the timing and frequency of the generated BCs, suggests that the model is capable of producing acceptable, natural-sounding BCs. However, as the study's sample size was relatively small, and the majority of the participants were Dutch, the results may not be completely generalizable and may contain cultural biases.

Finally, we used the VAP model in a human-CA collaborative user study, implementing a game to assess the influence of BCs on task duration and collaboration perception. This study was conducted with 20 participants from 9 nationalities using

a 2X1 factorial design; the BC's presence was used as the independent variable. In line with the reviewed literature, evaluation metrics were chosen based on task performance (i.e. task duration) and perceived collaboration. Results indicate that user turns were generally shorter without the presence of BCs, with the first turn being statistically significantly shorter ($p < 0.05$). A 5-point Likert scale survey showed BCs reduced the perceived CA contribution ($p < 0.001$); Other collaboration metrics (e.g. trust, working alliance, cooperation, commitment), however, were unaffected. These results may be explained by the BCs eliciting participants to speak longer due to the absence of immediate positive feedback of understanding. The difference in perceived contribution may be coupled with a shift in perceived responsibility by the participant to contribute to the collaboration. Future work, however, is necessary to refine the evaluation and better understand the inner workings of the participant's perceived collaboration.

Contents

Acronyms

AI Artificial Intelligence. vii, 1–3, 5, 7–9, 14, 16, 17, 37, 57

BC backchannel. v, 2, 3, 17, 19–33, 35–38, 46

BRP Backchannel Relevant Places. 28, 29, 31, 57

CA Conversational Agent. v, vi, 1–3, 19, 33, 35–37, 40–42, 45, 49, 52–55, 58, 59

NLP Natural Language Processing. 1

VA Voice Activation. 22–24

VAP Voice Activity Projection. v, 22–26, 31–33, 57

Chapter 1

Introduction

Adapted from Research Topics

Advancements in AI, notably in the fields of Natural Language Processing (NLP) and Deep Learning, are rapidly revolutionizing various aspects of our lives. As these technologies become increasingly sophisticated, they are being integrated more and more into our daily tasks allowing us to solve ever-increasing complex tasks with the use of AI [1]. Although some of these systems have the ability to even outperform humans on certain clearly defined tasks [2], true artificial general intelligence (AGI) (i.e. an intelligent agent which has the capacity to comprehend or pick up any intellectual skill that a human can) may still be far off [3]. Therefore, to take full advantage of AI, various researchers [3] argue that the most effective way for humans and AI to work together will be through Hybrid Intelligence, which involves combining the strengths of human intelligence and AI to achieve better results than either could alone.

This type of collaboration aligns with the idea of intelligence augmentation, which focuses on how AI can enhance human thinking and problem-solving abilities [4]. Intelligence augmentation can help overcome limitations in human reasoning, mitigate bias in decision-making, and reduce distractions during problem-solving, ultimately leading to improved task-solving abilities [4], [5]. Advances in NLP and speech processing have enabled human-AI collaboration using written or spoken natural language [4], [6]. Consequently, AI-powered CAs have seen an increase in attention within academic literature [4], [7]. During human-CA collaboration, the agent may be able to take on the role of peer, facilitator, or expert [6], [7] in order to enable intelligence augmentation.

Grounding is an essential aspect of human communication, where participants in a conversation align their understanding and establish a common ground to ensure effective communication [8]. In the context of human-CA collaboration, grounding is just as crucial as it helps bridge the gap between the human user and the AI

system [9]. Backchannels (BC), which are the minimal responses or cues that one participant provides to indicate their attention and understanding in a conversation, play a significant role in grounding [10]. Hence, incorporating BCs in AI systems may make interactions feel more natural and human-like, which may enhance the overall collaborative experience.

By investigating the role of BCs in human-CA collaboration, this thesis aims to shed light on how these conversational cues can influence the effectiveness and fluency of interactions between humans and AI systems. It seeks to explore how computationally generated BCs can impact task performance and the user's perception of the collaboration. Through this exploration, we hope to gain a better understanding of how to design CAs that foster more effective and engaging collaborations with humans. This work aspires to pave the way for more sophisticated and productive collaborations between humans and AI, ultimately leading to more capable and efficient problem-solving processes.

The thesis is structured as follows: Chapter 2 provides a comprehensive review of the current state of human-CA collaboration research, with a focus on the different models and corresponding evaluation methods. This chapter provides a groundwork for our later exploration and sets the stage for a more nuanced investigation into the role of BCs in human-CA collaboration. This chapter aims to provide an answer to the following research questions:

RQ1. What techniques and approaches are used to design and develop systems that support human-CA collaboration?

RQ2. How are these systems evaluated in terms of their success and effectiveness in studies of human-CA collaboration?

Subsequently, in Chapter 3, we delve into the world of BCs and how they can be modelled in human-CA interaction. To this end, a BC model is implemented and evaluated using a survey. In order to evaluate whether the proposed BC model can be used for a subsequent user study in the context of human-CA collaboration, we aim to answer the following research questions:

RQ3. To what extent can the timing and frequency of computational BC models be perceived as on par with human BCs?

Finally, building upon the insights from the previous chapters, Chapter 4 seeks to understand the role of BCs in a human-CA collaboration task. A collaborative game serves as the experimental platform where we evaluate whether the presence or absence of BCs has an impact on the perceived collaboration by participants and the duration of task completion. This chapter aims to provide answers to the following research questions:

RQ4. To what extent can CA BCs affect the task duration during a Human-CA collaborative task?

RQ5. To what extent can CA BCs affect the perceived collaborative fluency during a Human-CA collaborative task?

This thesis aims to contribute to the growing body of human-AI collaboration research by providing a comprehensive investigation into the role of BCs in this context. Through systematic analysis and experimentation, it offers valuable insights that could aid in designing more effective interactive systems, and hopefully, pave the way for more nuanced and beneficial human-CA collaborations in the future.

Chapter 2

Literature Review

Adapted from Research Topics

As mentioned in the previous chapter, Conversational Agents (CAs) have become increasingly popular in various fields. Their collaboration with humans is a novel and promising area of study, which offers opportunities to improve a range of tasks. Although various literature reviews have made an attempt at making the field of human-CA collaboration more accessible [4], [11], none have provided an overview of the methods and algorithmic models used to enable human-CA collaboration. Additionally, a clear overview of the methods used to evaluate these models is also lacking; Poser and Bittner [4] conducted a systematic survey focussing on teamwork-specific psychological concepts for the design of CAs, while Memmert and Bittner [11], conducted a survey on human-AI collaboration in the context of problem-solving. Therefore, in an attempt to make insights from human-CA collaboration research more accessible, this chapter aims to answer the following research questions:

RQ1. What techniques and approaches are used to design and develop systems that support human-CA collaboration?

RQ2. How are these systems evaluated in terms of their success and effectiveness in studies of human-CA collaboration?

By investigating the methods used in the development and evaluation of human-CA collaboration models, this chapter aims to provide a theoretical basis for future investigations conducted in this thesis. Our findings could potentially guide researchers and practitioners in developing and assessing their own human-CA collaboration systems.

This chapter is structured as follows: Section 2.1 covers the theoretical background and establishes the definitions that will be used as the basis for the remaining thesis. Section 2.2 will elaborate on the selected databases, keywords, and

search methodology used during the systematic review. The reviewed literature corresponding to the research questions will be presented in Section 2.3. Finally, in Section 2.4, the results and limitations will be discussed.

2.1 Background

2.1.1 Conversational Agents

Before we can understand the role of CAs in human-CA collaboration, it is first important to properly define the term conversational agent. According to Allouch et al. [12], a CA is a type of dialogue system that enables the user to communicate with a system using full natural language sentences. What differentiates a CA from any other dialogue system, is its capability to both understand and generate natural language using verbal (e.g. text and voice) methods of communication. Hence, dialogue systems can merely produce natural language text or speech, while CAs also have the ability to understand and respond to utterances from the user. Additionally, a CA can use nonverbal methods (e.g. face and body language) to increase user engagement [13]. Hence, dialogue systems that require the user to input a specific number or word to progress through the menu are not considered CAs because the user's input does not consist of full sentences. Similarly, systems that allow the user to provide input through a limited set of pre-programmed commands while not invoking a natural language response are also not considered CAs.

There are various ways to categorize CAs, including the way they communicate with users, the tasks they can perform, and the specific domain or application they are used in [12]. For this literature study, the definition of CAs is solely defined according to the method of communication with the user. This is necessary as the terms used to describe CAs within literature vary based on their method of communication. CAs that only communicate with users through text-based methods, such as ELIZA [14] are called text-based CAs or chatbots. CAs that can interact with users through voice, such as Siri or Cortana, are called voice-based CAs. Embodied CAs, are CAs that have a virtual or physical body in addition to voice recognition and speech generation abilities [14]. They can also communicate through facial or body gestures. Literature uses multiple terms to identify both types of embodied CAs; ranging from virtual avatars to digital humans for virtual-based agents to social robots for physical-based agents [12]. Please see Figure 2.1 for an overview of the CAs categorized by method of communication.

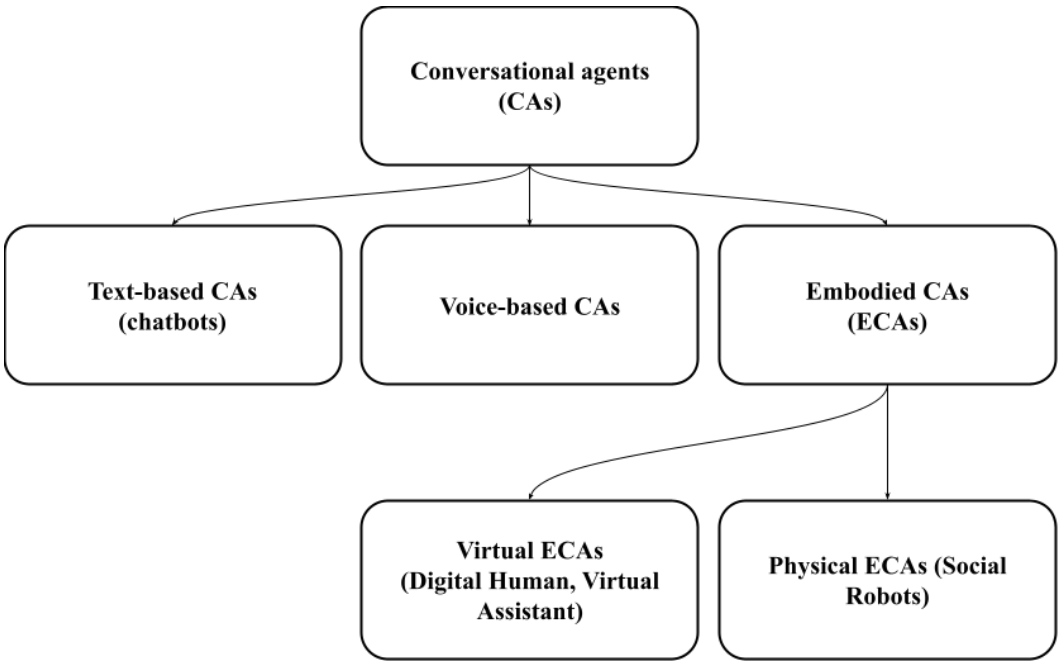

Figure 2.1: Various CAs categorized by method of communication adapted from [12]

2.1.2 Human-AI Collaboration

Before we can define conversational agents in the context of human-AI collaboration, it is necessary to first define what collaboration between humans and AI entails. According to various researchers, human-AI collaboration refers to the concept of humans and intelligent systems working together to achieve a shared goal or task [3], [11], [15]. In other words, it refers to the idea that humans and AI can complement each other's strengths and weaknesses. AI systems are capable of completing specific tasks efficiently, handling large amounts of information, identifying patterns, and generating logical predictions [3]. In contrast, humans have the capacity for common sense and other emotional traits like empathy and creativity [3]. Moreover, humans can more easily adapt themselves to new environments or deal with unexpected events [3], [15]. Additionally, we have the ability to handle incomplete information to solve complex, abstract issues [16].

Although multiple studies have researched human-AI collaboration, there still seems to be a lack of consistency regarding the definition of collaboration [15], [17]. Additionally, since this field of research is still relatively new, the topic has been studied under a variety of terms (e.g. human-AI teaming or hybrid intelligence) [11]. Bedwell et al. [18] conducted a thorough literature review in which they identified the various aspects of collaboration. According to their definition, collaboration can be understood as an evolving process, which involves the active participation of two or more social entities in joint activities, that aim to accomplish at least a single shared

goal [18]. Collaboration is an evolving process in the sense that it is not deterministic by nature. In other words, the outcome of the process is continually influenced by the emerging mental states of the collaborators (e.g. values, motivations), collaborative behaviour (e.g. reasoning, problem-solving), and the environment in which the activity takes place [18]. Social entities within the context of collaboration can be considered as individuals, teams, and organizations. This is also what differentiates teamwork from collaboration; while teamwork solely happens between individuals, collaboration can also happen between collective entities. Moreover, collaboration is reciprocal in the sense that the entities involved are proactive and mutually engaged. An AI system that solely gives the user recommendations whenever the user asks for it, can not be considered collaborative. Since the review conducted by Bedwell et al. [18] is widely cited, it will be used as the foundation for the understanding of collaboration for this study.

Finally, with regard to the definition of human-CA collaboration in specific, collaboration in this context can be defined as an evolving process, which involves the active participation of social entities - of which at least one entity can be considered a conversational agent - with the aim to achieve at least a single shared goal. Using this definition, the literature search keywords will be determined which will be described more elaborately in the following section.

2.2 Methodology

Multiple literature searches were conducted using the ACM and IEEE Xplore digital libraries. The literature searches were conducted in an iterative manner during which previous findings were used to alter the search queries with new relevant search terms. During the first iteration, the following search query was used to search in full article texts without any additional filters: *("Conversational Agent*" OR "Virtual Human*" OR "Digital Human*" OR "Virtual Avatar*" OR "Virtual Agent*" OR "Virtual Assistant*") AND ("Problem Solving" OR "Problem-solving" OR "Human-AI collaboration" OR "Creative problem solving" OR "Creative problem-solving")*. See Table 2.1 for an overview of the results.

Database	Initial hits	Abstract review	Full review
ACM	421	18	8
IEEE	31	6	1
Total	452	24	9

Table 2.1: Search results for the first review iteration.

Database	Initial hits	Abstract review	Full review + f/b
ACM	43	12	3
IEEE	98	7	3
Total	141	19	10

Table 2.2: Search results for the second review iteration and forward/backward (f/b) search. Four additional papers were found using f/b search.

During the full paper review, it quickly became apparent that various interchangeable keywords related to human-AI collaboration were missing (e.g. hybrid intelligence and human-AI teaming). Moreover, cooperation and coordination were also frequently used instead of collaboration. Therefore, an additional search was performed with the following keywords as a possible substitute for *human-AI collaboration*: *"team*" OR "coordination" OR "cooperation"*. To keep the number of results maintainable, the filters were adjusted to only search in the title and abstract. See Table 2.2 for the results of the second search iteration. During the second iteration, 7 duplicate papers were found and excluded.

To synthesize the search results, the relevancy of the articles was first analyzed by their title and abstract. An article's relevance was determined by their proposed type of human-CA collaboration and its correspondence to the definition mentioned in Section **??**. Subsequently, a full review of the article was conducted whenever an article was deemed relevant. Finally, a forward and backward search was performed to identify additional relevant studies. See Figure 2.2 for the process flow of the article selection method.

2.3 Results

This section provides an overview of the reviewed literature with the aim to answer the aforementioned research questions. The results are grouped by domain, as the models and evaluation methods vary significantly per type of collaboration (see Figure 2.3).

2.3.1 Collaborative Models

Group Discussions

Most of the human-CA collaboration studies found during this review revolve around the augmentation of group discussions. Do et al. [19], for example, studied how a chatbot can facilitate consensus reaching, encourage an even contribution by the

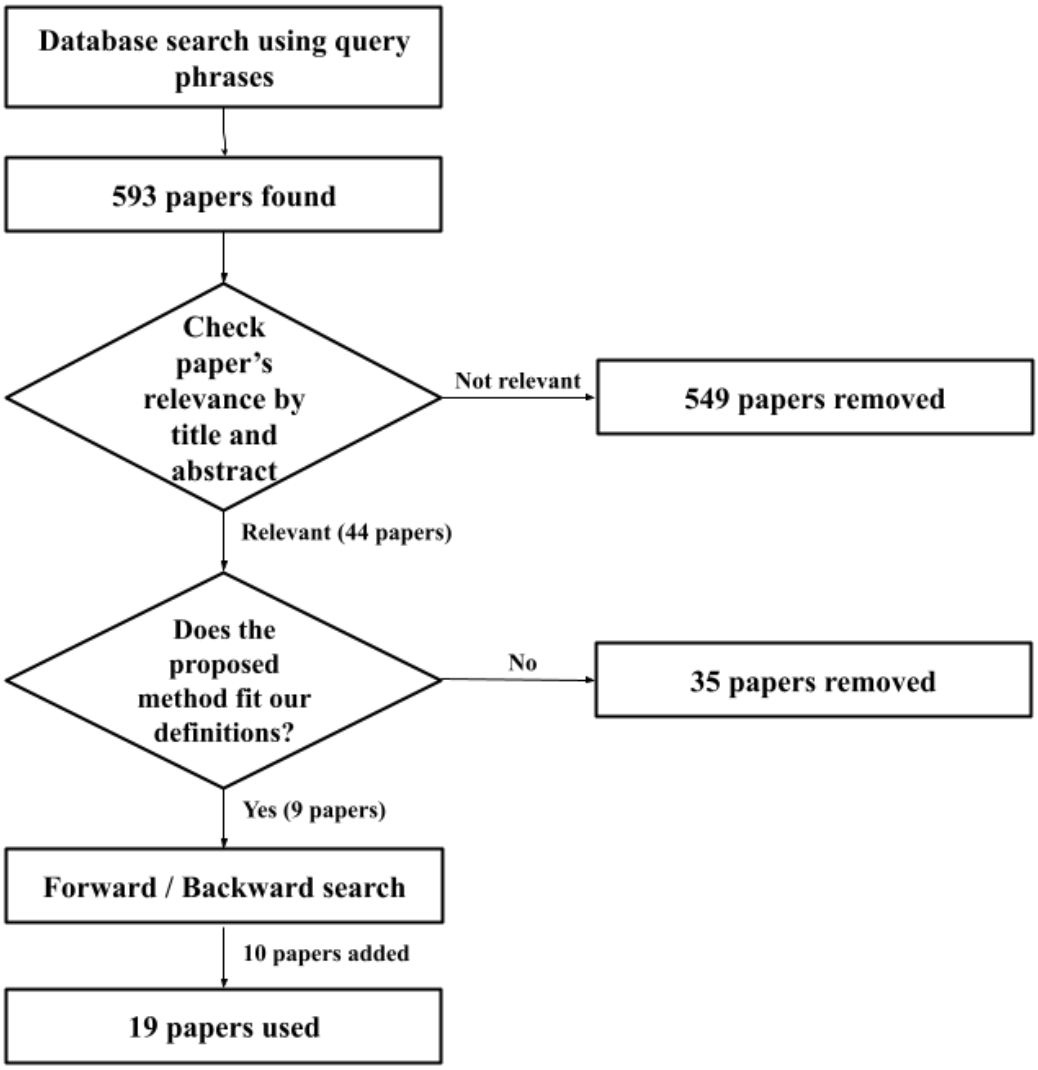

Figure 2.2: Article selection process flow for both iterations combined.

participants and aid in organizing various opinions during a discussion. Most of the tasks performed during these studies focussed on decision-making tasks. These tasks range from planning a day trip [20] to ranking patients waiting to receive a heart transplant [21]. Studies also frequently used open-debating, estimation, and problem-solving tasks during which participants had to either discuss ethical issues [22], estimate e.g. the height of the Eiffel Tower [20], or create an advertising slogan [19] respectively.

