

Time-Buying in Task-Oriented Spoken Dialogue Systems

María Soledad López Gambino



Dissertation eingereicht
bei der Humanwissenschaftlichen Fakultät
der Universität Potsdam

2022

This work is protected by copyright and/or related rights. You are free to use this work in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s).
<https://rightsstatements.org/page/InC/1.0/?language=en>

First supervisor and referee: Prof. Dr. David Schlangen
Second supervisor: Prof. Dr. Manfred Stede
Second referee: Prof. Dr. Petra Wagner

Date of final exam: 12-12-2022

Published online on the
Publication Server of the University of Potsdam:
<https://doi.org/10.25932/publishup-59280>
<https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-592806>

Declaration

I declare that this thesis was written by myself and that the work contained therein is my own, except in those cases where it is explicitly stated otherwise in the text. This work has only been submitted to Potsdam University and has not been submitted to another degree or professional qualification.

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which was funded by the German Research Foundation (DFG).

María Soledad López Gambino

To my parents

Abstract

This dissertation focuses on the handling of time in dialogue. Specifically, it investigates how humans bridge time, or “buy time”, when they are expected to convey information that is not yet available to them (e.g. a travel agent searching for a flight in a long list while the customer is on the line, waiting). It also explores the feasibility of modeling such time-bridging behavior in spoken dialogue systems, and it examines how endowing such systems with more human-like time-bridging capabilities may affect humans’ perception of them.

The relevance of time-bridging in human-human dialogue seems to stem largely from a need to avoid lengthy pauses, as these may cause both confusion and discomfort among the participants of a conversation (Levinson, 1983; Lundholm Fors, 2015). However, this avoidance of prolonged silence is at odds with the *incremental* nature of speech production in dialogue (Schlangen and Skantze, 2011): Speakers often start to verbalize their contribution before it is fully formulated, and sometimes even before they possess the information they need to provide, which may result in them running out of content mid-turn.

In this work, we elicit conversational data from humans, to learn how they avoid being silent while they search for information to convey to their interlocutor. We identify commonalities in the types of resources employed by different speakers, and we propose a classification scheme. We explore ways of modeling human time-buying behavior computationally, and we evaluate the effect on human listeners of embedding this behavior in a spoken dialogue system.

Our results suggest that a system using conversational speech to bridge time while searching for information to convey (as humans do) can provide a better experience in several respects than one which remains silent for a long period of time. However, not all speech serves this purpose equally: Our experiments also show that a system whose time-buying behavior is more varied (i.e. which exploits several categories from the classification scheme we developed and samples them based on information from human data) can prevent overestimation of waiting time when compared, for example, with a system that repeatedly asks the interlocutor to wait (even if these requests for waiting are phrased differently each time). Finally, this research shows that it is possible to model human time-buying behavior on a relatively small corpus, and that a system using such a model can be preferred by participants over one employing a sim-

pler strategy, such as randomly choosing utterances to produce during the wait—even when the utterances used by both strategies are the same.

Acknowledgements

First and foremost, I'd like to thank my advisor, Prof. Dr. David Schlangen, for all these years of support and patience. Thank you so much, David, for trusting me right from the start. For your guidance, for instilling a spirit of collaboration and curiosity in the Bielefeld Dialogue Systems Group, and for sharing your vast knowledge with us. Thanks for continuing to support me during and after the transfer to Potsdam University, and for always finding some time for my work in spite of your very busy schedule.

Secondly, I'd like to thank Potsdam University for admitting me halfway through my PhD, and especially Prof. Dr. Manfred Stede for agreeing to be my second advisor.

Thanks to Bielefeld University for hosting my work for four years, and to the Cognitive Interaction Technology Excellence Cluster (CITEC) and the DSG for providing me with the funding to carry out my research. Thanks to Claudia Muhl and all the CITEC staff for their help navigating the formalities of doing a PhD, and for the academic retreats and colloquia.

My gratitude also goes to my colleagues and friends Mariano Felice, Nazia Attari and Renaud Mousnier-Lompere for reviewing a part of my dissertation. Thank you so much for your time and your insightful comments. You rock! :)

During my time in the Bielefeld Dialogue Systems Group, I had the pleasure to work with some really inspiring and talented colleagues. Very special thanks go to Sina Zarriß and Casey Kennington, wonderful co-authors and mentors during this journey, from whom I learned so much. Also thanks to Simon Betz, Julian Hough, Ting Han, Spyros Kousidis, Nazia Attari (again :)), Nikolai Ilinykh, Ramesh Radhakrishna, Iwan de Kok, Birte Carlmeyer and Matthias Priesters for shared discussions, seminars, journal clubs, lunches, dinners and world cup matches. Thanks to all the fantastic student assistants who worked in the DSG at some point, especially those who helped with my studies: Oliver Eickmeyer, Michael Bartholdt, Gerdis Anderson, Angelika Maier, Ayten Tüfekçi, and probably others that I'm forgetting now (apologies!). And of course, thanks to all the participants who made the experiments presented in this work possible, and to Sarah Schopper, Patryk Wainaina, Jonas Rohnke, Yvonne Flory and Dan McCarthy for participating in the dry-run for the final experiment.

Last but not least, thanks to my parents, Jorge López and Carolina Gambino, to whom I owe everything I am and have ever achieved, and to my partner, Giovanni

Carrabs, for his solid trust and support to pursue my dreams, his relentless encouragement in the face of adversity, and for every vacation and trip he gave up because I was too busy writing the pages which follow.

Relevant Publications

Portions of this thesis are based on previously published material, specifically on the following publications:

- Betz, S., & López Gambino, M.S. (2016). Are We All Disfluent in Our Own Special Way and Should Dialogue Systems Also Be? In O. Jokisch (Ed.), *Studientexte zur Sprachkommunikation, 81. Elektronische Sprachsignalverarbeitung (ESSV) 2016*, pp. 168–174. TUD Press. <http://www.essv.de/paper.php?id=337>.
- López Gambino, M.S., Zarriß S., Kennington, C., & Schlangen D. (2018). Learning to Buy Time: A Data-Driven Model for Avoiding Silence While Task-Related Information Cannot Yet Be Presented. In L. Prévot, M. Ochs, & B. Favre (Eds.), *Proceedings of 22nd Workshop on the Semantics and Pragmatics of Dialogue (Semdial 2018)*. Aix-en-Provence, France. <http://semdial.org/anthology/papers/Z/Z18/Z18-3018/>.
- López Gambino, M.S., Zarriß, S., & Schlangen, D. (2017). Beyond On-Hold Messages: Conversational Time-Buying in Task-Oriented Dialogue. In K. Jokinen, M. Stede, D. DeVault, & A. Louis (Eds.), *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (Sigdial)*, pp. 241–246. Saarbrücken, Germany. <https://aclanthology.org/W17-5529/>.
- López Gambino M.S., Zarriß S., & Schlangen D. (2019) Testing Strategies for Bridging Time-To-Content in Spoken Dialogue Systems. In: L. D’Haro, R. Banchs, H. Li (Eds.): *9th International Workshop on Spoken Dialogue System Technology. Lecture Notes in Electrical Engineering, 579*, pp. 103–109. Springer. https://doi.org/10.1007/978-981-13-9443-0_9.

Other publications by the author are:

- Kennington, C., López Gambino, M.S., & Schlangen, D. (2015). Real-World Reference Game Using the Words-As-Classifiers Model of Reference Resolution. *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue (Semdial 2015)*, pp. 188–189. Gothenburg, Sweden. http://semdial.org/anthology/Z15-Kennington_semdial_0036.pdf.

- López Gambino, M.S., Kennington, C., & Schlangen, D. (2017). Silence, Please! Interrupting In-Car Phone Conversations. In A. Cafaro, E. Coutinho, P. Gebhard, & B. Potard (Eds.), *Proceedings of the First Workshop on Conversational Interruptions in Human-Agent Interactions (CIHAI 2017)*, pp. 9–18. Stockholm, Sweden. <http://ceur-ws.org/Vol-1943/WCIHAI-17-02.pdf>.

Contents

1	Introduction	16
1.1	Motivation	16
1.2	Aims	18
1.3	Structure	20
1.4	Contributions	21
2	Background	23
2.1	Situated communication	23
2.2	Silence in dialogue	25
2.3	Fluency, “disfluency” and grounding	26
2.4	Types of disfluencies	28
2.5	Dialogue acts	30
2.6	Types of dialogue systems	31
2.7	Incrementality in language generation	34
2.8	Dialogue systems and human-likeness	36
2.9	Evaluation of dialogue systems	37
2.10	Summary	38
3	Eliciting Time-Buying in Human-Human Dialogue	39
3.1	Introduction	39
3.2	Data collection	40
3.2.1	Setup and participants	40
3.2.2	Task	41
3.2.3	Information display	43
3.3	Data	46

<i>CONTENTS</i>	13
3.3.1 Segmentation of the phenomenon	49
3.3.2 Annotation	51
3.4 Analysis	51
3.4.1 Utterances	52
3.4.2 Pauses	57
3.5 Discussion	59
3.6 Summary	60
4 Experiment 1: Speech, Silence and Time-Buying	62
4.1 Introduction	62
4.2 The waiting experience	64
4.3 Experiment	65
4.3.1 Method	65
4.3.2 Results	70
4.4 Discussion	74
4.5 Summary	77
5 Experiment 2: Time-Buying in Spoken Dialogue Systems	78
5.1 Introduction	78
5.2 The system	80
5.2.1 Decision-making	80
5.2.2 Architecture	85
5.3 Experiment	88
5.3.1 Method	89
5.3.2 Results	93
5.4 Discussion	95
5.5 Summary	97
6 Experiment 3: Testing Strategies for Modeling Time-Buying	99
6.1 Introduction	99
6.2 Strategies	99
6.2.1 Random strategy	100
6.2.2 Flat strategy	100
6.2.3 Hierarchical strategy	101
6.3 Intrinsic evaluation: Perplexity	104

6.4	Extrinsic evaluation: Experiment	106
6.4.1	Method	106
6.4.2	Results	111
6.5	Discussion	115
6.5.1	Observations	115
6.5.2	Comparison with Chapter 5	117
6.5.3	Related work	118
6.5.4	Further work	119
6.6	Summary	119
7	Conclusions	121
7.1	Introduction	121
7.2	Overview of findings	121
7.3	Lessons learned	123
7.4	Further work	124
	Appendices	126
A	Data collection (Chapter 3)	127
A.1	Instructions for participants	127
B	Experiment 1 (Chapter 4)	129
B.1	Survey	129
C	Experiment 2 (Chapter 5)	133
C.1	Experiment instructions	133
D	Experiment 3 (Chapter 6)	135
D.1	Dialogues for each strategy	135
	References	137

Chapter 1

Introduction

1.1 Motivation

Situated communication takes place in a certain space and time. The fact that spoken dialogue occurs in a specific time setting which is normally common to all participants results in certain constraints, which largely determine how speech develops (Rickheit and Wachsmuth, 2006; Clark, 2002, 1996). One of these constraints is related to the need for an appropriate balance between speech and silence. Such balance is central to communication, and failure at this level can interfere with the normal flow of the interaction in different ways. Pauses which are perceived as excessively long may result in misunderstandings: They may lead the interlocutor to believe that there are problems in the communication channel, or with the content of the dialogue (Levinson, 1983), or they can be interpreted as signaling a non-cooperative attitude, for example if one speaker asks for a favor and the other one takes too long to answer (Roberts and Francis, 2013; Kohtz and Niebuhr, 2017). Moreover, lengthy pauses can be mistaken as a signal that the speaker is relinquishing the turn, and that the latter is available for other speakers to take (Sacks et al., 1974), which could result in overlaps.

What happens, then, when a speaker “has nothing to say”? Or more specifically: What happens when a speaker has nothing to say *yet*, but knows that they soon will? It is possible to imagine situations in which one of the speakers in a dialogue needs to delay a contribution that the rest of the participants are expecting. One such case is when the contribution depends on external information not available at the moment, for instance when one of the speakers asks for a third person’s phone number and the

other speaker needs to look it up within a list of contacts. Another case may be when cognitive demands, such as word-retrieval problems or other utterance-planning considerations, prevent the speaker from producing the desired utterance right away. Fortunately, humans are generally able to handle most of these situations gracefully, keeping the silence-speech balance unharmed. We achieve this by intermingling pauses with utterances which do not convey specific information about the main topic of the dialogue but instead communicate, more or less directly, the need for additional time. As an example, consider the following phone conversation between a travel agent and a customer:¹

- | | | |
|----------|---|-----|
| CUSTOMER | I'm looking for a flight to Bucharest in the first week of August... | (1) |
| | with Lufthansa, if possible. | (2) |
| AGENT | A flight to Bucharest... beginning of August... | (3) |
| | Please wait a second, the flights are still being loaded... | (4) |
| | Hmm... | (5) |
| | I'm looking into my list... | (6) |
| | There's a Lufthansa flight departing from Stuttgart on 3 August, | (7) |
| | departure time is 9:35, arrival time is 12:20... | (8) |

Until the travel agent announces the departure airport and date on line 7, none of the utterances provide information related to specific flights, which is the topic of the interaction. Instead, the speaker buys time in different ways: by uttering a filler (line 5), openly requesting extra time (line 4), echoing part of the customer's request (line 3) or explaining the reasons that prevent them from giving flight information, such as the behavior of the system (line 4) or the activity in which they are engaged at the moment (line 6). This distinction corresponds to Clark's two communicative tracks: a primary track which deals with the "official business" of the interaction and a collateral track which aims at creating conditions for successful communication (Clark, 1996). In our example, the time-buying utterances mentioned would correspond to the latter, since they are used to make the state of the interaction more transparent: They signal the need for extra time before presenting a result, thus preserving the communicative exchange in the face of possible misunderstandings. On the other hand, the last two lines of the example belong to the primary track, since they convey information about

¹This is a constructed example, for illustration purposes.

the subject matter of the conversation, namely flights.

In contrast to the smoothness with which humans handle these situations, most current dialogue systems do not (to the best of my knowledge) incorporate a comparable functionality. Systems such as virtual personal assistants in mobile phones are often not equipped with provisions to deal with delays in a conversational manner. Automatic systems for telephone applications, on the other hand, normally do employ filler strategies to cover up waiting periods, given that these can be rather lengthy. A traditional approach has been to play music, sometimes also interspersed with explicit requests to wait (such as *Please hold on; your call is important to us.*) or with information about the corresponding company and its products (Tom et al., 1997; Munichor and Rafaeli, 2007; Antonides et al., 2002). It is evident that none of these strategies are really conversational, nor do they reflect the variety and flexibility observed in human time-buying behavior.

A first question which arises from the considerations above is: Would it be possible to develop a dialogue system which buys time in a smooth and flexible way, exploiting a variety of resources and maintaining an appropriate silence-speech balance, in a similar way to that of humans? A second question is: How useful would this be? Would users feel more at ease when interacting with a system that can communicate the need for extra time in a more human-like way? These two questions constitute the main motivation for the present work.

1.2 Aims

As machines enhance their interactive capabilities and thus expectations about naturalness rise, it will become increasingly more important that systems project their internal processing state in a way that is understandable to users. A dialogue system which can communicate the need for more time conversationally while it looks for information would have the effect of increasing common ground regarding the state of the task, making it less likely for the interlocutor to attempt to take the turn or to terminate the interaction. Nevertheless, to the best of my knowledge, there have not been any systematic studies of the strategies used by humans in order to buy time. Therefore, the first step in developing such a system should be observation and analysis of this phenomenon in human-human conversation. With the insights derived from this analysis, it would be possible to develop a dialogue system which exhibits similar behavior. Fi-

nally, this system would need to be tested with humans, and their perception of it must be compared to that of more standard systems.

Therefore, the aims of this work are:

1. to identify, analyze and describe the strategies that human speakers employ in order to buy time in a task-oriented setting,
2. to model these strategies computationally and
3. to explore how applying these strategies in a system impacts users' experience.

During this process, the following questions (among others) shall be answered:

- Human speakers are sometimes expected to talk while they still lack enough information to convey. What do they do while they search for this information, or while they plan an utterance to convey it?
- Do speakers frequently produce successions of resources of the same type, or do they try to provide variety?
- Do different types of time-buying resources combine in predictable patterns? Are there types of time-buying resources which seem to co-occur particularly often?
- Once the customer has uttered a request, how long does the travel agent take to start speaking? Does there seem to be a general "maximum tolerable silence" before starting to buy time (or is this highly speaker-specific)?
- How long are the silences between time-buying resources?
- How would listeners experience a system which buys time in a similar way to humans? Would they find it more human-like? Or on the contrary, would they perceive it as too artificial, given that they do not expect this kind of behavior from a system?
- Would humans be more willing to interact with this system than with one that cannot buy time, or that fills up waiting time differently (e.g. by explicitly asking the user to wait)?

- How does the time-buying strategy used by a system affect humans' perception of waiting time? More specifically: If a system buys time in a natural, conversational way, will humans perceive the wait as shorter?
- Which elements of human time-buying contribute the most towards achieving a satisfactory time-buying strategy for a spoken dialogue system?

1.3 Structure

This work is organized as follows:

Chapter 2 provides an overview of previous work on related topics. After introducing the concept of situated communication, we discuss some basic considerations connected to silence and its role in human-human conversation. This is followed by a section on disfluencies such as fillers (*uh*, *uhm*, *mm*, *etc.*). Afterwards we present the concept of task-oriented dialogue system and we discuss the importance of human-likeness, an issue which has been and remains controversial. The chapter ends by describing different approaches to dialogue system evaluation.

Chapter 3 describes the process of collecting dialogue data from people buying time. We present the method that we used in order to elicit this phenomenon and the characteristics of the resulting corpus. Afterwards, we discuss the challenges involved in pre-processing and analyzing the data, such as segmentation of the phenomenon, which was not as straightforward as originally expected, given that speakers would often transition smoothly from the time-buying stage to the information presentation stage without a clear separation. We also describe the taxonomy of time-buying utterances that we developed in order to classify the instances found in the corpus. Finally, we present and discuss the results of our analysis, which help answer some of the questions enumerated above with respect to time-buying in human-human dialogue.

Chapter 4 starts with some considerations related to silence in the specific context of human-machine interaction, including a brief discussion of the waiting experience and the factors which influence the perception of elapsed time in humans. This is followed by a small study in which two different time-buying strategies were evaluated by human listeners. We discuss the ratings obtained and suggest possible reasons for one strategy being preferred over the other, as well as the interplay between the preferred

strategy and the quality of the voice used for the system.

In **Chapter 5** we experiment with different strategies to bridge time, and we ask humans to interact with the system and evaluate each strategy. The strategies are: a) repeatedly asking the user to wait (using different utterances), b) randomly selecting time-buying actions from the taxonomy proposed in Chapter 3, and c) sampling time-buying actions from this taxonomy based on their frequency in the corresponding stage of the wait (right after the interlocutor's request, after one time-buying action, after two time-buying actions, etc). We find that the third strategy is perceived as more human-like and pleasant to interact with than the other two. It is also perceived as capable of finding a result in a more appropriate amount of time than the strategy which only asks the user to wait.

In **Chapter 6**, we attempt to model humans' sequencing of time-buying actions using the data presented in Chapter 3. We compare two models based on trigrams against a strategy which selects time-buyers randomly. Both trained models reduce the perplexity of the random one by 50% when testing on unseen data. Moreover, human listeners reported a preference for one of the trained models over the random one for their own use. However, we found no difference in estimation of waiting time across conditions. Coupled with the data from Chapters 4 and 5, this suggests that exploiting a variety of dialogue acts to bridge time contributes to avoiding overestimation of waiting time, regardless of how these dialogue acts are sequenced.

Finally, **Chapter 7** presents the conclusions of this research, as well as questions and directions which remain open for the future.

1.4 Contributions

This work expands the existing knowledge in the fields of dialogue and dialogue systems, and the findings here presented can help developers build more natural conversational systems, mitigating the negative effect of making users wait for information. More concretely, this dissertation represents a contribution to the scientific community for the following reasons:

- It increases the existing knowledge about human dialogue, in particular with respect to the speech resources that humans use in order to bridge time until they can provide task information. It provides a taxonomy under which these

resources can be classified, and information about the frequency in which each category was used and how categories were sequenced in a role-playing dialogue task. It shows, for example, that explicit requests for extra time, such as *please wait*, were relatively uncommon, and that speakers normally preferred to signal the need for extra time in subtler ways, such as echoing words from the interlocutor's previous turn, or producing fillers (*uh, uhm, äh*) (see Chapter 3).

- It expands the existing knowledge about people's preferences when waiting for information. It reconfirms the preexisting claim that humans tend to perceive unfilled silent periods of time as longer (Hirsch et al., 1956; Tom et al., 1997) (see Chapter 4).
- It increases knowledge about human-machine dialogue, in connection to resources that automatic systems can use to buy time before they can provide task information and humans' preferences regarding these resources. It suggests that humans prefer a system which produces a variety of time-buying resource types over one which produces different utterances with similar meanings. In addition, it highlights the impact of voice quality on the strength of users' preference for different time-buying strategies (see Chapters 4 and 5).
- Finally, it shows that human time-buying behavior can be modeled computationally using a relatively small corpus, and that this model can enable a dialogue system to perform more satisfactorily—in terms of users' preferences—than a system which employs more traditional approaches (see Chapters 5 and 6).

Chapter 2

Background

2.1 Situated communication

Communication between humans is situated. As Jurafsky and Martin (2018, p. 21) explain:

Words don't appear out of nowhere. Any particular piece of text that we study is produced by one or more specific speakers or writers, in a specific dialect of a specific language, at a specific time, in a specific place, for a specific function.

Researchers have studied the relationship between communicative interactions and the physical world for years. Speakers usually *refer* to the objects surrounding them in their dialogues (Clark and Wilkes-Gibbs, 1986; Clark, 1996; Kennington, 2016). Therefore, shared knowledge of the physical environment often contributes to successful communication. In addition, conversations can also be influenced by outside events in unforeseen ways and force speakers to adjust their interactions. As an example, if a conversation is interrupted by a sudden noise, the active speaker might stop and resume speech once the noise disappears (Buschmeier et al., 2012; Villing, 2015). A change in the state of the environment might also render the speaker's statement not valid anymore and make it necessary to modify it (Raux and Nakano, 2010).

Another dimension of this connection is that of embodiment (Casell et al., 2000; Kopp and Wachsmuth, 2009). When speakers are co-located, they interact not only through speech but also through other channels. Gestures are widely used during spoken dialogue and they can contribute to its meaning in various ways (McNeill, 1992;

Kendon, 2004). Gaze also plays a key role in conveying meaning, as well as in turn-taking (Staudte and Crocker, 2018; Amati and Brennan, 2018; Kousidis and Schlangen, 2015; Sekicki and Staudte, 2018).

Space is not the only contextual element with an influence on dialogue. The fact that speech occurs *in time* is also crucial. Clark (1996) highlights the importance of time in dialogue in his *temporal imperative* (p. 267):

In a joint action, the participants must provide a public account for the passage of time in their individual parts of that action.

The author claims that as long as the speakers keep talking, they are providing a satisfactory account for the passage of time during their turn. In contrast, when speakers produce a lengthy pause, they no longer have a public justification for their actions, therefore they are *breaking the temporal imperative*. In this respect, speakers remain under the pressure of the temporal imperative throughout the duration of the dialogue.

This requirement, however, is at odds with the fact that *speakers often do not have their contributions completely planned before they start speaking*. Instead, they build their utterances *incrementally* (i.e. in sub-utterance chunks) while they talk. This frequently results in speakers temporarily “running out of words” halfway through an utterance and needing extra time to plan the next steps. Clark hints at this problem in his *formulation imperative*:

Speakers cannot present an expression before they have formulated it.
(Clark, 1996, p. 267)

Speakers are torn between the formulation imperative and the temporal imperative whenever they hold the conversational floor. Therefore, in order to avoid breaches, they make extensive use of certain speech devices known as *disfluencies*. The term “disfluency” encompasses phenomena such as repetitions, fillers and lengthening, among others (Shriberg, 1994; Clark, 2002; Ginzburg et al., 2014; Hough, 2015; Lickley, 2015). These phenomena allow the active speaker to continue the delivery while signaling to the interlocutor that they are in “planning mode” and intend to retain the conversational floor. We deal with some aspects of disfluencies in Section 2.3. The next section elaborates further on the topic of silence in conversations.

2.2 Silence in dialogue

A certain amount of silence is often inevitable—and even desirable—in speech. There are various reasons why speakers produce silence in conversations, and not all instances of silence constitute pauses. Silence is a part of some speech units (phonemes), such as English voiceless plosives /p/, /t/ and /k/ during their closure stage (Cruttenden, 2001). Silence can also stem from physiological factors, such as the speaker's need to breathe or to bring the speech organs into the right position in order to articulate a sound (Lickley, 2015). Cognitive factors may also play a role, for example when the speaker pauses in order to plan how to continue an utterance (Zellner, 1994).

The silence that occurs within a speech sound, or while breathing, is typically very short—sometimes imperceptible—so we do not normally regard this kind of phenomenon as a pause. For other kinds of silence in dialogue, Sacks et al. (1974) distinguish between three categories: *pauses*, *gaps* and *lapses*. Pauses are silences that occur within the turn of one speaker. A pause which occurs after a speaker's turn is a gap, unless the speaker has explicitly nominated another speaker for the next turn, in which case it counts as a pause within the turn of the next speaker. Finally, if a gap gets extended, it becomes a lapse, and discontinued talk arises (Sacks et al., 1974, p. 715).

As stated in Section 2.1, Clark's (1996) *temporal imperative* states that participants in a dialogue are accountable for the time elapsed during their turn. Thus, whenever speakers remain silent for too long, they are not properly accounting for their use of dialogue time. On the other hand, Levinson (1983) suggests that pauses which are too long, or which appear in unexpected contexts, gain further significance. Therefore, the listener might interpret them as a sign of trouble understanding, or as preceding the introduction of a problematic topic into the conversation (Lundholm Fors, 2015).

If a pause can be “too long”, what is appropriate pause length? This seems to be determined by both individual and cultural factors (Lundholm Fors, 2015). Stivers et al. (2009) analyzed pauses in 10 languages and found avoidance of overlaps and minimization of silence between turns to be common across all of them. However, they also found that absolute pause durations differed between languages, which suggests that there might not be a universal concept of a pause that is “normal”. On the other hand, the linguistic context of a pause may also determine whether its length is perceived as appropriate. As an example, pauses which are placed at syntactic or prosodic boundaries are normally licensed longer durations than those placed inside a syntactic

phrase or an intonational unit (Lickley, 2015; Moniz et al., 2010).

Although there is no straightforward answer to the question of when a pause becomes too long, certain trends seem to exist. Jefferson (1983, 1989) claims that most pauses in dialogue are shorter than 1200 ms. The author analyzed pause duration in a corpus of conversations and found a large number of pauses within the 900-1200 ms. duration interval, followed by a sharp drop in the next interval (1300-1800 ms.). Furthermore, the ratio between the number of 900-1200 ms. pauses and the number of all longer pauses in the same corpus was 3 to 1 (see Table 2.1). On the other hand, Campione and Véronis (2002) analyzed corpora in five languages containing a total of 6000 pauses and found a trimodal distribution of short (200 ms. or less), medium (200 to 1000 ms.) and long (more than 1000 ms.) pauses. In addition, Kohtz and Niebuhr (2017) found that, when someone asks for a favor, an answer coming after a pause which is longer than 600 ms. can be interpreted as indicating lack of willingness, even if the answer is affirmative. The authors ran a study in German in order to replicate Roberts and Francis's (2013) previous results for English and found that, when testers listened to the following dialogue:

Requester: *Kannst Du mich nachher zur Uni fahren?*

(Can you take me to the university later?)

Interlocutor: *Ja, natürlich.*

(Yes, of course.)

there was a significant decrease in perceived willingness when the pause between both turns was longer than 600 ms., in spite of the affirmative nature of the answer. All these considerations suggest that, although accepted pause duration is subject to individual variation, some patterns can be detected across speakers of the same language, or even across languages.

2.3 Fluency, “disfluency” and grounding

Traditionally, the yardstick used to make judgments regarding fluency has been a theoretical construct known as *ideal delivery of information*, which assumes that speech is more fluent when it contains as few pauses as possible. According to this view, information ought to be delivered at a constant pace, avoiding stops and interruptions.

Duration in ms.	Number of pauses
900-1200	951
1300-1800	92
1900-2200	92
2300-2800	72
2900-3200	32
3300-3800	17
3900-4200	8
4300-4800	5
4900-5200	5
>5200	6

Table 2.1: Numbers of pauses in a corpus of conversational speech, clustered by duration (Jefferson, 1983).

Moreover, utterances should be semantically dense and exclude “empty material” such as hedges, or hesitations such as *uh...* (Linell, 2004). From this perspective, speaking fluently could be equated to “filling time with talk”, in a similar way to that of a radio speaker or sports commentator (Fillmore, 1979). This notion of fluency has been regarded as universally valid—meaning that it was applied in the same way to all situational contexts—and objective, given that little or no attention was paid to individual, social and cultural factors.

In more recent years, however, this conception has been revised and alternative notions have been proposed. More emphasis has been placed on the subjectivity of fluency, its individual and socio-cultural dimensions (Lundholm Fors, 2015). The idea of context-dependence has also been highlighted, in order to point out that what is considered fluent in some situations can be considered non-fluent in others and vice-versa (Lennon, 2003). More importantly, the notion of ideal delivery has been repeatedly challenged. It is now widely accepted that speech can include pauses of various lengths and yet be perceived as continuous, so long as such pauses are produced under appropriate circumstances (Moniz et al., 2010). Similarly, repetitions, repairs and fillers are accepted as natural components of spontaneous speech, useful for creating a perception of “even flow”.

