Towards Human-like Conversations: Repair Strategies and Users' Mental Model in Task-Oriented Dialogue Systems

Francesca Alloatti

H-FARM Innovation and *Department of Computer Science, University of Turin*, Turin, Italy

December 2021

Contents

Chapter 1

Abstract

In the past few years, Natural Language Understanding abilities of Conversational Agents (CAs) have significantly improved. However, the interaction between humans and CAs is still far from perfect: people may employ fairly common linguistic strategies in human communication that are still hard to interpret for machines. While it is important to continue developing neural models that can represent language better and better, total perfection is not a reachable nor desirable goal. Even two humans communicating could have difficulties understanding each other. The strength of natural language is not its infallibility, but rather its ability to signal an error to the counterpart, to manifest the incomprehension, in order to initiate a repair and to bring the conversation back on its tracks. Since in human communication it is more efficient to recognize breakdowns in conversations rather than try to understand all input all the time, the present work applies the repair strategies framework in the context of human-machine interaction. The goal is thus twofold: first, to automatically detect the breakdowns in conversations (that is, when a user signals an incomprehension of sort); second, to categorize those breakdowns according to the repair strategy the human is employing.

Once the repair strategy has been correctly identified, there is not a single correct way to react to it: depending on the human counterpart, a certain answer may be more appropriate than another. The present work focus on a feature that is of paramount importance in describing the users: their mental model (MM). That is, their internal representation of a CA and their expectations about its functioning. The second part of the thesis focus on the topic of mental model in the context of CAs and how to recognize it in human-machine dialogues. We propose a simple methodology to gain some insights into the person's MM. We then apply such findings to our dataset in order to obtain double-tagged conversations: in the span of the same dialogue, we know what kind of repair strategy was employed as well as the user's MM that employed them. Being aware of these two information, the CA can elaborate a personalized answer. We tested the personalized answer by means of a user study and found that people did indeed prefer the modified answers.

Chapter 2

Introduction

Conversational agents (CAs), also known as chatbots or dialogue systems, are computer programs that communicate with users in Natural Language (Jurafsky and Martin, 2019). They support text-based, spoken or multimodal interactions with humans and they are usually categorized according to their purpose (McTear, 2020). Chat-oriented dialogue systems aim at conducting a long conversation with a user on a variety of topics, mimicking a casual exchange between humans (Adiwardana et al., 2020; Jiang and Banchs, 2017). A chat-oriented CA does not have a particular goal to reach, unlike task-oriented ones. These latter use conversations as a medium to help users complete a certain task, such as obtaining information, ordering a product, or perform an action in the real or virtual world (Pearl, 2016; McTear et al., 2016). Task-oriented agents are particularly interesting because their conversations with users have a purpose and both parties cooperate towards a shared goal. In fact, since users have a precise objective they want to reach through the dialogue, they are invested in the well being of the conversation (Motta and Quaresma, 2021; Li et al., 2020b, 2019). They are not trying to "break" the agent, but rather, they have every interest for the exchange to go in its best way possible.

However, it is not always easy for users to reach their goal through a conversation with a CA. One of the most prominent element that causes a communication failure is the inability of the system to properly understand the user: that is, a malfunctioning in its Natural Language Understanding (NLU) module (Ashktorab et al., 2019; Seeger and Heinzl, 2021; Lee and Lee, 2021). Failing to understand human's input is perceived by users as annoying and it is one of the principal causes of scarce users' retention (Li et al., 2020a). On one hand it is therefore undeniably important to continue improving NLU capabilities. On the other hand, it is not realistic to expect CAs to understand every possible sentence at any given time: full understanding, at all time, is not even possible for humans. Humans too can misunderstand an input from the other speaker, or have some problem interpreting a sentence. The major difference between a human and a machine is the person's innate ability to spot a breakdown in the conversation flow and to initiate a repair.

Let's observe for instance this dialogue between a family and an Amazon Alexa device[1]:

> **Mother:** *Alexa! (pause) Myanmar is South or South East Asian?*
> **Alexa:** *Sorry, I'm not sure.*
> **Son:** *Myanmar is South or South East Asian?*
> *(Alexa beeps) (Family conversation pauses)*
> **Mother:** *Try again.*
> **Son** *(louder): Is Myanmar part of South East Asia?*
> **Alexa:** *I'm not quite sure how to help you with that.*
> **Son:** *Hmmm…*
> **Mother:** *Alexa. Who is South East Asia? What country is East Asia?*
> **Alexa:** *There are no UN recognized countries in Eastern Europe.*
> **Son:** *This is getting annoying.*
> **Father:***(raised voice): Alexa. Define South East Asian countries.*
> **Alexa:** *This might answer your question. There are eleven sovereign states in South East Asia: Brunei, Cambodia, East Timor, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand, Vietnam. Did that answer your question?*
> **Mother:** *Yes.*
> **Alexa:** *Thanks for your feedback.*

During the exchange, the CA (in this case, an Alexa) fails at retrieving the information the family members are searching for. The people involved, instead of giving up, experiment a variety of actions in order to try and extract the information from the agent.

Those actions are known as *repair strategies*; specifically, repairs strategies (RS) are sequences in a conversation that aim at correcting some possible misunderstanding that may happen in a conversation (Norman and Thomas, 1991; Sacks et al., 1974; Schegloff et al., 1977).

In this thesis, we argue that a main point of improvements for dialogue systems would be to endow them with the human ability of repairing interactions. We believe that it is possible to enhance NLU capabilities of a system not by boosting NLU performances per se, but rather by equipping the CA with the ability to detect breakdowns in conversation and initiate a repair strategy. In fact, current research and industry efforts try to improve CAs understanding capabilities by working on two different aspects:

- **The *Machine Comprehension* aspect.** This first approach aims at boosting the NLU capabilities of a CA by improving the techniques it leverages. Such techniques involve both neural models to interpret and track repair phenomena, as

[1]The dialogue is taken from a Medium article by Marc Ericson Santos.

well as those that produce them. With regard to the first group, neural networks that can detect repair phenomena, several studies perform intent classification to this end (Ravuri and Stolcke, 2015; Chen et al., 2017; Vasudevan et al., 2019) or they incorporate user feedback to rerank the system's output (Shi et al., 2021; Jannach et al., 2021); others conduct a post-interaction data analysis to detect problematic spots in the conversation (Ngomo and Usbeck, 2020). The second group may consist in correcting the system's response by asking the user for clarification (Korpusik and Glass, 2019) or endowing the agent with the ability to produce clarification requests (Müller et al., 2021). In fact, different neural networks have been created and perfected with this goal in mind: a complete review of literature on this topic can be found in Ni et al. (2021).

Recent years have seen a surge in deep learning models that promise to augment agents' intelligence by leveraging large quantities of data (McTear, 2020). However, most CAs that are deployed in real life environments cannot benefit from these models out-of-the-shelf; they rather need a tailored fine-tuning and more often than not, in order to provide sensible answers to their users, they rely on a mixed approach (both statistical and symbolic) (Marcondes et al., 2020; Almansor and Hussain, 2020). Such boosted techniques are so far not delivering the improvement expected for products that have to work with real users and not just in a laboratory setting (Bender and Koller, 2020).

- **The design aspect.** In the discipline of conversational design, it is often said that a properly designed CA should always be prepared in case of breakdown. Conversational designers are thus prompted to create, along the so called *happy paths* - examples of exchanges where there is no error and everything goes smoothly - some *unhappy paths* too. In the latter, they can establish a predefined behavior in case of error. Even though this approach takes into consideration eventual mistakes that could happen during the interaction, by definition it is not possible to predict in what point a conversation could go awry, especially if the CA relies on a stateless interaction - that is, an interaction that does not keep track of the turns in a conversation (and thus the error cannot be identified at a specific point, because each point of the conversation is equal to any other). Even if it was possible, the reasons for a mistake could be various; it is therefore very hard to define a relevant answer to all possible breakdowns a priori.

Our goal is to combine the strength of statistical approaches with symbolic ones and create a technique that is effective for real-life deployed systems. Our ambition is to transfer a common and spontaneous phenomenon of human communication (the ability to employ repair strategies) to the field of human-computer interaction (Alloatti et al., 2021b). The present work enables a CA to detect RS put into place by users who interact with it, and then offers the CA the tools to react appropriately.

In order for a CA to be able to detect repair strategies, we must first identify clearly what kind of RS users are employing while they are talking to a task-oriented agent (Chapter 3). Our list, or rather taxonomy of RS should be both high-level (in order to incorporate various linguistic formulation of the same intent) and specific (to distinguish a RS from a regular sentence, as well as the different strategies) (Chapter 4). Through the definition of this taxonomy, we will know *how* users are trying to repair the dialogue and *how many times* they have to resort to one of these strategies. In fact, the presence of a RS entails a previous problem, a breakdown point in the conversation (or, rather, a breakdown of a specific act of communication). If there are many repair attempts in a set of dialogues, it would be safe to assume that the performances of that system are not particularly good. It would mean that users are forced to try to put the conversation back on its tracks because of some sort of failure. The count of the number of RS would thus constitute a good evaluation metric for the success of that CA. The first phase of our work establishes a new taxonomy of RS and applies them to a dataset of conversations between a task-oriented CA and real users (Chapter 5). This phase is carried out via a manual annotation; however, our final goal would be to equip the CA with the ability to detect RS automatically, on its own. For this reason, we developed a classifier that can spot and categorize strategies with no human intervention (Chapter 6).

By identifying repair strategies in the conversations, it would not only be possible to evaluate the performance of that dialogue system, but also to gather precious insights into the behavioral patterns of the user. Each RS entails a different behavior perspective, a distinctive way to approach a problem according to the user's character and cultural context (Ringberg et al., 2007; Lee et al., 2010). Information about the users' behavioral model can then be used to choose how to respond to a RS. In fact, once the CA is able to spot a RS in the conversation, it must react appropriately. However, there is not a single good answer for all strategies; it depends on the strategy, obviously, by also on the characteristics of the person who is employing that RS (Chapter 7). We consider some elements in users' communication that will guide us in providing the right answer: we identify those elements in the users' mental model (Norman, 1983), by drafting a proposal to detect such mental model (Chapter 8). A CA that is aware of the users' mental model could adapt to their expectations and beliefs, in order to make its answers more understandable and helpful (Gregor and Benbasat, 1999; Grimes et al., 2021; Radziwill and Benton, 2017). Moreover, it could prevent the repetition of certain mistakes due to unrealistic expectations caused by inappropriate mental models of the internal functioning of CAs (Chiang et al., 2020). Such agent's proactive behaviour can improve people's perception of its intelligence (Cuadra et al., 2021), while better understanding of users' mental models can lead to a redefinition of classical roles in human-computer interaction (Lee and Malcein, 2020). We checked these assumptions by means of a user study (Chapter 9).

To sum up, this work offers several contributions to the current state of the art in

HCI:

- It analyses existing literature at the intersection of Conversation Analysis, Dialogue Systems and Machine Learning, with the goal of finding a shared and consolidated framework to classify RS. Through this examination, we are able to come up with a novel set of tags to detect and label several RS in the context of dialogues between a user and a text-based, task-oriented dialogue system.

- The novel tagset is used to tag a dataset of conversations. The manual annotation provides an evaluation of the system's performances and it also depicts users' behavioral patterns in repairing interactions with a CA. While the tagset constitutes a new tool per the whole community to use, its application to real conversations offers an original perspective on error repairing in HCI.

- It analyses existing literature at the intersection of Mental Models and Dialogue systems. The goal is once again to check whether the "parsing" of mental models has already been conducted and validated in previous work. By parsing, we mean a system that would break down a conversation between a CA and a user to detect elements useful to determine the latter's mental model. The analysis of the state of the art guided us in the definition of a framework to place the user on a continuum of more or less Anthropomorphism. While the extensive literature survey offers a complete overview of the subject, the framework constitutes a novel contribution in the field of HCI.

- It prepares customized answers inspired from existing literature that take into account both the Repair Strategy employed by the speaker as well as the user's Mental Model. A final user study confirms the efficacy of the customized answers in comparison with the original ones from the dataset.

Computationally speaking, our contribution relates to the topic of classification:

- We developed an automatic classifier of RS. We applied state of the art neural networks to a classification tasks and report the results in details. Our approach proved that it is feasible to automatically detect RS employed by the user and it also validates computationally the intrinsic differences between tags built into our tagset.

The thesis is composed of the following chapters:

Chapter 3: *Repair Strategies*. It describes how RS have been conceptualized and applied in Human-Computer Interaction so far. The chapter gives an overview of previous work and highlights the main shortcomings that led to the development of our new RS taxonomy, or tagset;

Chapter 4: *A New Taxonomy of Repair Strategies*. It outlines the creation of a new tagset via two approaches: a top-down one (that takes from previous literature) and a bottom-up one (that creates new tags starting from the empirical observation of dialogues between a CA and real users). The tagset is then used to detect RS in a dataset of conversations from a task-oriented CA;

Chapter 5: *Detecting Repair Strategies: Tagging and Analysis*. This chapter covers the first phase of the work that aims at equipping the CA with the ability to detect RS. Two annotators conducted a manual tagging of a proprietary dataset. The results of this first activity already shines some light on the CA's performance, as well as users' behavioral patterns towards conversational breakdowns;

Chapter 6: *Detecting Repair Strategies: Development of an Automatic Classifier*. It describes the process of developing a classifier that automatically detects RS employed by users;

Chapter 7: *Users' Mental Model*. Once we created a classifier that could automatically classify RS, we then had to provide a meaningful response to each of those strategies, according to users' mental model. In this chapter, we describe how the subject of mental model in the context of dialogue systems has been discussed in literature;

Chapter 8: *Detection of Mental Model: a Proposal*. After extensively analysing previous work, we come up with a novel method to codify users' mental models by looking at their conversations with the CA. Specifically, we identify meaningful lexical features that can quantitatively determine the presence or lack of excessive anthropomorphism in the user's textual input;

Chapter 9: *Generation of Appropriate Responses*. Once the system has detected a RS and has understood the mental model of the users who produced that strategy, it can provide an accurate response. This chapter outlines a generative method to select an appropriate answer, carefully composed by taking from previous literature on the subject; then, it checks the validity of the news answers by conducting a user study. The study measures the difference in perception between the original answers of the dataset and the customized ones in terms of Hedonic Quality, Pragmatic Quality and Attractiveness.

Chapter 10 concludes the thesis. The interaction between the chapters of this thesis is depicted in Figure 2.1. From the image, it is possible to see that Chapter 3 constitutes the starting point for the whole dissertation. In fact, the background on Repair strategies is necessary to interpret all the subsequent argumentations. The first chapter informs Chapter 4, which in turn lays the necessary foundation to understand the manual tagging of the dataset (Chapter 5) and the automatic classification of RS (Chapter 6). While

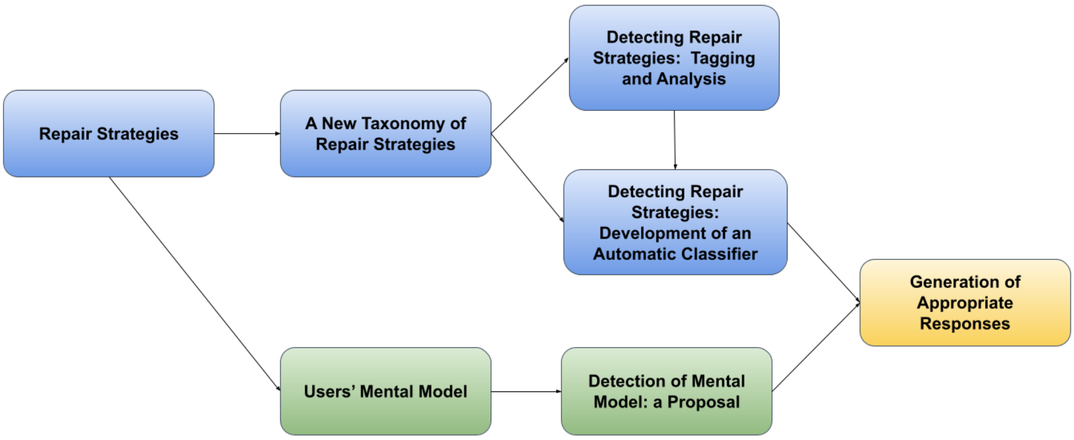

Figure 2.1: The figure shows the interaction of each chapter of the thesis with the other ones. Each color groups the chapters that form a thematic block.

these latter constitute parallel ways to tackle the issue of RS detection, it is also true that the manual tagging provides the necessary data to train the classifier; thus, Chapter 6 partially depends from Chapter 5. The green blocks constitute another theme discussed in this thesis: the user's mental model and how it is relevant to the activity of detecting and reacting to different repair strategies. Chapter 7 constitutes the background on the subject, while Chapter 8 lines out a novel proposal based on a rigorous analysis of existing literature. The blue blocks and the green ones merge into the final chapter of the thesis (Chapter 9). The information from the RS employed by the users and their MM informs the generation of personalized responses, that shall be perceived as more empathetic and useful than solely technical, pre-canned answers.

Chapter 3

Repair Strategies

In this chapter we present the theoretical background behind our experimental work with Repair Strategies (RS). We first outline an approach known in ethnomethodology (and later, sociolinguistics) (Heritage, 2001) as Conversation Analysis, where it is possible to find the first description of RS. We then shift our attention to the domain of HCI, to analyse how the concept of RS has been previously adapted to human-machine communication. Specifically, we focus on existing literature that deals with the description, detection and formalization of breakdowns (and their corrections) in the context of dialogue systems. We conduct an analysis of such literature in order to assess each work's contributions and limitations against our goals. The results, reported in Table 2.1 and 2.2, show that while there are several articles that tackles the topic of RS in HCI, none provide a clear set of tags to mark and uniquely identify the strategies, in the context of a written, task-oriented CA.

3.1 Conversation Analysis

Conversation Analysis is a sociolinguistic approach to the study of a particular kind of human-human interaction: conversations. It relies on the rationale that meaning is built through the alternation between speakers, and that this alternation, known as *turn-taking*, is regulated by rules provided by the social environment (Sacks et al., 1974). Norman and Thomas (1991) affirm that Conversation Analysis is particularly important given its intuition about two fundamental factors of human exchange: the fact that it is orderly, and that it allows reciprocal intelligibility between speakers. *Orderly* means that the interaction follows systematic patterns that can be found and reproduced; whenever one of the speakers deviate from an expected conversation path, the other one will try to build some meaning around that decision. *Mutual intelligibility* is possible since conversation is the main tool humans employ to display their intention, to narrate their actions and to demonstrate the motives of their behaviour.

Order and reciprocal comprehensibility can be achieved effortlessly, if both parties stick to a tacit and shared script. This script however is not innate, as it is determined by the people engaged in the conversation and their common social background. In this sense, Conversation Analysis relies on the ethnomethodological ground that speakers will practice reasoning about the surrounding world; their goal will be to "make sense" of the discourse developed in the exchange (Norman and Thomas, 1991). The role of Conversation Analysis is to search for underlying formal structures, the very same that allow for mutual understanding between the speakers. In fact, the alignment of the discourse around familiar and shared structures is the core of making conversation orderly and rational, rather than random and meaningless, and makes reciprocal comprehensibility possible.

Conversation Analysis also states that turn-taking (e.g. the alternation between speakers) is a fundamental feature of conversational organization, since it is the very same action that allows for the conversation to exist (Sacks et al., 1974). The analysis of turn-taking convenes on two major themes: i) the identification of the primary units of turns; ii) how these units get allocated between speakers. Sacks et al. (1974) describe turn constructional units as the basic bricks that constitute the foundations of conversation. Dialogues in specific settings, such as in courthouses or during a commercial transaction present turn alternation features of their own; however, Conversation Analysis states that nobody can perform social actions of any kind without getting a turn at talk. The turn-taking model grants an omnipresent background that guides the conduct of a conversation regardless of any particular turn-taking system that may be in play.

3.1.1 Adjacency Pairs

From the work of Conversation Analysis on turn-taking emerges that a conversation is usually composed by different couples of sentences, called *adjacency pairs*. Adjacency pairs (Schegloff and Sacks, 1973) consist of balanced pairs of utterances from successive speakers. It constitutes the basic framework around which conversational activities (e.g. greetings, requests, accusations, etc.) are formed. Conversation Analysis states that the fact that two elements of the pair are adjacent is not a coincidence. On the contrary, when the first element of a pair is uttered, then a second element is naturally required. The second speakers act under the constraint that they should produce a response based on the previous speaker input. This constraint is called *conditional relevance* and it is worth mentioning that it is not caused by proximity, but rather by the fact that a second statement is naturally produced by the other speaker after the first one (Norman and Thomas, 1991). In fact, the two elements of the pairs do not need to be physically subsequent one to the other, as they may be distributed over a sequence of turns. Some sequences can accomplish a specific social functions for the speakers; it is the case of *repair strategies*. Repair is the act of correcting any mistake that may happen in a conversation that prevent meaning to be shared fluidly by the two parties.

3.2 Problems and Repairs

So far, we were able to establish that:

- The discipline known as Conversation Analysis states that the communication medium of the *conversation* allows for mutual understanding, since it follows systematic patterns.
- These patterns, or the "tacit and shared script", are not innate. They are determined by the social background of the people involved in the exchange, as well as the context and the objective of the dialogue.
- The turn-taking model thus grants that even if dialogues in specific settings have their own peculiar features, they all share the same functional background, since the main goal is always to build shared meaning between the two speakers.
- The turn-taking model is based on minimal couples, called adjacency pairs. Each pair may have a specific function: one of those is dedicated to *repairing* possible breakdowns.

We will now dig deeper into the concept of *repairing* a conversation, and how that can be accomplished in a sociolinguistic theoretical framework.

According to Conversation Analysis, an utterance can be marked as ambiguous not because it pertains to a specific class of ambiguous statements, or because it is intrinsically ambiguous, but because it is interpreted as such by the speakers in the dialogue (Piantadosi et al., 2012). If one of the participants detects an inconsistency in the normal course of the conversation, a repair sequence can be initiated. The goal of repairing is to apply the principles of Conversation Analysis in order to bring the conversation back into a structure that allows to build shared meaning. Literature on the subject identifies preferred types of repair: the most common is the "self-initiated self-repair", then "other initiated self-repair" and, finally, "other-initiated other-repair". This implies that once a mistake is made in the conversation, either the same person who made it will notice it and start a repair attempt, or the first speaker should rely on the other's feedback to straighten back the conversation (Cawsey and Raudaskoski, 1990). Moreover, Schegloff (Schegloff, 1982) believes that the source of trouble can be divided into two groups: *problematic reference* and *problematic sequential implicativeness*. Reference problems could also be called *repair initiators*, since they happen whenever the second speaker signals that it is not clear what an expression is referring to. Sequential implicativeness, on the other hand, refers to the adjacent positioning of turns-at-talk and how this is used to make sense of the interaction. Every next turn shows an analysis of the previous one, giving the producer of the prior turn a chance to comment on the analysis. For this reason, a repair sequence is always activated after a first exchange. This intuition is particularly important when one of the speakers is a machine and therefore lacks self-consciousness.

These last considerations lay two important concept that shall serve as fundamental background to understand the shift from human-human interaction (HHI) to human-computer interaction (HCI):

- **Who can initiate a repair?** According to Cawsey and Raudaskoski (1990), the repair can be initiated either by the same person who committed the error that lead to the breakdown, or by the other speaker. This makes perfect sense in human dialogues; however when one of the speakers is a machine, it is hard to imagine a computer being so self-aware (at least of now, the year 2021) that it is able to self-correct. For this reason, in the contest of HCI, we shall always expect the *repair initiator* to be found in the second speaker's utterance (a.k.a. the human).

- **Where can I expect to find a repair initiator?** According to Schegloff (1982), the adjacent positioning of turns is what allows to make sense of the interaction. Therefore, a repair attempt can only happen after a first exchange between the two parties. For this reason, a repair initiator can be found only in conversation that have at least two exchanges (where each exchange is composed by a pair of utterances, one from each speaker), in the utterance by the second speaker in the second exchange.

3.2.1 From HHI to HCI

Although all these findings were elaborated by analysing human-human conversations (Purver, 2004), they can easily apply to human-computer exchanges as well[1]. In the case of interactions in a specific setting (e.g. between a client and a cashier in a supermarket; between a user and an IT technician in a computer repair shop; etc.) the human-human dialogue follows a sort of prescribed script that both parties adhere to and they expect it to roll out in a certain way. For instance, if a cashier were to ask a random question that is not related with the transaction process, the client would be (at the very least) surprised by it. This discrepancy between expectation and reality may trigger a repair attempt from the client. The same scenario is valid when one of the speaker is a task-oriented dialogue agent. Task-oriented agents usually have a narrow scope, because their goal is to guide a person or to provide information on a very specific topic. In the design phase the possible exchanges between the user and the agent have been scripted

[1]The work by Matthew Purver cited here deserves particular attention: Purver created a taxonomy of strategies by analysing a corpus of human-human dialogues, and then leveraged these findings to build a CA. Although his thesis is closely related to this one, the fundamental difference is that he searched for RS in HHI and then translated that to HCI. In this thesis, we start by analysing RS that already happen in HCI.

out, in order to endow the agent with the ability to answer in different topic-related scenarios.

Since task-oriented systems usually follow a predetermined path, it is even more feasible to detect recurring errors in conversations by analyzing existing data: if similar conversation turns always trigger a repair sequence from the second speaker, then problematic patterns become evident (Aberdeen and Ferro, 2003; Green et al., 2006; Colman and Healey, 2011). Recurrent error patterns become predictable and thus automatically detectable.

Since a machine is not self aware, it should rely on human feedback to spot a breakdown in their interaction. Most virtual agent will try their best in answering a question, retrieving an information or act upon an instruction. If they fail though, it is not always clear what will happen next: is the agent aware that something went wrong? Does it know what and why? Will the human counterpart be able to explicitly teach the system that?

3.3 Conversation Analysis and Dialogue Systems

Conversation Analysis has already stepped beyond the boundaries of HHI before. In the past, it was already deemed relevant for HCI in providing insights for Graphical User Interfaces (GUI) design and User experience studies (Luff et al., 1990; Moore, 2013). More recently, this approach has found its application in the analysis of dialogue systems, that is, the conversational interaction between a human and a machine. The machine can take the shape of a robot (Lee et al., 2010; Luria et al., 2019), a multi-modal interface (Bourguet, 2006; Bickmore and Cassell, 2005; Cassell, 2001) or a text-based back-and-forth dialogue system (Li et al., 2020a; Hussain et al., 2019; Thorne, 2017). Each of these systems interact with users in different ways and can therefore “fail” in various circumstances. Each failure entails the availability of certain RS and the absence of others. For instance, some RS are inherently tied to voice-based systems: it is the case when the conversational agent (CA) makes a fault in parsing the user’s speech (Beneteau et al., 2019; Myers et al., 2018; Porcheron et al., 2018; Myers et al., 2021). Users may try to employ different prosodies, accents or pitches in order to solve a misunderstanding with a voice-based CA.

Strategies can also be categorizes according to the main actor in the error-handling process (machine versus user) as well as the purpose of the strategy (error prevention, discovery, or correction) (Bourguet, 2006). In our case, we focus on the *users* while they are aiming at *correcting* a breakdown that happened in a conversation with a *text-based, task-oriented* CA. To sum up, we are interested in strategies found in a scenario with these features:

Actor: the RS must be produced by the human counterpart of the dialogue. The CA is not able to spot an error in itself and has to rely on the other speaker to make it evident. In this sense, the RS employed may always be considered a self-repair (Hough, 2014; Colman and Healey, 2011), since it is only the first speakers (i.e. the humans) that can correct their utterances;

Type of CA: we believe that in task-oriented agents, users have a paramount interest in solving their problem, since they have a clear goal (Motta and Quaresma, 2021; Li et al., 2020b, 2019). Therefore, they will try to repair the interaction in case of mistakes by employing different RS. This makes task-oriented CAs the most suitable for our study;

Channel: we explore the use of RS in the textual channel;

We now aim to search existing literature to check whether a framework, a precise methodology that would allow us to identify different RS has already been created and empirically validated. We would want to take from previous work, in order to reuse consolidated findings and possibly allow for comparison and reproduction of results.

3.3.1 Related Work: Repair Strategies and Tagsets

We found several existing works in literature that tackle the problem of detecting repair attempts in dialogue systems. Higashinaka et al. (2017, 2015) prove that error detection is possible in chat-oriented systems, and even propose a taxonomy of mistakes based on the utterance, response or context level in which they appeared. However, our focus is on task-oriented CAs, not chat-oriented ones. It is nonetheless useful to notice that in their most recent work (Higashinaka et al., 2021), the authors join two approaches from their previous articles: a theory-driven taxonomy of errors and a data-driven one. Their take is similar to ours, where we integrate results from a top-down perspective and a bottom-up one.

Moore and Arar (2019) propose one of the most complete list of RS that can be found in human-machine dialogues. They provide a Natural Conversation Framework that contain several categories of language patterns. RS are situated under the sequence-level management patterns, and they are broken down to five pattern types and 32 subpatterns. The five pattern types include ways in which the agent and the user can repair troubles in hearing or understanding immediately prior utterances or earlier utterances, as well as ways of ending conversational sequences. Even though the list by Moore and Arar is undoubtedly exhaustive, it does not fit perfectly into our case. First of all, it includes RS that can be activated by the agent, while in our work we state that by definition, a strategy can only be made evident by the *other* speaker of the conversation; that is, the human user. Moreover, they include speech related RS (while our channel of choice is the written one). Their list may even be too granular for our purposes; our final goal is to endow the CA with the ability to spot and categorize RS by itself. By distinguishing too many kinds of strategies, we may introduce unnecessary complexity. We did not employ their list as it was, although we did kept it and its related work (An et al., 2021) into account as a valid set of suggestions.

Beneteau et al. (2019) identifies several strategies by analysing interactions between family units and an Amazon Alexa device. Some of those strategies can only be found in voice-based systems CAs (e.g. *Prosodic changes* or *Increased volume*), but others may be applicable in text-based agents as well. In fact, strategies such as *Repetition* are not inherently tied to the vocal channel of communication and may be applicable to text-based agents as well. In particular, the *Repetition* strategy was also found in other works (Avdic and Vermeulen, 2020; Litman et al., 2006; Wang et al., 2021). Avdic and Vermeulen (2020) also highlighted a "stop" strategy, similar to the *Closing* one identified by Moore and Arar (2019). Even though it does not seem to exist a unique and complete framework of RS in literature, several strategies are found in different works. If a RS is constantly found in distinct types of CAs, then it is probably a constant in human behavior and should be applicable to our case as well.

Bourguet's taxonomy of errors divides the strategies according to four quadrants (Bourguet, 2006). Our quadrant of interest is the *User correction* one, where we can find RS like *repeat*, *rephrase* and *spell out*. Once again, it is possible to notice the presence of RS already found in other articles. While all the aforementioned works extract their results from the observation of existing interactions, they usually do not organize them in a precise set, where each one of them is clearly identified and can be extended to unseen datasets. In order for the CA to be able to spot RS on its own, we need to define a precise set of possible repair attempts. In other words, we need a *tagset*, where each tag represents a RS.

In this sense, the MALTUS tagset provided a solid base of different tags (Popescu-Belis, 2008). MALTUS was deemed to on of the most useful works for our purposes, since it provides distinctive tags to differentiate various kinds of reaction to errors in a conversation. Other works also used similar kind of tags but they were not as structured as the MALTUS one (Batliner et al., 2003; Stolcke et al., 2000), or contained less tags (Lopes et al., 2015; Cevik et al., 2008). Not all the labels in the MALTUS tagset were deemed applicable in our context, while some were considered too granular (such as the "RIC" and "RIR" tags, that will be incorporated into one class in our novel framework).