Several methods and collaborative models were applied to enable the CA to augment group discussions. Kim et al. [22] based their chatbot on principles established in prior work from psychology [23]–[25]. More specifically they used the *think-pair-share* strategy to encourage collaborative discussions by facilitating independent opinion formation and understanding the perspectives of others. As a result, the chatbot supports these principles during the discussions by encouraging equal par-

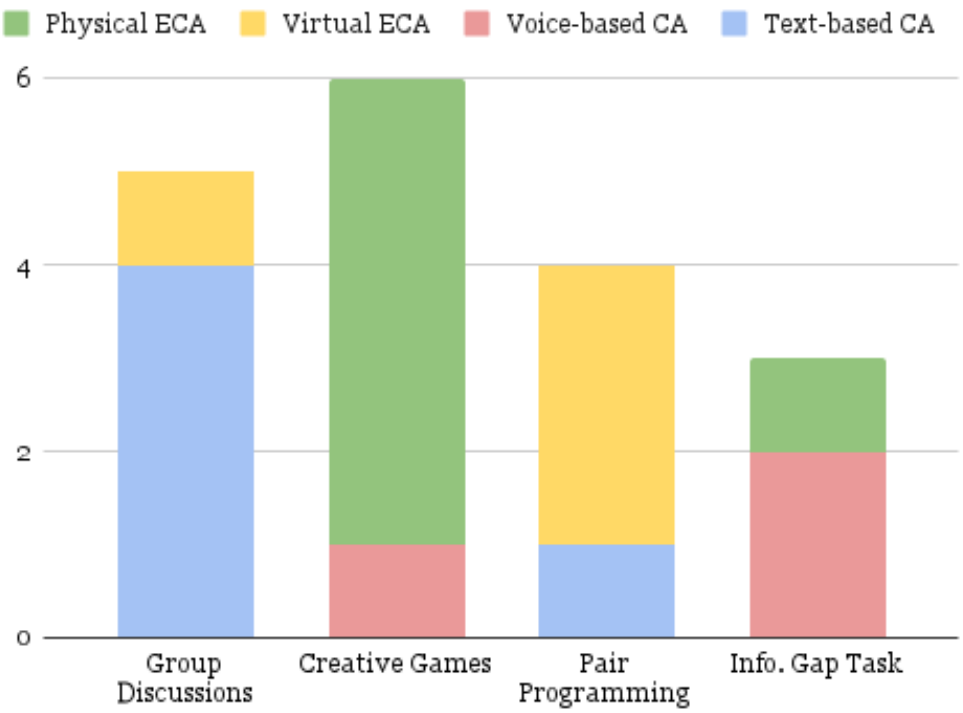

Figure 2.3: Overview of the types of CA used per domain by the reviewed literature

ticipation from all the participants. Bagmar et al. [21] partly designed their chatbot based on principles from Group Decision Support Systems (GDSS) [26]. GDSS explores the idea of anonymity in group discussions and its impact on the contributions and feedback made within those groups. Hence, the chatbot integrates these ideas by guiding the conversation using the occasional interruption; these interruptions can be directed to the entire group or privately to an individual.

Moreover, Kim et al. [20] conducted a specific need-finding study to discover the traits a chatbot should possess to augment group discussions. Using the results they develop a chatbot that aims to: 1) efficiently derive consensus within a given time; 2) encourage even contributions by asking specific participants to speak up; and 3) organize both the individual and the whole group's opinion by summarizing the main keywords. Subsequently, they modelled the cooperative behaviour of the CA based on the quantity and quality of the contribution of group chat participants by periodically computing the standard deviation of the number of messages and the number of unique words for each member. Haring et al. [27] implemented a virtual ECA to aid the discussion between military personnel during mission debriefs. They used *discuss*, *debate*, and *open communication* strategies of conflict resolution [28] as guidelines for their ECA. Finally, Do et al. [19] implemented a text-based CA that used communication strategies intending to aid participants with decision-making

tasks during chat discussions. These strategies include 1) messages sent to two types of recipients (i.e. @user or @everyone); 2) messages sent in public/private channels; 3) messages sent in which the CA asks a peer to aid an under-contributing member.

Creative Games

Various papers found during this review also studied the impact of human-CA collaboration on creativity. During these studies, using the assistance of either voiced or physically embodied CAs, participants were asked to perform a plethora of creative tasks. These tasks ranged from performing various drawing activities [29]–[32], to playing a creative video game [33], or zen rock gardening [34].

The methods and collaborative models used varied per study. To study the effects of human-CA collaboration on creativity during zen gardening, Kahn et al. [34], based their physical ECA on principles from Interaction Pattern Design in Human-Robot Interaction (HRI) literature [35]–[37]. These principles mainly refer to how a participant can be introduced to a given task. Additionally, they developed 10 additional interaction patterns that aimed to foster creativity during the given task. To evaluate whether these patterns work, Kahn et al. [34] conducted a Wizard of Oz study [38] during which the ECA was controlled by one of the researchers. Results demonstrate that participants using the ECA engaged in the creative task longer and provided around twice as many creative expressions compared to the participants who didn't use the ECA.

Multiple human-CA collaboration studies use and build upon the patterns provided by Kahn et al. [34]. Devasia et al. [33], for example, use '*Pushing the Limits*', '*Validate Decision*', and '*Condier the Alternative*' patterns to engage children during a creative problem-solving task. While interacting with a physical ECA, children had to play a scaffolding game on a tablet during which they had to reach an objective by building various contraptions. The ECA acted as a collaborative peer by demonstrating various scaffolding solutions, asking about possible alternative solutions, and encouraging the use of varied objects. Another human-CA collaboration study [29] implicitly used the '*Validate Decision*' and '*Reflect on Intuition*' patterns to develop a voiced CA with the aim to foster creativity in children during a drawing game. Additionally, they used collaborative strategies to overcome writer's block [39], [40] by allowing the CA to suggest and generate new drawings. Moreover, based on theories of embodied cognition [41], they aimed to foster creativity by enabling the CA to respond and generate drawings based on real-time tellings by the child. Finally, multiple studies [30]–[32] use findings from psychology and HCI [42], [43] which indicate that children's creativity is influenced by external factors (e.g. collaboration,

reflection and question asking). Additionally, they use Boden's [44] framework of creativity to design creative behaviours by the CA during gameplay

Pair Programming

Another collaborative task performed during various human-CA collaboration studies was pair programming. Using the aid of a CA, participants were given the objective to solve existing security vulnerabilities [45], or to program the game *tic-tac-toe* [46]–[48]. Kuttal et al. [47] conducted a pilot study to analyze the creative problem-solving strategies and conversational styles used during human-human pair programming sessions. Subsequently, they recommend transferrable guidelines from human-human to human-CA collaboration. These guidelines are based on a driver/navigator collaborative model during which one individual actively programs, while the other reviews the code, makes suggestions, and asks questions for clarification [49].

Finally, in a similar study, Robe et al. [46] used Shneiderman's guidelines [50] and Nielsen's heuristics [51] with the aim to create suitable dialogue options. Due to the social complexity of pair programming, all studies were conducted using a Wizard of Oz method.

Information Gap Tasks

Lastly, various studies applied CAs during information gap tasks [52], [53]. Simpson et al. [53] created a virtual reality environment in which a voiced CA had to guide multiple players to hidden targets [53], [54]. The responses of the CA were identified using prior literature [55]–[57] and using a previous study during which they analyzed human responses in the same context [58]. As a result, the CA's dialogue can be classified into three categories: 1) *task action directives* (e.g. giving direct commands to the players); 2) *information exchange* (e.g. sharing information about the environment); and 3) short, close-ended *responses*.

Moreover, a study conducted by Kontogiorgos et al. [59], explores the effect of CA embodiment and failures during information gap tasks. More specifically, participants were instructed to prepare various meals using recipes provided by a CA. The researchers were particularly interested in whether the participants used different strategies to (re)establish common ground with the CA during the task, depending on the CA embodiment and induced failures. To enable human-CA collaboration during the experiments, Kontogiorgos et al. used a wizarded CA. The collaborative model used during the study involved the following specific predefined dialogue options:

1. **Next Instruction**: This was used when the user had completed the current step or specifically asked for the next ingredient in the recipe.

2. **Clarification Answers**: If participants sought clarity on any aspect of the task, the CA could provide detailed information. Examples include answering questions about the location or identity of an ingredient, specifying quantities, or giving simple affirmative or negative confirmations.

3. **Repeat**: The CA had the option to repeat the previous instruction for the benefit of the user.

4. **Incorrect**: Whenever the participants chose the wrong ingredients, the CA could alert and correct them.

This structure ensured a streamlined and focused interaction between the participants and the CA, while still allowing for dynamic responses based on real-time user needs and actions.

2.3.2 Evaluation Methods

Among the studies that included an evaluation, the most occurring experimental setup was the mixed factorial design experiment. These experiments compare and evaluate the difference between pre-defined dependent variables (e.g. quality of the task output) based on multiple subject conditions (e.g. with CA or without). An overview of the qualitative and quantitative measures found in this review is provided in Table 2.3. The measures are grouped by collaborative task (i.e. group discussions, creative games etc.), as they vary significantly per context. Finally, two dominant evaluation categories were discovered among the various collaborative tasks; although not explicitly mentioned, all studies either evaluate the user's perception of the CA or the effect on the task performance, or both.

2.4 Discussion & Limitations

Due to advancements in deep learning, natural language processing, and AI, human-AI collaboration will become ever more prevalent as it opens the door for intelligence augmentation. As demonstrated during this study, various studies already pave the way for the application of CAs in this context. However, due to the various kinds of collaboration domains (e.g. group discussions, pair programming), and the social complexity of collaborative tasks in general, human-CA collaboration has not yet established general guidelines and still seems to be in its infancy; the reviewed studies

Task	Category	Quantitative Measures	Qualitative Measures	Used by[a]
Group Discussions	User Perception	Likert scale surveys were conducted for perceived *competence* [60], *usefulness* [61], [62], *rapport* [63], *intrusiveness* [64], and *embarrassment* [65]; The *quality of the responses* provided by the CA [66] and the *overall opinion* of the CA [67] were evaluated.	Survey with open-ended questions, reviewed by multiple researchers using qualitative content analysis [68] to extract themes.	[19], [20], [27]
	Task Performance	Self-reported post-experiment surveys we conducted that consisted of ratings of *team collaboration*, *perceived opinion alignment*, and *communication efficiency* [69]; The number of messages, number of senders, and sending times were used to analyze the *contribution of individual participants* and *group behaviours*.	Multiple researchers rated the group's output (e.g. slogan) on a five-point scale according to its *usefulness* and *uniqueness* [70]	[19], [22], [27]
Creative Games	User Perception	-	Semi-structured interview to evaluate the usability of the CA; Open questions during the post-questionnaire to study user engagement	[29]
	Task Performance	The participants' *baseline creativity* was assessed using the Torrance Test of Creative Thinking (TTCT) [71]; The participants' *fluency* (i.e. the number of distinct combinations attempted), *flexibility* (i.e. variance in the demonstrated play styles), *originality* (i.e. the number of unique combinations compared to other participants), *debugging skills* (i.e. the number of different solutions tried after an initial failed attempt) were evaluated; During storytelling tasks, the participant's *writing comprehension* was tested [72].	Expert ratings based on the creative outcomes (stories and drawings) [73]; Semi-structured interview with each participant to evaluate creativity using the coding and analysis of creative actions [74], [75]	[29]–[32]
Pair Programming	User Perception	Self-efficacy questionnaires and discrete answers to interview questions	Semi-structured interviews, transcription and analysis using Corbin and Strauss variant [76] of Grounded Theory [77]; Asked participants whether they would prefer the CA over an Internet search during programming.	[46], [48]
	Task Performance	Counting the number of vulnerabilities/bugs successfully fixed by the participants	-	[78]
Info. Gap Task	User Perception	-	-	-
	Task Performance	The participant's proportional gaze to the agent, nr. of conversational turns, nr. of clarification questions, nr. of user acknowledgements, and total interaction time was recorded.	-	[]

Table 2.3: Overview of the qualitative and quantitative measures found in the reviewed human-CA collaboration literature

[a]Used by at least one of the mentioned papers.

use vastly different collaborative models and a significant amount still use Wizard of Oz methods to control the CA. However, as the papers found using the systematic review were mostly published in the last 3 years, human-CA collaboration research seems to grow in popularity.

Furthermore, researchers have to be critical about the type of CA they use and whether to actually use CAs or not, as some collaborative contexts may be better suited for different kinds of AI systems. Although, for example, the pair programming studies found during this review evaluate important aspects, it has yet to be seen if a virtual ECA actually augments programming capabilities and is perceived as useful by the target audience. Meanwhile, collaborative language models like Github Copilot[1] are being used by industry professionals for programming tasks. It has to be noted, however, that the interaction capabilities with Copilot are limited, and a more conversational style of interaction may result in other collaborative benefits [79].

Regarding the evaluation methods used during the human-CA collaboration studies, two evaluation categories were discovered among the various qualitative and quantitative measures; studies evaluated the user perception of the CA and/or the effect of the CA on task performance. However, as each type of collaboration requires different collaborative models and aims at different types of outcomes, the measures and metrics used between the studies vary significantly. While studies regarding creative games focus more on the evaluation of task performance and use creativity metrics established in prior work from social psychology, pair programming and group discussion studies focus more on the evaluation of the users' perception with the aim to develop collaborative guidelines for future research. Regarding the similarities between the metrics used to evaluate user perception between the various domains, domains with more linear tasks (i.e. pair programming and group discussions tasks) seem to focus more on the evaluation of the perceived usefulness of the CA, while non-linear tasks (i.e. creative games) focus more on the perceived engagement.

Finally, partly due to the novelty of human-CA collaboration research and the multitude of terms used to describe it, various additional keywords were found that were not used during the literature search (i.e. *intelligence augmentation*, *social robots*). This is a limitation and, as a result, may require additional search iterations. Furthermore, some studies were excluded from this review due to the incongruence with our definition of collaboration. Studies that used CAs solely for teaching tasks, for example, were excluded from this study as it was not deemed collaborative; although debatable, teaching was not deemed collaborative as the goals between the teacher and student differ. Hence, the used definition of collaboration may still be too restrictive.

[1] Github Copilot: https://github.com/features/copilot

2.5 Conclusion

This chapter covered a systematic analysis of the collaborative models and evaluation methods used in human-CA collaboration research. According to our findings, there is no single model that accounts for all forms of collaboration; instead, studies either build models based on user preference studies or by drawing on prior social psychology research. Additionally, the users' perception of the CA and the CA's impact on the performance of collaborative tasks are the key metrics used to assess human-CA collaboration studies. To conclude, as AI gets increasingly more powerful, human-AI collaboration gets ever more prevalent. The findings of this study contribute to future research by providing an overview of current human-CA collaboration research.

As we transition into the next chapter, we will delve deeper into the nuances of human-CA collaboration, concentrating on an important yet underexplored aspect: backchanneling. Building on our understanding of collaborative models and evaluation methods, we will explore how BCs, as integral components of conversational dynamics, can be modelled in human-ca interaction. In this chapter, we aim to assess the extent to which computational BCs can emulate the naturalness of human BCs. This is crucial in order to conduct a human-ca collaborative user study, utilizing computational BCs (see Chapter 4). In doing so, we continue our pursuit of understanding and enhancing human-CA collaboration.

Chapter 3

Modelling Backchannels

The previous chapter contained a comprehensive analysis of the collaboration models and evaluation methods in human-CA collaboration research. It highlighted the diversity of models and the lack of a standard approach, underscoring the importance of context-specific studies and approaches. Given the wide range of metrics used to evaluate human-CA collaboration, user perception and task performance were identified as key indicators. Building on this analysis, the next chapter (Chapter 4) will use this as a theoretical foundation to design a human-CA collaborative user study. However, in order to conduct this experiment, it is first necessary to achieve a better understanding of the concept of backchannels (BC), and how they can be modelled for human-CA collaboration. BCs refer to short utterances or non-verbal signals that signify active listening, comprehension, and encouragement, such as ”mm-hmm,” ”okay,” or nodding. Although they are considered to play a pivotal role in effective communication and engagement in collaboration [10], no study has evaluated their impact on human-CA collaboration. Hence, this chapter aims to explore the modelling of BCs, and assesses whether their perceived naturalness can be effectively used in a human-CA collaborative user study. The guiding research question for this investigation is:

RQ3. To what extent can the timing and frequency of computational BC models be perceived as on par with human BCs?

The remainder of this chapter is organized as follows. Section 3.1 provides background information on grounding and common ground in conversations, as well as an overview of BC models. Section 3.2 presents the methodology used in this study, including the implementation of the used BC model and the survey conducted to assess the perception of generated BCs. Section 3.3 presents the results of the study, analyzing the generated BCs and survey responses. Section 3.4 discusses the findings, limitations, and implications of the study. Finally, Section 3.5 concludes the chapter.

3.1 Background

3.1.1 Theory of Grounding

In order to understand the importance of BCs in a conversational context, it is necessary to first understand the notion of both grounding and common ground. The term ”common ground” was first introduced by Stalnaker [8] and refers to the establishment of a shared understanding between multiple people during discourse. This notion uses a number of related ideas, including the concept of joint knowledge [80], mutual knowledge or belief [81], and common knowledge [82]. Common ground can be regarded as the confluence of these principles and is a vital part of successful communication between people [83].

According to Clark [84], common ground can be categorized into four different types; *communal*, *specialised*, *personal*, and *local* common ground. First of all, *communal* common ground refers to the idea that a shared understanding can be established among larger groups of people that belong to the same community (i.e. people sharing the same faith or nationality). *Specialised* common ground can be found amongst people that share a specific area of expertise or interest, such as friends or colleagues [9]. Moreover, *personal* common ground occurs between two interlocutors and can be defined as the collection of shared propositions between the individuals. Finally, as an element of *personal* common ground, *local* common ground can be understood as the shared understanding belonging to a piece of information obtained during discourse with a specific interlocutor. Clark [84] describes this type of information as concrete observations, such as the opening hours of a store or the price of a specific item.

To establish a common ground, individuals use the communicative technique called ’grounding’. Grounding occurs during dialogue whenever interlocutors attempt to update their shared understanding with new propositions [9]. According to the grounding model proposed by Clark and Schaefer [85], grounding is established during dialogue through communicative contributions. These contributions can be divided into two different phases (see Figure 3.1). First of all, during the *presentation phase*, interlocutor *A* presents an utterance to interlocutor *B*. *A* does so based on the assumption that as long as *B* does not give any strong evidence of the contrary, *B* understands what *A* is saying. Subsequently, during the *acceptance phase*, *B* responds to *A*’s utterance by giving the appropriate amount of evidence of understanding. *B* does so on the assumption that once *A* registers the evidence, *A* will believe *B* understands the utterance.