Clark’s two-track model of dialogue is useful to conceptualize the distinction be-

Speaker	Track 1	Track 2	
Roger	now,-urn do you and your husband have a j-car		(1)
Nina		-have a car?	(2)
Roger		Yeah.	(3)
Nina	No.		(4)

Table 2.2: Dialogue illustrating primary and secondary communicative tracks, from Clark (1996), p. 242.

tween utterances such as *uh...* or *I mean...* and utterances which are richer in content (Clark, 1996). The author postulates the existence of a *primary track of communicative acts*, on which speakers deal with the main business of the conversation, and a *secondary track of metacommunicative acts*, which is used for managing the interaction, solving any potential problems in the communicative channel, tracking the progress of the conversation, etc. He illustrates this idea through an exchange between two speakers, Roger and Nina, shown in Table 2.2. Roger’s question on line 1 is aimed at obtaining information. However, on line 2, Nina seeks to ensure that she is in possession of the right information before the conversation about the topic from line 1 can continue. Once Roger provides confirmation, both speakers can go back to the main topic of the exchange, since they have completed *grounding*. According to Clark, *grounding* is the process through which the participants in a conversation establish the information that has been conveyed as part of their shared knowledge or common ground (Clark and Brennan, 1991; Clark, 1996). For this purpose, Nina needs to ensure that she has understood Roger. Other grounding actions are *acknowledgments*, such as *mm-hm* or *I see*, which can be used to mean “I understand what you’re saying” (Clark, 1996, p. 231). In Clark’s words (1996, p. 242), “talk about talk is still talk”.

2.4 Types of disfluencies

Speech phenomena such as repairs, repetitions, filled pauses and reformulations have been studied extensively in the last decades. Shriberg (1994) (based on Levelt (1983)) formalizes the structure of disfluencies as displayed in Figure 2.1. Disfluencies can be seen as made up of:

- A *reparandum*, which contains the mistake that needs to be repaired

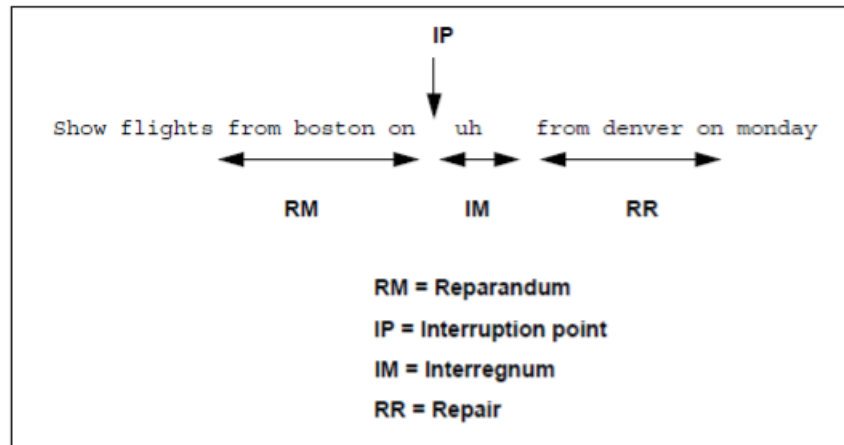


Figure 2.1: Structure of disfluencies, as represented by (Shriberg, 1994, p. 8)

Backward-looking	false start repairs	<i>It's fir- I could get it where I work.</i> <i>...any health cov- any health insurance.</i>
Forward-looking	fillers repetitions	<i>um, we're fine.</i> <i>...have the, the tools.</i>

Table 2.3: Examples of backward-looking and forward-looking disfluencies, from Shriberg (1999)

- An *interruption point*
- An optional *interregnum* or editing phrase
- The *repair* to replace the undesired content from the reparandum

Ginzburg et al. (2014) establish a distinction between *forward-looking* and *backward-looking* disfluencies (see Table 2.3). Forward-looking disfluencies are those which arise when the speaker is having problems planning the upcoming contribution. Some examples are **fillers**, (*uh...*, *um...*), **repetitions** (*I mean the- the red one*) and **lengthening** (*I mean theeeee... red one*) (Betz and Wagner, 2016). The other kind, *backward-looking disfluencies*, are those phenomena which relate to the previous contribution and are used to amend the content presented in it. Examples are **reformulations** and **repairs**. The former type, forward-looking disfluencies, is the most relevant for the discussion of “time-buying” which follows in the next chapters.

2.5 Dialogue acts

Modeling human dialogue is challenging. One of the main difficulties lies in the need to understand the kind of contributions that specific utterances represent in the interaction, in relation to the common ground shared by the participants and to their individual intentions and beliefs. For this purpose, several annotation frameworks have been developed throughout the years. One of the most widely adopted schemes for general purpose annotation of dialogue acts has been DAMSL, *Dialogue Act Markup in Several Layers* (Core and Allen, 1997). This scheme builds up on Searle's (1969) classification of Speech Acts, but it also allows for multi-functionality of utterances. DAMSL classifies utterances into three broad categories:

- **Forward communicative function:** utterances which affect a later portion of the dialogue. Included in this group are statements, information requests, etc.
- **Backward communicative function:** utterances which relate to a previous portion of the dialogue, such as an answer to a question. This category includes utterances to signal understanding, agreement or disagreement, among others.
- **Utterance features:** used for specifying other aspects of the utterance content or form, such as whether it deals with the actual task or with the communication process (in this respect, utterance features are related to Clark's (1996) communication tracks, mentioned in Section 2.3).

Within these broader categories are further classification levels, which allow tagging of utterances with more specific information. Some later annotation schemes have taken elements from DAMSL and sometimes merged them with other schemes. Such is the case of DIT++, which combined the DIT scheme for information dialogues with DAMSL (Bunt, 2009). DIT++ eventually resulted in the ISO Dialogue Act Annotation Standard 24617-2. This standard—which also specifies a markup language, DiAML—allows for annotation of the semantic content of the utterance (the *dimension* label) as well as addition of qualifiers to specify uncertainty, sentiment, or rhetorical relations between different dialogue acts in the interaction (Bunt et al., 2010, 2017).

2.6 Types of dialogue systems

A dialogue system is an artificial agent which can interact verbally with humans or other artificial agents. Dialogue systems can be classified into different types depending on their purpose, their architecture, the channels humans use to interact with them, etc. Below are the main categories into which dialogue systems are normally grouped.

Rule-based vs. data-driven

Early dialogue systems were based on handwritten rules. The first widely known example of a dialogue system was ELIZA, a simple rule-based chatbot which represented a Rogerian psychiatrist able to converse with patients through writing as early as 1966 (Weizenbaum, 1966). ELIZA employs *decomposition rules* and *reassembly rules*. The first set of rules is used to obtain some level of parsing of the user's input text, identify key words and minimal context. Afterwards, reassembly rules help the system re-use components from user input and effect certain transformations to generate an appropriate answer. In an example dialogue provided by the author, a user writes *My father is afraid of everybody*. On receiving this input, the system applies a rule which detects the pattern "X is Y", and another rule which reassembles it into the form "What else comes to your mind when you think of X?", resulting in the response *What else comes to mind when you think of your father?* Within such an approach, no understanding of the meaning of X is necessary for the system to be able to provide an answer.

Although this and other rule-based approaches can be effective for domains which, like Rogerian psychiatrist-patient interactions, follow a more or less predictable pattern, attempting to devise rules covering every possible conversation in any domain seems less than practical. For this reason, later conversational agents have increasingly relied on data in order to model dialogue. These data-driven approaches leverage large dialogue corpora for the system to learn how to select the best action for each turn (Lemon and Pietquin, 2012; Serban et al., 2018). Some of these systems compute semantic similarity between the user's input and utterances in the corpus, and return the most similar utterance they can find in the latter (Ritter et al., 2011; Banchs and Li, 2012). Other systems use an encoder-decoder architecture which allows them to produce the most suitable reply given the user's input (Lowe et al., 2017; Serban et al., 2017).

Conversational vs. task-based

Based on their purpose, dialogue systems can be classified into two large groups: those whose aim is to entertain the user, and those aimed at assisting the user with a specific task—such as booking tickets for a concert, helping fix a technical issue, etc. The former are often identified as *conversational dialogue systems* or *chatbots*, whereas the latter are referred to as *task-oriented dialogue systems* (Jurafsky and Martin, 2018). In the past decades, numerous conversational systems have emerged, some of which are able to participate in dialogues on a wide variety of topics. Examples are Cleverbot, Eugene Goostman, Elbot, JFred and UltraHal (Shah et al., 2016).¹ Well-known examples of task-oriented dialogue systems are the commercial digital assistants Apple Siri, Amazon Alexa and Google Home, which provide information about topics such as weather and news, manage calendar entries, play music on request, etc. (Bellegarda, 2013; Ehrenbrink et al., 2017; Awadallah et al., 2018; Lopatovska et al., 2019).

Text-based vs. spoken vs. multimodal

It is also possible to classify dialogue systems according to the channel used to interact with them. Some dialogue systems are purely text-based, some work through voice interaction (*spoken dialogue systems*, or *SDS*), and some support both channels. Other systems, sometimes referred to as *Multimodal Dialogue Systems*, also exploit further modalities, such as gestures and gaze (Wahlster, 2003; Matsuyama et al., 2016; Mitev et al., 2018).

End-to-end vs. modular

Dialogue systems employ two main types of architecture. Numerous modern systems are *end-to-end*, which means that they use an encoder-decoder neural architecture to produce the best answer given the user’s previous turn (Ritter et al., 2011; Vinyals and Le, 2015; Sordoni et al., 2015; Lowe et al., 2017). This approach seems particularly suitable for conversational, non-task-oriented scenarios, in which it is often enough to rely solely on the recent dialogue context (Jurafsky and Martin, 2018). Other scenarios, in particular task-oriented ones, require a higher degree of control over the flow of

¹<https://www.cleverbot.com/>, <https://www.elbot.com/>, <https://www.zabaware.com/ultrahal/>

the dialogue as a whole, often including awareness of the information shared by the dialogue participants at different points in time and its relation to the objective of the dialogue, the progress of the task, etc. In other words, it is necessary to *track the state of the dialogue* at all times (Williams et al., 2016). Many dialogue systems dealing with such scenarios have a *modular* architecture. In contrast to end-to-end dialogue modeling, in which a unique function is used to return the best answer given the user's input, in modular systems, processing is divided into different tasks, each handled by a dedicated component (Lowe et al., 2017). An example architecture is shown in Figure 2.2. Typical components of a dialogue system are:

Automatic speech recognition (ASR) Turns speech input from the user into a hypothesis of what has been said, usually in the form of text.

Natural Language Understanding (NLU) Takes the output of the ASR module and converts it into an abstract representation of its meaning which the system can process.

Dialogue Manager (DM) Decides which action to perform next, based on the output of the NLU component and any additional information (such as environmental factors). In some systems, it is divided into two components: one which tracks the state of the dialogue, and another one which selects the next action (Williams et al., 2016).

Natural Language Generation (NLG) Turns the action output by the DM into text.

Text-to-speech synthesizer (TTS) Converts the text from the NLG component into speech.

Partially Observable Markov Decision Processes (or POMDPs) have often been used in combination with supervised or reinforcement learning to train a *decision-making policy* for the DM, i.e. a strategy for making the best decision in every possible state with a view to maximizing an overall long-term reward (Williams and Young, 2007; Young et al., 2010; Gasic et al., 2012; Young et al., 2013). These models also allow taking into account the *uncertainty* caused by factors such as incorrect ASR hypotheses, or unpredictable user behavior and beliefs.

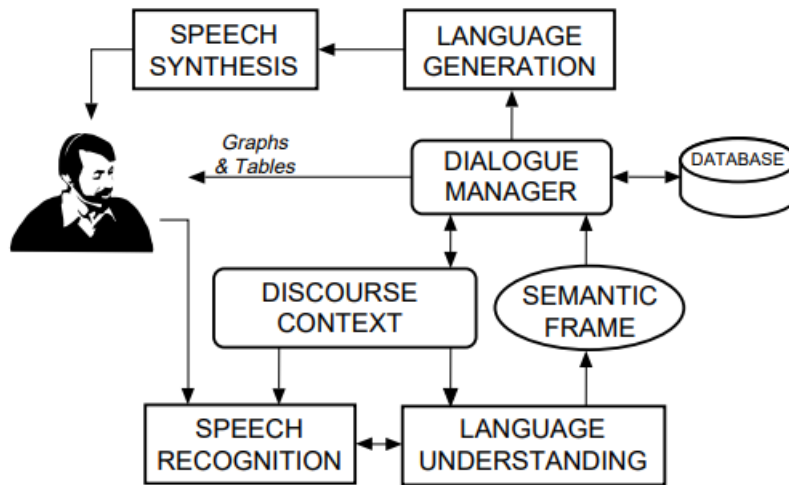


Figure 2.2: Example architecture of a modular spoken dialogue system, from Glass (1999). Besides the modules listed in this section, Glass’ scheme includes a discourse context and a semantic frame module separate from NLU. It also shows the Dialogue Manager interacting with a database.

2.7 Incrementality in language generation

In Section 2.1, we mentioned that participants in a dialogue build their utterances *incrementally* while they talk. This means that speakers do not wait until their contribution is fully planned before starting to produce it. Instead, they start talking and plan the rest of their turn while they speak (Schlangen and Skantze, 2011). Thus, utterances are not planned as a whole, but in sub-utterance chunks, which enables speakers to start talking as soon as the first chunk is ready. This kind of dynamic may sometimes result in *overcommitment*: The production of the utterance started too soon, and the speaker runs out of content before the next part of the utterance is planned. This is sometimes due to internal constraints, such as the cognitive load imposed by simultaneous speaking and planning, and sometimes to external ones, e.g. the speaker needs to find certain information, such as an address or a telephone number, in order to provide it to the interlocutor. Humans often avoid long pauses in such situations by producing fillers (Section 2.4). In addition, as well as planning the utterance while producing it, speakers will monitor it, revise it and modify it in response to changes in the environment

and/or feedback from their interlocutor (Levelt, 1983).

Given that human language production is incremental, efforts have been made to endow spoken dialogue systems with the same capability. This is not only motivated by a desire to base system design on the way dialogue works in nature, but it also brings practical advantages: e.g. it reduces latency, as the system is able to start speaking sooner (Schlangen and Skantze, 2011). One such effort is described in Skantze and Hjalmarsson (2013), in which a dialogue system bridges the gap before information presentation either through fillers (e.g. *eh*) or by playing beginnings of utterances such as *It costs...* or *Here is a...* Once the information is ready to be presented, it is used to complete the started utterances. When comparing this strategy with a non-incremental one—in which the system remains silent until the response is ready—the authors found that the incremental version required less time to complete its utterances.² This also seems to have been noticed by users, who rated it as more efficient than the non-incremental version.

Similarly, Baumann and Schlangen (2013) tested a strategy which uses open-ended utterances, and subsequently extends them as new information comes in. In a car-racing scenario, when the car is about to turn but the direction is yet unknown, the system will begin to say *and then turns...*, and complete with *right* or *left* once the car has started turning in that direction. In addition, filled pauses are introduced to compensate for long pauses resulting from overcommitment. Listeners preferred this strategy over a non-incremental one which always produces full, non-extensible utterances. This preference held despite poor quality of the filled pauses synthesized in the incremental strategy.

Buschmeier et al. (2012) combined incremental Natural Language Generation with incremental speech synthesis in a system which reacts to (simulated) noise interruptions from the environment. When the system receives a noise signal, it pauses its speech and, after the noise has stopped, it re-generates the interrupted chunk. In line with the studies mentioned above, this system was rated more natural by listeners than a non-incremental one. Finally, Tsai et al. (2018) present a movie recommendation

²This may appear obvious, due to the fact that the utterances start sooner in the incremental condition. However, the non-incremental system can start providing task-related information as soon as it becomes available, whereas the incremental one might be in the middle of producing a filler or the beginning of an utterance at that moment, and would have to finish the utterance before starting to provide such information.

system which starts synthesizing replies before the movie to recommend has been chosen, and continues searching for an appropriate result while speaking. Testers did not only rate this system as more responsive than one which waits to find a movie before starting to speak, but they also rated the movie recommendations as better.

2.8 Dialogue systems and human-likeness

Along the years, researchers and developers have striven to create more human-like systems, both with respect to the answers provided and to the voices used to synthesize them. Human-likeness has been a controversial topic in the literature on dialogue systems (Reichman, 1985; Larsson, 2005; Traum, 2018). Right at the dawn of discussions on artificial intelligence, Turing (1950) suggested the existence of a link between human-likeness and intelligence when introducing what is today widely known as the *Turing Test*. According to the Turing Test, a machine displays intelligent behavior if it can have a conversation in natural (written) language and lead humans to believe that it is a human, not a machine. Although the Turing Test remains an important concept in the field of dialogue systems, it has been challenged throughout the years. It has been proposed that human-likeness is not only extremely difficult to achieve but also not strictly necessary, and that developers ought to aim at usefulness instead (Dahlbäck et al., 1993). On the other hand, Edlund et al. (2008) suggest that focusing on human-likeness is a design choice which must be made at the start of the process of creation of the system, since what matters is not whether the system is human-like or not, but whether it is internally consistent. The authors present two metaphors of dialogue system design: The *interface metaphor* corresponds to systems meant to be perceived as tools rather than conversational partners. In these cases, speech is only a substitute for keyboard or mouse interaction. On the other hand, the *human metaphor* corresponds to systems which are meant to be viewed as interlocutors, thus speech is used with them not as a substitute, but as the natural interaction channel. The choices made in the design of the system should match the metaphor selected; for example, a system conceived under the *human metaphor* will be more likely to greet the user at the beginning of the interaction than a system under the *interface metaphor*.

2.9 Evaluation of dialogue systems

To date, the field of dialogue systems lacks consensus about the best approach to evaluation. Unlike other applications of language technology in which the output of the system can be contrasted with a gold standard—and success thus measured as the degree of similarity with the latter—dialogue may develop in many different and sometimes unpredictable ways.

Popular metrics for the evaluation of machine translation, such as BLEU and METEOR (Papineni et al., 2002; Banerjee and Lavie, 2005), have been employed for evaluating dialogue systems (Venkatesh et al., 2018). This relies on the assumption that, in dialogue, a good reply is one with a high amount of token overlap with its preceding question. In practice, however, responses might not repeat any tokens present in the preceding turn, yet be perfectly appropriate. Another approach to objective evaluation of dialogue system performance is to consider measures related to task success, such as task completion rate, task duration, etc. (Ferguson et al., 1996; Walker et al., 1997, 2000). An obvious drawback of this approach is that it is not valid for conversational dialogue systems, which are not meant to fulfill any specific tasks (other than entertaining the user). Moreover, even in the case of task-oriented systems, these measures are useful only to throw light on transactional aspects of dialogue, and they ignore the interactional dimension of system performance which, if faulty, can also be detrimental to user experience.

In order to assess aspects of performance which are dependent on the user's perception—rather than on observable phenomena—MOS (mean opinion score) tests have been widely used. In these tests, users rate different aspects of their experience with the system on a Likert scale, usually from 1 to 5 or 7 (Higashinaka et al., 2018; Lee et al., 2018; Sakai et al., 2018; Kageyama et al., 2018; Lubis et al., 2018). These tests are frequently used to obtain insight into aspects such as perceived human-likeness, friendliness, willingness to help, probability of recommending, etc.

Finally, different evaluation frameworks have been proposed in the last decades which integrate a variety of metrics in order to assess general experience with the system. For task-oriented dialogue systems, an example is PARADISE (Walker et al., 1997, 2000). This framework integrates measures of task success with cost minimization aspects (such as dialogue duration, number of utterances needed to complete the task, etc.), and other measures such as agent response delay. An example for non-task-

oriented (or conversational) dialogue systems has been proposed by Venkatesh et al. (2018). Since task-success metrics cannot be obtained here, the authors explore ways of anchoring abstract dimensions of user experience (which are normally difficult to measure) in quantitative variables. Thus, dialogue duration and number of turns are seen as providing a window into the degree of engagement of the user, coherence of system responses is measured in terms of number of responses with the same topic as the preceding question, etc.

2.10 Summary

When two or more people are involved in a conversation, they are accountable for the use of time during their turn. One of the unspoken norms of dialogue is to not remain silent for too long, although the exact meaning of “too long” in this context has not yet been determined. In order to avoid lengthy pauses and lapses, speakers employ resources such as fillers (*uh... um...*), repetitions and lengthening. Clark (1996) classifies these and other comparable actions as belonging to the *secondary track of metacommunicative acts*, as opposed to the *primary track of communicative acts*, which deals with the main business of the interaction.

An artificial agent which can engage in dialogue is called a *dialogue system*. Dialogue systems whose main channel of interaction with users is speech are called *spoken dialogue systems*, or SDS. Some SDS generate language incrementally, i.e. in sub-utterance units, and are able to start speaking before the plan for the whole utterance or turn is complete, in the same way as humans. To date, the field of Dialogue Systems lacks a standard approach to evaluation, although various techniques and frameworks have been proposed, aimed at assessing both task-related and interaction-related aspects.

Chapter 3

Eliciting Time-Buying in Human-Human Dialogue

3.1 Introduction

In order to endow dialogue systems with smoother ways of buying time, it is necessary to achieve a better understanding of how humans handle this process. For this reason, we set out to investigate the behavior that human speakers exhibit when they need to postpone the delivery of primary task information. Such behavior includes the utterances that they produce and the ways in which they combine them, as well as the duration of the pauses between them. Therefore, this chapter aims at exploring the following set of questions (from the list in Chapter 1):

- Human speakers are sometimes expected to talk while they still lack enough information to provide. What do they do while they search for this information, or while they plan an utterance to convey it?
- Do speakers frequently produce successions of resources of the same type, or do they try to provide variety?
- Do different types of time-buying resources combine in predictable patterns? Are there types of time-buying resources which seem to co-occur particularly often?
- Once the customer has uttered a request, how long does the travel agent take to

start speaking? Does there seem to be a general “maximum tolerable silence” before starting to buy time (or is this highly speaker-specific)?

- How long are the silences between time-buying resources?

In order to answer these questions, we set up an experiment in which a human information provider was shown the information to convey only in a delayed and incremental manner, which systematically created situations where the speaker had the turn but could not provide task-related information. Analysis of the data collected shows that 1) information providers bridged the gap before they were able to find an answer for the search by exploiting task- and grounding-related communicative actions (such as echoing the user’s request or uttering *mm-hm*, see Section 2.3), and that 2) information providers combined these actions productively to ensure an ongoing conversation.

The next section (3.2) describes the data collection process, followed by a description of the data obtained, its segmentation and annotation (Section 3.3). Afterwards, we present our findings with respect to how speakers in the corpus buy time (Section 3.4) and attempt to provide an answer to the questions above (Section 3.5).

3.2 Data collection

3.2.1 Setup and participants

Data collection was conducted at Bielefeld University, in a room of approximately 4x3 meters, divided by a wooden panel. There was a desk with a computer on each side of the panel: one for the participant, and one for their interlocutor (a confederate). It was important to ensure that the speakers did not see each other, since the presence of visual cues might have rendered buying time through speech less important: For instance, if the participant had known that the confederate was able to see them, looking up and down the computer screen with a concentrated facial expression might have been enough to convey the fact that the speaker was trying to find a flight, thus the state of the interaction might have been clear without the need for speech. In addition, there was another computer on the participant’s side, where the recording program (Avid ProTools) was executed.¹ The experimenter sat in front of this computer. Speakers

¹<http://www.avid.com/pro-tools>

communicated using headphones and a microphone connected to an Avid M-Box. The participants were ten, five female and five male, all of them German speakers, recruited through a university mailing list.

3.2.2 Task

Each participant played the role of a travel agent who received calls from potential customers. The role of the caller was played by the confederate (a student assistant), who asked for a flight matching certain criteria. These criteria were displayed on the confederate's screen at the beginning of each call. Examples are: *departing airport/city, destination airport/city, date, preferred time of day, preferred airline, direct flight, low price*, etc. Thus, for one of the flights, the criteria might be:

{	Origin:	near Bielefeld	}
	Destination:	St. Petersburg	
	Date:	Nov. 27, 28 or 29	
	Airline:	not Finnair	

The full list of criteria is shown in Table 3.1. Each call started with the sound of a phone ringing, followed by an automatic greeting message. After the message, only the confederate spoke: The participant was instructed to remain silent at this stage. The confederate made the request, using the criteria displayed on the screen but forming utterances spontaneously rather than just reading out the information (see Table 3.2 for an example with the above criteria). At the end of the request, the confederate pressed the “enter” key. This triggered a beep indicating that the participant was allowed to start speaking, and also instructed the system to start displaying flights on the participant's screen. We established this turn-taking structure in order to introduce a clear separation between the information request and information-providing phases, thus facilitating later analysis. Additionally, it helped ensure that the confederate would provide all the information in the request in the same turn, which we hoped would increase the difficulty of the task, as the participant had to remember all the details while looking through the list of flights. After the beep, both speakers were allowed to talk and manage their turns freely.

The dialogue ended either when the confederate accepted one of the flights offered by the participant, or when both speakers agreed that none of the flights available was

Episode	Criteria	English translation
trial A	Hannover–Izmir, 4. Woche im Juli	Hannover to Izmir, 4th week of July
trial B	nach Rom, Ende November, Nachmittag	to Rome, end of November, afternoon
1	nach Bristol, Erste Hälfte August, Sonntag	to Bristol, first half of August, Sunday
2	nach Korfu, Anfang August, nach 10 Uhr morgens	to Korfu, beginning of August, after 10 a.m.
3	nach Malaga, zwischen 3. und 9. August, so günstig wie möglich	to Malaga, between 3rd and 9th August, as cheap as possible
4	nach Dubai, im Juli, ab 25., nicht zu früh	to Dubai, in July, from the 25th, not too early
5	Köln-Bonn–Lissabon, Ende November, nicht Werktag, am Abend	Cologne-Bonn to Lisbon, end of November, not workday, evening
6	nach Bukarest, Erste Woche August, KLM	to Bukarest, first week in August, KLM
7	Hannover–Helsinki, August, vor dem 10.	Hannover to Helsinki, August, before the 10th
8	nach St. Petersburg, 27, 28 oder 29. November, nicht Finnair	to St. Petersburg, 27th, 28th or 29th November, not Finnair
9	nach Sidney, November, nach 26., Werktag, direkt	to Sidney, November, after 26th, workday, direct
10	nach Quito, August, ab 07., direkt	to Quito, August, from 7th, direct

Table 3.1: List of criteria used by the confederate (the “caller”) to request the flight for each episode.

RING + GREETING	CALLER'S REQUEST	BEEP	TRAVEL AGENT WAITS FOR INFO DISPLAY/SEARCHES FOR FLIGHT (TIME-BUYING STRETCH)	TRAVEL AGENT'S OFFER (+ NEGOTIATION)	CALLER'S DECISION
-----------------	------------------	------	---	--------------------------------------	-------------------

Figure 3.1: Stages of a standard dialogue between a customer (played by the confederate) and a travel agent (played by the participant).

suitable. In the former case, the participant pretended to transfer the customer to an automatic booking system in order to complete the purchase. The confederate then pressed “enter” and the next call started. The instructions for the participant are shown in Appendix A.1. The general structure of an episode is illustrated in Figure 3.1, and Table 3.2 shows one of the resulting dialogues in the collected corpus, together with its English translation.

Only the confederate was able to control the GUI (by pressing “enter” after the confederate’s request or at the end of the episode). The participant could only see the list of flights displayed on the screen, but could not interact with it, i.e. it was not possible to filter or sort the flights in any way. It was also not possible for the participants to write down any information. This, again, was done in order to increase the cognitive load imposed by the task, so that the participants would have to remember the information seen before while looking through the list. A session with a participant lasted approximately 20 minutes and was made up of two trial calls, a pause after which the participant could ask the experimenter questions, and 10 more calls. For each participant, the confederate was the same for every call, but different confederates interacted with different participants (there were five confederates for ten participants in total).

3.2.3 Information display

Information about available flights was displayed to the participant as a list with columns and sorted by price (lowest to highest, see Figures 3.2 and 3.3). For certain episodes, the list included some flights matching the search criteria, whereas for others, none of the flights displayed was suitable for the request.

As the aim of the data collection procedure was to elicit time-buying phenomena, the way in which information was presented was of central importance. We focused on creating situations which led speakers to buy time by combining three strategies:

C: Hallo, äh, ich bin in Bielefeld und ich würde gerne äh nach Sankt Petersburg fliegen, ähm, am liebsten dann natürlich von einem nahegelegenen äh Flughafen aus, und zwar, ähm, am 27., 28. oder am 29. November. Und, äh, ich würde ungern mit der Linie Finnair fliegen, also lieber eine andere Fluggesellschaft, wenn das geht.	C: Hello, uh, I'm in Bielefeld and I'd like to fly to St. Petersburg, uhm, of course preferably from a nearby airport and, uhm, on November 27, 28 or 29. And, uh, I'd prefer not to fly with the airline Finnair, so another airline would be better, if possible.
T.A.: Mhm. Ähm, da habe ich eine, ah, die Liste wird noch vervollständigt, Moment. Ah, da gibt's sehr viele Angebote. Einmal von... von Frankfurt, das wäre am 29.11. Ist Ihnen das zu weit weg?	T.A.: M-hm. Uhm, here I have a... uh, the list is being displayed, one moment. There are a lot of offers. From... from Frankfurt, that would be on November 29. Is that too far away?
C: Ähm, ja näher ist natürlich besser, aber es ist, es wäre noch in Ordnung, denke ich.	C: Uhm, yeah, closer would be better, of course, but it's- it would still be okay, I think.
T.A.: Hannover hätte ich noch im Angebot, das ist dann aber deutlich teurer.	T.A.: I also have an offer from Hannover, but that one's significantly more expensive.
C: Mh, ja, mh.	C: Mh, yes, mh.
T.A.: Also der Flug von Frankfurt würde 28undnein, 287 Euro und von Hannover 684 Euro kosten.	T.A.: So the flight from Frankfurt would cost 28no, 287 euros and the one from Hannover, 684.
C: Gut, dann würde ich natürlich eher von Frankfurt fliegen wollen.	C: Well, then of course I prefer to fly from Frankfurt.
T.A.: Okay, dann wäre das am Frank- äh von Frankfurt nach Sankt Petersburg, am 29.11., Abflug um 13.35 Uhr, Ankunft um 16.15 Uhr. Sie fliegen nicht mit Finnair, sondern mit Emirates und der Preis liegt bei 287 Euro.	T.A.: Okay, then that is on Frank- from Frankfurt to St. Petersburg, on November 29, departure at 13:35, arrival at 16:15. You're not flying with Finnair but with Emirates, and the price is 287 euros.
C: Ja, das hört sich doch sehr gut an. Ja, den Flug nehm ich denn.	C: Yes, that sounds very good. Yes, I'm getting the flight.
T.A.: Gut.	T.A.: Good.
C: Alles klar, Dankeschön!	C: All right. Thanks.

Table 3.2: Example dialogue from the corpus: original on the left, English translation on the right. *C*: customer, *T.A.*: travel agent.