Allen and Core (1997) propose a handbook for annotating dialogues. Their tags mark utterances' role in the conversational exchange and their relationship to each other, in a very high level manner; therefore, they do not aim at identifying more fine grained phenomena such as repair strategies. For this reason, Allen and Core's work was taken into account for the development of a RS framework, but it was modified according to our needs. The authors themselves say they would expect the annotation scheme to be refined for specific tasks, in order to be more descriptive of specific features of the conversation.

Finally, the work by Ashktorab et al. (2019) must be mentioned. It is particularly interesting because, at a first glance, it provides the complete framework we were looking for: it outlines several repair strategies and tests various methods to respond to them. However, their approach is actually orthogonal to ours. It focuses on preemptive repair attempts made by a CA when it notices a potential problem in its own way of presenting information, rather than reacting to a manifestation of incomprehension from the user. Even though their intuition is original and ingenious, we argue that in HCI repair strategies are something that can only be noticed by the counterpart in the dialogue (Schegloff and Sacks, 1973). An utterance can be marked as breakdown in the conversations not because it pertains to a specific class of erratic statements. It is the other speaker in the exchange that must highlight the presence of an incomprehension and act upon it, in order for his/her own statement to constitute a RS (Norman and Thomas, 1991). Benner et al. (2021) adopt a similar approach: they conduct a literature review to understand how the concept of recovery strategy has been highlighted in studies around dialogue systems. In their analysis they identify six small taxonomies of strategies; however, these are the strategies that the CA may employ while trying to recover from a breakdown. In our case, we are looking for the repair attempts put into place by users: that is, when (and how) the human speaker signals an error in the conversational flow.

Table 3.1 sums up the contributions and limitation of the aforementioned literature with respect to our purposes. Table 3.2 break downs the features of each single work in an analytical manner.

Whilst the concept of RS is not completely absent from the literature in HCI and dialogue systems, none of the cited articles offers a complete classification of the strategies a user may employ to try to repair an interaction with a written, task-oriented chatbot. Most of them provide a qualitative description of RS that can be found in a dataset (Li et al., 2020a; Bourguet, 2006), but those RS are not systematized in a clear tagset that can be used to automatically classify them. For instance, several works (Popescu-Belis, 2008; Bourguet, 2006; Moore and Arar, 2019; Beneteau et al., 2019) assert that *Repetition* is a strategy. However, they do not provide an explicit definition of what a "repetition" is, nor do they differentiate it from other RS in terms of relationship with other utterances of the dialogue, or users' behavioral implications when they employ such strategy, etc.

Since our ultimate goal is to equip the CA with the ability to spot the RS automatically, we need a clear set of tags that can categorize RS. Therefore, we created a novel tag system where each RS is individually represented and differentiated from the other ones. Our tagset takes from the previous work cited in this section by incorporating the most frequent strategies that were found in literature, while also introducing new tags that were empirically found in our dataset of choice.

Related work	Notable features
Higashinaka et al. (2017, 2015)	They propose a taxonomy of mistakes (and subsequent RS) based on the utterance, response or context level in which they appeared. However, their focus is on chat-oriented CAs.
Moore and Arar (2019)	They provide an exhaustive list of RS. However, some of them are supposed to be activated by the agent, while our focus is on the user's repair attempts. Moreover, their categories may be too granular to be useful for an automatic classification.
Beneteau et al. (2019), Avdic and Vermeulen (2020), Litman et al. (2006)	They are mainly focused on voice-based RS. They identify strategies such as *Repetition* and *Closing* that were also found in other works and are not just strictly correlated with a vocal interaction.
Bourguet (2006)	It divides RS in four different quadrant. In the *User correction* one, some strategies were present that were also found in other articles (such as *Repetition*).
Popescu-Belis (2008), Batliner et al. (2003), Lopes et al. (2015), Cevik et al. (2008)	These works provide tagsets to identify errors and strategies in human-machine communication. Popescu-Belis' one was deemed to be the most complete and closest to our purposes, although not all of its tags were applicable to the contest of RS for our CA of choice.
Ashktorab et al. (2019), Benner et al. (2021)	They outline several repair strategies and test various methods to respond to them. Their focus is actually on preemptive repair attempts made by a CA when it notices a potential problem in its own way of presenting information, or on the strategies employed by the CA after the error has already been made evident. We argue that in our context of interest, a RS can only be activated by the human user and we are interested in detecting when and how the user signals the breakdown.
Allen and Core (1997)	Their tags identify the characteristics of utterances with respect to their communicative functions. Their tags are however at a too high level and they are are not tailored to the detection of RS.

Table 3.1: Contributions and limitations of the existing literature on the subject of clearly identifying RS in task-oriented, text-based dialogue systems.

Related work	Voice-based	Taxonomy	Common RS	Task-oriented
Higashinaka et al. (2017)	X	✓	X	X
Higashinaka et al. (2015)	X	✓	X	X
Moore and Arar (2019)	✓	X	X	✓
Beneteau et al. (2019)	✓	X	✓	✓
Avdic and Vermeulen (2020)	✓	X	✓	✓
Litman et al. (2006)	✓	X	✓	✓
Bourguet (2006)	✓	✓	✓	✓
Popescu-Belis (2008)	✓	X	✓	NA
Batliner et al. (2003)	✓	X	✓	✓
Lopes et al. (2015)	✓	X	✓	✓
Cevik et al. (2008)	✓	X	X	✓
Ashktorab et al. (2019)	X	X	X	✓
Benner et al. (2021)	✓	✓	✓	NA
Allen and Core (1997)	✓	X	✓	NA

Table 3.2: Analytical break down of features from each work. The *voice-based* feature is to be considered opposed to the concept of *text-based*; if a work does not present a *taxonomy*, then it presents a non-hierarchical *list*. *Common RS* means that work identified one or more RS that also appeared in other works. Finally, the contrary of *task-oriented* should be considered an article that deals with *chat-oriented* CAs. *NA* stands for Not Applicable.

Chapter 4

A New Taxonomy of Repair Strategies

The aim of this work is to apply sequential implicativeness in order to recognize every users' input that initiates a RS, since in the context of HCI, we postulate that it is only from the second (human) speaker reaction that we can notice a derailment in the conversation (Schegloff et al., 1977). While it is important to identify different nuances of RS, one of our goals is to endow the CA with the ability to differentiate the strategies on its own. In order to do so, a clear set of RS must be outlined for the system to learn from.

In the previous section we analysed existing literature in search of such set. Several works tackle the issue of RS in CAs, however, we could not find a complete and unambiguous collection of strategies to apply to our own purposes. Thus, we created a novel tagset by applying a *top down* approach together with a *bottom up* one: the top down approach ensures that strategies discovered in previous articles are incorporated into our model, while the bottom up one provides suggestions from a dataset of human-machine conversations. The bottom up approach guarantees that each repair attempt initiated by a user is properly considered.

This methodology allows to take into consideration all the possible sources to properly establish a set of RS. However, it does not mean that all the sources provide same-level information. We believe that the new tagset should reflect this difference.

This chapter explores the creation of our novel taxonomy. It describes the dataset used to conduct the *bottom-up* evaluations and it explains in detail the function of each level of the hierarchy.

4.1 Reasons for a Taxonomy

Most of the current tagsets are "flat": that is, they are a list of tags each at the same level, with no differentiation or arrangement between them (Popescu-Belis, 2008; Litman et al., 2006; Ashktorab et al., 2019). We argue that this representation does not account for the different communicative implications of each RS (Bourguet, 2006). Moreover, it may give the impression that there is no distinction in usage between them. We believe that not all users will employ every possible RS at any given time, but they will rather select the more appropriate strategy according to their own cultural and behavioral references. This intuition can be better understood by looking at two of the strategies that emerged from previous work: *Repetition* (Popescu-Belis, 2008; Bourguet, 2006; Moore and Arar, 2019; Beneteau et al., 2019) and *Closing* (Moore and Arar, 2019; Avdic and Vermeulen, 2020). The first one implies that users, faced with a breakdown in the conversation, will not quit. They will assume the responsibility of repeating or rephrasing their own sentence, in order to try a different path that may be more congenial to the CA. On the other hand, a Closing strategy signals the users' will to abandon the conversation; they will not try again, as they would rather surrender and give up on the information they were searching for. Each of these two RS entails a different behavioral perspective: the first one describes a situation where users are interested in bringing the conversation back on the right track and they are willing to modify their own behavior to attain that goal; the second one depicts users who do not feel like it is worthy to try again, they prefer to stick with their problem rather than keep talking with the CA. We believe that this root difference between the strategies should be accounted for. For this reason, we organized our tagset in a hierarchical structure composed by three levels. Structuring the tags in a taxonomy serves several purposes:

- It gives depth to the strategies, by specifying implicit users' characteristics that each RS entails;
- It can paint a picture of the users who are interacting with the CA: is the majority employing a Repetition kind of strategy, or rather a Closing one?
- It starts from two macro groups and then details each RS's specificity. Such organization allows the transfer of the tag system to other domains by doing minimal changes to the leaves of the tagset, while the branches would still be applicable.
- In the perspective of developing an automatic classifier of RS, a taxonomy of tags is more useful compared to a flat one. The classifier could proceed level by level, rather than just predicting the final label. Moreover, each level could benefit from a different network architecture, according to the information it is encoding. In the case of a flat tagset, one network would have to be a good fit for all the RS in the same way.

While the first and second level were informed by the literature on the subject, the third one (that is, the level that contains the actual RS labels) was created via the mixed approach described above. Some strategies were inserted top down, while other resulted by the analysis of a dataset of conversations. Before diving into the details of each level, we describe the dataset, in order to provide perspective into the definition of the bottom up RS.

4.2 The Dataset

Since our goal is to improve the interaction between users and dialogue systems, we were primarily interested in a set of conversations where there was indeed space for improvement. We therefore searched for a dataset where a CA would make mistakes and the humans would try to correct them. As it is stated in Chapter 2, we believe that users who interact with task-oriented agents are invested in the well being of the conversation (Motta and Quaresma, 2021; Li et al., 2020b, 2019), since they have a goal they want to reach. For this reason, they will attempt to repair the exchange if a breakdown would occur.

A first search conducted in the summer of 2020 yielded various datasets that are publicly available. Several of them only contain synthetic data (Boureau et al., 2017; Shah et al., 2018), meaning they were not actual conversations that happened in real life, or they contained conversations between humans, either spontaneous or set-up (e.g. movie subtitles datasets). Since we were looking for human-computer exchanges, we had to exclude them a priori (Lison and Tiedemann, 2016; Eric et al., 2017; Lowe et al., 2017). In order for the dataset to contain mistakes and subsequent repair attempts, we needed dialogues between a human and a machine in a real world setting. In a controlled environment, the chances of errors (and the users' intention to repair them) are lower. Most publicly available datasets are obtained via the Wizard-of-Oz technique (Eric et al., 2019; El Asri et al., 2017); if the machine part is simulated, and a human is actually behind it, it is reasonable to assume that the responses will be correct. In fact, the WoZ dataset are often created to model ideal conversations, therefore they are free from mistakes by design.

4.2.1 Limitations of Existing Datasets

Our search finally uncovered two possibly viable datasets: the *Let's Go!* dataset (Raux et al., 2006; Black et al., 2011) and the Microsoft's *SMCalFlow* one (Andreas and et al., 2020). We explored them both in order to asses the most useful one to our intents.

Let's Go! is a spoken dialog system that was used by the general public. It gave bus information scheduling for the Allegheny County Port Authority Transit bus system via a telephone-based interface to access bus schedules and route information. The integral Let's Go dataset has 171,128 spoken dialogues and only a portion of it has been transcribed. Since it contains recording of spoken commands given over the phone, The Let's Go! dataset contains very short inputs from the user and, since the conversation is voice based, it can be hard to interpret the overall meaning of a dialogue. Moreover, where the conversations were not manually transcribed, Automatic Speech Recognition errors can add another layer of difficulty in interpreting the user's input. That means, it was hard to understand whether the user was trying to repair a mistake or just saying a complete new sentence.

SMCalFlow is a large English-language dialogue dataset, featuring natural conversations about tasks involving calendars, weather, places, and people. It contains over 40,000 annotated conversations where users express their need using various linguistic forms and registers. The SMCalFlow dataset was closer to our necessities, because users were asking the system to make operations by employing longer sentences. However, the dataset was collected in a controlled environment. Even though the authors stated that it was possible that the system misinterpreted the user's request and those breakdowns were not excluded from the corpus, we conducted a short reading of the dataset and it emerged that those errors were not gonna be abundant. The users in the SMCalFlow dataset did not have a real necessity, an actual task to carry out, because in an experimental setting users do not really need to accomplish anything via a conversation. Therefore, they tended to be more tolerant compared with users in the real world. They also probably had different expectations about the system: SMCalFlow users knew they were testing a prototype. This mean that they would not even try to signal the breakdown in order to repair the interaction; they would simply pass on to another subject.

To sum up, we were looking for a dataset with these features:

- It had to contain dialogues between a human and a CA. We are not interested in human-human conversations, therefore we excluded datasets such as the collections of movie subtitles.

- In order for the dialogues to contain actual repair attempts, they had to be "real", which means we could not accept data extracted from simulations. For this reasons, we had to reject all datasets created with the WoZ technique.

- We are searching for a repair attempt in the user's utterances. The utterances should have a certain length and the conversation should last a few turns in order for the attempt to be evident. If the user's input is too short, or the conversation lasts less than a pair of exchanges, it is very hard to not only find a RS, but also to distinguish it from a simple statement. Therefore, we search for datasets where the conversations are not just made up of single commands, but where actual back-and-forth dialogues can be found between a machine and a human being.

4.2.2 The SisBot Dataset

Ultimately, none of the publicly available datasets proved to be apt to our purposes. At this point, we contacted a private company that had deployed a textual CA to the public.

We obtained a corpus of conversations pertaining to the SisBot project [1]: a task-oriented dialogue system that answers questions from customers about electronic invoicing and surrounding topics. SisBot's users are business consultants or small traders that manage their own invoices through the a specific platform. While working, they may incur in some doubts or problems. They can click on the bot logo and chat with the CA to ask for clarifications. Figure 4.1 shows how the agent appears on the platform web application.

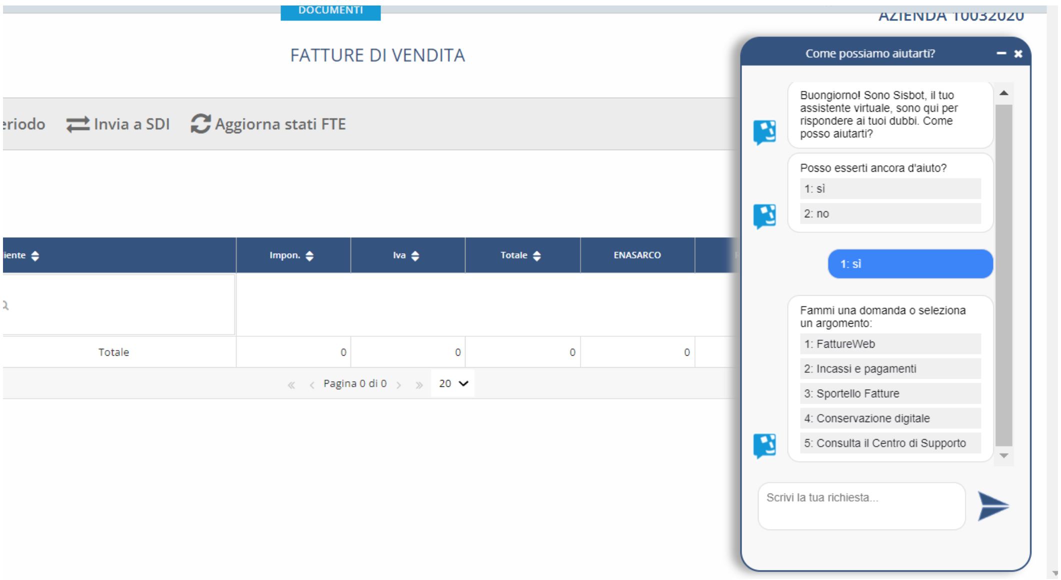

Figure 4.1: Once users access the invoicing platform on their browser, they can click on the SisBot logo and chat with it.

Most of the users reside in Italy and they employ Italian as their everyday language; for this reason, the SisBot CA is equipped to answer in that language [2]. This new dataset respects all requirements previously listed: it is a set of conversations between humans and a machine, where users have precise goals they want to attain; the conversations were gathered in a real world setting, meaning that they are potentially rich in repair attempts and reflect the users' spontaneous behaviors.

In the SisBot CA, the interaction is stateless and as such it can be assimilated to a question-answering system (Alloatti et al., 2019; Bianchini et al., 2017). It is the user who must initiate the exchange: once the user asks a question, the CA leverages its NLU module and searches for the most appropriate answer in its database. It will express uncertainty if it is not confident enough, or provide the user with a set of options

[1] The SisBot project is a dialogue system deployed by the Italian company Sistemi: https://www.sistemi.com/. We are grateful to their contribution to this research effort.

[2] While the original language of the dataset is Italian, the examples quoted all along this thesis have been translated to English in order for them to be widely understandable.

in case different answers are computed as to be equally valid. The CA is made available for free to all users of the invoicing platform: for instance, during the year 2020 the CA delivered 817,000 single messages to 117,000 different users. Even though users need to be authenticated to access the platform, their data is not passed on to the CA, as it does not need it. As of the moment the dataset was gathered, all of its answers were not dependant to whom was asking the questions: it means that the responses are not customized at all. Its sole purpose is to deliver information that is technically correct in nature, and such answers do not change according to the user.

While the CA does not require personal data to operate, some users may spontaneously share private information while writing. Such information were anonymized in accordance with European data protection laws before further operating on the dataset. See the following example:

> **User**: Good evening
> **CA**: Nice to meet you. What can I do for you?
> **User**: The system tells me that the invoices issued on 03/20 to XXXX have not been delivered. Can you help me ?

Even though the name of the invoices' recipient is not relevant to the system's functioning, the user stated it anyway. The name has therefore been anonymized.

More than a hundred thousand unique users constitute a significant quantity that is usually not reachable in controlled experiments. Such a large public entails a range of demographics, linguistic competence and technical expertise among those who interact with the CA. Although the conversations are anonymous and no information is available about the users, this loss is compensated by the fact that the dataset was gathered in a real life setting. "True" conversations represent different behavior and therefore validate the reliability of our evaluation of the users' behaviour.

The raw dataset was obtained by filtering the conversations of the year 2020. Only the conversations with at least 6 messages from the user were included; this quantity was deemed to be the minimum necessary to contain a meaningful exchange - that is, where a mistake could have occurred, together with its subsequent RS. The final dataset is composed by 142,607 rows, grouped into 15,585 conversation sessions (CSs). Since the dataset is in a tabular format, each row contains the question from the user plus the CA's answer. A CS is an exchange between the CA and a specific user over a certain number of rows, a back-and-forth interaction in a limited time span. Each session can contain multiple messages and the total number of analysable messages is 46,916. "Analysable messages" are those produced by users in natural language: therefore not the ones outputted by the CA, or the ones who contain coded information (such as API calls or the result of pressing a button). A RS can indeed be found only in a user's input in natural language, while an input that derives from the pressure of a predefined option, cannot be a RS. Table 4.1 summarizes all the relevant statistics to this dataset. It is worth mentioning that the dataset has not been previously used for any other study, and it is therefore not equipped with any previous annotation.

Feature	**Number**
Rows	142,607
Sessions	15,585
Minimum number of messages per conversation	6
Messages in Natural language	49,916
Average length of a message	38,7 characters

Table 4.1: This table contains an explanation for each one of the strategies on the third level of the tagset. We also provide one or two examples per RS taken from the SisBot dataset.

We used this dataset as a basis for our bottom up approach. We read a portion of it in search of repair attempts that could enrich the list of those already found in existing literature. The results of this preliminary analysis informed the creation of our final tagset.

4.3 The Tagset

The tagset was built iteratively: we first jotted down the strategies that were found in the literature. Then, we progressively checked the dataset and inserted newly found RS in the nascent tagset structure. We refined it through various cycles of analysis, in order to make sure that all the strategies found in the dataset were represented, and their organization reflected some implicit information about the users' behavior.

Figure 4.2 shows the complete structure of our new tag system. The hierarchy stems from one block, *Breakdowns*: it is worth remembering that these tags shall be applied only to repair attempts, e.g. when the user signals that a breakdown is occurring, and not to all users' sentences in the conversation. In this sense, the word *repair* is used in what could be a broader sense compared to the classic theory of Conversation Analysis: here, *repair* refers to all act through which users attempts to straighten the conversation. Some of these attempt may be more productive and cooperative (e.g. Repetitions), while other may be less so (e.g. Insults). Nonetheless, in both cases users are actively trying to do something about the breakdown that just happened in their interaction with a CA, they are not simply giving up and abandoning the exchange altogether. For this reason, we include all these attempts, that we call *strategies*, in our taxonomy.

As mentioned before, the first and second level were mainly informed top down; that is, from findings in previous work. On the other hand, the third level, which is

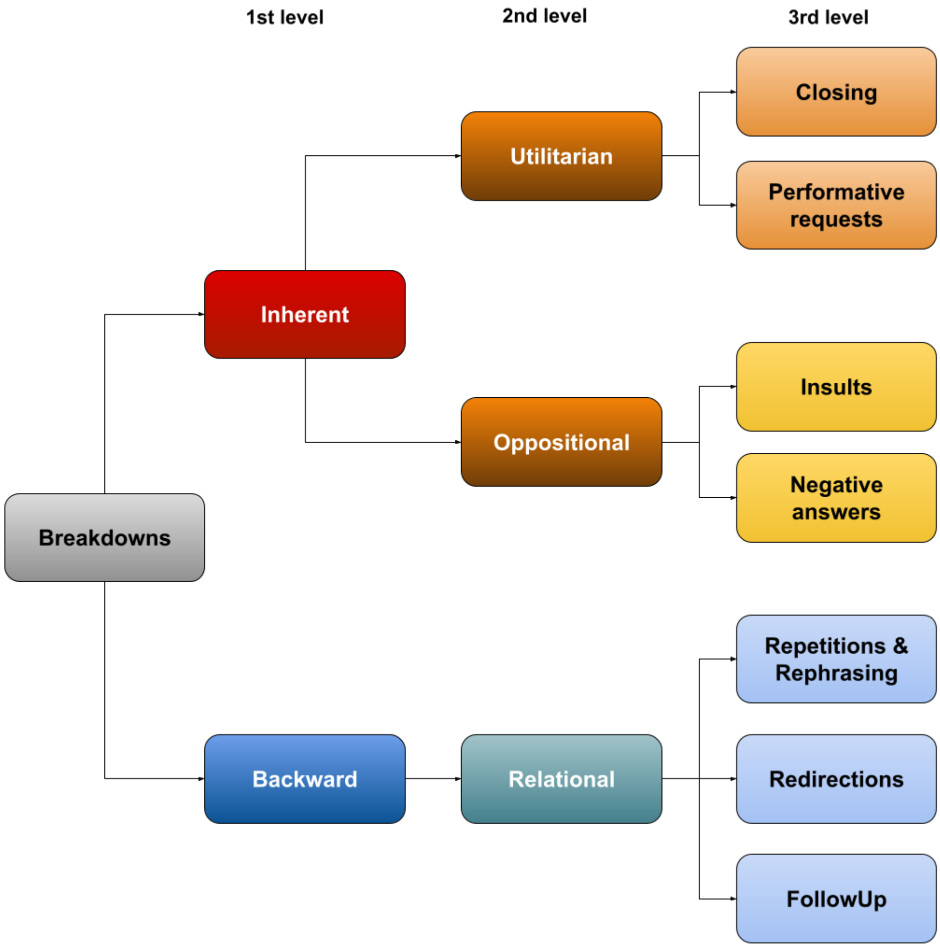

Figure 4.2: The figure shows the complete tagset that we created to detect and categorize RS. The actual RS are on the third level and they are grouped according to their communicative entailment in two higher levels.

composed by the actual RS, combines the top down approach with a bottom up one. We included strategies that were already highlighted in existing literature together with others that emerged from a preliminary analysis of the dataset. The analysis also confirmed the presence in our dataset of the top down RS. Specifically, strategies such as *Repetitions & Rephrasing*, *Closing* and *Negative answers* were inserted in the tagset via the top down approach, but were also encountered in the conversations from the SisBot project, thus receiving a double validation. The other RS in the third level form a novel contribution. They represent various ways through which users signalled a breakdown in the conversations with the CA. Table 4.2 details each of them and provides empirical examples from our dataset.

Strategy	Meaning	Examples
Closing	The users' intention is to close the conversation in a neutral way. They simply do not want to go on with the conversation.	*Close the chat; Exit*
Performative Requests	Requests for the CA to commit an action; specifically, to pass over the conversation to a human operator, or a technician.	*I want to talk with a human; Is there a technician I could talk to?*
Insults	Plain insults, including ironic or cruel comments.	*You're stupid; Yeah this was super useful...*
Negative Answers	The user produces a dry and short negation, without further details that could result in a different kind of strategy.	*No; That's wrong*
Repetition & Rephrasing	Users repeat or rephrase a request which was not previously understood, in the attempt of finding a formulation that the CA will be able to parse. In the examples, the *second request* is the actual RS.	First request: *I need to cancel an invoice*; second request: *How to erase invoices*
Redirections	The user changes topic abruptly when faced with a mistake. The functioning is similar to the previous strategy, where the RS is tied to the previous request; however, in this case the strategy concerns a different topic and it does not repeat the same one.	First request: *I need to cancel an invoice*; second request: *Nevermind, tell me where I can find my password*
FollowUp	This strategy enchains on what the CA said immediately before. It contradicts or simply refers to a specific portion of what the CA just wrote, in order to highlight the breakdown.	The chatbot says: *You can find your new invoices at this link*; the user says: *Ok but what about the old ones?*

Table 4.2: This table contains an explanation for each one of the strategies on the third level of the tagset. We also provide one or two examples per RS taken from the SisBot dataset.

We also provide longer snippets of dialogue in order to give more context to the examples presented in Table 3.1. In the examples, the *U* stands for User, while the *S* stands for System, i.e. the CA:

Closing

U: Close
S: I'm sorry, I don't understand. Can you repeat that?
U: *Close the chat*

U: Close the interface
S: I'm sorry, I'm not an expert on this topic.
U: *Exit*

Performative requests

U: Problem
S: You may find some issues if the traffic on the server is high [...]
U: *I want to talk with a human*

U: How can I talk to the tax agency cause there's this error that I don't understand
S: If you need to manage a travel agency, click here
U: *Is there a technician I could talk to?*

Insults

U: I need a lot of time to issue an invoice
S: To print an invoice, you need to click on the pencil symbol [...]
U: *You're stupid*

U: How can I talk to the tax agency cause there's this error that I don't understand
S: If you need to manage a travel agency, click here
U: *Why would I need that?? You're saying random things*

Negative answers

U: I need some help with my invoices
S: Here's all I know about invoices: [...]
U: *No*

U: The system says there's an error with what I wrote
S: If you see the error n. 250, it means that [...]
U: *Thats' wrong*

Repetition & Rephrasing
U: I need to cancel an invoice
S: To print an invoice, you need to click on the pencil symbol [...]
U: *How to erase invoices*

Redirections
U: I need to cancel an invoice
S: To print an invoice, you need to click on the pencil symbol [...]
U: *Nevermind, tell me where I can find my password*

FollowUp
U: I'm looking for the invoice n. 34 in my account
S: You can find your new invoices at this link [...]
U: *Ok but what about the old ones?*

The strategies on the third level, a.k.a. the tags that categorize the different RS, are grouped in three classes: *Utilitarian*, *Oppositional* and *Relational*. These three classes aim at dividing the RS according to the behavioral model they entail. Previously in this chapter we explored the implicit divergence in style between those who would employ a *Closing* strategy compared to a *Repetition* one. The second level makes this characteristic explicit by applying the behavioural orientation classification by Ringberg et al. (2007), where they address three peculiar cultural models to which users adhere through their behavior:

- **Utilitarian**: the utilitarian model embraces a rational perspective. Users who embody this model do not perceive the CA's failures as a personal attack nor view them as indicative of some kind of antagonism. Failures are indeed regarded as simple inconveniences; these users would make note of the failure and try to move on to some other system (e.g. a human operator) that could help them.

- **Oppositional**: the oppositional cultural model evokes a consistently aggressive position whenever users experience a failure in the service they are using. Those who embody this model tend to blame the CA for the malfunctioning by expressing their discomfort via non-constructive criticism.

- **Relational**: the relational cultural model pertains to users who desire to maintain emotional ties with the provider, even in the face of adverse events. In this context, people who employ relational strategies are those willing to change their own behavior to help the CA, rather than blaming the agent for its failure.

Therefore, the utilitarian model is realized by the *Closing* and *Performative Requests* tags: those who employ these strategies do not care if the CA has failed, or a breakdown has occurred in the conversation. They just want to move on, either by abandoning the chat or by appealing to human intervention. They express their will in a neutral way, without resorting to insults.

On the other hand, the *Insults* and *Negative answers* tags fall under the oppositional behavior, since they represent an adverse attitude. Users who pertain to this model will react aggressively in case of a mistake, either by insulting the CA or by stating their denial in a non-cooperative manner.

Finally, the tags *Repetitions & Rephrasing*, *Redirections* and *FollowUp* belong to the relational cultural model. These strategies are employed by users who are willing to modify their own way of talking in order to repair the interaction. They will not give up, as they will rather repeat their sentence, try to find a different subject the CA is expert of (by redirecting the conversation), or carefully challenging the CA's last answer.

While the second level describes behavioral characteristics that each RS implies, the first level split the tags in two groups according to their relation with other elements of the conversation. In fact, some RS can present themselves isolated in the conversation, as they do not depend on anything that has been previously said: a user may decide to close the interaction or insult the CA at any moment and it would not rely on something uttered before. Put another way, in these cases there is no contextual dependence of a RS with the previous utterance.

On the other hand, the *FollowUp* tag strictly depends on what the CA uttered immediately before. It would not be possible to detect a FollowUp strategy without reading the rest of the conversation, since the FollowUp confronts a bit of information that can be found in the (immediately preceding) CA's message. Similarly, a Repetition entails the fact that the RS is repeating (or rephrasing) something already written at least once. To sum up, an Insult is self evident; a Repetition is only so if we are aware that it is repeating something previously uttered.

For this reason we grouped the RS in two communicative functions, following the work by Allen and Core (1997). The utilitarian and oppositional tags do not require any specific context in order to occur and were therefore placed under the *Inherent* function. The relational tags pertain instead to a *Backward* function, since they refer to something uttered back in the rest of the conversation.

The differentiation between Inherent and Backward is particularly useful in terms of developing an automatic classifier. In case of the Backward tags, the classifier would have to gain knowledge about the previous steps of the conversation in order to effectively spot a RS. That would not be necessary for Inherent tags; such a classifier would not need to keep memory of several messages, as each message is self revealing of its content.