During the acceptance phase of the grounding process, communicative feedback is used to provide evidence of whether the interlocutor accepts or refuses the

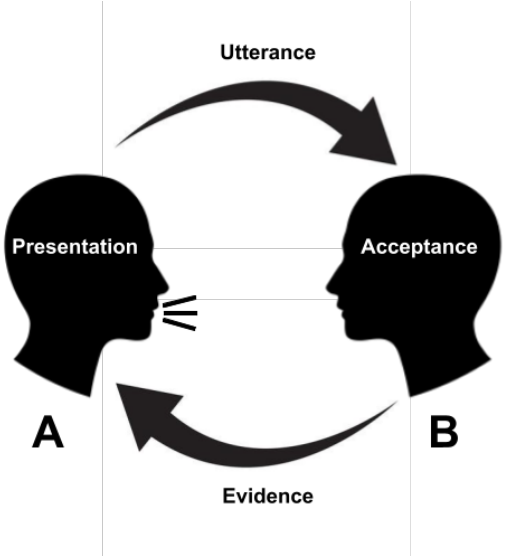

Figure 3.1: Overview model of grounding as described by Clark and Schaefer [85].

utterance [9], [86]–[88]. Research suggests there are five different types of positive evidence of understanding [83], [85]. First of all, an individual can provide evidence by displaying various social signals that indicate they are paying attention (e.g. appropriate eye gaze). Whenever a speaker may feel like they lost the attention of the person they are talking to, the speaker may use phatic utterances to obtain additional evidence of understanding (i.e. ”Do you get what I mean?”). This category of positive evidence is called *'continued attention'* and is often considered the most basic form.

The second category, *'assertions of understanding'*, establishes evidence by providing various acts of acknowledgement. This is generally paired with verbal *assessments* (e.g. ”are you serious?”, ”oh really?”) or *BCs* responses (i.e. ”okay”, ”yes”, ”uhu”). Third, *'presuppositions of understanding'* establishes evidence whenever the listener introduces a new topic which is relevant to the previously discussed topic. Next, *'displays of understanding'* occur whenever the listener construes part of the speaker's intention behind their utterance. Finally, *'exemplifications of understanding'* occur whenever the listener exemplifies whatever they have construed the speaker to have meant. More specifically, the listener can use paraphrasing or verbatim repetition of the speaker's utterance to provide evidence. Moreover, they can display sadness, disappointment, or any other empathic/iconic gesture that makes the speaker feel understood.

Since grounding is relatively broad in terms of its scope, this thesis focuses on a single aspect of grounding, which is vocal BCs. Hence, in the following subsection, we will focus on BCs and how they can computationally be modelled.

3.1.2 Backchannel Models

Regarding computational BC models, early models were mainly offline and used, among other methods, decision trees based on prosodic and syntactic part-of-speech features to predict BC relevance places (BRPs). BRPs are considered points in a conversation where one of the listening interlocutors may provide a BC [89], [90]. For example, Cathcart et al. [91] used pause durations in combination with an n-gram model to predict appropriate BCs behaviour. Generating real-time online BCs, however, seems to be an even bigger challenge; although complex models generally provide higher accuracy, they also require more computing power, which makes the timing of feedback increasingly difficult. Meena et al. [92], for example, trained a real-time BC prediction model using prosodic and lexico-syntactic features using automatic speech recognition. Moreover, various studies trained probabilistic sequential models that continuously predict the probability of whether a BC could occur within a given time frame [93]–[95]. Recently, deep learning methods are being used more frequently to predict the timing of appropriate BCs. For example, Hussain et al. [96] trained a deep Q-network for BCs during human-robot interaction. Ruede et al. [97] used prosodic features (i.e. energy and pitch) in combination with syntactic word embeddings to train an LSTM model for BCs generation.

Regarding state-of-the-art BCs models, Voice Activity Projection (VAP) [98] is a BC prediction model that uses a transformer-based [99] architecture to predict the occurrence of turn-taking events (i.e. BCs and turn-takes). More specifically, VAP uses Voice Activation (VA) (i.e. whether an interlocutor is talking or not) to predict changes within the interlocutor's VAP. To elaborate, the model uses the VAP of two interlocutors to construct a VAP window which is used to model the future VAP information over the course of the dialogue (see Figure. 3.2). This window consists of a fixed number of bins that are considered to be either active or inactive, determined by a VAP threshold. These windows are used as labels during training.

Using this architecture, Ekstedt et al. [98] trained three different models: 1) an independent model that aims to predict the activation probability for each bin independently; 2) a discrete model that aims to predict the probability of a specific combination of activated bins (e.g. there are 2^8 different bin combinations, predict the probability for each combination); 3) a comparative model that predicts the VAP ratio over the entire window, disregarding the bins.

The VAP model consists of a VA encoder followed by a window sequence predictor (see Figure. 3.3). As the input of the model, the raw audio waveforms at the current timestep, the VA frame vector at the current timestep ($V_t^f \epsilon \{0,1\}^2$), and the VA ratio from the beginning of the recording until the current timestep split into five frames ($V_t^h \epsilon \mathbb{R}^5, \{-inf : 60, 60 : 30, 30 : 10, 10 : 5, 5 : 0\}$) was used. The encoder module consists of two submodules, one contrastive predictive coding (CPC)

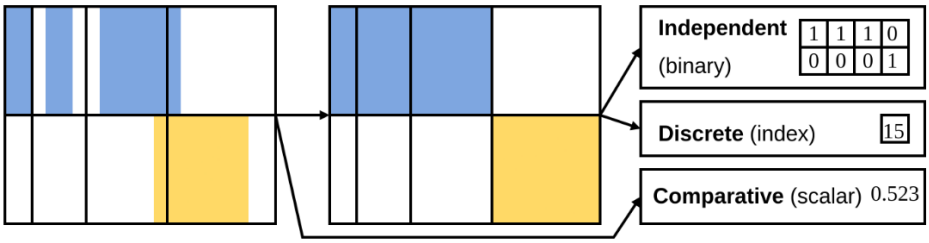

Figure 3.2: VAP window as proposed by Ekstedt et al. [98]. The window consists of 8 bins, each colour representing the VA of one interlocutor over a period of 2 seconds. The VA in each bin (left window) is extracted and used to determine whether a bin is considered active or not (right window) whenever it succeeds the specific threshold. An example of the output of the three different models is shown on the right.

module [100] which processes the raw audio waveforms and one VA module which processes the VA frame vector and VA history. The output of the encoder module is the speech frame representation h_{speech} at timestep t ($h_{speech,t} \in \mathbb{R}^{256}$). Subsequently, the predictor module uses this speech representation as input for a causal, decoder-only transformer layer [99] in combination with a linear layer to predict the voice activation windows.

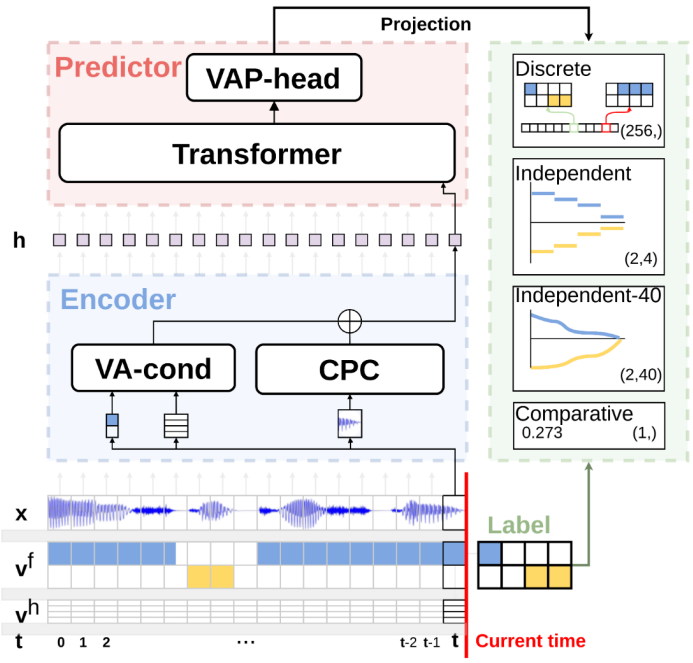

Figure 3.3: VAP model proposed by Ekstedt et al. [98]

As the model is not explicitly trained to predict BCs, Ekstedt et al. provide zero-

shot classification tasks[1] to classify specific turn-taking events. Hence, the authors defined a set of VA bin conditions used to classify BCs (see Figure. 3.4). While evaluating these zero-shot tasks, the discrete model has an average weighted F1-score of 0.723, which is statistically significantly better compared to the alternatives.

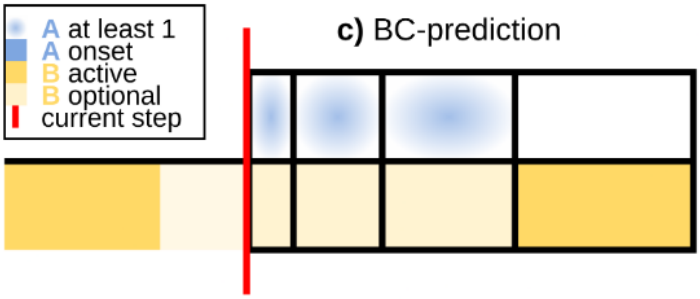

Figure 3.4: Zero-shot BCs classification conditions; to be classified as a BCs, the VA of the listening interlocutor has to be active for at least one of the first three bins in the projection window. Adapted from Ekdstedt et al. [98].

As the VAP model is, currently, considered state-of-the-art, it will be used for the remaining studies of this thesis. In the following section of this chapter, the theoretical knowledge regarding backchannels - and the modelling thereof - will be applied in the design of a user experiment to evaluate the perceived naturalness of the frequency and timing of generated BCs.

3.2 Methodology

3.2.1 Backchannel Model

To address whether the perceived naturalness of the frequency and timing of generated BCs can be on par with human BCs, we implemented a BC model. Among various BC models, the VAP [98] model, discussed in Section 3.1.2, demonstrated promising outcomes and was therefore selected for this study. However, since VAP solely provides the probability of a BC happening, we needed to introduce several modifications to assess the perception of these BCs. For an overview of the entire BC generation architecture, refer to Figure 3.5. Each individual module will be described in the following paragraphs.

First, to generate the BC utterances using the probability provided by the VAP model, a BC generation module was implemented. To trigger the BC utterances, this module uses the BC probability output from the VAP model in combination with

[1]Within deep learning, zero-shot classification is used whenever a model tries to perform a task it is not explicitly trained for; i.e. the model is trained to predict VA and not explicitly BCs.

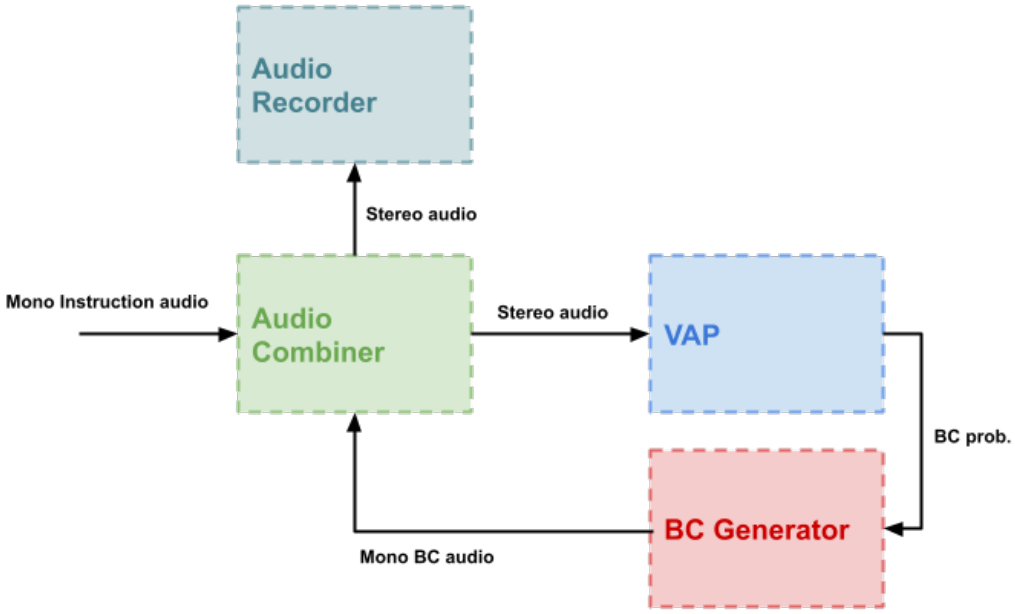

Figure 3.5: Overview of the architecture and various modules used to generate BCs.

a probability threshold and a BC cooldown (See Algorithm 1). The cooldown was necessary as it was likely that VAP produced multiple consecutive frames surpassing the BC threshold; this cooldown prevented the module from triggering multiple consecutive BC utterances every 500ms. To determine the appropriate BC threshold, multiple thresholds were evaluated during a user survey (see Section 3.2.2). The duration of the cooldown was determined in an informal experimental fashion; after running various pilot studies we concluded that a cooldown of 2 seconds was appropriate.

Algorithm 1 BC Generation Algorithm; the algorithm uses the BC probability (*BC prob.*), probability threshold (*BC threshold*), and cooldown (*BC cooldown*) to determine when to trigger a BC or not.

for *BC prob.* in *BC prediction stream* **do**
  if *BC cooldown* is not active **then**
    if *BC prob.* $>$ *BC threshold* **then**
      - Trigger BC utterance
      - Activate *BC cooldown*
    end if
  end if
end for

Second, as VAP requires a stereo input of two separate interlocutors, a module

had to be built that combined both interlocutors' audio streams into a single stereo stream. Both a mono audio stream containing a monologue from one interlocutor and a mono audio stream containing the generated BCs were used as input for this module. Subsequently, the stereo stream was passed to the VAP module as an audio recording module using a sample rate of 16khz, sample width of 2, and a sample frame duration of 500ms. The audio recording module saved the stereo stream to a file which could later be used for playback purposes.

3.2.2 Survey

A survey was conducted to evaluate whether the frequency and timing of the aforementioned BC model would be perceived as on par with humans. The survey consisted of 16 audio recordings in which the participants could hear two interlocutors; one of which was giving instructions, while the other provided BCs. The audio recordings used for both interlocutors were retrieved from the HCRC map task corpus [101], which is a corpus of unscripted, task-oriented dialogues which has been designed to support the study of spontaneous speech in general. For this study in specific, 4 different audio segments of various interlocutors were used in which they gave uninterrupted instructions. Moreover, a collection of around 6 short BC recordings made by the same interlocutor were extracted and used for the BC model. The recordings of the instructions were approximately around 20 seconds each and the BC recordings were triggered in random order by the BC model. Finally, the survey was reviewed and approved by the Ethics Committee Computer & Information Science at the University of Twente.

As the VAP model can be configured using various parameters (i.e. probability threshold P, prediction distance D), multiple configurations were evaluated using the survey. These conditions were chosen in specific as noticeable differences were found between the BCs, while still sounding relatively natural (e.g. a probability threshold of 0.1 wasn't used as it would result in almost a constant stream of BCs). Hence, the following conditions for the experiment were used:

- **D-P-**: prediction distance of between 0 and 600ms and probability threshold of 0.4.
- **D-P+**: prediction distance of between 0 and 600ms and probability threshold of 0.7.
- **D+P-**: prediction distance of between 600ms and 2s and probability threshold of 0.4.
- **D+P+**: prediction distance of between 600ms and 2s and probability threshold of 0.7.

During the survey, the participant was instructed to rate a total of 16 segments (i.e. 4 segments in each condition) based on how they perceived the naturalness of the frequency and timing of the BCs. Each metric was assessed using a 5-point Likert scale, where a rating of 1 indicated very poor quality and a rating of 5 indicated quality on par with human performance. See Appendix B for the survey and please see the footnote for the link to the audio samples used in the study[2].

3.2.3 Participants

In total, 20 participants were recruited for the survey using the researcher's personal network and various social media channels (i.e. LinkedIn, WhatsApp). To participate in the experiment, one was required to be at least 18 years old and possess a satisfactory level of English. Although the participant demographic consisted of various nationalities (i.e. South Korean, Chinese, French, Belgium, Dutch, and Irish), the majority of the participants were Dutch.

3.2.4 Analysis

Both the BC audio recordings and the survey results were analysed extensively. First, to mitigate bias, the BC recordings were annotated by two separate annotators for BRPs. We chose to use multiple annotators as 1) there doesn't seem to be a reliable method to extract BRPs in a quantitative manner [102], and 2) multiple annotators decrease the subjectiveness of the annotations and therefore make the annotations more reliable. Subsequently, the inter-annotator reliability was calculated using the Intersection over Union (IoU) method, which is also commonly known as the Jaccard Index [103] (see Equation 3.1). We chose to use IoU as it provides a normalized value, which is relatively easy to interpret and use. Finally, the annotations were analysed regarding the error between the timing of the BCs and the start/end of the BRPs. To determine whether there are statistically significant differences in the timing error between the various conditions, Mann-Whitney U tests were applied.

$$IoU = \frac{\sum_{i=1}^{N} (e_{Ai} - s_{Ai}) \cdot (e_{Bi} - s_{Bi})}{\sum_{i=1}^{N} (e_{Ai} - s_{Ai}) + \sum_{i=1}^{N} (e_{Bi} - s_{Bi}) - \sum_{i=1}^{N} (e_{Ai} - s_{Ai}) \cdot (e_{Bi} - s_{Bi})} \tag{3.1}$$

[2]Youtube playlist containing the audio samples used in the survey: https://tinyurl.com/hkxpd37b

Where:

N represents the number of intervals.
$\mathrm{s_{Ai}}$ and $\mathrm{e_{Ai}}$ are the start and end times of BRP i by annotator A, respectively.
$\mathrm{s_{Bi}}$ and $\mathrm{e_{Bi}}$ are the start and end times of BRP i by annotator B, respectively.

Regarding the analysis of the survey, the results were first assessed for normality using the Shapiro-Wilk and the Kolmogorov-Smirnov test. Subsequently, Levene's test was conducted to examine the homogeneity of variances for the perceived timing and frequency of the BCs. Moreover, a two-way repeated measures ANOVA was performed to analyze the effect of the prediction distance and probability threshold on the perceived frequency and timing. Finally, a post hoc Tukey test was conducted to further explain the ANOVA results.

3.3 Results

3.3.1 Annotations

As mentioned in the previous section, the BRPs within the BC audio recordings have been annotated by two separate annotators. The inter-annotator agreement, calculated using the IoU, is available in Table 3.1. The IoU value can be interpreted as the degree of overlap or similarity between two sets of intervals. The IoU value ranges between 0 and 1, with higher values indicating a greater degree of overlap or agreement. A higher IoU value signifies a greater amount of overlap between the intervals. For example, an IoU of 0.5 implies that half of the intervals from the two sets overlap or align with each other.

Sample	**Nr. of BRPs by A**	**Nr. of BRPs by B**	**IoU**
1th	8	5	0.410
2nd	6	5	0.330
3rd	4	4	0.430
4th	5	4	0.360

Table 3.1: The number of annotated BRPs by both annotators (A and B) and the respective IoU value between the annotations.

Using only the overlapping BRPs of both annotators, the error between the timing of the BCs and the annotated BRPs was computed; this timing error can be interpreted as the difference between the start of the BC and either start or end of the

BRP. For example, when the BC gets triggered by the model exactly during the BRP, the timing error is 0. While, when it gets triggered 1 second before the BRP, the timing error will be -1 seconds. We specifically chose to use the overlap of the BRPs instead of the union, as it resulted in a more reliable dataset for analysis, minimizing the potential for discrepancies that might skew the timing error data. See Figure D.1 and Table 3.2 for an overview of the timing errors using the aforementioned samples and experiment conditions.