Hannover	Nach:	Izmir	Datum:	4. Woche im Juli	Abflug:	--	Ankunft:	--
ID	VON	NACH	DATUM	ABFLUGSZEIT	ANKUNFTSZEIT	FLUGLINIE	ANSCHLUSSE	PREIS
1	Düsseldorf	Izmir	Mo_20.07	7:35	10:40	Lufthansa	0	288
2	Bremen	Izmir	Di_21.07	11:35	14:40	Alitalia	0	288
3	Hannover	Izmir	Mi_22.07	19:20	22:20	Ryanair	0	362
4	Bremen	Izmir	Do_23.07	19:55	22:20	Alitalia	1	367

Figure 3.2: MINIMAL information display mode

Nach:	Bristol	Datum:	erste Hälfte August, Sonntag	Abflug:	--	Ankunft:	--	
ID	VON	NACH	DATUM	ABFLUGSZEIT	ANKUNFTSZEIT	FLUGLINIE	ANSCHLUSSE	PREIS
1	Hannover	Bristol	Di_04.08	18:20	21:35	Ryanair	1	221
2	Münster-Osnabrück	Bristol	Sa_08.08	9:15	12:35	Alitalia	0	240
3	Bremen	Bristol	Sa_08.08	14:20	17:20	Lufthansa	0	261
4	Bremen	Bristol	Mi_05.08	7:15	10:20	Alitalia	1	290
5	Bremen	Bristol	Fr_07.08	9:55	12:15	Ryanair	1	328
6	Düsseldorf	Bristol	Do_06.08	19:15	22:20	Scandinavian	0	395
7	Frankfurt	Bristol	So_09.08	18:20	21:20	Ryanair	0	446
8	Frankfurt	Bristol	Mo_03.08	15:40	18:55	Alitalia	0	503
9	Frankfurt	Bristol	Do_06.08	7:55	10:15	Scandinavian	1	534
10	Hannover	Bristol	Do_06.08	7:55	10:40	Ryanair	0	550
11	Münster-Osnabrück	Bristol	Mi_05.08	8:20	11:35	Lufthansa	0	601
12	Bremen	Bristol	Mo_03.08	19:40	22:55	Alitalia	0	667
13	Münster-Osnabrück	Bristol	So_09.08	19:15	22:55	Lufthansa	1	702
14	Hannover	Bristol	Do_06.08	12:20	15:55	Scandinavian	0	748
15	Bremen	Bristol	So_09.08	15:15	18:40	Air_Berlin	0	807
16	Münster-Osnabrück	Bristol	Fr_07.08	6:15	9:15	Ryanair	1	834

Figure 3.3: IMMEDIATE information display mode

--	Nach:	Rom	Datum:	Ende November	Abflug:	Nachmittag	Ankunft:	--
ID	VON	NACH	DATUM	ABFLUGZEIT	ABKUNFTZEIT	FLUGLINIE	ANZAHLE	PREIS
1	Essen	Rom	Fr_27.11	8:15	11:15	Air_Berlin	0	230
2	Münster-Osnabrück	Rom	Sa_28.11	19:40	21:55	Ryanair	1	294
3	Frankfurt	Rom	Mi_26.11	19:20	19:40	Ryanair	0	331
4	Hannover	Rom	Mo_23.11	14:40	17:20	Lufthansa	1	351
5	Hannover	Rom	Sa_28.11	15:05	19:40	Alitalia	0	405
6	Hannover	Rom	Mo_23.11	8:20	11:55	Scandinavian	1	418
7	Düsseldorf	Rom	Mo_23.11	15:55	18:20	Scandinavian	0	494
8	Essen	Rom	Fr_27.11	16:40	19:55	Air_Berlin	0	522
9	Düsseldorf	Rom	Sa_29.11	10:20	13:40	Scandinavian	1	553
10	Essen	Rom	Mi_25.11	10:40	13:55	Ryanair	1	625
11	Hannover	Rom	Mo_23.11	11:40	14:15	Alitalia	1	640
12	Hannover	Rom	Sa_28.11	8:15	11:15	Lufthansa	1	689
13	Hannover	Rom	So_28.11	19:55	21:55	Alitalia	0	737
14	Hannover	Rom	Do_26.11	12:15	15:40	Scandinavian	1	778
15	Münster-Osnabrück	Rom	Mi_25.11	15:40	18:55	Lufthansa	1	804
16	Frankfurt	Rom	Fr_27.11	8:15	11:55	Alitalia	0	894

(a) BLOCKS display mode, at second 2

(b) BLOCKS display mode, at second 8

Figure 3.4: BLOCKS information display mode

- temporarily withholding task information, i.e. not always showing the information needed to solve the task on the screen right from the start,
- maintaining the difficulty of the task high enough that participants would not be able to find a result immediately, and
- generating an uncertain dynamic environment, in which participants could never be completely sure whether they were in possession of the final information.

In practice, we implemented these strategies by exposing participants to four different modes of information presentation, described in Table 3.3. The two calls in the trial phase were presented using the MINIMAL (Figure 3.2) and BLOCKS (Figure 3.4) modes respectively, whereas in the second phase (the experiment proper), there were two more occurrences of the MINIMAL mode, two of the BLOCKS mode, three of the IMMEDIATE mode (Figure 3.3) and three of the DELAYED mode (Figure 3.5), presented in random order.

3.3 Data

The corpus collected comprises 2 hours, 31 minutes and 6 seconds of speech (after removing the ring and greeting message at the beginning of each dialogue, see Section 3.2.2). For the analysis, we excluded the two trial episodes corresponding to each

Nach: Bristol		Datum: erste Hälfte August, Sonntag			Abflug: --		Ankunft: --	
ID	VON	NACH	DATUM	ABFLUGSZEIT	ANKUNFTSZEIT	FLUGLINIE	ANSCHLASSE	PREIS
1	Hannover	Bristol	Di_04_08	18:20	21:35	Ryanair	1	221

(a) DELAYED information display mode, at second 2

Nach: Bristol		Datum: erste Hälfte August, Sonntag			Abflug: --		Ankunft: --	
ID	VON	NACH	DATUM	ABFLUGSZEIT	ANKUNFTSZEIT	FLUGLINIE	ANSCHLASSE	PREIS
1	Hannover	Bristol	Di_04_08	18:20	21:35	Ryanair	1	221
2	Münster-Osnabrück	Bristol	Sa_08_08	9:15	12:35	Alitalia	0	240
3	Bremen	Bristol	Sa_08_08	14:20	17:20	Lufthansa	0	261

(b) DELAYED information display mode, at second 4

Nach: Bristol		Datum: erste Hälfte August, Sonntag			Abflug: --		Ankunft: --	
ID	VON	NACH	DATUM	ABFLUGSZEIT	ANKUNFTSZEIT	FLUGLINIE	ANSCHLASSE	PREIS
1	Hannover	Bristol	Di_04_08	18:20	21:35	Ryanair	1	221
2	Münster-Osnabrück	Bristol	Sa_08_08	9:15	12:35	Alitalia	0	240
3	Bremen	Bristol	Sa_08_08	14:20	17:20	Lufthansa	0	261
4	Bremen	Bristol	Mi_06_08	7:15	10:20	Alitalia	1	290

(c) DELAYED information display mode, at second 5

Figure 3.5: DELAYED information display mode

MODE	DESCRIPTION	# IN TRIAL	# IN MAIN
MINIMAL	Only four flights are presented, immediately after the caller's request (see Figure 3.2).	1	2
IMMEDIATE	Sixteen flights are presented, immediately after the caller's request (see Figure 3.3).	—	3
DELAYED	Sixteen flights are presented. These start being displayed between 5500 and 7500 ms. after the caller's request. They are displayed one by one, with delays of random duration from 500 to 2500 ms. between them (see Figure 3.5).	—	3
BLOCKS	Sixteen flights are presented, in two blocks of 8 flights each. The first block is presented immediately after the caller's request. The second block starts being displayed after approximately 6 seconds (the exact duration is a randomly selected number between 5200 and 6500 ms.). These flights are presented one by one, separated by pauses of random duration between 200 and 1500 ms. During display, two of the flights which were presented as available become grayed out, indicating that they are now sold out, and not available for booking anymore (see Figure 3.4).	1	2

Table 3.3: Modes of information display used for the participant's screen. The last two columns represent the number of times each mode is used for each participant in the initial trial phase and in the main experiment respectively.

participant. In the case of one participant, we were only able to collect three dialogues due to technical problems during recording. For another participant, time constraints made it possible to collect only nine dialogues (apart from the two trials) instead of 10. Therefore, the resulting corpus without the trial episodes includes 92 dialogues, with a total duration of 2 hours, 12 minutes and 36 seconds. The typical structure of a dialogue is illustrated in Figure 3.1 (an example of a full dialogue from the corpus can be found in Table 3.2).

3.3.1 Segmentation of the phenomenon

Our main aim was to explore participants' speech (and silence) behavior before they were able to provide concrete flight information. The kind of flight information provided depended on the results displayed: If no matching flights were on the list, the travel agent said so and apologized; otherwise, they offered the caller suitable options and the latter decided whether to buy one of these flights or not. Therefore, for each episode, we concentrated on the part between the time when the travel agent was first allowed to speak (marked by the beep) and the time at which they either offered a specific flight or declared that no suitable flights were available. For practical purposes, we will refer to this segment as *time-buying stretch*.² Below is an example from the corpus, with the time-buying stretch in bold:

- (1) **CUSTOMER:** Ähm, ich hätte gern einen Flug von Köln-Bonn nach Lissabon, ähm, sollte am Ende November losgehen aber nicht werktags, äh, und ich muss abends, äh, ähm, in Lissabon angekommen sein [...]

Uhm, I'd like a flight from Cologne-Bonn to Lisbon, uhm, it should leave at the end of November but not on a weekday, uh, and I need to be, uh, uhm, in Lisbon in the evening [...]

TRAVEL AGENT: Ein Flug von Köln-Bonn nach Lissabon, Abflug Ende November, äh, ein Moment bitte, die Flüge werden noch gesucht... Zur Verfügung steht ein Flug ab Stuttgart am 28.11 um 19:20.

A flight from Cologne-Bonn to Lisbon, departure end of November, uh, one moment please, the search for flights is in progress... there is an available flight from Stuttgart on 28-11 at 19:20.

²This term does not indicate that this is the only part of the dialogue where information is postponed, but is rather chosen, for lack of a better one, to refer to the specific stage in the interaction while no concrete response to the customer's request has been offered yet.

Delimiting the time-buying stretch was not always straightforward. The passage from this stage to the information presentation stage was sometimes smooth, with both stages “blending” together, as in the following example from the corpus, in which the travel agent starts providing some information about the flights on the screen, without committing to a specific offer:

- (2) **CUSTOMER:** Ja, guten Morgen, äh, ich würde gerne nach Korfu fliegen und, äh, was noch schön wäre, wenn ich mit den Emirates fliegen könnte [...]
Yes, good morning, uh, I'd like to fly to Korfu and, uh, it would be nice if I could fly Emirates [...]

TRAVEL AGENT: Dann werde ich mir das mal anschauen. Wir haben... wir haben leider keinen aus der Fluglinie, Emira- Emirates. Es kommen zwar neue rein, ich schau mal eben. Ah, es ist vor allem sehr viel Air Berlin, danach Lufthansa.
I'll have a look then. We have... unfortunately we don't have any with Emira- Emirates airline. Two new flights are coming in, I'm having a look. Ah, it's mainly a lot of Air Berlin (flights), then Lufthansa.

CUSTOMER: Ja.
Yes.

TRAVEL AGENT: Aber wir haben noch... Anfang August sagten Sie?
But we also have... did you say beginning of August?

The lack of a clear division between time-buying stretch and information presentation suggests that sometimes speakers utilized the flight information that was available to them to buy time, even knowing that this information was not complete or suitable enough to offer the caller a result. They achieved this by reading out general information from the screen without offering any results in particular, or by mentioning flights which did not suit the request completely (such as in the example above, in which the speaker announces the availability of flights with airlines other than the one requested by the customer). Thus, our annotation scheme (described in 3.3.2) takes this into account by including categories for such occurrences. In addition, and for the sake of clarity in delimitation of the phenomenon, we established that only the offering of one or more specific flights constitutes the end of the time-buying stretch, rather than a general statement about a number of flights which might be available, without providing any details (e.g. *Several flights are available*, without offering any specific ones).

There are 92 instances of time-buying stretches in the corpus, one per call. These stretches differ in duration, since the latter depends on several factors such as when the relevant information becomes visible and when the speaker finds it, among other considerations. Time-buying stretches are between 3.7 and 49.1 seconds long, with a mean of 17.5 ($SD=10.68$).

3.3.2 Annotation

As mentioned above, time-buying stretches consist of speech interspersed with pauses of varying length. We will refer to speech chunks occurring within the time-buying stretch as *time-buying actions*, or simply *time-buyers*.³ In order to annotate them, we started out from the general DAMSL scheme (Core and Allen, 1997) but —somewhat contrary to our expectations— found that the dialogue moves in our data correspond to various backward and forward-looking actions coded in different parts of the DAMSL hierarchy. Thus, we opted for a flat scheme which allowed us to label conversational actions specific to our domain. The categories are shown together with their DAMSL equivalent and with examples in Table 3.5. It is important to note here that, just like DAMSL, we allow for multi-functionality of the dialogue moves. Moves in the ECHOING category, for example, also have a conversational grounding function (Clark, 1996; Bunt, 2011); however, our focus is on their function to avoid giving task information or being silent.

All 92 time-buying stretches were segmented and annotated by the author of this work. An independent second annotator also labeled a randomly selected set of 20% of the time-buyers, using the information from Table 3.5 as a guideline. For these segments, we calculated Cohen's $\kappa = 0.93$, which indicates that the categories are well-recognizable. The number of time-buyers per time-buying stretch in the corpus is between 1 and 26, with a mean of approximately 5 ($M = 5.38$, $SD=3.7$).

3.4 Analysis

We analyzed the 92 time-buying stretches from the corpus in order to find answers for the questions listed in Section 3.1. The insights derived from this analysis are presented

³We originally called them *time-buying utterances*, but then opted for *time-buyers* because the chunks often contain more than one phonetic utterance.

Category	%
echoing	21
filler	19
agent/system state	10.4
acknowledgment	9.4
commitment	8.8
incomplete	6.7
wait request	6.3
confirmation/expansion/ repetition request	5.9
availability	5.1
other	3.5
partial match	2.2
temp. non-availability	1.6

Table 3.4: Distribution of time-buyer categories

in the following subsections.

3.4.1 Utterances

General observations

There are 490 time-buyers in the time-buying stretches, which shows that subjects did not wait for information in silence. Furthermore, all episodes include at least one time-buyer, i.e. participants never provided a result immediately after the customer's request, even in the cases in which only four flights were displayed and it was relatively easy to find a result promptly.

Secondly, the frequency of the time-buying categories defined in Section 3.3.2 in the corpus is shown in Table 3.4. As can be seen, ECHOING occurs frequently, and so do FILLERS. On the other hand, direct requests to wait, which are the resource of choice in a number of telephony systems, are comparatively rare. It must be noted that ECHOING was the most common resource despite the fact that confirmation of the search parameters was not necessary for the travel agent, given that these were displayed on the screen. This reinforces the idea that this resource may have been used for time-buying rather than for information purposes.

There is considerable variation between speakers in their distribution of time-buyer

Category	Description	DAMSL	Examples
acknowledgment	signaling understanding of the request/ acceptance of task	Signal Understanding → Acknowledge	C: I want to fly to Bristol. A: <i>Okay</i>
echoing	repeating the request or part of it	Signal Understanding → Repeat / Statement → Reassert	C: I'm looking for a flight to Izmir at the beginning of August. A: <i>A flight to Izmir . beginning of August</i>
confirmation/ expansion/ repetition request	A asks C to clarify, repeat or expand on request	Influencing addressee future action → Directive → Info-Request	<i>Did you say Lufthansa?</i>
filler	conventional hesitation sound	---	<i>Uh, uhm, mm, etc.</i>
wait request	A asks C to wait	Influencing addressee future action → Directive → Action-Directive / Information Level → Task Management	<i>One moment, please</i>
agent/system state	providing information about factors which prevent A from offering information	Information Level → Task Management	<i>The search for flights is still in progress. I'm not sure if Emirates flies this route.</i>
commitment	expressing that A is (still) engaged in performing the task	Committing Speaker Future Action → Commit	<i>Let's have a look...</i>
availability	announcing the existence of information without presenting it	Statement → Assert / Committing Speaker Future Action → Commit	<i>I could offer you a number of flights... Hmm, you said Quito, is that correct?</i>
partial match	presenting information which only matches the request partially	Statement → Assert / Signal-Understanding → Repeat	<i>There's a flight to Sidney on 2/8 at 07:15, but you would prefer to fly after lunchtime, so let's keep looking...</i>
temporary non-availability	announcing lack of information at the current moment	Statement → Assert	<i>Until now I haven't found any flights for your request, let's keep looking...</i>
incomplete	partial utterance	Communicative Status → Abandoned	<i>Maybe I can find...</i>

Table 3.5: Time-buyer categories (C: Customer, A: Agent)

	ack.	a/s state	avail.	commit.	c/e req.	echo	filler	inc.	other	partial match	temp n/a	wait
P 1	7	7	5	0	4	3	14	1	1	0	0	3
P 2	1	2	0	0	7	4	7	1	1	0	0	9
P 3	1	5	1	11	5	15	0	1	2	2	0	0
P 4	4	6	2	1	0	24	4	1	0	0	0	9
P 5	7	6	1	11	0	5	3	7	0	0	1	2
P 6	0	13	6	9	8	15	12	12	10	5	4	0
P 7	11	4	4	9	0	29	5	1	1	3	0	0
P 8	1	0	2	1	2	1	10	2	0	1	0	1
P 9	8	2	4	1	3	7	22	4	1	0	2	7
P 10	6	6	0	0	0	0	16	3	1	0	1	0

Table 3.6: Number of time-buyers of each category produced by each participant

categories, in particular for ECHOING and FILLER, which can occur very frequently or rarely depending on the speaker. This is shown in Table 3.6. In addition, it is interesting to notice that the percentages of ECHOING and FILLER produced by each participant correlate negatively ($r(18) = -0.77, p < 0.01$).

Time-buyers over time

Preference for some of the resources also varies along the time-buying stretch, i.e. as time passes and the speaker does not offer an answer for the request. Here again, an example is ECHOING, a phenomenon which is expected to occur more often in the first utterances of the stretch, namely in the proximity of the words being echoed: It would be unusual to echo the caller's request a long time after it has finished. Similarly, announcements of TEMPORARY NON-AVAILABILITY would not be expected right at the beginning of the time-buying stretch, but rather after the search for flights has been in progress for some time. Our results are in line with this expectation. Figure 3.6 shows the frequency of all time-buyers along the first seven slots of the time-buying stretch. Each slot corresponds to one time-buyer: *tb_1* is the first time-buyer produced when the travel agent takes the turn after the customer's request, *tb_2* is the second, etc. We only plot until slot seven because less than 25% of the episodes have more than seven time-buyers.

The graph shows that ECHOING is markedly more frequent in the second and third

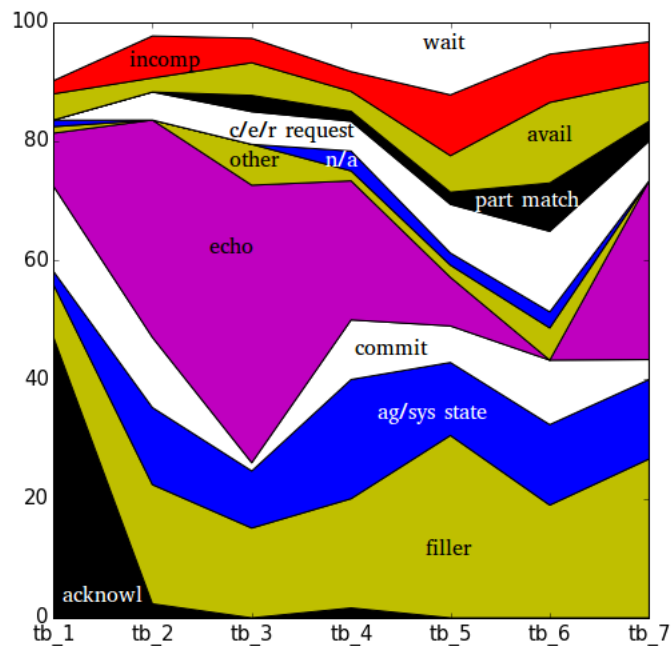


Figure 3.6: Percentage of each time-buyer category for the first seven time-buyers after the travel agent takes the turn

slots than in previous and following ones. As much as 47% of all the time-buyers produced in the third position are instances of echoing, whereas this percentage drops sharply in the immediately subsequent slots. An interesting remark can be made regarding how speakers normally begin to buy time, right after the beep. We see that the category ACKNOWLEDGMENT (*ja, okay*, etc. as reply to caller speech) occurs almost exclusively in the first position, as can be expected, in direct connection to the customer's request. The use of this time-buyer drops from almost 47% to 2% for the second slot and is practically null thereafter. Another interesting case is COMMITMENT, i.e. utterances such as *I'll have a look*. In principle, instances of commitment could also be expected to occur mostly at the beginning: Once the customer has uttered a request, the agent can opt to explicitly signal agreement to take on the task. The data show that, although this resource is indeed most frequent in the first slot, it does not virtually disappear afterwards (in the way acknowledgement does) but there are occur-

Hm, I'd like a flight from...	BEEP	A flight from Köln Bonn	to Lisbon	departure end of November	uh	one moment, please	the search for flights is in progress	There is an available flight...
CALLER'S REQUEST		ECHO: origin	ECHO: destination	ECHO: date	FILLER	WAIT REQUEST	SYSTEM STATE	ANSWER (flight offer)

Figure 3.7: Example interaction with successive instances of ECHOING (gray: caller, white: travel agent)

rences of commitment later, especially from the fourth slot onwards. This seems to suggest that, at times, it may not be enough to signal engagement only at the uptake, but that occasionally speakers may also choose to renew their commitment explicitly later. This could perhaps be related to a need to reassure the interlocutor after a prolonged period of time has elapsed without a satisfactory resolution of the task, or also after additions or modifications to the request.

Based on the above considerations, it is possible to postulate a general tendency for time-buyers in our corpus to be sequenced in the following way: First, taking over the floor (and accepting the task) is acknowledged; then some time is filled, generally with echoing parts of the request; once some information is available, clarification/expansion requests and announcements of partial or full availability become more frequent. Other dialogue acts such as fillers, announcements of system state, or direct wait requests, are available at any time, but they are most relevant after the initial grounding has been performed and before information (partial or full) is available for presentation.

Time-buyer patterns

In the corpus, we observe a tendency for some speakers to produce repeated instances of ECHOING: Out of all bigrams of time-buyers, 10% are instances of ECHOING - ECHOING (40 instances out of 399 bigrams). This is a relatively high percentage considering that there are 144 possible combinations (12 categories in two slots). These speakers echo different parts of the request in a row, as can be seen in the example on Figure 3.7.

The next most frequent bigrams were ACKNOWLEDGMENT - ECHOING (4.3%) (e.g. *Okay, a flight to Bristol...*) and COMMITMENT - ECHOING (3.5%, e.g. *Let's have a look. A flight to Bristol...*), most of them occurring at the beginning of the time-buying stretch. This, again, reinforces the idea postulated in Section 3.4.1 that time-

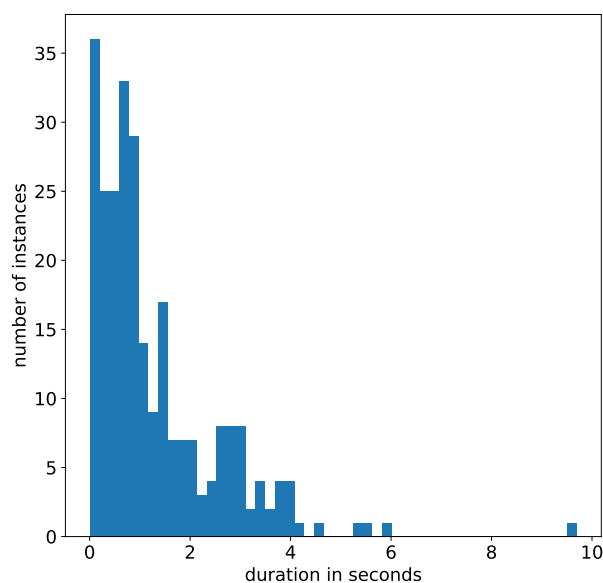


Figure 3.8: Duration of pauses in the time-buying stretch

buying stretches often start with an element expressing either acknowledgment of the request or acceptance of the task, followed by repetition of the query (or a part of it). On the other hand, 49 of combinations did not occur at all, and half of the categories never appeared twice in a row: ACKNOWLEDGMENT, COMMITMENT, INCOMPLETE, PARTIAL MATCH, WAIT and TEMPORARY NON-AVAILABILITY.

3.4.2 Pauses

Figure 3.8 shows the duration of the pauses which occur inside the time-buying stretch, within participants' turns (i.e. we excluded gaps between turns by different speakers). Pause durations range between 0.03 and 9.7 seconds, with a mean of 1.29 seconds ($SD=1.25$). For non-initial pauses (i.e. pauses in slot 2 or higher), mean duration is considerably shorter ($M=0.88$ seconds, $SD=0.8$). In line with Jefferson's (1989) findings (see Chapter 2), there is a large number of pauses with a duration of 1200 ms. or less —however, the ratio is less disparate in our data: Whereas in Jefferson's corpus, there were 3 pauses below 1200 ms. for every pause above this duration, in our data,

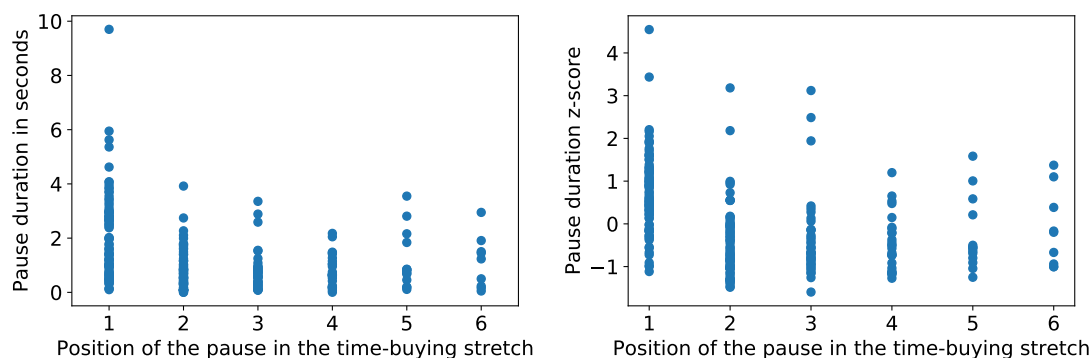


Figure 3.9: Pause durations of the first six pauses in all time-buying stretches: absolute duration (in seconds) on the left, duration normalized by speaker on the right. For the first pause (number 1 on the x -axis) there are 92 data points, and this number decreases gradually for each additional pause. This is because the length of the time-buying stretches varies: Therefore, there is a first pause in all the stretches, but not all of them contain a second/third/etc. pause.

63% of the pauses last 1200 ms. or less, whereas 37% are longer. This difference might perhaps be related to our scenario, since the nature of our task rendered participants unable to provide task information at all at certain points, and this may have resulted in longer pauses than those which normally occur in more generic conversational settings.

The first pause of the stretch (immediately after the customer's request) tends to be relatively long: The mean duration of the initial pause in the time-buying stretch is 2.1 seconds, ($SD=1.58$), whereas the mean pause duration for subsequent slots is much shorter (for slots 2-6, durations in ms. are 678, 601, 792, 958 and 1109 respectively). The longest pauses produced by individual speakers exhibit wide variation (e.g. one speaker did not produce any pauses longer than 1.05 seconds whereas, for another speaker, the longest pause is 9.7 seconds). All initial pauses except for one are shorter than six seconds, and 77% of them are shorter than three seconds.

Although participants normally paused between time-buyers, there are also cases in which these are produced consecutively, without a pause between them. The number of pauses separating time-buyers is 268, whereas speakers produced two time-buyers without pausing between them in 169 cases. Instances of the ACKNOWLEDGMENT and ECHOING categories are often immediately followed by another time-buyer. In 69% of the cases, acknowledgments are immediately followed by more speech, without a

pause. This is not surprising, given that these moves normally occur at the beginning of the interaction, and only uttering *okay* or *mm-hm* may not be enough to clearly signal the speaker's commitment to the task. In the case of ECHOING, 61% of the instances are directly followed by more speech, which is also not surprising, since we have already established speakers' tendency to produce several echoing instances in a row (see Section 3.4.1).

3.5 Discussion

In the introduction, we presented a list of questions regarding the strategies used by speakers to buy time. Below we review these questions and attempt to answer them in the light of the above analysis.

- Human speakers are sometimes expected to talk while they still lack enough information to provide. What do they do while they search for this information, or while they plan an utterance to convey it?

The speakers in the collected corpus use a wide repertoire of resources in order to postpone the delivery of task information. Table 3.5 shows an attempt at systematizing this variety. Speakers bridge the gap until they can provide information by “re-purposing” a variety of task- and grounding-related communicative actions (e.g. echoing the user's request, signaling understanding, asserting partially relevant information, etc.), rather than remaining silent.

- Do speakers frequently produce successions of resources of the same type, or do they try to provide variety?

In general, speakers produce a variety of resources and repetition tends to be avoided, with the exception of the ECHOING category, since it is not uncommon for several instances of this resource to occur one after the other.

- Do different types of time-buying resources combine in predictable patterns? Are there types of time-buying resources which seem to co-occur particularly often?

ECHOING - ECHOING is one of the most common bigrams. Further common combinations are ACKNOWLEDGMENT - ECHOING and COMMITMENT - ECHOING, especially at the beginning of the time-buying stretch. Aside from these combinations, it is difficult to identify any patterns, since speakers seemed to combine resources in many different ways. However, we detected a tendency for certain resources to be more frequent at particular moments. Participants frequently reply to customers' requests with an instance of ACKNOWLEDGMENT, such as *okay*. Afterwards, it is common to echo parts of the request. Later, instances of CLARIFICATION/EXPANSION/REPETITION REQUEST, AVAILABILITY and PARTIAL MATCH become more frequent. In addition, FILLER, AGENT/SYSTEM STATE and WAIT REQUESTS are produced all throughout the time-buying stretch.

