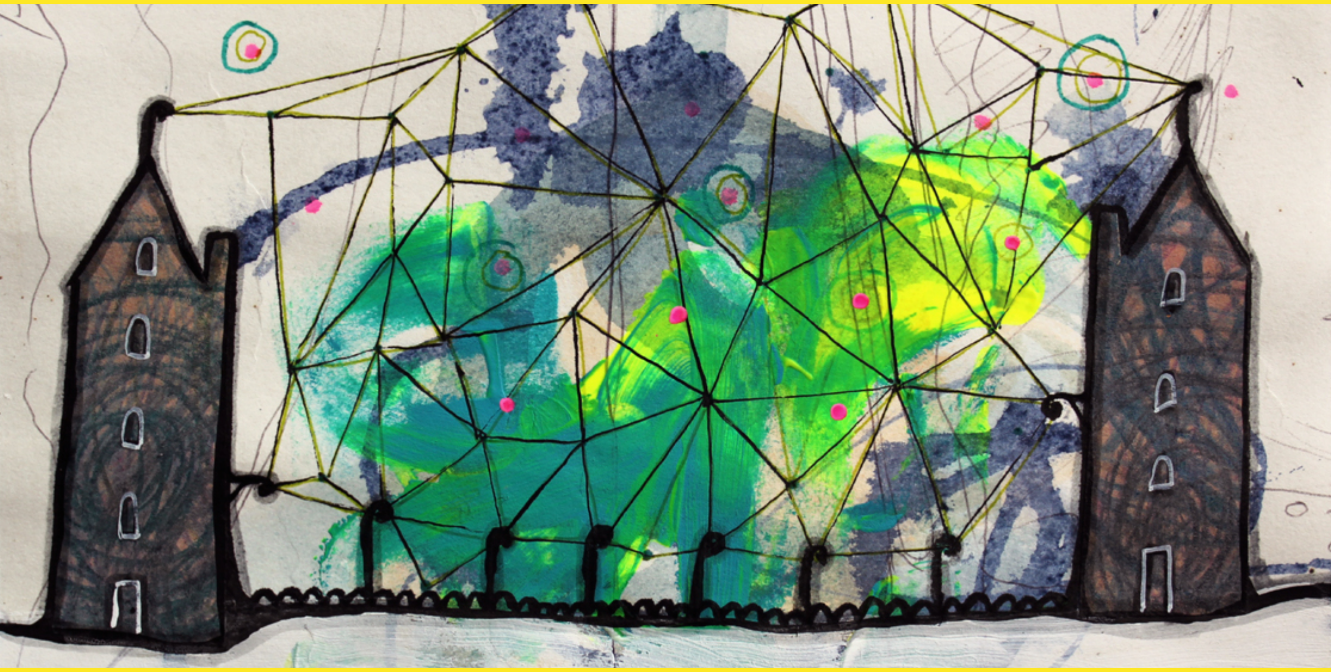
ATTENTIVE SPEAKING

From Listener Feedback to Interactive Adaptation



HENDRIK BUSCHMEIER

ATTENTIVE SPEAKING

From listener feedback to interactive adaptation

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doktor der Ingenieurwissenschaften (Dr.-Ing.)
at the Faculty of Technology at Bielefeld University

by

HENDRIK BUSCHMEIER

May 2017

THESIS COMMITTEE

Prof. Dr.-Ing. Stefan Kopp
Bielefeld University

Prof. Dr. Dirk Heylen
University of Twente

Prof. Dr.-Ing. Britta Wrede
Bielefeld University

Dr.-Ing. Basil Ell
Bielefeld University

Hendrik Buschmeier

Social Cognitive Systems Group, CITEC and Faculty of Technology,
Bielefeld University, PO Box 10 01 31, 33501 Bielefeld, Germany, EU

hbuschme@techfak.uni-bielefeld.de
https://purl.org/net/hbuschme
https://orcid.org/0000-0002-9613-5713

URN: urn:nbn:de:0070-pub-29182956 (e-book)

Von der Technischen Fakultät der Universität Bielefeld zur Erlangung des Grades eines Doktor der Ingenieurwissenschaften (Dr.-Ing.) genehmigte Dissertation.

Thesis committee:
– Prof. Dr.-Ing. Stefan Kopp (Bielefeld University)
– Prof. Dr. Dirk Heylen (University of Twente)
– Prof. Dr.-Ing. Britta Wrede (Bielefeld University)
– Dr.-Ing. Basil Ell (Bielefeld University)

Date of submission: 2017-05-09
Date of defense: 2017-11-28

Typeset in Minion Pro, Lato, and `Inconsolata` by pdfTeX. Created with LaTeX 2$_\varepsilon$, biber, BibLaTeX, booktabs, csquotes, hyperref, memoir, microtype, minted, tabu, Ti*k*Z, et al.

Cover artwork by Jessica Koppe, *(a home for) all the possible structures,* mixed-media collage on paper, 2014. (detail). All rights reserved. Used with permission.

Generated from commit 4f29fdf on 2018-03-06 at 09:03.

∞ The paper used in this publication meets the requirements for
permanence of paper for documents as specified in ISO 9706:1994.

Understanding involves inference that is not only
occasionally defeasible: We almost always *fail.*
Yet we almost always nearly succeed:
This is the paradox of communication.

William J. Rapaport
(2003, p. 402; lightly edited)

TABLE OF CONTENTS

III Evaluation

Appendices

ABSTRACT

Dialogue is an interactive endeavour in which participants jointly pursue the goal of reaching understanding. Since participants enter the interaction with their individual conceptualisation of the world and their idiosyncratic way of using language, understanding cannot, in general, be reached by exchanging messages that are encoded when speaking and decoded when listening. Instead, speakers need to design their communicative acts in such a way that listeners are likely able to infer what is meant. Listeners, in turn, need to provide evidence of their understanding in such a way that speakers can infer whether their communicative acts were successful. This is often an interactive and iterative process in which speakers and listeners work towards understanding by jointly coordinating their communicative acts through feedback and adaptation. Taking part in this interactive process requires dialogue participants to have 'interactional intelligence'.

This conceptualisation of dialogue is rather uncommon in formal or technical approaches to dialogue modelling. This thesis argues that it may, nevertheless, be a promising research direction for these fields, because it de-emphasises raw language processing performance and focusses on fundamental interaction skills. Interactionally intelligent artificial conversational agents may thus be able to reach understanding with their interlocutors by drawing upon such competences. This will likely make them more robust, more understandable, more helpful, more effective, and more human-like.

This thesis develops conceptual and computational models of interactional intelligence for artificial conversational agents that are limited to (1) the speaking role, and (2) evidence of understanding in form of communicative listener feedback (short but expressive verbal/vocal signals, such as 'okay', 'mhm' and 'huh', head gestures, and gaze). This thesis argues that such 'attentive speaker agents' need to be able (1) to probabilistically reason about, infer, and represent their interlocutors' listening related mental states (e.g., their degree of understanding), based on their interlocutors' feedback behaviour; (2) to interactively adapt their language and behaviour such that their interlocutors' needs, derived from the attributed mental states, are taken into account; and (3) to decide when they need feedback from their interlocutors and how they can elicit it using behavioural cues. This thesis describes computational models for these three processes, their integration in an incremental behaviour generation architecture for embodied conversational agents, and a semi-autonomous interaction study in which the resulting attentive speaker agent is evaluated.

The evaluation finds that the computational models of attentive speaking developed in this thesis enable conversational agents to interactively reach understanding with their human interlocutors (through feedback and adaptation) and that these interlocutors are willing to provide natural communicative listener feedback to such an attentive speaker agent. The thesis shows that computationally modelling interactional intelligence is generally feasible, and thereby raises many new research questions and engineering problems in the interdisciplinary fields of dialogue and artificial conversational agents.

ACKNOWLEDGEMENTS

No man is an Iland, intire of it selfe;
every man is a peece of the Continent, a part of the maine;
—John Donne (1624/1923, 17. Meditation, p. 98)

Could I have done the research that led to this thesis entirely by myselfe? No. And I certainly would not have wanted to. I am a peece of the Continent, a part of the maine.

First and foremost, I want to thank my teacher and advisor Stefan Kopp for his guidance, his inspiration, his knowledge, his motivation, his kindness, and his seemingly endless patience. I am very very grateful for his continued support on my way from an undergraduate student to the person I am today. Stefan made everything possible I enjoyed working on these past years: he has given me continued employment in interesting research projects, he has allowed me to pursue my own ideas and side projects, and he has build a lovely research group with a truly positive and supportive atmosphere. From the start he has made me feel being a peer, has given me a lot of freedom, and has — I think — sometimes even valued *my* advice. Thank you, Stefan!