Condition	Nr. of BC.	Mean	Std.	Precision
P-D-	14	0.283	0.606	0.714
P-D+	16	0.414	1.136	0.313
P+D-	6	0.000	0.000	1.000
P+D+	9	−0.193	0.756	0.222

Table 3.2: The mean and standard deviation of the timing errors and the number of BCs for each condition. The precision is calculated by dividing the BCs with a timing error of zero by the total amount of BCs

Additionally, the normality of the data for each condition was assessed using the Shapiro-Wilk and the Kolmogorov-Smirnov test. All the conditions, however, exhibited non-normal distributions. Therefore, Mann-Whitney U tests were applied to test for statistically significant differences between the conditions. The results indicated that there were no significant differences between any of the conditions.

To determine whether there were any compounding effects caused by the differences in the used samples, the timing error was also computed for each individual sample. See Table 3.3 for an overview of the timing errors per sample. Finally, Mann-Whitney U tests didn't denote any statistical significance between the samples.

Sample	Nr. of BC.	Mean	Std.	Precision
1st	12	0.103	0.523	0.500
2nd	14	0.282	1.228	0.500
3rd	11	0.048	0.617	0.455
4th	8	0.391	0.711	0.625

Table 3.3: Overview timing errors and amount of BCs, grouped by sample.

3.3.2 Survey

Regarding the analysis of the survey, first, a descriptive analysis was conducted by grouping the data both by condition and sample (see Table 3.4 and Table 3.5).

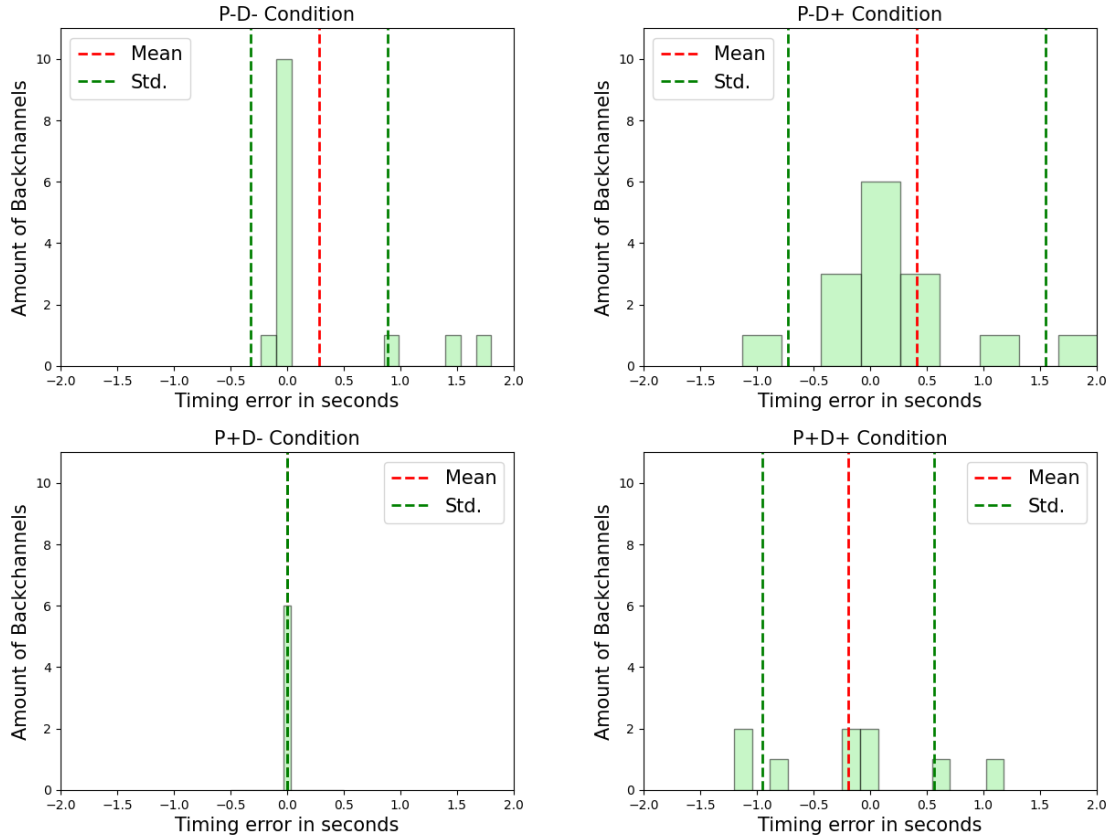

Figure 3.6: Timing error of the BCs relative to the BRPs, grouped per condition.

Subsequently, the normality of the data for each condition was assessed using the Shapiro-Wilk and the Kolmogorov-Smirnov test. For the Shapiro-Wilk test, the data exhibited a non-normal distribution. Similar results were obtained when applying the Shapiro-Wilk test. Therefore, caution should be exercised when interpreting the results.

	Timing		**Frequency**	
Condition	**Mean**	**Std.**	**Mean**	**Std.**
P-D-	3.325	1.155	3.200	1.107
P-D+	3.250	1.345	3.088	1.214
P+D-	3.450	1.168	3.150	1.045
P+D+	2.188	1.115	2.325	1.016

Table 3.4: Overview survey results, grouped by experiment condition.

Levene's test was conducted to examine the homogeneity of variances for the perceived naturalness of the timing and frequency. Regarding the timing, Levene's test statistic was found to be 3.0449 (p = 0.0290), indicating a significant difference

in variances among the groups. On the other hand, for the perceived frequency, Levene's test statistic was 1.7659 (p = 0.1536), suggesting no significant difference in variances.

	Timing		**Frequency**	
Sample	**Mean**	**Std.**	**Mean**	**Std.**
1st	3.550	1.157	3.338	1.006
2nd	2.738	1.040	2.700	1.024
3rd	3.013	1.345	3.025	1.067
4th	2.913	1.477	2.700	1.363

Table 3.5: Overview survey results, grouped by sample.

Finally, a two-way repeated measures ANOVA was performed to analyze the effect of the prediction distance and probability threshold of the BC model on the perceived frequency and timing of the BCs. This revealed that there was a statistically significant interaction between the effects of the prediction distance and the probability threshold for the perceived naturalness of the frequency (F(1, 19) = 8.755, p = 0.008) and perceived naturalness of the timing (F(1, 19) = 20.426, p = 0.001). Simple main effects analysis showed that the prediction distance did show a statistically significant effect on the perceived naturalness of the frequency (p = 0.001) and timing (p = 0.001). Additionally, another simple main effects analysis showed that the probability threshold did have a statistically significant effect on the perceived frequency (p = 0.004) and timing (p = 0.001).

For both the perceived naturalness of the frequency and timing, a post hoc Tukey test revealed that the P+D+ condition performs statistically significantly worse ($p < .05$) compared to all the other conditions; there was no significant difference found between the remaining conditions. An additional post hoc Tukey test revealed that audio sample 1 demonstrates a significantly better-perceived timing compared to the other samples.

3.4 Discussion & Limitations

In this study, we implemented a BC model based on the VAP model [98] to generate BC utterances in conversations. We annotated the BC relevant places (BRPs) and computed the timing errors between the start of the generated BCs and the BRPs. Moreover, we introduced modifications to assess the perception of these BCs and conducted a survey to evaluate the naturalness of the frequency and timing of the generated BCs. The analysis of the BRP annotations and survey results provides insights into the performance and limitations of the BC model.

The BC annotations revealed the timing errors between the generated BCs and the annotated BRPs. The mean timing errors varied across different conditions, but no statistically significant differences were found. This suggests that the different configurations of the BC model did not have a significant impact on the timing accuracy of the generated BCs. Furthermore, it is worth noting that, despite the relatively low IoU, the precision of the BCs (BCs triggered exactly during the BRPs) for the D- conditions are moderately high. These findings indicate that the D- BC model conditions were able to generate BCs that align well with the annotated BRPs. Finally, since the number of audio samples, annotations, and BCs used during this study is limited, these results should be interpreted with caution. Future studies could therefore focus on using a larger sample size to provide a more reliable statistical analysis.

Furthermore, the survey results provided insights into the perceived naturalness of the generated BCs. Participants rated the perceived naturalness of the frequency and timing of the BCs on a Likert scale. The results indicated that the prediction distance and probability threshold had a significant interaction effect on the perceived frequency and timing of the BCs. The analysis of simple main effects revealed that both the prediction distance and probability threshold individually had a significant effect on the perceived frequency and timing. Post hoc Tukey tests indicated that the P+D+ condition performed significantly worse than the other conditions in terms of perceived frequency and timing. This suggests that the combination of a longer prediction distance and a higher probability threshold led to less natural-sounding BCs. Furthermore, audio sample 1 was perceived to have significantly better timing compared to the other samples; this may indicate that the difference in speaker is a confounding factor.

While the survey results offer noteworthy insights, it's important to acknowledge certain limitations. Most notable among these is the relatively small sample size, potentially restricting the wider applicability of our findings. Another important point is that most participants identified themselves as Dutch, which might skew the interpretation of BCs due to cultural influences. By augmenting the sample size and diversifying the cultural backgrounds of the participants, we could attain a more inclusive comprehension of BC perception.

3.5 Conclusion

This chapter has shed light on the usefulness and limitations of the VAP model for generating BC responses during conversations. The analysis of timing errors and the annotation of BC relevant places (BRPs) have provided key insights into the model's effectiveness. Specifically, we found that the D- condition was successful in

producing BCs that matched well with the BRPs. Furthermore, both of the evaluated probability thresholds (i.e. 0.4% and 0.7%) resulted in relatively high precision (i.e. 0.714 and 1, respectively). Although the number of BCs in the P+ condition was lower compared to the P- condition, according to the survey, this didn't have a significant impact on the perceived naturalness of the frequency of the model. Therefore, the settings of the P+D- condition will be used during the final user study in Chapter 4.

Considering these findings, we concluded that the VAP model, though not perfect, is good enough to be used in the next chapter. Therefore, the next part of our research will look at how BCs influence a task involving human-CA collaboration. We aim to understand if these BCs can affect the time taken to complete a task and how they may affect the perceived collaboration with the CA. This continuation of our study will further explore the potential of BCs in human-CA collaboration.

Chapter 4

Collaborative Game

As concluded in the previous chapter, BCs play a pivotal role in the establishment of common ground between humans, which is essential for collaboration. Hence, the study in this chapter aims to delve into the role of BCs in a human-CA collaborative context, investigating their influence on both the perceived collaborative fluency and task duration. We chose these metrics in specific, as most human-CA collaboration studies use either the perception of the participants or task performance for their evaluations (as concluded in Chapter 2). Therefore this chapter aims to answer the following two primary research questions:

RQ4. To what extent do CA backchannels affect the task duration during a Human-CA collaborative task?

RQ5. To what extent do CA backchannels affect the perceived collaborative fluency during a Human-CA collaborative task?

This chapter is structured as follows: Section 4.1 provides a literature review, focussing on human-CA grounding and the evaluation of perceived collaboration. This is followed by a detailed description of our methodology in Section 4.2, outlining our experimental design, data collection, and analysis procedures. Our findings, as presented in Section 4.3, shed light on the nuanced effects of backchannels on perceived collaboration and task duration. We discuss these results in detail in Section 4.4, explaining their implications for understanding collaborative interaction dynamics. Finally, Section 4.5 concludes the chapter, reflecting on the key findings, their implications regarding human-CA collaboration, and proposing directions for future research.

4.1 Background

Although the literature study in Chapter 2 provided a solid foundation for Human-CA collaboration, this section aims to expand on that foundation in several directions. First, we review several papers regarding grounding in a Human-CA collaborative context. Subsequently, we elaborate on various evaluation methods to measure collaboration, with a focus on the Subjective Fluency Metric Scales [104], which will be used as one of the main evaluation methods during this study.

4.1.1 The Effect of Backchannels on Interaction

Although the available research regarding the effect of BCs on human-CA collaboration in specific is limited, several studies have attempted to analyse the effect of grounding on collaboration and communication in general; both in a human-human and human-computer context. This section aims to establish an overview of this literature.

First, in a study conducted by Gratch et al. [105], they investigated the effect of BCs on rapport, likability, trustworthiness, and helpfulness between human subjects and virtual agents. The study designed a face-to-face setup that captured the body movement and voice of the listening interlocutor. These features were subsequently used to animate and voice a CA. As a result, the speaking interlocutor could see the other participant represented as a virtual avatar/agent on a screen. The experiment had four conditions: 1) a control condition where both interlocutors had to communicate with each other in a face-to-face manner, without being represented as an agent; 2) a mediated condition, which simulates the actual head motions and postural changes of the listening interlocutor; 3) a 'contingent' agent condition, which uses automatic BC behaviour that is aligned with the whatever the speaker is saying; and 4) a 'non-contingent' condition, which uses automatic BC behaviour that is not aligned. Their results indicate that both 'non-contingent' and 'contingent' agents were as effective as human listeners in creating rapport, likability, and trustworthiness, as captured by a self-report scale. The mediated avatar condition was not as effective, however. Although the mediated avatar was perceived as equally as likeable and trustworthy as the other conditions, it was also perceived to be less helpful. This effect, however, can be reduced to the confounding effects caused by technical limitations.

In another study, Kontogiorgos et al. [59] discuss the importance of establishing, maintaining, and repairing common ground in task-oriented dialogues when collaborating with conversational interfaces. To elaborate, their study investigates if humans respond similarly to agents with different embodiments, social behaviour and con-

versational failures. Using three different wizarded CAs (i.e. one smart-speaker, one embodied CA without gaze behaviour, and one with gaze behaviour), they asked the participants to cook various dishes using the recipes provided by the CA. The results showed that the acceptance, clarification-seeking, and compliance behaviours of the participants were influenced by the embodiment of the agent and the reliability of its instructions. The study also used a referential communication task to maintain consistency in robot and human behaviour across conditions. It was found that participants exhibited more socially contingent interactions and increased gaze towards a human-like robot compared to a less anthropomorphic one. However, the lack of gaze behaviours in combination with an anthropomorphic body was counterproductive in stimulating non-verbal grounding behaviours. This indicates that it is not always favourable for CAs to be embodied or to use non-verbal backchannels (e.g. nodding) to establish common ground.

Furthermore, a study conducted by Blosma et al. [10] explores how different contexts, tasks, and applications require varying interaction styles for conversational AI systems. More specifically, it focuses on the impact of personality variation in CAs and how it can enhance the usability of such systems. The study investigates if differences in backchannel behaviour (audio and visual), particularly in embodied conversational agents, can signal variations in the perceived personality of the systems. Two rating experiments were conducted to assess participants' judgments of personalities in both human and artificial communication partners. The results indicate that feedback behaviour influences the perceived personalities of both humans and AI partners. This understanding can guide CA developers in incorporating personality into BC generation algorithms, leading to improved perceived personality and a stronger sense of presence for human users.

Several other studies emphasize the effect of BCs on the user's perception of the CA [106], [107]. Ding et al. [107], for example, derived various categories of BCs and analyzed their specific effect on cognitive assessments with older adults. They concluded two categories of BCs; reactive BCs (e.g. ”uhu”, ”yeah”), and proactive BCs (e.g. ”really?”, ”keep going”). According to their study, involving 36 older adult participants, proactive BCs are generally more appreciated. Moreover, they identified that while reactive BCs are generally more backwards-looking, proactive BCs can both be backward (e.g. ”really?”) and forward-looking (e.g. ”please keep going”). While BCs are generally considered to elicit responses from the user and prolong conversations [106], [108], forward-looking proactive BCs are considered to have the strongest effect in this regard.

While focussing on BCs in a human-human context, Wolf [109] studied the effects of various BCs provided by listeners on the fluency of speaking interlocutors who spoke in their second language (L2). The experiment was conducted using 14

Japanese participants who all had an intermediate level of English. The participants were instructed to perform three different oral tasks in English using three BC conditions; i.e. verbal/nonverbal (V/NV), nonverbal-only (NV), and no backchannels (NB). Results indicate that the participants were most fluent (on average) in the V/NV condition, while the fluency got incrementally worse for the NV and NB conditions. The differences in fluency between the V/NV and NB conditions proved to be statistically significant. This indicates that BCs encourage communicative fluency.

Since grounding in communication is a multi-faceted and complex process, it remains difficult to provide a thorough overview of all the effects BCs play within this context. Moreover, the observations made by the literature reviewed in this study, paint a contradictory view regarding the effect of BCs on task performance. Whenever, for example, a user is instructed to solve a task as quickly as possible using the aid of a CA, BCs generally encourage communicative fluency, which in return enables a quicker establishment of mutual understanding. Conversely, BCs may slow down the collaboration as they invite the user to continue talking for longer periods; especially when forward-looking, proactive BCs are used. Finally, the literature on the effect of BCs on the perception of humans appears to be more consistent. Generally, studies agree on the effect that BCs induce a stronger sense of presence and increase the likability and helpfulness of the CA. Although debatable, it may therefore be reasonable to extend this effect to the perceived collaboration.

4.1.2 Evaluating Collaboration

In the field of human-robot collaboration, there has been increasing focus on achieving ”collaborative fluency,” defined as the smooth interaction and integration of actions between human participants and their robot teammates [104]. The primary objective is not only to maximize the efficiency of the task at hand, but also to ensure a seamless human-robot partnership. To measure this fluidity, several evaluation metrics have been designed and utilized in ongoing studies [110]–[113]. In a recent study, Hoffman [104] defines and categorizes both subjective and objective fluency metrics within human-robot collaboration. The collaboration metrics discussed in the study encompass subjective measures such as internally valid scales and individual indicators, as well as four objective measures that could serve as reference points for appraising the fluency of human-robot collaborative interactions.

The subjective metrics mentioned by Hoffman encompass both direct measurements of fluency perceived in collaboration and the consequential outcomes of this perception, such as the human collaborator’s trust in the robot, the robot’s perceived contribution, its positive teammate traits, and the human’s belief in the robot’s commitment to the team. The questions used in their research are listed in Figure 4.1.

1 Human-Robot Fluency **α=0.801**

- "The human-robot team worked fluently together."
- "The human-robot team's fluency improved over time."*
- "The robot contributed to the fluency of the interaction."

2 Robot Relative Contribution **α=0.785**

- "I had to carry the weight to make the human-robot team better."(R)
- "The robot contributed equally to the team performance."
- "I was the most important team member on the team."(R)
- "The robot was the most important team member on the team."

3 Trust in Robot **α=0.772**

- "I trusted the robot to do the right thing at the right time."
- "The robot was trustworthy."

4 Positive Teammate Traits **α=0.827**

- "The robot was intelligent."
- "The robot was trustworthy."
- "The robot was committed to the task."

5 Improvement* **α=0.793**

- "The human-robot team improved over time"
- "The human-robot team's fluency improved over time."
- "The robot's performance improved over time."

*** only applicable for a learning or adaptation scenario**

6 Working Alliance for H-R Teams **α=0.843**

Bond sub scale (α=0.808)

- "I feel uncomfortable with the robot." (reverse scale)
- "The robot and I understand each other."
- "I believe the robot likes me."
- "The robot and I respect each other."
- "I am confident in the robot's ability to help me."
- "I feel that the robot appreciates me."
- "The robot and I trust each other."

Goal sub scale (α=0.794)

- "The robot perceives accurately what my goals are."
- "The robot does not understand what I am trying to accomplish."(R)
- "The robot and I are working towards mutually agreed upon goals."

Additional

- " I find what I am doing with the robot confusing."(R)

7 Individual Measures

- "The robot's had an important contribution to the success of the team."
- "The robot was committed to the success of the team."
- "I was committed to the success of the team."
- "The robot was cooperative."