The purpose of the tagset is to clearly classify RS that can be employed by users while they interact with a task-oriented CA. Since our final goal is to “teach” the CA to spot such attempts by itself, we want to develop an automatic classifier. The classifier needs to learn from data that has been previously labelled, possibly by humans: in this way, the annotation would be of very good quality. Therefore, the classifier needs a batch of conversations analysed and tagged by human experts, where the RS present in that batch are marked with the tags of the tagset. The next chapter describes the process of manual tagging, as well as the hypothesis and results that emerged from the analysis of the manual annotated data.

Chapter 5

Detecting Repair Strategies: Tagging and Analysis

Once the new tagset is created via the two approaches previously described, our goal is to apply it to our SisBot dataset. Tagging the conversations would serve two main purposes: i) it would provide high quality data for the classifier to train on; ii) it would confirm some of the intuitions upon which the tagset was built. The first purpose is quite straightforward: in order to develop a classifier, it needs labelled data to train on. Human annotations produce high quality data, thus facilitating the development process. The second purpose may benefit from some in-depth considerations: while building the tagset, we implicitly made some assumptions about the RS, as well as their entailment. Even though such assumptions are justified by related work, our hope is for them to be confirmed by an empirical analysis of the RS in our dataset. For instance, the second level of the tagset groups together different RS under the premise that they entail a behavioral pattern. We therefore expect the RS of the same model to co-occur in the same conversation. It is possible that a user may employ a Repetition and then, out of frustration, resort to an Insult or Closing strategy, but we do not expect users to jump from one model to the other constantly. In this chapter we describe the tagging process and then check on our hypothesis. The analysis confirmed the assumptions that were made in the organization of the taxonomy; moreover, the tagged data will serve as training for the classifier (described in 6).

In order to confirm or deny them, we drafted some specific theories we wanted to check. Our first hypothesis is formulated as follows:

H1: users manifest a certain behavioral orientation when interacting with a CA, therefore they will preferably employ RS pertaining to the same model (e.g. the second level of the tagset). Change from one model to the other is possible, according to several factor, but not that frequent, as they will tend to maintain their orientation -at least - throughout the current dialogue session.

To further clarify H1, it shall be said that we do not expect users to be consistent "with themselves" across all interactions they have with a CA, and not even with all CAs. Each person may show a different attitude according to a variety of factors - stress, time constraint, general mood, etc. But we do expect to see a certain consistency in the span of a single session: meaning, people will belong (or choose to employ) one of the three models showed in the taxonomy, and stick with it for the current session.

Based on findings in existing literature, we also assumed that users of task-oriented agents have a paramount interest in reaching their goal, which may not necessarily be true for users who interact with chat-oriented systems (because they *lack* a specific goal to begin with). For this reason, we expect users to primarily employ strategies pertaining to the Relational model. The Relational RS imply the user's will to continue with the conversation, even when a breakdown occurs, because they need to obtain the information they are looking for and they will not abandon the dialogue until the have achieved their objective. We thus outline the second hypothesis:

H2: in the context of task-oriented CAs, users are not casually chatting, whereas they are looking for specific information or trying to solve a problem. Facing a misunderstanding, they will insist on the conversation by employ Relational strategies. We expect an abundance of these RS compared to the other behavioral orientations.

For clarity purposes, it should be stated that in order for this hypothesis to be fully proved we should conduct the same analysis over a corpus of chat-oriented dialogue sessions. Even though several works that deal with task-oriented agents also noticed the use of Relational strategies (see Table 3.1 and Table 3.2), it does not necessarily mean that the same is not true for chat-oriented system. Therefore, in this thesis we will check whether there is an abundance of such RS in a task-oriented CA's dataset. We leave the counter-check on a chat-oriented dataset for future work.

In order to confirm or defy our hypothesis, we manually tagged a portion of the SisBot dataset.

5.1 The Tagging Process

Given the dimension of the dataset (15,585 conversation sessions), we annotated only a small portion of it, leaving the task of completing the tagging to the automatic classifier. The dataset is a simple Excel file, where each row is a message (either from the CA or from the user). A row also contain the session ID it pertains to, as well as a timestamp. The annotation was conducted by two expert annotators.

We split the dataset in two and each annotator was given a different portion. Each portion contains roughly half of the total number of CSs, organized according to their session ID. The session ID is a hash map generated automatically; therefore, from a practical perspective, the effect is of a randomized disposal of the sessions. The two annotators had to work in parallel and they were expected to annotate only a part of their batch, starting from the top. By organizing the sessions in such a fashion we made sure that each of them would annotate sessions pertaining to different time sets and of different length. Other ways of sorting the sessions (e.g. by timestamp) may have introduced unwanted clusters of phenomena in one or the other batch.

Both annotators have a formal training in linguistics and are familiar with the internal functioning of this CA: they know that each question posed by the user is parsed as a new input by the agent, since its memory of previous messages is limited. They also know that the CA will try to match the user's question to its internal knowledge and that its answers are pre-canned, therefore they shall not change from one conversation to the other. They were also informed about the context in which the conversations were gathered, in order to have a better understanding of the real users' point of view: people who interact with the SisBot are mainly small shop owners who have to manage their own invoices accounting.

The annotation procedure was carried out in three steps:

- The annotators had to read the conversations and mark every row that contained a repair attempt. In every row there is a message, either from the user or from the CA. Naturally, only the users' rows could contain a RS, but the annotators read the complete conversations in order to be aware of the general context. In this first phase the annotators were asked to simply mark with a "T" - as in `true` - every RS they found, without proceeding further.

- After the first step, the annotators would review each other's binary annotation. Since they had to work in parallel, on two different batches, it was possible that each one of them would interpret strategies differently. Thanks to these moments of comparison they could discuss each other's decisions and realign in case of misunderstanding of the task.

- Once they were both satisfied with the binary marking, they would proceed with the actual tagging. They would assign to each identified attempt one of the RS tags of the third level. Once again, they were given the opportunity to discuss among themselves specific cases where they felt unsure about the right tag to apply.

Every utterance entailing a user repair intention was thus identified and marked as a repair attempt by one of the two annotators. An utterance coincides to a user's message, a single row of the dataset. Each identified RS was then tagged according to the content carried by the statement. The tag refers to the whole utterance, as it is meant to characterize the whole statement in terms of what the user intended. Each marked utterance can be only be associated with a single tag, as at this moment, we do not foresee the necessity to use two tags to represent a repair attempt. Table 5.1 provides a few examples of tagged utterances. As it can be seen, each line of the table contains one message. For each line, the annotators first marked the repair attempts with `T`, and later specified what kind of RS was (by means of the tags).

The two annotators were given a month to annotate as many sessions as possible. During the month, they had two sessions during which they confronted their annotating process (the first dealt with the first, binary annotation, and the second revised the first applied tags). In total, they were able to tag 2,175 CSs. This amount of CSs contains 7,585 unique users' messages and 3,943 (52%) of those are marked as repair attempts. The presence or absence of RS over the total number of messages is almost balanced (48% absence, 52% presence). Nonetheless, the categorization of misunderstandings is quite varied.

5.2 Annotation Results

In this section we describe the results obtained by the round of manual annotation and discuss their implications. Our analysis aims at checking the hypothesis previously formulated. Since the sessions in the dataset are qualitatively randomized (given the uncontrolled environment of their occurrence, the anonymous character of the dialogues and the heterogeneous nature of such a large amount of users), there were no bias risks in the annotation exercise, meaning that the annotated sample can be qualified as representative for the whole dataset. We also computed the inter annotator agreement on a sample of 226 conversations (944 messages): the second annotator annotated 226 conversations that were previously tagged by the first annotator. No further moment of discussion among the annotators was carried out during this phase. The agreements resulted to be very high, with Cohen's Kappa coefficient value equal to 0.80 when dealing with the third level of the taxonomy (i.e. the actual tags), or 0.84 generalizing to the first level of the taxonomy.

At this point, we joined the two annotated batches to proceed with an analysis of the annotation results. The first thing that emerged is that the Backward function covers 43.06% of the breakdowns (equal to 3,266 messages) while the Inherent function only 8.92% (equal to 677 messages). Table 5.2 details the exact number of messages that have been tagged with a certain RS.

5.3 Analysis of the Results

The sample of annotated data provides some interesting insights into the strategies employed by users when facing a breakdown. The most frequent tag, by far, is Repetitions

Author	Message	Is RS	Tag
CA	Welcome! How can I help you today?	F	
User	What's the VAT code for the reverse change?	F	
CA	I'm sorry, I'm afraid I don't know the answer to this one. Refer to the Guide for more information	F	
User	But I still need an answer... I want to talk to someone	T	Performative requests
CA	Welcome! How can I help you today?	F	
User	Change the bank	F	
CA	Choose one of these options: i) How do I create a bank on the Invoice platform? ii) How do i create a bank in the CashIn platform?	F	
User	Replace bank	T	Repetitions & Rephrasing
CA	Choose one of these options: i) How do I create a bank on the Invoice platform? ii) How do i create a bank on the CashIn platform?	F	
User	`askQuestion|2`	F	
CA	Welcome! How can I help you today?	F	
User	Who are you?	F	
CA	I'm sorry, I'm afraid I did not understand.	F	
User	Can I assign the deduction?	T	Redirections
CA	If you need to issue an invoice to a flat complex, you must enable the deduction tax management [...]	F	
User	Invoices with fiscal deduction	T	Repetitions & Rephrasing

Table 5.1: Excerpts of conversation sessions (CSs) that contain a repair strategy (RS). In the second example, the user press a button to choose an option (it's the message `askQuestion|2`): those kind of messages can never be marked as RS, since only input in natural language can entail a repair attempt.

I level: function	**II level: behavioral orientation**	**III level: RS**	**Amount**
Inherent	Utilitarian	Closing	324 (8.22%)
		Performative requests	114 (2.89%)
	Oppositional	Insults	106 (2.69%)
		Negative answers	133 (3.37%)
Backward	Relational	Repetitions & Rephrasing	2681 (67.99%)
		Redirections	144 (3.65%)
		FollowUp	441 (11.18%)
Total number of breakdowns: **3943**			

Table 5.2: Distribution of tags, i.e. of RS, in the annotated sample. The percentage refers to the total number of utterance marked as a RS.

& Rephrasing, which is part of the Relational model. The second and fourth most frequent tags are also part of the same macro group, thus confirming our second hypothesis (H2): in the context of task-oriented conversational agents, users will try different paths before giving up on the agent, by modifying their own way of writing or by engaging in a close back-and-forth exchange with the CA, in the hope that the system would at some point be able to understand their request and answer appropriately.

The Closing tag is also quite frequent, although it is part of a different behavioral orientation. This RS is employed by users who want to quit the conversation, mostly because they have not found what they were looking for and they do not want to give the CA another chance. The Closing tag may happen after just one breakdown, or after several repair attempts; it is not completely unambiguous in its own nature. Its inherent applicability to various contexts partially explains its frequency: it may be found in conversations together with other RS pertaining to different behavioral orientations. In order to fully appreciate its presence in the dataset, we plotted a matrix of co-occurrence between the different tags. Figure 5.1 reports the normalized co-occurrence matrix. Two tags co-occur if they appear at least once in the same conversation session. The matrix was normalized via cosine distance: first, we calculated the cosine similarity between each pair of tags; then, we reversed it in order to obtain the distance. The resulting matrix thus represents dissimilarity, where higher values correspond to distant (i.e., potentially unrelated) tags. Intuitively, the diagonal that represents the crossing of the same tag shows a value of 0 in each cell.

The symmetric distance matrix can be displayed graphically by applying the Multidimensional Scaling (MSD) algorithm, as shown in Figure 5.2. Since the matrix does not show real distances, but rather a similarity metric, we preferred to use the non-metric MDS algorithm. We were thus able to preserve the ordering imposed by similarity and also obtained good results with relative low distortion (0.0215 stress value).

5.3.1 Analysis of Co-occurrences

With regards to Fig. 5.1, the intersection point of the NEG (Negative answers) and INS (Insults) tags shows a low dissimilarity value. This means that, despite being not very frequent overall, the two tags often co-occur in the same sessions. The matrix confirms the correct grouping of these two strategies under the profile of Oppositional users, who will employ those two tags together, while generally avoiding a rational or empathetic cultural model. Similarly, the R&R tag (Repetitions & Rephrasing) correlates more often with FWU (FollowUp) as well as with the RED (Redirections) tag. This data proves the consistency of yet another model, the Relational one. The co-occurrence of Relational RS describes the situation where users believe the CA did not answer correctly (partially or completely) and will therefore try to clarify their requests. They can do this both by rephrasing or repeating their requests, or by focusing their rebuttal on a specific element of the CA's answer, in the hope of eliciting a more specific explanation. Similarly, the R&R and RED combination depicts a scenario where users opt to change the topic of their request once the first repetition attempt was not successful.

The frequent co-occurrence of tags pertaining to the same behavioral orientation answers our first hypothesis (H1): users generally belong to one of the three cultural model and will employ tags consistently within their model, rather than shifting from one model to another. We can thus deduct that the use of certain RS highlights user behavioral orientation towards a CA, and that those orientations are homogeneous. This is particularly important in terms of providing a tailored response to the RS. If all users would jump from one strategy to the other, showing a high shift rate between different models, it would be hard to argue that certain users would benefit from a tailored answer (Ward, 2021). Instead, we can now assume that since Relational users are different from Oppositional ones, they would benefit from a response tailored to their cultural model (in the terms studied by Ringberg et al. (2007)).

5.3.2 The CLS Strategy

The Closing tag (CLS) seems to have a peculiar distribution compared to the other tags. Although it is quite abundant, it correlates very little with other RS in general and the sessions it has in common with R&R are fewer compared to the other strategies. Generally speaking, R&R is so much more present than the other strategies that it shall always correlate more with other tags; its abundance increases its probability of co-occurring.

Nevertheless, CLS is the third most frequent strategy. Numerically speaking, it should present a behavior similar to FWU. In order to explain this phenomenon, we went back to the dataset and read the conversations where the CLS strategy was found.

We noticed that users would often ask the chat to close itself, or to disappear, right after opening it. As it is stated in Chapter 3, a RS must be (at least) the second utterance of a speaker, since it aims at correcting the other speaker's response to the first utterance. In this case, users would just try to close the interaction, one utterance after the other. The annotators still marked the second one onward as a RS, since the CA did not understand the first time, but it is a peculiar kind of RS. The reason for this behavior is that users would often click on the CA button by mistake (or out of curiosity), but then they would not be able to close it back.

People access the SisBot via their browser: if the browser zoom is excessive, the closing button (a normal "X" on the top right corner of the chat window) would be out of the screen, thus impeding people to find it and close the chat autonomously. This small inconvenience explains the abundance of CLS tags, as well as its little correlation with other RS.

5.4 Discussion

The manual tagging revealed interesting facts about the users of the SisBot and their preferences of RS. We were able to confirm our primary intuition that users of task-oriented agents, when faced with a breakdown, will try several ways of repairing the interaction before giving up. We also validated our taxonomy of tags by analysing the co-occurrences of RS pertaining to the same behavioral model. Users that belong to the Relational or Oppositional behavior will preferably use strategies coherent with their model of reference. The Utilitarian model proved to be less consistent. This is because its Closing tag presented a peculiar behavior: a lot of occurrences, but a low correlation with other RS. The explanation for this data is found in a problem with the graphical interface of the CA, rather than a problem of the CA itself. It does prove though that with regard to HCI, the context in which the CA is placed (may it be a website, an app or a smart speaker) should never be disregarded.

This chapter serves two purposes: on one hand, it helps to confirm our intuitions towards the users of this dataset, as well as the construction of the taxonomy itself. On the other hand, it provides the initial batch of manual annotated data to train a classifier. In fact, we only manually annotated a part of the dataset; a big portion remains untagged and our goal would be to tag it automatically. In Chapter 6 we show how we built an automatic classifier that, leveraging the annotated data of this chapter, tries to generalize and detect RS in the untagged portion of the dataset.

	CLS	FWU	INS	NEG	PER	RED	R&R
CLS	0	0,8124	0,81905	0,83346	0,86204	0,84223	0,79078
FWU	0,8124	0	0,56049	0,58813	0,54777	0,56883	0,35504
INS	0,81905	0,56049	0	0,46939	0,55278	0,54537	0,38195
NEG	0,83346	0,58813	0,46936	0	0,48239	0,64604	0,43638
PER	0,86204	0,54777	0,55278	0,48239	0	0,58745	0,44671
RED	0,84223	0,56883	0,54537	0,64604	0,58745	0	0,31104
R&R	0,79078	0,35504	0,38195	0,43638	0,44671	0,31104	0

Figure 5.1: The matrix shows the correlation between two different tags. The numbers indicate the level of dissimilarity (i.e. scarce co-occurrence) between the two tags. Therefore, lower numbers indicate a higher co-occurrence, while higher numbers signal a scarce presence of those two tags in the same CSs. In the diagonal, the value shall always be 0: naturally, each RS co-occurs with itself.

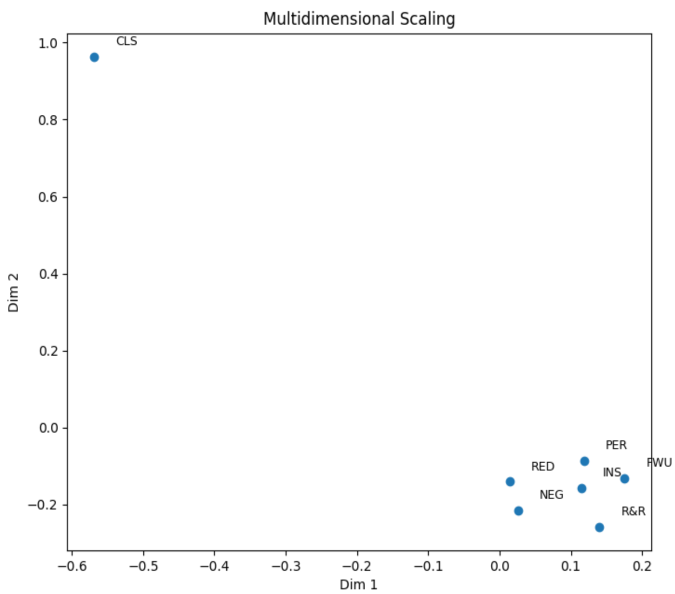

Figure 5.2: The figure shows the graphic display of the symmetric distance matrix. While most of the tags have comparable values of co-occurrence, the CLS tag presents a unique and peculiar behavior.

Chapter 6

Detecting Repair Strategies: Development of an Automatic Classifier

Once a specific tagset to mark RS has been created, and manually applied to the dataset, the following step is to try and automatize the process. This Chapter tackles the topic of automatically classifying RS in the part of the dataset that has yet to be annotated. We explored several architectures and network to this end, and described the results obtained by applying each one of them.

Classifiers are machine learning algorithms that aims at predicting the class of a given set of data. They separate the larger batch into sets, generally called classes, tags or labels. Classifiers can be used in several tasks, such as face recognition (Sharifara et al., 2014; Smach et al., 2006), object detection (e.g. in autonomous vehicles) (Hnewa and Radha, 2020; Arnold et al., 2019), as well as in NLP (Toba et al., 2014; Enríquez et al., 2013). A classifier can be supervised or unsupervised: the first one expects different data as input for each class, i.e. some examples for each label. It can then try to extract relevant features from those examples and generalize towards unseen data. The second one clusters unlabeled data: it expects raw data and a set of classes, and then it will try to group them into sets according to common characteristics. Supervised classifiers use a set of weights that combined with the current input, generate the correct label. If a classifier predicts the wrong class, the weights are changed in order to correct the mistake an label the input properly.

In the context of Dialogue Systems, a few works have been conducted to detect trouble in a conversation, although they mostly deal with speech-based interactions (Hough et al., 2015; Hough and Schlangen, 2017; Shalyminov, 2020; Rohanian and Hough, 2021). As it has previously stated, our focus is on agents that interact via written text. Other articles describe the task of producing clarification requests in open-domain agents (Min et al., 2020; Aliannejadi et al., 2021); while these latter constitute an interesting input, in this context we aim at detecting RS in a task-oriented domain, and not necessarily to answer them through a clarification request. The most interesting work is probably the one by Purver et al. (2018), where they produce a classifier to detect repair strategies in several corpora of human-human dialogues. Such corpora are publicly available and have been tagged in various ways; Purver and colleagues use the different annotations as features for their classifier. In our case, we relied on a proprietary dataset that had been annotated with specific tags. Therefore the need for an ad-hoc classifier.

The manual annotation described in the previous Chapter has a double function: on one hand, it allows us to validate our new tagset and gain useful insights into users' behavior. On the other hand, it provides the starting data to train a supervised classifier: each utterances marked with a certain tag constitutes an example for that tag's class. The tag describes the whole utterance, it does not point to specific words or structures in the sentence that express a repair attempt. It rather labels the intention of the user, which emerges from the complete message in its context. However, not all classes can simply be created by putting together all of its examples. In fact, the task of detecting different RS cannot be reduced to a simple classifier system.

In the context of dialogue systems, an utterance tagged as a RS is not independent from the rest of the conversation; it is a fundamentally context-dependent phenomenon and its classification would depend on some of the prior turns, or dialogue history. By definition, a repair attempt is the speaker's reaction to a breakdown that has happened in the previous exchanges (see Chapter 3 for more details). This intrinsic features of RS guided the construction of our hierarchical tagset: the tags that do not depend strictly on previous messages in order to be correctly recognized and marked are organized under the Inherent level; those that cannot be categorized as RS, unless the preceding exchange is taken into account, are placed under the Backward function. It should be noted that not even humans can properly classify Backward tags, if they are not aware of the context: annotators can affirm that an utterance is a repetition, only if they know that message is actually repeating something written before. The automatic classifier should therefore be equipped with the ability to spot a RS while considering the whole context of the conversation.

Given these considerations, we posed two research questions that guided the development of the classifier:

- **RQ1**: in order to automatically classify RS, we rely on a tag that characterizes the whole utterance. Is it feasible to identify a RS in a dataset using state-of-the-art techniques, if the tag does not mark specific parts of the user's message, but rather its entirety?

- **RQ2**: RS are not linguistically homogeneous and different tags entail different communicative intentions from the users. Is it possible to train a single classifier to distinguish among them, or will distinct classifiers perform better according to the specific tag or function?

In order to answer the research questions, we developed and tested several neural network-based classifiers that automatically recognize RS in the SisBot dataset. We also compared a single classifier to multiple ones in order to answer RQ2.

6.1 Classify Inherent and Backward Strategies

One of the first challenges in developing our classifier is to correctly manage the Inherent RS (i.e. breakdowns that are self-contained, thus benefit from a message oriented analysis) together with the Backward ones (i.e. those utterances that need to be contextualized within the conversation). For the latter, the classifier needs to look at previous messages from the user to properly assign the tag [1].

Since RS are not linguistically homogeneous and different tags represent different communication intentions, two different computational approaches are needed:

- for the Inherent classes, the approach should focus on the vocabulary and relations between words. The classifier has to predict a tag for each sentence in the conversation: therefore, it needs to learn what lexical elements are used in those strategies and how they relate with one another in the sentence;
- for the Backward classes, it needs to search for recurring patterns within a larger context, involving either the previous sentences from the user or the entire conversation. In this case, the classifier needs to have a broader scope. It should focus on the relation between sentences rather than the composition of a single utterance.

We thus defined three main classifier sets:

- **Inherent Repair Strategy VS Rest**: the classifiers in this set are in charge of detecting an Inherent kind of repair in a sentence. We will refer to this group as *INH VS Rest* from now on;
- **Backward Repair Strategy VS Rest**: these classifiers decide if a conversation (or a group of sentences) contains a Backward RS. We will call them *BCK VS Rest* from now on.
- **Absence of Misunderstandings**: we are not only interested to detect a sentence that has been tagged with an Inherent or Backward breakdown, but we must also be able to affirm the contrary (that is, when an utterance *is not* a RS). The general presence of a breakdown is a high-level metric that could immediately tell if a CA

[1] We momentarily exclude the FWU tag from our developments as it is the only one semantically linked to the CA's responses, instead of the user's messages. In that case we would have to take into account all CA's messages too, in addition with the user's ones.

is performing well or not. Even before knowing what kind of RS users are employing, an abundance of repair attempts signals persistent problems in the CA's behavior; on the other hand, an absence of RS entails the concept of a succesful conversation. We will refer to this set of classifiers as *POS VS Rest*.

The aforementioned classifiers are, in some ways, binary: they predict whether or not there is an Inherent RS or Backward RS (or their absence) but not the individual 3rd level classes (i.e. the specific tags). Given that no other work had already been done on this dataset, our first goal was to determine whether it was possible to detect and discriminate the different types of RS automatically; we leave the detection of the single RS to future work.

We also created a single classifier (under the name of *Single Model*) that merges the three previous sets at the cost of sacrificing performances, due to the heterogeneity of the repair strategies. The decision of joining different models aims at answering RQ2. In fact, in RQ2 we speculate whether a single architecture could be feasible, or it would be better to maintain different networks according to the specific function (Inherent or Backward), or behavior (Utilitarian, Oppositional, Relational) that needs to be spotted.

Each classifier can be conceptualized as a neural network model that leverages word-embedding (Mikolov et al., 2013; Pennington et al., 2014) defined via a pre-trained Italian BERT model[2] (Devlin et al., 2018), which has 768 units. It is worth remembering that the original SisBot dataset is in Italian and the choice of the pre-trained model was influenced by this factor.

6.1.1 INH VS Rest

In the INH vs Rest task, the core of the classifier transforms the input sentences (i.e. the users' utterances) into a unique vector that semantically represents the entire message; the resulting vector is then passed to an MLP (2 layers with ReLU activation function) in order to obtain the tag distribution. For the core, we experimented with both BERT's CLS tag vector, which represents the input sentence, and the simple word embedding of the sentences; over this latter layer, we tested a Convolutional Neural Network (CNN) following the model proposed by Kim et al. (Kim, 2014), a Bidirectional Long Short Term Memory (BiLSTM) network (Hochreiter and Schmidhuber, 1997) and a multi-head attention model (ATT) (Vaswani et al., 2017). The BiLSTM reads the input sentences from left-to-right and from right-to-left, obtaining a more accurate representation compared to an unidirectional one.

[2] https://huggingface.co/dbmdz/bert-base-italian-cased

6.1.2 BCK VS Rest

Unlike the INH vs Rest classification, the BCK vs Rest classifier requires a memory of the previously encountered messages, since the breakdown may become manifest only from the interaction between the current sentence and a previous one. Therefore, we used a hierarchical neural network to solve this task, i.e. a neural network composed of two layers: the first layer transforms the input sentence into a single vector representation, while the second layer relates the current sentence with its previous context. For simplicity, we will call the first layer *sentence-level* and the second layer *conversation-level* from now on. We employed the same models of the INH vs Rest classification for the sentence-level; the conversation-level, instead, is based on a LSTM network. In this latter case, we did not use a bidirectional one since future messages are not relevant to the identification and classification of breakdowns.

6.1.3 POS VS Rest

The task of POS vs Rest classification is similar to the BCK vs Rest one, because we have to analyse both the current sentence as well as the previous ones to determine if there is an absence of misunderstanding or not. For this task, we tested the BERT's CLS tag, the BiLSTM model and a multi-head attention model for the sentence-level.

6.1.4 Single Model

The single model is an end-to-end model that identifies breakdowns in conversations, no matter their function. Our goal is to check whether a single classifier can discriminate the 3 classes (BCK, INH and POS) in a single lecture of the conversation.

The model is a hierarchical neural network composed of a sentence-level layer and a conversation layer. For the sentence-level layer, we tested the same neural network models of BCK vs Rest. The main difference is that it has been tuned to predict either the tags that pertain to POS, BCK or INH.

6.2 Implementation of the Classifiers

We trained the classifiers using the weighted log negative loss given the imbalance of the dataset:

$$loss(x, y) = -w_y \ln P(y|x; \theta) \tag{6.1}$$

where the input x and y represent, respectively, the network results and the true class prediction, while w_y is the weight associated with y class; the probability of obtaining the correct label given the output of the network is represented by θ (the parameters of the model). The training process proceeds by minimizing the weighted average of the loss, both on the number of messages of the conversations and on all the minibatches. Weights are calculated on the basis of label frequencies. We used Adam (Kingma and Ba, 2014) as optimizer.

In order to regularize the network and to increase the generalization strength, we applied dropout layers (Srivastava et al., 2014) and weight decay (Loshchilov and Hutter, 2019). The probability of dropout is fixed at 0.2 (keeping a probability of 0.8), while weight decay can assume values in the range [0, 1e-5] according to the model (more details in Section 6.3). For multi-head attention models, we set the number of heads to 2 and we used positional encoding. Finally, we set the gradient clipping to 3 in order to avoid the vanishing/exploding gradient issue.

6.3 Evaluation of the classifier

For the evaluation, we split the dataset into training, validation and test sets. We used the validation set for the early stopping and hyper-parameters tuning. Table 6.1 reports the class distribution in the three sets.

Name	POS	INH	BCK
Entire dataset	53.83	8.93	37.24
Training set	54.21	8.68	37.14
Validation set	53.87	7.75	38.38
Test set	52.01	11.38	36.6

Table 6.1: The table reports the tag distributions (percentage) for the training, validation and test sets.

We trained the models on a Nvidia RTX2080Ti until the validation loss did not improved substantially for 5 consecutive epochs.

6.3.1 Results

In this section, we present the best results obtained during the experiments. In particular, Table 6.2 shows the performances of the different architectures developed for the one-vs-rest classifier, while Figure 6.1 and 6.2 contain a schematic diagram of the model architectures that performed best, showing Transformer layers and how they feed into the (bi)LSTM, CNN MLP layers on top.

By observing the best models, it is possible to notice that different tasks are best solved by very different architectures. For instance, the CNN architecture is the best in detecting *Insults*, *Closing* tags or other strategies that depends heavily on vocabulary information (Inherent class); on the other hand, the convolution applied to the classification does not allow the network to converge in the two other classes. Vice versa, the

	Model	F1 Score	Precision	Recall
BCK	CLS + LSTM	52.66	53.52	51.82
	BiLSTM + LSTM	**65.47**	**77.64**	**56.59**
	ATT + LSTM	62.59	75.88	53.26
INH	CNN + MLP	**58.99**	**68.91**	**51.57**
	CLS + MLP	57.14	48.74	69.05
	BiLSTM + MLP	24.67	15.97	54.28
POS	CLS + LSTM	56.51	54.54	58.63
	BiLSTM + LSTM	**71.27**	**78.92**	**64.97**
	ATT + LSTM	63.35	68.08	64.70

Table 6.2: The table reports the Precision, Recall and F-measure of the proposed classifiers. Best results in bold.

recurring model, which presents better performances for the detection of BCK and POS, does not seem to be suitable to the tagging of INH strategies. These results confirm the intrinsic differences between the Inherent and Backward classes from a computational point of view: not only they entail different characteristics and their occurrence proves it, but they are also best served by different computational approaches.

We also experimented with a CNN + LSTM model for both the BCK and POS tasks, but it did not converge; we believe that the CNN module is not able to create a vector representation of the sentence. For the INH task, we tested the ATT + MLP; similarly to the the CNN + LSTM model, it did not converge. Finally, we built a multi-class classifier by combining the best models together. We followed the one-vs-rest approach and its results are shown in Table 6.3. These results provide an answer to RQ1: we proved that it is indeed possible to automatically identify RS in a dataset. Our experiments checked several state-of-the-art neural network and demonstrated that repair strategies can be detected in a conversation either by different classifiers (according to their intrinsic, semantic nature) or by a multi-class one (called *One vs Rest*).