Second, I would like to thank my scientific collaborators. Joint work on sidetrack projects may be one reason why finishing this thesis took a bit longer ☺ than everyone expected. These collaborations were rewarding and memorable social interactions as well as fantastic learning experiences. They pushed me forward professionally and personally. I regret nothing and would like to say thank you to ▹ Zofia Malisz, Marcin Włodarczak, and Petra Wagner from Bielefeld's Phonetics and Phonology group for years of collaboration on the ALICO-corpus that serendipitously grew out of a bi-weekly reading group on 'backchannels, barge-ins, and turn-taking'. ▹ Ramin Yaghoubzadeh for close collaboration on IPAACA, the KDS-1-corpus, the dialogue engine framework, our 'informal conversations', and numerous other activities. ▹ Timo Baumann, David Schlangen, Casey Kennington, and Spyros Kousidis from Bielefeld's Dialogue Systems group for joint work on incremental and adaptive output generation in dynamic situations. ▹ Zofia Malisz, Petra Wagner, and Joanna Skubisz for joint work on and in the SFB 673 sub-project 'Rhythm and timing in dialogue' (B4). ▹ All colleagues from the KOMPASS-project, especially my close collaborators Farina Freigang, Ramin Yaghoubzadeh, Katharina Cyra, Christiane Opfermann, and Lars Schillingmann. ▹ Kirsten Bergmann for joint work on interactive alignment in natural language and gesture generation that grew out of my undergraduate Diploma thesis. ▹ Marcin Włodarczak for joining me in creating and maintaining TextGridTools. ▹ Kristina Striegnitz from Union College for joint work in the GIVE domain investigating the effect of instructions being spoken. ▹ Kirsten Bergmann, Iwan de Kok, Jan Pöppel, Matthias Priesters, Amir Sadeghipour, Herwin van Welbergen, and Ramin Yaghoubzadeh for joint teaching and/or student supervision. ▹ My students Benjamin Dosch and Sebastian Rottschäfer for work on adaptive, incremental natural language generation and online Lombard-adaptive incremental speech synthesis.

Third, I want to thank the members of the Social Cognitive Systems/Sociable Agents research group — past and present. I have always felt very comfortable in your company. Special thanks go to Dagmar Philipp — for being on top of things, knowing how to handle all things administration, and for always having cookies — , and to Farina Freigang — for conversations on life, the university, and everything, as well as for organising fun running activities.

Fourth, I want to thank ▹ Dirk Heylen, Britta Wrede, and Basil Ell for serving on my thesis committee. ▹ Sebastian Rottschäfer and Katharina Suhre who, as part of their Hilfskrafttätigkeit, helped with the evaluation of the attentive speaker agent. ▹ Karolina Buschmeier, Konstantin Buschmeier, Rainer Buschmeier, and Farina Freigang for reading and commenting on parts of this thesis. ▹ David DeVault for giving privileged access to the SPUD$_{coref}$ source code and for extended conversations on NLG and dialogue whenever we met. ▹ David Schlangen for his series of 'Semantik und Pragmatik des Dialogs' seminars, which I tremendously enjoyed attending in 2008; 2011–2013. ▹ Ipke Wachsmuth for his inspiring lectures that sparked my interest for conversational agents. ▹ Filip Szufnarowski for weekly lunch meetings and PhD-student talk. ▹ Daniel Sieger for insightful exchanges about typography and LaTeX.

Fifth, and on a more formal note, I would like to acknowledge support by the German Research Foundation (DFG) via the Cluster of Excellence EXC 277 'Cognitive Interaction Technology' (CITEC) and via the Collaborative Research Centre SFB 673 'Alignment in Communication'. I would further like to acknowledge support by the German Federal Ministry of Education and Research (BMBF) via the project 'KOMPASS' (FKZ 16SV7271K). I am also grateful to my alma mater Bielefeld University, whose mission statement — 'transcending boundaries' — lives up to my personal experience as a student and researcher.

Finally, I want to thank my family — especially my dearest ones, Karo and Leopold.

PRELIMINARY REMARKS

Speakers and listeners Speaking and listening roles in dialogue are neither fixed nor clearly discriminable. Just a moment ago Lieselotte had the role of the 'speaker', and now the turn has already changed and she is the 'listener', attending to and trying to make sense of what Simon has to say. While he is speaking, Lieselotte is neither inactive nor silent, but continues contributing to the dialogue. She is following Simon's gaze, nodding when she understands his words, and displaying a puzzled face when she finds an argument inconclusive. She is uttering *uh-huh*s, *okay*s, or *huh?*s, throwing in a word or phrase here and there, and making short remarks. Simon, while being the speaker, is paying attention to what Lieselotte does, immediately assessing the consequences of her actions and adapting his language and speech. In dialogue, listeners listen and speak, and speakers speak and listen.

This observation is a central theme in this thesis and it makes writing about persons who temporarily occupy one of the roles complicated. In the course of this thesis I often write about communicative acts whose 'sender' is seen as a listener and whose 'receiver' is seen as a speaker. To avoid confusion, I sometimes introduce abstract persons (speakers and listeners) with random names, such as *Simon* and *Lieselotte* in the example above. The names of a speaker always start with letter 'S' and the names of a listener with letter 'L'.

Previously published material and co-authorship Parts of the work presented in this thesis have already been published as peer reviewed workshop, conference, or journal papers. Use of such material is indicated at chapter (or sometimes section) beginnings in special footnotes marked '☆'.

All publications are co-authored by my doctoral advisor Stefan Kopp, whoses co-authorship is not further indicated. When presenting work published with additional co-authors, it is described which parts were contributed by me.

Usually co-authorship meant that individual authors were responsible for writing specific parts of a text. There was, however, a liberal policy on editing. All authors were allowed — even encouraged — to edit any part of the text. Consequently parts that lay in my responsibility may have been altered in ways that are difficult or even impossible to unravel.

In addition, some research was done and some texts were written in a highly collaborative style — multiple authors being co-present in a room with a whiteboard and a computer. These ideas and texts are the result of 'distributed cognition' (Gureckis and Goldstone 2006), to which traditional concepts of authorship cannot be applied.

CHAPTER 1
INTRODUCTION

We begin by briefly introducing the essential background of this dissertation. Simultaneously we make a case — rooted in research on conversational interaction in the fields of conversation analysis, linguistics, philosophy, and psychology — for a shift of focus in artificial conversational agent research: from natural language processing to interactional intelligence. Following this, we describe the objective and scope of this thesis and provide an overview of the text.

1.1 CONTEXT AND MOTIVATION

Artificial conversational agents — such as spoken dialogue systems (McTear 2002), embodied conversational agents (Cassell et al. 2000), or sociable robots (Fong et al. 2003) — are computational artifacts that apply natural language processing and artificial intelligence (AI) techniques in order to be able to interact with humans using spoken language (possibly accompanied by non-verbal communicative acts). The abilities of these agents range from reacting to a limited set of spoken commands to taking part in face-to-face dialogues — in restricted domains — that may result in fairly coherent discourses.

With automatic speech recognition now performing as good as humans in recognising conversational speech (Xiong et al. 2017), it may seem that scaling artificial conversational agents to operate in larger, possibly open ended, domains is mainly a question of improving natural language processing techniques. Natural language processing, however, is commonly thought to be an 'AI-complete' problem[1]. Understanding and generating natural language requires linguistic knowledge on all levels of

[1] AI-complete problems are, rather informally, defined as requiring an artificial general intelligence ('the synthesis of a human-level intelligence', [Raymond 2003, ll. 4734–4751]), perhaps even 'strong' artificial intelligence — the attainability of which is questionable (cf. Russell and Norvig 2010, § 26.2).

language processing (phonetic, morphological, lexical, syntactic, prosodic, semantic, and basic pragmatic knowledge), it requires computational models (algorithms and representations) for these processes, as well as, and perhaps most importantly, extensive world knowledge. All this makes it unlikely that a general solution to natural language processing will be found in the foreseeable future (Jurafsky and Martin 2000, pp. 702–703).

An additional problem for artificial conversational agents is that 'speaker meaning' (what the speaker intents to communicate; Grice 1957; 1975) does not, in general, correspond to the 'literal meaning' of communicative acts. Relying on syntax, semantics, and basic pragmatics does not get listeners far in understanding what the speaker means. 'Code models' of communication (e.g., Shannon's [1948] mathematical model of communication) — often implicitly assumed in modelling artificial conversational agents — do not even consider this. The only explanation they offer for problems in communication is a 'noisy channel'. Reasons for the difference between speaker meaning and literal meaning, however, are that the former is only implicated, that communicative acts contain multiple intentions (a communicative intention and an informative intention; Sperber and Wilson 1986/1995, §§ 1.11, 1.12), and that dialogue participants are individuals with private beliefs and subjective conceptions of language and the world — that is, ultimately they do not share 'the code'. In comparison, the problem of the noisy channel seems almost negligible.

Using lexical semantics, syntax, and discourse relations to derive the semantics of an utterance, or to classify its speech/dialogue act (these are core natural language processing tasks) plays an important role in deriving speaker meaning (i.e., 'understanding' the speaker). Yet, it yields merely one source of information among several (the context, such as common ground, discourse context, situation, expectations, beliefs about the interlocutor, and others). Deriving speaker meaning from communicative acts using a variety of sources of information is called 'pragmatic reasoning'. It is based on the principle that interlocutors in conversation are fundamentally cooperative (Grice 1975) and it relies on ostension[2] — on the side of the speaker — and inference — on side of the addressee — (Sperber and Wilson 1986/1995, § 1.10).