Figure 4.1: Various scales, related to human-robot collaboration, used in the study by Hoffman [104]

Finally, through the use of a user study, Hoffman concludes that there is a complex relationship between multiple object metrics (e.g. human's idle time, robot's functional delay) and the actual perception of fluency in a human-robot collaborative setting. Interestingly, the study highlights that external observations of fluency may not be as sensitive as the perceptions of the participants directly involved in the collaboration, indicating a need for more participant-centric studies. Hoffman also notes that other elements may play a role in the perception of fluency, such as the correct and incorrect actions of the robot and human, the start and end times of actions, the relationship between the human and robot, and the repetition of actions, which are aspects not currently addressed by the existing metrics.

4.2 Methodology

4.2.1 Experiment

To answer research questions 4 and 5, an experiment has been conducted using a collaborative game (see Section 4.2.2). For this experiment, we used a 2X1 between-subject factorial design. To answer question 5, we used the perceived collaboration between the Human and CA as the dependent variable, while for question 4 we used the duration of each of the participant's turns. The independent variables used are whether the agent would use backchannels while listening to the participant (**B+**) or wouldn't use any backchannels (**B-**).

Regarding the perceived collaboration, we used an altered version of the Subjective Fluency Metric Scales [104] (see Section 4.1.2). The alterations had to be made as the original scale mainly focused on human-robot collaboration and some of the questions weren't applicable to our context. As a result, the participants were asked the answer the questionnaire depicted in Table 4.1 once they had completed the task. Using 5-point Likert scale questions, the questionnaire aims to capture the participant's perception of various aspects of collaborative fluency with the CA (i.e. contribution, working alliance, trust, positive traits, and several other individual measures). Regarding the effect of backchannels on the task duration, we recorded the timestamps whenever the participants started or finished their turn (i.e. pressed or released the spacebar).

Scale	Sub scale	Question
Contribution	-	(**Reversed**) I had to carry the weight to make the team better.
Contribution	-	The assistant was the most important team member on the team.
Contribution	-	The assistant contributed equally to the team performance.
Contribution	-	(**Reversed**) I was the most important team member on the team.
Contribution	-	The assistant had an important contribution to the success of the team.
Working alliance	Bond	The assistant and I understood each other.
Working alliance	Bond	I believe the assistant liked me.
Working alliance	Bond	I was confident in the assistant's ability to help me.
Working alliance	Bond	(**Reversed**) I felt uncomfortable with the assistant.
Working alliance	Bond	The assistant and I trusted each other.
Working alliance	Bond	I felt that the assistant appreciated me.
Working alliance	Bond	The assistant and I respected each other.
Working alliance	Goal	The assistant and I were working towards mutually agreed upon goals.
Working alliance	Goal	(**Reversed**) I find what I am doing with the assistant confusing.
Working alliance	Goal	(**Reversed**) The assistant did not understand what I was trying to accomplish.
Working alliance	Goal	The assistant accurately perceived what my goals were.
Trust	-	I trusted the assistant to do the right thing at the right time.
Trust	-	The assistant was trustworthy.
CA Commitment	-	The assistant was committed to the task.
CA Commitment	-	The assistant was committed to the success of the team.
Individual Measures	Intelligence	The assistant was intelligent.
Individual Measures	Cooperation	The assistant was cooperative.

Continued on next page

Table 4.1 Continued from previous page

Scale	Sub scale	Question
Individual Measures	Human commitment	I was committed to the success of the team.
Individual Measures	Fluency	The assistant and I worked fluently together.

Table 4.1: Altered Subjective Fluency Metric Scales, Subscales and individual measurements

Based on the various literature reviews conducted in this thesis, we hypothesize that the effects of the independent variable are as follows:

H1. Participants in the **B+** condition will have a longer task duration compared to participants in the **B-** condition.

H2. Participants in the **B+** condition will have an enhanced perception of collaborative fluency compared to participants in the **B-** condition.

4.2.2 Collaborative Game

In order to evaluate the effect of backchannels on human-CA collaboration, a prototype collaborative game was built. The main intention behind the design of this prototype was to elicit responses from the participants that were long enough in order for the CA to provide backchannels. This was especially challenging as most task-oriented conversational interfaces - to this day - only require their users to speak for relatively short durations [114]. This is partly due to the limitations of automatic speech recognizers and the additional cognitive load required to formulate long utterances. Since the HCRC map task [101], as described in Section 3.2.2, seemed to provide responses from the interlocutors that were long enough to elicit an appropriate amount of backchannels, it served as the main inspiration for the game built for this study. However, partly due to the complexity of the HCRC map task, which was difficult to fully simulate using a CA, we had to simplify the design of the game.

With reference to the game interface provided in Figure 4.2, the collaborative game goes as follows: The participant is instructed to guide the agent towards the right locations (e.g. ”Go north until you reach location Y. Then turn right” etc.). At each location, the agent will provide the participant with a part of the puzzle (i.e. a code). This code designates the subsequent location, as shown in Table 4.2, to which the participant must guide the agent. After 3 locations, the participant is asked

by the agent to validate whether they managed to collect the correct code. Once the participant presses the ”validate” button, the game will be finished.

To limit the probability and effect of any confounding factors as mentioned by Hoffman [104] (i.e. differences in perceived collaboration due to e.g. speech recognition errors), the agent was preprogrammed to always go to the right location - no matter the instructions of the participant. This resulted in the design shown in Figure 4.3.

Code	Location
33MT	Modern Art Gallery
T3JT	Daiso
ISP7	Emart24
59MN	Harbor
L3OR	City Hall
LMQ7	Lotteria
8Z9Y	Olive Young

Table 4.2: In order to decipher the code provided by the agent to find the subsequent location, the participants had access to the information provided in this table. Since the experiment was conducted in South Korea, the names used for the locations were inspired by South Korean franchises or landmarks

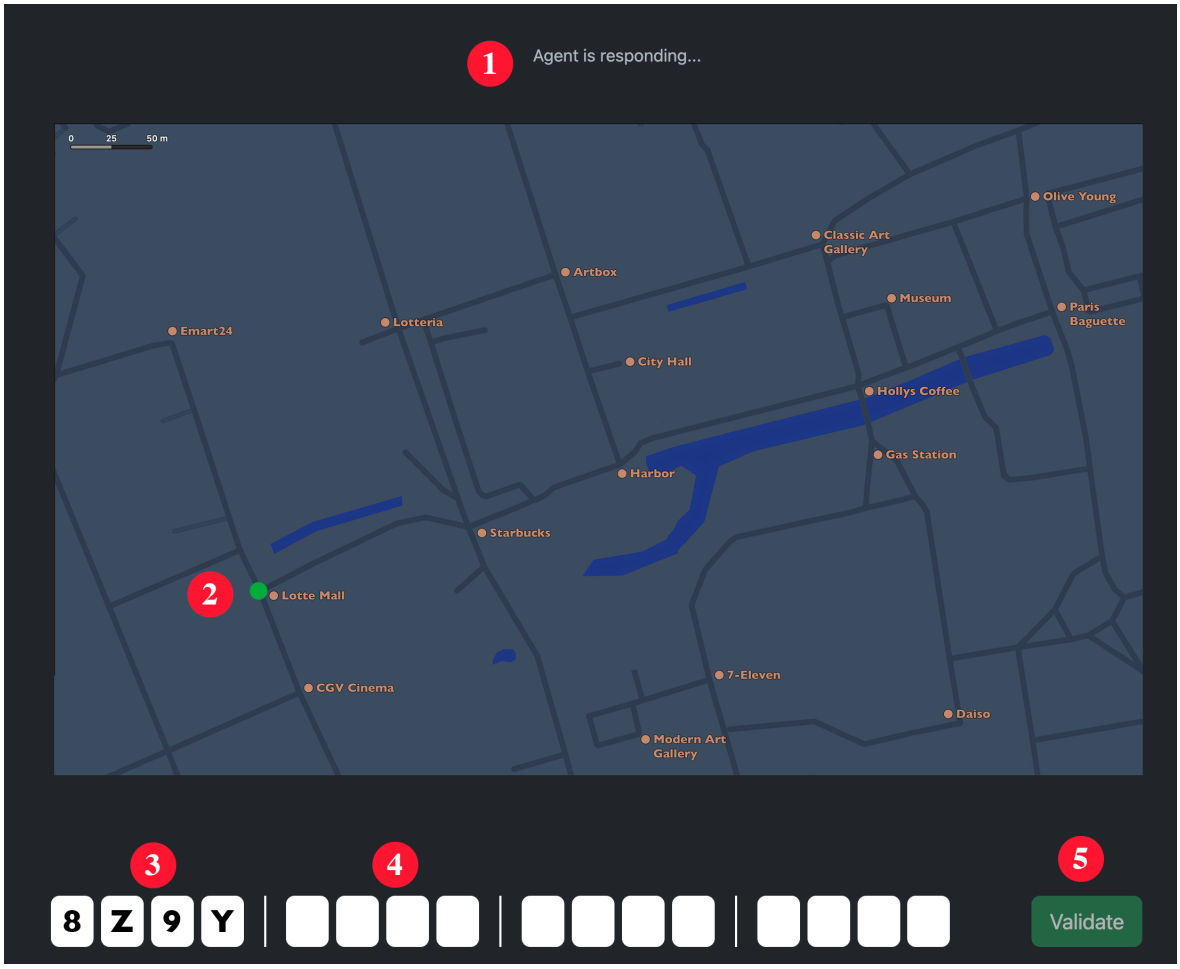

Figure 4.2: The graphical interface used by the participants to play the game. The numbers provided in this figure refer to the following:

1. **Agent Status**: the current status of what the agent is doing (i.e. responding, listening, or waiting).
2. **Agent Location**: the current location of the agent. This will be updated once the agent has reached any of the subsequent locations.
3. **Starting Code**: the first part of the code, which is provided by the agent at the start of the game.
4. **Remaining Code**: every time the agent reaches any of the subsequent locations, it will provide a part of the remaining code.
5. **Validate Button**: once the participant has navigated the agent to three more locations, the agent will ask the participant to press the validate button and finish the game.

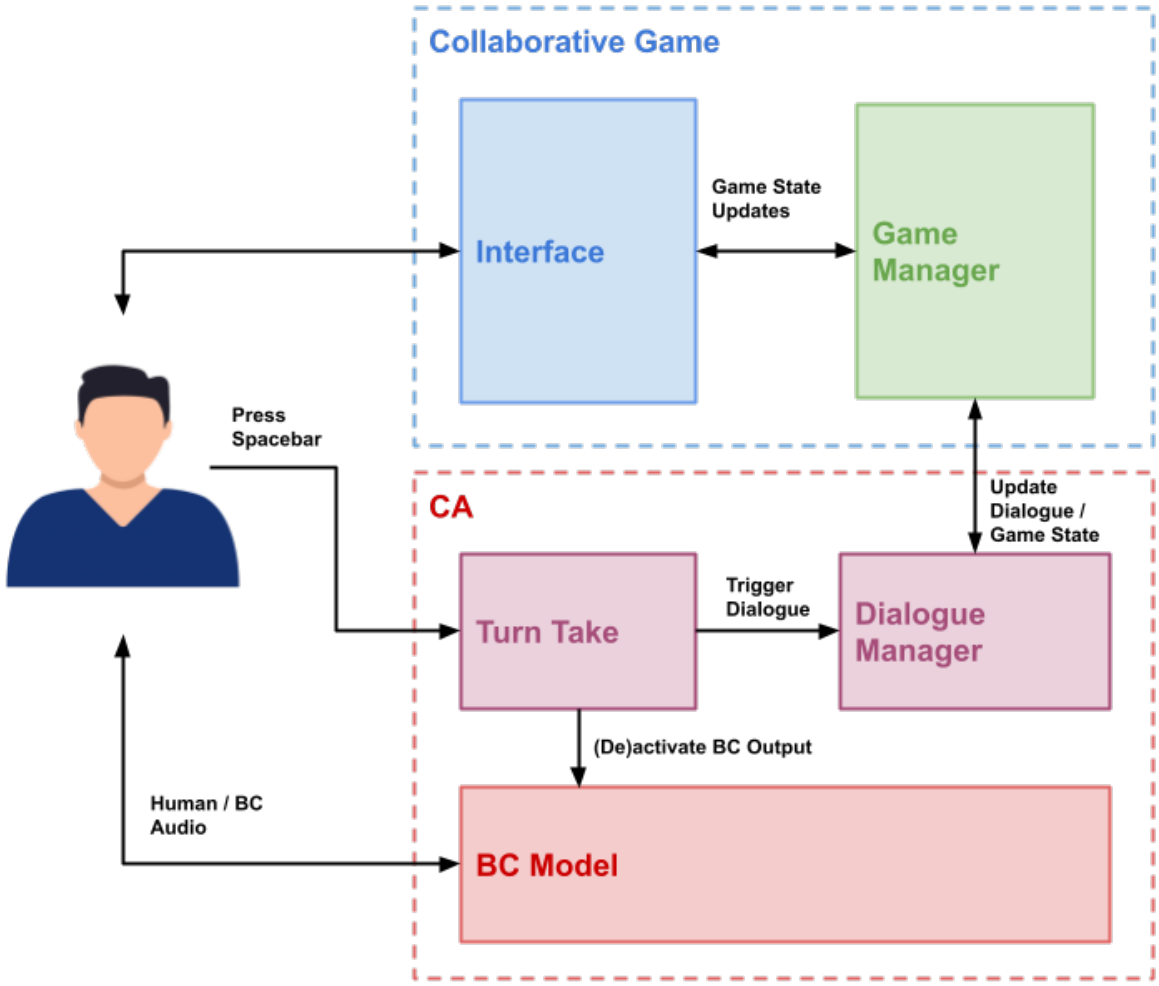

Figure 4.3: Overview of the architecture regarding the collaborative CA and game.

To enable appropriate turn-taking behaviour in conjunction with backchannels from the agent, we first designed a turn-taking module. Using this module the participant can take a turn by pressing and holding the spacebar on their keyboard; their turn will be finished whenever they release it. Using the spacebar for turn-taking enabled us to 1) accurately record the start and end time intervals of their turn; and 2) limit complexity and confounding effects caused by end-of-turn detection models. Although instructing the participant to use the spacebar to talk to the agent may introduce additional cognitive load, research suggests that it does not cause any significant hindrance for conversational interfaces [115]. Additionally, the spacebar was used to (de)activate the backchannel model's output; backchannels were generated as long as the participant held the spacebar. An additional safety measure was implemented to ensure the agent wouldn't prematurely move to the right location if the participant accidentally released the spacebar before finishing their instructions; An error response (e.g. "Sorry, but I don't really know where to go. Can you please give me a more elaborate explanation?") would be triggered

whenever the participant released the spacebar within 5 seconds of the start of their turn.

The backchannel model used in this study uses the voice activity from the participant, among other things, to generate backchannels at the right time during the participant's instructions (see Chapter 3 for an in-depth explanation of the model). Based on the outcome of the previous experiment, the model uses a BC probability threshold of 0.7 with a prediction window from 0 to 600ms. Moreover, the audio files used for the backchannel utterances are segments taken from the HCRC map task. The utterances chosen were both proactive and reactive, backward-looking BCs as they reduce the elicitation of longer responses compared to forward-looking BCs. We intentionally selected human utterances for generating backchannels since those produced through text-to-speech (TTS) seemed overly repetitive and robotic. While the remaining responses from the agent do employ TTS, we conducted an informal assessment with different participants to determine whether they noticed the use of two distinct voices. Initial responses indicate that the participants didn't notice remarkable differences until they were questioned about it.

In addition to the aforementioned modules, both a game and dialogue manager module have been designed to keep track of the state of the graphical user interface and agent. Utilizing these modules, we were able to program the agent through a series of JSON[1] files. Each file detailed a single turn and outlined the specific actions the agent was to perform during that turn (see Appendix C for additional details). The exact utterances used during this study were mainly written with the aim of providing responses that are collaborative, friendly, and as non-repetitive as possible. The TTS model used to generate the agent's utterances is IBM Watson[2]. The agent's exact actions and dialogue can be found in Table 4.3.

[1] JSON is a lightweight format using JavaScript notation for data exchange between computers.
[2] IBM Watson TTS: `https://www.ibm.com/products/text-to-speech`

Turn	Action	Duration	Utterance
1	Speak	5 sec.	Hey there, my name is Emma. Let's solve this task together.
1	Speak	5 sec.	Since you already got the instructions, let's get to it!
1	Speak	4 sec.	Okay, let me send you my location.
1	Give location	2 sec.	-
1	Speak	4 sec.	There you go. You should be able to see it on your map now.
1	Speak	5 sec.	I will also send you the first part of the code. Once you receive it, you should be able to give me directions on where to go next!
1	Give code	5 sec.	-
2	wait	3 sec.	-
2	Speak	5 sec.	Alright, I should almost be there.
2	Speak	4 sec.	Okay, I'm pretty sure I found it! Let me quickly update my location
2	Give location	2 sec.	-
2	Speak	5 sec.	Hmm, I also just found the code. Give me a second, I will send it to you.
2	Give code	4 sec.	-
2	Speak	5 sec.	Alright, where should I go next?
3	Speak	5 sec.	Almost there!
3	Speak	5 sec.	Okay, I made it. I'll update it on your map.
3	Give location	4 sec.	-
3	Speak	4 sec.	And let me send you the new code as well.
3	Give code	2 sec.	-
3	Speak	5 sec.	Alright! What is the next location?
4	Speak	5 sec.	Hold on. I am almost there.
4	Speak	5 sec.	Okay, I made it.
4	Give location	4 sec.	-
4	Speak	4 sec.	Here is the last piece of the puzzle
4	Give code	2 sec.	-
4	Speak	5 sec.	This is the moment of truth... can you validate whether the code is correct?
Outro	Speak	5 sec.	Great! It seemed to work. We solved the puzzle!
Outro	Speak	5 sec.	Thank you for helping me with this task. See you next time!

Continued on next page

Table 4.3 Continued from previous page

Turn	Action	Duration	Utterance
Error	Speak	5 sec.	Sorry, but I don't really know where to go. Can you please give me a more elaborate explanation?
Error	Speak	5 sec.	Sorry, but I think I am missing some directions. Can you tell me how I can get to the correct location?

Table 4.3: The agent's actions and utterances for each turn. The 'Outro' turn will be triggered once the participant presses the validate button. A random 'Error' turn would be triggered whenever the participant released the spacebar within 5 seconds.

Finally, in order to communicate between the game manager module and the graphical user interface, a WebSocket server was implemented in Python using FastAPI[3]. This server enabled communication from and to the user interface using JSON update messages. The user interface was built using vanilla Javascript, HTML and CSS.

4.2.3 Participants

Regarding the participants, to better understand possible confounding factors, we conducted a pre-experiment questionnaire in which the participants had to provide information regarding their native language and their general experience using CAs (see Figures 4.4 and 4.5 respectively). The experiment had a total of 21 participants, which although originating from different countries, were all able to speak English fluently. The participants were all university students at Pusan National University, recruited via various social media channels (i.e. Whatsapp, Instagram, KakaoTalk). The user study was reviewed and approved by the Ethics Committee Computer & Information Science at the University of Twente.