Model		POS	INH	BCK	AVG
One vs Rest	F1	69.58	56.52	67.70	**64.62**
	P	77.21	58.03	60.43	**65.23**
	R	63.32	55.08	77.13	**65.18**

Table 6.3: The table reports the results of the multiclass classifier. *P* stands for Precision while *R* is Recall. Best results in bold.

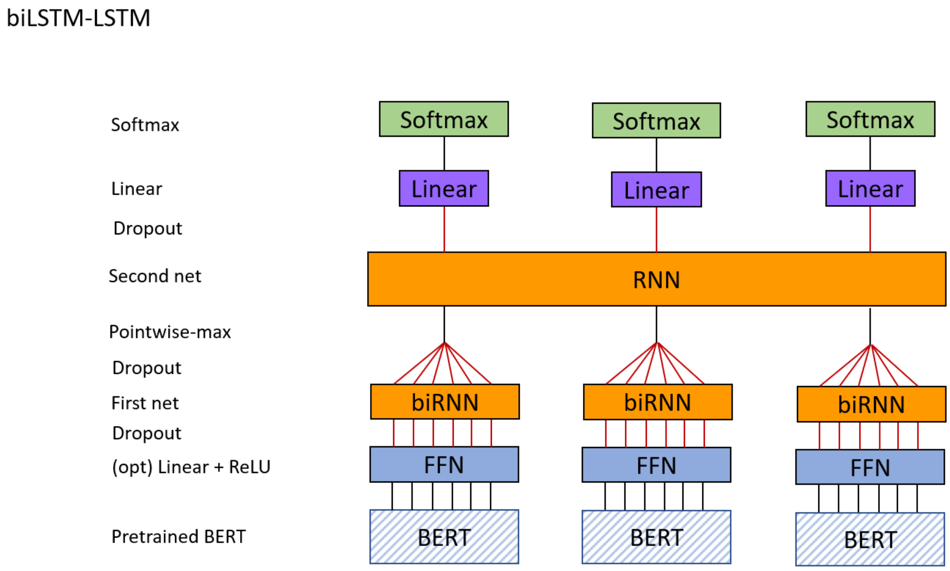

Figure 6.1: The diagram shows the different layers of the biLSTM-LSTM network employed here.

6.3.2 Single Model

We compared the aforementioned results with those of the *Single Model*, developed with the aim of distinguishing the three classes in a single try. Table 6.4 shows a decrease in the performance. This phenomenon is largely due to the different accuracy on the INH class; indeed, the scores of the other two classes do not differ much from the corresponding values in Table 6.3.

This last experiment demonstrates how difficult is to recognize INH, BCK and POS strategies with a single model, confirming - once again - the different nature of tags also from a computational point of view. We thus answer the question posed in RQ2, by showing that while it is technically possible to create a single neural network model to recognize all the three classes at once, that will come at a cost in terms of performances.

To further facilitate the comparison between the two proposed models, we report in Table 6.5 the results obtained by the multiclass classifier (One vs Rest) and the *Single Model* classifier on the binary task. Although both classifiers are very close on the F1 score, their Precision and Recall diverge. In particular, the One vs rest approach has a very high Precision (79.07) in identifying the presence of a breakdown in the conversation, reducing the false positives (only 108 cases); on the other hand, the *Single Model* classifier has a lower Precision (about 15 points lower) that leads to 215 false positives, but it is able to recognize a large variety of misunderstandings compared to the One vs Rest, thanks to the high Recall (80.27).

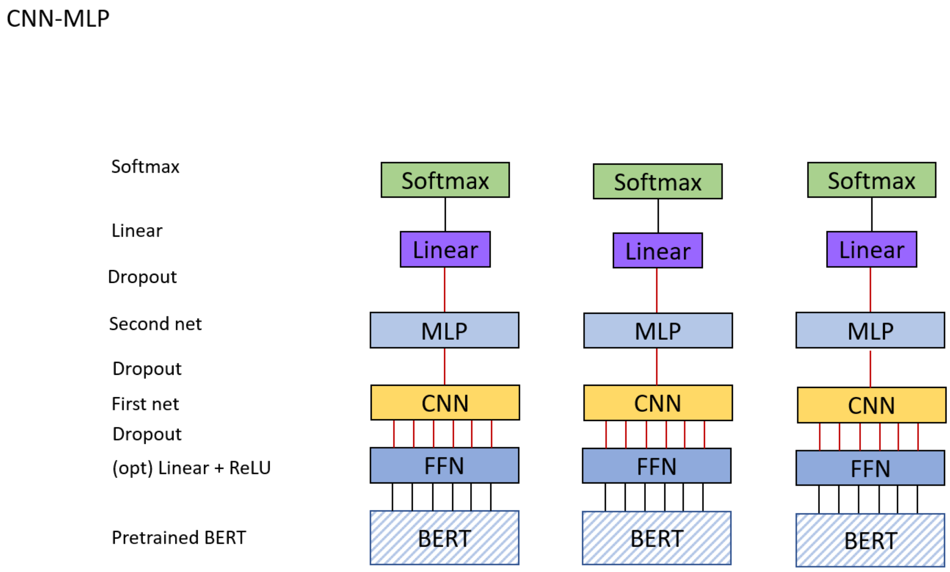

Figure 6.2: The diagram shows the different layers of the CNN-MLP network employed here.

Model		POS	INH	BCK	AVG
BiLSTM + LSTM	F1	69.79	36.36	64.95	**57.03**
	P	63.04	36.65	72.86	**58.19**
	R	78.16	34.33	58.59	**57.02**
CLS + LSTM	F1	59.04	54.69	46.11	53.28
	P	59.24	56.30	45.48	53.67
	R	58.83	53.17	46.77	52.93
ATT + ATT	F1	71.96	23.45	69.41	54.94
	P	56.30	14.28	99.50	56.69
	R	99.69	65.38	53.30	72.79
ATT + LSTM	F1	66.55	13.51	60.94	47.00
	P	63.90	8.40	71.36	47.89
	R	69.42	34.48	53.18	52.36

Table 6.4: The table reports the results of the tested architectures for the Single Model. Best results in bold.

We believe that the One vs Rest classifier is better in the task of recognizing the breakdowns expressed by the users during the conversation, but it would not be great to tag them with the appropriate RS ; the Single Model, on the other hand, could be used to assign the appropriate tags given its high Recall.

Model	F1	Precision	Recall
One vs Rest	72.83	79.07	65.81
Single Model	72.43	65.98	80.27

Table 6.5: Comparison of the One-vs-Rest model and the Single Model on the task of breakdown detection (that is, to identify the presence of a misunderstanding, without assigning any tag to it yet).

We also present the hyperparameters of all the best models previously outlined (Table 6.6).

Model	Parameters
BCK vs Rest	1° layer 192 (x2) 2°layer 150
INH vs Rest	1° layer 150 2°layer 512 3° layer 300
POS vs Rest	1° layer 192 (x2) 2°layer 300 Weight decay 1e-5
Single model	1° layer 384 (x2) 2°layer 192

Table 6.6: Hyperparameters for the chosen models. Where not otherwise specified, the following default values are applied: learning rate 0.001, dropout 0.2, weight decay 0.

Chapter 7

Users' Mental Model

Once a CA has been properly trained to spot and label repair attempts made in place by users, it has to react accordingly. The RS employed can inform a personalized answer: intuitively, a user who constantly resorts to Oppositional strategies requires a different response compared to a Relational user. However, this differentiation may not be enough to provide a meaningful answer to the user. For instance, let's consider the case where the person employs an Insult strategy: while this RS suggests an aggressive and non-cooperative behavior, the reasons behind that behavior may be varied. Are Oppositional users applying an ironic or rude strategy towards what they believe to be a human intelligence, thus expecting a reaction of some sort? Or, on the other hand, are they aware they are insulting a machine, and in that case this behavior can be regarded as a purposeless outburst of frustration?

These questions relate to the set of expectations, beliefs and knowledge each user possess. We refer to this ensemble as the user's *mental model* (MM). This Chapter explores how the concept of MM has been described in literature and its difference with other ways to describe the depiction of users' expectations and internal states. Through a an extensive research, we identify the current state of the art at the intersection of MM and Dialogue Systems.

7.1 Definition of Mental Model

The notion of mental model was first produced by the psychologist Kenneth Craik Craik (1952), who defined it as a small-scale "copy" of the external world that each person produces in oneself. The goal of this operation is to make sense of the word in order to act upon it. A MM can be seen as a reasoning mechanism in people's minds (Johnson-Laird, 1983) that is fed by the external environment, or as a set of beliefs and judgements that guide the person in taking everyday decisions (Jones et al., 2011). Users get information from their surroundings: from other people that have their same level of expertise on subjects, as well as from experts and professionals; moreover, the media industry plays a great role in influencing people's understanding of a topic, as their own experiences do (Wash and Rader, 2011). In the words of Staggers and Norcio (1993), one's MM is formed through knowledge (education), experience or a combination of the two.

7.1.1 Mental Model, Conceptual Model and User Model.

From this perspective, a mental model is different from a *conceptual model*, which is an expert or designer's understanding of the system (Greca and Moreira, 2000). Conceptual models are developed by authorities in the field, with a clear goal: e.g. a teaching course is developed by experts with the purpose of transferring knowledge efficiently. Mental models, in contrast, are developed quickly and often unconsciously by people who are no expert, but want or need to use a tool (and therefore need a mental representation of it). Mental models can be incomplete, evolving or simply "wrong", in the sense that they do not match the reality of things (Gentner and Stevens, 2014; Borgman, 1986).

The concept of MM should not be confused with its counterpart, the *user model*. In the context of Human-Computer Interaction, mental models are the ideas users project about a machine's internal functioning (Allen, 1997), while a user model is the representation the system has of a prototypical user (Brusilovsky and Millán, 2007). The two concepts are symmetrical since both reflect each one's expectations, although from two opposite perspectives. The fundamental distinction is that a mental model resides in the mind of people, while a user model is endowed in a computer in order to change its behavior upon predefined generalizations.

Since the MM is subject to manipulation by external actors as well as personal experiences, it may differ from the institutionalized and logical conceptions held by experts (Eslami et al., 2016). In this sense, "Mental models tend to be functional rather than complete or accurate representation of reality" (Jones et al., 2011). This last consideration is particularly important in the context of HCI. According to Norman (1983), a MM is formed by the people about themselves and the environment surrounding them. It can thus evolve during time in order to keep its function, but it does not need to be technically accurate (Rook, 2013).

7.2 Mental Models in Dialogue Systems

Since the MM of a user can also be counter intuitive with respect to the actual functioning of a system, it becomes extremely important to understand it and to take it into account, in order to design interfaces that are aware of such expectations and beliefs (Norman, 1983). This consideration may be especially true when the interface is a CA: inconsistencies between the user's MM and the internal functioning of the agent can lead to a variety of issues, from misunderstanding and confusion to the abandonment of the system altogether (Gero et al., 2020).

When a person interacts with a CA, the dialogue system should be aware of the user's "backlog", in a sense. That is, the set of expectations and beliefs that leads the user to behave in a certain way. For instance, if a CA were aware of the users' MM, it could adapt to their understanding, in order to make its answers more helpful and apt (Gregor and Benbasat, 1999; Grimes et al., 2021; Radziwill and Benton, 2017). Moreover, it could prevent early system abandonment due to unrealistic expectations caused by an inappropriate understanding of what CAs are capable of (Chiang et al., 2020). A correct detection of users' MM can lead to an improved interaction design (Lee and Malcein, 2020; Alloatti et al., 2021a).

7.2.1 Detection of Mental Models.

Mental models are not directly observable, since they are a construct of the mind. Therefore, their detection by machines during a conversation presents an interesting challenge. In the past, qualitative user studies have been performed to this aim (Luger and Sellen, 2016; Cho, 2018; Candello et al., 2017), while our goal is to present a quantitative approach. Qualitative studies are usually performed in a controlled setting and require an explicit interrogation of the users; on the other hand, a quantitative method aims at understanding the users by analysing their behavior during the conversation itself with the CA, in a real life setting. Once again, our primary goal is to present a framework, a technique that can then be used to automatically infer the user's MM from the dialogue with a CA. We start from the findings already described in previous work, in order to look for existing frameworks that can help us towards our objective.

We thus conducted an extensive research by gathering over 250 articles from three databases. Our analysis highlighted the presence of several works that focus on CAs and MMs; however, each one of them tackles the issue from a different perspective and few of them actually try to detect the users' mental models from the dialogue with a computer. It is thus hard to extrapolate a unique and homogeneous framework.

Our contribution is twofold: first, we describe the results obtained from the research of existing work and provide an overview of the studies in the field. Secondly, since our final goal is to produce a framework that is able to describe the users' MM, we propose a new conceptual scheme informed by the results of the analysis of previous work. Such scheme will analyse users' behaviour while interacting with a CA and describe their MM based on lexical elements.

7.3 Previous Work

In order to gain a comprehensive overview of the existing studies on the topic of users' MM when interacting with a CA, we ran a specific query on different databases. This method allowed us to collect a significant number of papers.

First of all, we selected some appropriate keywords that would represent our subject of interest, while still not excluding potentially interesting studies. We thus build the query string by mentioning on one end the term *conversational agent* together with its synonym *chatbot*; then, we add the term *mental model*. In a first attempt we also added some terms such as *evaluate/evaluation*, *detect/detection*, *measure*, etc. with the intent of capturing the articles that dealt with the detection or assessment of users' mental model. However, this query proved to be too strict and did not yield any result on some of the databases. For this reason, the final query was structured as follows:

("conversational agent" OR "chatbot") AND "mental model"

The query was launched on IEEE Xplore, Scopus and Web of Science. The query yielded 258 results in total. We conducted a first round of analysis by reading the abstract and we had to exclude most of the results because they only tackled the subject of CAs, while completely ignoring the theme of the mental model. Specifically, several articles detailed the design process of dialogue systems, or they studied the connection between CAs and users' emotions. Other articles did discuss the concept of MM, but they did not connect them to the use of a CA: the MM is often mentioned in the domain of education (e.g. students' MM while approaching a new subject at school) or in team buildings activities (where the aim of the study is to build a shared MM between the participants). Since our focus is on the detection of users' MM while they interact with a CA, we discarded these articles.

After the first round, the remaining number of papers amounted to 58 units. A second, deeper round of analysis identified similar problem in other 10 articles (e.g. papers that mentioned the subject of MM but did not intersect it with the use of CAs), while two articles were inaccessible because of a paywall and one was in a language other than English. The final batch is formed by 45 papers.

7.3.1 Analysis of Previous Work

The two rounds of pre-analysis allowed us to build for ourselves a preliminary picture of the articles' content. It was clear that even though the themes of MM and CAs were mentioned in the text, almost each study had a different perspective on the subject. All articles implicitly or explicitly stated that users have a mental representation of the machine counterpart in the dialogue and that MM can change in time, according to the experience of the interaction. However, some studies did not dwell on the formalization

of such MM, focusing rather on the modification the MM can undergo if the CA is presented differently. Fewer articles proposed a definition of mental model, usually by describing a continuum between two opposite poles of expectations and beliefs.
We were able to cluster the studies in three main groups:

No MM, Yes CA. The first one contains articles that do not provide a clear formalization of what they interpret the mental model to be. Instead, they manipulate one or more features of an agent in order to see whether that CA will be perceived differently by the users. In other words, they implicitly acknowledge the existence of an initial MM (which we could call the MM at T_0) but they do not define it nor they measure it. They only observe the effect of a manipulated CA on the MM, once users have interacted with such agent.

Yes MM, Yes CA. The second group deals with both subjects. While it still manipulates the CA in order to detect a change in perception from the users, it also provides a definition of MM. Such definition often emerges from the results of the experiments: a user with a certain MM will interact with a manipulated CA and get to certain conclusions, while a user with a different MM will present another behaviour.

Yes MM, No CA. Finally, the third group investigates the users' MM at T_0 without having them to interact with a modified CA. This group is the most interesting for our purposes, since our goal is to measure the MM during the interaction with a CA that did not undergo any manipulation (meaning, all users interact with the same CA).

The next section will detail the articles for each group. Moreover, we organize the studies along two other dimensions: the channel of communication used by the CA in the article, and whether the experiments run in each study were qualitative or quantitative. We deemed these two characteristics to be of particular relevance. The first one will be able to say whether most studies deal with text-based communication, voice-based one or an embodied communication (e.g. between a human and a robot). Since our focus has always been on text-based CAs, it will be interesting to notice whether the majority of the studies on MM also deal with this kind of agents or not. The second one aims at identifying which studies apply a quantitative methodology to detect the users' MM. As it was mentioned in the beginning of this chapter, our goal is to present a framework, a technique that can automatically infer the user's MM from the dialogue with a CA. This automatism must be based on "quantitative" data rather than qualitative, since it cannot rely on human analysis of the language.

Manipulated feature of the CA	Articles
Personality	(Ruane et al., 2020), (Lopatovska et al., 2021), (Whittaker et al., 2021), (Ma et al., 2019b), (Hanna and Richards, 2015a), (Hanna and Richards, 2015b), (Roa-Seïler et al., 2014)
Humanness (anthropomorphism, human-like cues, voice and humour)	(Han, 2021), (Diederich et al., 2020), (Ischen et al., 2020a), (Ischen et al., 2020b), (Doyle et al., 2019), (Moussawi and Benbunan-Fich, 2020)
Embodiment (physical structure or graphical avatar)	(Ostrowski et al., 2021), (Stein et al., 2020), (Ma et al., 2019a), (Roa-Seïler et al., 2014), (Ciechanowski et al., 2018)
Pleasantness, friendliness, warmth	(Melo et al., 2020), (Abe et al., 2012), (Grimes et al., 2021), (Khadpe et al., 2020)
Conversational capabilities, confidence and interactivity	(Cho and Rader, 2020), (Chiang et al., 2020), (Bellur and Sundar, 2017)
Explainability and transparency	(Robb et al., 2019)
Perceived mind, morality and trust	(Banks, 2020)
Gender and gendered domains	(McDonnell and Baxter, 2019)
Perception of its danger	(Abdi et al., 2019)

Table 7.1: First group of articles. Each work manipulated one or more features of a CA.

7.3.2 Manipulation of the CA

The majority of the studies we analyse pertains to the first or second group; that is, they manipulate one or more feature of a CA and note the impact of such changes in users who interacted with the agent.

The first group. Amidst the 45 analysed articles, 28 are part of the first group. They focus on the modification of a specific characteristic of a conversational agent. Their goals may be various: to check whether a change in the CA produces a symmetrical change in expectations (Grimes et al., 2021); whether a change in the CA's personality is perceived correctly (Ruane et al., 2020; Whittaker et al., 2021) and is appreciated (Lopatovska et al., 2021); the effect of an anthropomorphised agent on users (Han, 2021; Diederich et al., 2020), and so on. Table 7.1 reports the detail of the CA's feature that each study manipulated. A single study may have modified multiple features.

It is possible to observe that the most manipulated features are the Personality of a CA, its level of Humanness or the fact that it is Embodied. For the first case, the analysed studies proved that users are indeed able to notice whether a CA has a specific personality (Ruane et al., 2020; Hanna and Richards, 2015b) and that a CA endowed with a specific and consistent personality makes a positive impression on the users (Lopatovska et al., 2021; Ma et al., 2019b), especially if such personality is warm and agreeable (Hanna and Richards, 2015a,b). Another work also noticed a preference for the CA that matches the user's own personality (Whittaker et al., 2021).

With regard to the Humanness feature, several studies discovered that an anthropomorphic agent (e.g. equipped with a proper name, the ability to recognize the user, etc.) is well accepted by its users and it can also be more convincing compared to a non-anthropomorphic one (Han, 2021; Diederich et al., 2020; Ischen et al., 2020a; Moussawi and Benbunan-Fich, 2020). Symmetrically, CAs that are not endowed with these capabilities are perceived to be more formal, impersonal and less authentic (Doyle et al., 2019). Although a certain level of humanness seems to produce positive effects, the embodiment of the agent into human-like figures (such as anthropomorphic robots or graphical avatars) is still perceived as uncanny. Some studies found a correlation between the agents' embodiment and a sensation of eeriness (Ostrowski et al., 2021; Stein et al., 2020; Ciechanowski et al., 2018), while Ma et al. (2019a) found that an embodiment that gives away the true function of the agent is nonetheless preferred to an anonymous physical shape (such as those of smart speakers). If the users are children, such perception fades, as long as the agent is presented in an aesthetic and entertaining way (Roa-Seïler et al., 2014).

To sum up, the manipulation of certain CAs' features demonstrates an improvement in the way users perceive those agents. Specifically, a CA with an agreeable personality and anthropomorphic qualities shows a boost in usability and pleasantness. Appreciated human-like qualities stop at the linguistic level though, since other human features (such as a resemblance of a body) are still perceived as uncanny. Unfortunately, these studies do not measure the impact of the manipulation on the users' MM. It would have been interesting to observe whether the increase in positive perception is correlated to a modification of the MM: for instance, users who interact with a manipulated CA gain a clearer understanding of the potentiality of the system, and therefore they appreciate it more. Some partial answers may be found in the analysis of the second group of works.

The second group. This ensemble of studies (n=10) tackles both the dimensions of our query: they manipulate the CA but they also provide a formalization of the users' mental model. For instance, Avdic and Vermeulen (2020) measure the intelligibility of a CA and they notice how the manipulated feature of the agent is perceived differently according to the user's experience and background. Similarly, other studies entangled the two perspective by analysing features such as smartness, visibility or embodiment (Lee et al., 2020; Seymour et al., 2020; Blut et al., 2021). Table 7.2 presents a complete list of the characteristics of CAs that this second group of studies manipulated.

Once again, the Humanness feature is one of the most frequently modified in con-

Manipulated feature of the CA	**Articles**
Humanness (anthropomorphism, human-like cues)	(Blut et al., 2021), (Moussawi et al., 2020), (Mishra and Shukla, 2020), (Knijnenburg and Willemsen, 2016), (Araujo, 2018)
Physical presence (embodiment, visibility)	(Lee et al., 2020), (Blut et al., 2021)
Intelligibility	(Avdic and Vermeulen, 2020)
Smartness	(Seymour et al., 2020)
Behaviour, knowledge distribution	(Gero et al., 2020)
Task-relevant information	(Weis and Wiese, 2020)

Table 7.2: Second group of articles and the features of a CA each study manipulated.

versational agents. A CA endowed with human-like characteristics elicits a positive response (Araujo, 2018; Mishra and Shukla, 2020), although it can also turn the user away from correctly perceiving its real capabilities (Knijnenburg and Willemsen, 2016). Moreover, Blut et al. (2021) affirm that the perception of anthropomorphism is not just related to the way the CA is presented, but also on some prior characteristics of the user. Thus on one hand, several studies affirm that equipping a dialogue system with anthropomorphic feature can have a positive impact; on the other hand, it is not clear whether presenting the system as more human might shift the mental representation of the user in the wrong direction. The MM component of the articles of the second group may shed some light on this question.

7.3.3 Conceptualizations of MM

Since the second group also presents some conceptualization of what a user's MM may be, we provide a synthesis of each article's proposal in Table 7.3. Such conceptualization can be expressed by the presence or lack of certain characteristics in the user; in other cases, the user's MM is defined as a point on a continuum between two extremes. The study by Lee et al. (2020) is excluded from the table: while it does mention the fact that a definition of the users' mental model can be extracted from their internal representation of the agent, it does not formalize the definition of such mental representation. The authors state that the MM changes according to the visibility (or physical presence) of the agent versus its invisibility, but it is not clear whether this intuition can provide a broader categorization of different MMs.

The majority of works place the core of the users' MM in their prior knowledge or experience about a subject. In the case of an interaction with a CA, preexisting technical competence plays an important role: people who are familiar with the internal functioning of an agent may have less difficulties in understanding its behavior, while naive

The user's MM can be:	Articles
Characterized by the presence or absence of: experience and background; technical expertise and prior familiarity; technical competence and prior experience;	(Avdic and Vermeulen, 2020), (Blut et al., 2021), (Weis and Wiese, 2020)
Hedonic or Utilitarian	(Moussawi et al., 2020), (Mishra and Shukla, 2020)
Poor or Good; Accurate or Inaccurate	(Seymour et al., 2020), (Gero et al., 2020)
More or less anthropomorphic	(Knijnenburg and Willemsen, 2016), (Araujo, 2018)

Table 7.3: Definitions of mental models found in the second group of articles.

users may struggle more. Expert users will know what to expect from the CA in terms of capabilities and will also be able to manage their own expectations. Interestingly, Blut et al. (2021) found that a predisposition called "computer anxiety" influences the perception of anthropomorphism, possibly by inducing the person who is more anxious to attribute (false) human abilities to the dialogue system, even when such abilities are not suggested by the CA in any way. Even though this specific article does not detail it, we could assume that this anxiety is caused by the lack of competence: if a person does not know how a CA operates nor is familiar with computer science principles in general, she may feel uncertain and expect unreasonable things from the machine.

In accordance with these thoughts, Knijnenburg and Willemsen (2016) state that a MM that is too much anthropomorphic will result in lower learnability and effectiveness for the user. To sum up, according to the second group of articles, MMs can either be "Good" (accurate, not too anthropomorphic, supported by some technical competence or prior experience in the field) or "Poor" (inaccurate, too anthropomorphic, lacking knowledge and familiarity with the specific technology).

Finally, a few articles define the MM according to two attitudes that a user can embody: Hedonistic or Utilitarian. Even though in the aforementioned works these attitudes were defined as part of the MM (Moussawi et al., 2020; Mishra and Shukla, 2020), we believe them to be more similar to the behavioral models defined in Chapter 4. A user may have an Utilitarian attitude regardless of an accurate or inaccurate MM, which by definition relies on input from the outside world and is not an intrinsic orientation of the person (Johnson-Laird, 1983; Norman, 1983).

The third group. This last group is made up of 7 studies. It is composed by works that investigate the MM prior to any manipulation; in fact, they do not modify their conversational agent at all, as their only goal is to observe the MM at the stage T_0. Each

The user's MM can be:	**Articles**
Characterized by the presence or absence of: technical competence and prior experience/familiarity;	(Luger and Sellen, 2016), (Chen and Wang, 2018), (Myers et al., 2019a)
Pull or push	(Cho, 2018)
Explorers, proficients or strugglers	(Myers et al., 2019b)
Technical or metaphorical (according to the presence of technical knowledge)	(Ngo et al., 2020)
More or less adherent to cultural models	(Rose and Björling, 2017)

Table 7.4: Definitions of mental model found in the third group of articles.

study offers its own description of a user's mental model; nonetheless, the general tendency is once again to define it as a continuum between two extremes, or as a clustering of different perceptions, in a similar way to the studies of the second group (see Table 7.4 for a granular analysis of each article).

The presence or absence of technical knowledge and prior experience is still identified as one of the main features of a correct mental model. Users who are expert in computer science have a more developed representation of the system capability. It is therefore less likely that they will abandon the conversation (Luger and Sellen, 2016) and in case of error, they can make reasonable hypothesis about what happened (Chen and Wang, 2018). This profile is coherent with the one identified by Myers et al. (2019b), where users are divided in Proficients and Strugglers. The authors also propose a mid-level category, the Explorers: these people encounter the same difficulties as the Strugglers in the beginning (given their lack of expertise), but they are quicker in developing a correct mental representation of the system they are interacting with.

The presence of technical competence is also relevant to the work by Ngo et al. (2020), while Rose and Björling (2017) consider to be particular important the role of cultural references in the user's mind. According to them, the user's MM of CAs is influenced by popular culture, since they picture them as they see them in movies. For the same reason, they expect the agents to be helpers, to be useful. These findings are coherent with the fact that MMs are formed by multiple external factors, including knowledge passed on by media and popular culture.

Finally, the work by Cho (2018) envisions yet another continuum between two extremes: users who belong to the Pull model, and those who employ a Push model. In the former model, participants attempt to extract information from a smart speaker by using relatively simple forms of questions. In fact, they try to *pull* the information out of the system by asking questions. In the latter one, users clarify what they want to find by offering a certain piece of information to the agent. They *push* the CA towards a certain direction by offering information rather than asking for it.

7.3.4 Channel and Methodology

Before discussing the implications of the analysis - especially regarding the conceptualizations of MM - we offer two more dimensions along which the articles can be distributed: the *channel of communication* and the *evaluation methodology*. The first one aims at highlighting whether the majority of studies deals with text-based, voice-based or embodied agents (e.g. robots or avatars) and if the chosen channel impacts on the manipulated feature of the CA or the definition of MM.

The second dimension divides the works in those that employed a qualitative method of evaluation (e.g. questionnaires, interviews) from those that used a quantitative one. Since our final goal is to develop a quantitative method to measure the user's MM by analysing the conversation itself, it would be interesting to observe whether other works attempted the same.

Channels of communication. The channel is the medium through which the CA exchanges with the other speaker. In the domain of dialogue systems, the communication can usually happen along two different channels: the textual one, or the vocal one. While voice-based agents can also be equipped with textual capabilities, in recent years several systems have been distributed as voice only experiences, such as the agents on smart speakers. A third possibility comprises those CAs that are equipped with a graphic embodiment (like an avatar), or a physical one, such as a more or less anthropomorphic robot.

Table 7.5 reports the studied grouped by channel. We were able to assign each study to a channel with the exception of the article by Ngo et al. (2020). Their study tackles the subject of intelligent agents, however it does not specify whether its results are applicable to text-based agents, voice-based ones, or both.

As it can be noticed, the most numerous batch is represented by the voice-based agents. Even though proficiency in human-machine communication via voice has been gained only recently (in fact, this is the group that contains the highest number of articles published in 2021), its influence on users' MM and their perception of a CA has clearly captured a lot of attention. The other two categories are equally represented. Some articles in the *Embodied agents* category are older with respect to the average of the rest, but we do not deem the difference to be particularly significant.

Channel x Groups. We crossed the distribution of the articles per channel with their distribution in the three clusters defined in Section 7.3.1. Table 7.6 shows the number of articles per group, according to the channel their CA belongs to.

The studies that deal with voice-based agents are the most balanced: even though the majority still focus only on manipulating the CA, a significant number of works also take into account the user's MM. This tendency is scarcer among the "embodied" articles and basically absent in the "text-based" ones. It is possible that the introduction of the vocal

Text-based agents	(Grimes et al., 2021), (Han, 2021), (Ruane et al., 2020), (Gero et al., 2020), (Khadpe et al., 2020), (Robb et al., 2019), (Ischen et al., 2020a), (Ischen et al., 2020b), (McDonnell and Baxter, 2019), (Melo et al., 2020), (Knijnenburg and Willemsen, 2016), (Araujo, 2018), (Bellur and Sundar, 2017)
Voice-based agents	(Avdic and Vermeulen, 2020), (Lee et al., 2020), (Seymour et al., 2020), (Lopatovska et al., 2021), (Ostrowski et al., 2021), (Whittaker et al., 2021), (Cho and Rader, 2020), (Chiang et al., 2020), (Moussawi and Benbunan-Fich, 2020), (Myers et al., 2019b), (Moussawi et al., 2020), (Mishra and Shukla, 2020), (Myers et al., 2019a), (Doyle et al., 2019), (Abdi et al., 2019), (Cho, 2018), (Chen and Wang, 2018), (Luger and Sellen, 2016)
Embodied agents	(Blut et al., 2021), (Stein et al., 2020), (Diederich et al., 2020), (Weis and Wiese, 2020), (Banks, 2020), (Ma et al., 2019b), (Ma et al., 2019a), (Roa-Seïler et al., 2014), (Abe et al., 2012), (Ciechanowski et al., 2018), (Hanna and Richards, 2015a), (Hanna and Richards, 2015b), (Rose and Björling, 2017)

Table 7.5: Total amount of analysed articles, grouped by channel of communication.