- Once the customer has uttered a request, how long does the travel agent take to start speaking? Does there seem to be a general “maximum tolerable silence” before starting to buy time (or is this highly speaker-specific)?

The high inter-speaker variability in our data makes it difficult to provide a definite answer to this question. Overall, almost all first pauses (except for one) are shorter than six seconds, and most of them (77%) are shorter than three seconds. The mean duration of first pauses is 2.1 seconds.

- How long are the silences between time-buying resources?

The average pause duration in the time-buying stretch is 1.29 seconds ($SD=1.25$) when including initial pauses (right after the customer's request), and 880 ms. ($SD=0.8$) when excluding them.

3.6 Summary

It is often difficult to systematically elicit conversational phenomena in human-human dialogue, at least to an extent that can support robust data-driven systems for conversational dialogue (Gustafson and Merkes, 2009). In this chapter, we have presented an experiment designed to investigate conversational strategies used to bridge time or to say something before fully knowing what to say. These phenomena were triggered on

the one hand, by maintaining the difficulty of the task at an appropriate level (challenging enough that it would not be possible to solve it too fast) and, on the other hand, by manipulating and delaying the information that the participant needs to communicate.

Subsequently, we defined the *time-buying stretch* as the phase of the interaction when the information provider cannot yet offer task information, and focused our analysis on the phenomena which occur during this stretch. We observed the speech resources which participants produced (which we refer to as *time-buyers*) and proposed a taxonomy to classify them. Certain categories (such as ECHOING and FILLER) are clearly more frequent than others overall, although individual preferences for some or other resources are also noticeable. Some of these resources are normally preferred at specific times, whereas others are equally frequent all throughout the time-buying stretch. It is also possible to find pauses of various lengths, whose general distribution resembles descriptions found in the literature (Jefferson, 1989; Campione and Véronis, 2002; Lundholm Fors, 2015).

This analysis constitutes a first step towards understanding a set of time-buying behaviors that could be incorporated into spoken dialogue systems, enabling them to better cope with the demands of real-time interaction. In the following chapters, we attempt to model various aspects of these time-buying behaviors computationally, embed them into systems and assess the resulting impact on humans' perceptions.

Chapter 4

Experiment 1: Speech, Silence and Time-Buying

4.1 Introduction

The previous chapter addressed the topic of how humans “fill up” silent time in conversation while they look for content to provide, such as an answer to a question that has just been asked. In this chapter, we extend this idea to dialogue between humans and spoken dialogue systems. Should computers also “buy time”?

If we wish to design systems which emulate human behavior as closely as possible, our first answer might be *yes* —after all, humans do buy time. In Chapter 2 we proposed that one of the main reasons for this is the need to avoid awkwardly long pauses. However, replies that come too soon may also, under certain circumstances, be considered unnatural for human standards. Imagine the following interaction between two humans:

Speaker A: What’s the square root of 67,937?

Speaker B: 260.64

Most human conversational partners would not be able to provide this kind of response instantly. Therefore, a computer which announces the result right away cannot be seen as acting in a human-like manner.

On the other hand, the role of human-likeness in artificial intelligence is a controversial issue (as discussed in Section 2.8), and the level of resemblance of human-human dynamics that is preferred may differ across applications and across users. Therefore, “because humans do it” may not be considered a strong enough reason to justify time-buying capabilities in dialogue systems.

A more practical consideration is that *lengthy silences can confuse users*. Lack of feedback regarding a user’s question, for example, might lead this user to think that the system has either not received the incoming speech, is still processing it, or has simply crashed. In contrast, a system which buys time while searching for information or planning a reply provides a certain degree of insight into its own processing state, which could help avoid such confusion. Speech feedback can be more informative for this purpose than a progress bar or a blinking light (it makes it possible, for example, to explain the reasons for the delay). This kind of visibility gains even further significance when considering the pragmatic implications of lengthy silences in human-human dialogue. For instance (as mentioned in Section 2.2) when someone asks for a favor and the interlocutor takes too long to answer, this can be interpreted as a sign of lack of willingness to grant that favor, even if the answer is affirmative (Roberts and Francis, 2013; Kohtz and Niebuhr, 2017). Given that human users are known to sometimes attribute personality traits to spoken dialogue systems, this is a legitimate consideration in the context of human-computer dialogue (Fink, 2012).

In this chapter, we seek to answer the following questions from Section 1.2:

- How would listeners experience a system which buys time in a similar way to humans? Would they find it more human-like? Or on the contrary, would they perceive it as too artificial, given that they do not expect this kind of behavior from a system?
- Would humans be more willing to interact with this system than with one that cannot buy time, or that fills up waiting time differently (e.g. by explicitly asking the user to wait)?
- How does the time-buying strategy used by a system affect humans’ perception of waiting time? More specifically: If a system buys time in a natural, conversational way, will they perceive the wait as shorter?

To answer these questions, we conducted a study in which participants rated two

information systems: one which asked the interlocutor to wait and then remained silent while looking for the information to present, and another one which produced utterances of various types during the wait. We found that participants perceive the time elapsed between the interlocutor's request and the system's response as longer in the first condition, even though the actual time elapsed is the same. In addition, if the synthesized voice is relatively human-like, the system producing utterances throughout the wait is also perceived as more willing to help, better understanding of the user's request, and more human-like.

This chapter is structured as follows. In Section 4.2, we discuss briefly how humans experience waiting under different conditions. Section 4.3 describes the experiment mentioned in the previous paragraph and its results. We discuss the latter and attempt to answer the questions listed above in Section 4.4.

4.2 The waiting experience

Research about humans and waiting has been carried out in various areas such as transportation services, system usability and customer satisfaction. Part of this work centers around humans' emotional responses to waiting, and how the environmental conditions of the wait influence such responses. Studies have shown that long waits are normally associated with negative feelings such as anger and frustration (Friman, 2010; Taylor, 1994). While this is not surprising, it is rather difficult to define what exactly constitutes a long wait, since this is subject to a number of factors, some of them at least partly subjective. The waitee's expectations with regard to the duration of the wait play a key role, which is why researchers have tested different strategies to manage these expectations, such as providing an estimated duration for the wait or offering a "maximum wait guarantee" (Antonides et al., 2002; Kumar et al., 1997). It has also been claimed that the negative emotions triggered by lengthy waits are less marked when the latter takes place *in-process* (such as during a purchase) rather than before the process (Friman, 2010). As can be expected, waiting in a visually attractive environment also has been shown to mitigate the negative effects of waiting (Pruyn and Smidts, 1998).

In addition, researchers have experimented with different ways of *filling up time* in order to make the waiting experience less frustrating. Taylor (1994) reported that passengers who waited to board a plane at an airport felt less angry and less uncertain when they reported having performed other activities during the wait. Similarly,

customers calling a service line and waiting to be assisted by a human operator found this wait more enjoyable when music was played than when they had to wait in silence (Tom et al., 1997).

An important concept in connection to the waiting experience is that of *perceived duration*, i.e. the time that the waitee *thinks* has elapsed while waiting, as opposed to the actual duration of the wait. Decades ago, Hirsch et al. (1956) found that auditory stimulation has an impact on perception of elapsed time. Specifically, the author observed that when participants heard a sound over background noise, they tended to underestimate its duration. Relatedly, Tom et al. (1997) observed in one experiment that customers perceived elapsed time as shorter when the wait was “filled” with music; however, the author did not find this effect in a second, similar experiment. It is also possible that the kind of filler used may play a role on the perception of the duration of the wait. In this respect, Antonides et al. (2002) found overestimation of waiting time to be less when waitees received information about waiting time in advance than when they heard music or information about location in the queue. Munichor and Rafaeli (2007), in turn, claim that information about location in the queue resulted in a more positive experience for participants than playing music or issuing apologies.

4.3 Experiment

In the following subsections, we describe an experiment in which we compared testers’ perception of two German-speaking systems: one which buys time using some of the time-buying actions we found in our human-human corpus, and one which asks the user to wait and then remains silent.

4.3.1 Method

Design

The main factor was WAIT vs. TIME-BUYING:

- **WAIT:** The system asks the customer to wait by producing an utterance such as *Bitte einen kleinen Moment Geduld* (“Please be patient for a moment”), and then remains silent until it announces having found the flight.

- **TIME-BUYING** The system produces a variety of utterances separated by short pauses, thus buying time until it has found a flight.

We conducted two runs of the study, with two different speech synthesizers: the first one more easily identifiable as a machine and the second one sounding (subjectively) more human-like (see *Materials* below). Participants listened to four recordings, two for each condition, in random order.

Participants

Recruitment was carried out on the crowdsourcing platforms Amazon Mechanical Turk and Crowdfunder, and limited to workers in Germany. Forty-two subjects participated in the first run (16 female and 26 male, aged 20 to 69) and 39 in the second run (15 female and 24 male, aged 21 to 63). The study was published in the form of a questionnaire on the online platform SoSciSurvey.¹

Materials

Stimuli: The full dialogues are shown in Table 4.1. The human customer asked for a flight meeting certain criteria and the system pretended to look for an option which satisfied the customer’s needs (see Figure 4.1). After a while, the system announced having found an appropriate flight. The time between the end of the customer’s request and the system’s announcement was approximately 12 seconds.²

Time-buying sequence generation: In order to produce the sequences for the time-buying condition, we implemented a simple “time-buying generator” which produced a sequence of five time-buying actions and then announced having found a flight. Since there were only two stimuli for this condition, and they were relatively short, we did not use the whole range of time-buying categories from the DSG Corpus but only the most frequent ones: ECHOING, FILLER, ACKNOWLEDGMENT and AGENT/SYSTEM STATE. At each step, the system chose one of

¹<https://www.mturk.com/>, <https://www.crowdfunder.com> (now part of Appen), <https://www.soscisurvey.de/>

²We considered 12 seconds to be a realistic waiting period a relatively lengthy lookup might take, yet not so long that the WAIT strategy would obviously be disadvantaged.

these categories and produced one out of a set of canned utterances belonging to that category (taken from the DSG-Travel Corpus, the corpus of human interactions described in Chapter 3). The choice of category depended on: a) the previous system utterance and b) the number of time-buyers already produced since the beginning of the time-buying stretch. Using bigram probabilities for a and unigram probabilities for b (both calculated using the DSG-Travel Corpus), the system produced a probability distribution over all possible categories given the number of time-buyers produced before, and sampled from it to select the next action. The time-buying actions were interspersed with pauses of random duration between 500 and 1500 ms.

Voices: For the first run, the system's utterances were synthesized using MaryTTS, whereas Cereproc was used for the second run.³ For MaryTTS we chose a Hidden Semi-Markov Model (HSMM) voice, which resulted in a (subjectively) less natural sound than the second one, a commercial Unit Selection voice (Cereproc Alex). The dialogues were the same in both runs, and all participants were presented with all the dialogues.

Procedure

The participants first provided demographic data and completed a brief German language check, in order to verify that they understood German. The check consisted in listening to a short recording in this language and answering a question about the contents of this recording. Results from participants who did not pass this check were excluded from the analysis. Subsequently, the task instructions were displayed, followed by an example trial.

After this initial phase, the task started. Participants listened to recordings of enacted phone conversations between a human customer and an automatic system at a travel agency.⁴ After each recording, participants rated the corresponding system between 1 and 5 (5 meaning "strongly agree") with respect to five statements (here in translation):⁵

³<http://mary.dfki.de/>, <https://www.cereproc.com/>

⁴The customers' utterances were taken from the DSG-Travel corpus, from the confederates' turns.

⁵Images of the survey can be found in Appendix B.1

Example	
C: Äh, guten Tag, ich würde gern... nach Izmir fliegen von Hannover aus und zwar in der vierten Woche im Juli, wenn das geht.	C: Uh, good morning, I'd like to fly... to Izmir from Hanover in the fourth week of July, if that's possible.
S: Hallo, ähm... ein Flug nach Izmir, von Hannover aus, im Juli... Ich habe einen Flug für Sie gefunden, das wäre am...	S: Hello, uhm... a flight to Izmir, from Hanover, in July... I've found a flight for you. That would be on the...
Dialogue 1	
C: Äh, guten Tag. Ich würde gerne Ende November, äh, nach Rom fliegen und zwar bin ich in Bielefeld, ich weiß jetzt nicht welcher Flughafen da am nächsten ist, äh und wenn möglich würde ich gerne nachmittags abfliegen.	C: Uh, good morning. I'd like to fly to Rome, uh, at the end of November and I'm in Bielefeld; I'm not sure which airport is closest, uh, and if possible, I'd like to depart in the afternoon.
S: Okay, ein Flug nach Rom, ich schaue mal eben. Ende November, ähm... Ich habe einen Flug für Sie gefunden, das wäre am...	S: Okay, a flight to Rome, I'll have a look. At the end of November, uhm... I've found a flight for you. That would be on the...
Dialogue 2	
C: Ja, hallo. Ich suche einen möglichst günstigen Flug nach Malaga, und zwar zwischen dem Dritten und Neunten August.	C: Yes, hello. I'm looking for a cheap flight to Malaga if possible, between August third and ninth.
S: Ich bitte um ein wenig Geduld. [SILENCE] Ich habe einen Flug für Sie gefunden, das wäre am...	S: Please bear with me. [SILENCE] I've found a flight for you. That would be on the...
Dialogue 3	
C: Äh, ich hätte gerne einen Flug nach Bucharest, und zwar in der ersten Augustwoche, und ich möchte, äh, auf jeden Fall mit der Fluglinie KLM fliegen.	C: Uh, I'd like a flight to Bucharest, in the first week of August, and I'd like, uh, to fly KLM by all means.
S: Ich schaue mal eben. Das System arbeitet gerade, ähm... Ein Flug nach Bukarest, mit KLM... Ich habe einen Flug für Sie gefunden, das wäre am...	S: I'll have a look. The system is working, uhm... a flight to Bucharest, with KLM... I've found a flight for you. That would be on the...
Dialogue 4	
C: Hallo, ich möchte von Köln-Bonn nach Lissabon fliegen, ähm, Ende November und ich möchte nicht an einem Werktag fliegen.	C: Hello, I'd like to fly from Cologne-Bonn to Lisbon, uhm, at the end of November and I don't want to fly on a weekday.
S: Bitte einen kleinen Moment Geduld. [SILENCE] Ich habe einen Flug für Sie gefunden, das wäre am...	S: Please bear with me for a moment. [SILENCE] I've found a flight for you. That would be on the...

Table 4.1: Dialogues presented to participants. The first one is the example dialogue and the following four are the experiment dialogues, which were presented in random order. C: customer, S: system. Dialogues 1 and 3 correspond to the TIME-BUYING condition, and dialogues 2 and 4, to the WAIT condition.

CUSTOMER	SYSTEM					
Ich würde gerne Ende November von Köln nach Rom fliegen	Bitte einen kleinen Moment Geduld		...		Ich habe einen Flug für Sie gefunden	
	<i>Please hold on a second</i>		...		<i>I've found a flight for you</i>	
WAIT strategy						
<i>I'd like to fly from Cologne to Rome at the end of November</i>	Okay	Ein Flug nach Rom	Ich schaue mal eben	Ende November	äh	Ich habe einen Flug für Sie gefunden
	<i>Okay</i>	<i>A flight to Rome</i>	<i>I'll have a look</i>	<i>end of November</i>	<i>uh</i>	<i>I've found a flight for you</i>
TIME-BUYING strategy						

Figure 4.1: Example dialogue for each of the two experiment conditions (original utterances in German in bold; English translation below in italics)

1. The system understood the caller well.
2. The system took an appropriate amount of time to find a flight.
3. The system sounds as if willing to help.
4. The system acts the way I would expect a person to act.
5. If I had to buy a flight on the phone, I would use this system.

We composed these statements ourselves based on the aspects that we wanted to evaluate, and we asked other research group members to review them. However, similar statements or questions have been included in papers by other authors:

The system understood the caller well. Harms and Biocca's (2004) Networked Minds Social Presence Measure includes the statements (*My partner*) *found it easy to understand me* and (*My partner*) *had difficulty understanding me*.

The system took an appropriate amount of time to find a flight. Skantze and Hjalmarsson (2013) include a differential *faster response - slower response*. Note, however, that "appropriate" does not necessarily mean "fast", as responses which come too fast might also be perceived as inappropriate or unnatural (see Section 4.1).

The system sounds as if willing to help. Bergmann et al. (2012) include “helpful” as one of the dimensions to rate for their virtual agents. In principle, “helpful” and “willing to help” seem related; however, they are not necessarily the same (someone could be helpful without wanting to, or someone might be willing to help and yet not be able to do it).

The system acts the way I would expect a person to act. A number of papers include the notion of human-likeness in their questionnaires, e.g. Bartneck et al. (2009); Skantze and Hjalmarsson (2013).

If I had to buy a flight on the phone, I would use this system. The SUS usability scale (Brooke, 1996) includes the statement *I think that I would like to use this system frequently*. Shamekhi et al. (2016) asked participants to rate how willing they were to continue interacting with the agent.

4.3.2 Results

We compared the ratings between the WAIT and TIME-BUYING strategies. Since Likert data are ordinal, they cannot be normally distributed, which is why we report median as measure of central tendency (instead of mean) and interquartile range for dispersion (instead of standard deviation). For the same reason, we use a non-parametric test, Wilcoxon signed rank, in order to evaluate significance of differences (Wilcoxon, 1945). We used Bonferroni-adjusted alpha levels to correct for multiple comparisons: Given that testers rated five statements per stimulus, the alpha levels used were $.05/5 = .01$, $.01/5 = .002$, $.001/5 = .0002$). In the first run, in which the Mary-TTS voice was used, raw scores for TIME-BUYING were higher than for WAIT for all five statements. However, the difference only proved significant in the case of statement 2, “The system took an appropriate amount of time to find a flight” ($W = 244.5$, $p < .002$). Results are displayed in Table 4.2, and the distribution of ratings for each statement is shown in Figure 4.2.

In the second run, in which Cereproc Text-to-Speech was used, the TIME-BUYING strategy was rated better for each of the five statements, and differences were highly significant in all cases. Results are displayed in Table 4.3, and the distribution of ratings for each statement is shown in Figure 4.3.

Statement	Total sum	Total sum	<i>Mdn</i>	<i>Mdn</i>	<i>IQR</i>	<i>IQR</i>	Wilcoxon
	WAIT	TB	WAIT	TB	WAIT	TB	
1	324	337	4	4	2	1	$W=473.5, p>.01$
2	311	342	4	4	1	1	$W=244.5, p<.002$
3	313	324	4	4	1	1	$W=259.5, p>.01$
4	280	301	4	4	1	1	$W=321.5, p>.01$
5	252	269	3	3	2	1	$W=128, p>.01$

Table 4.2: Results for each statement: HSMM voice run. Columns 2 and 3 show the sum of the ratings per statement for each condition. TB stands for TIME-BUYING, and *IQR*, for *interquartile range*.

Statement	Total sum	Total sum	<i>Mdn</i>	<i>Mdn</i>	<i>IQR</i>	<i>IQR</i>	Wilcoxon
	WAIT	TB	WAIT	TB	WAIT	TB	
1	305	349	4	5	2	1	$W=111, p<.0002$
2	250	342	3	5	2	1	$W=52, p<.0002$
3	260	310	4	4	1	1	$W=163.5, p<.0002$
4	236	289	3	4	2	1	$W=248, p<.0002$
5	222	267	3	4	1	1	$W=132, p<.0002$

Table 4.3: Results for each statement: Unit Selection voice run. Columns 2 and 3 show the sum of the ratings per statement for each condition. TB stands for TIME-BUYING, and *IQR*, for *interquartile range*.

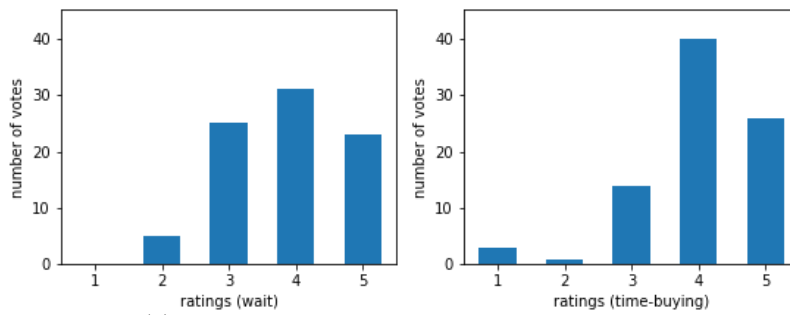
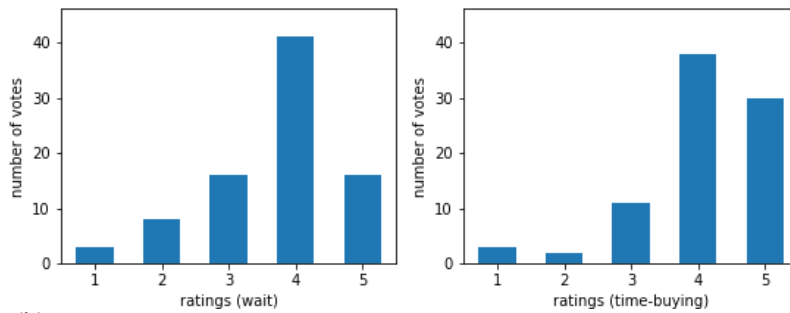
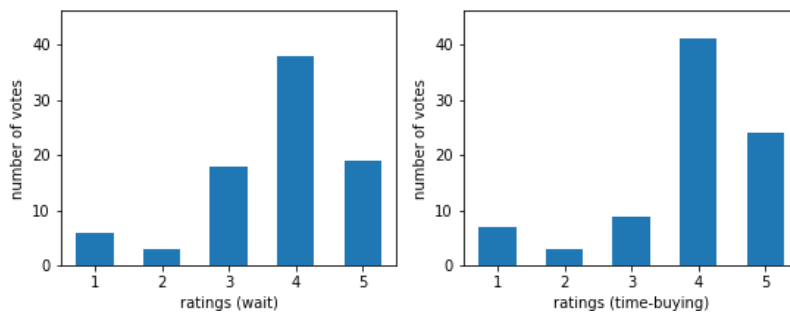
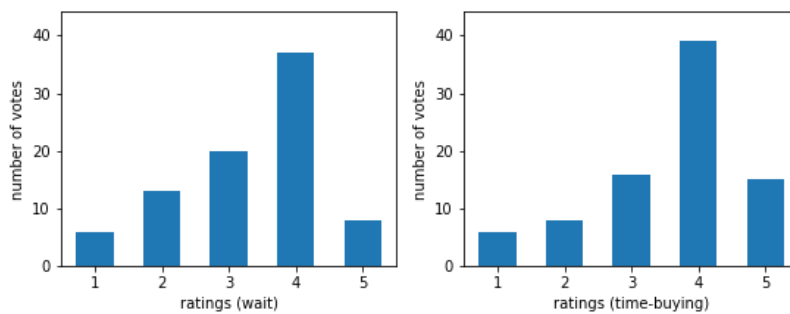
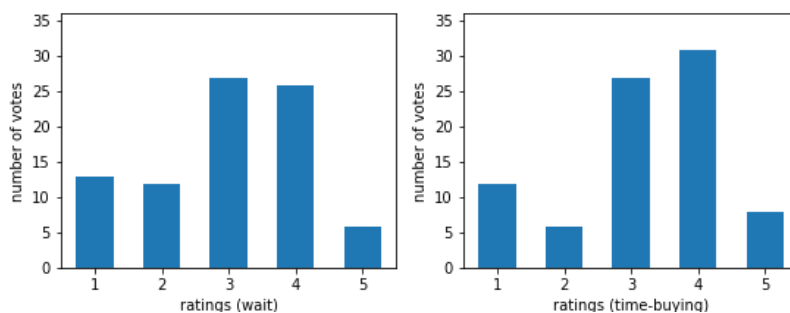
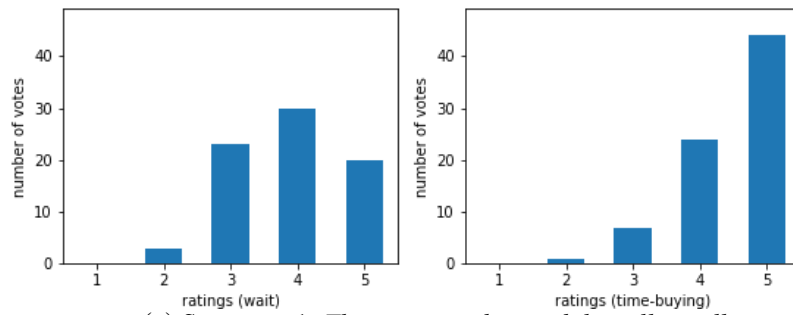
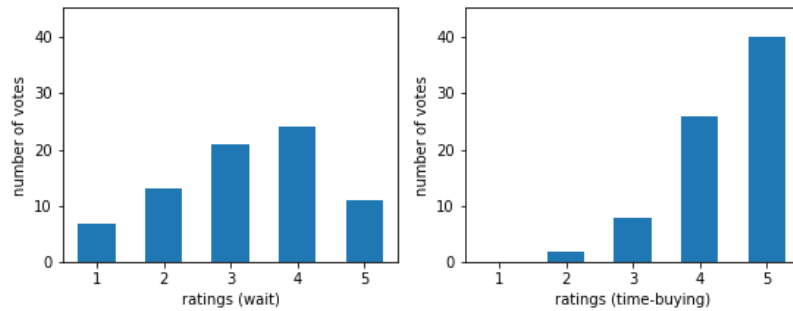
(a) Statement 1: *The system understood the caller well.*(b) Statement 2: *The system took an appropriate amount of time to find a flight.*(c) Statement 3: *The system sounds as if willing to help.*(d) Statement 4: *The system acts the way I would expect a person to act.*(e) Statement 5: *If I had to buy a flight on the phone, I would use this system.*

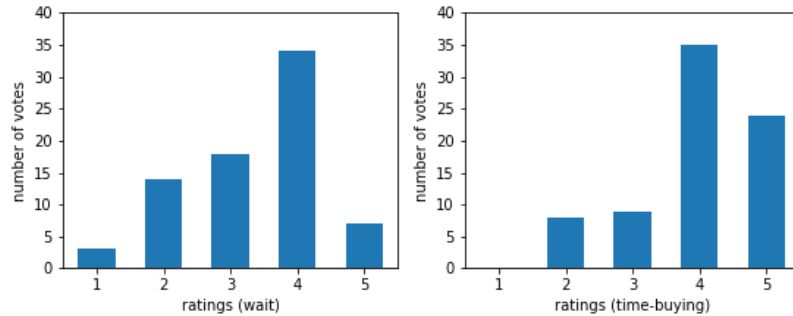
Figure 4.2: Distribution of ratings for the first study run (HSMM voice)



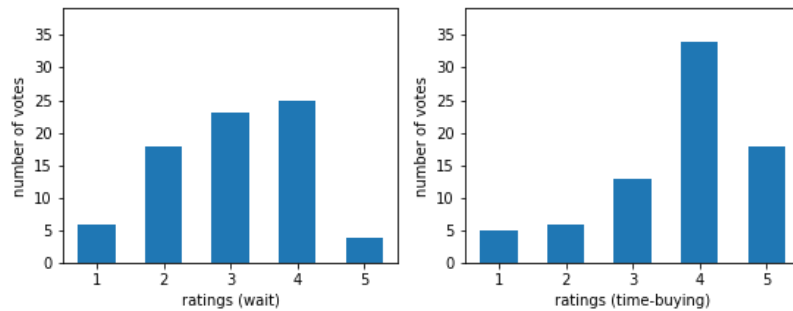
(a) Statement 1: *The system understood the caller well.*



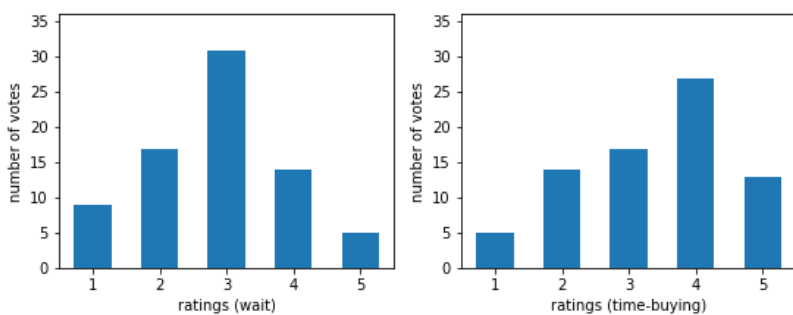
(b) Statement 2: *The system took an appropriate amount of time to find a flight.*



(c) Statement 3: *The system sounds as if willing to help.*



(d) Statement 4: *The system acts the way I would expect a person to act.*



(e) Statement 5: *If I had to buy a flight on the phone, I would use this system.*

Figure 4.3: Distribution of ratings for the second study run (Unit Selection voice)

4.4 Discussion

The results presented above show that an information-providing dialogue system which uses speech to avoid long gaps after an interlocutor's request —similarly to what humans usually do— can make a better impression on overhearers than a system which asks the user to wait and then remains silent until it can provide an answer. In the first run of our study, participants found waiting times to be more appropriate in the TIME-BUYING system than in the WAIT one, even though the actual times remained constant across conditions. Additionally, the second run revealed that listeners also perceived the TIME-BUYING system as more willing to help, better understanding of the user's request, and more human-like than the WAIT system. Finally, participants preferred the former over the latter for their own use. These results suggest that dialogue systems could benefit from the incorporation of time-buying capabilities.

On the other hand, the differences between the results of both study runs open up questions regarding the interplay of voice quality and time-buying strategy. While the TIME-BUYING system was already perceived as better in study run 1 in one respect (time needed to find a flight) without detriment to the other four aspects, the results of the second study run showed a more marked preference for the TIME-BUYING system, since it was rated significantly better for all five aspects considered (see Table 4.4). We can view this from two perspectives. On the one hand, the fact that *there is more speech overall* in the TIME-BUYING samples than in the WAIT samples might have intensified the impact of voice quality on listeners' experience. In particular, when rating human-likeness, the fact that the HSMM voice —we assume— is perceived as less human-like might be more salient whenever the system produces more speech (i.e. when buying time conversationally), and this may counterbalance the fact that the time-buying behavior is indeed more human-like. A similar logic might apply for the other statements: For example, in *I would use this system*, participants might in principle favor a strategy which buys time conversationally, but also prefer one that speaks less if the voice is not human-like (and therefore unfamiliar to them), which would also give some small advantage to the WAIT strategy. However, this is only speculation, as more data would be required to draw clear conclusions.