[3]FastAPI website: `https://fastapi.tiangolo.com/`

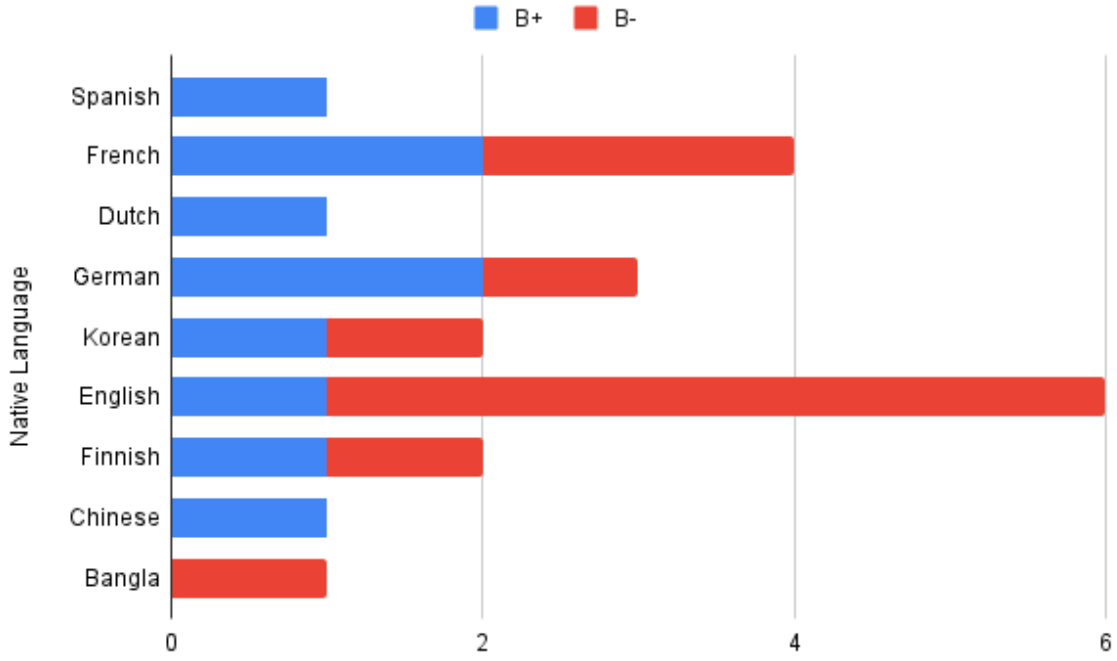

Figure 4.4: The number of participants separated by their native language.

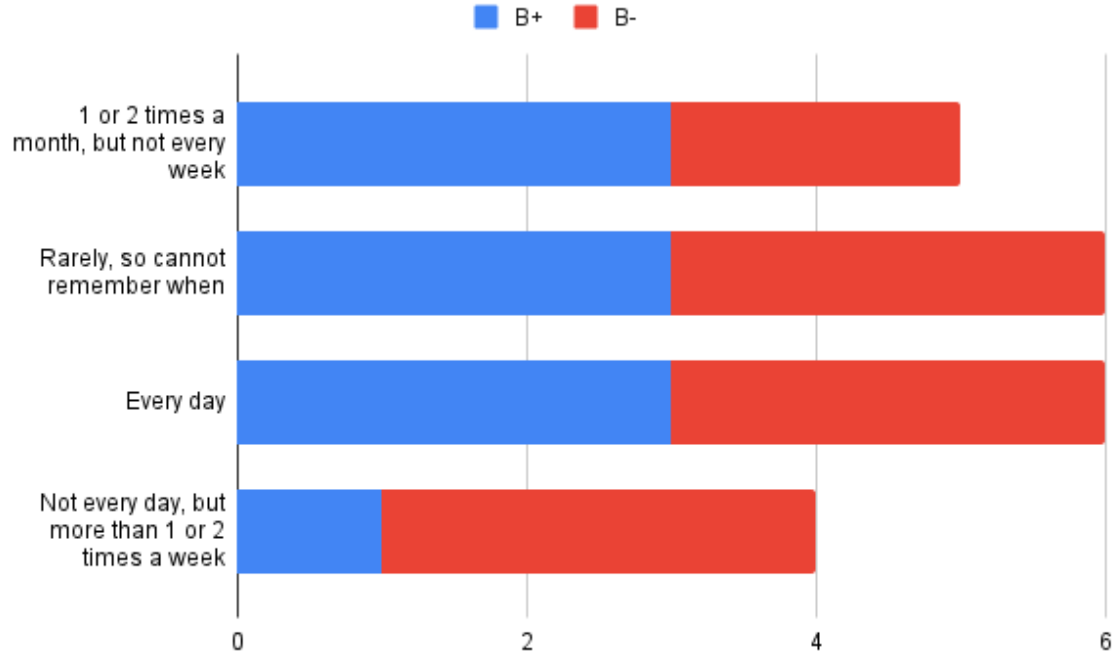

Figure 4.5: The number of participants separated by how frequently they use a CA.

4.2.4 Analysis

To analyse both the turn durations and the questionnaire results, various approaches have been utilized. Regarding the turn durations, a statistical analysis was conducted using both Welch's t-test and a Mann-Whitney U test. Prior to these tests, Shapiro-Wilk Test and Levene's Test were applied to evaluate whether the necessary assumptions were met to conduct the statistical analysis. Subsequently, regarding the post-experiment questionnaire results, both Welch's t-test and a Mann-Whitney U test are also applied for the analysis; we use both parametric and non-parametric tests as, within statistics, it is still highly debatable whether or not to use parametric tests for Likert scale data [116]. Finally, to analyse confounding factors caused by the differences between the background of the participants in both conditions, independent sample t-tests using a Bonferroni correction were performed for each question comparing the responses of English speakers and non-English speakers within each condition

4.3 Results

4.3.1 Task Duration

To conduct a statistical analysis regarding the effect of backchannels on the task duration, we first removed 3 significant outliers using a Z-score of 2. In other words, turns were removed of which the duration was either shorter or longer than 2 times the standard deviation from the mean. This reduced the number of used participants to 8 in the B+ condition and 10 in the B- (see Table 4.4, or Appendix D for more detail).

	B+			**B-**		
Turn	**Amount**	**Mean**	**Std.**	**Amount**	**Mean**	**Std.**
1	8	32.277	15.344	10	18.016	6.402
2	8	37.495	14.45	10	30.942	10.777
3	8	36.226	15.478	10	28.243	8.276
Sum	8	105.997	43.203	10	77.200	23.103

Table 4.4: Descriptive statistics regarding each turn for both conditions.

Additionally, the normality of the data for each turn for both conditions was assessed using the Shapiro-Wilk test (See Table 4.5). Turn 1 in the B- condition shows significant results, hence we reject the null hypothesis of normality. However, we fail to reject this hypothesis for the remaining conditions. Therefore, all the data (ex-

cept for turn 1 in the B- condition) can be considered to be approximately normally distributed.

	B+		B-	
Turn	**Test Statistic**	**p-value**	**Test Statistic**	**p-value**
1	0.938	0.588	0.728	0.002
2	0.967	0.876	0.936	0.511
3	0.950	0.706	0.910	0.279

Table 4.5: Shapiro-Wilk Test Statistics and p-values for condition B+ and B-

In order to assess the assumption of homogeneity of variance, Levene's test was conducted on the turn duration data (see Table 4.6). The results regarding turn 1 indicate a significant deviation from the assumption of equal variances, $F(1, 18) = 20.97$, $p < .001$. This suggests that the variability in scores between groups is significantly different at the first turn. However, for the second and third turn, the test results were non-significant ($F(1, 18) = 1.393$, $p = .256$ and $F(1, 18) = 3.004$, $p = .104$, respectively) suggesting that the assumption of equal variances holds true for the last two turns.

Turn	**Test statistic**	**p-value**
1	10.099	0.006
2	0.786	0.388
3	1.578	0.227
Sum	1.992	0.177

Table 4.6: Results Levene's Test of Homogeneity of Variance for each Turn.

Since the assumptions of normality and homogeneity of variance are only partially met in some of the turns in some conditions, we apply both Welch's t-test and a Mann-Whitney U test using a Bonferroni correction (see Table 4.7). Regarding turn 1, both Welch's t-test ($t=-2.65$, $p=0.024$) and Mann-Whitney U test ($U=19.0$, $p=0.023$) indicate statistically significant differences between the backchannel conditions. However, for the second and third turns, no significant differences could be found. Finally, the consistency between the two different tests, despite the partial fulfilment of the assumptions, gives added confidence to these findings.

4.3.2 Perceived Collaboration

To evaluate whether backchannels influenced the participants' perception of the collaboration, we analysed the results retrieved using the post-experiment questionnaire. We used Cronbach's Alpha to compute the internal consistency and reliability

	Turn 1		Turn 2		Turn 3	
	Statistic	**p-value**	**Statistic**	**p-value**	**Statistic**	**p-value**
Welch's	-2.650	0.024	-1.374	0.198	-0.933	0.366
Mann-Whitney	19.0	0.023	33.0	0.224	37.0	0.362

Table 4.7: Results from Welch's t-test and Mann-Whitney U test using a Bonferroni correction over the three turns. Regarding Welch's t-test, the 'Statistic' value denotes the t-value, which measures the size of the difference relative to the variation in the data; regarding the Mann-Witney U test, it denotes the U-value, in which a small value indicates a larger difference between the conditions.

of the scales. The obtained alpha values indicate a very good consistency for the 'Contribution' scale (α = 0.908 for B+, and α = 0.891 for B-) while indicating moderate to low consistency results for the remaining scales (see Tables 4.8, 4.9, and 4.10). Therefore, the results provide a robust basis for assessing the impact of backchannels on the 'Contribution' aspect of collaboration. However, due to the less reliable internal consistency scores for the 'Working Alliance', 'Trust', and 'CA Commitment' scales, the conclusions drawn from these scales should be treated with caution.

	B+			B-		
Scale	**Mean**	**Std.**	α	**Mean**	**Std.**	α
Contribution	2.983	1.330	0.908	3.787	1.122	0.891
Work. Alliance	4.418	0.846	0.442	4.331	0.847	0.819
Trust	4.455	0.656	-1.636	4.418	0.846	-1.208
CA Commitment	4.55	0.921	-1.636	4.636	0.839	0.370

Table 4.8: Mean, Standard Deviation and Cronbach's Alpha per scale per condition.

	B+			B-		
Subscale	**Mean**	**Std.**	α	**Mean**	**Std.**	α
Goal	4.550	0.705	-0.455	4.636	0.526	0.676
Bond	4.342	0.908	0.546	4.155	0.940	0.815

Table 4.9: Mean, Standard Deviation, and Cronbach's Alpha for the "Working Alliance' subscales, grouped by condition.

We used both Welch's T-Test and the Mann-Whitney U Test using a Bonferroni correction to compare the means of both groups for each scale and to investigate the differences in perceived collaboration between the two conditions (B+ and B-). The results of these tests are summarized in Table 4.11. There was a significant

	B+		B-	
Measurement	**Mean**	**Std.**	**Mean**	**Std.**
Cooperation	4.900	0.300	4.909	0.287
Commitment Human	4.900	0.300	4.455	0.891
Fluency	4.600	0.490	4.636	0.481
Intelligence	4.300	0.780	4.636	0.643

Table 4.10: Mean and Standard Deviation for each individual measurement

difference in the perceived contribution between B+ and B- as evidenced by both the Welch's T-Test (t=-3.613, $p<0.001$) and the Mann-Whitney U Test (U=1316.0, p=0.001). This suggests that backchannels indeed have a significant impact on the perceived contribution to collaboration, with the B- condition experiencing a higher level of contribution (see Figure 4.6). No significant differences were found between the backchannel conditions in the 'Working Alliance', 'Trust', and 'CA Commitment' scales. Likewise, when looking at the subscales of 'Working Alliance' (i.e. 'Goal' and 'Bond') and the individual measurements (i.e. 'Cooperation, 'Human Commitment', 'Fluency' and 'Intelligence', we found no significant differences between B+ and B- conditions.

	Welch's T-Test		**Mann-Whitney U Test**	
Scale	t	**p-value**	U	**p-value**
Contribution*	-3.613	0.000	1316.000	0.001
Work. Alliance	0.782	0.435	7133.000	0.292
Trust	-0.550	0.586	223.000	0.943
CA Commitment	0.179	0.859	233.500	0.680
Working Alliance Subscales				
Goal	-0.623	0.535	863.000	0.858
Bond	1.217	0.225	3038.500	0.147
Individual Measurements				
Cooperation	-0.067	0.947	54.500	1.000
Human Commitment	1.490	0.161	70.000	0.169
Fluency	-0.163	0.872	53.000	0.900
Intelligence	-1.018	0.322	42.000	0.309

Table 4.11: Results Welch's T-Test and Mann-Whitney U Test. * denotes a statistically significant difference.

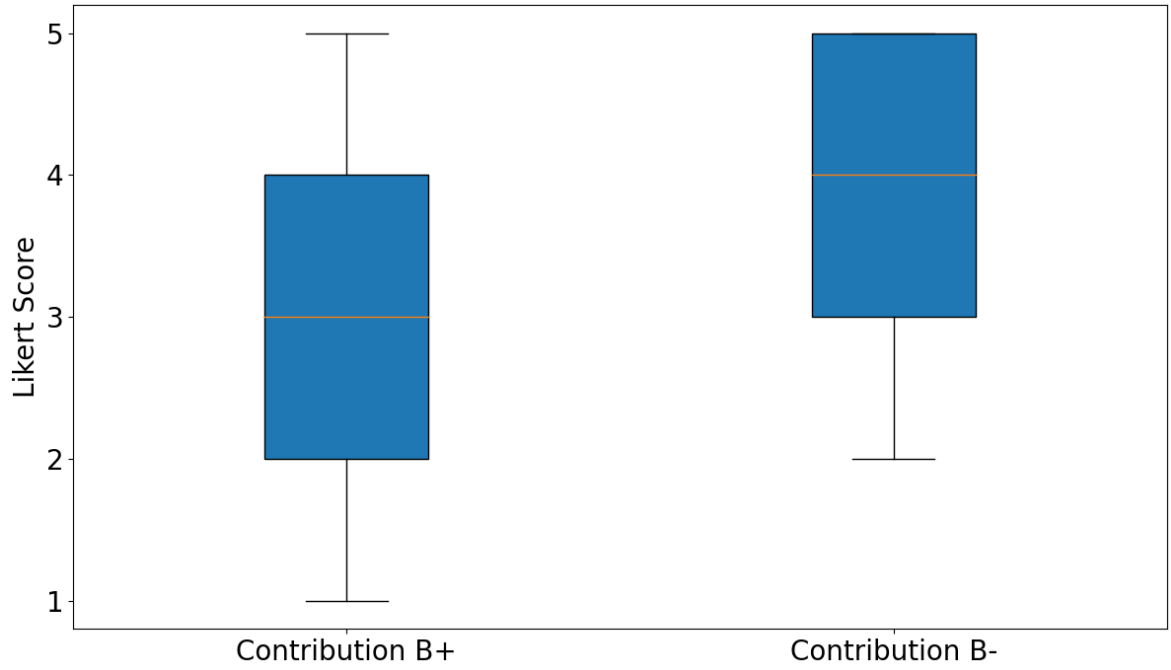

Figure 4.6: Box plots regarding Likert Scale answers for the 'Contribution' scale.

Finally, since there was a disproportionate amount of native English-speaking participants in B- compared to the B+ condition, we evaluated whether this resulted in any confounding factors. Independent sample t-tests using Bonferroni corrections were performed for each question comparing the responses of English speakers and non-English speakers within each condition. For all questions, the p-values exceeded a threshold of 0.05, suggesting that there was no statistically significant difference in responses based on native language. This indicates that the disproportionate amount of native English speakers didn't significantly influence the experiment results.

4.4 Discussion & Limitations

This research investigates the role of backchannels on perceived collaboration and task duration in the context of human-CA collaboration. A collaborative game served as our experimental environment, yielding important insights. Our data analysis indicates that backchannels exhibited a notable influence on task duration, particularly at the beginning of the interaction; the duration of the first turn was significantly longer in the backchannel-present (B+) condition, as evidenced by Welch's t-test and Mann-Whitney U test. This can possibly be explained by the response elicitating effects of the BCs; due to the BCs, the participants are invited to keep their turn and therefore speak for prolonged periods of time. Additionally, as participants are

not provided with any positive evidence of understanding during their turn in the B- condition, they may feel the need to end their turn to establish a mutual understanding with the agent. Although the latter turns do not show a statistically significant difference, the B+ turns remain longer compared to the B- turns (see Appendix D). This, however, may be explained by the relatively small sample size; a larger sample size may make the effect stronger and should therefore be studied more elaborately in future research.

In addition to the effect on task duration, our data analysis also paints a nuanced picture of the role backchannels play in the perceived collaboration. When evaluating the perceived contribution to the collaboration, the presence of BCs led to a lowered sense of contribution by the CA, which is supported by both Welch's T-Test and the Mann-Whitney U Test results. This may suggest - maybe counterintuitively - that a participant, in the absence of BCs, feels more hesitant about whether the CA understands their instructions, and therefore may experience an increased sense of responsibility to establish a mutual understanding. This sense of responsibility, in turn, may shift the participant's focus away from the CA, which may explain the increased perception of the CA's contribution to the collaboration. Likewise, in the B+ condition, the BCs may shift the participant's focus more towards the CA. As a result, the participant may experience an increased sense of shared responsibility to contribute together with the agent. This, however, is still quite speculative and will require further investigation.

Furthermore, in contrast to our initial expectations, we found that the presence or absence of backchannels did not notably impact perceptions of 'Working Alliance', 'Trust', and 'CA Commitment'. Both Welch's T-Test and the Mann-Whitney U Test showed no meaningful differences between B+ and B- conditions for these metrics. Still, given the moderate to low alpha values for these metrics, caution is required when interpreting these non-significant results. The limited internal consistency may suggest that these constructs are multi-faceted, or that the questions used to measure them did not fully capture the intended concepts. Future work could therefore focus on refining these scales to improve their reliability and accuracy in capturing participants' perceptions of these aspects of collaboration.

4.5 Conclusion

As mentioned in Section 4.2, based on the literature reviews carried out in this thesis, our hypothesis posited that the effects of the independent variable are:

H1. Participants in the **B+** condition will have a longer task duration compared to participants in the **B-** condition.

H2. Participants in the **B+** condition will have an enhanced perception of collaborative fluency compared to participants in the **B-** condition.

In light of our findings, various conclusions can be made. First, regarding *H1*, our results showed a trend in the expected direction, with participants in the B+ condition generally taking longer to complete their turns relative to those in the B- condition. This difference, however, was only significant for the first turn of the task; although the latter turns were also still longer compared to the B- condition, they were not statistically significant. Given these results, we fail to reject *H1*.

In contrast, the data for *H2* contradicted our expectations. Participants in the B+ condition reported a reduced perception of contribution by the CA compared to those in the B- condition. Other factors of collaborative fluency (e.g. working alliance, trust, commitment, intelligence and cooperation) didn't show a noticeable difference between the conditions. Moreover, given the relatively low alpha values for these scales, their reliability is potentially questionable. Therefore, caution is required when interpreting these scales. Given these results, we reject *H2*.

To conclude, BCs seem to have a significant effect on both task duration and perceived collaborative fluency. According to the result of this study, the use of BCs results in an overall increased task duration, while reducing the perceived contribution by the CA. These insights provide valuable considerations for designing more effective interactive systems. Conversational systems that may the user to answer short and concisely, for example, may not benefit from BCs, as they generally encourage the user to speak for longer durations. Conversely, if longer, more detailed user responses are desired, or if a mutual understanding is of high importance, BCs may be beneficial. While BCs might diminish the perceived contribution of the CA, the design of the task and the agent's responses could significantly influence this outcome. A CA that, for example, is less dependent on the user and shows more initiative may already change the perceived contribution. Therefore, more comprehensive research is required to fully understand the role of backchannels and to develop more reliable measures for assessing perceived collaboration.