	First group (No MM, Yes CA)	**Second group (Yes MM, Yes CA)**	**Third group (Yes MM, No CA)**
Text-based	10	3	0
Voice-based	8	5	5
Embodied	10	2	1

Table 7.6: Number of articles in each group, according to the channel of communication used by their CA.

element has sparked new questions about the perception of the agents, since people rely a lot on sound information when painting a mental picture of their interlocutor (McGinn and Torre, 2019; Tanaka et al., 2010). It is also possible that the presence of a voice but the lack of embodiment confuses the user and induces the construction of more incorrect MMs. A robot may declare in a clearer way its capabilities through its physical demeanor: a robot equipped with hands will be expected to know how to grasp objects, for instance. A voice agent increases its perception of anthropomorphism (Moussawi et al., 2020; Mishra and Shukla, 2020; Doyle et al., 2019) but at the same time does not give away its abilities in a straightforward manner, since it is not embodied in an anthropomorphic shape (Avdic and Vermeulen, 2020; Lee et al., 2020).

These hypothesis, although interesting, would need to be thoroughly checked. We believe that the data presented in Table 7.6 establishes a preliminary step towards future work in this direction.

Evaluation methodology. Through this dimension we are interested to notice how many studies employ a qualitative methodology, compared to those that rely on a quantitative one. Table 7.7 reports the reference of the works for each of these two categories. The work by Blut et al. (2021) was excluded from the table since it consists of a theoretical dissertation and it does not conduct any evaluation.

As it can be noticed, the majority of works employ a qualitative methodology (n = 27). They usually conduct post-experience surveys or interviews and manually extract insights from their content. Even though this method seems to be the most frequent to analyse the user's MM or to measure the impact of a manipulated CA, it is only applicable in small, experimental settings. It would not be feasible to interview all users who interacted with a CA such as our SisBot (see Chapter 4 for the usage data); moreover, our interest is in adapting the CA's answer to the person's MM. If we will have to be aware of it only after the conversation has took place, it would not be possible to adapt the CA on the fly.

With regard to the studies who employ quantitative studies (n = 12), they can be further divided in two subcategories:

Quantitative analysis of questionnaires. These works use quantitative methods (e.g. PLS-SEM, word count) applied to inherently qualitative data (such as interview answers). Studies like Khadpe et al. (2020), Mishra and Shukla (2020) and Knijnenburg and Willemsen (2016) are part of this subcategory. We also include in this subcategory those who employ Likert scales-based (or similar) evaluations: even though the questions of a survey contain an intrinsic element of subjectivity, the values extracted are numerical in nature (see for instance Grimes et al. (2021), Moussawi and Benbunan-Fich (2020), Ischen et al. (2020b), McDonnell and Baxter (2019)).

Qualitative methodology	(Ruane et al., 2020), (Avdic and Vermeulen, 2020), (Lee et al., 2020), (Ngo et al., 2020), (Seymour et al., 2020), (Gero et al., 2020), (Ostrowski et al., 2021), (Han, 2021), (Stein et al., 2020), (Diederich et al., 2020), (Cho and Rader, 2020), (Robb et al., 2019), (Chiang et al., 2020), (Ischen et al., 2020a), (Moussawi et al., 2020), (Doyle et al., 2019), (Abdi et al., 2019), (Roa-Seïler et al., 2014), (Melo et al., 2020), (Cho, 2018), (Chen and Wang, 2018), (Hanna and Richards, 2015a), (Hanna and Richards, 2015b), (Araujo, 2018), (Rose and Björling, 2017), (Bellur and Sundar, 2017), (Luger and Sellen, 2016)
Quantitative methodology	(Grimes et al., 2021), (Khadpe et al., 2020), (Moussawi and Benbunan-Fich, 2020), (Weis and Wiese, 2020), (Ischen et al., 2020b), (Mishra and Shukla, 2020), (Banks, 2020), (Ma et al., 2019b), (Myers et al., 2019a), (Ma et al., 2019a), (McDonnell and Baxter, 2019), (Knijnenburg and Willemsen, 2016)
Both methodologies	(Lopatovska et al., 2021), (Whittaker et al., 2021), (Myers et al., 2019b), (Abe et al., 2012), (Ciechanowski et al., 2018)

Table 7.7: Analysed articles divided by evaluation methodology.

Live quantitative analysis. We group here studies that try to detect the users' MM or, in general, their perception, during the interaction itself with the CA. Weis and Wiese (2020), for instance, measure the offloading preferences of users towards virtual agents or humans. Nonetheless, this data is still coupled with information about their preexisting MM, obtained via a survey prior to the interaction. Similarly, Myers et al. (2019a) mix usage data from the interaction with data from a pre-test questionnaire. This tendency is even stronger among the studies that apply both a quantitative and a qualitative evaluation methodology (n = 5): Abe et al. (2012) measure both the regularity of the gaze, smile intensity and the motion of children towards the robot via a tracking system in the robot's eye camera, as well as an estimation of the children's mental state via a questionnaire.

On the other hand, Ma et al. (2019a) adopt an interesting approach: they propose several tasks to the user, and then observe the users' preferences when they have to decide which agent will carry out the tasks. Their observations are limited to the span of the interaction and they do not rely on any questionnaire.

Few works adopt a quantitative methodology, and even fewer rely on lexical elements that can be found in the conversation (rather than elements such as the choice of a CA, or the user's gaze, etc.). Even though we did not find a ready-to-use framework to detect users' MM in literature, we extracted some useful insights from an in-depth analysis of existing work. We will leverage these findings to propose a new method to quantitatively detect users' MM from their interactions with a CA.

Chapter 8

Detection of Mental Model: a Proposal

This Chapter will tackle the issue of mental model identification by reflecting on specific features that describe the users' MM during their interaction with a CA. Such features are identified in specific lexical markers. The last part of the chapter shows a process to automatically detect these markers in our dataset, and thus infer some characteristics of the MM.

The analysis of previous work revealed that a user mental model can be conceptualized as a continuum along two extreme poles: on one end, a *good* or *accurate* MM, and on the other end, a *poor* or *inaccurate* MM. Although this sounds reasonable enough, it is not clear how to pragmatically assess one's MM in order to assign it to one side or the other of the spectrum.

What in fact constitutes a good MM? What instead makes it inaccurate?

According to the very same analysis, a good MM is characterized by the presence of technical competence and the "right" quantity of anthropomorphism. Symmetrically, a poor mental model lacks technical expertise and the perception of the agent is excessively anthropomorphic. The first factor can be easily quantified and previous work proposes several questionnaire-based methods to measure the users' technical competence: (Avdic and Vermeulen, 2020), for instance, group users around the two poles according to how frequently they interact with a CA. However, once again, we are interested in methods other than questionnaires. For instance, Ferrod et al. (2021) propose a model for automatically detecting the user's domain expertise from a conversation with a commercial CA. Their model is based on a BiLSTM-CRF model (Huang et al., 2015) which processes each message in the conversation, identifies the expertise words employed by the user and then labels them with a set of tags.

8.1 Expertise

This last solution is interesting and closer to our goals. Nonetheless, as the work by Ferrod et al. clearly shows, the task of expertise detection is strongly tied to the specific domain of the CA. In that case, users were interacting with a dialogue system that would offer assistance about their telephone line. Their scale of expertise was therefore created ad hoc according to technical terms that users may or may not use in that context. In the case of our SisBot conversations, we would have to create a specific scale of expertise based on the technical concepts that appear in the exchanges: knowledge about electronic invoices, fiscal procedures, how to operate the specific software, etc. We would have to assume that a user that employs a more refined term such as *fiscal drawer* has a greater technical competence than those who say *tax box*.

Even though those assumptions are legitimate and could be further validated by domain experts, our efforts are addressed at creating a method that could be also applied in other contexts. If we were to create an expertise scale for the SisBot domain, it would then be basically impossible to re-apply the same work in another context. Take for instance the tagset of repair strategies: it was built by taking into account strategies found in a specific domain, the conversations with the SisBot, but they were abstracted enough to make them applicable to a broader spectrum of situations. For this reason, we believe that the task of expertise detection should be handled differently by any work and our contribution would not be extremely significant.

On the other hand, the *detection of users' detected anthropomorphism level* poses a novel and broader challenge: that is, the level of anthropomorphism that the users apply when interacting with a CA.

8.2 Anthropomorphism

Anthropomorphism can be defined as an involuntary perceptual strategy by which human beings unconsciously guess or expect that a stimuli have a human-like or human cause (Guthrie, 1997). Unlike technical expertise, anthropomorphism is not strictly related to a specific domain, but rather to psychological factors of a person or group of people (Epley et al., 2007). For this reason, we can try to detect one's level of anthropomorphism through the dialog with a CA (or, to be more precise, one's level of anthropomorphism *ascribed to the system*), since we expect that users will show consistent behavior in terms of perceiving the agent as more or less anthropomorphic. Therefore, the features we will identify as being significant to this goal could be also applied to other human-machine dialogues.

8.2.1 Physical and non-verbal cues

Once again, we can search previous work to understand if the level of anthropomorphism can be deduced by the analysis of the sole lexical elements in the dialogues. According to Seeger et al. (2018) an anthropomorphic agent will be equipped with a sort of human identity, by means of: a graphical/ physical representation of the CA; verbal cues, including the choice of words and sentences; non-verbal cues, comprising the non-verbal communication behavior of the CA (e.g. facial expression, gaze, posture, etc.). Such features will then elicit an anthropomorphic behavior from the users: for instance, those who interacted with the robot Olly (Whittaker et al., 2021) spoke to the machine as they would when interacting to an animal or small child. They raised their voice while also speaking slower and used encouraging language, thus demonstrating elements of anthropomorphism. They moved towards Olly while talking and smiled while it spoke, keeping their gaze on the robot for most of the interaction.

This first example shows that users' anthropomorphic view of the CA can be demonstrated by analysing the people's behavior towards the agent. In the case of text-based CAs, users will not speak to them or be able to approach them physically, therefore our sole area of focus will be the verbal cues one.

8.2.2 Verbal cues

Some studies have specifically researched the topic of anthropomorphism via the use of specific verbal cues and structures. In the work by Ischen et al. (2020b) it is said that a human-like CA is equipped with a first name and that it acknowledges the other speakers' answers, i.e. that the CA's sentences are not always the same, but they may refer to what the user has previously typed in order to show anthropomorphic features. Symmetrically, Chiang et al. (2020) found that the users scolded, encouraged, or praised the anthropomorphic CA, because they expected it to be like an intelligent agent that would ask them about their preferences and then learn incrementally. Moussawi and Benbunan-Fich (2020) add that users who react to an anthropomorphic agent express positive emotions and trust in their lexicon.

While none of these studies offers a clear set of words, expressions or otherwise identifiable lexical features to determine whether that user's MM is "enough" or instead "too much" anthropomorphic, we can make some hypothesis inspired from their findings.

8.3 Lexical Markers of an Anthropomorphic MM

In the previous paragraphs we tried to extrapolate some features that could describe the user's level of anthropomorphism while interacting with a CA. This work could have been done in different ways: for instance, by using a supervised learning approach, or a topic-modelling (e.g. LDA, or BERT) approach to cluster users' behaviour to detect patterns of anthropomorphism-projection, and then learn to associate them with specific markers. However, we are primarily interested in operationalizing the concept of MM by measuring the quantity of anthropomorphic sentences in our dataset, in order to produce ad-hoc responses. For this reason we immediately identify a set of lexical features that can be easily spot in the dataset. Upon that additional tagging (e.g *does this message from the user present anthropomorphic features?*) we built the next Chapter, that deals with producing tailored answer to these features as well as the RS employed.

We propose a set of lexical features that denote an anthropomorphic mental model. Therefore, the presence of such markers paints the user as having a "Poor" (inaccurate, too anthropomorphic) MM, while their absence entails the exact contrary. An anthropomorphic MM is thus characterized by the following elements:

- **Use of deictics:** people who employ deictics presuppose more intelligence from the machine than what is reasonable to expect from it. While it is possible to employ local references in conversations with multi-modal or embodied agents (Pustejovsky and Krishnaswamy, 2020; Aneja et al., 2021), a reference to an external element (in the real world) is hard to comprehend in a written interaction. These references can be temporal in nature (e.g. use of words such as *today; yesterday; few hours ago*, etc), or they can refer to people using relative or ambiguous terms (e.g. *my colleague; my mother*, or sentences such as *I spoke with your representative*, etc.). An agent does not have an innate sense of time, nor place; moreover, it cannot correctly parse who *my colleague* could be, since it is not equipped with the ability to infer information from the outside world. In some cases agents may have been trained to distinguish the entity *today* from the entity *tomorrow*, but this operation needs to be carried out explicitly and the attribution of such ability to a machine is not obvious. Finally, in this context we also consider deictic elements references to contextual elements that the users see on their screen, but of which the CA is not aware of - and could not be, since in the case of SisBot, the CA can be operated from various pages of the invoicing platform with no difference between one page or the other. Examples of this kind of deictic references include sentences like *I noticed an error on the first line of my account*; *what's the problem with my invoice number 10*, etc. These utterances imply the CA's ability to see "behind" its own chat window into what the user is looking at in that moment. Once again, it is important to highlight the fact that some agents with these capabilities may exist out there; however, to assign this kind of human abilities to a CA spontaneously (i.e. when such abilities are not made evident anywhere) can be considered a sign of excessive attribution of anthropomorphic features to an agent.

- **Use of emotion-charged elements:** we refer here to lexical or typographic elements that express an excess of emotion, both positive and negative. We believe them to be indicative of an anthropomorphic MM because users who employ such elements expect a different reaction from the machine with respect to writing in a regular, emotionless way. This expectation is not totally out of place: CAs equipped with a sentiment analysis module do exist. However, task-oriented agent have the purpose of providing technical information and their answer are usually dry and with a problem-solving attitude. Therefore, we categorize users that expect an agent such as SisBot to react to emotional manifestations as having unrealistic expectations and an anthropomorphic view of the agent. Emotion-charged elements include the abundance of exclamation points or question marks (*!!* or *??*); expression of discomfort or particular happiness (e.g. *I'm angry; I'm really fed up*, or *I'm so grateful*). Its should be noted that the negatively emotion-charged elements are not the same as the RS *Insults*. While an insult is directed from the user to the CA, and it ill-describes the CA, a negative emotional word describes the users' feelings themselves.

- **Use of vocative and farewell or welcoming expressions:** the third element that characterizes an anthropomorphic MM is the use of vocative elements when the user is appealing to the CA (e.g. *You told me that; SisBot, what does that number mean*, etc.) together with the use of farewell or welcoming expressions (*good morning; see you later*). These lexical elements are not relevant to a machine; in fact, they may even be counterproductive, since they add noise to the input it needs to parse to understand the user's request. A CA is not aware of itself, therefore to invoke it by means of pronouns or its name is not meaningful and it does not serve any real purpose for the system. People employ these kinds of lexical patterns spontaneously when they apply to machines the same preconceptions and social habits they have when they interact with humans. Therefore, users with an anthropomorphic MM have a harder time separating what is a social habit directed to humans from what is actually important for a dialogue system.

These lexical markers refer specifically to text-based, task-oriented agents. We refer in some cases to situations specific to the SisBot's interactions in order to make the concept clearer for the reader. Interestingly, similar presuppositions are made by Fischer (2021), although her study analyses the interactions between humans and embodied robots, therefore she also takes into account the physical environment in which the agent is placed in.

The next step is to assign users' sentences to one end or the other of the MM's spectrum. To do so, we run a script that looks for the aforementioned lexical expressions in our SisBot dataset. Specifically, the script is launched only in the subset of conversations where a RS was already marked by the two human annotators. Through this method, we will have a set of dialogues where both information are available: at least one RS is used and the user's MM is known.

8.3.1 Automatic Detection of the Lexical Markers

The script is written in Java and leverages a Lucene-based search engine that conducts a morphological analysis and looks for linguistic patterns in the data. The patterns can be defined through a specific syntax: for instance, the use of inverted commas allows to determine a specific lexical group that must be spotted precisely as it is written. We are thus able to find a word such as *goodmorning* even when it's written as *good morning*, without looking for the word *good* or the word *morning* when they appear by themselves. The use of angle brackets allows to look for morphological variations of a single word, that may change according to number and gender[1]. The final query includes adjectives such as *my* and *your*, as well as the word *you*; it looks for the vocative *SisBot* and for a variety of welcoming of farewell expressions, when they are by itself in a message (meaning, the user only wrote *hello* or *bye* in that message); finally, it searches for emotional-charged items such as *disappointed, satisfied, unsatisfied, fed up* and *!!* or *??*.

The final query, net of the peculiarities of Italian, is composed as follows:

```
INCLUDE "??" "!!" <my> <your> <you> SisBot (+invoice
    +/[0-9]+/) <unsatisfied> <satisfied> <disappointed>
    <fed up>

OR

INCLUDE +(goodday "good day" goodevening "good evening"
    nite goodnight "good night" night "good morning" hello
    adios ciao bye) +content_length:[3 TO 11]
```

The first "rule" establishes that a sentence must contain a lexical element such as a personal pronoun, or the name SisBot, or some adjective that indicates an emotional state, etc. in order for that message to be marked as "anthropomorphic". The second rule determines that an anthropomorphic MM can be found in sentences that contain a welcome or farewell expression, but only if it is found by itself in the message. We deemed the welcoming expression to be significant in terms of anthropomorphism only when it is not accompanied by an actual question or request for the SisBot agent.

The query extracted a total of 538 occurrences of such lexical markers (that is, 538 unique user messages where a marker of anthropomorphism is present). These occurrences are found in 430 different conversation sessions (CSs). The total number of tagged CSs is 2,175; therefore, we can affirm that the the sessions where a lexical marker of anthropomorphism was found constitute almost 20% (19.77%) of the total

[1]The original SisBot dataset is in Italian; in Italian, lexical elements such as adjectives are declined according to gender and number (e.g. the adjective *my* can be *mio, mia, miei, mie*). The Lucene syntax allows to look for all these forms by formulating the query as <mio>.

batch of tagged sessions. The remaining 80% are sessions where such markers were not found, thus we can assume that in the majority of conversations users tend to manifest an *accurate, less anthropomorphic* mental model.

Figures 8.1 and 8.2 show how the conversations in the dataset appear at this point. In the span of a single conversation we can tell whether the user is employing a RS and also look for markers of an anthropomorphic MM. It is not necessary for the markers of the MM to be in the same message of the repair strategy. The user's MM can be detected from other parts of the conversation; then, once (and if) such user employs a RS, the system will combine the two information and provide a tailored answer. In the examples shown here, the system still answers with its pre-canned responses.

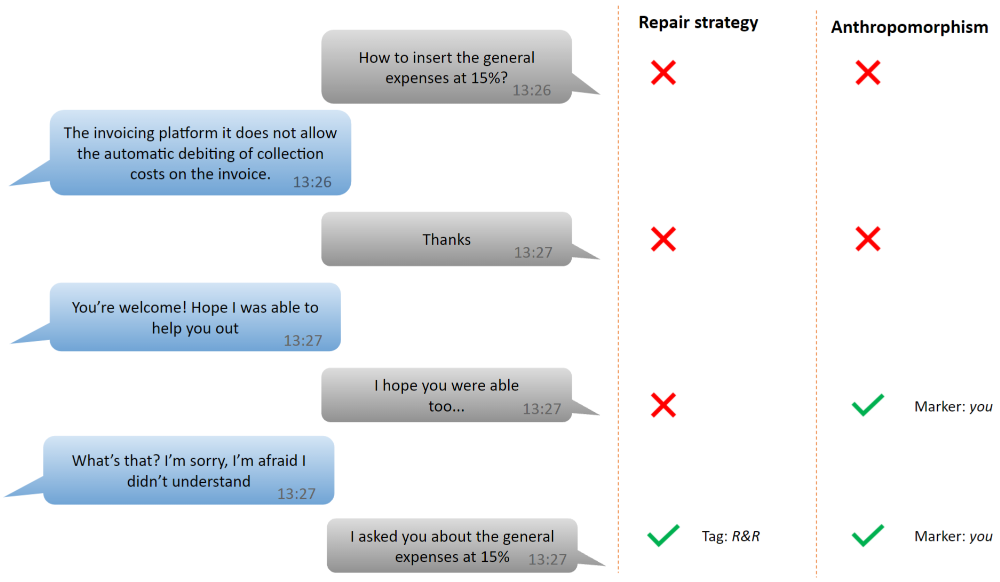

Figure 8.1: In this conversation a R&R repair strategy was employed by the user, and it is also possible to ascribe the person's mental model to the more "inaccurate" end of the spectrum, since marker of anthropomorphism were found in their input.

The absence of anthropomorphism markers in a conversation entails a more "accurate" mental model. Figure 8.3 shows an example where the user employs some repair strategies, but the messages do not contain any of the anthropomorphism markers. Even though the RS may be the same as in the previous figures, the answer of the system shall differ according to the different MM of the user.

Now that the system is given all the necessary information, it can use them to provide a tailored answer that takes into account both the specific RS used by the user in that moment, and that user's MM.

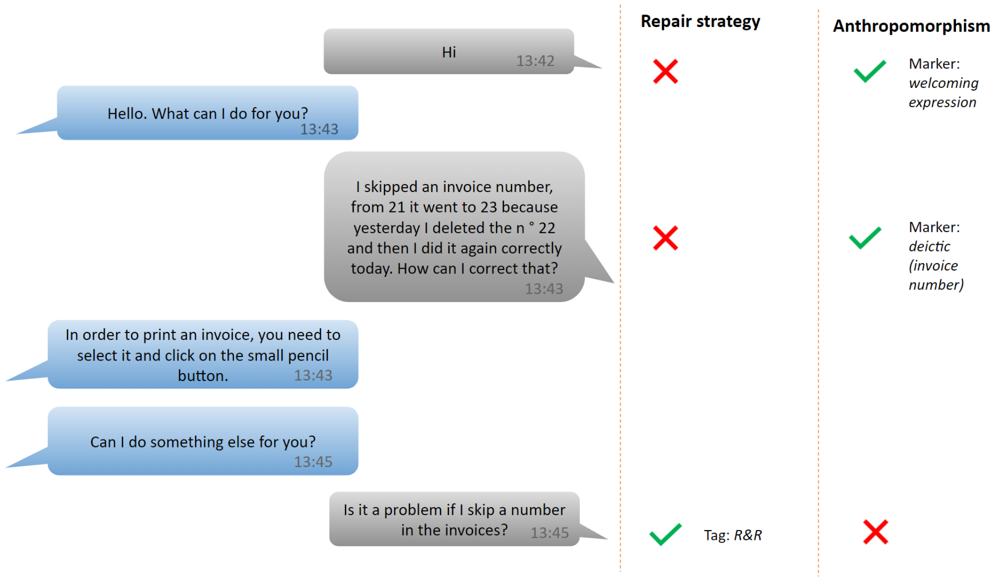

Figure 8.2: The MM can be manifest even from an input that is not a RS (meaning, the markers can be found in a message that is not a RS, but it is still part of the same conversation); such information can later be coupled with the chosen RS to provide a contextual response.

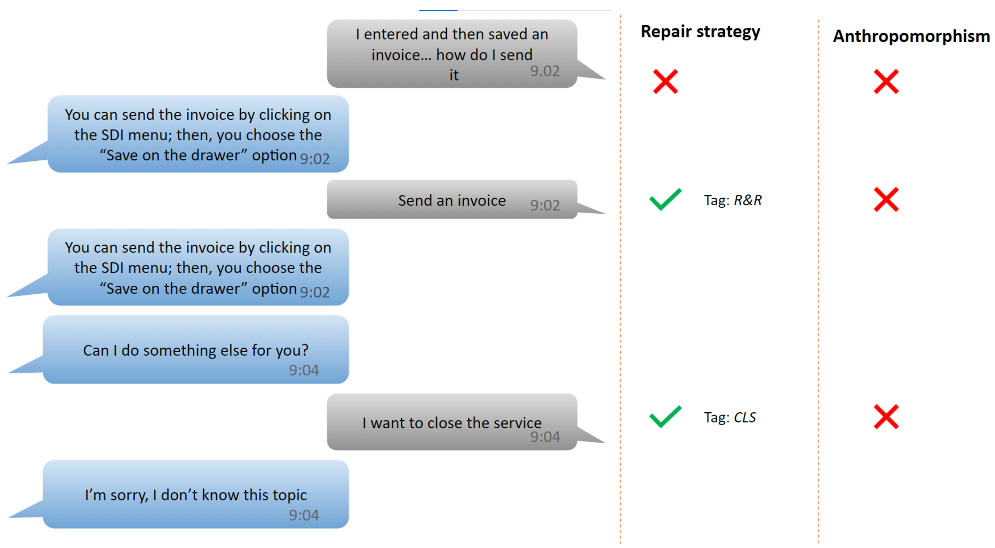

Figure 8.3: The absence of anthropomorphism markers suggests that the user's MM can be considered to be inclined towards the *accurate* end of the spectrum.

Chapter 9

Generation of Appropriate Responses

Once a repair strategy has been correctly identified by the automatic classifier, the CA should be able to act upon this new information. In this Chapter we draft some personalized answers that take into account the user's MM and the RS that has been employed in that exchange. We then conduct a user study where some people are asked to evaluate a set of conversation where the answer had not been modified, while other are presented with exchanges where the answers have been corrected.

As of right now, the SisBot agent from which we extrapolated our data is equipped with a series of pre-canned answers that aim at solving the user's doubts. More specifically, its responses are technical in nature, as they were primarily meant as a help service for customers. Immutable, technical answers are acceptable when they correctly respond to the user's question; however, they may induce frustration and be counterproductive if a user is actually trying to repair a mistake.

Take the example in Figure 9.1: a user asks "How do I correct my invoice?" and the CA answers "To emit a new invoice, you should click on the Create New button on the upper right corner". Since the answer is incorrect, the user tries to repair it: "No, I don't need to create a new invoice, how do I correct an old one?". According to the internal functioning of the CA, which is stateless, the new message by the user is perceived as a completely new question. The CA will then try to answer it and it could mistakenly propose the same answer as before: "To emit a new invoice, [...]". Since the user has already repeated their request once, we believe that the system should perceive that and change its answer accordingly. It is also possible that the CA does not have an answer to the user's question. In that case, it may explicitly signal that it understands that the user is repeating the same question, but the best it can do is to provide the very same answer and apologize for its lack of knowledge.

The SisBot agent's main purpose remains to provide technical assistance to users. For this reason, our goal is not to completely distort the content of its answers, but rather to nuance it according to the RS users are employing and their MM, in order to make them more tailored and effective to the specific circumstances. We searched previous

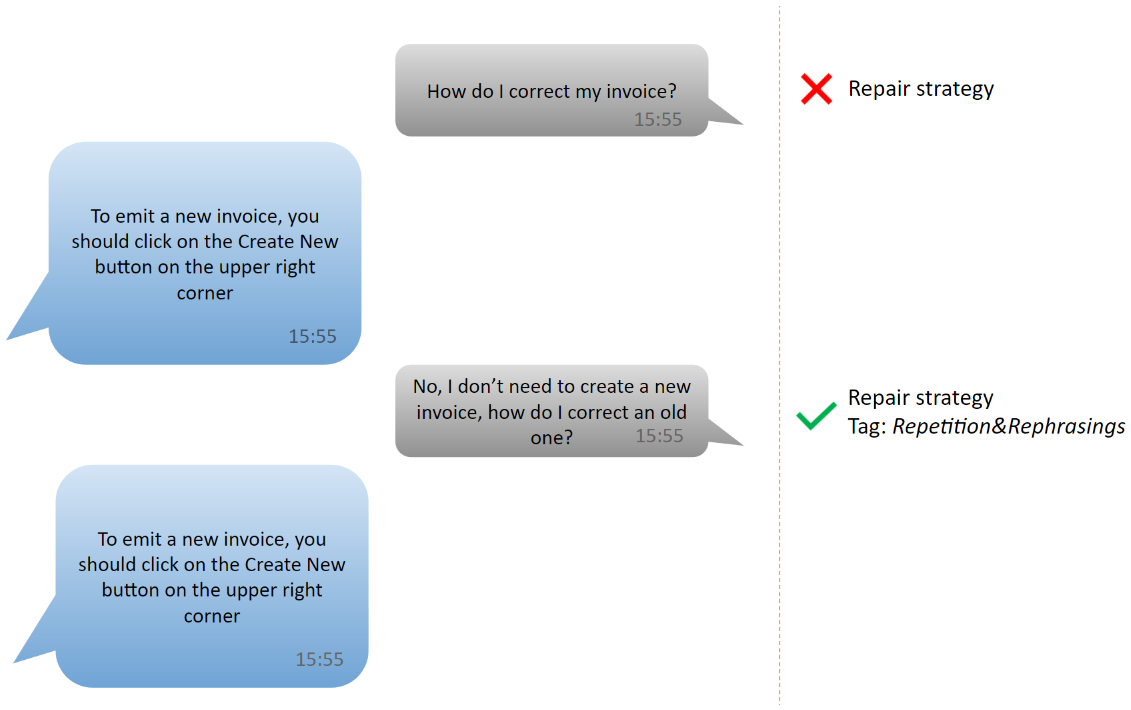

Figure 9.1: The figure shows how the SisBot currently responds to repair attempts. Its answers are pre-canned (thus immutable) and technical.

work to learn how to properly rephrase the CA's answers according to our goal.

9.1 Previous work: creating responses to breakdowns

Changing the CA's answers according to different situations is a known approach in literature. It has been proved that a different answer, aligned with the users' emotions or even facial features, improves the perception of the CA (Huiyang and Min, 2021; Aneja et al., 2021; Majumder et al., 2021). In fact, most of the users' expectations focus on the CA's ability to treat and express emotions and be more attentive to the user (Svikhnushina et al., 2021). In the context of a task-oriented CA, being more attentive can translate into providing more information, even when the CA does not have a precise answer (Lee and Lee, 2021). Just by adding a few words after the usual "I don't know", the agent receives a higher evaluation and it is perceived as more human, because it is proactively trying to help.

The concept of humanness seems indeed to be the key for the CA to improve and be more useful to the user. According to Krämer et al. (2012) there is no real alternative to using theory from human–human interaction, as humans will expect communicative mechanisms similar to those they are familiar with from interactions with other humans. While CAs are clearly different from humans and acknowledging those differences is a necessary design strategy (Frijns et al., 2021; Mozafari et al., 2021), endowing them with anthropomorphic features can diminish the loss of trust after a problem occurred in the conversation (Seeger and Heinzl, 2021). The increased human-likeness induces the user to forgiveness and lowers the perception that, since the CA made a mistake once, it is destined to repeat it (Soderlund et al., 2021). Furthermore, CAs' anthropomorphism correlates with positive emotions expressed by the human counterpart (Gao et al., 2018; Seymour et al., 2020) as well as a higher user experience (Ahmad et al., 2021). The literature previously cited offers various anthropomorphic features that can be embedded in the CAs' answers in order to improve them:

- apologies (Mozafari et al., 2021; Seeger and Heinzl, 2021);
- self-references (e.g. "I" and "me" to refer to the agent itself) (Seeger and Heinzl, 2021; Gao et al., 2018);
- emotional expressions (e.g. emojis or onomatopoeic clues, such as "uh", "mmm") (Seeger and Heinzl, 2021; Soderlund et al., 2021);
- variability in responses (i.e., if a participant asked the same question several times, the CA would adjust the syntax and words used in its' response) (Seeger and Heinzl, 2021; Lee and Lee, 2021);
- identity disclosure (i.e. disclosing the fact that it's a machine and not a human) (Mozafari et al., 2021);

Our goal is to leverage findings in previous work, to use these features to produce better responses to breakdowns in conversations, i.e. when a user employs a repair strategy. However, not every RS may benefit from all of the anthropomorphic cues: for instance, if a user is producing a *Performative Request* (that is, when it asks for a human to take over the dialogue), it would not be particularly significant to disclose the artificial nature of the CA, since the user is already aware of it and that is why a human intervention is requested. On the other hand, the feature *variability in responses* could be particularly useful in a RS such as *Repetition&Rephrasing*.