Yet, in order for interlocutors to 'understand' each other, abilities even beyond natural language processing and pragmatic reasoning are needed. An inferential process based on a multitude of information sources — which are not completely shared among interlocutors — is basically guaranteed to draw wrong conclusions, i.e., to come to an interpretation of a communicative act that differs from what the speaker

[2] With ostensive behaviour (a technical term from relevance theory; Sperber and Wilson 1986/1995, p. 49), communicators go beyond 'making manifest' *what* they intent to communicate (the informative intention), by simultaneously making manifest *that* they want to communicate (the communicative intention).

meant. This is why 'understanding involves inference [that] is not only occasionally defeasible' and directly leads to the 'paradox of communication' (Rapaport 2003, p. 402; see this thesis' epigraph on page iii): In trying to establish understanding, interlocutors in dialogue

> [...] almost *always* fail [...]. Yet [they] almost always nearly succeed.

It is paradoxical that although it should theoretically be highly improbably for interlocutors to understand each other, our everyday experience actually suggests the opposite. Participants in conversation commonly seem to understand what the other 'means'. As a resolution Rapaport proposes that:

> Misunderstandings, *if small enough,* can be ignored. And those that cannot be ignored can be minimized through negotiation.

In order for pragmatic reasoning to draw conclusions that are 'less wrong', the information sources that the speaker uses to produce ostensive actions and the information sources that underlie the listener's inference process should be as similar as possible. This is the reason why interlocutors in conversation are, to a large extent, concerned with establishing 'common ground' (Clark 1996) and making beliefs shared.

Conversation is an 'interactive' endeavour in which interlocutors 'coordinate' the information sources underlying ostensive–inferential communication. Such coordination happens on multiple levels. Dialogue participants coordinate their behaviour,[3] but also their beliefs and attitudes (Kopp 2010) — relying, to some extent, on the results of behaviour coordination. Beliefs and attitudes are subjective mental states of interlocutors which, even though they are not directly observable, are crucial information sources for ostensive–inferential communication. For communication to be possible and efficient, interlocutors thus need to represent the beliefs and attitudes that their dialogue partners are likely holding. Coordination happens via communication itself, in an activity that can be compared to a negotiation (Rapaport 2003), or to 'hypothesis testing' (Brennan 1990). It involves multiple interacting processes within and among dialogue participants:

[3] Coordination mechanisms on the level of behaviour provide the basic infrastructure for interaction to be possible. These include, inter alia, establishing and maintaining contact, turn taking (i.e., who speaks when), regulating the flow of information (e.g., with backchannel-feedback), assessing and establishing gaze-based attention (focus, shared attention, joint attention), as well as 'alignment' (Pickering and Garrod 2004) of surface structure (e.g., words, syntactic structures) and non-verbal behaviours (e.g., speech-accompanying gesture).

- Forming a representation of others' beliefs and attitudes and updating them given new information (the general ability of 'mentalising' [Frith and Frith 2006]).
- Providing meta-communicative information about one's own beliefs and attitudes (for example by providing communicative feedback [Allwood et al. 1992], by requesting clarification [Purver 2004], or by producing a relevant response).
- Designing communicative acts against the background of the addressees' likely hold beliefs and attitudes, thereby increasing the probability that they will be able to infer speaker meaning (through 'audience/recipient design' [Clark and Marshall 1981; Clark and Brennan 1991] or 'monitoring and adjustment' [Horton and Keysar 1996]).

Coordination, using these processes, may result in meaning to be communicated iteratively over multiple turns. A speaker produces a communicative act, the listener responds, for example, by providing feedback of non-understanding, thereby revealing to the speaker that some of the act's presuppositions on common ground are not satisfied. The speaker refines his model of the listener's beliefs and attitudes by incorporating her feedback, and attempts a new (or adapted) communicative act, the design of which takes the updated representation into account. This loop of feedback and interactive adaptation continues, with the listener's success in inferring speaker meaning (i.e., understanding) gradually improving, until, at some point, both dialogue partners mutually believe that an understanding 'sufficient for current purposes' has been reached (Clark and Schaefer 1989).

The actual amount of coordination that is needed depends on many factors, e.g., on the complexity of the meaning that a speaker intends to communicate, on the extent of common ground between dialogue partners, and on the accuracy of the participants' partner models. Simple and highly conventionalised meaning, especially in situations where it can be expected, may not require much coordination as speaker meaning is derived easily. In general, however, conversation is a highly dynamic coordination process that involves both dialogue partners and requires that they are willing to jointly work towards understanding (Clark 1996).

This shows that in addition to being able to process natural language, having knowledge about language and the world, and being able to reason pragmatically, artificial conversational agents need to master these interactive coordination processes that are at work in conversational interaction. They need 'interactional intelligence'[4]

[4] Levinson (2006) postulates an 'interaction engine' which serves as the universal cognitive and behavioural foundation for interaction, including conversation. Recent theories of language evolution argue con-

(Levinson 1995), which consists of two general abilities (ibid., pp. 254–255) that were basically described above, namely, being able 'to attribute intention to other agents' actions [. . .], and to respond appropriately in interdigitated sequences of actions'. Not only is interactional intelligence necessary for being able to engage in conversation, in some cases it may even be sufficient. This is supported by the fact that successful (albeit less smooth and efficient) dialogical interaction is possible even in the absence of language, for example between people not sharing a common language, or between parents and infants (Levinson 2006, pp. 40–42).

Research on artificial conversational agents should therefore strive to focus more strongly on computational models of pragmatic reasoning[5] and interactional intelligence. Agents endowed with such skills will, potentially, be better communicators. Being able to engage in interactive and ostensive–inferential communication, they would likely be more robust in understanding others' communicative acts (especially in unplanned domains and situations), and also more effective in making their own communicative acts understood by others. When reaching the limits of their natural language processing capabilities, interactional intelligence would allow artificial conversational agents to try to establish understanding using interactive means such as the ones described above. Consequently, the natural language processing components of such agents need not be perfect, circumventing the problem of AI-completeness of natural language processing.

1.2 OBJECTIVE AND SCOPE

The general objective of this dissertation is to investigate how interactional intelligence can be modelled computationally for artificial conversational agents.

We approach this research question by focussing on a specific, seemingly small, but ubiquitous interactional phenomenon in dialogue: 'communicative listener feedback'. In this thesis we use this term to encompass short verbal/vocal expressions (such as *mhm*, *yeah*, or *huh?*; often called 'back-channels' elsewhere) as well as non-verbal, embodied signals (head gestures, facial expression, and gaze). Listeners in conversation produce communicative feedback in response to (and often in overlap with) utterances produced by their interaction partners in order to communicate — not necessarily

vincingly that interactional intelligence has served as the basis for the development of ostensive–inferential communication, which has then enabled the development of conventional signals and, based on that, combinatorial communication, i.e., language (Scott-Phillips 2015).

[5] Computational models of pragmatic reasoning were a topic in artificial intelligence and computational linguistics in the 1980s and 90s — then termed 'plan recognition', see, e.g., Cohen et al.'s (1990) edited collection 'Intentions in communication'.

intentionally — whether they are in contact and whether they are willing and able to perceive, understand, accept, and agree (Allwood et al. 1992).

The interactional relevance of feedback in conversation can be considered well-established in conversation analysis, linguistics, and psycholinguistics.[6] Research carried out in these fields has revealed that feedback is a nuanced, multi-dimensional mechanism that operates on all levels of dialogue coordination and serves important interactional functions.

Over the past two decades listener feedback has also been an active research topic in artificial conversational agents.[7] The majority of the computational work focusses on the generation of feedback behaviours in response to human speech: when is it appropriate to produce feedback, which form should it take, how does the interlocutor perceive the agent and its behaviour, etc. Endowing agents with feedback generation models is mainly seen as a means to make them more human-like (Edlund et al. 2008) and believable interaction partners that are pleasant to interact with, as well as to make them appear attentive, thus encouraging their interlocutors to continue. Interactional functions that exceed the level of behaviour coordination are usually not considered in these models.

The research we present in this thesis is different in that it aims to conceptually and computationally model the processes that play a role in interactive feedback-based coordination on the levels of belief and attitude. We focus on these aspects of communicative listener feedback in order to use it as an interactional means for efficiently establishing understanding and evaluating acceptance and agreement in conversational interaction.

In contrast to most of the computational work on feedback in dialogue, which develops models that make artificial conversational agents come across as 'attentive listeners', the work in this thesis models the processes that are in operation when artificial conversational agents hold the turn, that is, models for 'attentive speaking'. An 'attentive speaker agent'[8] is an artificial conversational agent that has the 'desire' to

[6] For example, Stolz and Tannenbaum (1963), Yngve (1970), Kraut et al. (1982), Schegloff (1982), Heritage (1984), Ehlich (1986), Goodwin (1986), Allwood et al. (1992), Clark (1996), Bavelas et al. (2000), and Bunt (2012).

[7] For example, Novick and Sutton (1994), Cassell and Thórisson (1999), Ward and Tsukahara (2000), Cathcart et al. (2003), Heylen et al. (2004), Gratch et al. (2006), Kopp et al. (2008), Bevacqua (2009), Morency et al. (2010), Wrede et al. (2010), Gravano and Hirschberg (2011), Poppe et al. (2011), Reidsma et al. (2011), Schröder et al. (2012), Inden et al. (2013), de Kok (2013), Neiberg et al. (2013), Skantze et al. (2014), and Oertel et al. (2016).