Another (although related) perspective from which we could view the differences between the two experiment runs is that participants may have found the more human-like voice in the second run a *better match* for the more human-like behavior of the

	Run 1 (MARY TTS)	Run 2 (Cereproc)
Q1: Understanding	No significant difference	TIME-BUYING
Q2: Time elapsed	TIME-BUYING	TIME-BUYING
Q3: Willingness to help	No significant difference	TIME-BUYING
Q4: Human-likeness	No significant difference	TIME-BUYING
Q5: I would use it	No significant difference	TIME-BUYING

Table 4.4: Summary of the results of both experiment runs. The table shows which system (WAIT vs. TIME-BUYING) participants preferred for each statement.

TIME-BUYING system. This could be interpreted in the light of Edlund et al.’s (2008) *metaphors* in human-agent interaction (already discussed in Section 2.8). The authors draw a distinction between the *interface metaphor* (in which the system is perceived as a machine) and the *human metaphor* (in which the system is viewed as an interlocutor with whom speech is the natural interaction channel) and highlight the need for internal coherence between the metaphor selected and the behavior of the system. Here again, one can only speculate, since the data at hand do not provide enough insights as to the actual rootcause of the differences.

From this perspective, one could argue that a system seeking to buy time like humans should use a voice as similar as possible to that of a human. However, deciding what kind of voice is best for a dialogue system is not always so straightforward, and other considerations also need to be taken into account. One of them is flexibility. Many commercial TTS systems sound relatively human-like but do not offer many options for acoustic modification (other than SSML tags for coarse-grained pitch, speed or loudness adjustment, or general emotion tags).⁶ Systems like MaryTTS, on the other hand, offer both unit selection and HSMM voices, and the latter grant the possibility, for example, to adjust the frequency and duration of each phone to specific values (Schröder and Trouvain, 2003). It is therefore necessary to take this trade-off between human-likeness and flexibility into account, and prioritize depending on the aims and specificities of the dialogue system under construction.⁷

⁶This was the case of the voice used for the second study run, see <https://www.cereproc.com/files/CereVoiceCloudGuide.pdf>

⁷It must be mentioned that, although at the time of the study, HSMM-based and Unit Selection synthesis were the main approaches in TTS systems available to the public, neural technologies have been introduced since then, which render much more human-like speech (van den Oord et al., 2017;

This experiment was a preliminary attempt towards understanding the effects of conversational time-buying on human's perception of a spoken dialogue system. However, such a system is ultimately aimed at interacting with humans: Therefore, it should ideally be tested in an interactive setting, where the users of the system are also its evaluators. Betz et al. (2018) observed differences in the results of an evaluation of speech synthesis between interactive and non-interactive scenarios. The authors ran two MOS evaluations of synthesis quality for artificial speech with and without hesitations. In one of them, the ratings were provided by the users of the system, whereas the other one involved crowd workers who listened to the system without interacting with it. Results reveal a significant difference between ratings for both synthesis conditions in the interactive experiment, but not in the crowdsourced one (although the authors attribute these differences partly to the presence/absence of interaction and partly to specific characteristics of their experimental settings). Analogously, evaluating time-buying in an interactive setting might produce different results from those obtained in an overhearer study (like the one in this chapter), since participants who are engaged directly with the system might focus on different aspects of its performance than those acting as mere listeners.

Below we attempt to answer the questions asked in Section 4.1:

- How would listeners experience a system which buys time in a similar way to humans? Would they find it more human-like? Or on the contrary, would they perceive it as too artificial, given that they do not expect this kind of behavior from a system?

Our results suggest that listeners may find a system which buys time conversationally more human-like than one which asks the user to wait and then remains silent, as long as its voice is also relatively human-like.

- Would humans be more willing to interact with this system than with one that cannot buy time, or that fills up waiting time differently (e.g. by explicitly asking the user to wait)?

Similarly to the previous question, participants expressed more willingness to interact with a system which buys time conversationally than with one that cannot buy time

Latorre et al., 2018).

when the voice of the system was relatively human-like. A comparison with a system which fills up waiting time by explicitly asking the user to wait is presented in Chapter 5.

- How does the time-buying strategy used by a system affect humans' perception of waiting time? More specifically: If a system buys time in a natural, conversational way, will they perceive the wait as shorter?

In this case, the answer was *yes* regardless of the system's voice: Humans perceived the wait for task information as shorter when the system bought time conversationally than when it asked the user to wait and then remained silent.

4.5 Summary

In this chapter, we took a preliminary look at how humans perceive a spoken dialogue system which buys time conversationally. We reviewed some of the literature on waiting, and discussed how humans experience this process in different conditions. We described an experiment where we compared human listeners' perceptions of a system that buys time conversationally with that of one that asks its interlocutor to wait and then remains silent until it can provide an answer to the user's request. We conducted two identical runs of the experiment—one with a subjectively more human-like voice than the other one—and found that, although waiting time was perceived as shorter in both runs when the system bought time conversationally, participants' preference for the time-buying system was clearer when the voice used for both systems was more human-like, in which case it was also perceived as more willing to help, more human-like, better able to understand the user's request, and preferred for future use over the system which only asked to wait. In the next chapters, we dive deeper into humans' impressions of time-buying systems as we try different strategies for selection of time-buying actions. We also experiment with participants being the users of the evaluated system, rather than mere overhearers.

Chapter 5

Experiment 2: Time-Buying in Spoken Dialogue Systems

5.1 Introduction

In the previous chapter, we discussed the importance of conversational time-buying in spoken dialogue systems. By “conversational time-buying”, we refer to the use of a variety of dialogue acts for the purpose of bridging time while searching for task-related information to convey. Based on the results of an experiment where participants listened to enacted phone conversations, we claimed that listeners prefer a system with such time-buying behavior over one which asks the user to wait and then remains silent until it can provide task-related information.

In this chapter, we attempt to find out more about this preference. In Chapter 2, we mentioned human speakers’ avoidance of long pauses in dialogue, which may lead us to wonder whether the preference in the experiment above stems from the conversational nature of time-buying in the preferred condition, or if this strategy was preferred merely because it contained less silence. If we were to choose one single dialogue act (such as *wait*) and “fill up silent time” by producing different utterances corresponding to that dialogue act, would this be enough? Or would we need a variety of dialogue acts similar to the one found in human time-buying for listeners to perceive this system as human-like and/or to prefer it? And if so, does variety *per se* suffice? Can we, for example, produce any time-buying action at any time, as long as we use a variety of actions overall? Or do we need to take other aspects into account, such as how long the

system has been buying time, and base the selection on this information? Therefore, the questions from the Introduction chapter that we seek to answer here are:

- Would humans find interacting with a system which can buy time more pleasant than interacting with a system which cannot do this, or which fills up waiting time differently (e.g. by explicitly asking the user to wait)?
- Which elements of human time-buying contribute the most towards achieving a satisfactory time-buying strategy for a spoken dialogue system?

In the next sections, we explore modeling conversational time-buying behavior in a spoken dialogue system, and we run an evaluation with human participants, this time not as overhearers but as users of the system.¹ We focus on modeling the variety of human time-buying as well as the way human speakers distribute time-buyers along the time-buying stretch.

To evaluate the system, we had participants interact with it and with two baseline systems. The first one bridges the gap between the user's request and presentation of a result by repeatedly asking the user to wait. The second system uses the same utterances as the one based on human behavior but selects them randomly, without considering any sequencing information. After each dialogue, participants were asked to rate the system with which they had just interacted. Our system was rated as more human-like and more enjoyable to interact with than the other systems. In addition, it was perceived as capable of finding a result in a more appropriate amount of time than the system which used explicit requests to wait, although the actual time elapsed before announcing a result was the same for all three systems.

Below, we describe the system and its evaluation. We begin by explaining the model that we trained to decide which time-buyer to use at each time (Section 5.2.1), and outlining the architecture of the system (Section 5.2.2). Afterwards, we present the experiment and its results in Section 5.3, followed by a discussion of the results in Section 5.4.

¹It must be noted, however, that the participants' use of the system was simulated, since we employed a Wizard-of-Oz setup: Participants believed the system could understand them, but its answers were actually triggered by a human pressing a key (see Section 5.3 for more details).

5.2 The system

5.2.1 Decision-making

Below, we describe the model that we trained to select time-buying actions. Due to technical problems, the strategy we tested was not exactly the same as the one we planned (a fact of which we became aware during post-experiment analysis). Therefore, in *Selection of high-level action (as planned)*, we explain the strategy that we intended to use, and in *Selection of high-level action (as executed)*, the one that was actually tested.

States and actions

The time-buying strategy in the experiment in Chapter 4 used only the four most frequent time-buying categories from the DSG-Travel Corpus. In this experiment, we intended to include all 11 —however, we kept only seven, due to several reasons:

- Utterances corresponding to the categories FILLER and INCOMPLETE were difficult to synthesize with the right prosody.
- Including category CONFIRMATION/EXPANSION REQUEST would have introduced the risk of the user producing new content which we could not handle within our Wizard-of-Oz setup (see Section 5.3).
- Finally, we merged category PARTIAL MATCH under TEMPORARY NON-AVAILABILITY, since we did not find enough variety of non-availability utterances in the corpus and the functions of both categories are relatively similar.

In addition, in order to further reduce the action and state spaces given the small size of the training data, we grouped these seven categories into two larger classes: *grounding* (see Section 2.3) and *task state*.

- Within *grounding* we included those actions which, besides buying time, help to increase common ground by acknowledging understanding, confirming uptake of the task, etc. (Clark and Brennan, 1991; Clark, 1996).

Action	Category	Example
PRODUCE GROUNDING UTTERANCE	acknowledgment	C: I want to fly to Bristol. A: <i>Okay.</i>
	commitment	<i>Let's have a look.</i>
	echoing	C: I'm looking for a flight to Izmir at the beginning of August. A: <i>A flight to Izmir, beginning of August, let me see...</i>
PRODUCE TASK STATE UTTERANCE	agent/system state	<i>The search for flights is still in progress.</i>
	availability	<i>We have a few choices to offer you. So far I only see evening flights.</i>
	temporary non-availability	<i>Until now I haven't found any morning flights.</i>
	wait request	<i>Please hold on.</i>

Table 5.1: Actions and utterance categories

- *task state* comprises any utterances which convey information about the state of the flight searching task, such as the suitability of the flights found so far, or the existence of delays in the search interface.

Table 5.1 shows the seven categories chosen and how they were grouped.

On the other hand, we wanted our system to resemble, to some extent, human speakers' pausing behavior. Therefore, we explicitly included *pausing* in the action space. The resulting space thus consisted of four actions: **produce grounding utterance** and **produce task state utterance** (as in Table 5.1) along with **produce long pause** and **produce short pause**. We labelled pauses shorter than 1200 ms. as SHORT PAUSE and those longer than 1200 ms. as LONG PAUSE, following Jefferson's (1989) observation of 1200 ms. as the approximate maximum duration of an unmarked pause in conversation.

As for the state space, the state variables were the last two actions produced by the system: a_{t-2}, a_{t-1} . Given the four actions available, this resulted in 16 possible states.

Selection of high-level action (as planned)

In each of the 16 possible dialogue states, one of four actions can be selected. For example, in the state $s = (a_{t-2} = \textit{grounding}, a_{t-1} = \textit{long pause})$, the system can

choose to produce any of four high-level actions: *grounding*, *task state*, *short pause*, *long pause*. To train a model that would enable the system to make this decision, we used OpenDial (Lison, 2015; Lison and Kennington, 2016).² OpenDial makes it possible to define a factored joint distribution (in the form of *probabilistic rules*), structured as sets of conditions (the states) together with the effects which may take place given those conditions (the actions). As an example, the rule for the condition $s = (a_{t-2} = \textit{grounding}, a_{t-1} = \textit{long pause})$ mentioned above is structured as follows:

```

if  $a_{t-2} == \textit{grounding}$  and  $a_{t-1} == \textit{long pause}$ :
    decision = grounding (util = theta_grounding)
    decision = task state (util = theta_task_state)
    decision = long pause (util = theta_long_pause)
    decision = short pause (util = theta_short_pause)

```

The aim of the training is to learn a function which accounts for the utility of selecting each action at each state. The four parameters starting with *theta* are the utility values which will be learned. The training data were the 92 time-buying stretches in the DSG-Corpus dialogues, structured as 801 sequences of actions (a_{t-2}, a_{t-1}, a_t) representing the speaker’s decision at time t and the two immediately previous decisions.

The initial prior of the utility function was modeled with a Gaussian distribution. OpenDial applies Bayesian learning to estimate the posterior distribution $P(\theta|\mathcal{D})$, where \mathcal{D} is the set of state-action pairs in the training data and θ represents the rule parameters. This distribution can be expressed as below, following Lison (2015):

$$P(\theta|\mathcal{D}) = \eta P(\theta) \prod_{\langle \mathcal{B}_i, a_i \rangle \in \mathcal{D}} P(a_i|\mathcal{B}_i; \theta)$$

where $P(a_i|\mathcal{B}_i; \theta)$ is the probability of action a_i being selected in the state \mathcal{B}_i with rule parameters θ , and η is a normalization factor. Thus, at each iteration, the parameters are updated as follows:

$$P(\theta_{(i+1)}) = \eta P(\theta_{(i)}) P(a_i|\mathcal{B}_i; \theta_{(i)})$$

It must be noted that OpenDial is normally used to train Partially Observable Markov Decision Process (POMDP) models, in which both a utility function and a probability function are learned. The probability function makes it possible to model

²<http://www.opardial-toolkit.net>

states where part of the information is not certain (Young et al., 2013). However, in our case, the states are made up of the last two system actions, and are thus fully observable. For this reason, our model is best described as a simple Markov Decision Process (MDP), in which the probability function is near-uniform, and the system's choices were thus controlled by the utility function.

Selection of high-level action (as executed)

The strategy described in the previous subsection relies on the two previous actions as context for the decision, resulting in a sequence of three actions a_{t-2}, a_{t-1}, a_t . However, in practice, an extra time-buyer was inserted between a_{t-1} and a_t due to a threading issue. As a result, if $a_{t-2} = \textit{grounding}$ and $a_{t-1} = \textit{task state}$, and the system's current decision was $a_t = \textit{long pause}$, the sequence produced should be

- *grounding - task state - long pause.*

Instead, if the extra time-buyer inserted due to the bug was *short pause*, the sequence produced ended up being

- *grounding - task state - short pause - long pause.*

In the next step, the system selects another action given the context as $a_{t-2} = \textit{task state}$ and $a_{t-1} = \textit{short pause}$, since it considers the context to be the last action of the previous context (a_{t-1} from the previous step) plus the action produced after it (in this case, the intruding action). If the selection it makes based on this context is $a_t = \textit{grounding}$, the cumulative sequence produced so far will be

- *grounding - task state - short pause - long pause - grounding*

and the system will then continue to make the next decision with $a_{t-2} = \textit{short pause}$ and $a_{t-1} = \textit{long pause}$.

Selection of specific time-buyer

At runtime, the system makes a decision in two steps:

	1	2	3	4	5	6
acknowledgment	0.67	0.05	0	0.04	0	0
commitment	0.2	0.23	0.03	0.29	0.43	1
echoing	0.13	0.72	0.97	0.67	0.57	0

Table 5.2: Frequency distributions for grounding utterance categories in the first six slots of the time-buying stretch

Step 1: One of the four high-level actions is selected, using the model described above. If the action is *produce short pause* or *produce long pause*, the system produces a pause of 2 or 4 seconds respectively.³

Step 2: If the high-level action is not a pause, a specific time-buyer for that category is selected: e.g. if the high-level action selected is *produce grounding utterance*, the system will select between ACKNOWLEDGMENT, COMMITMENT and ECHOING. In order to make this selection, it considers the frequency distribution, in the human-human data, of the available time-buying categories in the corresponding position in the interaction. For instance, if the decision received from the Action Selection module is *grounding* and the system has already produced three time-buying utterances, it will consider all the *grounding* utterances which appear in the data in the fourth position of the time-buying phase, together with their respective categories. Since the distribution for this position is *acknowledgment: 0.04, commitment: 0.29, echoing: 0.67*, the Utterance Selection module will sample from this distribution in order to select the next category (Tables 5.2 and 5.3 show all the frequency values used in the system). Due to the reduced size of the corpus, only the frequencies of the first six positions are considered: Starting from the seventh utterance, the module alternates between the probabilities for the fifth and sixth slots (if we only used the probabilities for the sixth slot, *grounding* actions would always be COMMITMENT, as can be seen in Table 5.2).

Resulting strategy

To summarize, the decision-making process consisted of two stages:

³We originally chose 500 and 3000 ms. as pause durations, keeping the length of the short pause under the 1200 ms. threshold proposed by Jefferson (1989), and that of the long pause above it. However, we perceived the resulting production as too rushed, which is why we extended pause durations to 2000 and 4000 ms. respectively.

	1	2	3	4	5	6
agent/system state	0.13	0.74	0.54	0.57	0.38	0.38
availability	0.25	0.13	0.31	0.1	0.19	0.38
temporary n/a	0.06	0	0	0.1	0.06	0.08
wait request	0.56	0.13	0.15	0.23	0.37	0.16

Table 5.3: Frequency distributions for task state utterance categories in the first six slots of the time-buying stretch

- a Selection of high-level category (using the trained model).
- b Selection of specific time-buying action (using unigram probabilities) for the high-level category chosen in *a*.

The technical issue mentioned above resulted in step *a* not working as expected. However, in step *b*, the specific time-buying action was still sampled based on probabilities from the corpus for that stage of the wait, as planned. As a result, we decided to shift the focus of our analysis: Whereas we originally intended to observe the impact of considering the recent context (the two previous time-buying actions) when choosing what to say, we now focus on the impact of basing the selection on the probabilities at the stage of the time-buying stretch when it takes place. In other words, if the system must choose the fourth time-buying action since it began to buy time, and it has selected *grounding* in step *a*, the action will be sampled from the distribution of grounding actions in the fourth slot of all the episodes in the corpus (and the same applies for *task state*). Given that in Chapter 3 we established that some time-buyers are more likely to occur at certain stages of the wait than at others (e.g. immediately after the interlocutor’s request vs. some seconds into the wait), we hypothesize that a strategy which accounts for the distribution of actions along the time-buying stretch in this way will be perceived as better than a strategy which does not consider this dimension.

5.2.2 Architecture

A full architecture for a task-oriented spoken dialogue system would require an Automatic Speech Recognition (ASR) component which receives speech input from the

user and produces a hypothesis of the user's utterance. It would also involve a component which can perform a search in a database with relevant information (in our scenario, a database of flights) and identify matches for the user's query. In addition, it would include a dialogue management component which selects the best action to execute based on the availability of matches, the dialogue history, the time elapsed since the beginning of the search, etc. Finally, it would require two more components: one for language generation and another one for speech synthesis.

For our experiment, time constraints made it impossible to develop a full system, so we simulated received queries through keyboard input, and we did not perform a real database search. The resulting simplified architecture is illustrated in Figure 5.1. The system was developed using InproTK_s, a toolkit for building incremental dialogue systems with a modular architecture (Kennington et al., 2014).⁴ InproTK_s is based on the IU model of incremental processing (Schlangen and Skantze, 2011), and it allows construction of dialogue systems conceived as a series of interconnected modules. These modules have a *left buffer* and a *right buffer*. Each module receives data packaged as IUs (Incremental Units) through its left buffer, processes the data in some way, and then sends the output through its right buffer into the left buffer of the next module. In the case of our system, as can be seen in Figure 5.1, the Environment Module receives an IU coming from keyboard input through its left buffer, processes this input and sends the output into the left buffer of the Action Selection Module, etc.

Environment Module:

This module has two functions:

- It receives keyboard input from the Wizard. The Wizard can press different keys depending on whether the query is complete or whether the participant forgot to mention any of the search criteria. The Environment Module checks which key was pressed and forwards the corresponding IU to the Action Selection module. If the “query is complete” key has been pressed, the IU sent will instruct the Action Selection module to start buying time. Otherwise, it will tell it to ask a clarification question to find out the missing information (e.g. departure airport, airline, destination city, etc).

⁴<https://bitbucket.org/inpro/inprotk>

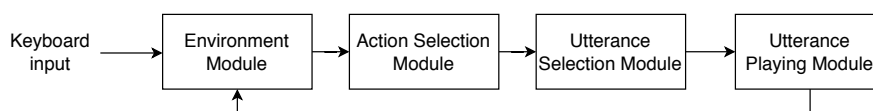


Figure 5.1: System architecture

- Twenty seconds after time-buying has started, the Environment Module sends a timeout signal to the Action Selection module, so that the system announces that it has found a flight.

Action Selection Module:

This module selects one of the time-buying actions available to the system. This selection is based on the learned strategy explained in Section 5.2.1. The decision is then forwarded to the Utterance Selection Module. After the corresponding utterance has been played, the Action Selection module chooses a new time-buying action, and this process continues until 20 seconds have passed since the beginning of the time-buying phase (at which point the system announces that it has found a matching flight). We chose 20 seconds as the duration of the time-buying stretch because this is close to the average duration of the time-buying stretches in the human-human corpus (17.5 seconds), and also long enough to showcase the differences between strategies in the experiment described below (see Section 5.3).

Utterance Selection Module

This module has two main functions. The first one is choosing a time-buying category based on the high-level decision received from the Action Selection module, following the mechanism described in Section 5.2.1. For example, if the decision received is *grounding*, it will choose between *acknowledgment*, *commitment* and *echoing*; otherwise, if the decision received is *task state*, the choice will be between *agent/system state*, *availability*, *temporary non-availability* and *wait request* (See Table 5.1).

Once a category has been selected, the second task of the Utterance Selection module is to choose a specific utterance to send to the Utterance Playing module. Some of the utterances were taken from the human-human corpus, and others were based on it and modified as needed (see Table 5.4 for the full list). The module looks at the

four utterances for the selected category and chooses the first one that has not been used yet (if all four have been used, the selection starts again at the beginning of the list). Finally, the utterance is forwarded to the Utterance Playing module, which plays an audio file with the synthesized utterance.⁵ If the decision received from the Action Selection module is not an utterance but a pause, the task of the Utterance Selection module is simply limited to forwarding this decision to the Utterance Playing module.

In summary, to make a decision about a particular act at a given point, the system first checks whether information can already be presented (Environment Module); if not, it selects a high-level act based on the learned strategy (Action Selection module) and, based on that, an actual utterance (Utterance Selection module), which it then realizes (Utterance Playing module). The division of the decision-making process between the Action Selection and Utterance Selection modules was due to the reduced size of the training data: Grouping all utterances into two broad categories in the Action Selection module (*grounding* and *task state*) and refining the decision in the Utterance Selection module made it possible to keep the state space smaller for learning the parameters of the action selection rules (see Section 5.2.1).

5.3 Experiment

In order to evaluate the strategy described above without developing a full dialogue system, we employed a *Wizard-of-Oz* environment. This is a widespread technique in Human-Computer Interaction and related fields, in which a human (the “Wizard”) controls all or part of the system’s responses, usually without the participant’s knowledge (Kelley, 1984; Dahlbäck et al., 1993).⁶ In our case, the Wizard was a student assistant whose task was to press a key whenever she judged that the participant’s request was complete. The system then acknowledged the request by producing *mm-hm* and started to buy time. Participants were told that they interacted with a fully automated system. The Wizard could also trigger clarification requests when the participant for-

⁵InproTK_s is integrated with the speech synthesizer MaryTTS to offer real-time, incremental speech synthesis. However, we chose to play canned utterances in audio files to be able to use the Cereproc voice instead, as we obtained more interesting results with this voice than with a MaryTTS voice in the experiment described in Chapter 4.

⁶This technique was originally named “the Oz Paradigm”, and the term “Wizard” came to be associated with it later on (Kelley, 1984).

got to mention any of the search criteria or the request had not been expressed clearly. Below are the questions asked for each criterion:

origin *Haben Sie eine Präferenz des Abflughafens?* (“Do you have a preferred departure airport?”)

airline *Könnten Sie bitte die Fluglinie wiederholen* (“Could you please repeat the airline?”)

city *Könnten Sie bitte nochmal sagen, wohin Sie fliegen möchten?* (“Could you please repeat where you want to fly to?”)

time *Um wie viel Uhr würden Sie am liebsten fliegen?* (“At what time would you prefer to fly?”)

date *Könnten Sie bitte nochmal das Datum sagen?* (“Could you please repeat the date?”)

This was an important consideration since, as mentioned above, the system’s utterances were canned. Therefore, if an instance of ECHOING was produced for a criterion that the participant had not mentioned, the latter might realize that the system was not taking speech input into account.

Below we provide details about the experiment design, procedure and participants.

5.3.1 Method

Design

There were three experimental conditions: LEARNED, RANDOM and FIXED. The difference between the conditions was the strategy used by the system to bridge the gap between the user’s request and the moment when it announces finding a flight (see Figure 5.2 for examples):

FIXED The system bridges the gap by explicitly asking the user to wait, through utterances such as *please wait; one moment, please; give me a second*, etc.⁷ The utterances are separated by four-second intervals.

⁷Note that this is different from the *wait* condition in Chapter 4, which only asked for waiting once and then remained silent.

Category	Utterance	English translation
acknowledgment	<i>Gut.</i> <i>Okay.</i> <i>Sehr gern.</i>	<i>Good.</i> <i>Okay.</i> <i>Gladly.</i>
commitment	<i>Ich schaue gerade mal in meine Liste.</i> <i>Da gucken wir doch mal.</i> <i>Schaue ich gerade einmal nach.</i> <i>Ich muss mal gucken.</i> <i>Schauen wir doch mal.</i>	<i>I'm looking in my list.</i> <i>Let's have a look.</i> <i>I'm having a look.</i> <i>I need to check.</i> <i>Let's have a look.</i>
echoing	<i>Ab Düsseldorf.</i> <i>Nach Rom.</i> <i>Im Juni.</i> <i>Nachmittags.</i>	<i>From Düsseldorf.</i> <i>To Rome.</i> <i>In June.</i> <i>In the afternoon.</i>
agent/system state	<i>Die Flüge werden noch gesucht.</i> <i>Ich warte noch auf die Liste.</i> <i>Die Flüge kommen langsam rein.</i> <i>Ach, mein System braucht gerade ein wenig länger.</i> <i>Ich muss hier eben auf die Daten warten.</i>	<i>The flights are still being searched.</i> <i>I'm still waiting for the list.</i> <i>The flights are slowly appearing.</i> <i>Argh, my system needs a bit more time.</i> <i>I need to wait for the data.</i>
availability	<i>Da haben wir was im Angebot.</i> <i>Zur Verfügung stehen verschiedene Flüge.</i> <i>Ich habe hier ein paar Möglichkeiten.</i> <i>Da haben wir einige Flüge für Sie.</i> <i>Es gibt hier eine ziemlich grosse Auswahl.</i>	<i>Here we have something to offer you.</i> <i>There are various flights available.</i> <i>Here I have a couple of options.</i> <i>We have a few flights for you.</i> <i>Here we have quite a large selection.</i>
temporary non-availability	<i>Ich sehe gerade nichts ab Düsseldorf.</i> <i>Bisher habe ich nur ab Köln-Bonn.</i> <i>Im Moment sehe ich gar nichts nachmittags.</i> <i>Bisher nur vormittags</i>	<i>I don't see anything from Dusseldorf.</i> <i>Until now I only have from Cologne-Bonn.</i> <i>At the moment I don't see anything in the afternoon.</i> <i>Until now only in the morning.</i>
wait request	<i>Einen kleinen Moment, bitte.</i> <i>Sekunde noch.</i> <i>Einen kleinen Moment.</i> <i>Warten Sie bitte noch einen Augenblick.</i> <i>Ich bitte um ein wenig Geduld.</i>	<i>One moment, please.</i> <i>One second.</i> <i>One moment.</i> <i>Please wait a little longer.</i> <i>Please bear with me.</i>

Table 5.4: System utterances. For ECHOING and TEMPORARY NON-AVAILABILITY, we include a set of examples, as the actual utterance depends on the search parameters for that episode.

PARTICIPANT: Ich würde gerne von Frankfurt nach Sydney fliegen, am 3. August und zwar vormittags.
I'd like to fly from Frankfurt to Sydney, on August 3 in the morning.

SYSTEM (FIXED):

Mm-hm	einen kleinen Moment		warten Sie bitte noch einen Augenblick		Sekunde noch		Augenblick, bitte		Ich habe einen passenden Flug gefunden...
Mm-hm	one moment, please		please wait a little longer		one more second		one moment, please		I have found a matching flight.

SYSTEM (RANDOM):

Mm-hm	vormittags	da haben wir was im Angebot	da gucken wir doch mal		okay	schaue ich gerade einmal nach		Ich habe einen passenden Flug gefunden...
Mm-hm	in the morning	we have something to offer you	let's see		okay	I'm having a look		I have found a matching flight.

SYSTEM (LEARNED):

Mm-hm	einen kleinen Moment, bitte	nach Sydney	am 3. August	Sekunde noch	ich schaue gerade mal in meine Liste		Ich habe einen passenden Flug gefunden...
Mm-hm	one moment, please	to Sydney	on August 3	one more second	I'm having a look in my list		I have found a matching flight.

Figure 5.2: Examples of the three time-buying strategies employed by the system (original utterances in German in bold; English translation provided below). The gray intervals represent pauses: The wider ones last four seconds and the narrower ones, two seconds.

RANDOM The system bridges the gap by randomly selecting from a set of utterances similar to those found in the DSG-Corpus. In between utterances, the system can also randomly choose to produce a four-second pause, a two-second pause or no pause at all. The system's utterances are displayed in Table 5.4.

LEARNED The system employs the learned strategy described in 5.2.1. The utterances are the same as in the RANDOM strategy.

Participants were presented with each of these conditions four times, in random order.

Participants

Thirty participants were involved in the study, 19 female and 11 male, recruited through flyers left at the cafeteria in Bielefeld University, by email or on the university Facebook group.

Procedure

Each participant played the role of a secretary at a company, who had the assignment of calling a travel agency to book flights for some executives. Participants were told that they would be speaking to an automatic system which could understand speech, and they received a handout with a list of items. Each item contained the criteria defining a flight that the participant should request, e.g. *Frankfurt-Sydney, May 24, Lufthansa*. The structure of each call was as follows:

1. First, the system greeted the participant.
2. After the greeting, the participant asked for one of the flights on the list.
3. Following this request, the Wizard pressed a key for the system to produce an acknowledgment (*mm-hm*), signaling reception of the participant's request, after which it started buying time (if the request was incomplete, the Wizard triggered a clarification request instead, and only produced the acknowledgment once all the search parameters had been mentioned).
4. After 20 seconds, the system announced having found a flight and told the participant that the flight details would be sent to the company by email.⁸
5. Finally, if the participant said "goodbye", the Wizard pressed a key for the system to say "goodbye" as well.