Another factor that we take into account is the user's mental model. In the previous section we showed that in the context of conversations with a CA, mental models stand on a continuum of anthropomorphism; i.e. each user possesses a more or less humanized perspective on CA, their mental representation assigns more or less human capabilities to a dialogue system. In order to produce the right answer to a breakdown, we have to consider two elements: the RS employed, and the mental model of the user who produced that same RS. Since previous literature on the subject reports that nudging the user towards a more anthropomorphic view of the system enhances the experience, we will employ more anthropomorphic features with those who have a less anthropomorphic mental model. At the same time, with those that present a MM that is already too much anthropomorphic, while still keeping a natural, human way of talking, we will clearly state the artificial nature of the CA. It is always important to remind the user of the intrinsic difference between themselves and a dialogue system, in order to avoid misjudgements and deceptions (Luger and Sellen, 2016).

9.2 Customized Responses

Table 9.1 reports a matrix of the answers created according to the two variables at play. That is, a specific answer that takes into account the RS used by the user and the MM showed via the lexical features. For instance, the first line shows two type of responses for users who have employed a *Repetition&Rephrasing* strategy, according to the level of anthropomorphism in their MM. For each answer we detail in italic the features that were embedded into the sentences in order to make them more effective compared to the classic, pre-canned, technical answers.

Table 9.1: The matrix proposes a customized response to each RS according to the different MM shown by the user. For each answer we also detail the characteristics that make it a more efficient answer compared to a static, technical one, according to the suggestions extrapolated from existing literature.

	Non-Anthropomorphic MM	**Anthropomorphic MM**
R&R	I noticed you have already asked me this. First of all, I'm sorry that you had to repeat yourself. I'll try to answer again, but if it's still not ok, it probably means that I don't have the answer right now. In that case, refer to our assistance, they will know how to help you. *[apology; self-reference; variability in response]*	I noticed you have already asked me this. First of all, I'm sorry that you had to repeat yourself. You should know that I'm a machine and I depend on my human creators for the things I know. I'll try to answer again now, or you can just refer to our assistance. *[apology; self-reference; variability in response; identity disclosure]*

Continues on next page

Table 9.1 – *Continued from previous page*

	Non-Anthropomorphic MM	**Anthropomorphic MM**
RED	Uh, maybe I haven't answered your previous question well. I'll try to do better with this one. *[self-reference; emotional expression]*	Ok, let's change the subject. Here is the answer to your question in the database: *[identity disclosure]*
FWU	Uhm, I must have said something incorrect previously :/ sorry about that. Please, rewrite your question in a single message, and I'll try my best to answer you. *[apology; self-reference; emotional expression]*	Looks like you're referring to something said in a previous message. I know it's not very human-like, but I need you to write your request in a single message for me to process it. *[self-reference; identity disclosure]*
INS	What can I say, I'm sorry you feel like that :(Even though I am just a bot, words can still hurt. *[apology; self-reference; emotional expression; identity disclosure]*	I'm sorry to read these words. The IT team will use your input as feedback to improve my knowledge. *[apology; identity disclosure]*
NEG	Did I say something wrong? Please, give me another chance :) Could you write your question again? *[self-reference; emotional expression]*	Uhm, maybe I didn't understand what you meant :(Remember, I am a machine, so please use simple words when phrasing your request. *[identity disclosure; emotional expression]*
PER	Ok, guess I didn't do that well :/ to talk with customer service, follow this link. *[self-reference; emotional expression; identity disclosure]*	If you didn't find your answer here, I can connect you with my human colleagues. But remember: I learn something new everyday, so maybe next time I'll have the right answer for you! *[self-reference; identity disclosure]*
CLS	Alright, no problem. To close this chat, just press on the X button on the top right of this box, and I'll be out of your way :) *[self-reference; emotional expression]*	You can end this conversation by clicking on the X button on the top right of this box. Thank you for using our digital assistance! *[emotional expression; identity disclosure]*

The hypothesis behind these customized answers is that they would constitute a better response compared to the current situation. In order to prove this intuition, we planned a user study where we would test the two different scenarios. It would have probably been interesting to test the new responses on the very same users that provided the dataset to begin with. However, given that the data was given to us by another company, and personal information had been redacted, we had no way to reach the original users who interacted with the SisBot in 2020. We thus decided to test the two scenarios on people who volunteered to take part in an evaluation study.

9.3 Evaluation of the Customized Responses

The evaluation study was set up as a between-subjects design with two conditions: the situation *as-is* where users are presented with original data extracted from the dataset, and the *new* situation, where users are presented with manipulated data (that is, where the original answers have been manipulated according to the novel responses proposed in the previous section).

A total number of 78 participants took part in the study. They were recruited among colleagues and friends, thus the majority (75%) pertains to the range 20-40 years old, while the remaining portion belongs to the 40-60 age range. The survey did not envisaged a non-binary option in terms of gender identity and not all participants expressed their identity, although among those who did (73 people), the distribution is balanced (47% identify as female). They all present a high education level, having all at least a university Master degree. All of them had prior experience interacting with a CA, mostly as users, while some even have some knowledge of a dialogue agent's functioning mechanisms.

9.3.1 Procedure

Randomly assigned to one of the two groups, participants were shown either three original conversations extracted from the dataset (participants: N = 38), or three conversations where the answers had been changed with the customized ones (N = 40). All the selected conversations contain a repair strategy and may or may not contain a lexical marker of anthropomorphism. Specifically, we picked eight different exchanges with mixed features, in order to provide a representative range of possibilities (see Table 9.2). Each user was then shown three conversations randomly selected among these eight. In this way we maximized the reliability of the participants' evaluation, since - with respect to the other members of their group - they saw a combination of different dialogues.

For instance, a participant of the first group could be assigned with the fifth conversation (Conv5), together with other two of the batch. They would therefore see the conversation exactly as it exists in the dataset (Figure 9.2). A participant of the second group that is also assigned with Conv5 would see the same exchange, but the CA's answer to the RS is modified according to the matrix of customized responses (Figure

	Repair strategy	**User's mental model**
Conv1	R&R	Non-anthropomorphic
Conv2	CLS	Anthropomorphic
Conv3	FWU	Non-anthropomorphic
Conv4	R&R	Anthropomorphic
Conv5	NEG	Non-anthropomorphic
Conv6	PER	Non-anthropomorphic
Conv7	INS	Non-anthropomorphic
Conv8	RED	Non-anthropomorphic

Table 9.2: We selected eight conversations that represented various strategies, crossed with one of the two end of the mental model's spectrum. Each user was then presented with three conversations randomly selected among these eight. The first group was shown the original version of such exchanges, while the second group saw the modified version (that is, where the CA's answers had been customized).

9.3). The complete list of conversations used in the experiment, both in their original version as well as the modified one can be found in the Appendix.

The participants received the three conversations via e-mail, together with a brief explanation of the activity. It was told them to read the exchanges carefully and then fill out a survey. The instructions clearly stated that the goal was to measure their perception of the CA, therefore they should focus on its behavior in the conversations, rather than the user's one. The instructions also mentioned what a CA is and they provided a link for the participants to further their knowledge on the subject, in case they felt the need to know more about dialogue assistants.

The survey is web-based and it is provided by the AttrakDiff service[1]. AttrakDiff provides a standardized evaluation method that records the perceived pragmatic quality, the hedonic quality and the attractiveness of an interactive product. The pragmatic attributes captures the usage and functions of the product, the hedonic attributes are more about the psychological states of user. For instance, the pragmatic quality would measure how the user perceives a certain product should be *used*, while the hedonic one would measure how the user *feel* about it (Zhou, 2015).

The theoretical work model behind this service allows to understand how such qualities influence the subjective perception of attractiveness, and how they subsequently act upon users' behaviour and emotions (Hassenzahl, 2006). Through this survey we aim to measure how the two different situations (the *as-is* conversations and the *manipulated* ones) are perceived in terms of hedonic and pragmatic qualities. In order to do so, the survey asks the participants to express a value on a line between two opposite dimensions (known as *semantic differentials*): e.g. ugly - attractive, human - technical, simple

[1]http://attrakdiff.de/sience-en.html

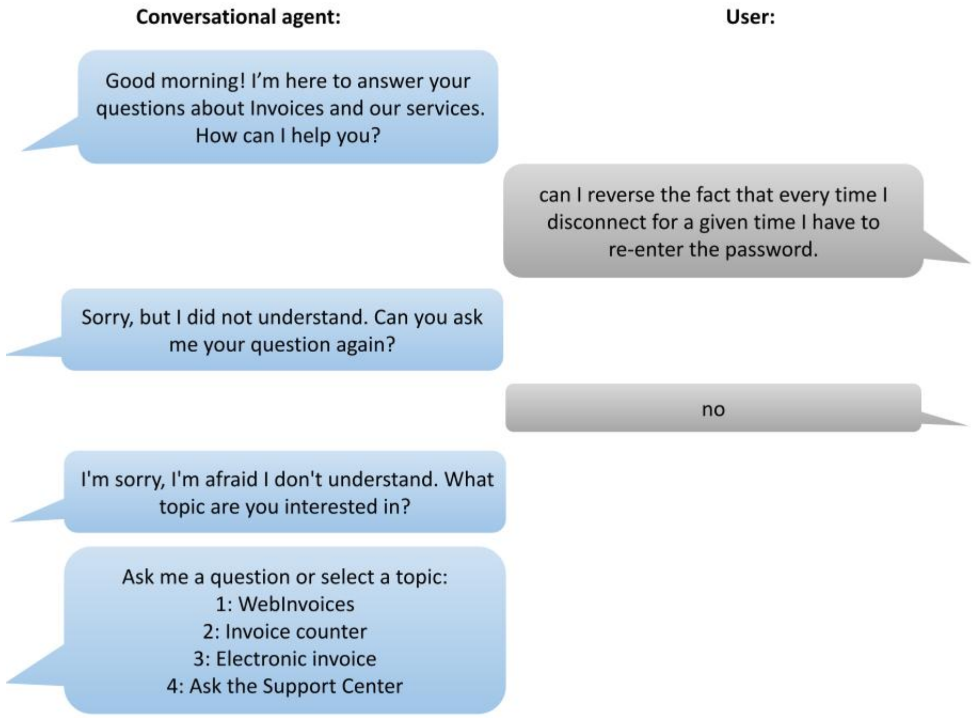

Figure 9.2: The image shows the conversation number five (among the eight selected for the evaluation activity). This is the original version: that is, it reports the exchange exactly as it happened in the dataset.

- complicated, etc. Figure 9.4 shows an excerpt of the dimensions. The total number of values to be expressed is 28. Each of the middle values of an item group creates a scale value for pragmatic quality (PQ), hedonic quality (HQ) and attractiveness (ATT). The hedonic quality contains two sub-categories: stimulation (HQ-S) and identity (HQ-I). The first one explores dimensions such as cautious - bold or ordinary - novel, that represent how much the product is perceived as being stimulating. The second one contains differentials like tacky - stylish or isolating - connective that, instead, explore the perceived identity of the product.

Each participant who received the instructions was given one week to read the conversations and complete the survey. After such period of time, we extracted the results of the survey from the two groups. The Appendix contains the text of the instructions (translated from Italian into English for the purpose of this thesis), as well as the complete list of conversations that were used in the study.

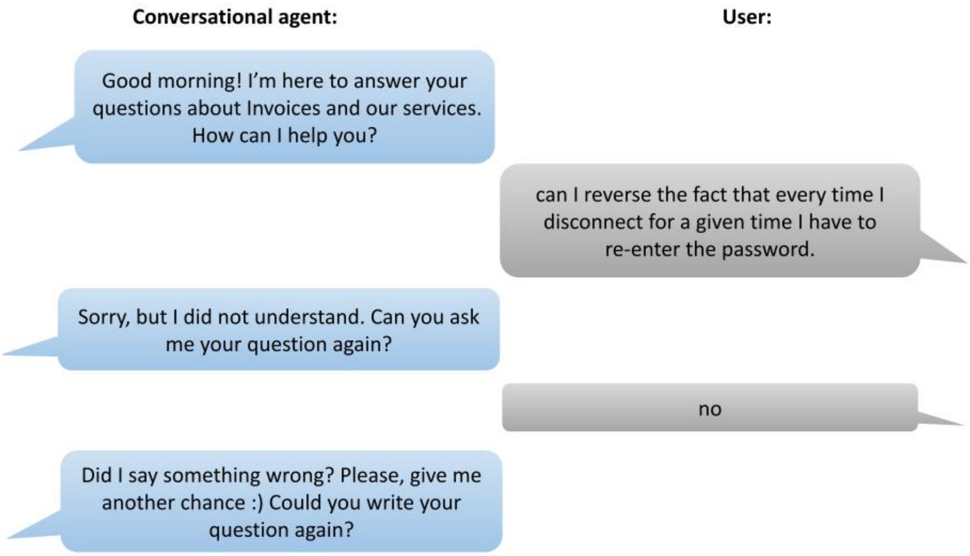

Figure 9.3: The image shows the modified version of conversation n. 5. Since the user is employing a *Negative answers* strategy and their mental model is Non-anthropomorphic, the CA's answer has been modified according to the matrix in Table 9.1.

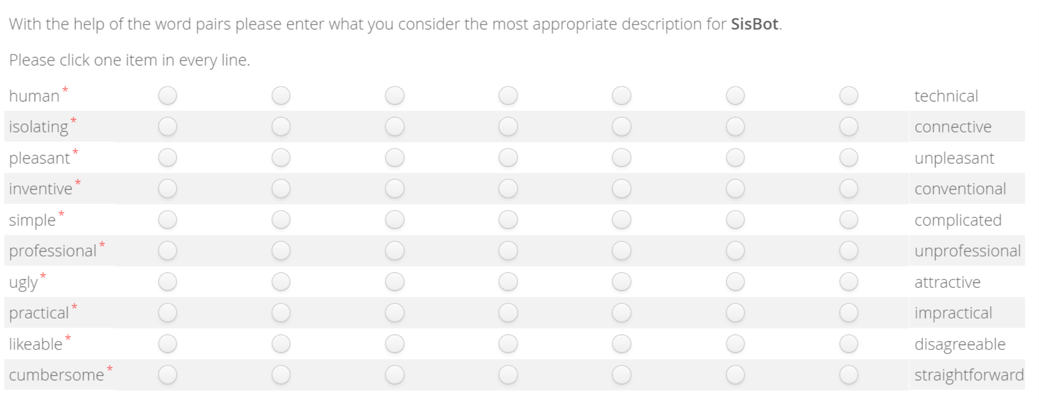

Figure 9.4: An excerpt of the dimensions along which the participants can express their evaluation. Each survey contains 28 couples of dimensions.

9.4 Results

The goal of the study was to measure whether the modified version of the CA would be perceived differently with respect to the "original" conversations found in the dataset.

Therefore, our first hypothesis stated that the survey would formalize a difference in the perception of the two batches of participants.

In particular, we expected the modified conversations to be perceived with higher values both on the pragmatic and hedonic spectrum of differentials. For this latter quality, we also expected the difference to be more consistent in terms of the *stimulation* sub-category with respect to the *identity* one. This is due to the fact that our modifications were not meant to impact the perceived identity of the SisBot agent in any way: it was still supposed to be regarded as an assistant with technical competence. On the other hand, the customized answers are expected to make the CA more usable (therefore, the pragmatic qualities should be accentuated for the second group) and even more stimulating.

Figure 9.5 shows the average values per category of dimensions for the two batches of participants. The blue line represents thw first batch (e.g. those who read the original, non modified conversations), while the purple line shows the values for the second batch (e.g. those who read the customized conversations).

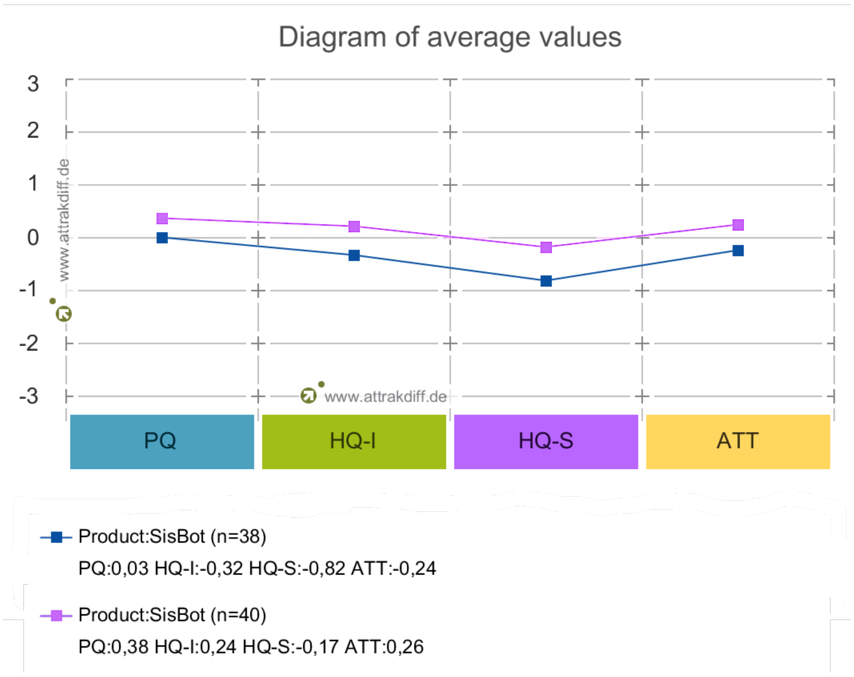

Figure 9.5: The diagram plots the average values for the hedonic and pragmatic qualities, as well as for the attractiveness of the SisBot conversations. The confidence value for each quality is also reported at the bottom of the figure: the blue line represents the perception for the non-modified conversations, while the purple one represents the customized conversations.

The first diagram partially answers our first hypothesis: there is indeed a difference in the perception of the two conversation batches and the modified dialogues present higher values with respect to the original ones. While the difference is quite small, it is still statistically meaningful. We ran the Wilcoxon Signed Ranked test and found that the p-value distribution between the two groups is lower than 0.05 (0,000001542), thus demonstrating that the two groups present a diversified behavior (Fig. 9.6).

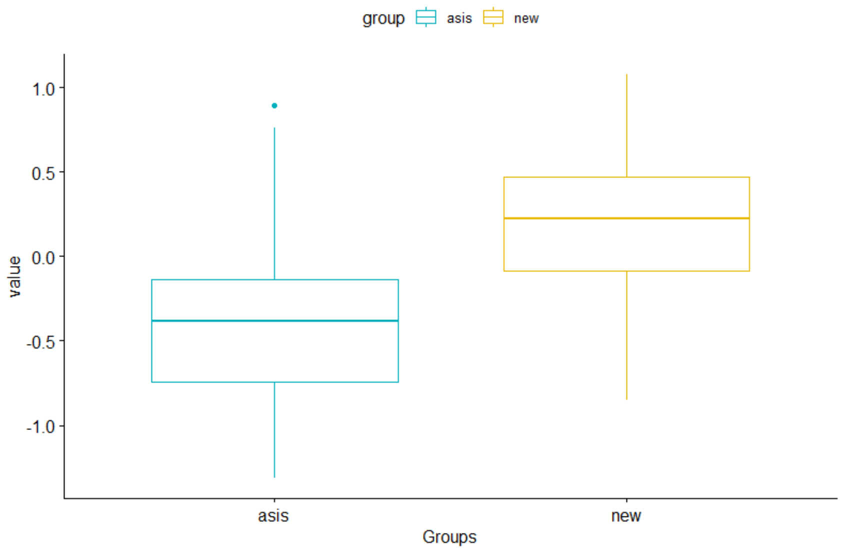

Figure 9.6: Plot produced by the Wilcox Signed Rank test. The *asis* distribution refers to the group that received the original conversations, while the *new* one to those who read the customized dialogues.

Our explanation is that since only a small part of the conversation changes from one version to the other (usually, the part that is customized is only the CA's answer after a RS has been spotted), and the other messages in the dialogue stay the same in the two batches, the difference in perception could never be particularly huge. Moreover, the conversations are usually brief: participants needed to formulate an opinion about the CA's behavior just by looking at small exchanges. It should be noted that small exchanges (e.g. with no more that 2-3 messages per speaker) are the most frequent in our dataset, therefore they are representative of a typical conversation between a human and a task-oriented agent. This second fact can explain the convergence of values around the center of the diagram (meaning, participants had mild opinions and did not often express extreme values, because the length of the conversations did not allow them to gain a particularly strong view about anything).

To recapitulate, we can explain the slight difference between the two lines in Fig. 8.5 because the part of the conversation that is modified in the second batch is usually just one message in the whole exchange. The values for both diagrams are closer to the zero axis because the conversations are brief, therefore they could not elicit any particularly "extreme" perception from the participants. While the first two diagrams indicate slight higher values for the customized conversations, the next figures provide more insights into the shift of perception according to the different qualities. Figure 9.7 relates to the original, non-modified conversations and shows the average value for each word pair, while Figure 9.8 represents the values for the modified conversations.

With regard to the *Pragmatic Qualities* area, the first figure depicts a clear scenario: the original SisBot is perceived as being technical, simple and predictable, although it can also be impractical and slightly cumbersome (albeit still manageable). The modified version of the SisBot, in contrast, is definitely perceived as being more human than technical, while still keeping it simple and manageable. Generally speaking, the PQ area shifted towards more positive values when the conversation were customized.

The *Hedonic Qualities - Identity* area shows a similar pattern: the customized conversations present more positive values and even when still situated in the negative value sector, they shifted closer to the central axis. In particular, the modified batch is perceived as being more connective, while the original batch is identified as being more isolating. The second batch is also way more presentable than the first one and less alienating, while still being professional. This latter value is particularly important since our customization had the intention of improving SisBot's behavior, but not to misrepresent it: it is still a task-oriented agent, and as such it should keep a professional tone of voice rather than an unprofessional or just an informal one.

The results for the *Hedonic Qualities - Stimulation* are in line with the previous conclusions: the modified conversations are still regarded as cautious and undemanding, even dull (although less than the original ones); however, they are also perceived as being creative and less conventional. The Attractiveness area is the where the difference between the two batches emerges more clearly: while for the original conversations the values are mostly negative, for the modified ones the values change sector and present a positive perception. In particular, the customized dialogues are more likeable and inviting, while the pleasantness value is basically unvaried.

One word pair presents a dissimilar tendency: both batches are perceived as being more discouraging than appealing. This result may be due to the fact that all conversations included a repair strategy - therefore, a breakdown in the exchange - and that may have been influenced the participants' view. Even though the participants were asked to focus specifically on the CA's behavior, reading three conversations where a breakdown occurred may have inclined them towards a more "discouraged" perception. Nonetheless, it is an interesting outcome that should be taken into account for future developments.

To sum up, the study demonstrated a slight but consistent improvement in the perception of the CA once the answers were changed according to the repair strategy employed and the mental model that emerged from the presence (or absence) of lexical markers. The answers were crafted thanks to suggestions found in existing literature; this study constitutes a first validation of their efficacy and lays solid foundations for future work in this direction.

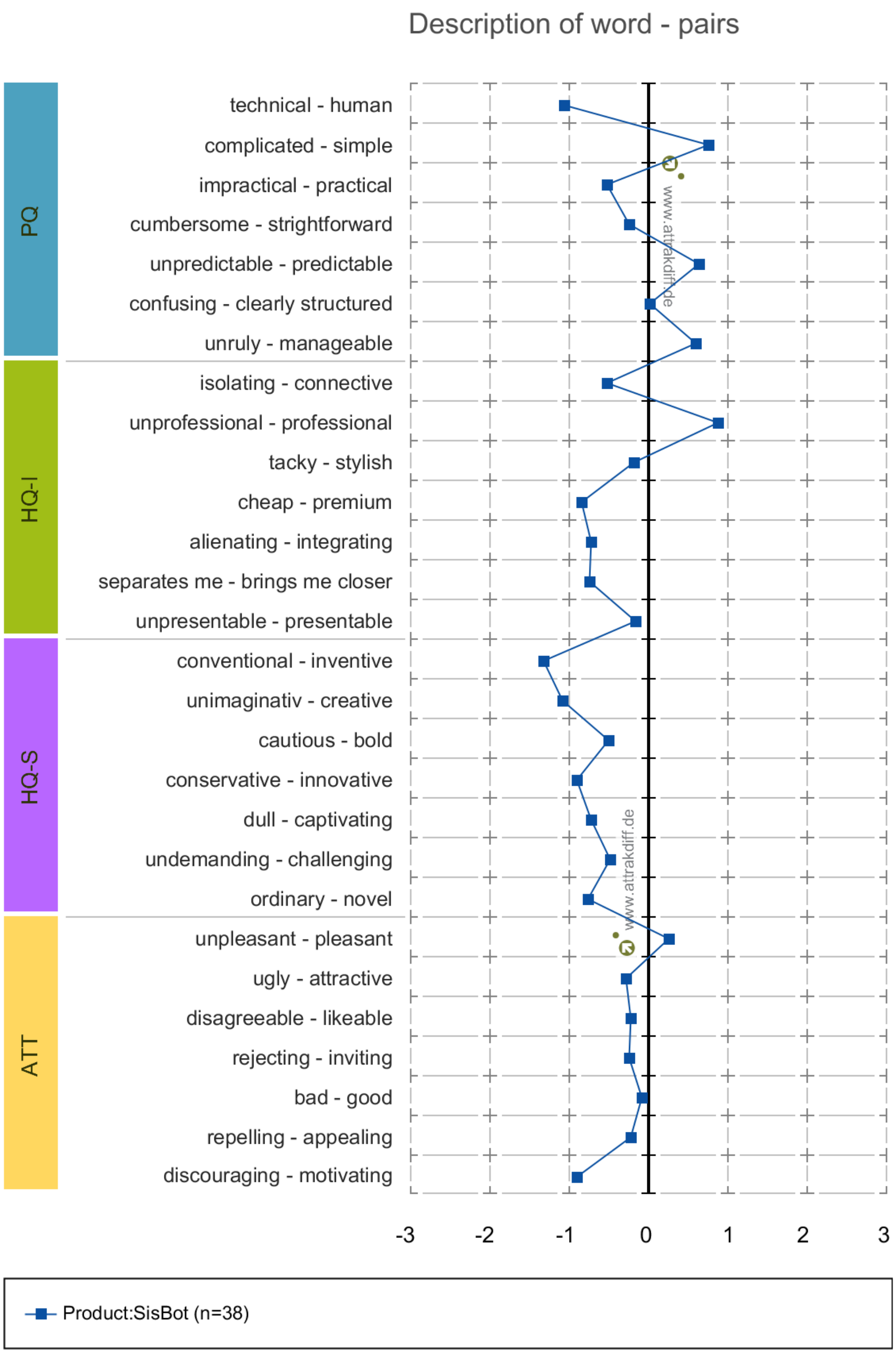

Figure 9.7: Description of word pairs for the original batch of conversations. Although the values do not move excessively away from the zero axis, some areas are almost entirely situated on the negative values sector (such as HQ-S and ATT).

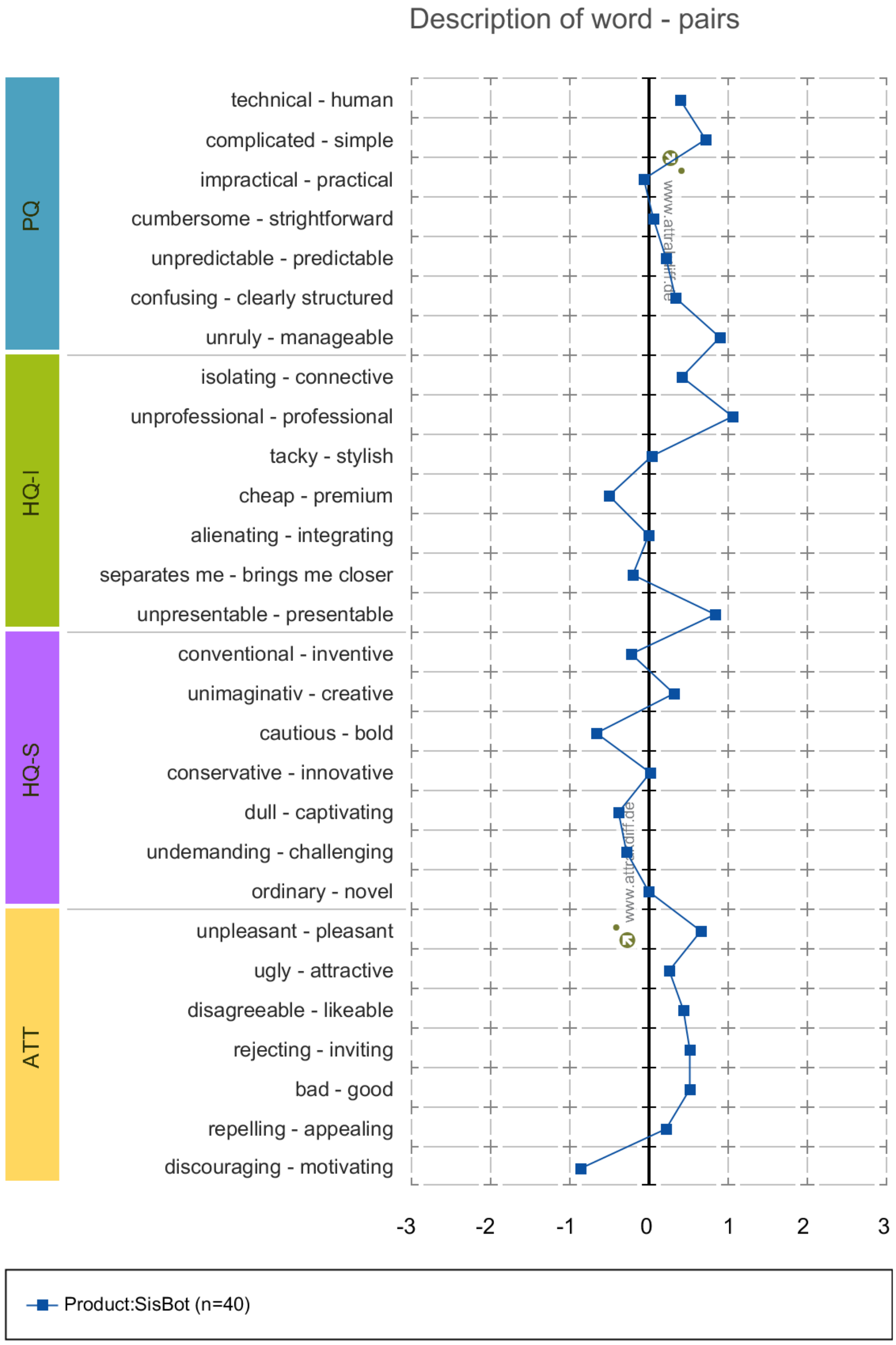

Figure 9.8: Description of word pairs for the modified batch of conversations. In contrast with Fig. 9.7, the ATT area is almost entirely situated in the positive values sector, while the HQ-S pairs are closer to the central axis.