[8] A term simultaneously developed by Reidsma et al. (2011), who investigated communicative listener feedback from the perspective of a speaker as well.

be understood[9] by its interlocutors. This desire translates to a general willingness to work towards being understood (sufficiently well for current purposes), and to make extra efforts — if necessary — to achieve this. This dissertation spells out conceptually and computationally the implications of this in terms of three interacting processes:

Listener state attribution An attentive speaker agent should theorise about the dynamic mental states of understanding (and other listening-related mental states, namely contact, perception, acceptance, and agreement) of its interlocutors. To achieve this, the agent should engage in probabilistic inferential attribution of these states (so called mentalising), based on evidence in form of communicative feedback signals provided concurrently by the interlocutors in response to the agent's ongoing behaviour. Importantly, the attribution process should factor in the agent's utterance and its representation of dialogue context, thus embodying a 'probabilistic pragmatics' approach (Goodman and Frank 2016; Franke and Jäger 2016) to the problem of what received feedback signals mean.

Interactive adaptation An attentive speaker agent should make efforts to communicate its meaning as effectively as possible until it is sufficiently well understood by its interlocutors. To achieve this, the agent should adapt its natural language production processes, on different levels of processing (e.g., dialogue management, natural language generation, speech synthesis), with the goal of generating communicative acts specifically designed in the light of the listening-related mental states currently attributed to its interlocutors. Adaptations should not be limited to reacting to problems in understanding, but also include the ability to spare efforts if understanding is achieved without problems.

Feedback elicitation An attentive speaker agent should make efforts to lead its interlocutors to provide as much communicative feedback as is needed to be well informed about their listening-related mental states. To achieve this, the agent should be able to decide when and to know how to produce behavioural cues that elicit communicative feedback.

[9] Attentive speaker agents are also interested in their interlocutors' attitudes towards what was meant and understood. Does the interlocutor accept what the agent meant? Does the interlocutor agree with the agent? Artificial attentive speakers agents do not try to persuade their interlocutors of their proposals or opinions, though. (They prefer to leave such delicate activities to humans.)

In the form described above, the three processes encompass the two abilities claimed to underlie interactional intelligence (Levinson 1995, pp. 254–255): attribution of mental states (here based on listener feedback), and responding appropriately (here through interactively adapted behaviour). Our research hypothesis thus is that attentive speaker agents endowed with computational models of these processes will be able to engage in a simple[10] and effective form of interactive coordination — Rapaport's (2003) 'negotiation' — on the levels of belief and attitude. Measurable correlates of this hypothesis, to be evaluated in interactions between an attentive speaker agent and a human interlocutor, should be:

- The attentive speaker agent and its interlocutors iteratively establish understanding (and evaluate acceptance and agreement) in a loop of feedback and continuous adaptation of the agent's communicative behaviours.
- Interlocutors notice that the attentive speaker agent is interested in, and able to, infer their mental state of listening and responds appropriately.
- Interactions with an attentive speaker agent will be better than interactions with non-attentive speakers in that higher understanding will be reached, in a more efficient way.

As is often the case when developing embodied conversational agents, implementing a system that humans can interact with is a challenging engineering task that spans multiple disciplines (Isbister and Doyle 2004).

As outlined above, communicative listener feedback is a topic in conversation analysis and various areas of linguistics (phonetics, semantics, pragmatics, dialogue, psycholinguistics). The inference-based mentalising processes of the attentive speaker agent fall into the fields of cognitive science and artificial intelligence. Adaptive natural language production on multiple levels falls into different areas of computational linguistics (dialogue systems, natural language generation, speech synthesis), as does the principle of 'incremental processing' — which is applied throughout the agent in order to enable timely reactions to interlocutor feedback and to allow adaptations to ongoing utterances. The generation of multimodal behaviour for the attentive speaker agent falls into the field of intelligent virtual agents.

In developing conceptual and computational models of the processes that play a role in interactive feedback-based coordination and in implementing them in an attentive speaker agent, this dissertation makes an interdisciplinary contribution and cuts across the boundaries of the fields mentioned above. The venues in which parts of the research have already been published and the thesis' bibliography reflect this.

[10] Because it is restricted by the expressiveness of feedback which is limited in contrast to language itself.

1.3 OVERVIEW

Having described the aims and scope of the research described in this dissertation, we now continue with an overview of its structure.

This thesis consists of three parts. Following this introductory chapter, part I, 'Dialogue Coordination in the Human and the Machine' provides the theoretical, cognitive, and computational background.

Chapter 2, 'Communication, Dialogue, and Coordination', deepens the perspective on communication in dialogue for three aspects that are key to this thesis: understanding in dialogue, common ground and grounding, and adaptation in language production.

Following this, chapter 3, 'Communicative Feedback', introduces communicative listener feedback by describing the origins of the theoretical concept of feedback and how it relates to the dialogue phenomenon. It then illustrates the rich form of feedback signals, their pragmatic functions, and the timing of feedback signals, closing with a review of computational models of feedback processing in artificial conversational agents.

Part II, 'A Computational Model of Attentive Speaking', describes the conceptual modelling of the three computational processes that are at work within the attentive speaker agent.

Chapter 4, 'Interactional Intelligence for Attentive Speaking', introduces the second part of the thesis and argues, given the background, that attentive speaking is a subset of the processes of general interactional intelligence for artificial conversational agents that is both interesting and feasible to model.

Following this, chapter 5, 'Mental State Attribution Based on Communicative Listener Feedback', develops the probabilistic inferential computational model for feedback interpretation based on attribution of listening-related mental states.

Chapter 6, 'Interactive Adaptation', then describes levels and mechanisms for adaptation of speech, and develops models for incremental and adaptive natural language generation that takes the representation of attributed listener state into account when making adaptive generation decisions.

Following this, chapter 7, 'Feedback Elicitation', develops a model for feedback elicitation based on a concept of the agent's information needs with regard to the attributed listener state.

Part III, 'Evaluation', describes the implementation and evaluation of the attentive speaker agent.

Chapter 8, 'Bringing it Together: An Attentive Speaker Agent', describes the IU-model for incremental processing in dialogue and discusses limitations that standard approaches to behaviour generation face when producing incremental real-time adaptive behaviour. Following this, the chapter then describes the implementation and integration of the models developed in part II of this thesis in a novel incremental behaviour generation architecture for artificial conversational agents.

Chapter 9, 'Evaluation of the Attentive Speaker Agent', then develops an evaluation strategy for the attentive speaker agent. It describes the study (which uses a semi-autonomous Wizard-of-Oz paradigm) and analyses and discusses the results.

Finally, chapter 10, 'Conclusion', concludes this thesis by briefly summarising its results, discussing its contributions and their implications, as well as its limitations and future research directions.

The conclusion chapter is followed by two appendices. Appendix A, 'Model Parametrisation from Implicit Representation', describes the approach and algorithm underlying the Attributed Listener State model developed in chapter 5. Appendix B, 'Study Materials', contains material used in the evaluation study, namely the written as well as one exemplary oral instruction given to participants.

The back matter contains, in addition to the Bibliography, a list of Accompanying Resources (e.g., model parameters published in form of data publications, software packages).

PART I

DIALOGUE COORDINATION IN THE HUMAN AND THE MACHINE

CHAPTER 2
COMMUNICATION, DIALOGUE, AND COORDINATION

In this chapter we further develop the perspective on communication in dialogue[11]and conversation that we sketched in section 1.1 of the introduction. We will elaborate on three aspects that are particularly relevant in the context of this thesis: the notion of understanding and its opposites misunderstanding and non-understanding, the importance of common ground as a shared basis for understanding and how it comes about, and different approaches to adaptation in dialogical interaction. We then analyse how far these concepts have been integrated into artificial conversational agents.

2.1 UNDERSTANDING, MISUNDERSTANDING, AND NON-UNDERSTANDING

The central goal of conversation and dialogue is 'understanding'. Speakers producing an utterance, a communicative act, want to achieve a communicative effect in their addressees. They want them to understand the 'speaker's meaning' (Grice 1957; 1975). Addressees, on the other hand, have a 'duty of understanding' (Dascal and Berenstein 1987), i.e., they are obliged to '[find] out [...] what is the speaker's meaning' (ibid., p. 140).

Focussing on the conversation, speakers want to make themselves understood more generally. Telling a story, they want their addressees to follow. In an argument,

[11] In this dissertation we generally consider 'dialogue' to be a dyadic, primarily language-based, face-to-face interaction between situated individual agents. Using the term dialogue in its 'concrete, empirical sense' (Linell 2009, p. 4). That the focus is on dyadic interactions is a pragmatic and not a principled choice. Theories and models developed for dialogues with two participants may — with some adjustments — very well scale to polyadic dialogues involving more participants.

they might want to convince their addressees of their position, or at least make their position clear to them. In a negotiation, speakers may want to reach a compromise that both parties can live with or make a joint decision of some sort. All this involves understanding on the utterance, the sub-utterance, and the super-utterance level. This raises the questions of what understanding means in principle, i.e., when a listener can be considered having understood.

2.1.1 UNDERSTANDING

Since antiquity, 'understanding' in language and communication has been thought of as extracting exactly the 'message' from an utterance that the speaker intends to convey (see Taylor 1986).[12] As Schlesinger and Hurvitz (2008, p. 569) put it, a listener l understands the message of a speaker s, if

$$msg(s, X_s) \wedge msg(l, X_l) \wedge X_l = X_s, \qquad (2.1)$$

where X_s is assigned the message intended to be conveyed by the speaker and X_l is assigned the message extracted by the listener. According to this view, understanding thus means that the listener recovers exactly the intended message that the speaker conveyed.