After every call, participants were given some time to rate the system. For each of the statements below, they chose an option from 1 (completely disagree) to 5 (completely agree):

- It was pleasant to interact with this system.
- The system provided an answer within an appropriate amount of time.
- The system acts the way I would expect a person to act.

⁸We asked participants to pretend that the system already had the contact details of the company, the latter being a frequent customer.

These statements were inspired by the ones in the experiment in Chapter 4. We removed the ones related to perceived willingness to help and ability to understand, in order to focus only on perceived human-likeness and appropriateness of waiting time. We also removed the question on willingness to interact with the system again and asked about pleasantness of the interaction instead. After these statements, there was an optional field for further comments.

Once the participant had completed the assessment, the next call started, with the system greeting the customer as before. Each participant completed 14 calls: 2 test calls for making sure they had understood the instructions and 12 experiment calls. Participants were instructed to include only one flight per call.

5.3.2 Results

We compared the ratings given to each of the strategies (FIXED, RANDOM and LEARNED) for each of the three statements rated. We obtained 120 sets of ratings per strategy, as participants were exposed to each strategy four times (30 participants x 4 repetitions). As there were three statements to rate for each stimulus, the total number of individual ratings per strategy was 360. Here, as in Chapter 4 above, we report median as measure of central tendency (instead of mean) and interquartile range for dispersion (instead of standard deviation), given that Likert data are ordinal and thus not normally distributed. We tested significance of differences through the Wilcoxon signed-rank test and applied Bonferroni correction due to the multiplicity of statements per stimulus, which resulted in the following significance levels: $.05/3 = .017$; $.01/3 = .003$; $.001/3 = .0003$.

No significant differences were found between the FIXED and RANDOM strategies. In contrast, the LEARNED strategy was rated significantly better than the FIXED strategy for all three statements ($W=356, p<.0003$; $W=475, p<.003$ and $W=800, p<.0003$ respectively). Additionally, LEARNED was rated significantly better than RANDOM for statement 1, *It was pleasant to interact with this system* ($W=652, p<.017$), and 3, *The system acts the way I would expect a person to act* ($W=904, p<.0003$). Figure 5.3 shows the total sum of the ratings assigned to each condition in each statement, as well as the median score and the interquartile range.

St.	FIXED (total sum)	FIXED (median)	FIXED (IQR)	RANDOM (total sum)	RANDOM (median)	RANDOM (IQR)	LEARNED (total sum)	LEARNED (median)	LEARNED (IQR)
1	427	4	1	452	4	1	486	4	1
2	460	4	2	471	4	2	496	4	1
3	376	3	1	402	3	1	456	4	2

Figure 5.3: Ratings received by each strategy, by statement (column *St.*). Statements are 1) *It was pleasant to interact with this system*, 2) *The system provided an answer within an appropriate amount of time*, 3) *The system acts the way I would expect a person to act*. IQR stands for interquartile range. (* $p < .017$, ** $p < .003$, *** $p < .0003$)

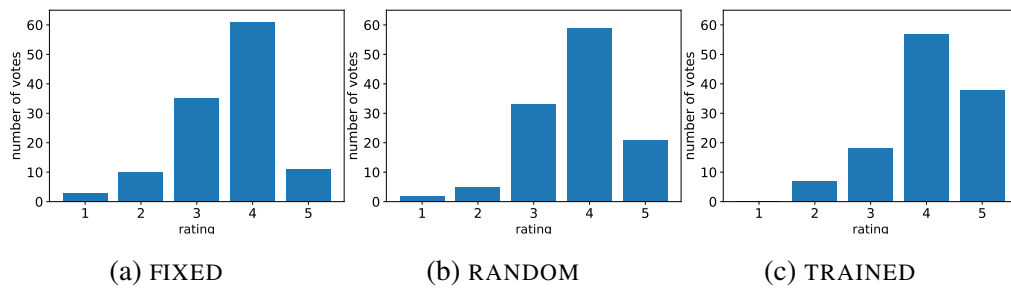


Figure 5.4: Distribution of ratings for statement 1, for each condition

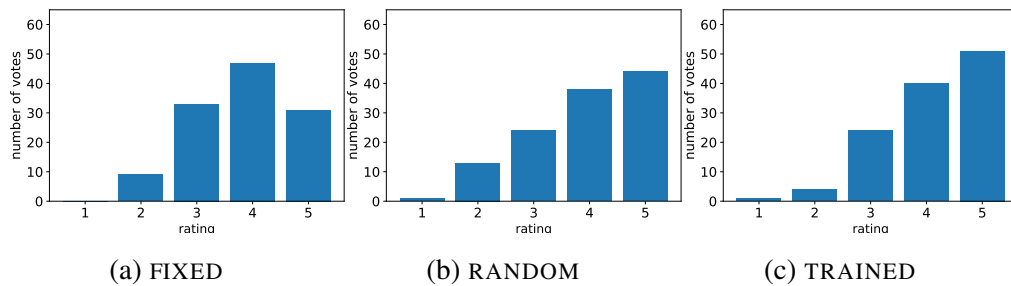


Figure 5.5: Distribution of ratings for statement 2, for each condition

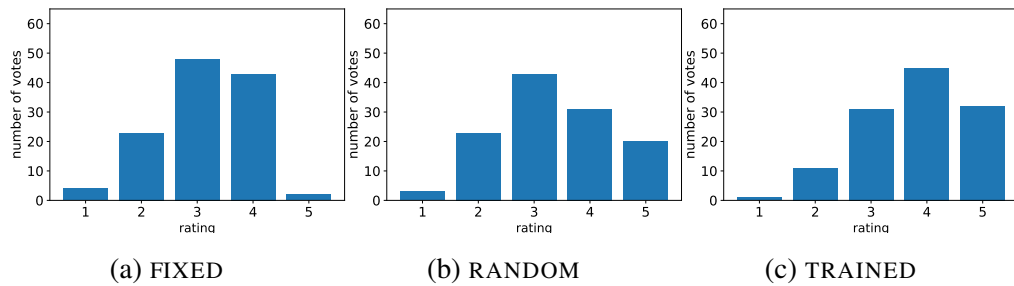


Figure 5.6: Distribution of ratings for statement 3, for each condition

5.4 Discussion

In this experiment, we tested three time-bridging strategies involving speech, with a view to identifying the characteristics that this speech must have in order to render the interaction human-like and pleasant. In particular, we focused on two aspects: *variety* and *distribution along the time-buying stretch*. By *variety*, we refer to the use of time-buyers corresponding to several dialogue acts —as opposed to, for example, asking the interlocutor once and again to wait. By *distribution along the time-buying stretch*, we mean recreating the frequency with which these dialogue acts occur in human dialogue at different moments of the time-buying stretch: right after the request for information is received, some time into the search, etc.

In the FIXED condition, no attention is paid to either of these aspects, since all utterances realize the same dialogue act (requesting extra time). The RANDOM condition includes a variety of utterances representing different dialogue acts, but their distribution along the time-buying stretch is random. Finally, the LEARNED condition considers knowledge about both the variety observed in human time-buying strategies and their distribution along the time-buying stretch. Our results show that considering both aspects (variety and distribution along the time-buying stretch) when developing a time-buying strategy for a system leads to a better user experience, since the LEARNED condition received higher ratings than the other two strategies.

This leads to an obvious question: What is the decisive factor? Is it variety, distribution along the time-buying stretch, or both? It is clear that modeling distribution along the time-buying stretch contributed to the improvement in user experience, given that the *only* strategy that accounted for this aspect (the LEARNED strategy) was the preferred one. On the other hand, the role of variety of time-buying acts is less clear.

At first sight, the results seem to suggest that variety did not play a role: The RANDOM strategy used a variety of time-buying acts, yet it was not significantly preferred over the FIXED strategy. This comes as a surprise, as we had expected that a system producing a variety of utterances—even if randomly selected—would be rated higher than one which can only ask the interlocutor to wait (especially since such explicit waiting requests are relatively uncommon in human dialogue, as shown in Chapter 3). An alternative explanation could be related to the coherence of the information provided in different utterances: The RANDOM strategy could, for example, select an utterance from the *availability* category, such as *We have several flights to offer you*, immediately followed by one from the *temporary non-availability* category, such as *Right now we don't have any flights available from [city of origin]*. Such contradictory statements may have penalized the RANDOM strategy, whereas the FIXED strategy did not run the risk of contradicting itself, since all it did was ask the interlocutor to wait. It is possible, therefore, that this may have compensated for the better experience provided when including variety, which in turn may have led to a lack of significant differences between both systems.

While pleasantness and human-likeness were rated higher for the LEARNED strategy than for all other conditions, there was no significant difference between the LEARNED and RANDOM conditions for statement 2: *The system provided an answer within an appropriate amount of time*. A possible explanation is that repetition of the same dialogue act in the FIXED condition might have led to users' annoyance and, consequently, to a perception of waiting time as longer, something which did not happen in the other two conditions, in which moves were more varied and potentially more “entertaining”, thus causing a similar impression in terms of time elapsed. In addition, in the fixed condition, users were continuously reminded of the fact that they needed to continue waiting and, as discussed in Section 4.2, lengthy waits can result in feelings of frustration, whereas the other two strategies provided other kinds of information as well.

In the face of these results, we attempt to provide an answer to the questions in the Introduction section of this chapter.

- Would humans find interacting with a system which can buy time more pleasant than interacting with a system which cannot do this, or which fills up waiting time differently (e.g. by explicitly asking the user to wait)?

In our experiment, a system which buys time conversationally, exploiting a range of utterances similar to the one found in our human-human corpus and leveraging frequency information from the same corpus to select them was rated more pleasant to interact with, more human-like, and able to find a flight in a more appropriate amount of time than a system which fills up silent time by repeatedly asking the user to wait.

- Which elements of human time-buying contribute the most towards achieving a satisfactory time-buying strategy for a spoken dialogue system?

Our results suggest that modeling both the variety found in human time-buying behavior as well as the way humans distribute these dialogue acts along the time-buying stretch leads to a perception of increased human-likeness and pleasantness when compared to a system which only asks users to wait. In addition, variety of time-buying dialogue acts may also contribute to perception of waiting time as more appropriate, although further data would be needed to support this claim.

It must be noted that our initial intention was to model not only the distribution of time-buyers along the waiting stretch, but also their sequencing —i.e. which time-buyer is more likely to occur given the actions produced immediately before. In reality, the strategy that we tested only recreated the former aspect, since the sequencing of the utterances got distorted as a result of the bug described in Section 5.2. However, as mentioned above, the results of our experiment show a clear preference for a strategy which employs information from human data to sample time-buying actions based on their location in the time-buying stretch. Therefore, all things considered, we believe that the tested system sufficiently represents the general ideas in our research questions, and thus allows us to shed light on some of the aspects relevant to this work.

5.5 Summary

In this chapter, we tested three strategies for task-oriented spoken dialogue systems to bridge the time between the user’s request and the moment when the system can present a result. We found that a system whose time-buying behavior is modeled from human data (focusing on the range of time-buying acts used as well as how they are distributed along the time-buying stretch) was rated as more pleasant and human-like than both a system which uses a variety of utterances inspired from human data but

selects them randomly and a system which only asks the user to wait. Furthermore, the system modeled from human data was perceived as able to find a result in a more appropriate amount of time than the system which asked the user to wait, although the actual time needed by both systems was the same. In the next chapter, we compare different ways of modeling the sequencing of time-buying acts in human dialogue, and assess how humans perceive these strategies in a dialogue system.

Chapter 6

Experiment 3: Testing Strategies for Modeling Time-Buying

6.1 Introduction

In this chapter, we explore the question of how a system can combine a series of time-buying actions to sound as human-like as possible and offer a satisfactory experience to its human interlocutor. We trained two time-buying strategies and compared them with a random baseline. Both trained strategies are essentially n-gram models: One of them is a pure trigram model, and the other one is a trigram model enriched with information on number of time-buying actions produced after the human's turn. We refer to the former as *flat model* or *flat strategy*, and to the latter as *hierarchical model* or *hierarchical strategy*, as the decision is made in two steps and one of these steps is dependent on the other. We present these strategies (Section 6.2), and evaluate them intrinsically, by calculating the perplexity of each model (Section 6.3) and extrinsically, in an experiment in which participants listened to time-buying sequences generated with each model and rated them (Section 6.4). We discuss the results of the evaluations and propose further work in Section 6.5.

6.2 Strategies

We start from the following list of actions:

$A = \{\textit{temporary non-availability, greeting, response to caller intervention, filler,}$

agent/system state, commitment, echoing, confirmation/expansion request, acknowledgment, availability, wait }

Note that GREETING and RESPONSE TO CALLER INTERVENTION are not part of the taxonomy presented in Chapter 3, as they are direct responses to specific contributions by the other dialogue participant (in our data collection scenario, the customer). We include them here to enable a more comprehensive model of speakers' dialogue behavior while they wait to be able to provide task information. We also added LONG PAUSE and SHORT PAUSE to the list of possible actions the system can select at any point. Therefore, for modeling purposes, we considered 13 actions (for experiment purposes, as in previous chapters, we only used a subset of them: See Section 6.4.1). We compared three ways in which a system can select actions from this set. These strategies, *random*, *flat* and *hierarchical*, are described below.

6.2.1 Random strategy

Our baseline strategy rests on the assumption that the main consideration when buying time is to produce *some* content to prevent silent intervals from becoming too long —*which* content is produced does not matter so much. Therefore, in this strategy, each action is selected randomly from the above list, without considering the actions produced before, or the time elapsed since the time-buying started. If V is the number of actions available for the system to select, the probability of any action to be selected at any given moment is $1/V$ (or $1/13$ in our case, 13 being the number of possible actions).

6.2.2 Flat strategy

It is possible that a specific time-buying action might sound more or less natural depending on what was said immediately before. In order to take the recent context into account, we used a trigram model: Given the small size of our corpus, a larger value of N would have led to data sparsity, and a bigram model would have reduced the context too much —sometimes to just a pause. For the trigram model, the decision on which time-buying action to select at any time depends on the two preceding actions, therefore the probability for each action to be selected at a certain point in time t is

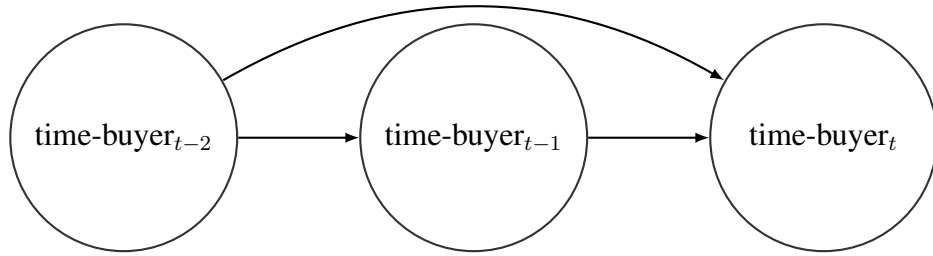


Figure 6.1: Selection of time-buyer in the flat strategy.

calculated as follows:

$$P(a_t | a_{t-2}, a_{t-1}) = \frac{C(a_{t-2}, a_{t-1}, a_t)}{C(a_{t-2}, a_{t-1})}$$

where a is an element from set A in Section 6.2 above. This is illustrated in Figure 6.1. At each step, the model considers the two preceding actions and makes a selection by sampling from a probability distribution over all possible time-buyers.

6.2.3 Hierarchical strategy

The flat strategy described above (like all n-gram models) is based on the Markov assumption, i.e. the idea that the most recent context is enough to predict the next decision (Jurafsky and Martin, 2018; Sutton and Barto, 1998). However, the data presented in Chapter 3 show that the probability of some types of time-buying actions varies based on the time elapsed since the speaker started buying time. As an example, ACKNOWLEDGMENT instances occur most frequently at the beginning, whereas ECHOING instances become more numerous as third and fourth time-buying actions in the sequence. For this reason, we tried a variation of the flat strategy in which the action is selected in two steps, similar to the strategy used in the previous chapter and described in Section 5.2.1. The second step considers how many time-buyers have been produced so far, and uses unigram probabilities to inform the selection. These probabilities represent the choices found in our corpus in the same slot, e.g. as first/second/third time-buyer in the sequence, etc.

We sorted the available time-buying actions into three broader categories: the two categories used in the previous experiment and a new one, *interaction*. Below is the list of categories, together with the subcategories they comprise:

grounding: greeting, acknowledgment, commitment, echoing

task state: agent/system state, availability, temporary n/a, wait, filler

interaction: confirmation/expansion request, response to caller intervention

long pause

short pause

Within *grounding* we included those actions which, besides buying time, help to increase common ground by acknowledging understanding, confirming uptake of the task, etc. (Clark and Brennan, 1991). *Task state* comprises any utterances which convey information about the state of the flight-searching task, such as the suitability of the flights found so far, or the existence of delays in the search interface. *Interaction* refers to actions in which the travel agent engages in dialogue with the customer, either trying to elicit information about the request or responding to a new contribution. For the experiment in the previous chapter, we removed instances of interaction before training the model. For the current experiment, we decided to maintain them during training, so as to assess the perplexity of each model with the full range of options, and subsequently remove them for the experiment with humans.

The selection of a time-buying action is done as follows:

Step 1: The model chooses between *grounding*, *task state*, *interaction*, *long pause* and *short pause*. For this purpose, we trained a trigram model similar to the one in the flat strategy, but using these five broader categories instead of the specific ones. The selection is made by sampling from a probability distribution based on the two previous time-buying categories.

Step 2: Once a broad category has been selected:

- If the selection is *long pause* or *short pause*, this is the final action produced.
- If the selection is *grounding*, *task state* or *interaction*, a subcategory needs to be selected—for example, if the chosen category is *grounding*, it is necessary to select between *greeting*, *acknowledgment*, *commitment* and *echoing*. For this purpose, just like in the TRAINED strategy in Chapter 5, we

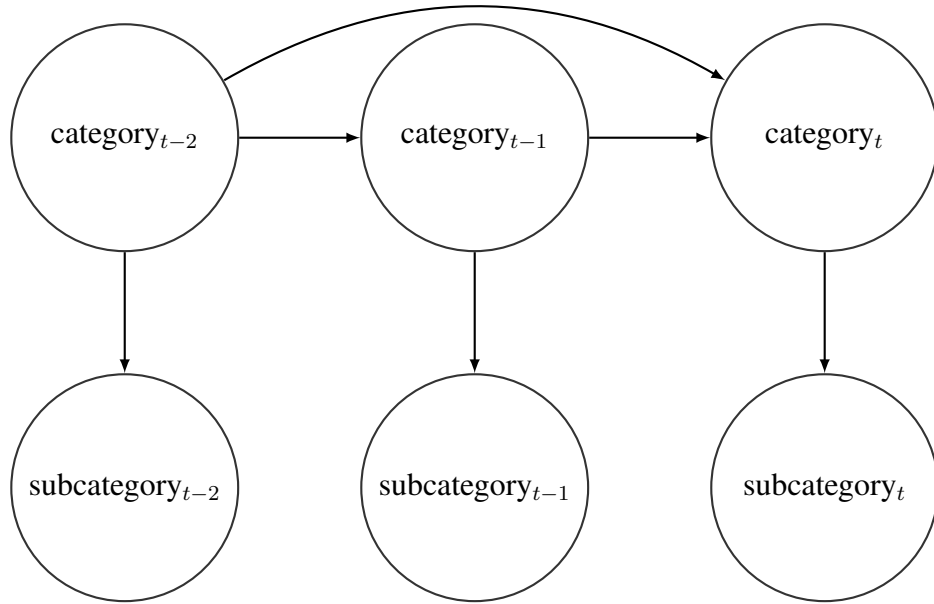


Figure 6.2: Selection of time-buyer in the hierarchical strategy.

take into account the number of actions involving speech (i.e. not pauses) which were produced since we started buying time. This is illustrated in Figures 6.3a and 6.3b. Here, aside from pauses, two time-buying actions have been produced, so we can consider the current decision to be in slot 3. If the category selected is *grounding*, we produce a probability distribution based on all the grounding acts which occur in slot 3 in the corpus. Thus, the probability for each subcategory of grounding at this particular moment will be:

$$P(sub) = \frac{C(sub_{slot3})}{C(cat_{slot3})}$$

where *sub* is the subcategory, and *cat* is the broader category which has been selected (in this example, *grounding*).

In summary, the probability of any given time-buyer being selected at a certain point in time t is represented as follows:

$$P(sub_t | sub_{t-2}, sub_{t-1}, cat_{t-2}, cat_{t-1}, cat_t) = P(sub_t | cat_t) * P(cat_t | cat_{t-2}, cat_{t-1})$$

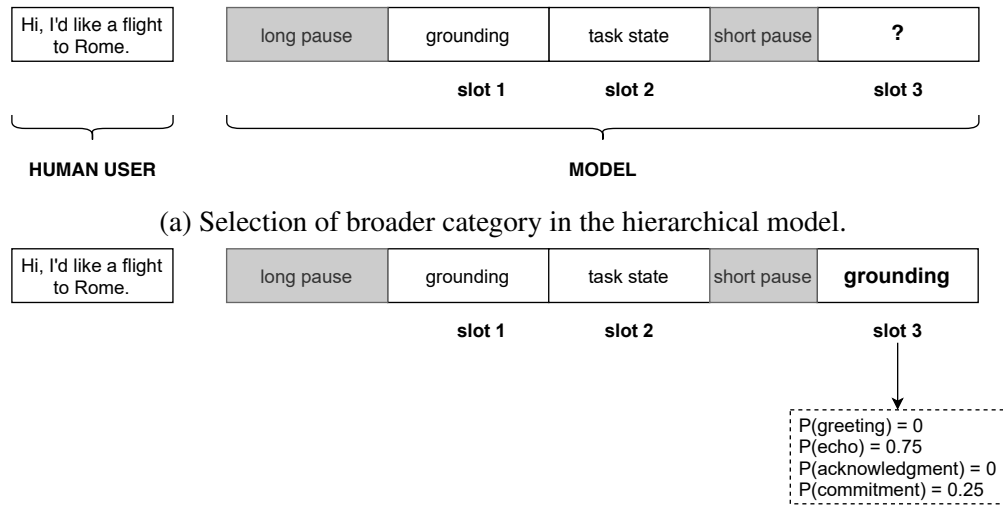


Figure 6.3: Action selection in the hierarchical model.

Figures 6.2 and 6.3 illustrate this process.

6.3 Intrinsic evaluation: Perplexity

We calculated the perplexity of each of the three models. Perplexity is a metric commonly used to compare language models, and it is defined as the inverse probability of a test set, normalized by the number of words. It is calculated as follows, for a trigram model and given a test set of actions $A = a_1 a_2 \dots a_N$ (Jurafsky and Martin, 2018):¹

$$PP(A) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(a_i | a_{i-2}, a_{i-1})}}$$

The test set consisted of 10 episodes —one from each participant— selected randomly (and excluded from the training). It contained 119 actions, which amounts to 10% of the data approximately. In order to avoid zero probabilities, we applied Laplace

¹Jurafsky and Martin (2018) present the formula for bigram models; here we have adjusted it for application to trigrams. Also, authors often use W instead of A and w instead of a , since perplexity is frequently used for scenarios where each action is the selection of a word.

(add-k) smoothing with $k = 0.1$ during probability calculation. For a bigram model, add-k smoothing is applied as follows (Jurafsky and Martin, 2018):

$$P_{Add-k}(a_t|a_{t-1}) = \frac{C(a_{t-1}, a_t) + k}{C(a_{t-1}) + k * V}$$

Analogously, for our trigram model, we calculated:

$$P_{Add-k}(a_t|a_{t-2}, a_{t-1}) = \frac{C(a_{t-2}, a_{t-1}, a_t) + k}{C(a_{t-2}, a_{t-1}) + k * V}$$

For the hierarchical model, we applied smoothing twice for each action, namely at the two levels at which decisions are made: selection of general category (such as *grounding*) and selection of specific action (such as ACKNOWLEDGMENT). Probability calculation for the hierarchical model resembles the Viterbi algorithm, since the probability of reaching the previous state is multiplied with the probability of transitioning from the previous state to the current one, and with the likelihood of observing the selected subcategory given the selected category (Jurafsky and Martin, 2018, p. 155). It differs from Viterbi in that we calculate only one path and we do not need to back-trace, as the sequence produced is fully observable (i.e. we are not searching for the most likely sequence but calculating the probability of the *actual* sequence).

Table 6.1 shows the results obtained. Both trained models were better than the random strategy, which almost doubled the perplexity of the former. On the other hand, the perplexity of the flat model was slightly lower than that of the hierarchical one. This hints at the idea that accounting for the time elapsed since the speaker started buying time does not bring about an advantage over a strategy which uses only the most recent context to make a decision.

In order to know whether these findings are also reflected in human listeners' perceptions, we carried out an overhearer experiment, described in the next section.

RANDOM	FLAT	HIERARCHICAL
13	6.82	6.87

Table 6.1: Perplexity of different time-buying strategies: one in which actions are selected randomly, and the two trained models described in Sections 6.2.2 and 6.2.3, with add-k smoothing ($k=0.1$)

6.4 Extrinsic evaluation: Experiment

6.4.1 Method

Design

We compared three conditions: RANDOM, FLAT and HIERARCHICAL. There were four audio clips available for each condition, and each participant was presented with two of the four, selected randomly. In total, every participant listened to six audio clips (2 x 3 conditions). The order of the conditions was also randomized.

Materials

The survey was hosted on Soscisurvey, the platform used for the experiment in Chapter 4.² We synthesized the utterances using Cereproc voice Alex, the male German unit selection voice used for the experiments in previous chapters.³ The full audio clips, as presented to the participants, consisted of a human customer's request for a flight, followed by the system buying time and finally announcing that it found a match. While starting to give the details of the match, the voice fades out and the recording ends before any flight information can be heard. We did this to ensure that the participants would not rate the samples based on the flight offered. The initial part of the audio clips, in which the customer requests the flight, is also the same for all samples — the only part differing across samples is the time-buying section, as this is the part on which we wanted participants to concentrate. After each clip, we presented the following questions (here in translation) for testers to rate between 1 (*strongly disagree*) and 5 (*strongly agree*):

1. The system behaves in a human-like way.
2. The system is intelligent.
3. The system found a flight quickly enough.
4. I would use this system.

²www.soscisurvey.de

³www.cereproc.com

The statements above were inspired by the ones we used in Chapter 4, as well as by Skantze and Hjalmarsson (2013), where participants assessed systems on eight aspects. We took the four aspects that were relevant to our purposes, *human-like*, *intelligent*, *faster response* and *preferred*, and embedded them in statements.⁴⁵ The visual layout of the questionnaire was the same as in Chapter 4 (see Appendix B.1) except for the modified statements and some minor adjustments in the instructions to improve clarity.

The customers' utterances were taken from the DSG-Travel corpus, described in Chapter 3. Since the model focused on dialogue acts and did not make any decisions at the lexical level, we excluded some utterances which were appropriate only in a specific part of the dialogue: For example, a travel agent would say *sehr gern* ("of course", literally "very gladly") mostly after the customer has spoken, in order to signal acknowledgment, but would be unlikely to say it later in the interaction (unless the customer makes a new contribution). However, our model does not handle this type of information. Therefore, we only used utterances which were somewhat flexible in terms of where in the time-buying stretch they might occur. For the same reason, we excluded the GREETING category, since the speakers in our data often used greetings after acknowledgments such as "ja" (yes), but not after acknowledgments such as "sehr gern".

As in Chapter 5, we also removed *filler* (since it is difficult to synthesize fillers with the right prosody) and all categories subsumed under *interaction* (to avoid needing to simulate an exchange between the system and the human during time-buying). The resulting list of actions for the experiment was the same as in Chapter 5: *temporary non-availability*, *agent/system state*, *commitment*, *echoing*, *acknowledgment*, *availability*, *wait*, *short pause*, *long pause*. The dialogues are shown in Appendix D.1.

We wanted to keep the duration of the samples homogeneous to be able to com-

⁴In Skantze and Hjalmarsson (2013), participants used a slider to compare two systems, one positioned on the left and one on the right, so they were only presented with isolated words or phrases. In our evaluation, testers rated each system individually, and there was no slider but discrete options between "strongly agree" and "strongly disagree", therefore we framed the dimensions as full statements.

⁵We initially intended to reuse a standard questionnaire from the field of Human Computer Interaction but, for each option we considered, only one, or at most two items were suitable for our purposes (Horvath and Greenberg, 1989; Harms and Biocca, 2004; Bartneck et al., 2009; Bergmann et al., 2012; Shamekhi et al., 2016; Fitriani and Richards, 2019). The dimensions in Skantze and Hjalmarsson (2013) were the most relevant set we were able to find.

pare perceived waiting time, therefore we ensured that the flight announcement always occurred between 17 and 18 seconds after starting to buy time —since 17.5 seconds is the mean duration of the time-buying stretch in our corpus. For this purpose, we initially generated several samples (between 6 and 10 as needed) for each condition, and we used the first four which were long enough for our purposes. We trimmed the end of those which were too long, and we ensured that there was always at least one second of silence between the end of the last time-buying utterance and the beginning of the flight announcement (since a direct transition between these stages would have sounded unnatural, and the models did not include flight announcements, thus they could not predict when these should occur). We also inserted a one-second pause before the system starts buying time, to avoid sudden transitions between caller and system speech. Figure 6.4 shows an example sequence for each condition.

Procedure

Participants first provided demographic data. Afterwards, they did a brief German language check in which they listened to a short recording and answered a question about its content (results from participants who did not pass this check were excluded from the analysis). Later they read the task instructions and were shown an example, after which the actual task started. During this stage, they listened to six audio clips (two for each condition) and, after each clip, rated the corresponding system between 1 and 5 (1 meaning “strongly disagree” and 5, “strongly agree”) with respect to the four statements listed in 6.4.1. Finally, they were asked about any technical issues during the experiment and for further (optional) comments, and were shown the code to enter in the crowdsourcing page to prove they had finished the questionnaire and receive their payment.