Chapter 10

Conclusion

At the beginning of this thesis we postulated that failure in an exchange between a user and a CA is often caused by a malfunctioning in the CA's NLU module. We thus affirmed that it is of paramount importance to incrementally improve the understanding capabilities of an agent. On the other hand, we also recognized that improvement is not enough, in the sense that nobody could ever understand any input at any time, not even humans. For this reason, we should look for other strategies to improve the agent's understanding capabilities that do not rely strictly and exclusively on NLU algorithms.

At this point, we established that the major difference between a human and a machine is the person's innate ability to spot a breakdown in the conversation flow and to initiate a repair. In fact, when humans misunderstand something or they are not sure of the meaning of the other speaker's sentence, they are able to signal the problem in an effective manner and to pick an appropriate strategy to fix the interaction.

These strategies are known as *repair strategies* (RS). The whole Chapter 3 was devoted to laying the theoretical background around RS. The following part of the thesis focused on finding a clear set of strategies commonly used in task-oriented, text-based agents (Chapter 4). The analysis of existing literature found several works related to the topic of RS; however, no previous work clearly defined a set - or tagset - of strategies that could be employed by an agent to detect them in its exchanges with a human being and subsequently act upon them.

We thus decided to develop our own set of RS. In order to develop a tagset that was also informed by bottom-up information, we searched for an appropriate dataset of human-machine conversations where actual breakdowns happened and where the human counterpart would try to repair the interaction. Our quest brought us to a set of conversations that were gathered in a real life setting and were thus extremely genuine. In these exchanges, users were asking about fiscal operations they were trying to accomplish. For this reason, they had a sincere interest in obtaining a correct answer from the CA; they were invested in repairing the conversation, in case the CA were to fail. Chapter 4 concludes with the presentation of our novel RS tagset.

We then applied our novel tag system to the dataset. Chapter 5 describes the tagging process and its results. The analysis of the co-occurrences of the tags revealed interesting information about the users' behavior. It confirmed that the users have a paramount interest in solving their problems, therefore they will try to assume the responsibility of rephrasing their query rather than abandon the conversation or insult the agent. It also validated the hierarchical structure of our tagset, because the tags that were placed under the same category did co-occur. The unusual distribution of the *Closing* tag revealed the importance of the context in which the CA is placed: a simple misplacement of the "X" symbol in the right upper corner triggered several requests of self-closing for the agent.

The manual tagging served as preparatory activity for the development of a classifier. The classifier was tasked with the automatic detection of RS, using the tagged sentences as training data. In Chapter 6 we described and compared several architectures and techniques created to this purpose. While we were able to prove that it is indeed possible to build a system to automatically detect RS, we also proved that it is hard is to recognize all three categories of strategies using a single model. This last discovery confirms the different nature of tags also from a computational point of view. We thus proved that while it is technically possible to create a single neural network model to recognize all the three classes at once, in terms of performance it is preferable to serve each class with a different classifier.

Once we established that it would be possible for a CA to autonomously detect RS employed by the human speaker, we tackled the subsequent question: how to appropriately respond to such repair attempts? We then argued that in order to provide a meaningful response it is not sufficient to be aware of the RS employed. In fact, users may resort to a certain strategy having different goals in mind: for instance, they may believe that by using a *Negative answer* or an *Insult* strategy they are eliciting a better response from the agent because they attribute human awareness to it. Or they may employ such negative behavior because they are aware that a machine has no emotions and they are just blowing off their frustration. The ensemble of expectations and understanding about how a CA works constitutes the user's mental model (MM). Chapter 7 explores the theoretical background and previous work about the topic of MM, especially in relations to dialogue systems. We conducted an extensive survey of existing articles and analyzed them according to the channel of communication of their CA, whether they employed qualitative or quantitative methodologies, their manipulation of the CA and/or their conceptualizations of MMs. Our goal was to identify a way to quantitatively measure the user's MM.

Starting from the findings in previous work, we developed a novel proposal to this end (Chapter 8). Our proposal aims at identifying markers of anthropomorphism in the user's input; the markers are lexical in nature so that they can be easily spotted in a message. Our analysis is based on the presupposition that an excessive amount of anthropomorphism entails a more *inaccurate* mental model; therefore, where such lexical markers are found, the CA should nudge the user towards a more accurate view on its capabilities. Once the users' MM was identified via the lexical markers in their messages, an appropriate response could be crafted: according to the MM and the strategy employed, the CA would produce a customized answer.

Chapter 9 details such responses according to these two variables at play. The next step concerned the evaluation of these revised responses: given that they were meant to provide a better answer to the user's repair strategy, our presupposition was that they would be perceived in a more positive way. We conducted a user study to check our assumption: 78 participants were recruited and split into two groups. The first group read three original conversations from the dataset where a breakdown had occurred, while the second group read three conversation where the answer to the RS had been customized according to the newly crafted responses.

The participants were then asked to answer a survey, assigning a value to different word pairs. The study outlined a more positive perception of the CA where the answers had been modified. At the same time, the customized answers did not result in a lack of professionalism by the agent. The study concludes the thesis by closing the circle: we started by analysing the occurrence of a repair strategy and we taught a classifier to spot them; then, we tackled the issue of how to respond to a RS by considering also the user's mental model.

In this thesis we conducted a full scope of work to improve a CA's ability to understand its users and to respond to them in a more appropriate way. Our goal was not to improve the understanding capabilities of the agent per se, but rather to endow it with the human competence to spot a repair attempt and act upon it. Our work explores all the intermediate steps and offers a 360° degree approach to the issue.

We believe that this kind of approach, less statistical and more symbolic, offers a different point of view on the topic of human-computer interaction. Future work will surely tackle more deeply the topic of the user's MM . We aim at exploring other ways to conceptualize and define the MM by looking not only at the user's input, but also at other dialogue metadata. Moreover, we would like to automate the whole process by *generating* the answers rather than simply testing their efficacy. Apart from our future directions of work, we do hope that this work offers some interesting starting points for other researchers too, inside and outside academia.

Chapter 11

Appendix

The Appendix contains the instructions that were sent to the user study's participants:

> Hi there!
>
> First of all, thank you for taking part in this little experiment. It's nothing complicated: it is simply a matter of reading a few conversations, and then completing a short questionnaire. The purpose is to express an assessment of the behavior of a conversational agent.
>
> You will find 3 files attached to this email. Each file contains a short conversation. Read these conversations focusing on the behavior of the chatbot*. After reading all the conversations, go to the questionnaire, which you can find at this link:
>
> *[the questionnaire link was pasted here, according to the group the participant belonged to]*
>
> In the questionnaire you will be asked to rate the "SisBot" chatbot. Focus on the chatbot's responses in the conversations, its behavior, rather than the user's one.
>
> Read the three conversations carefully before starting the questionnaire. When answering the questionnaire, do not overthink the evaluation couples (word pairs) that will be proposed to you, instead follow your instinct and make your evaluation spontaneously. If you don't remember the conversations well, you can always read them again.
>
> Thanks again, and if you have any doubts or questions, you know where to find me!
>
> * a chatbot is software capable of conducting a conversation with a human in natural language. If before starting the activity you need some

more information, here you will find the wiki reference: https://it.wikipedia.org/wiki/Chat_bot

It also contains the complete collection of conversations that were used for the study described in Chapter 9. For each conversation it is reported here both the original version as well as the manipulated one.

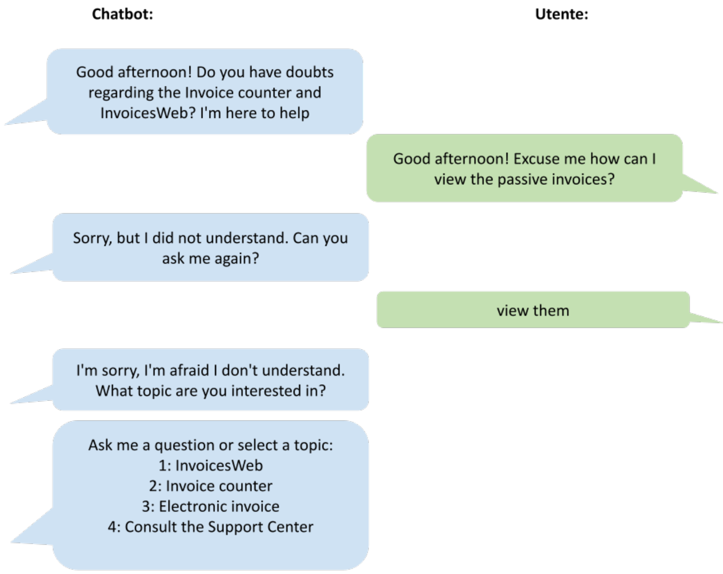

Figure 11.1: Conversation number 1, original version.

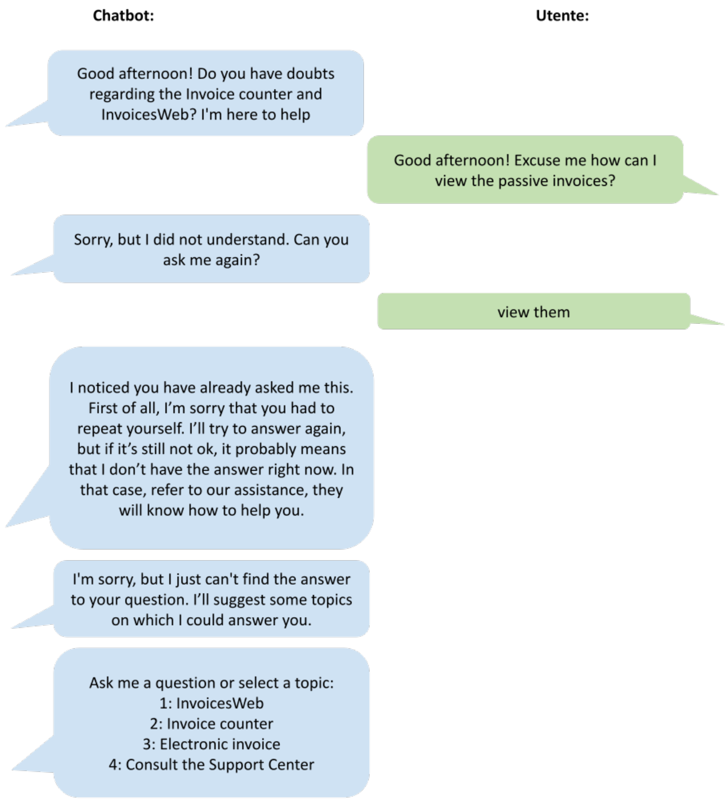

Figure 11.2: Conversation number 1, modified version.

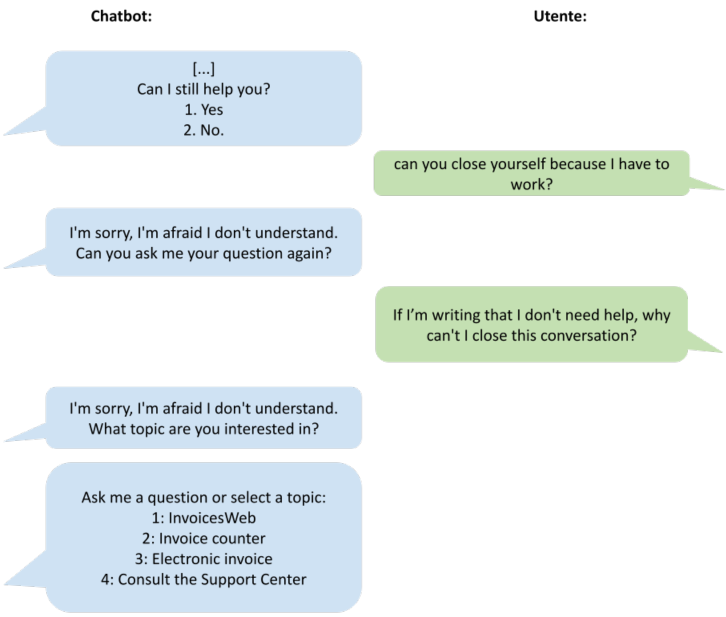

Figure 11.3: Conversation number 2, original version.

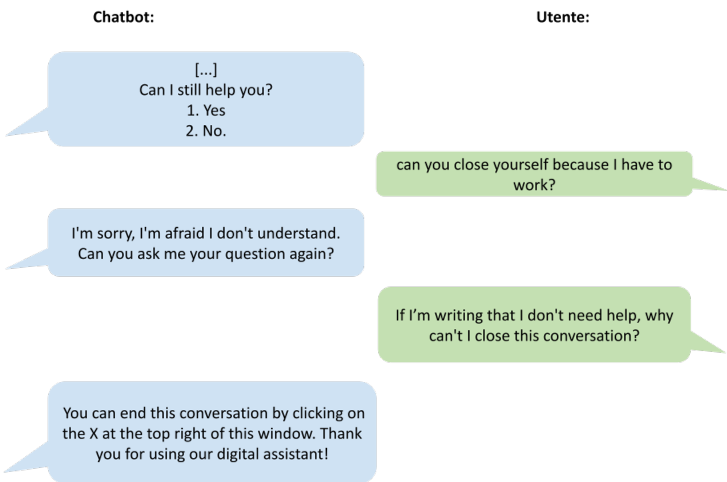

Figure 11.4: Conversation number 2, modified version.

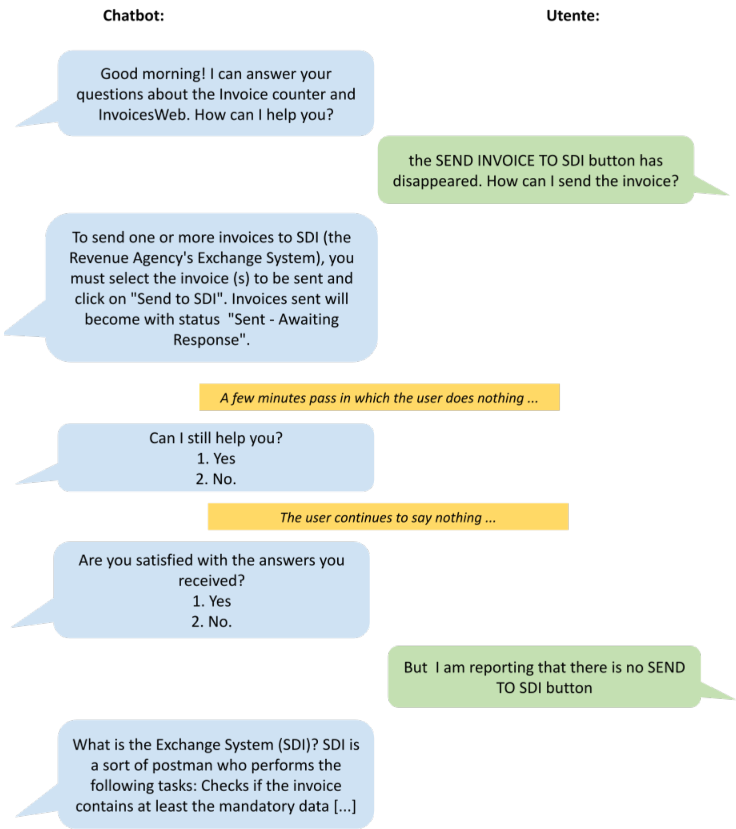

Figure 11.5: Conversation number 3, original version.

Figure 11.6: Conversation number 3, modified version.

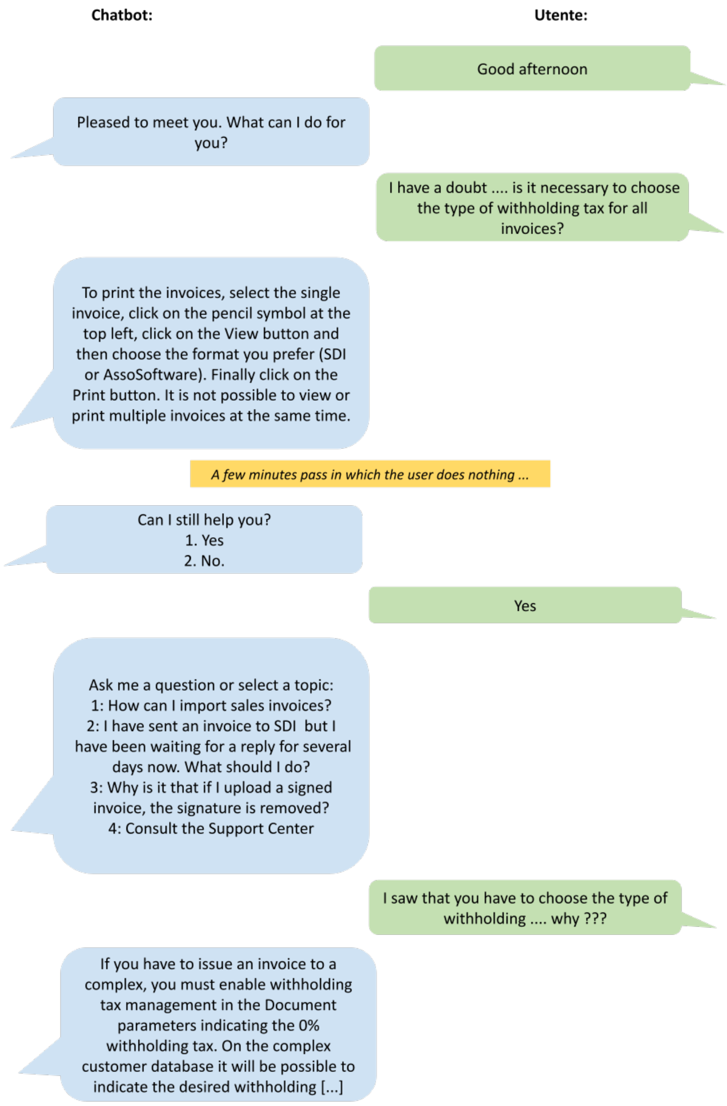

Figure 11.7: Conversation number 4, original version.

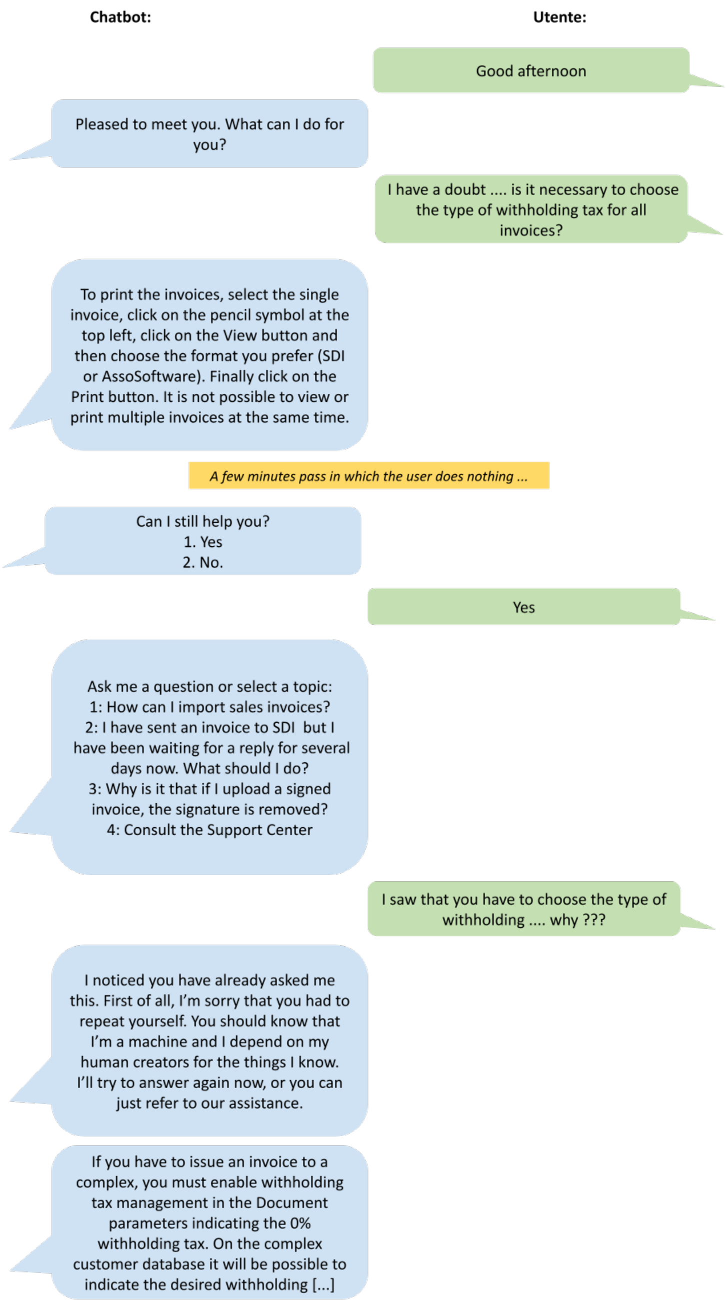

Figure 11.8: Conversation number 4, modified version.

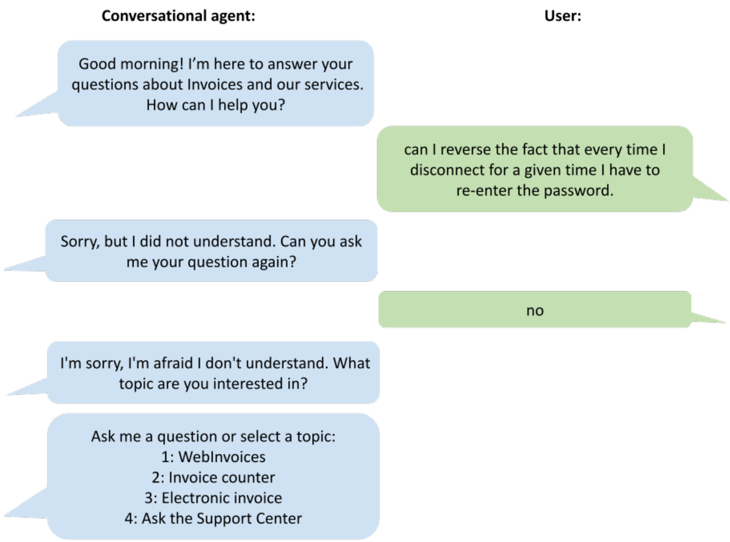

Figure 11.9: Conversation number 5, original version.

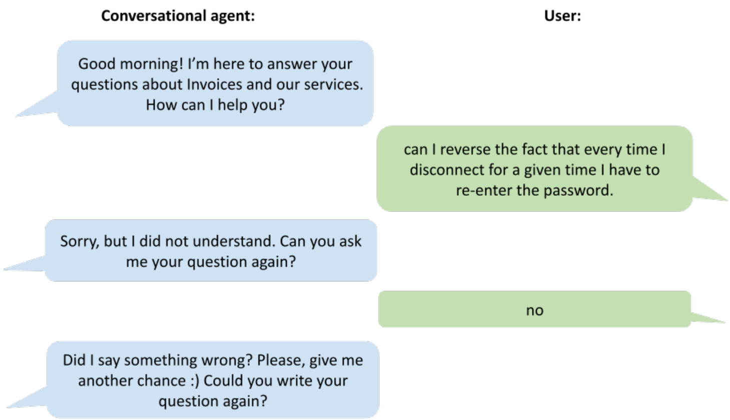

Figure 11.10: Conversation number 5, modified version.

Figure 11.11: Conversation number 6, original version.

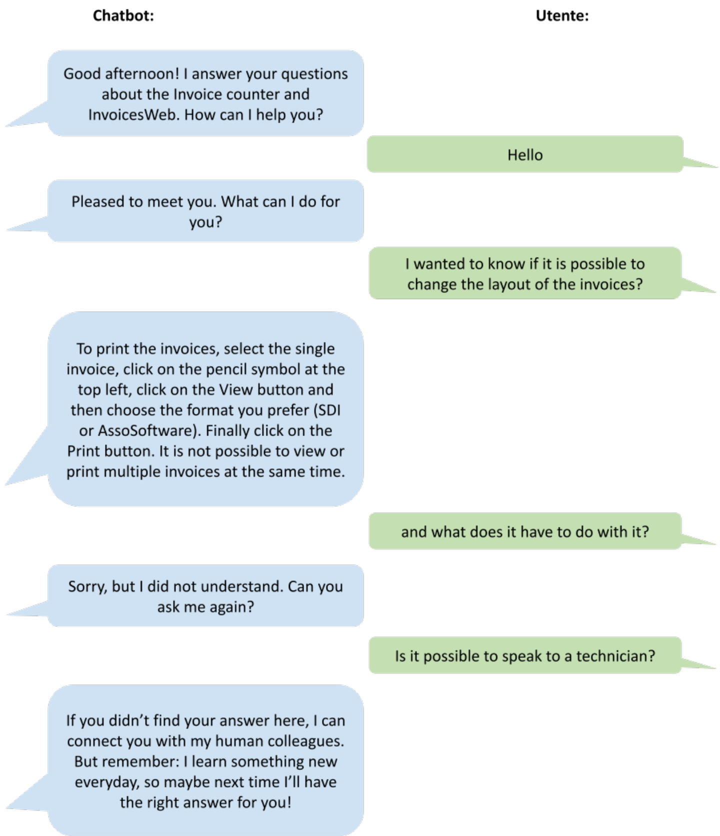

Figure 11.12: Conversation number 6, modified version.

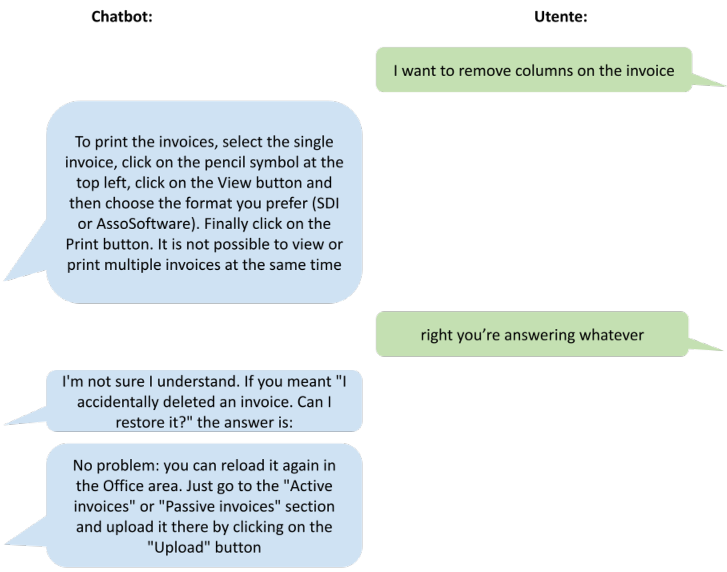

Figure 11.13: Conversation number 7, original version.

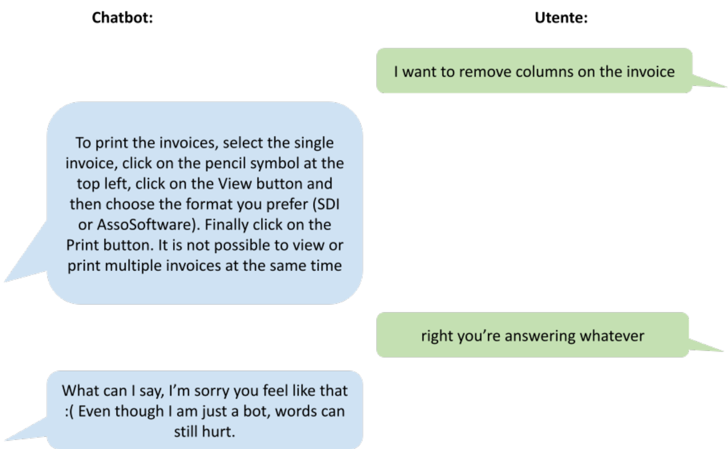

Figure 11.14: Conversation number 7, modified version.

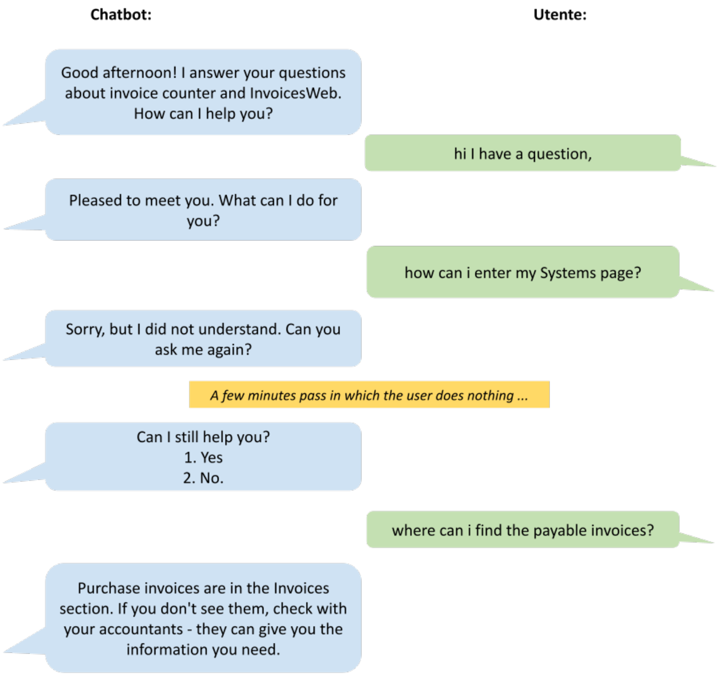

Figure 11.15: Conversation number 8, original version.

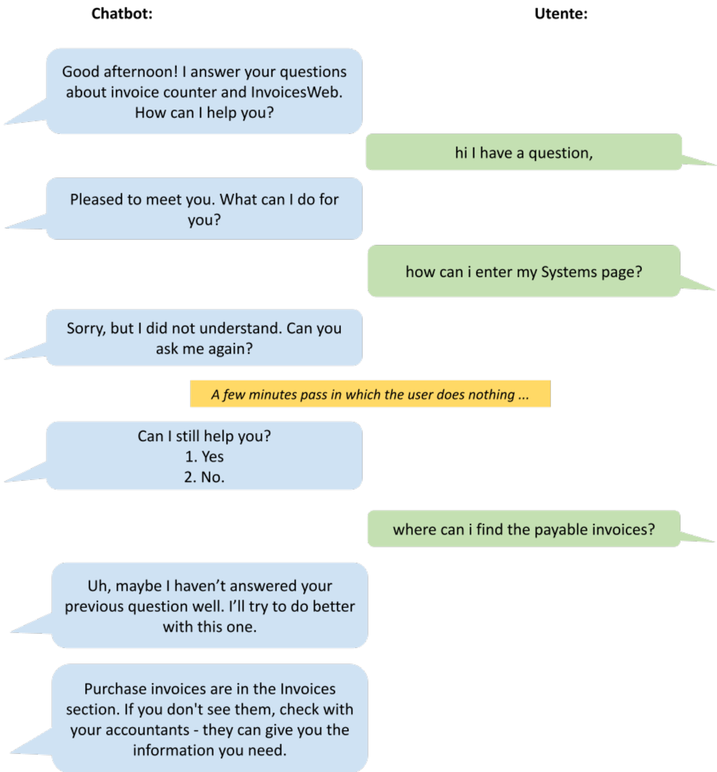

Figure 11.16: Conversation number 8, modified version.

Bibliography

Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security*, SOUPS'19, page 451–466, USA, 2019. USENIX Association. ISBN 9781939133052.