Starting from this, Schlesinger and Hurvitz broaden the definition to include cases in which speakers intend to be misunderstood, which listeners may notice or not. A speaker may, for example, have a 'speaker's-intention' $s.int(s, msg(l, X_{s.int}))$ that the listener extracts a message from the utterance that differs from the message $msg(s, X_s)$ that she actually intends her utterance to convey — with $X_{s.int} \neq X_s$. This is for example possible when the utterance is ambiguous, or contains an implicature that is achieved by leaving something unsaid. From a listener's perspective, this intended misunderstanding $l.persp(msg(s, X_s), s.int(s, msg(l, X_{s.int})))$ is either noticed, i.e., $X_s \neq X_{s.int}$, or not, i.e., $X_s = X_{s.int}$. Furthermore, the listener may falsely attribute an 'intention to be misunderstood' to the speaker.

With this in mind, Schlesinger and Hurvitz define 'understanding' in terms of three properties (ibid., p. 577):

1. 'The message recovered by the [listener] is identical to the one the speaker intended the utterance to convey', see eq. (2.1).

[12] In the code model of communication — which combines information theory, the conduit metaphor of communication, and de Saussure's speech circuit (Blackburn 2007) — this concept is still very prominent today, both in linguistics as well as in dialogue system research.

2. 'The [speaker's intention] does not differ from what the [listener] actually recovered':

$$s.int(s, msg(l, X_{s.int})) \wedge msg(l, X_l) \wedge X_{s.int} = X_l \tag{2.2}$$

3. 'There is no discrepancy between the [listener perspective] and the corresponding' message the speaker intended to convey and the speaker's intention:

$$\begin{gathered} l.persp(msg(s, X_s), s.int(s, msg(l, X_{s.int}))) \wedge \\ msg(s, X_s) \wedge s.int(s, msg(l, X_{s.int})) \wedge X_s = X_{s.int} \end{gathered} \tag{2.3}$$

Schlesinger and Hurvitz argue that these properties are jointly necessary and sufficient to say that a listener understands a speaker's utterance. If one or more of these properties do not hold in a communicative interchange between a speaker and a listener the result is a 'misunderstanding' (ibid., p. 577).

2.1.2 STRONG AND WEAK CONCEPTS OF UNDERSTANDING

The classic view reflected in Schlesinger and Hurvitz's definition is what we want to call 'strong' understanding[13], a theoretic ideal that most possibly can never be achieved in actual communication between humans. Taylor (1986; 1992) shows that the 'utopian' nature of the concept of strong understanding was already recognised by Locke (1690/1979), who sees ideal communication as 'telementation' — the flawless transfer of 'ideas and thoughts' from the speaker's mind to the listener's mind (Taylor 1986, p. 171). Telementation supports a strong concept of understanding such as presented in eq. (2.1). Examining the implications of a strong concept of understanding for successful language-based human communication, Locke (1690/1979, pp. III, ix, 6) notes that

> To make words serviceable to the end of communication, it is necessary [...] that they excite in the hearer exactly the same idea they stand for in the mind of the speaker. Without this, men fill one another's heads with noise and sounds; but convey not thereby their thoughts, and lay not before one another their ideas [...].

[13] Based on the distinction between the strong (machines can — in principle — think) and weak (machines can only behave intelligently) artificial intelligence hypotheses (Russell and Norvig 2010, pp. 1020–1033).

As an empiricist, Locke is, however, aware that language-based communication has a severe limitation: the 'imperfection of words'. According to Locke, words signify 'private' ideas and it cannot be taken for granted that the private idea a speaker signifies with a certain word (or utterance) evokes the same — or a qualitatively indistinguishable — idea in the listener (though Locke claims that it is not impossible per se; Taylor 1992, pp. 33–36).

When analysed further, the imperfection of words consists of two independent subproblems: the 'conformity of representation' and the 'conformity of intersubjectivity' (ibid., p. 28). According to the conformity of representation, people commonly assume that the concepts they hold are adequate — objective — representations of the things they are concepts of. Concepts are, however, shaped through the experiences an individual makes in life (e.g. Rosch 1978; Barsalou 1999; Hampton 1999), that is, highly subjective and private. Hence, identity, sameness, or qualitative indistinguishableness of concepts across individuals are not plausible requirements for understanding. The conformity of intersubjectivity takes on the link between words and the concepts they are linked to. Again, people commonly hold the assumption that the concepts they and others signify with certain words correspond. As is the case in the formation of concepts, these associations, the lexical semantics, are, formed through language use and experience of the individual speaker (Tomasello 2003), which means that they are subjective as well.

The imperfection of words poses a fundamental problem for strong understanding and Locke's conclusion therefore is that language as a means of communication is insufficient for precisely conveying ideas and thoughts from a speaker to a listener. In particular, Locke thinks that communicators 'are mistaken to take for granted the belief that [they] ordinarily understand each other' (Taylor 1992, p. 36).

It should be noted that Locke (1690/1979), carrying out a general investigation into the nature of human knowledge, was primarily concerned with communication among 'philosophers' and less with 'civil and common conversation'. His writing suggests that such ordinary uses of language consist mostly of reference to objects and that a precise understanding of invisible properties of objects is, in general, not needed (ibid., pp. III, ix, 15).

For Locke, a 'weaker' form of understanding — in which the insufficiency of language does not pose an insurmountable problem — seems to be sufficient in everyday communication. Revisiting the description of the conformities of representation and intersubjectivity above, some concessions can be made. Even though concepts are subjective, they are acquired in a common natural and cultural environment. It can therefore be assumed that individuals acquire concepts that have some similarity to the concepts acquired by others. Although this similarity of concepts across individuals may be inadequate for achieving strong understanding among communicators — the concepts are not the same or qualitatively indistinguishable — it might be sufficient for a weaker form of understanding to be principally possible.

The same holds for the conformity of intersubjectivity. According to the socio-pragmatic theory of word learning (Tomasello 2003, § 3.3.3), meanings are learned through language use in social interaction, in which interlocutors share joint frames of attention. During these activities, which often follow certain interaction patterns and are repeated over and over, associations between words and concepts are established.[14] In this way, conventions of word meanings in a linguistic community are spread and acquired, as well as (implicitly) negotiated and established (Lewis 1969/2002). A conformity of the intersubjectivity of word–concept associations cannot be guaranteed of course, but it can nevertheless be assumed that these associations are sufficiently similar across individuals for a weak form of understanding.

These concessions to the imperfection of words do not change its nature as a fundamental problem for strong understanding. Strong understanding remains impossible, or at least highly unlikely. But the concessions made above weaken the premises of strong understanding and clear the way for a weaker concept of understanding. In parallel to eq. (2.1), we can say that a speaker *s* weakly understands a listener *l*, if

$$msg(s, X_s) \wedge msg(l, X_l) \wedge X_l \approx X_s, \tag{2.4}$$

that is, if the message $msg(l, X_l)$ that *l* extracts from the utterance is similar to the message $msg(s, X_s)$ that *s* intended the utterance to convey. In a similar way the two properties in eqs. (2.2) and (2.3) can be adapted to represent weak understanding.

2.1.3 NON-UNDERSTANDING AND MISUNDERSTANDING

Having considered the different reasons for problems in understanding as well as the fact that understanding is subjective, we can identify two different types of problems in understanding. The distinction is based on the listener's awareness of problems in understanding (Weigand 1999; Schlesinger and Hurvitz 2008).

If a listener identifies a problem, she beliefs that she does not understand (or has difficulties understanding) the speaker for a specific reason, e.g., because she did not perceive one of the words, does not know the meaning of this word, or cannot infer the pragmatic meaning of the utterance. These are cases of 'non-understanding' and, depending on how acute the understanding problem is, we can either speak of

[14] Although conceived for language acquisition in children, it can be assumed that similar mechanisms are at work over the whole span of life. Language use relies on patterns and repetitions as well.

'partial' or 'total non-understanding'. Total non-understanding is rare (Schlesinger and Hurvitz 2008) and needs clarification. Partial non-understanding can often be recovered — with some effort on the side of the listener that does not involve the speaker — by taking the context into account. If such a recovery is not possible, the listener may initiate repair on the side of the speaker (Schegloff et al. 1977), for example, by providing feedback (Allwood et al. 1992) or by requesting clarification (Purver 2004; Schlangen 2004).

If, on the other hand, the listener is not (immediately) aware that she only partially understands (or even does not understand at all), we speak of a 'misunderstanding' (Weigand 1999). In these cases the listener is confident that she understood what the speaker meant and it may (if at all) only become apparent later on in the dialogue that this assumption was false. Misunderstandings can be detected by the speaker, who can explicitly point them out to the listener or provide information so that the listener becomes aware of the misunderstanding herself. Listeners may also discover — coincidentally — that a misunderstanding occurred, e.g., when further information, which seems incompatible with her model of the discourse, becomes available. In such cases it becomes a mere non-understanding and the listener may then either revise her discourse model (DeVault 2008) — not necessarily making the temporarily present misunderstanding public — or may, again, initiate repair.