Participants

Sixty testers recruited through Amazon Mechanical Turk completed the questionnaire.⁶ We excluded the data from one of them, who reported having uncorrected hearing impairments. We also checked the time that participants spent in the rating pages, in order to detect those who had rated the statements too quickly. We defined the minimum time for each rating page as five seconds —one for each statement and one for moving the

⁶www.mturk.com

<i>Ich warte noch auf die Liste</i>	<i>ja</i>	<i>im Moment sehe ich gar nichts am Ende November</i>	<i>an einem Wochenende</i>	<i>ich schaue mal eben</i>	<i>bisher habe ich nur Anfang Dezember</i>	<i>nach Lissabon</i>	<i>ich sehe gerade nichts an einem Wochenende</i>
I'm waiting for the list	yes	right now I don't see anything at the end of November	on a weekend	I'll have a look	so far I only have beginning of December	to Lisbon	I don't see anything on a weekend
AGENT/SYSTEM STATE	ACK	TEMPORARY N/A	ECHOING	COMMITMENT	TEMPORARY N/A	ECHOING	TEMPORARY N/A

<i>Sekunde noch</i>		<i>Im Moment sehe ich gar nichts am Ende November</i>	<i>ich warte noch auf die Liste</i>		<i>einen kleinen Moment</i>	
One second		right now I don't see anything at the end of November	I'm waiting for the list		just a moment	
WAIT REQUEST		TEMPORARY N/A	AGENT/SYSTEM STATE		WAIT REQUEST	

<i>Okay</i>	<i>gut</i>	<i>Ende November</i>	<i>von Köln-Bonn</i>	<i>schauen wir doch mal</i>	<i>da stehen ein paar Flüge zur Auswahl</i>	<i>ich muss mal gucken</i>
Okay	good	end of November	from Cologne-Bonn	let's have a look	there are a couple of flights to choose from	I need to have a look
ACK	ACK	ECHOING	ECHOING	COMMITMENT	AVAILABILITY	COMMITMENT

Figure 6.4: Example of a sequence for each condition: 1) random, 2) flat, 3) hierarchical. ACK stands for ACKNOWLEDGMENT. The gray blocks represent pauses: wider ones are long pauses (four seconds) and narrower ones are short pauses (two seconds).

Frequency of IVA use	percentage of testers
Every day	25%
Several times a week	32%
Several times a month	11%
Once in a while	25%
Never	7%

Table 6.2: Frequency of use of IVAs reported by testers.

cursor towards the “submit” button and clicking on it. We excluded testers who spent less than five seconds in at least two of the rating pages. This left us with a total of 56 testers, i.e. 112 ratings for each condition (since each tester rated each condition twice). The majority of these participants reported being male: 47, vs. 9 who reported being female (no participants chose the option “other”). Testers were between 18 and 49 years of age, with a mean of 32 ($SD=7.5$). Fifty percent of them expressed that they usually prefer to talk to a person on the phone, 20% preferred a system, and 30% had no preference for either. Information about the frequency with which they used Interactive Virtual Assistants (IVAs) can be found in Table 6.2.

Hypotheses

The expected results were as follows:

- **H1: The flat system will receive better ratings than the random one.** Since the flat strategy is trained on data from humans and it captures information on how speakers sequence time-buyers, we expect it to be perceived as more human-like and more intelligent than the random one (statements 1-2). This will also lead testers to prefer it over the random system for their own use (statement 4). The random system, on the other hand, will provide a less familiar, more confusing experience, which will lead testers to perceive the time elapsed until finding a flight as longer than for the flat system (statement 3).
- **H2: The hierarchical system will receive better ratings than the random one.** Similarly to the flat model, the hierarchical model captures information on how speakers sequence time-buyers. Therefore, the expectations from the previous point regarding the flat model also apply to the hierarchical model.

St.	RAND. (total sum)	RAND. (median)	RAND. (IQR)	FLAT (total sum)	FLAT (median)	FLAT (IQR)	HIER. (total sum)	HIER. (median)	HIER. (IQR)
1	348	3	2	346	3	2	385	4	2
2	396	4	1	399	4	1	* 419	4	1
3	382	3	1	384	4	1	402	4	1
4	329	3	2	336	3	2	363	3	1

Figure 6.5: Ratings received by each strategy, by statement: 1) *The system behaves in a human-like way*, 2) *The system is intelligent*, 3) *The system found a flight quickly enough*, 4) *I would use this system*. The header *IQR* stands for *interquartile range*. (* $p < .0125$, ** $p < .0025$, *** $p < .00025$)

- H3: The hierarchical system will receive equal or better ratings than the flat one.** In addition to the sequencing information, the hierarchical model also includes information about preference for different time-buyers at different stages of the time-buying process (i.e. depending on the time elapsed since the speaker started buying time). We do not know, however, whether this information will further improve the listeners' perception, or whether they will just recognize the overall strategy as more natural than the random one but the extra information will provide no additional advantage. For this reason, we expect the hierarchical strategy to receive either equivalent or better ratings than the flat one.

6.4.2 Results

In order to find out whether the strategies were perceived differently by the participants, we compared median ratings for each question across conditions. These are displayed in Figure 6.5. As pointed out in Chapters 4 and 5, Likert data cannot be normally distributed, therefore we report median as measure of central tendency (instead of means) and interquartile range for dispersion (instead of standard deviation). For the same reason, we use a non-parametric test, Wilcoxon signed rank, in order to evaluate significance of differences (Wilcoxon, 1945): The results of these tests can be found in Tables 6.3 to 6.6. We use Bonferroni-adjusted alpha levels to correct for multiple comparisons: Given that testers rated four statements per stimulus, the alpha levels used are $.05/4 = .0125$, $.01/4 = .0025$, $.001/4 = .00025$. Figures 6.6 to 6.9 show the distribution of the ratings for each statement.

Statement 1	RANDOM	FLAT	HIERARCHICAL
RANDOM	—	$W=1260.5, p>.0125$	$W=1018.5, p>.0125$
FLAT	—	—	$W=1048.5, p<.0125$ (hierarchical wins)

Table 6.3: Results of Wilcoxon signed rank test for statement 1: *The system behaves in a human-like way.*

Statement 2	RANDOM	FLAT	HIERARCHICAL
RANDOM	—	$W=631.5, p>.0125$	$W=563.5, p>.0125$
FLAT	—	—	$W=625.0, p>.0125$

Table 6.4: Results of Wilcoxon signed rank test for statement 2: *The system is intelligent.*

Statement 3	RANDOM	FLAT	HIERARCHICAL
RANDOM	—	$W=1108.0, p>.0125$	$W=1003.5, p>.0125$
FLAT	—	—	$W=761.0, p>.0125$

Table 6.5: Results of Wilcoxon signed rank test for statement 3: *The system found a flight quickly enough.*

Statement 4	RANDOM	FLAT	HIERARCHICAL
RANDOM	—	$W=847.5, p>.0125$	$W=634.0, p<.0125$ (hierarchical wins)
FLAT	—	—	$W=461.0, p>.0125$

Table 6.6: Results of Wilcoxon signed rank test for statement 4: *I would use this system.*

Below we revisit our hypotheses from Section 6.4.1 in light of the results.

- H1: *The flat system will receive better ratings than the random one.*
 - Contrary to our expectations, testers did not find the flat system better than the random one for any of the statements. Possible reasons are discussed in Section 6.5.
- H2: *The hierarchical system will receive better ratings than the random one.*
 - This holds for the fourth statement: Testers were more willing to use the hierarchical system than the random one. However, they did not find it more human-like, more intelligent or able to find a flight faster, therefore it is not clear why they were more willing to use it. This point is discussed further in Section 6.5.
- H3: *The hierarchical system will receive equal or better ratings than the flat one.*
 - Testers found the hierarchical system more human-like than the flat one. However, they did not express more willingness to use it. It is interesting to note that, although they found it more human-like, they did not find it more intelligent. Nevertheless, the subset of testers who did not express a preference for talking to a human on the phone (i.e. those who preferred to talk to a system or had no preference for either, $N = 28$), did rate the hierarchical system as significantly more intelligent than the flat one ($W=46, p<.0125$).

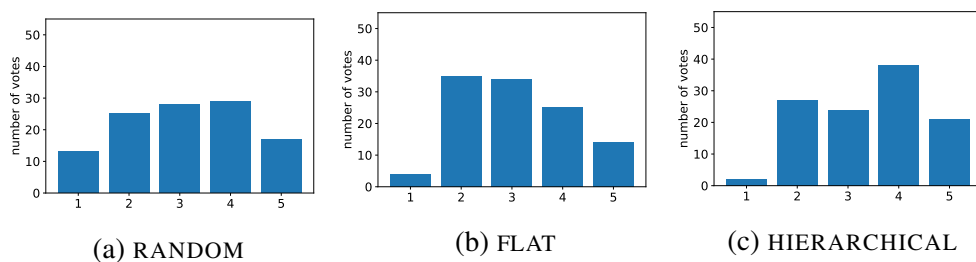


Figure 6.6: Distribution of ratings for statement 1, *The system behaves in a human-like way.*

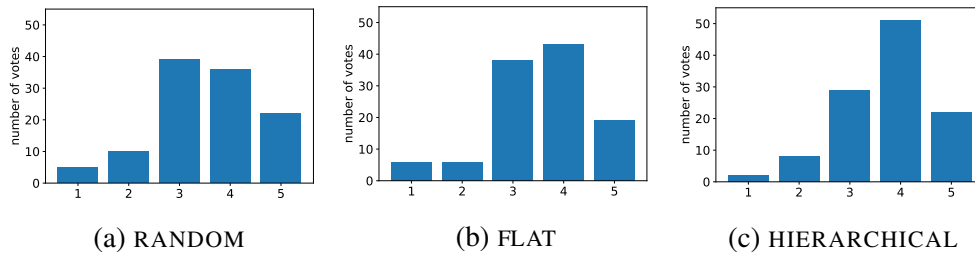


Figure 6.7: Distribution of ratings for statement 2, *The system is intelligent*.

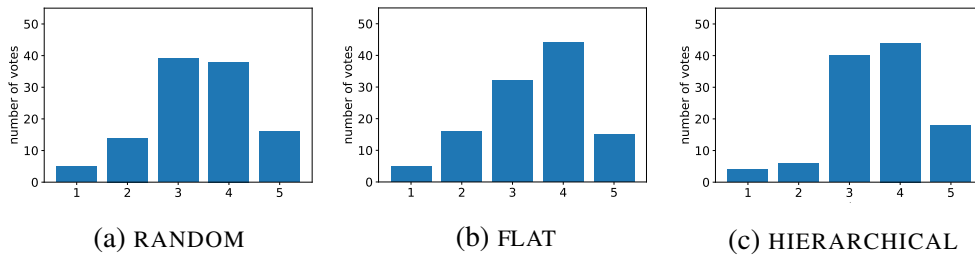


Figure 6.8: Distribution of ratings for statement 3, *The system found a flight quickly enough*.

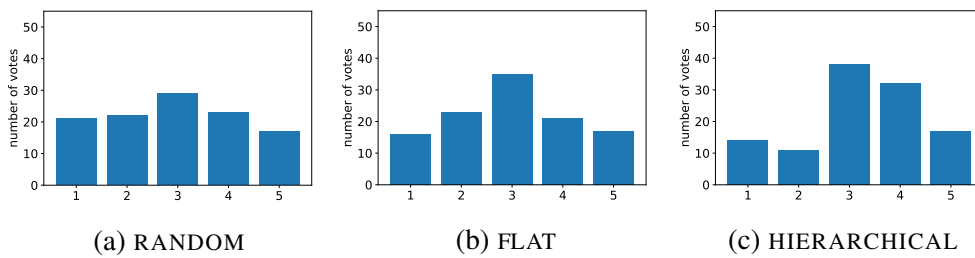


Figure 6.9: Distribution of ratings for statement 4, *I would use this system*.

6.5 Discussion

6.5.1 Observations

As mentioned in Section 6.4.2, for statement *The system found a flight quickly enough*, we detected no differences in ratings across conditions, either for the whole group of participants or for specific demographic groups. Coupled with the results from Chapter 4, which did reveal highly significant differences in how participants perceived elapsed time when the system buys time vs. when it remains silent, this suggests that even randomly selected dialogue acts can prevent users from overestimating waiting time, and that a series of time-buying utterances will be more useful in this respect than silence, without need to put too much thought into what these utterances express or how they are sequenced. This is consistent with the results presented in Chapter 5, in which no difference was found between ratings for a random and a learned strategy. This does not mean, however, that *any* speech will do, since the experiment in Chapter 5 also revealed that buying time by repeatedly asking the user to wait resulted in overestimation of waiting time compared to using a variety of time-buying acts.

For the rest of the statements, results for the hierarchical strategy in comparison to the others were only partially what we expected. As predicted in H3, ratings for this condition showed no difference from those for the flat strategy, except for human-likeness, where the hierarchical strategy was rated higher. This suggests that considering how many time-buying actions were produced since the system started buying time when choosing the next utterance has a positive impact on perception of human-likeness. It is somewhat surprising, therefore, that ratings for intelligence (statement 2) are not higher too, given the connection between human-likeness and intelligence. One interpretation might be that it is possible, in principle, to view an artificial system as something whose behavior mimics human traits while still perceiving it as an inanimate entity which does not possess intelligence in the human sense (on the other hand, as noted in Section 6.4.2, half of the participants did rate the hierarchical system as more intelligent than the flat system, namely those who did not prefer talking to a human on the phone).

When comparing the hierarchical and random strategies, testers were more willing to use the former than the latter in a future occasion. However, the reason is not immediately clear, as they did not rate the hierarchical strategy as more human-like,

intelligent or quick than the random one. This preference might be due to other reasons not covered in the rated statements (e.g. the hierarchical system sounds more polite, more willing to help, etc).⁷ On the other hand, it must be noted that the samples for the random condition had fewer pauses than those for the trained conditions (hierarchical and flat). This was due to the fact that, in the random strategy, each of the pause actions (*long pause* and *short pause*) had the same probability as each of the seven utterance actions, whereas in the trained strategies, utterances were regularly interspersed with pauses. Therefore, one might speculate that the preference for the hierarchical strategy could stem from a more natural pausing behavior, and that the random strategy may have been perceived as too much non-stop talk. However, although the pausing behavior in both trained strategies was similar, only the hierarchical strategy was preferred over the random one, suggesting that the preference is at least not exclusively due to pausing behavior. Therefore, a clearer understanding of this preference would require further data.

One of the most surprising aspects of the results was that the flat strategy was not rated higher than the random strategy, given that the former is trained on data from humans and that perplexity for the flat model is considerably lower than for the random one (see Section 6.3). This suggests that sequencing information does not improve humans' perceptions of a time-buying system. An alternative explanation could be connected to the stimuli used for this condition. Table 6.7 shows all the audio clips used in the experiment sorted by total sums of scores. Flat audio clips are split: two of them appear in the top half, whereas the other two are at the very bottom. We took a closer look at the latter. For one of them, the sequence of time-buyers generated was relatively unlikely: Since the model samples its choices from a probability distribution, it is always possible that it will generate a low-probability sequence. The other side of this consideration is that even the random strategy might, once in a while, produce a human-like sounding sequence by chance. We can also find such a case in Table 6.7: Although three of the random clips were rated in the bottom half, the remaining one

⁷It is worth noting that, whereas the Wilcoxon signed rank test does not suggest significant differences in perceived human-likeness ($W=1018.5, p=0.017$), a t-test does ($t(111)=-2.61, p=0.01$). Although we opted for the more conservative approach of using a non-parametric test due to Likert data not being normally distributed, it has been claimed that, for five-point Likert items, the t-test is relatively robust: de Winter and Doudou (2010) found the risk of Type I error to be similar to that of the Mann-Whitney test, which is the correlate of the Wilcoxon signed rank test for unpaired data (Cairns, 2019). This is not enough to draw any conclusions, but it is worth keeping in mind for further research.

AUDIO CLIP	SUM OF RATINGS
hierarchical_6	428
random_4	409
hierarchical_3	405
flat_4	396
flat_6	385
hierarchical_4	369
hierarchical_2	367
random_2	351
random_3	349
random_5	346
flat_5	344
flat_2	340

Table 6.7: Sum of all ratings for all questions for each stimulus.

contains a likely sequence, and it obtained the second-best score. The case of the other flat clip with low ratings is different: The generated sequence was not unlikely, but the last time-buyer was TEMPORARY NON-AVAILABILITY, which means that the system said it had not found any suitable flights so far, paused for one second, and then announced having found a flight. As mentioned in Section 6.4.1 above, the model does not predict when flight information should be announced: Instead, this happens after a fixed amount of time in the clips. This might perhaps explain the low ratings: Saying that no flight is available and offering a flight a second later may have been perceived as inconsistent.

6.5.2 Comparison with Chapter 5

Since there are considerable similarities between the experiment presented in this chapter and the one presented in Chapter 5, we list the differences between them below:

- **Strategies compared:** Chapter 5 focused on the effect of a “fixed” time-buying strategy which only uses explicit requests for waiting vs. two strategies which use a variety of resources to buy time. Since we found that participants favored variety, Chapter 6 did not include a fixed strategy: All strategies exploit a range of resources,

and we compare different ways of selecting them (randomly, using a trigram model, and using a hierarchical model which combines trigram and unigram probabilities).

- **Model training:** The “learned” model in Chapter 5 and the “hierarchical” model in Chapter 6 are similar in that they both arrive at the final decision in two steps. For the first step, in Chapter 5, we used a tool (OpenDial) to train a Markov Decision Process whereas in Chapter 6 we trained a trigram model only using Python, without a dedicated toolkit. The aim of the latter decision was to make the training process simpler and more transparent in order to a) better understand and be able to explain the model’s choices and b) easily detect any possible problems before running the experiment.

- **Interaction category:** In Chapter 6, we evaluated the model before the experiment (by calculating its perplexity) which we hadn’t done in previous experiments. Moreover, we introduced the *interaction* category, which includes requests for information and clarification, responses to the interlocutor’s speech, etc. In Chapter 5, we had removed this category before training because we were unable to handle this type of dialogue acts with our architecture. In Chapter 6, however, we trained the model using all the categories, and removed interaction dialogue acts before generating the samples for the crowdsourced experiment.

- **Preceding context:** As mentioned above, Chapter 5 focuses on the choice of a time-buying action given its placement in the time-buying stretch (e.g. right after the interlocutor’s request vs. some seconds into the search), regardless of the preceding actions, due to the bug mentioned in Section 5.2.1. In the experiment in Chapter 6, on the other hand, the choice of time-buying action depends on both factors: placement in the time-buying stretch, and the two last time-buyers produced.

6.5.3 Related work

In our experiment, we used trigrams to model the system’s time-buying behavior. Henderson et al. (2005) also used n-grams to model decision-making in dialogue within the flight booking domain. Although their main focus was on developing a user simulator, their system also included a mode which could be used to generate system decisions. The latter, however, differed from ours in that it chose the action with the best score, whereas the user simulation mode was stochastic, like our models. In later work, the authors introduced the concept of “advanced n-grams”, in which —similarly to our

hierarchical strategy— the n-gram models are augmented with additional information (Georgila et al., 2006). In their case, this information is the state of the flight-booking task (e.g. whether information has been provided and confirmed by the user). Their results show that advanced n-grams outperform their classic counterparts for $N = 2$ and $N = 3$, but this does not hold for higher values of N .

6.5.4 Further work

Future possible steps include addressing the shortcomings described above, to be able to confidently separate the potential of the models from the effects that the characteristics of individual audio clips can have on results. One option is to produce more audio clips for each condition (although a drawback of this approach is the need for a larger number of participants). Another strategy could be to use only the most likely sequence of time-buyers for each condition (and any one sequence produced by the random strategy), and only vary the utterances used for each time-buyer. Given that we trained a simple trigram model based on a relatively small amount of data, the risk of sampling sequences which are not so common in real life is always present. Gathering a larger corpus would enable us to train a more sophisticated model and potentially obtain better performance.

Finally, a crowdsourced overhearer test was the only option available during the COVID-19 lockdown but, as discussed in Section 4.4, since such a system is ultimately aimed at interacting with humans, it should ideally be tested in an interactive setting, such as in a lab with participants engaging with it (as in the experiment in Chapter 5). Betz et al. (2018) mention that, in their speech synthesis evaluations, the presence or absence of interaction was shown to impact results. There is a possibility that this could also generalize to studies like ours, since one might imagine that just listening vs. having a responsibility as co-drivers of the interaction could lead humans to focus their attention on different aspects of the system’s performance.

6.6 Summary

In this chapter, we attempted to model human time-buying behavior using trigrams. We developed two models: a pure, “flat” trigram model, and another one which uses trigram probabilities to make a high-level decision and subsequently refines this deci-

sion based on unigram probabilities for that specific stage of the wait. Both of these models reduced perplexity by half when compared to a strategy selecting time-buyers randomly. In addition, in a crowdsourced study where participants listened to samples of the three systems interacting with a human, the two-step trigram model was perceived as more human-like than the flat one, and participants preferred it for their own use over the random one. However, there were no differences in how long participants perceived the wait to be between any of the conditions.

The next chapter summarizes the contributions of this dissertation and lists possible improvements and further work.

Chapter 7

Conclusions

7.1 Introduction

We begin this final chapter by summarizing the findings from the experiments presented above (Section 7.2). Afterwards, we discuss ways in which the work that led to these findings could be improved (Section 7.3). Finally, we present ideas for further research (Section 7.4).

7.2 Overview of findings

In the introductory chapter of this dissertation, we presented its aims as follows:

1. to identify, analyze and describe the strategies that human speakers employ in order to buy time in a task-oriented setting,
2. to model these strategies computationally and
3. to test whether a system which applies these strategies leads to a better user experience.

Chapter 3 centers around the first point. It describes an experiment in which we employed a setup especially designed to elicit time-buying data from humans. The resulting corpus shows that speakers exploit a variety of task- and grounding-related dialogue acts to bridge the gap before they can provide task information. Based on these data, we proposed a time-buyer classification scheme comprising 11 categories.

Speakers combined these categories in various ways —however, we detected a tendency for some categories to occur at specific moments: ACKNOWLEDGMENTS, such as *okay*, were very frequent at the beginning of the time-buying period and became more sparse thereafter, ECHOING of the interlocutor’s request was most frequent in second and third position, and other resources such as FILLERS (*äh...*, *ähm...*) were produced all throughout. ECHOING and FILLERS were the most frequent time-buyers overall, and most participants used them extensively, although it was possible to observe a preference for one or the other for some speakers.

The subsequent chapters focus on aims 2 and 3. **Chapter 4** compares a system which buys time conversationally (by using utterances from the corpus in Chapter 3) with one which does not buy time (it asks the human user to wait and then remains silent until it can provide task information). When presented with the time-buying system, listeners perceived waiting time as more appropriate, even though the actual time was the same for both systems. In addition, when the voice of the system was a commercial voice that we perceived as sounding relatively human-like, they also found the time-buying system more human-like, willing to help and better at understanding the interlocutor than the other system, and they expressed more willingness to interact with it again.

The results above suggest that humans prefer a system that buys time through speech over one which waits in silence. This opens up the question of whether any speech will do as long as it “fills up the gaps”, or if time-buying which is perceived by listeners as satisfactory exhibits specific characteristics. The results of the experiment described in **Chapter 5** suggest the latter. Speech repeating the same dialogue act (request to wait) over and over was dispreferred when compared to a more varied strategy exploiting a number of communicative actions, and in which the probability of a time-buying action being selected at a certain point during the searching period was modeled on human data.

The fact that not all time-buying strategies are perceived equally by human users leads to the question of what is the best way to model time-buying in a system. In **Chapter 6**, we experimented with two ways of modeling humans’ sequencing of time-buyers using trigrams. Both of these models reduced perplexity by 50% in comparison to a strategy which selects time-buying actions randomly, and one of them was preferred by listeners for their own use over the random strategy (although this preference was not as strong as the ones we observed in previous chapters). Interestingly,

we found no differences in perception of waiting time between any of the conditions tested. Coupled with the findings from the previous chapter, in which there were also no differences between the random and learned conditions, this suggests that exploiting a variety of dialogue acts—as humans do—contributes to avoiding overestimation of elapsed time, regardless of how these dialogue acts are sequenced: Participants only perceived elapsed time as longer when the time-buying speech was repetitive (in the *FIXED* condition in Chapter 5) and when there was no speech at all (in the *WAIT* condition in Chapter 4).

7.3 Lessons learned

In retrospect, it is possible to identify a few points in the chapters above where different choices might have been beneficial. As an example, the experiments in Chapters 4, 5 and 6 all use different (though overlapping) sets of statements. Using the same statements for all the experiments would have increased comparability across studies. Ideally, a standard, validated questionnaire from the field of Human-Computer Interaction would have been employed (something we attempted for the experiment in Chapter 6 without success, since none of the questionnaires that we found covered enough of the aspects we intended to focus on, and some aspects were not covered by any of the questionnaires).

On the other hand, in Chapter 6, we highlighted the potential impact of individual stimuli generated by a stochastic model on results. A model trained on human data may generate an unlikely sequence, and a model which selects actions randomly will, from time to time, produce a likely one. In our experiment, different stimuli for the same condition obtained rather disparate ratings, which leads us to wonder whether all stimuli were representative of their corresponding system. To ensure that the stimuli presented to participants represent model performance adequately, one could generate a larger number of samples per condition (and increase the number of participants accordingly). An alternative approach would be to test only the most likely sequence of actions produced by the trained model (varying the surface utterances) instead of several different sequences.

In addition, in the experiment in Chapter 5, the intended behavior of the trained model was different from the actual one due to a technical error discovered during post-experiment analysis. This resulted in the exclusion of one of the aspects that

we planned to test (sequencing of time-buying actions). Therefore, in the analysis, we focused on another relevant aspect, namely frequency of time-buying actions at different points during the wait. It would be instructive to re-run the experiment with the system behaving as initially planned, and observe any differences in results which arise when adding the sequencing dimension.

7.4 Further work

As stated in the introduction of this dissertation, speakers' need to buy time is a result of the situated nature of dialogue. Participants in a conversation are accountable for the use of time during their turn (Clark, 1996), and avoid prolonged silent periods and communicate their state to the user by engaging in time-buying behaviors. Therefore, the effects of time-buying in spoken dialogue systems should be best studied in interaction with human users. Although interaction can be simulated through a Wizard-of-Oz setting (as in Chapter 5), this approach has its limitations. We can simulate reacting to user input when we know in advance what the speaker will say—as when we echoed part of the user's request in Chapter 5, given that the parameters were scripted. However, our human-human corpus shows that, while speakers wait for task information, they do not only monologue, but also sometimes encourage the interlocutor's participation by requesting additional information, confirmation and repetition of the search parameters. In order to cover the full range of behaviors displayed by human speakers, we would need a system that can react to the interlocutor's input while buying time, even when this input deviates from the plan. Another resource which we excluded from our tests due to practical limitations is fillers, as we were unable to generate natural-sounding instances with the artificial voice we used. The next step towards understanding time-buying in dialogue systems and its effect on users should overcome this limitation, given the high frequency of this resource in human time-buying speech.

Another aspect worthy of further research is the interplay between time-buying and system voice. The experiment in Chapter 4 showed that the characteristics of the voice used for the system can have an impact on how differently participants perceive a strategy that buys time conversationally from one which remains mostly silent. As a possible explanation, we postulated that the voice in the second run of the experiment (the Unit Selection voice) may have been perceived as more human-like than the voice in the first run (the HSMM voice), and thus deemed a better match for a system which

buys time in a more human-like way. It would be interesting to run the same experiment with other kinds of voices (higher vs. lower pitched, faster vs. slower, younger vs. older-sounding) and see if the strength of listeners' preferences for the time-buying system also differs. In addition, one could ask participants to rate how much they like the system's voice and check if this correlates with the strength of their preference for the silent or the time-buying system. One could also investigate whether participants who dislike the voice prefer the silent system over the time-buying one, or whether they still prefer the time-buying one but just less strongly—in other words, whether the more familiar, more human-like behavior of the time-buying system compensates for their disliking of the voice.

One of the main underlying themes of this dissertation is human-likeness in human-computer dialogue. Therefore, the work presented above focuses on time-buying behaviors which can occur in human-human interaction, which is why we did not test a strategy that plays music during the wait, despite the widespread use of this resource in telephony applications.¹ From the perspective of user experience, however, it would be worth comparing this kind of strategy with our conversational time-buying models, to find out whether the latter might help reduce customer frustration while waiting. In addition, although we have focused on interaction where speech is the only channel, it would be possible to expand the concept of time-buying to include other modalities: One might, for example, test an embodied agent that communicates its need for extra time through a hand gesture, or by staring at the results with a concentrated expression while moving its eyes from left to right. These strategies could be compared with the use of visual cues normally associated with machines, such as showing a spinning wheel, a progress bar or an hourglass.

Finally, an aspect which is not addressed in this work but deserves further research is the relation between time bridging and estimated time until task content is available or, in other words, whether the characteristics of the speech used to buy time are somehow influenced by the predicted length of the information delay.

¹Anecdotally, although humans cannot play pre-recorded music by themselves while buying time in the way systems can, the author of this work has observed speakers in real life humming a melody or rhythmic pattern while waiting for task information to be available.

Appendices

Appendix A

Data collection (Chapter 3)

A.1 Instructions for participants

In diesem Experiment geht es darum, die Dialogkomponente eines Systems zu verbessern, das Auskunft über Flugreisen geben können soll.

Das System kann bereits Anrufe entgegennehmen, die Anfragen eines Anrufers verstehen, und (mehr oder weniger) passende Flüge zu dieser Anfrage anzeigen.

Wir wollen die Dialogkomponente so weiterentwickeln, dass sie einem Anrufer auch mündlich bestimmte Flüge anbieten kann.

Deswegen nehmen wir Dialoge zwischen Ihnen und einem Anrufer auf.

Ihre Aufgabe besteht darin, aus der Liste der Flüge, die das System vorschlägt, einen bestimmten Flug herauszusuchen, der zur Anfrage des Anrufers am besten passt.

Die Dialoge werden wie folgt ablaufen:

1. Sie hören einen Telefonklingelton. Es meldet sich das System mit einer automatischen Begrüßung.
2. Der Anrufer teilt dem System mit, nach welchem Flug er sucht.
3. Das System nimmt die Anfrage auf, und spielt einen Piepton. Bis zu diesem Piepton müssen Sie einfach zuhören und nicht mit dem Anrufer sprechen.