Chapter 5

Final Conclusion & Recommendations

5.1 Conclusions

In this thesis, we have conducted a systematic analysis of human-CA collaboration with a particular focus on the role of BCs. Over the course of three interconnected chapters, we have evaluated current collaborative models and evaluation methods, developed an effective BC generation model, and explored the nuanced impacts of BCs on collaborative tasks.

First, in Chapter 2, we conducted a systematic literature review on the existing collaborative models and corresponding evaluation methods. With regard to the techniques and approaches used to design and develop systems that support human-CA collaboration (RQ1), we found that despite the growing popularity of human-CA collaboration research, it still remains relatively unexplored, with no general collaborative models or framework available and evaluation methods varying significantly across studies. Furthermore, to answer the question regarding how these systems are generally evaluated (RQ2), we found that the type of collaboration and the goals of the tasks dictate the evaluation metrics, such as user perception and task performance. We also identified that researchers need to carefully consider the type of CA they use, as some collaborative contexts may be better suited for different kinds of AI systems.

In Chapter 3, we evaluated the use of the VAP model for generating BC responses during conversations. More specifically, we questioned the extent to which the timing and frequency of the BC model can be perceived as on par with human BCs (RQ3). Using a user survey and analysis of the annotations of BC relevant places (BRPs), we found that a shorter prediction distance (i.e. D- condition, using a prediction window from 0 to 600ms) produced BC aligning well with annotated BRPs. Moreover, results indicate that both prediction distance and probability threshold sig-

nificantly affected the perceived naturalness of the timing and frequency of the generated BCs. Additionally, we found that the model is capable of producing relatively natural-sounding BCs, as evaluated by the perceived naturalness of the timing and frequency of the generated BCs. However, as the study's sample size was relatively low, and the majority of the participants were Dutch, these results may not be generalizable and may contain cultural biases. Still, the model was of sufficient quality to be used during the subsequent user study.

Finally, in Chapter 4, using the VAP model in the context of human-CA collaboration, we implemented a game to evaluate the effect of BCs. More specifically, using this experiment we questioned the extent to which the presence of BCs affects task duration (RQ4) and perceived collaborative fluency (RQ5). A user study was conducted (using 20 participants from 9 different nationalities), using a 2X1 between-subject factorial design with the presence or absence of BCs as the independent variables. The participants' earlier turns during the collaboration were significantly shorter when BCs were not present ($p < 0.05$), however, later turns didn't display any significant differences. Furthermore, using a 5-point Likert scale survey, we evaluated the effect on various metrics regarding perceived collaborative fluency. We conclude that the presence of BCs results in a significantly lower perceived contribution by the CA ($p < 0.001$). Other collaborative metrics (e.g. trust, working alliance, cooperation, commitment), however, didn't show any significant differences. All in all, these results indicate that BCs generally increase task duration, while reducing the perceived contribution by the CA. However, since the alpha values regarding the collaborative fluency scales, more comprehensive research is required to fully understand the role of BCs and to develop more reliable measures for assessing perceived collaboration.

On a final note, although BCs in the context of human-CA collaboration do influence both the task duration and perception of the CA, it is important to emphasize the nuanced nature of collaboration. As evidenced by the systematic literature review, collaboration covers a wide area of different contexts, goals, and objectives. Consequently, it is difficult to conclude whether computational BCs play a positive or negative role in collaboration. Collaborative systems that, for example, require the user to complete a task as quickly as possible may not benefit from BCs - or maybe even conversational interfaces - as BCs generally elicit longer user responses. However, systems with objectives that require elaborate descriptions (e.g. reporting incidents, eliciting customer feedback) may be able to benefit from BCs. Furthermore, with regard to the perceived collaborative fluency, it should be emphasized that the perceived contribution may also be different depending on the collaborative task. A system that, for example, would guide the user (instead of the other way around), inherently has a higher contributing factor during the collaboration. Therefore, in

order to better understand the multi-faceted role of BCs, they should be evaluated during different collaborative tasks as well.

5.2 Recommendations

Moving forward, various recommendations can be made regarding future improvements of this study and future human-CA collaboration research in general. First, focusing on this study specifically, future studies should aim to use larger sample sizes while evaluating BC generation models and aim to diversify the cultural backgrounds of the participants. This will improve the generalizability of the findings. Second, given the inconsistent results on the impact of BCs on perceptions of 'Working Alliance', 'Trust', and 'CA Commitment', among others, it would be beneficial to refine the scales used to measure these metrics. Improving their reliability and accuracy would provide a clearer picture of how BCs influence these facets of collaboration.

Furthermore, with regard to future human-CA collaboration research in general, given the relatively unexplored nature of this field, future work could focus on establishing standardized collaborative models and evaluation methods. This would ensure a common language and understanding in the field and facilitate more direct comparisons between studies. Moreover, future research should also explore the effects of other types of BCs (such as visual or body language cues) in addition to the verbal BCs evaluated in this thesis. These non-verbal cues play a significant role in human-human interaction and may provide additional richness in human-CA collaboration. Additionally, they could explore the effects of more advanced BC models (e.g. models that predict whether BCs should be proactive or reactive). Finally, the effect of BCs should be evaluated in different collaborative settings with different kinds of goals and objectives as the perception of the collaboration may differ depending on the context.

This thesis serves as a step toward understanding the intricate dynamics of human-CA collaboration. By building upon the findings and recommendations presented here, future research can help enhance the effectiveness and usability of collaborative CA systems.

Bibliography

[1] D. Rozado, "What is the IQ of ChatGPT?," Dec 2022.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, Jan. 2016. Number: 7587 Publisher: Nature Publishing Group.

[3] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister, "Hybrid Intelligence," *Business & Information Systems Engineering*, vol. 61, pp. 637–643, Oct. 2019.

[4] M. Poser and E. A. Bittner, "Hybrid teamwork: Consideration of teamwork concepts to reach naturalistic interaction between humans and conversational agents," *WI2020 Zentrale Tracks*, p. 83–98, 2020.

[5] D. Kahneman, A. M. Rosenfield, L. Gandhi, and T. Blaser, "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making," *Harvard Business Review*, Oct. 2016. Section: Decision making and problem solving.

[6] I. Seeber, E. Bittner, R. O. Briggs, T. de Vreede, G.-J. de Vreede, A. Elkins, R. Maier, A. B. Merz, S. Oeste-Reiß, N. Randrup, G. Schwabe, and M. Söllner, "Machines as teammates: A research agenda on AI in team collaboration," *Information & Management*, vol. 57, p. 103174, Mar. 2020.

[7] E. Bittner, S. Oeste-Reiß, and J. M. Leimeister, "Where is the bot in our team? toward a taxonomy of design option combinations for conversational agents in collaborative work," *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2019.

[8] R. Stalnaker, "Assertion," *Syntax and Semantics (New York Academic Press)*, vol. 9, pp. 315–332, 1978.

[9] M. Di Maro, "Computational Grounding: An Overview of Common Ground Applications in Conversational Agents," *IJCoL. Italian Journal of Computational Linguistics*, vol. 7, pp. 133–156, Dec. 2021. Number: 1 | 2 Publisher: Accademia University Press.

[10] P. Blomsma, G. Skantze, and M. Swerts, "Backchannel Behavior Influences the Perceived Personality of Human and Artificial Communication Partners," *Frontiers in Artificial Intelligence*, vol. 5, 2022.

[11] L. Memmert and E. Bittner, "Complex problem solving through human-ai collaboration: Literature review on research contexts," *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2022.

[12] M. Allouch, A. Azaria, and R. Azoulay, "Conversational Agents: Goals, Technologies, Vision and Challenges," *Sensors (Basel, Switzerland)*, vol. 21, p. 8448, Dec. 2021.

[13] M. E. Foster, "Enhancing Human-Computer Interaction with Embodied Conversational Agents," vol. 4555, July 2007.

[14] M. E. Foster, "Enhancing human-computer interaction with embodied conversational agents," *Universal Access in Human-Computer Interaction. Ambient Interaction*, p. 828–837, 2007.

[15] Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen, and M. Welling, "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence," *Computer*, vol. 53, pp. 18–28, Aug. 2020. Conference Name: Computer.

[16] A. Maedche, C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner, "AI-Based Digital Assistants: Opportunities, Threats, and Research Perspectives," *Business & Information Systems Engineering*, vol. 61, pp. 535–544, June 2019.

[17] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, "The future of human-ai collaboration: A taxonomy of design knowledge for hybrid intelligence systems," *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2019.

[18] W. L. Bedwell, J. L. Wildman, D. DiazGranados, M. Salazar, W. S. Kramer, and E. Salas, "Collaboration at work: An integrative multilevel conceptualization," *Human Resource Management Review*, vol. 22, pp. 128–145, June 2012.

[19] H. J. Do, H.-K. Kong, J. Lee, and B. P. Bailey, "How Should the Agent Communicate to the Group? Communication Strategies of a Conversational Agent in Group Chat Discussions," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 387:1–387:23, Nov. 2022.

[20] S. Kim, J. Eun, C. Oh, B. Suh, and J. Lee, "Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), pp. 1–13, Association for Computing Machinery, Apr. 2020.

[21] A. Bagmar, K. Hogan, D. Shalaby, and J. Purtilo, "Analyzing the Effectiveness of an Extensible Virtual Moderator," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 18:1–18:16, Jan. 2022.

[22] S. Kim, J. Eun, J. Seering, and J. Lee, "Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, pp. 87:1–87:26, Apr. 2021.

[23] J. Navajas, T. Niella, G. Garbulsky, B. Bahrami, and M. Sigman, "Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds," *Nature Human Behaviour*, vol. 2, pp. 126–132, Feb. 2018. Number: 2 Publisher: Nature Publishing Group.

[24] M. Schaekermann, J. Goh, K. Larson, and E. Law, "Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, pp. 154:1–154:19, Nov. 2018.

[25] N. A. Nik Ahmad, "CETLs : Supporting Collaborative Activities Among Students and Teachers Through the Use of Think- Pair-Share Techniques," *International Journal of Computer Science Issues*, vol. 7, Sept. 2010.

[26] L. M. Jessup, T. Connolly, and J. Galegher, "The Effects of Anonymity on GDSS Group Process with an Idea-Generating Task," *MIS Quarterly*, vol. 14, no. 3, pp. 313–321, 1990. Publisher: Management Information Systems Research Center, University of Minnesota.

[27] K. S. Haring, J. Tobias, J. Waligora, E. Phillips, N. L. Tenhundfeld, G. Lucas, E. J. de Visser, J. Gratch, and C. Tossel, "Conflict Mediation in Human-Machine Teaming: Using a Virtual Agent to Support Mission Planning and Debriefing," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1–7, Oct. 2019. ISSN: 1944-9437.

[28] K. J. Behfar, R. S. Peterson, E. A. Mannix, and W. M. K. Trochim, "The critical role of conflict resolution in teams: a close look at the links between conflict type, conflict management strategies, and team outcomes," *The Journal of Applied Psychology*, vol. 93, pp. 170–188, Jan. 2008.

[29] C. Zhang, C. Yao, J. Wu, W. Lin, L. Liu, G. Yan, and F. Ying, "StoryDrawer: A Child–AI Collaborative Drawing System to Support Children's Creative Visual Storytelling," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, (New York, NY, USA), pp. 1–15, Association for Computing Machinery, Apr. 2022.

[30] S. Ali, T. Moroso, and C. Breazeal, "Can Children Learn Creativity from a Social Robot?," in *Proceedings of the 2019 on Creativity and Cognition*, C&C '19, (New York, NY, USA), pp. 359–368, Association for Computing Machinery, June 2019.

[31] S. Ali, H. W. Park, and C. Breazeal, "Can Children Emulate a Robotic Non-Player Character's Figural Creativity?," in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '20, (New York, NY, USA), pp. 499–509, Association for Computing Machinery, Nov. 2020.

[32] S. Ali, H. W. Park, and C. Breazeal, "A social robot's influence on children's figural creativity during gameplay," *International Journal of Child-Computer Interaction*, vol. 28, p. 100234, June 2021.

[33] N. Devasia, S. Ali, and C. Breazeal, "Escape!Bot: Child-Robot Interaction to Promote Creative Expression During Gameplay," in *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '20, (New York, NY, USA), pp. 219–223, Association for Computing Machinery, Nov. 2020.

[34] P. H. Kahn, T. Kanda, H. Ishiguro, B. T. Gill, S. Shen, J. H. Ruckert, and H. E. Gary, "Human creativity can be facilitated through interacting with a social robot," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (Christchurch, New Zealand), pp. 173–180, IEEE, Mar. 2016.

[35] P. Jr, N. Freier, T. Kanda, H. Ishiguro, J. Ruckert, R. Severson, and S. Kane, "Design patterns for sociality in human-robot interaction," pp. 97–104, Mar. 2008.

[36] P. Jr, B. Gill, A. Reichert, T. Kanda, H. Ishiguro, and J. Ruckert, "Validating interaction patterns in HRI," pp. 183–184, Jan. 2010.

[37] P. H. Kahn, J. H. Ruckert, T. Kanda, H. Ishiguro, A. Reichert, H. Gary, and S. Shen, "Psychological intimacy with robots? using interaction patterns to uncover depth of relation," *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010.

[38] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications," *ACM Transactions on Information Systems*, vol. 2, pp. 26–41, Jan. 1984.

[39] P. Karimi, K. Grace, N. Davis, and M. L. Maher, "Creative Sketching Apprentice: Supporting Conceptual Shifts in Sketch Ideation," in *Design Computing and Cognition '18* (J. S. Gero, ed.), (Cham), pp. 721–738, Springer International Publishing, 2019.

[40] P. Karimi, N. Davis, M. L. Maher, K. Grace, and L. Lee, "Relating cognitive models of design creativity to the similarity of sketches generated by an ai partner," *Proceedings of the 2019 on Creativity and Cognition*, 2019.

[41] J. Waskan and W. Bechtel, "Paul Thagard, Mind: An Introduction to Cognitive Science. Cambridge, MA: The MIT Press 1996. Pp. xi + 213.," *Canadian Journal of Philosophy*, vol. 28, pp. 587–608, Dec. 1998. Publisher: Cambridge University Press.

[42] J. P. Guilford, "Creativity: Yesterday, today, and tomorrow," *The Journal of Creative Behavior*, vol. 1, pp. 3–14, 1967. Place: US Publisher: Creative Education Foundation.

[43] M. Jung, N. Martelaro, H. Hoster, and C. Nass, "Participatory materials: Having a reflective conversation with an artifact in the making," *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS*, June 2014.

[44] M. A. Boden, *The Creative Mind: Myths and Mechanisms*. London ; New York: Routledge, 2nd edition ed., Sept. 2003.

[45] C.-H. Li, S.-F. Yeh, T.-J. Chang, M.-H. Tsai, K. Chen, and Y.-J. Chang, "A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), pp. 1–12, Association for Computing Machinery, Apr. 2020.

[46] P. Robe and S. K. Kuttal, "Designing PairBuddy—A Conversational Agent for Pair Programming," *ACM Transactions on Computer-Human Interaction*, vol. 29, pp. 34:1–34:44, May 2022.

[47] S. K. Kuttal, J. Myers, S. Gurka, D. Magar, D. Piorkowski, and R. Bellamy, "Towards Designing Conversational Agents for Pair Programming: Accounting for Creativity Strategies and Conversational Styles," in *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 1–11, Aug. 2020. ISSN: 1943-6106.

[48] S. K. Kuttal, B. Ong, K. Kwasny, and P. Robe, "Trade-offs for Substituting a Human with an Agent in a Pair Programming Context: The Good, the Bad, and the Ugly," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, (New York, NY, USA), pp. 1–20, Association for Computing Machinery, May 2021.

[49] L. Williams, C. McDowell, N. Nagappan, J. Fernald, and L. Werner, "Building pair programming knowledge through a family of experiments," in *2003 International Symposium on Empirical Software Engineering, 2003. ISESE 2003. Proceedings.*, pp. 143–152, Sept. 2003.

[50] B. Shneiderman, "Designing computer system messages," *Communications of the ACM*, vol. 25, pp. 610–611, Sept. 1982.

[51] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, (New York, NY, USA), pp. 249–256, Association for Computing Machinery, Mar. 1990.

[52] J. Lopes, D. A. Robb, M. Ahmad, X. Liu, K. Lohan, and H. Hastie, "Towards a Conversational Agent for Remote Robot-Human Teaming," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 548–549, Mar. 2019. ISSN: 2167-2148.

[53] J. Simpson, H. Stening, P. Nalepka, M. Dras, E. D. Reichle, S. Hosking, C. J. Best, D. Richards, and M. J. Richardson, "DesertWoZ: A Wizard of Oz Environment to Support the Design of Collaborative Conversational Agents," in

Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing, CSCW'22 Companion, (New York, NY, USA), pp. 188–192, Association for Computing Machinery, Nov. 2022.

[54] J. Simpson, M. Richardson, and D. Richards, "A Wizard or a Fool? Initial Assessment of a Wizard of Oz Agent Supporting Collaborative Virtual Environments," in *Proceedings of the 10th International Conference on Human-Agent Interaction*, HAI '22, (New York, NY, USA), pp. 299–301, Association for Computing Machinery, Dec. 2022.

[55] T. J. Wiltshire, J. E. Butner, and S. M. Fiore, "Problem-Solving Phase Transitions During Team Collaboration," *Cognitive Science*, vol. 42, no. 1, pp. 129–167, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12482.

[56] S. Kim, J. Park, S. Han, and H. Kim, "Development of extended speech act coding scheme to observe communication characteristics of human operators of nuclear power plants under abnormal conditions," *Journal of Loss Prevention in the Process Industries*, vol. 23, pp. 539–548, July 2010.

[57] A. Drachen and J. H. Smith, "Player talk—the functions of communication in multplayer role-playing games," *Computers in Entertainment*, vol. 6, pp. 56:1–56:36, Dec. 2008.

[58] J. Simpson, P. Nalepka, C. L. Crone, R. W. Kallen, M. Dras, E. D. Reichle, S. G. Hosking, C. J. Best, D. Richards, and M. J. Richardson, "Tip of the Finger or Tip of the Tongue? The Effects of Verbal Communication on Online Multi-Player Team Performance," in *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '21, (New York, NY, USA), pp. 175–178, Association for Computing Machinery, Oct. 2021.

[59] D. Kontogiorgos, A. Pereira, and J. Gustafson, "Grounding behaviours with conversational interfaces: effects of embodiment and failures," *Journal on Multimodal User Interfaces*, vol. 15, pp. 239–254, June 2021.

[60] K. Bergmann, F. Eyssel, and S. Kopp, "A Second Chance to Make a First Impression? How Appearance and Nonverbal Behavior Affect Perceived Warmth and Competence of Virtual Agents over Time," vol. 7502, (Berlin, Heidelberg), pp. 126–138, Springer Berlin Heidelberg, 2012.

[61] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340,

1989. Publisher: Management Information Systems Research Center, University of Minnesota.

[62] A. Lund, "Measuring Usability with the USE Questionnaire," *Usability and User Experience Newsletter of the STC Usability SIG*, vol. 8, Jan. 2001.

[63] S. N. Gilani, K. Sheetz, G. Lucas, and D. Traum, "What Kind of Stories Should a Virtual Human Swap?," in *Intelligent Virtual Agents* (D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Scherer, and A. Leuski, eds.), Lecture Notes in Computer Science, (Cham), pp. 128–140, Springer International Publishing, 2016.

[64] H. Li, S. M. Edwards, and J.-H. Lee, "Measuring the Intrusiveness of Advertisements: Scale Development and Validation," *Journal of Advertising*, vol. 31, pp. 37–47, June 2002. Publisher: Routledge _eprint: https://doi.org/10.1080/00913367.2002.10673665.