Depending on the severity of an unresolved misunderstanding it may cause a sequence of inconsistencies in the discourse models of both interlocutors and subsequent occurrences of non-understanding. This may lead to a discovery of the misunderstanding. The important point here is that misunderstandings may persist over extended periods of time until they are finally noticed and identified. If not crucial for the ongoing dialogue, misunderstandings may even remain undetected (though Weigand [1999, p. 770] claims that they are usually resolved over the course of the dialogue), in which case we can speak of 'miscommunication'.

2.1.4 WAYS OUT: APPROACHING UNDERSTANDING

As an intermediate conclusion, we can say that for both dialogue participants it is uncertain whether a listener understands — even weakly — a speaker. As we have seen, the reasons for this are manifold, but hinge on the central problem of subjectivity and privateness of ideas and thoughts. Only an omniscient observer would be able to determine with certainty whether strong understanding is present or not. For actual interlocutors in a dialogue (as well as for people in other listener roles such as bystanders, overhearers, or researchers) this information is out of reach. The uncertainty

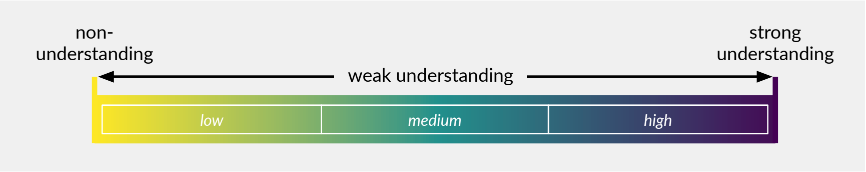


Figure 2.1: Non-understanding and strong understanding define the opposite extremes of an understanding continuum, with various (possibly infinite) degrees of weak understanding in between.

whether understanding is present or not cannot even be completely eliminated for weak concepts of understanding. Interlocutors can only try to quantify and reduce it.

Quantification of understanding is, in principle, possible as the conceptual distinction between strong and weak understanding and the possibility of non-understanding readily suggest that understanding can be seen as a gradual quality (Bazzanella and Damiano 1999). Given this view, non-understanding and strong understanding form the extreme ends of an understanding continuum. Grades of weak understanding — being measurable due to the use of similarity instead of equality in its definition eq. (2.4) — occupy the space in between (fig. 2.1 illustrates this continuum of understanding).

Weigand (1999, p. 765) proposes to see misunderstanding as 'the standard case' in communication, that is, *not* to

> regard cases of misunderstanding or miscommunication as deviant[, but to consider] language use as inherently problematic, and miscommunication not as a failure but as part and parcel of the act of communication.

Here Weigand takes Locke's (1690/1979) insight that interlocutors should not take understanding for granted (see section 2.1.2) and derives a strong premise for communication from it: interlocutors expect non-understanding and misunderstanding to be inevitable. Weigand argues that linguistic interaction in dialogue is not, in the first place, about the cognitive state of 'understanding', but about the process of 'coming to an understanding' (Weigand 1999, p. 769). Interlocutors allow the occurrence of non-understandings and misunderstandings to happen because they can be confident that these problems will be detected and corrected through socially interactive means. Language-based communication works well not despite, but due to the fact that problems arise frequently. Since participants expect nothing more than to eventually come to an understanding at some point in the future and know that difficulties will likely arise on the way, dialogical interaction is inherently robust. If dialogue would be more similar to telementation, as code models of communication basically suggest, the occurrence of an understanding problem would result in a breakdown of communication.

In the following two sections we will describe the process of coming to an understanding.

2.2 COMMON GROUND AND GROUNDING

Coming to an understanding — even communication in general — is impossible without something which both dialogue partners have a conceptualisation of and can relate to. They need some prior 'shared basis' which gives rise to 'common ground' (Clark 1996, p. 92; Stalnaker 2002, p. 701).

Humans, for example, fundamentally share their humanness, i.e., they share the same basic needs (for food, sleep, shelter, etc.), and the same basic experiences from their ontogenesis (such as being cared for, interacting with other humans and objects). They also have similar bodies, which they can control, and similar sensory systems with which they perceive their environment. Such a minimal shared basis is enough to bootstrap communication. Humans do not even need to speak the same language or have a similar cultural background for being able to share, at least some, meaning (Levinson 2006, pp. 254–255) — although this helps of course.

A shared basis, regardless of its characteristics, can be considered a necessary condition for communication. It is, however, not a sufficient one. Even if a shared basis is 'objectively present' — i.e., a hypothetical omniscient observer is able to identify a set of relevant beliefs that are present in all agents involved in an interaction — individual agents might be ignorant of its presence (cf. DeVault and Stone 2006, p. 141). In this case the shared basis is of no use for communication. It is important that interacting agents are also 'subjectively aware' of this shared basis (Clark 1996), i.e., that they have good reasons to belief that a certain belief is also held by their interaction partners and that the interaction partners are aware of this as well. Based on Lewis (1969/2002, p. 56), Clark (1996, p. 94–96; here in an adapted form) develops a definition of 'common ground', in which 'a proposition p is common ground [for two agents S and L] if and only if:

1. [S and L] have information that basis b holds;

2. *b* indicates to [*S* and *L* that both] have information that *b* holds;
3. *b* indicates to [*S* and *L*] that *p*.'

Clark includes the joint situation of the two agents as an important aspect of the shared basis. Both agents need to be aware of their shared basis *b*. Common ground, once established, can serve as a shared basis in the process of establishing new common ground as well.

2.2.1 GROUNDING

The process of updating the common ground (for example by adding propositions) is called 'grounding' (Clark and Brennan 1991; Clark and Schaefer 1989; Clark 1996, p. 221) and is achieved by dialogue partners by making 'contributions' to the discourse (Clark and Schaefer 1989). The central property of contributing to discourse is that contributions are not actions of an individual interlocutor, but joint actions of both dialogue partners. A contribution consists of two phases: a 'presentation phase' and an 'acceptance phase' (ibid., p. 265–266). In the presentation phase, one of the interlocutors (*S*) makes an utterance in which she makes an attempt of presenting the content that she intends to contribute. In the acceptance phase, the dialogue partner (*L*) is then supposed to provide evidence of understanding of the presentation. When making her presentation, *S* does not know whether *L* is able to understand (or not) what she means and therefore needs to wait for *L* to provide evidence of understanding. Similarly, *L* knows that he needs to provide this evidence of his understanding in order for *S* to belief that he understands what she meant.

Strictly speaking, the act of accepting a presentation by providing evidence of understanding is a presentation phase itself, the content of which is the evidence that is provided. This presentation then needs to be accepted by the interlocutor, which, again, is a presentation phase, and so on. The question thus is how strong the evidence of understanding needs to be such that a contribution can be considered completed.

Clark and Schaefer (ibid., p. 262) propose that this is the case when the 'grounding criterion' is reached, which is specified such that it dynamically adjusts to the current situation and needs of the interlocutors. A contribution can be considered grounded when both *S* and *L* share the belief that *L* has 'has understood what [*S*] meant to a criterion *sufficient for current purposes*' (emphasis added). The strength of this definition of the grounding criterion is that it is sensitive to different aspects that play a role in a dialogue (Clark and Brennan 1991). When it is crucial that the content of the presentation phase is understood — e.g., because the information is very important — the grounding criterion could be set high. When the content is likely easy to understand and somewhat expected, the grounding criterion can be set to a rather low level. When the situation in which the dialogue takes place is difficult — e.g., when the environment is noisy — the grounding criterion should take this into account as well.

Importantly, this definition of the grounding criterion explains why not every acceptance phase needs to be accepted itself. The grounding criterion for presentations of evidence of understanding (i.e., of acceptance) is lower than for presentations of actual content. For acceptance phases it is often low enough such that no evidence of understanding is needed at all (the 'strength of evidence principle'; Clark and Schaefer 1989, p. 268). Nevertheless, in contrast to other theories of discourse, in which a lack of evidence of non-understanding is sufficient for updating common ground, the theory of grounding put forward by Clark and his colleagues generally requires evidence of understanding (see Clark 1996, p. 228).

Another question is what can be considered 'evidence of understanding' and how it varies in strength. According to Clark and Schaefer (1989, p. 267) there are five ways of giving evidence of understanding. A listener can (i) continue to attend, (ii) make a relevant next contribution, (iii) acknowledge the presentation (by providing communicative listener 'feedback'; see chapter 3), (iv) demonstrate that the presentation has been understood, or (v) display (verbatim) that the presentation has been understood. These are roughly ordered from weakest to strongest and also from least costly to most costly. Depending on the presentation and the grounding criterion, different ways of providing evidence of understanding may be appropriate. Sometimes it is enough if a listener simply continues to attend, sometimes a listener needs to actually demonstrate his understanding. It is generally the case that strategies that are weak in strength are often the preferred way to provide evidence as they are less costly to produce (see Clark and Brennan [1991, pp. 230–231] for an overview of costs that play a role contributing to discourse). The strength of evidence of understanding that is sought by speakers and the way this evidence is provided by listeners is determined collaboratively guided by a principle according to which they try to reduce the collaborative effort spend on contributing and grounding (Clark and Wilkes-Gibbs 1986; Clark and Brennan 1991). Often it is sufficient when listeners provide communicative feedback as evidence of understanding (Clark and Brennan 1991, p. 224), which seems to be a trade off between strength of evidence and production costs and thus satisfies speakers and listeners alike.