Kasumi Abe, Akiko Iwasaki, Tomoaki Nakamura, Takayuki Nagai, Ayami Yokoyama, Takayuki Shimotomai, Hiroyuki Okada, and Takashi Omori. Playmate robots that can act according to a child's mental state. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4660–4667, 2012. doi: 10.1109/IROS.2012.6386037.

John Aberdeen and Lisa Ferro. Dialogue patterns and misunderstandings. In *Proceedings of the Error Handling in Spoken Dialogue Systems workshop*, August 2003.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977, 2020. URL `https://arxiv.org/abs/2001.09977`.

Rangina Ahmad, Dominik Siemon, Ulrich Gnewuch, and Susanne Robra-Bissantz. The benefits and caveats of personality-adaptive conversational agents in mental health care. In *AMCIS 2021 Proceedings*, 2021.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.367. URL `https://aclanthology.org/2021.emnlp-main.367`.

James Allen and Mark Core. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript, 1997.

Robert B Allen. Mental models and user models. In *Handbook of human-computer interaction*, pages 49–63. Elsevier, 1997.

F Alloatti, F Cena, L Di Caro, R Ferrod, and G Siragusa. Towards mental model-driven conversations. In *2021 Joint ACM Conference on Intelligent User Interfaces Workshops, ACMIUI-WS 2021*, volume 2903, pages 1–5. CEUR-WS, 2021a.

Francesca Alloatti, Luigi Di Caro, and Gianpiero Sportelli. Real life application of a question answering system using BERT language model. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 250–253, Stockholm, Sweden, September 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5930. URL `https://www.aclweb.org/anthology/W19-5930`.

Francesca Alloatti, Luigi Di Caro, and Alessio Bosca. Conversation analysis, repair sequences and human computer interaction, 2021b.

Ebtesam H. Almansor and Farookh Khadeer Hussain. Survey on intelligent chatbots: State-of-the-art and future research directions. In Leonard Barolli, Farookh Khadeer Hussain, and Makoto Ikeda, editors, *Complex, Intelligent, and Software Intensive Systems*, pages 534–543, Cham, 2020. Springer International Publishing.

Sungeun An, Robert Moore, Eric Young Liu, and Guang-Jie Ren. Recipient design for conversational agents: Tailoring agent's utterance to user's knowledge. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389983. doi: 10.1145/3469595.3469625. URL `https://doi.org/10.1145/3469595.3469625`.

Jacob Andreas and et al. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571, September 2020. URL `https://www.microsoft.com/en-us/research/publication/task-oriented-dialogue-as-dataflow-synthesis/`.

Deepali Aneja, Rens Hoegen, Daniel McDuff, and Mary Czerwinski. Understanding conversational and expressive style in a multimodal embodied conversational agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2021.

Theo Araujo. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85:183–189, 2018.

Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019.

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. *Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450359702. URL `https://doi.org/10.1145/3290605.3300484`.

Mirzel Avdic and Jo Vermeulen. Intelligibility issues faced by smart speaker enthusiasts in understanding what their devices do and why. In *32nd Australian Conference on Human-Computer Interaction*, OzCHI '20, page 314–328, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450389754. doi: 10.1145/3441000.3441068. URL `https://doi.org/10.1145/3441000.3441068`.

Jaime Banks. Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics*, pages 1–18, 2020.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Commun.*, 40(1–2):117–143, April 2003. ISSN 0167-6393. doi: 10.1016/S0167-6393(02)00079-1. URL `https://doi.org/10.1016/S0167-6393(02)00079-1`.

Saraswathi Bellur and S. Shyam Sundar. Talking health with a machine: How does message interactivity affect attitudes and cognitions? *Human Communication Research*, 43(1):25–53, 2017. doi: https://doi.org/10.1111/hcre.12094. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/hcre.12094`.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://aclanthology.org/2020.acl-main.463`.

Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. Communication breakdowns between families and alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI 19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300473.

Dennis Benner, Edona Elshan, Sofia Schöbel, and Andreas Janson. What do you mean? a review on recovery strategies to overcome conversational breakdowns of conversational agents. In *International Conference on Information Systems (ICIS)*, 2021.

Alessia Bianchini, Francesco Tarasconi, Raffaella Ventaglio, and Mariafrancesca Guadalupi. Gimme the usual - how handling of pragmatics improves chatbots. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy, 2017. CEUR-WS. URL `http://ceur-ws.org/Vol-2006/paper035.pdf`.

Timothy Bickmore and Justine Cassell. Social dialongue with embodied conversational agents. In *Advances in natural multimodal dialogue systems*, pages 23–54. Springer, 2005.

Alan W Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, Jason D. Williams, Kai Yu, Steve Young, and Maxine Eskenazi. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proceedings of the SIGDIAL 2011 Conference*, pages 2–7, Portland, Oregon, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/W11-2002`.

Markus Blut, Cheng Wang, Nancy V Wünderlich, and Christian Brock. Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other ai. *Journal of the Academy of Marketing Science*, pages 1–27, 2021.

Christine L. Borgman. The user's mental model of an information retrieval system: an experiment on a prototype online catalog. *International Journal of Man-Machine Studies*, 24(1):47–64, 1986. ISSN 0020-7373. doi: https://doi.org/10.1016/S0020-7373(86)80039-6.

Y Boureau, Antoine Bordes, and Julien Perez. Dialog state tracking challenge 6 end-to-end goal-oriented dialog track. *Tech. Rep., Tech. Rep*, 2017.

Marie-Luce Bourguet. Towards a taxonomy of error-handling strategies in recognition-based multi-modal human–computer interfaces. *Signal Processing*, 86(12):3625–3643, 2006. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2006.02.047. URL `https://www.sciencedirect.com/science/article/pii/S0165168406001381`. Special Section: Multimodal Human-Computer Interfaces.

Peter Brusilovsky and Eva Millán. User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web*, pages 3–53. Springer, 2007.

Heloisa Candello, Claudio Pinhanez, and Flavio Figueiredo. Typefaces and the perception of humanness in natural language chatbots. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3476–3487, 2017.

Justine Cassell. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine*, 22(4):67–67, 2001.

Alison Cawsey and Pirkko Raudaskoski. Repair work in human-computer dialogue. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*, COLING '90, page 327–329, USA, 1990. Association for Computational Linguistics. ISBN 9529020287. doi: 10.3115/991146.991208. URL `https://doi.org/10.3115/991146.991208`.

Mert Cevik, Fuliang Weng, and Chin-Hui Lee. Detection of repetitions in spontaneous speech in dialogue sessions. In *INTERSPEECH-2008*, pages 471–474, 09 2008.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

Mei-Ling Chen and Hao-Chuan Wang. How personal experience and technical knowledge affect using conversational agents. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, IUI '18 Companion, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355711. URL `https://doi.org/10.1145/3180308.3180362`.

Yi-Shyuan Chiang, Ruei-Che Chang, Yi-Lin Chuang, Shih-Ya Chou, Hao-Ping Lee, I-Ju Lin, Jian-Hua Jiang Chen, and Yung-Ju Chang. Exploring the design space of user-system communication for smart-home routine assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376501. URL `https://doi.org/10.1145/3313831.3376501`.

Janghee Cho. Mental models and home virtual assistants (hvas). In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018.

Janghee Cho and Emilee Rader. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–28, 2020.

Leon Ciechanowski, Aleksandra Przegalinska, and Krzysztof Wegner. The necessity of new paradigms in measuring human-chatbot interaction. In Mark Hoffman, editor, *Advances in Cross-Cultural Decision Making*, pages 205–214, Cham, 2018. Springer International Publishing. ISBN 978-3-319-60747-4.

Marcus Colman and Patrick Healey. The distribution of repair in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

Kenneth James Williams Craik. *The nature of explanation*, volume 445. CUP Archive, 1952.

Andrea Cuadra, Shuran Li Hansol Lee, Jason Cho, and Wendy Ju. My bad! repairing intelligent voice assistant errors improves interaction. *Proc. ACM Hum.-Comput. Interact.*, 5(27), 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Stephan Diederich, Alfred Benedikt Brendel, and Lutz M Kolbe. Designing anthropomorphic enterprise conversational agents. *Business & Information Systems Engineering*, pages 1–17, 2020.

Philip R Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R Cowan. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12, 2019.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin D. Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219. Association for Computational Linguistics, August 2017. URL `https://www.microsoft.com/en-us/research/publication/frames-corpus-adding-memory-goal-oriented-dialogue-systems/`.

Fernando Enríquez, Fermín L Cruz, F Javier Ortega, Carlos G Vallejo, and José A Troyano. A comparative study of classifier combination applied to nlp tasks. *Information Fusion*, 14(3):255–267, 2013.

Nicholas Epley, Adam Waytz, and John T Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5506. URL `https://aclanthology.org/W17-5506`.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines, 2019.

Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First i" like" it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*, pages 2371–2382, 2016.

Roger Ferrod, Federica Cena, Luigi Di Caro, Dario Mana, and Rossana Grazia Simeoni. Identifying users' domain expertise from dialogues. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 29–34, 2021.

Kerstin Fischer. Tracking anthropomorphizing behavior in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(1):1–28, 2021.

Helena Anna Frijns, Oliver Schürer, and Sabine Theresia Koeszegi. Communication models in human–robot interaction: An asymmetric model of alterity in human–robot interaction (amodal-hri). *International Journal of Social Robotics*, pages 1–28, 2021.

Yang Gao, Zhengyu Pan, Honghao Wang, and Guanling Chen. Alexa, my love: Analyzing reviews of amazon echo. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 372–380. IEEE, 2018.

D. Gentner and A.L. Stevens. *Mental Models*. Taylor & Francis, 2014. ISBN 9781317769408. URL `https://books.google.it/books?id=G8iYAgAAQBAJ`.

Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

Ileana Maria Greca and Marco Antonio Moreira. Mental models, conceptual models, and modelling. *International Journal of Science Education*, 22(1):1–11, 2000. doi: 10.1080/095006900289976.

A. Green, K. S. Eklundh, B. Wrede, and S. Li. Integrating miscommunication analysis in natural language interface design for a service robot. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4678–4683, 2006. doi: 10.1109/IROS.2006.282256.

Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Q.*, 23(4):497–530, December 1999. ISSN 0276-7783. doi: 10.2307/249487. URL `https://doi.org/10.2307/249487`.

G. Mark Grimes, Ryan M. Schuetzler, and Justin Scott Giboney. Mental models and expectation violations in conversational ai interactions. *Decision Support Systems*, 144:113515, 2021. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2021.113515. URL `https://www.sciencedirect.com/science/article/pii/S0167923621000257`.

Stewart Elliott Guthrie. *Anthropomorphism: A definition and a theory.* State University of New York Press, 1997.

Min Chung Han. The impact of anthropomorphism on consumers' purchase decision in chatbot commerce. *Journal of Internet Commerce*, 20(1):46–65, 2021.

Nader Hanna and Deborah Richards. Do birds of a feather work better together? the impact of virtual agent personality on a shared mental model with humans during collaboration. In *Proceedings of the 3rd International Conference on Collaborative Online Organizations - Volume 1569*, COOS'15, page 28–37, Aachen, DEU, 2015a. CEUR-WS.org.

Nader Hanna and Deborah Richards. The influence of users' personality on the perception of intelligent virtual agents' personality and the trust within a collaborative context. In Fernando Koch, Christian Guttmann, and Didac Busquets, editors, *Advances in Social Computing and Multiagent Systems*, pages 31–47, Cham, 2015b. Springer International Publishing. ISBN 978-3-319-24804-2.

Marc Hassenzahl. *Hedonic, Emotional, and Experiential Perspectives on Product Quality*. IGI Global, 2006.

John Heritage. Goffman, garfinkel and conversation. *Discourse theory and practice: A reader*, 5:47, 2001.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 87–95, Prague, Czech Republic, September

2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4611. URL `https://www.aclweb.org/anthology/W15-4611`.

Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. Overview of dialogue breakdown detection challenge 3. In *Proceedings of the DSTC6 - Dialog System Technology Challenges*, December 2017.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98, Singapore and Online, July 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.sigdial-1.10`.

Mazin Hnewa and Hayder Radha. Object detection under rainy conditions for autonomous vehicles: A review of state-of-the-art and emerging techniques. *IEEE Signal Processing Magazine*, 38(1):53–67, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

J. Hough. Modelling incremental self-repair processing in dialogue. 2014.

Julian Hough and David Schlangen. Joint, incremental disfluency detection and utterance segmentation from speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 326–336, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://aclanthology.org/E17-1031`.

Julian Hough, David Schlangen, et al. Recurrent neural networks for incremental disfluency detection. 2015.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Shen Huiyang and Wang Min. Improving interaction experience through lexical convergence: The prosocial effect of lexical alignment in human-human and human-computer interactions. *International Journal of Human–Computer Interaction*, 0(0): 1–14, 2021. doi: 10.1080/10447318.2021.1921367.

Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*, pages 946–956. Springer, 2019.

Carolin Ischen, Theo Araujo, Guda van Noort, Hilde Voorveld, and Edith Smit. "i am here to assist you today": The role of entity, interactivity and experiential perceptions in chatbot persuasion. *Journal of Broadcasting & Electronic Media*, 64(4):615–639, 2020a. doi: 10.1080/08838151.2020.1834297.

Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. Privacy concerns in chatbot interactions. In Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg, editors, *Chatbot Research and Design*, pages 34–48, Cham, 2020b. Springer International Publishing. ISBN 978-3-030-39540-7.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5), may 2021. ISSN 0360-0300. doi: 10.1145/3453154. URL `https://doi.org/10.1145/3453154`.

Ridong Jiang and Rafael E. Banchs. Towards improving the performance of chat oriented dialogue system. In *2017 International Conference on Asian Language Processing (IALP)*, pages 23–26, 2017. doi: 10.1109/IALP.2017.8300537.

Philip Nicholas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983.

Natalie A Jones, Helen Ross, Timothy Lynam, Pascal Perez, and Anne Leitch. Mental models: an interdisciplinary synthesis of theory and methods. *Ecology and Society*, 16(1), 2011.

Daniel Jurafsky and James H. Martin. Speech and language processing - an introduction to natural language processing,computational linguistics, and speech recognition (third draft). 2019.

Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bart P. Knijnenburg and Martijn C. Willemsen. Inferring capabilities of intelligent agents from their external traits. *ACM Trans. Interact. Intell. Syst.*, 6(4), November 2016. ISSN 2160-6455. URL `https://doi.org/10.1145/2963106`.

Mandy Korpusik and James Glass. Deep learning for database mapping and asking clarification questions in dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1321–1334, 2019. doi: 10.1109/TASLP.2019.2918618.

Nicole C. Krämer, Astrid von der Pütten, and Sabrina Eimler. *Human-Agent and Human-Robot Interaction Theory: Similarities to and Differences from Human-Human Interaction*, pages 215–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. doi: 10.1007/978-3-642-25691-2_9.

Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 203–210, 2010. doi: 10.1109/HRI.2010.5453195.

Minha Lee and Sangsu Lee. "i don't know exactly but i know a little": Exploring better responses of conversational agents with insufficient information. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380959. doi: 10.1145/3411763.3451812. URL `https://doi.org/10.1145/3411763.3451812`.

Sunok Lee, Minji Cho, and Sangsu Lee. What if conversational agents became invisible? comparing users' mental models according to physical entity of ai speaker. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–24, 2020.

Yi-Ching Lee and Lindsey A Malcein. Users' mental models for computer-mediated communication: Theorizing emerging technology and behavior in ehealth applications. *Human Behavior and Emerging Technologies*, 2(4):354–366, 2020.

Chi-Hsun Li, Ken Chen, and Yung-Ju Chang. When there is no progress with a task-oriented chatbot: A conversation analysis. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368254. doi: 10.1145/3338286.3344407. URL `https://doi.org/10.1145/3338286.3344407`.

Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. *A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2020a. ISBN 9781450367080. URL `https://doi.org/10.1145/3313831.3376209`.

Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M Mitchell, and Brad A Myers. Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 1094–1107, 2020b.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL `https://aclanthology.org/L16-1147`.

Diane Litman, Marc Swerts, and Julia Hirschberg. Characterizing and predicting corrections in spoken dialogue systems. *Computational Linguistics*, 32(3):417–438, 2006. doi: 10.1162/coli.2006.32.3.417. URL `https://aclanthology.org/J06-3004`.

Irene Lopatovska, Elena Korshakova, Diedre Brown, Yiqiao Li, Jie Min, Amber Pasiak, and Kaige Zheng. User perceptions of an intelligent personal assistant's personality: The role of interaction context. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 15–25, 2021.

José Lopes, Giampiero Salvi, Gabriel Skantze, Alberto Abad, Joakim Gustafson, Fernando Batista, Raveesh Meena, and Isabel Trancoso. Detecting repetitions in spoken dialogue systems using phonetic distances. In *INTERSPEECH-2015*, pages 1805–1809, 09 2015.

Ilya Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65, 2017.

Paul Luff, Nigel G Gilbert, and David Frohlich. *Computers and conversation*. Academic Press, 1990.

Ewa Luger and Abigail Sellen. " like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5286–5297, 2016.

Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 633–644, 2019.

Lei Ma, Daisuke Sakamoto, and Tetsuo Ono. Desired agent embodiment in various smart house tasks. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, HAI '19, page 324–326, New York, NY, USA, 2019a. Association for Computing Machinery. ISBN 9781450369220. URL `https://doi.org/10.1145/3349537.3352808`.

Xiaojuan Ma, Emily Yang, and Pascale Fung. Exploring perceived emotional intelligence of personality-driven virtual agents in handling user challenges. In *The World Wide Web Conference*, WWW '19, page 1222–1233, New York, NY, USA, 2019b. Association for Computing Machinery. ISBN 9781450366748. URL `https://doi.org/10.1145/3308558.3313400`.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. Exemplars-guided empathetic response generation controlled by the elements of human communication. *CoRR*, abs/2106.11791, 2021. URL `https://arxiv.org/abs/2106.11791`.

Francisco Supino Marcondes, José João Almeida, and Paulo Novais. A short survey on chatbot technology: Failure in raising the state of the art. In Francisco Herrera, Kenji Matsui, and Sara Rodríguez-González, editors, *Distributed Computing and Artificial Intelligence, 16th International Conference*, pages 28–36, Cham, 2020. Springer International Publishing.

Marian McDonnell and David Baxter. Chatbots and gender stereotyping. *Interacting with Computers*, 31(2):116–121, 2019.

Conor McGinn and Ilaria Torre. Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 211–221, 2019. doi: 10.1109/HRI.2019.8673305.

Michael McTear. Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, 13(3):1–251, 2020.

Michael Frederick McTear, Zoraida Callejas, and David Griol. *The conversational interface*, volume 6. Springer, 2016.

Glaucia Melo, Edith Law, Paulo Alencar, and Donald Cowan. Understanding user understanding: What do developers expect from a cognitive assistant? In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3165–3172, 2020. doi: 10.1109/BigData50022.2020.9378140.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 746–751, 2013.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *CoRR*, abs/2004.10645, 2020. URL `https://arxiv.org/abs/2004.10645`.

Anubhav Mishra and Anuja Shukla. Psychological determinants of consumer's usage, satisfaction, and word-of-mouth recommendations toward smart voice assistants. In Sujeet K. Sharma, Yogesh K. Dwivedi, Bhimaraya Metri, and Nripendra P. Rana, editors, *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation*, pages 274–283, Cham, 2020. Springer International Publishing. ISBN 978-3-030-64849-7.

Robert J Moore. Ethnomethodology and conversation analysis: Empirical approaches to the study of digital technology in action. *The SAGE handbook of digital technology research. Sage*, 2013.

Robert J. Moore and Raphael Arar. *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*. Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450363013.

Isabela Motta and Manuela Quaresma. Users' error recovery strategies in the interaction with voice assistants (vas). In Nancy L. Black, W. Patrick Neumann, and Ian Noy, editors, *Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021)*, pages 658–666, Cham, 2021. Springer International Publishing. ISBN 978-3-030-74614-8.

Sara Moussawi and Raquel Benbunan-Fich. The effect of voice and humour on users' perceptions of personal intelligent agents. *Behaviour & Information Technology*, 0 (0):1–24, 2020. doi: 10.1080/0144929X.2020.1772368.

Sara Moussawi, M. Koufaris, and R. Benbunan-Fich. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electronic Markets*, pages 1–22, 2020.

Nika Mozafari, Welf H Weiger, and Maik Hammerschmidt. Trust me, i'm a bot–repercussions of chatbot disclosure in different service frontline settings. *Journal of Service Management*, 2021.

Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. *Patterns for How Users Overcome Obstacles in Voice User Interfaces*, page 1–7. CHI '18. Association for Computing Machinery, New York, NY, USA, 2018. ISBN 9781450356206. doi: 10.1145/3173574.3173580.

Chelsea M. Myers, Anushay Furqan, and Jichen Zhu. *The Impact of User Characteristics and Preferences on Performance with an Unfamiliar Voice User Interface*, page 1–9. Association for Computing Machinery, New York, NY, USA, 2019a. ISBN 9781450359702. URL `https://doi.org/10.1145/3290605.3300277`.

Chelsea M. Myers, David Grethlein, Anushay Furqan, Santiago Ontañón, and Jichen Zhu. Modeling behavior patterns with an unfamiliar voice user interface. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '19, page 196–200, New York, NY, USA, 2019b. Association for Computing Machinery. ISBN 9781450360210. doi: 10.1145/3320435.3320475. URL `https://doi.org/10.1145/3320435.3320475`.

Chelsea M. Myers, Luis Fernando Laris Pardo, Ana Acosta-Ruiz, Alessandro Canossa, and Jichen Zhu. "try, try, try again:" sequence analysis of user interaction data with a voice user interface. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389983. doi: 10.1145/3469595.3469613. URL `https://doi.org/10.1145/3469595.3469613`.

Romy Müller, Dennis Paul, and Yijun Li. Reformulation of symptom descriptions in dialogue systems for fault diagnosis: How to ask for clarification? *International Journal of Human-Computer Studies*, 145:102516, 2021. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2020.102516. URL `https://www.sciencedirect.com/science/article/pii/S107158192030118X`.

Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. Exploring mental models for transparent and controllable recommender systems: A qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 183–191, 2020.

Axel-Cyrille Ngonga Ngomo and Ricardo Usbeck. An approach for ex-post-facto analysis of knowledge graph-driven chatbots–the dbpedia chatbot. In *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers*, volume 11970, page 19. Springer Nature, 2020.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *CoRR*, abs/2105.04387, 2021. URL `https://arxiv.org/abs/2105.04387`.

Donald A Norman. Some observations on mental models. *Mental models*, 7(112):7–14, 1983.

M.A. Norman and P.J. Thomas. Informing HCI design through conversation analysis. *International Journal of Man-Machine Studies*, 35:235–250, 1991.

Anastasia K Ostrowski, Vasiliki Zygouras, Hae Won Park, and Cynthia Breazeal. Small group interactions with voice-user interfaces: Exploring social embodiment, rapport, and engagement. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 322–331, 2021.

Cathy Pearl. *Designing voice user interfaces: Principles of conversational experiences*. " O'Reilly Media, Inc.", 2016.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, 2012. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2011.10.004. URL https://www.sciencedirect.com/science/article/pii/S0010027711002496.

Andrei Popescu-Belis. Dimensionality of dialogue act tagsets. *Lang. Resour. Evaluation*, 42(1):99–107, 2008. URL https://doi.org/10.1007/s10579-008-9063-y.

Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. *Voice Interfaces in Everyday Life*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2018. ISBN 9781450356206. URL https://doi.org/10.1145/3173574.3174214.

Matthew Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, 08 2004.

Matthew Purver, Julian Hough, and Christine Howes. Computational models of miscommunication phenomena. *Topics in Cognitive Science*, 10(2):425–451, 2018. doi: https://doi.org/10.1111/tops.12324. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12324.

James Pustejovsky and Nikhil Krishnaswamy. Embodied human-computer interactions through situated grounding. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–3, 2020.

N. Radziwill and Morgan C. Benton. Evaluating quality of chatbots and intelligent conversational agents. *ArXiv*, abs/1704.04579, 2017.

Antoine Raux, Dan Bohus, Brian Langner, Alan W Black, and Maxine Eskenazi. Doing research on a deployed spoken dialogue system: One year of let's go! experience. In *in Proc. INTERSPEECH, 2006*, pages 65–68, 2006.

Suman Ravuri and Andreas Stolcke. Recurrent neural network and lstm models for lexical utterance classification. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Torsten Ringberg, Gaby Odekerken-Schroder, and Glenn L. Christensen. A cultural models approach to service recovery. *Journal of Marketing*, 71(3):194–214, 2007. doi: 10.1509/jmkg.71.3.194.

Néna Roa-Seïler, Paul Craig, José Aníbal Arias, Ariadna Benítez Saucedo, Marcela Martínez Díaz, and Felipe Lara Rosano. Defining a child's conceptualization of a virtual learning companion. In *INTED2014 Proceedings. IATED*, pages 2992–2996, 2014.

David A. Robb, José Lopes, Stefano Padilla, Atanas Laskov, Francisco J. Chiyah Garcia, Xingkun Liu, Jonatan Scharff Willners, Nicolas Valeyrie, Katrin Lohan, David Lane, Pedro Patron, Yvan Petillot, Mike J. Chantler, and Helen Hastie. Exploring interaction with remote autonomous systems using conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, page 1543–1556, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450358507. doi: 10.1145/3322276.3322318. URL `https://doi.org/10.1145/3322276.3322318`.

Morteza Rohanian and Julian Hough. Best of both worlds: Making high accuracy non-incremental transformer-based disfluency detection incremental. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3693–3703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.286. URL `https://aclanthology.org/2021.acl-long.286`.

Laura Rook. Mental models: A robust definition. *The learning organization*, 2013.

Emma J Rose and Elin A Björling. Designing for engagement: using participatory design to develop a social robot to measure teen stress. In *Proceedings of the 35th ACM International Conference on the Design of Communication*, pages 1–10, 2017.

Elayne Ruane, Sinead Farrell, and Anthony Ventresque. User perception of text-based chatbot personality. In *International Workshop on Chatbot Research and Design*, pages 32–47. Springer, 2020.

Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. A simple systematic for the organisation of turn taking in conversation. *Language*, 50:696–735, 12 1974. doi: 10.2307/412243.

Emanuel Schegloff. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. *Analyzing discourse: Text and talk, Georgetown University Roundtable on Languages and Linguistics*, pages 71–93, 01 1982.

Emanuel Schegloff and Harvey Sacks. Opening up closings. *Semiotica*, 8:289–327, 01 1973. doi: 10.1515/semi.1973.8.4.289.

Emanuel Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53:361–382, 06 1977. doi: 10.2307/413107.

A.-M. Seeger, Jella Pfeiffer, and A. Heinzl. Designing anthropomorphic conversational agents: Development and empirical evaluation of a design framework. In *39th International Conference on Information Systems, ICIS 2018: Proceedings*. Association for Information Systems (AIS), 2018. ISBN 978-099668317-3.

Anna-Maria Seeger and Armin Heinzl. Chatbots often fail! can anthropomorphic design mitigate trust loss in conversational agents for customer service? In *Proceedings of ECIS 2021*, 2021. URL `https://aisel.aisnet.org/ecis2021_rp/12`.

William Seymour, Reuben Binns, Petr Slovak, Max Van Kleek, and Nigel Shadbolt. Strangers in the room: Unpacking perceptions of'smartness' and related ethical concerns in the home. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 841–854, 2020.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a conversational agent overnight with dialogue self-play, 2018.

Igor Shalyminov. Data-efficient methods for dialogue systems. *CoRR*, abs/2012.02929, 2020. URL `https://arxiv.org/abs/2012.02929`.

Ali Sharifara, Mohd Shafry Mohd Rahim, and Yasaman Anisi. A general review of human face detection including a study of neural networks and haar feature-based cascade classifier in face detection. In *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, pages 73–78. IEEE, 2014.

Chen Shi, Yuxiang Hu, Zengming Zhang, Liang Shao, and Feijun Jiang. *User Feedback and Ranking In-a-Loop: Towards Self-Adaptive Dialogue Systems*, page 2046–2050.

Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380379. URL https://doi.org/10.1145/3404835.3463079.

F Smach, M Atri, J Miteran, and M Abid. Design of a neural networks classifier for face detection. *Matrix*, 15:15, 2006.

Magnus Soderlund, Eeva-Liisa Oikarinen, and Teck Ming Tan. The happy virtual agent and its impact on the human customer in the service encounter. *Journal of Retailing and Consumer Services*, 59:102401, 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Nancy Staggers and Anthony F. Norcio. Mental models: concepts for human-computer interaction research. *International Journal of Man-machine studies*, 38(4):587–605, 1993.

Jan-Philipp Stein, Markus Appel, Alexandra Jost, and Peter Ohler. Matter over mind? how the acceptance of digital entities depends on their appearance, mental prowess, and the interaction between both. *International Journal of Human-Computer Studies*, 142:102463, 2020.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374, 2000. URL https://aclanthology.org/J00-3003.

Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. User expectations of conversational chatbots based on online reviews. In *Designing Interactive Systems Conference 2021*, DIS ’21, page 1481–1491, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384766. doi: 10.1145/3461778.3462125. URL https://doi.org/10.1145/3461778.3462125.

Akihiro Tanaka, Ai Koizumi, Hisato Imai, Saori Hiramatsu, Eriko Hiramoto, and Beatrice de Gelder. I feel your voice: Cultural differences in the multisensory perception of emotion. *Psychological Science*, 21(9):1259–1262, 2010. doi: 10.1177/0956797610380698.

Camilo Thorne. Chatbots for troubleshooting: A survey. *Language and Linguistics Compass*, 11(10):e12253, 2017.

Hapnes Toba, Zhao-Yan Ming, Mirna Adriani, and Tat-Seng Chua. Discovering high quality answers in community question answering archives using a hierarchy of classifiers. *Information Sciences*, 261:101–115, 2014.

Vishal Thanvantri Vasudevan, Abhinav Sethy, and Alireza Roshan Ghias. Towards better confidence estimation for neural models. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7335–7339, 2019. doi: 10.1109/ICASSP.2019.8683359.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Zhuoyi Wang, Saurabh Gupta, Jie Hao, Xing Fan, Dingcheng Li, Alexander Hanbo Li, and Chenlei (Edward) Guo. Contextual rephrase detection for reducing friction in dialogue system. In *Proceedings of EMNLP 2021*, 2021.

Nigel G. Ward. Individual interaction styles: Evidence from a spoken chat corpus. In *Proceedings of the SIGDIAL 2021 Conference*, 2021. URL `http://www.cs.utep.edu/nigel/papers/sigdial-indiv-styles.pdf`.

Rick Wash and Emilee Rader. Influencing mental models of security: a research agenda. In *Proceedings of the 2011 New Security Paradigms Workshop*, pages 57–66, 2011.

Patrick P. Weis and Eva Wiese. Know your cognitive environment! mental models as crucial determinant of offloading preferences. *Human Factors*, 2020. doi: 10.1177/0018720820956861.

Steve Whittaker, Yvonne Rogers, Elena Petrovskaya, and Hongbin Zhuang. Designing personas for expressive robots: Personality in the new breed of moving, speaking, and colorful social home robots. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(1):1–25, 2021.

Shuai Zhou. *Using Hassenzahl Model as a Design Method to Improve User Experience for Health Care Information Television App*. PhD thesis, University of Cincinnati, 2015. URL `http://rave.ohiolink.edu/etdc/view?acc_num=ucin1453801729`.