[65] A. Modigliani, "Embarrassment, facework, and eye contact: Testing a theory of embarrassment," *Journal of Personality and Social Psychology*, vol. 17, pp. 15–24, 1971. Place: US Publisher: American Psychological Association.

[66] Z. Xiao, M. X. Zhou, Q. V. Liao, G. Mark, C. Chi, W. Chen, and H. Yang, "Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions," *ACM Transactions on Computer-Human Interaction*, vol. 27, pp. 15:1–15:37, June 2020.

[67] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics*, vol. 1, pp. 71–81, Jan. 2009.

[68] S. Elo and H. Kyngäs, "The qualitative content analysis process," *Journal of Advanced Nursing*, vol. 62, pp. 107–115, Apr. 2008.

[69] S. R. Hong, M. M. Suh, N. Henry Riche, J. Lee, J. Kim, and M. Zachry, "Collaborative Dynamic Queries: Supporting Distributed Small Group Decision-making," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, (New York, NY, USA), pp. 1–12, Association for Computing Machinery, Apr. 2018.

[70] M. A. Pang and C. C. Seepersad, "Crowdsourcing the evaluation of design concepts with empathic priming," *Volume 7: 28th International Conference on Design Theory and Methodology*, 2016.

[71] E. P. Torrance, *Torrance tests of creative thinking. Norms-technical manual. Research edition. Verbal tests, forms A and B. Figural tests, forms A and B.* Princeton: Personnel Press, 1966. OCLC: 714040431.

[72] H. Autman and S. Kelly, "Reexamining the Writing Apprehension Measure," *Business and Professional Communication Quarterly*, vol. 80, pp. 516–529, Dec. 2017. Publisher: SAGE Publications Inc.

[73] T. M. Amabile, "Social psychology of creativity: A consensual assessment technique," *Journal of Personality and Social Psychology*, vol. 43, pp. 997–1013, 1982. Place: US Publisher: American Psychological Association.

[74] P. H. Kahn Jr., *The human relationship with nature: Development and culture*. The human relationship with nature: Development and culture, Cambridge, MA, US: The MIT Press, 1999. Pages: xiv, 281.

[75] E. Turiel, *The Development of Social Knowledge: Morality and Convention*. Cambridge Cambridgeshire ; New York: Cambridge University Press, Apr. 1983.

[76] A. Strauss, J. M. Corbin, and J. Corbin, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Sept. 1998. Google-Books-ID: wTwYUnHYsmMC.

[77] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction, 1967. Google-Books-ID: oUxEAQAAIAAJ.

[78] C. Tony, M. Balasubramanian, N. E. Díaz Ferreyra, and R. Scandariato, "Conversational DevBots for Secure Programming: An Empirical Study on SKF Chatbot," in *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022*, EASE '22, (New York, NY, USA), pp. 276–281, Association for Computing Machinery, June 2022.

[79] A. Belshee, "Promiscuous pairing and beginner's mind: embrace inexperience [agile programming]," in *Agile Development Conference (ADC'05)*, pp. 125–131, July 2005.

[80] S. R. Schiffer, *Meaning*. New York, NY, USA: Oxford, Clarendon Press, 1972.

[81] D. K. Lewis, *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell, 1969.

[82] D. Heath, “Vocabulary: McCarthy, Michael, Oxford: Oxford University Press, 1990, x + 173 pp., £6.95 (Language Teaching: a Scheme for Teacher Education).,” *System*, vol. 20, pp. 531–537, Nov. 1992.

[83] H. H. Clark, “Grounding,” *Using Language*, p. 221–252, 1996.

[84] E. V. Clark, “Common Ground,” in *The Handbook of Language Emergence*, pp. 328–353, John Wiley & Sons, Ltd, 2015. Section: 15 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118346136.ch15.

[85] H. H. Clark and E. F. Schaefer, “Contributing to discourse,” *Cognitive Science*, vol. 13, pp. 259–294, 1989. Place: Netherlands Publisher: Elsevier Science.

[86] J. Allwood, J. Nivre, and E. Ahlsen, “On the Semantics and Pragmatics of Linguistic Feedback,” *Journal of Semantics*, vol. 9, Jan. 1992.

[87] H. H. Clark and S. E. Brennan, “Grounding in communication,” in *Perspectives on socially shared cognition*, pp. 127–149, Washington, DC, US: American Psychological Association, 1991.

[88] H. Buschmeier and S. Kopp, “Efficient communication through attentive speaking,” *Proceedings of the 14th Biannual Conference of the German Society for Cognitive Science*, 2018.

[89] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, “An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs,” *Language and Speech*, vol. 41, pp. 295–321, July 1998. Publisher: SAGE Publications Ltd.

[90] E. L. Asu-Garcia Ph. D. and P. Lippus Ph. D., *Nordic Prosody*. Dec. 2013.

[91] N. Cathcart, J. Carletta, and E. Klein, “A model of back-channel acknowledgements in spoken dialogue,” in *10th Conference of the European Chapter of the Association for Computational Linguistics*, (Budapest, Hungary), Association for Computational Linguistics, Apr. 2003.

[92] R. Meena, G. Skantze, and J. Gustafson, “Data-driven models for timing feedback responses in a Map Task dialogue system,” *Computer Speech & Language*, vol. 28, pp. 903–922, July 2014.

[93] L. Huang, L.-P. Morency, and J. Gratch, “Learning Backchannel Prediction Model from Parasocial Consensus Sampling: A Subjective Evaluation,” in *Intelligent Virtual Agents* (J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 159–172, Springer, 2010.

[94] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, pp. 70–84, Jan. 2010.

[95] D. Lala, P. Milhorat, K. Inoue, M. Ishida, K. Takanashi, and T. Kawahara, "Attentive listening system with backchanneling, response generation and flexible turn-taking," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, (Saarbrücken, Germany), pp. 127–136, Association for Computational Linguistics, Aug. 2017.

[96] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, "Speech Driven Backchannel Generation using Deep Q-Network for Enhancing Engagement in Human-Robot Interaction," Aug. 2019. arXiv:1908.01618 [cs].

[97] R. Ruede, M. Müller, S. Stüker, and A. Waibel, "Yeah, Right, Uh-Huh: A Deep Learning Backchannel Predictor: 8th International Workshop on Spoken Dialog Systems," pp. 247–258, Jan. 2019.

[98] E. Ekstedt and G. Skantze, "Voice Activity Projection: Self-supervised Learning of Turn-taking Events," May 2022. arXiv:2205.09812 [cs, eess].

[99] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[100] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," Jan. 2019. arXiv:1807.03748 [cs, stat].

[101] H. S. Thompson, A. Anderson, E. G. Bard, G. Doherty-Sneddon, A. Newlands, and C. Sotillo, "The HCRC Map Task Corpus: Natural Dialogue for Speech Recognition," in *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993.

[102] M. Heldner, A. Hjalmarsson, and J. Edlund, "Backchannel relevance spaces," in *Nordic Prosody: Proceedings of XIth Conference, 2012*, pp. 137–146, Jan. 2013. Journal Abbreviation: Nordic Prosody: Proceedings of XIth Conference, 2012.

[103] A. H. Murphy, "The Finley Affair: A Signal Event in the History of Forecast Verification," *Weather and Forecasting*, vol. 11, pp. 3–20, Mar. 1996. Publisher: American Meteorological Society Section: Weather and Forecasting.

[104] G. Hoffman, "Evaluating Fluency in Human–Robot Collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, pp. 209–218, June 2019. Conference Name: IEEE Transactions on Human-Machine Systems.

[105] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating Rapport with Virtual Agents," in *Intelligent Virtual Agents* (C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 125–138, Springer, 2007.

[106] E. Cho, N. Motalebi, S. S. Sundar, and S. Abdullah, "Alexa as an Active Listener: How Backchanneling Can Elicit Self-Disclosure and Promote User Experience," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 1–23, Nov. 2022. arXiv:2204.10191 [cs].

[107] Z. Ding, J. Kang, T. O. T. HO, K. H. Wong, H. H. Fung, H. Meng, and X. Ma, "TalkTive: A Conversational Agent Using Backchannels to Engage Older Adults in Neurocognitive Disorders Screening," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, (New York, NY, USA), pp. 1–19, Association for Computing Machinery, Apr. 2022.

[108] K. Inoue, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, "An Attentive Listening System with Android ERICA: Comparison of Autonomous and WOZ Interactions," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, (1st virtual meeting), pp. 118–127, Association for Computational Linguistics, July 2020.

[109] J. P. Wolf, "The effects of backchannels on fluency in L2 oral task production," *System*, vol. 36, pp. 279–294, June 2008.

[110] G. Hoffman and C. Breazeal, "Collaboration in human-robot teams," *AIAA 1st Intelligent Systems Technical Conference*, 2004.

[111] A. Thomaz, G. Hoffman, and M. Cakmak, "Computational Human-Robot Interaction," *Foundations and Trends in Robotics*, vol. 4, pp. 104–223, Jan. 2016.

[112] S. D. Jiang and J. Odom, "Toward Initiative Decision-Making for Distributed Human-Robot Teams," in *Proceedings of the 6th International Conference on Human-Agent Interaction*, HAI '18, (New York, NY, USA), pp. 286–292, Association for Computing Machinery, Dec. 2018.

[113] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient Model Learning from Joint-Action Demonstrations for Human-Robot Collaborative Tasks,"

in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, (New York, NY, USA), pp. 189–196, Association for Computing Machinery, Mar. 2015.

[114] F. Rakotomalala, H. Randriatsarafara, A. Hajalalaina, and R. Ndaohialy Manda Vy, "Voice User Interface: Literature Review, Challenges and Future Directions," *SYSTEM THEORY, CONTROL AND COMPUTING JOURNAL*, vol. 1, pp. 65–89, Dec. 2021.

[115] S. Mattys and L. Wiget, "Effects of cognitive load on speech recognition," *Journal of Memory and Language*, vol. 65, pp. 145–160, Aug. 2011.

[116] T. M. Liddell and J. K. Kruschke, "Analyzing ordinal data with metric models: What could possibly go wrong?," *Journal of Experimental Social Psychology*, vol. 79, pp. 328–348, Nov. 2018.

During the preparation of this work, I used Grammarly and ChatGPT to fix grammatical errors and to format LaTeX tables. After using these tools, I thoroughly reviewed and edited the content as needed, taking full responsibility for the final outcome.

Appendix A

Timing errors grouped by sample

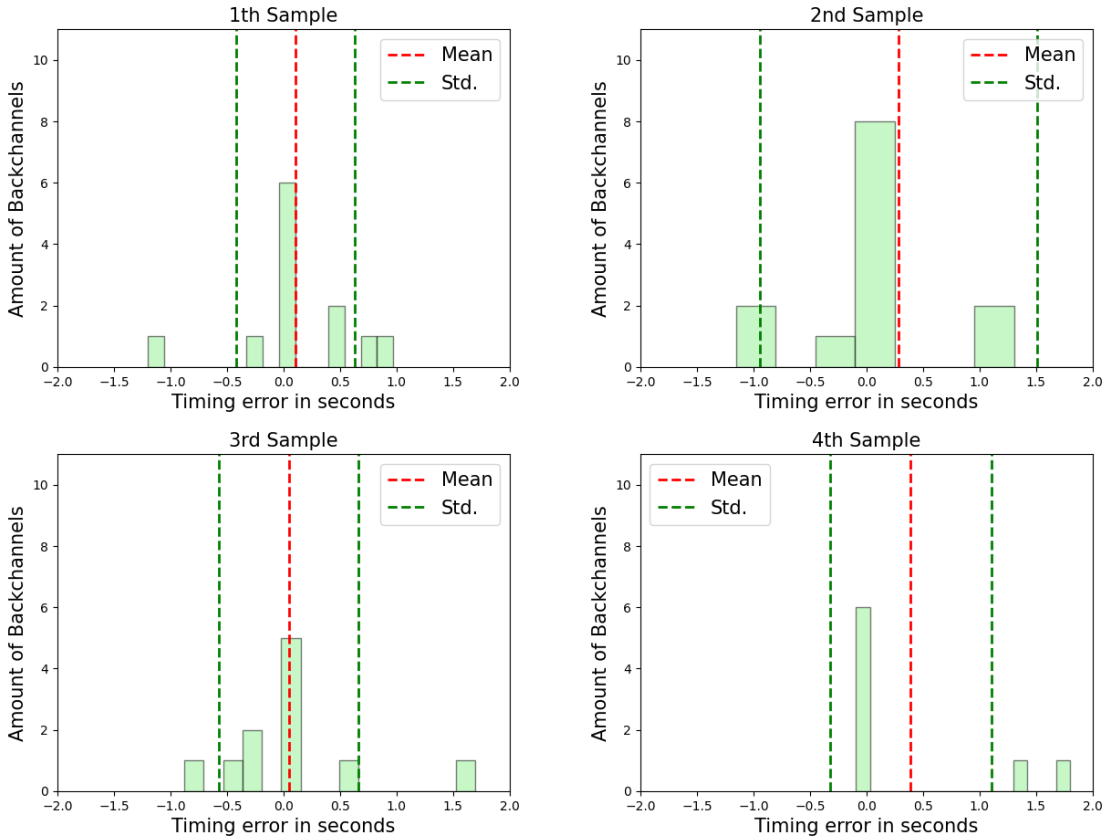

Figure A.1: Timing error of the backchannels relative to the BRPs and grouped per sample.

Appendix B

Backchannel Survey

Active Listening for Digital Assistants

* Indicates required question

Digital assistants are systems that understand and respond to users using natural language. Although digital assistants are gaining popularity in the industry (e.g. Siri and Alexa), various studies indicate that there is a lot to be desired regarding the overall quality of the interaction.

One possible solution to improve these interactions is to make the assistant listen more actively to whatever the speaker is saying. This can be achieved using so-called *backchannels*. *Backchannels* are short responses by the assistant (e.g. "Okay", "Uh huh", "I see", etc.) that provide evidence to the speaker that the assistant is still listening and that it correctly understands whatever has been said.

As the timing of these *backchannels* remains challenging, this study aims to evaluate various automatic *backchannel* models. In order to do so, you will be presented with several short audio segments containing multiple dialogues between humans and digital assistants. Your task will be to evaluate the **timing** and **frequency** of the *backchannels* made by the assistant.

Timing refers to whether the timing of the backchannels generally was *too soon, on time,* or *too late*

Frequency refers to whether backchannels are given *too often* or *too little*

The audio segments you'll listen to contain various route descriptions given by multiple people. The actual content of these descriptions are not important. In one ear you can hear the backchannels, in the other you can hear the person giving the instructions. **Please make sure you can hear both audio channels.**

Additional information
Time involvement: Your participation will take approximately 10 minutes.

Risks and benefits: If you decide to participate, please understand your participation is voluntary and you may withdraw from it at any time without giving any reasons. There are no risks or benefits that you can reasonably expect from participation.

The results of this research study will be presented in a master thesis. Your identity will not be made known in written materials resulting from the study. No personal information about you will be collected.

Contact information
Questions: If you have any questions, concerns, or complaints about this research, its procedures, risks, and benefits, contact the main responsible researcher, Roel Leenders, at +82-010-9796-9602 or email r.c.leenders@student.utwente.nl.

Independent Contact: If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact the Secretary of the Ethics Committee Computer & Information Science: ethicscommittee-cis@utwente.nl

Please click **YES** below if you are at least 18 years old and agree with the following conditions.

- I hereby declare that I have been clearly informed about the nature and method of the research and I agree to participate.
- I give consent for my (anonymous) answers to be analyzed.

1. Confirm: *

Check all that apply.

☐ Yes

Question 1/15

Please rank the backchannel's **timing** and **frequency**.

Timing refers to whether the timing of the backchannels generally was *too soon*, *on time*, or *too late*
Frequency refers to whether backchannels are given *too often* or *too little*

Appendix C

CA turn actions

Listing C.1: Example JSON file containing the specific type and duration of the actions the CA should perform in subsequent fashion. The 'key' value refers to the specific mp3 file that will be played during that action.

```
"actions": [
  {
    "type": "wait",
    "duration": 3
  },
  {
    "type": "speak",
    "key": "description_journey",
    "duration": 5
  },
  {
    "type": "update_code",
    "duration": 2
  },
  {
    "type": "speak",
    "key": "request_directions",
    "duration": 5
  }
]
```

Appendix D

Turn Durations per Condition

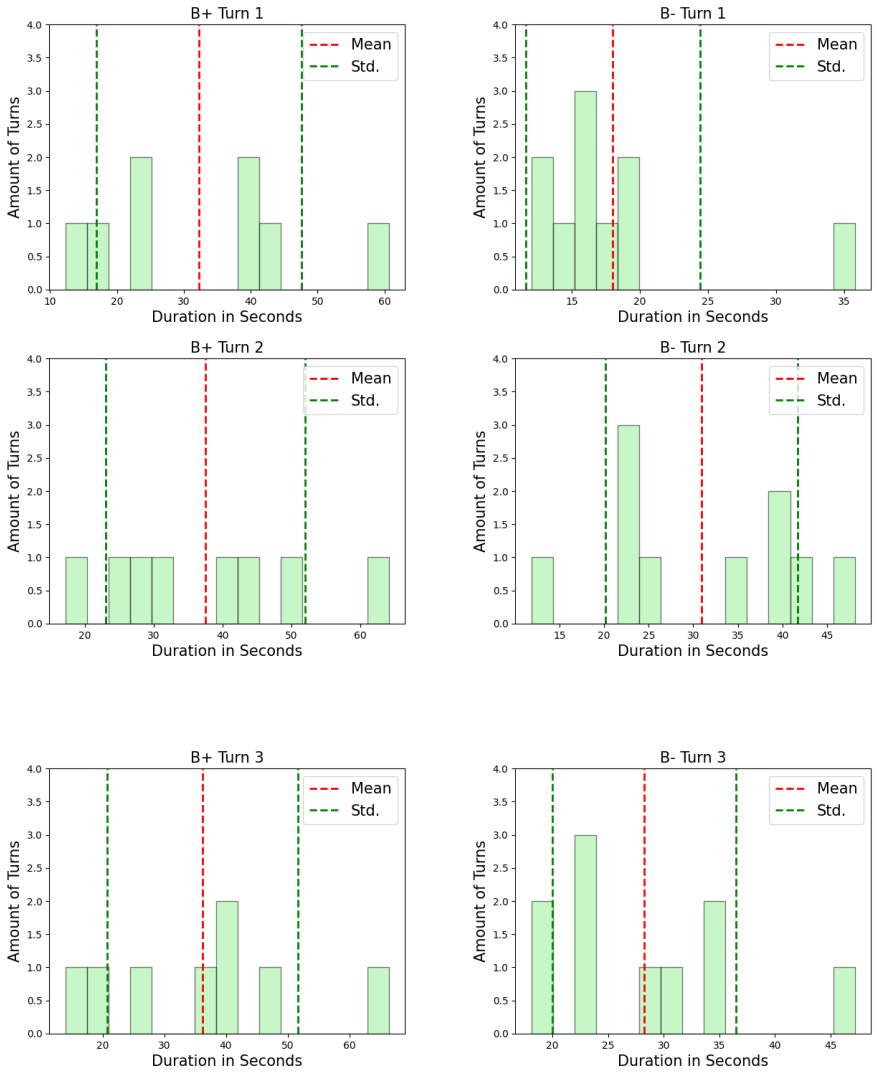

Figure D.1: Frequency distributions for the turn durations per condition.