2.2.2 COMPUTATIONAL THEORIES OF GROUNDING

The relevance of Clark and colleagues' grounding theory for modelling artificial conversational agents was recognised early on, with a first computational model of grounding for task-oriented dialogue developed by Traum (1994). In the following, we present a number of computational approaches to grounding.

The central idea in Traum's model is that grounding adheres to a 'protocol' and that this protocol can be modelled as a finite-state automaton. Drawing upon Clark and Schaefer's (1989) contribution model, Traum notes that it is impossible for an agent to decide when its interaction partner has ended its presentation phase — it might continue after a pause, be interwoven with practical actions, etc. — and thus difficult to recognise the current state of the contribution and which action to take next. As an alternative to contributions, Traum therefore proposes 'discourse units' as the central building-blocks of dialogue. These are similar to contributions in that they comprise presentation of new information as well as grounding of this information. In contrast to contributions, however, they do not consist of a nested sequence of unspecific presentation–acceptance phases that are alternately produced by the interacting agents, but of a sequence of specific 'grounding acts.'

Traum distinguishes seven different types of these grounding acts. Each discourse unit begins with an act of the type initiate. New information is added with continue acts. Acknowledgement (ack) acts signal understanding of what was said before, reqAck acts request an acknowledgement from an interaction partner. Problems in understanding are communicated in reqRepair acts and are addressed in repair acts. Finally, a cancel act may abandon the current discourse unit without grounding what was communicated.

Each utterance in a discourse unit can be classified in terms of grounding acts. This information and the sequence of grounding acts that occurred up to this point of the discourse unit can then be used to determine the grounding status of the information presented in the discourse unit. That is, whether a discourse unit can be regarded as being grounded or not depends on the sequence of grounding acts performed so far.

Traum proposes — based on insights from a corpus of task-oriented dialogues — that a finite state automaton, with states representing the status of a discourse unit, is an adequate model to describe 'valid' sequences of grounding acts.[15] In the model, the

[15] Knowing that repairs may form their own 'sub'-discourse units, which may be in need of repair themselves, Traum initially proposes to use a more expressive formalism, pushdown automata, for the model. As it is unclear whether recursion beyond a certain depth occurs in dialogue and, in particular, whether humans keep track of it and properly 'unwind' it, he opts for the simpler and more efficient finite state automata (Traum 1994, pp. 36–37) — even though these cannot explain, for example, occurrence of multiple acknowledgements in a discourse unit (ibid., pp. 36–37).

Table 2.1: Discourse unit transition table of the finite-state automaton central to Traum's (1994) computational theory of grounding.

	In state						
Next act	S	1	2	3	4	F	D
initiate^I	1						
continue^I		1			4		
continue^R			2	3			
repair^I		1	1	1	4	1	
repair^R		3	2	3	3	3	
reqRepair^I			4	4	4	4	
reqRepair^R		2	2	2	2	2	
ack^I				F	1	F	
ack^R		F	F			F	
reqAck^I		1				1	
reqAck^R				3		3	
cancel^I		D	D	D	D	D	
cancel^R			1	1		D	

Note: S is the initial, F the final, and D a 'dead' state. Acts marked with an ^{I} are performed by the initiator of a discourse unit, those marked with an ^{R} by the responder. *Source:* Reprinted with minor modification from Traum (1994, p. 41, tbl. 3.1).

seven types of grounding acts (additionally individuated by participant) are used as the automaton's input alphabet. If a discourse unit's sequence of grounding acts up to a point is accepted by the automaton — i.e., if starting from the initial state S the input leads to the final state F — the discourse unit is assumed to be grounded. If the input sequence ends in one of the intermediate states 1, 2, 3, or 4 the discourse unit is *not yet* grounded and additional action is required by the participants. If the sequence leads into the failure state D, the discourse unit is abandoned and will remain ungrounded. See Table 2.1 for the automaton's state transition table (reprinted from Traum 1994, p. 41, tbl. 3.1).

Several properties of Traum's grounding model are worth highlighting in the context of this thesis. (1) The model primarily embodies a subjective theory of grounding[16] that an individual conversational agent may possess. Estimation of groundedness

[16] Theory of grounding in the sense of theory of mind.

is based on introspection of the agent's own behaviour and based on observed surface behaviour of the interlocutor. This makes the model cognitively plausible as no omniscient perspective is taken. (2) The model is role-independent. It is valid and useful for the initiator of a discourse unit as well as for the responder. (3) The model processes the conversational actions of dialogue participants incrementally (a property shared with the bigger theoretical framework in which it operates; cf. Poesio and Traum 1997). While unfolding over time, utterances are segmented into units of the size of intonation phrases ('utterance units'; ibid., p. 317). An identified utterance unit is immediately classified into one of the seven types of grounding acts and then given as input to the finite state automaton. Information on the state (grounded or not grounded) of a developing discourse unit is thus constantly being updated and available at all times. (4) Discourse units are either grounded or not grounded. Finer grades of common ground — as suggested by degrees of 'strength of evidence' (Clark and Schaefer 1989) or the findings of Brown-Schmidt (2012) — can neither be computed nor represented. Traum (1994) briefly mentions the possibility of 'degrees of groundedness' and 'confidence' of common ground in the context of the discussion of the problem of multiple acknowledgements (ibid., p. 49; see fn. 15), but does not develop the idea further at that point.

The binary nature of common ground status in Traum's model is an obvious simplification of reality. The distinction between knowledge and belief alone suggest at least a quaternary nature of common ground (believed to be in the common ground or not, known to be in the common ground or not). Similarly, Clark and Marshall (1981, p. 58, emphasis added), concerning the concept of mutual knowledge, note

> [w]hich propositional attitude is appropriate — *knowledge, belief, assumption, supposition,* or even some *other term* — depends on the evidence [an interlocutor] possesses and other factors

thereby already suggesting a number of different degrees of common ground. In an annotated collection of his papers, Clark (1992, p. 6, emphasis added) emphasises this point again

> [...] people hold mutual beliefs with *greater or lesser conviction.* How *strongly* they hold a mutual belief depends on the evidence and assumptions it is based on

this time, however, without explicitly linking strength of common ground to categorical propositional attitudes. And even later, he (Clark 1996, p. 98; emphasis added) writes that

> [p]eople tacitly evaluate shared bases for quality, recognizing that pieces of common ground range in *likelihood from 0 to nearly 1*,

implying a continuous nature of common ground that may even be represented probabilistically.

Roque and Traum (2008) take up the idea of 'degrees of groundedness' mentioned in Traum (1994) and develop them into a computational model of grounding. Similar to the approach in Traum's model, utterance units are mapped onto grounding acts (that — following Clark and Schaefer [1989] — Roque and Traum [2008] call 'evidence of understanding'). Also similar to Traum's (1994) approach, while a discourse unit unfolds utterance unit by utterance unit, the status of the discourse unit is updated, this time however not just from not-grounded to grounded, but with interpretable intermediate steps.

The model defines nine degrees of groundedness. A discourse unit is yet unknown as long as its first utterance unit has not been produced. After that the status of the discourse unit is misunderstood (if a repair is requested), accessible, or, when there is a lack of response by the interlocutor, unacknowledged. If acknowledgement happens, the discourse unit is either agreed-signal, or, when additional evidence is present, agreed-signal+. If concrete content level evidence of understanding has been provided, the degree of groundedness is agreed-content or agreed-content+. If common ground can be assumed for other reasons, the discourse unit has the status assumed. The nine degrees form an ordinal scale with unknown being the least grounded and assumed the most grounded.

When developing the model, Roque and Traum were especially concerned with deciding on a set of degrees of groundedness 'worth modelling' (Roque and Traum 2008, p. 58). They are aware that they left out a great number of potential degrees such as the one that models double acknowledgement in contrast to single acknowledgement (see fn. 15). Their specific choice of degrees is empirically motivated, and reflects their particular dialogue domain (radio-communication), which is highly structured, uses a specific vocabulary, and is limited in the channel. Hence, the model in its concrete form is most probably not directly applicable to more natural forms of conversation such as face-to-face dialogue.

Roque and Traum also see the various degrees of groundedness they define as an operationalisation of Clark and Schaefer's (1989) grounding criterion. Which degree is 'sufficient for current purposes' (ibid., p. 291) may vary depending on the utterance and its context. In some situations, it may be sufficient if agreed-signal is reached, other situations may require a higher degree such as agreed-content.

DeVault and Stone (2006) propose a model of common ground that allows for uncertainty on the side of dialogue participants but evades the necessity to define graded shared belief in terms of probabilities. According to DeVault (2008, pp. 30–33), a probabilistic representation of common ground is methodologically difficult, because (i) it requires an adequate data base from which the necessary probabilities can be learned (which does not exist), (ii) it is difficult to estimate when something between two interlocutors is common ground, and (iii) a sound definition of probabilistic shared belief is difficult.

DeVault and Stone (2006) develop the view that common ground is 'objective' and 'normative'. Although interlocutors in dialogue might still try to achieve shared belief in their actions, it is not necessary that they actually represent shared belief with its many problems. Instead — this is the normative aspect — interlocutors try to represent the common ground as it is objectively the case. The implication of this view on common ground is that each agent maintains its own subjective ('private' in DeVault and Stone's terms) grounding status, together with an estimation of the probability of this being the objective context.