4. Nach dem Piepton werden auf dem Bildschirm Daten zur Anfrage des Anrufers und mehrere mögliche Flüge angezeigt. Jetzt sollen Sie das Gespräch mit ihm so nahtlos wie möglich übernehmen und aus der Liste einen Flug herausuchen, der gut zur Anfrage passt.
5. Sie sollten versuchen, möglichst kundenfreundlich und kooperativ mit dem Anrufer umzugehen. Nach dem Piepton brauchen Sie sich jedoch nicht weiter vorstellen oder den Kunden begrüßen. Führen Sie einfach das Gespräch fort, was vom System begonnen wurde.
6. Wenn der Anrufer einen für ihn passenden Flug gefunden hat, können Sie das Gespräch beenden. Wenn er noch andere oder zusätzliche Vorstellungen für seine Flugbuchung hat, führen Sie das Gespräch weiter, bis Sie einen entsprechenden Flug gefunden haben.
7. Insgesamt werden sie 2 Testgespräche und 10 richtige Gespräche mit einem Anrufer führen.

Anmerkungen:

Es kann vorkommen, dass das System die Ergebnisse für eine Anfrage verzögert anzeigt. Außerdem können während des Gesprächs bestimmte Flüge ausgebucht werden, diese werden dann in der Anzeige hellgrau.

Appendix B

Experiment 1 (Chapter 4)

B.1 Survey



0% ausgefüllt

Willkommen zur Studie "Reiseinformationssystem DSG - version 2"

In diesem Experiment hören Sie Ausschnitte von Telefongesprächen zwischen einem Kunden und einem automatischen System bei einer Reiseagentur.

Ihre Aufgabe besteht darin, nach jedem Gespräch einen kurzen Multiple-Choice-Fragebogen auszufüllen. Dabei wollen wir Ihre Meinung über das System erfahren.

Weiter

[Studie Reiseinformationssystem](#), Universität Bielefeld – 2017

Figure B.1: Online survey for the experiment in Chapter 4: part 1

Geschlecht:

Alter (in Jahren):

1. Wie würden Sie Ihre Deutschsprachkenntnisse beschreiben?

2. Haben Sie nicht-korrigierte Beeinträchtigungen der Hörfähigkeit?

Höchster Bildungsabschluss:

3. Wie oft benutzen Sie Virtuelle Assistenten (Siri, Cortana, Alexa, Google Now usw.)?


4. Wenn ich Informationen über das Telefon anfrage, spreche ich lieber mit:

einer Person

einem automatischen System

es ist mir egal

Figure B.2: Online survey for the experiment in Chapter 4: part 2



11% ausgefüllt

Danke! Jetzt hören Sie einen kurzen Ausschnitt aus einem Gespräch. Sie werden später eine Frage dazu beantworten.

Sie müssen darauf achten, **aus welchem Grund der Sprecher ab Köln-Bonn Flughafen abfliegen will**.

Das Gespräch fängt an, wenn Sie auf "Weiter" drücken.

[Studie Reiseinformationssystem](#), Universität Bielefeld – 2017

Figure B.3: Online survey for the experiment in Chapter 4: part 3

Danke schön! Bitte lesen Sie nun die Instruktionen für Ihre Aufgabe.

Sie werden vier Gespräche zwischen einem Kunden und einem System hören. Zuerst wird der Kunde nach einen Flug fragen. Danach spricht das System. Wir wollen erfahren, was Sie von dem System halten. Bitte hören Sie jedes Gespräch an und bewerten dann, inwieweit Sie den folgenden Äußerungen zustimmen:


	nicht einverstanden	vollkommen einverstanden
Das System hat den Kunden gut verstanden.	<input type="radio"/>	<input type="radio"/>
Das System hat in angemessener Zeit eine Antwort gegeben.	<input checked="" type="radio"/>	<input checked="" type="radio"/>
Das System klingt, als ob es helfen möchte.	<input type="radio"/>	<input type="radio"/>
Das System benimmt sich so, wie ich es von einer Person erwarten würde	<input checked="" type="radio"/>	<input checked="" type="radio"/>
Wenn ich einen Flug am Telefon kaufen müsste, würde ich dieses System benutzen.	<input type="radio"/>	<input type="radio"/>

Ihre Bewertungen sollen sich nur auf das Gespräch beziehen, *dass Sie gerade gehört haben* (nicht auf die Vorherigen).

Bevor Sie mit der Aufgabe anfangen, werden Sie ein Beispiel hören. Das Gespräch fängt automatisch an, wenn Sie auf die nächste Seite gehen. Sie hören es nur einmal. Nachdem Sie das ganze Gespräch gehört haben, klicken Sie bitte auf "Weiter".

Hinweis: Der "Weiter" Knopf erscheint erst nach Ende des Gesprächs.

Figure B.4: Online survey for the experiment in Chapter 4: part 4



53% ausgefüllt

Wie finden Sie das System, das Sie gerade gehört haben? Bitte bewerten Sie die folgenden Äußerungen:

	nicht einverstanden	vollkommen einverstanden
Das System hat den Kunden gut verstanden.	<input type="radio"/>	<input type="radio"/>
Das System hat in angemessener Zeit eine Antwort gegeben.	<input checked="" type="radio"/>	<input checked="" type="radio"/>
Das System klingt, als ob es helfen möchte.	<input type="radio"/>	<input type="radio"/>
Das System benimmt sich so, wie ich es von einer Person erwarten würde.	<input checked="" type="radio"/>	<input checked="" type="radio"/>
Wenn ich einen Flug am Telefon kaufen müsste, würde ich dieses System benutzen.	<input type="radio"/>	<input type="radio"/>

Weiter

Studie [Reiseinformationssystem](#), Universität Bielefeld – 2017

Figure B.5: Online survey for the experiment in Chapter 4: part 5 (repeated for each recording)

soSci
ofB - der onlineFragebogen

89% ausgefüllt

Vielen Dank! Auf der nächsten Seite bekommen Sie Ihren Umfrage-Code.
Wir hätten jetzt zwei letzte Fragen:

6. Haben Sie irgendwelche technische Probleme gehabt?

[Bitte auswählen] ▼

7. Haben Sie sonst andere Kommentare?

Weiter

[Studie Reiseinformationssystem](#), Universität Bielefeld – 2017

Figure B.6: Online survey for the experiment in Chapter 4: part 6

soSci
ofB - der onlineFragebogen

95% ausgefüllt

Vielen Dank!
Dies ist Ihr Umfrage-Code:
3221982962

Bitte tragen Sie diesen Code in das passende Feld auf der Crowdfunder Webseite ein.
Dies ist der Beweis, dass Sie die Aufgabe bis zum Ende gemacht haben.

Weiter

[Studie Reiseinformationssystem](#), Universität Bielefeld – 2017

Figure B.7: Online survey for the experiment in Chapter 4: part 7

Appendix C

Experiment 2 (Chapter 5)

C.1 Experiment instructions

SITUATION:

Du arbeitest als SekretärIn bei der Firma *Rees & Associates*. Dein Chef hat dir die Aufgabe gegeben, eine Liste von Flügen für einige Mitarbeiter der Firma zu buchen. Die Firma bucht Flüge immer telefonisch bei der DSG-Reiseagentur. Diese Agentur benutzt ein automatisches Dialogsystem mit Spracherkennungstechnologie, das deine Bestellung versteht und nach einem passenden Flug sucht.

INSTRUKTIONEN:

Gehe die Liste der Flüge Schritt für Schritt durch, und rufe für jeden Flug die DSG-Reiseagentur an. Für jeden Flug gibt es einige Kriterien, die erfüllt sein sollen (z.B. Start- und Zielflughafen) und die du in deiner Anfrage erwähnen sollst. Nachdem dich das System begrüßt hat, ist deine Aufgabe, das System nach einem Flug mit den entsprechenden Kriterien zu fragen. Wenn das System einen Flug gefunden hat, schickt es dir eine Email mit den Details des Fluges (deine Email ist schon in der Kundendatenbank der Firma registriert). Wenn es keinen passenden Flug finden kann, fragt es dich nach alternativen Kriterien. Du kannst pro Anruf nur ein Flug bestellen.

Jetzt gibt es zwei Training-Anrufe, damit du die Aufgabe üben kannst

- TEST ANRUF 1: Frankfurt - Sydney / 3. August / vormittags
- TEST ANRUF 2: Köln-Bonn - Lissabon / Ende November / Wochenende

Wenn du fertig bist, drehe bitte die Seite um.

Jetzt sind wir bereit, mit dem Experiment anzufangen. Du sollst zehn Flüge buchen, also zehn Anrufe durchführen. Nach jedem Anruf sollst du die letzte Interaktion in Bezug auf drei Kriterien bewerten. Die Kriterien sind:

- Es war angenehm, mit diesem System zu interagieren
- Das System hat in angemessener Zeit eine Antwort gegeben
- Das System benimmt sich so, wie ich es von einer Person erwarten würde

Nach jedem Satz sollst du eine Bewertung zwischen 1 (nicht einverstanden) und 5 (vollkommen einverstanden) abgeben (Zahl einkreisen).

Danke und viel Spaß !!!

ANRUF 1: Düsseldorf - Rom – 10. Juni – nachmittags

- | | | (nicht
einverstanden) | | (vollkommen
einverstanden) | |
|---|--|--------------------------|---|-------------------------------|-----|
| • Es war angenehm, mit diesem System zu interagieren: | | 1 | 2 | 3 | 4 5 |
| • Das System hat in angemessener Zeit eine Antwort gegeben: | | 1 | 2 | 3 | 4 5 |
| • Das System benimmt sich so, wie ich es von einer Person erwarten würde: | | 1 | 2 | 3 | 4 5 |

Kommentare (optional):

ANRUF 2: Berlin – Bristol -- 31. Mai -- abends

- | | | (nicht
einverstanden) | | (vollkommen
einverstanden) | |
|---|--|--------------------------|---|-------------------------------|-----|
| • Es war angenehm, mit diesem System zu interagieren: | | 1 | 2 | 3 | 4 5 |
| • Das System hat in angemessener Zeit eine Antwort gegeben: | | 1 | 2 | 3 | 4 5 |
| • Das System benimmt sich so, wie ich es von einer Person erwarten würde: | | 1 | 2 | 3 | 4 5 |

Figure C.1: Rating form for participants (first two calls)

Appendix D

Experiment 3 (Chapter 6)

D.1 Dialogues for each strategy

(LP stands for “long pause”, SP stands for “short pause”)

Customer’s request (same for all dialogues)

- Hallo, ich möchte von Köln-Bonn nach Lissabon fliegen, am Ende November und ich möchte nicht an einem Werktag fliegen.

System’s reply: random strategy

- Zur Verfügung stehen verschiedene Flüge. Da haben wir was im Angebot. Ende November. Schau ich gerade einmal nach. Bisher habe ich nur werktags. Gut, von Köln-Bonn. [LP] Da gucken wir doch mal.
- Ich warte noch auf die Liste. [SP] Ja, im Moment sehe ich gar nichts am Ende November. An einem Wochenende. Ich schaue mal eben. Bisher habe ich nur Anfang Dezember. Nach Lissabon. Ich sehe gerade nichts an einem Wochenende.
- Ach, mein System braucht gerade ein wenig länger. Einen kleinen Moment. [LP] Im Moment sehe ich gar nichts am Ende November. Ich schaue gerade mal in meine Liste.

- An einem Wochenende. Da gucken wir doch mal. [SP] Ich habe hier ein paar Möglichkeiten. Bisher habe ich nur Anfang Dezember. Nach Lissabon. Noch einen Moment. Ich sehe gerade nichts an einem Wochenende. Ende November, okay.

System's reply: flat strategy

- Sekunde noch [LP]. Im Moment sehe ich gar nichts am Ende November. Ich warte noch auf die Liste. [SP] Einen kleinen Moment. [LP].
- Okay. [SP] Nach Lissabon. Schau ich gerade einmal nach. [SP] Ich bitte um ein wenig Geduld. [LP] Bisher habe ich nur werktags.
- Schauen wir doch mal [LP]. Ich sehe gerade nichts an einem Wochenende. [SP] Ich schaue gerade mal in meine Liste.
- Bisher habe ich nur werktags. [LP] Im Moment sehe ich gar nichts am Ende November. [LP] Noch einen Moment.

System's reply: hierarchical strategy

- Mm-hm. [SP] Ich warte noch auf die Liste. [SP] Zur Verfügung stehen verschiedene Flüge. [SP] Da haben wir was im Angebot. [SP] Einen kleinen Moment, bitte.
- Okay. [SP] Von Köln-Bonn. [SP] Ich schaue gerade mal in meine Liste. [LP] Die Flüge kommen langsam rein. Ich sehe doch gerade mal in unserem System.
- Gut. [SP] Ende November. Von Köln-Bonn. Ich bitte um ein wenig Geduld. [SP] Ich sehe gerade nichts an einem Wochenende. [LP] Die Liste wird noch aktualisiert.
- Okay. Gut. Ende November [SP] Von Köln-Bonn. Schauen wir doch mal. [SP] Da stehen ein paar Flüge zur Auswahl. [SP] Ich muss mal gucken.

System's flight announcement (same for all dialogues)

- Ich habe einen Flug für Sie gefunden. Das wäre am... [FADEOUT]

Bibliography

- Amati, F. and Brennan, S. (2018). Eye gaze as a cue for recognizing intention and coordinating joint action. In *Advances in Interaction Studies*, volume 10. John Benjamins Publishing Company.
- Antonides, G., Verhoef, P., and van Aalst, M. (2002). Consumer perception and evaluation of waiting time: A field experiment. In *Journal of Consumer Psychology*, volume 12 (3), pages 193–202. Lawrence Erlbaum Associates, Inc.
- Awadallah, A. H., Kholly, A. E., and Zitouni, I. (2018). Hey Cortana! Exploring the use cases of a desktop based digital.
- Banchs, R. E. and Li, H. (2012). IRIS: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics.
- Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1:71–81.
- Baumann, T. and Schlangen, D. (2013). Open-ended, extensible system utterances are preferred, even if they require filled pauses. In *Proceedings of Short Papers at SIGdial 2013*.

- Bellegarda, J. R. (2013). Large-scale personal assistant technology deployment: the Siri experience. In *INTERSPEECH*, pages 2029–2033.
- Bergmann, K., Eyssel, F., and Kopp, S. (2012). A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In *Proceedings of IVA*, Santa Cruz, USA. Springer.
- Betz, S., Carlmeyer, B., Wagner, P., and Wrede, B. (2018). Interaction Hesitation Synthesis: Modeling and Evaluation. In *Multimodal Technologies and Interaction*, volume 2.
- Betz, S. and Wagner, P. (2016). Disfluent lengthening in spontaneous speech. In Jokisch, O., editor, *Studentexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2016*, pages 135–144. TUDpress, Dresden.
- Brooke, J. (1996). SUS – a quick and dirty usability scale. *Usability evaluation in industry*, 189(194).
- Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *In Proceedings of the AAMAS 2009 Workshop*.
- Bunt, H. (2011). The semantics of dialogue acts. In *Proceedings of the Ninth International Conference on Computational Semantics, IWCS '11*, pages 1–13, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A. C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., and Traum, D. (2010). Towards an ISO standard for dialogue act annotation. In *Proceedings of LREC 2010, the Seventh International Conference on Language Resources and Evaluation*.
- Bunt, H., Petukhova, O., Traum, D., and Alexandersson, J. (2017). Dialogue Act Annotation with the ISO 24617-2 Standard. In Dahl, D., editor, *Multimodal Interaction with W3C Standards*, pages 109–136. Springer.
- Buschmeier, H., Baumann, T., Dosch, B., Kopp, S., and Schlangen, D. (2012). Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of SIGdial 2012*, pages 295–303.

- Cairns, P. (2019). *Doing Better Statistics in Human-Computer Interaction*. Cambridge University Press, Cambridge, UK.
- Campione, E. and Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Proceedings of Speech Prosody 2002*. International Speech Communication Association (ISCA).
- Casell, J., Sullivan, J., Prevost, S., and Churchill, E., editors (2000). *Embodied Conversational Agents*. The MIT Press, Cambridge, Massachusetts.
- Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge, UK.
- Clark, H. (2002). Speaking in time. In *Speech Communication*, volume 36, pages 5–13. Elsevier Science.
- Clark, H. and Brennan, S. (1991). Grounding in communication. In Resnick, L. B., Levine, J. M., and Teasley, J. S. D., editors, *Perspectives on Socially Shared Cognition*. Elsevier.
- Clark, H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. In *Cognition*, volume 22, pages 1–39. Elsevier.
- Core, M. G. and Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56. Boston, MA.
- Cruttenden, A., editor (2001). *Gimson's Pronunciation of English, 6th ed.* Routledge.
- Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of Oz studies: Why and how. *Knowledge-Based Systems*, 6(4):258–266.
- de Winter, J. and Doudou, D. (2010). Five-point Likert items: t-test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research and Evaluation*, 15.
- Edlund, J., Gustafson, J., Heldner, M., and Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. In *Speech Communication*, volume 50, pages 630–645. Elsevier.

- Ehrenbrink, P., Osman, S., and Möller, S. (2017). Google Now is for the extraverted, Cortana for the introverted: Investigating the influence of personality on IPA preference. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, OZCHI '17, pages 257–265, New York, NY, USA. ACM.
- Ferguson, G., Allen, J., and Miller, B. (1996). Trains-95: Towards a mixed-initiative planning assistant. In *Proceedings of the Third Conference on Artificial Intelligence Planning Systems (AIPS-96)*, Edinburgh, Scotland. Association for the Advancement of Artificial Intelligence.
- Fillmore, C. (1979). On fluency. In *Individual Differences in Language Ability and Language Behavior*, pages 85–101. Elsevier.
- Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In Ge, S. S., Khatib, O., Cabibihan, J.-J., Simmons, R., and Williams, M. A., editors, *Social Robotics*, pages 199–208, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Fitrianie, S. and Richards, D. (2019). What Are We Measuring Anyway? A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In *Proceedings of IVA*, pages 159–161, Paris, France. Springer.
- Friman, M. (2010). Affective dimensions of the waiting experience. *Transportation Research Part F: Traffic Psychology and Behaviour*, 13:197–205.
- Gasic, M., Jurčićek, F., Thomson, B., and Young, S. J. (2012). Optimisation for POMDP-based spoken dialogue systems.
- Georgila, K., Henderson, J., and Lemon, O. (2006). User Simulation for Spoken Dialogue Systems: Learning and Evaluation. In *Proceedings of Interspeech 2006 - ICSLP*, pages 1065–1068, Pittsburgh, Pennsylvania, USA. International Speech Communication Association (ISCA).
- Ginzburg, J., Fernández, R., and Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7.
- Glass, J. (1999). Challenges for Spoken Dialogue Systems. In *Proceedings of 1999 IEEE ASRU Workshop*. IEEE.

- Gustafson, J. and Merkes, M. (2009). Eliciting interactional phenomena in human-human dialogues. In *Proceedings of SIGdial 2009*, pages 298–301. Association for Computational Linguistics.
- Harms, C. and Biocca, F. (2004). Internal consistency and reliability of the networked minds measure of social presence. In *Proceedings of 7th Annual International Workshop: Presence*, Valencia, Spain.
- Henderson, J., Lemon, O., and Georgila, K. (2005). Learning user simulations for Information State Update dialogue systems. In *Proceedings of Interspeech 2005*, Lisbon, Portugal. International Speech Communication Association (ISCA).
- Higashinaka, R., Mizukami, M., Kawabata, H., Yamaguchi, E., Adachi, N., and Tomita, J. (2018). Role play-based question-answering by real users for building chatbots with consistent personalities. In *Proceedings of SIGdial 2018*, Melbourne, Australia. Association for Computational Linguistics.
- Hirsch, I., Bilger, R., and Heatherage, B. (1956). The effect of auditory and visual background on apparent duration. In *American Journal of Psychology*, volume 69. University of Illinois Press.
- Horvath, A. O. and Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of Counseling Psychology*, 36:223–233.
- Hough, J. (2015). *Modeling Incremental Self-Repair Processing in Dialogue*. PhD thesis, Queen Mary University.
- Jefferson, G. (1983). Notes on a possible metric which provides for a standard maximum silence of approximately one second in conversation. In *Tilburg Papers in Language and Literature*.
- Jefferson, G. (1989). Notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In Roger, D. and Bull, P., editors, *Conversation: An interdisciplinary perspective*. Multilingual Matters, Clevedon, UK.
- Jurafsky, D. and Martin, J. (2018). *Speech and Language Processing, 3rd ed. draft*.

- Kageyama, Y., Chiba, Y., Nose, T., and Ito, A. (2018). Improving user impression in spoken dialog system with gradual speech form control. In *Proceedings of SIGdial 2018*, Melbourne, Australia. Association for Computational Linguistics.
- Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems*, 2(1):26–41.
- Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge, UK.
- Kennington, C. (2016). *Incrementally Resolving References in Order to Identify Visually Present Objects in a Situated Dialogue Setting*. PhD thesis, Bielefeld University.
- Kennington, C., Kousidis, S., and Schlangen, D. (2014). InproTKs: A toolkit for incremental situated processing. In *Proceedings of SIGdial 2014*, pages 84–88, Philadelphia, USA. Association for Computational Linguistics.
- Kohtz, L. S. and Niebuhr, O. (2017). How long is too long? How pause features after requests affect the perceived willingness of affirmative answers. In *Proceedings of the International Conference on Spoken Language Processing*.
- Kopp, S. and Wachsmuth, I., editors (2009). *Gesture in Embodied Communication and Human-Computer Interaction. Revised selected papers from the 8th International Gesture Workshop*. Springer.
- Kousidis, S. and Schlangen, D. (2015). The power of a glance: Evaluating embodiment and turn-taking strategies of an active robotic overhearer. In *Proceedings of the AAAI Spring Symposia 2015*.
- Kumar, P., Kalwani, M. U., and Dada, M. (1997). The impact of waiting time guarantees on customers' waiting experiences. *Marketing Science*, 16(4):295–314.
- Larsson, S. (2005). Dialogue systems: Simulations or interfaces? In *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2005)*, Nancy, France.

- Latorre, J., Lachowicz, J., Lorenzo-Trueba, J., Merritt, T., Drugman, T., Ronanki, S., and Viacheslav, K. (2018). Effect of data reduction on sequence-to-sequence neural TTS. <http://arxiv.org/abs/1811.06315>.
- Lee, K., Zhao, T., Black, A. W., and Eskenazi, M. (2018). Dialcrowd: A toolkit for easy dialog system assessment. In *Proceedings of SIGdial 2018*, Melbourne, Australia. Association for Computational Linguistics.
- Lemon, O. and Pietquin, O. (2012). *Data-Driven Methods for Adaptive Spoken Dialogue Systems*. Springer, New York.
- Lennon, P. (2003). The Lexical Element in Spoken Second Language Fluency. In Riggensbach, H., editor, *Perspectives on Fluency*, chapter 2, pages 25–42. University of Michigan.
- Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.
- Levinson, S. (1983). *Pragmatics*. Cambridge University Press, Cambridge, UK.
- Lickley, R. (2015). Fluency and Disfluency. In Redford, M., editor, *The Handbook of Speech Production*, chapter 20, pages 445–469. Wiley-Blackwell, West Sussex, UK.
- Linell, P. (2004). *The Written Language Bias in Linguistics: Its Nature, Origins and Transformations*. Routledge, Abingdon, UK.
- Lison, P. (2015). A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232 – 255.
- Lison, P. and Kennington, C. (2016). OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - System Demonstrations*.
- Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q., and Martinez, A. (2019). Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science*, 51(4):984–997.

- Lowe, R., Pow, N., Serban, I. V., Charlin, L., Liu, C.-W., and Pineau, J. (2017). Training end-to-end dialogue systems with the Ubuntu Dialogue Corpus. *Dialogue & Discourse*, 8:31–65.
- Lubis, N., Sakti, S., Yoshino, K., and Nakamura, S. (2018). Unsupervised counselor dialogue clustering for positive emotion elicitation in neural dialogue systems. In *Proceedings of SIGdial 2018*, Melbourne, Australia. Association for Computational Linguistics.
- Lundholm Fors, K. (2015). *Production and Perception of Pauses in Speech*. PhD thesis, University of Gothenburg.
- Matsuyama, Y., Bhardwaj, A., Zhao, R., Romeo, O., Akoju, S., and Cassell, J. (2016). Socially-aware animated intelligent personal assistant agent. In *Proceedings of SIGdial 2016*, pages 224–227, Los Angeles. Association for Computational Linguistics.
- McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press.
- Mitev, N., Renner, P., Pfeiffer, T., and Staudte, M. (2018). Towards efficient human-machine collaboration: effects of gaze-driven feedback and engagement on performance. *Cognitive Research: Principles and Implications*, 3.
- Moniz, H., Trancoso, I., and Mata, A. I. (2010). Disfluencies and the perspective of prosodic fluency. In *Proceedings of the Second International Conference on Development of Multimodal Interfaces: Active Listening and Synchrony*, COST'09, pages 382–396, Berlin, Heidelberg. Springer-Verlag.
- Munichor, N. and Rafaeli, A. (2007). Numbers or apologies? Customer reactions to telephone waiting time fillers. In *Journal of Applied Psychology*, volume 92 (2), pages 511–518. American Psychological Association.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

- Pruyn, A. and Smidts, A. (1998). Effects of waiting on the satisfaction with the service: Beyond objective time measures. *International Journal of Research in Marketing*, 15(4):321–334.
- Raux, A. and Nakano, M. (2010). The dynamics of action corrections in situated interaction. In *Proceedings of SIGdial 2010*, pages 165–174. Association for Computational Linguistics.
- Reichman, R. (1985). *Getting Computers to Talk Like You and Me*. The Massachusetts Institute of Technology, Cambridge, Massachusetts.
- Rickheit, G. and Wachsmuth, I. (2006). Introduction: Situated Communication. In Rickheit, G. and Wachsmuth, I., editors, *Situated Communication*, volume 166, pages 1–6. Mouton de Gruyter.
- Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 583–593, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Roberts, F. and Francis, A. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. In *Journal of the Acoustical Society of America*, volume 133.
- Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.
- Sakai, K., Higashinaka, R., Yoshikawa, Y., Ishiguro, H., and Tomita, J. (2018). Introduction method for argumentative dialogue using paired question-answering interchange about personality. In *Proceedings of SIGdial 2018*, Melbourne, Australia. Association for Computational Linguistics.
- Schlangen, D. and Skantze, G. (2011). A General, Abstract Model of Incremental Dialogue Processing. *Dialogue and Discourse*, 2(1):83–111.
- Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. In *International Journal of Speech Technology*, volume 6, pages 365–377.

- Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
- Sekicki, M. and Staudte, M. (2018). Eye'll help you out! How the gaze cue reduces the cognitive load required for reference processing. In *Cognitive Science*.
- Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2018). Survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9:1–49.
- Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3295–3301. AAAI Press.
- Shah, H., Warwick, K., Vallverdú, J., and Wu, D. (2016). Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior*, 58(C):278–295.
- Shamekhi, A., Czerwinski, M., Mark, G., Novotny, M., and Bennett, G. A. (2016). An exploratory study toward the preferred conversational style for compatible virtual agents. In *Proceedings of IVA*, Los Angeles, USA. Springer.
- Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California, Berkeley.
- Shriberg, E. (1999). Phonetic consequences of speech disfluency. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS-99)*, pages 619–622.
- Skantze, G. and Hjalmarsson, A. (2013). Towards incremental speech generation in conversational systems. *Computer Speech & Language*, 27:243–262.
- Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J., Gao, J., and Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. *CoRR*, abs/1506.06714.
- Staudte, M. and Crocker, M. (2018). Studies on the role of eye gaze in dialogue. In *Advances in Interaction Studies*, volume 10. John Benjamins Publishing Company.

- Stivers, T., Enfield, N., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J., Yoon, K.-E., and Levinson, S. (2009). Universal and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 106(26):10587–10592.
- Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts, USA.
- Taylor, S. (1994). Waiting for Service: The Relationship between Delays and Evaluations of Service. *Journal of Marketing*, 58(2):56–69.
- Tom, G., Burns, M., and Zeng, Y. (1997). Your life on hold: The effect of telephone waiting time on customer perception. In *Journal of Direct Marketing*, volume 11 (3), pages 25–31. John Wiley and Sons, Inc. and Direct Marketing Educational Foundation, Inc.
- Traum, D. (2018). Beyond dialogue system dichotomies: Principles for human-like dialogue. Presentation – Ninth International Workshop on Spoken Dialogue Systems Technology (IWSDS 2018).
- Tsai, V., Baumann, T., Pecune, F., and Casell, J. (2018). Faster responses are better responses: Introducing incrementality into sociable virtual personal assistants. In *Proceedings of the Ninth International Workshop on Spoken Dialogue Systems Technology (IWSDS 2018)*, Singapore.
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59:433–460.
- van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., van den Driessche, G., Lockhart, E., Cobo, L. C., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D., and Hassabis, D. (2017). Parallel wavenet: Fast high-fidelity speech synthesis. *CoRR*, abs/1711.10433.
- Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., Prasad, R., Cheng, M., Hedayatnia, B., Metallinou, A., Goel, R., Yang, S., and Raju, A. (2018). On evaluating and comparing conversational agents. *CoRR*, abs/1801.03625.

- Villing, J. (2015). *Towards Dialogue Strategies for Cognitive Workload Management*. PhD thesis, University of Gothenburg.
- Vinyals, O. and Le, Q. V. (2015). A neural conversational model. *CoRR*, abs/1506.05869.
- Wahlster, W. (2003). Towards symmetric multimodality: Fusion and fission of speech, gesture, and facial expression. In *KI 2003: Advances in Artificial Intelligence*, Heidelberg. Springer.
- Walker, M., Kamm, C., and Litman, D. (2000). Towards developing general models of usability with PARADISE. In *Natural Language Engineering*, volume 6 (3-4).
- Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. *CoRR*, cmp-lg/9704004.
- Weizenbaum, J. (1966). Eliza — a computer program for the study of natural language communication between man and machine. *Communications of ACM*, 9(1):36–45.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.
- Williams, J., Raux, A., and Henderson, M. (2016). The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7:4–33.
- Williams, J. and Young, S. (2007). Partially observable Markov decision processes for spoken dialogue systems. *Computer Speech & Language*, 21:393–422.
- Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., and Yu, K. (2010). The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24:150–174.
- Young, S., Gašić, M., Thomson, B., and Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.
- Zellner, B. (1994). Pauses and the temporal structure of speech. In Keller, E., editor, *Fundamentals of Speech Synthesis and Speech Recognition*, chapter 3, pages 41–62. John Wiley, Chichester.