Agents in dialogue therefore may have periods of 'transient uncertainty' (DeVault 2008, p. 33) about what is meant by the interlocutor. If this uncertainty is high and in need of clarification, the agent may take action, otherwise it might track several hypotheses and drop those that become unlikely over the course of the next utterances. Many coordination problems, for example concerning targets of ambiguous referring expressions, will solve themselves sooner or later either because alternative hypotheses do not make sense as soon as more information is produced or because the agent's own actions communicate their uncertainty and the interlocutor notices that understanding is not (yet) present.

Visser et al. (2014) present an incremental model of grounding. While an utterance unfolds, partials of this utterance, as well as overlapping verbal or non-verbal acts of an interlocutor, are classified in terms of the grounding acts defined in Traum's (1994) model. In this model, however, common ground units do not need to span complete utterances. One part of an utterance can be seen as already grounded, while the following part is still in the in the process of being grounded (state 1–4). In general, the model is able to account for interactive utterance production and is capable of dealing with utterances that contain multiple repairs as well as parts that are considered ungrounded and cancelled. The model is used for incrementally generating overlapping grounding behaviour of a conversational agent (that reflect its state of natural language processing) in response to an ongoing utterance of the interlocutor.

Li et al. (2006) present a grounding model that uses an augmented push-down automaton (instead of Traum's [1994] finite state automaton) in order to account for nested discourse structures. In this model, grounding acts do not need to relate to the immediate preceding act of the interlocutor. Ungrounded contributions are pushed onto the stack and taken from the stack once they are considered grounded.

Paek and Horvitz (2000b) present a decision theoretic model of Clark and Schaefer's (1989) grounding criterion. Based on the expected utility of grounding-related actions of dialogue participants, the model formalises situation specific thresholds p_L^* for sufficiency of understanding (from a listener's perspective) and p_S^* for a sufficiently high degree of belief in a listener's understanding (from a speaker's perspective). If p_L^* is exceeded, speakers should move on, if p_L^* is not reached, listeners should initiate repair (and vice versa). Thus a proposition is considered to be a shared belief (i.e., in the common ground) if the evidence of understanding an utterance, or increment, that communicates this proposition, exceeds both thresholds. This model is used in the 'Quartet' architecture for spoken dialogue systems (Paek and Horvitz 2000a) which models inference and decision making in dialogue as reasoning under uncertainty and pursues the objective to reduce uncertainty on multiple levels of processing. Another system, by Skantze (2007), uses the model in order to decided when to produce grounding actions, but instead of using a handcrafted parameters, learns the costs of actions from a corpus of actual interactions.

We argued that a shared basis as well as common ground is a prerequisite for conversation and we saw how — while the conversation unfolds — dialogue participant ground their utterances and the content that is communicated by seeking and showing evidence of understanding from and to their interlocutors. In the following section, we describe how dialogue participants, when producing utterances, make use of the information in the common ground. We discuss this in the broader context of adaptation to and coordination of dialogue partners.

2.3 ALIGNMENT, ADAPTATION, AND COORDINATION

It is accepted by researchers that adaptation between interlocutors in dialogue takes place. Still, discussions about adaptation in dialogue revolve around underlying mechanisms and their demand of cognitive resources. Perspectives differ on whether adaptation leads to an 'alignment' of underlying representations, and on how much effort interlocutors can actually spend on adaptations that happen concurrently to speech and behaviour production.

Over the course of an interaction, participants 'adapt' their behaviour to each other, that is, their behaviour patterns become similar and synchronous (Burgoon et al. 1995, p. 4). In dialogue or conversational interaction, their speech features ('amplitude, pitch, rate of articulation, pause structure, phonological features, and response latency before initiating a conversational turn' [Oviatt et al. 2004, p. 303]) become more similar, the levels of the rhythmic prosodic hierarchy become synchronised (Wagner et al. 2013), they use the same lexical items and referring expressions (e.g., Garrod and Anderson 1987; Brennan and Clark 1996), they tend to adopt each other's syntactic structures (e.g., Branigan et al. 2000; Reitter and Moore 2014 — but see Healey et al. 2014), and they use similar speech accompanying gestures (e.g., Bergmann and Kopp 2012; Mol et al. 2012) as well as other non-verbal behaviours (e.g., facial expressions, posture; see Chartrand and Bargh 1999). If such adaptations go beyond behaviour, that is, if the underlying mental representations of the interaction partners become similar, the effect of such an adaptation is called 'alignment' (Pickering and Garrod 2004, p. 172).

In addition to behaviour becoming more similar, adaptation in dialogue can also be seen from the perspective of 'recipient design' or 'audience design' (Clark and Carlson 1982). According to this view speakers adapt their speech in order to make themselves better understood to the interaction partner, e.g., by taking the common ground into account, but also by making choices on the level of discourse that take the interaction partners' needs into account.

These perspectives on adaptation are not incompatible, quite the contrary, they interact. Automatic adaptation on the behavioural level may also results in audience design, and audience design often results in behaviour becoming similar to the interaction partners'. This is captured in the concept of coordination on different levels of processing and awareness (Kopp 2010, see section 1.3). In the following, we present different theories of adaptive speech production, but start with a description of the interactive nature of adaptation.

2.3.1 INTERACTIVE ADAPTATION

Adaptations in language production can take place at virtually any point of time, even mid-utterance. The cognitive processes of speech production operate 'incrementally' on all levels of processing: articulation, formulation, and conceptualisation (Kempen and Hoenkamp 1987; Levelt 1989), see chapter 6 for further discussion.

Already at the lowest level of speech production — articulation — speakers adapt their speech to the situation and environment. In noisy environments, for example, the acoustic parameters of speech are adapted in order to increase speech intelligibility. This is achieved by increasing intensity, raising the fundamental frequency, shifting the spectrum, enhancing voiced sounds, and re-allocating energy in the sound spectrum (Cooke et al. 2014, § 4.1). This results in so-called 'Lombard speech'. Similar changes are also made in order to accommodate to interaction partners with known or assumed limitations in perception, e.g., hearing impaired persons, infants, or artificial conversational agents (Cooke et al. 2014, § 3.1).

On the level of utterance production, Clark and Krych (2004) showed that speakers actively monitor their addressees (especially for behaviours that pertain to groundedness) and incrementally take these insights into account. They observed, in a task-oriented setting, that speakers' and addressees' behaviour is highly coordinated and that speakers rapidly and interactively change the course of their actions depending on the addressees' verbal and nonverbal behaviours. The results of this interactive adaptation is that utterances that speakers produce are co-constructed by their addressees. Contrasting dialogues in which such interactive adaptation is possible with dialogues in which speakers and addressees could only coordinate to a lesser degree (due to limited visibility) — or not at all — Clark and Krych (ibid., pp. 67–69) showed that tasks were solved faster, more efficiently (using fewer words), and that fewer errors were made. They (ibid., pp. 76–78) conclude that (i) common ground is updated continuously and not just after turns are finished, (ii) increments of speech production are constructed jointly by speakers and addressees, (iii) addressees provide multimodal evidence of their understanding as soon as possible, (iv) speakers and addressees interact all the time, and (v) speakers plan their utterances opportunistically, i.e., they assume that repair will be necessary (see section 2.1.4).

Interactive adaptation to the interaction partner does not only happen on the level of utterance construction but is shaping the overall conversation. This might be regarded as trivial given that each interaction partner can, in general, propose topic shifts, but it also happens when one interaction partner is mostly listening and may only provide feedback. Bavelas et al. (2000) showed that speakers have difficulties telling stories (stories were told less well) when they face a lower number of a certain type of listener feedback responses by their interlocutors — which was the result of them not being fully attentive due to an experimental manipulation. The degradation in evidence of understanding left speakers uncertain about common ground and their feeling for the comprehensibility of their utterances. From this Bavelas et al. (ibid.) concluded that even when just listening and providing feedback, listeners 'co-narrate' the stories that speakers tell.

In the following we present models of adaptation in speech production that are generally compatible with the concept of interactive adaptation. They are continuously active and can immediately shape speech production when new information (e.g., evidence of understanding or changes in the environment) becomes manifest.

2.3.2 INTERACTIVE ALIGNMENT

Based on the findings about adaptation, on an assumption of parity between production and comprehension, and on an assumption that both of these processes rely on the same underlying linguistic representations, Pickering and Garrod (2004) present an 'interactive alignment' theory of dialogue, in which adaptation phenomena are explained by low-level 'priming' of the underlying linguistic representations. Perceiving a referring expression from an interlocutor, for example, primes and activates its lexical representations in a listener, which, in turn, makes it more readily accessible in speech production and increases the likelihood of using the same expression. Moreover, the interactive alignment model claims that activation of representations on one level activates representations on other levels, percolating through phonetic, phonological, lexical, syntactic, and semantic representations to the level of 'situation models' (Zwaan and Radvansky 1998) — an alignment of which leads to dialogue success.[17]

Instead of explicitly representing shared beliefs or common ground, the interactive alignment model posits that information that is present in the aligned representations of interlocutors serves as an 'implicit common ground' (Pickering and Garrod 2004, p. 178). Reaching understanding among interlocutors and adapting to each other is thus an automatic process which does not — in general — involve reasoning about common ground or planning adjustments to the interlocutor's needs. According to Pickering and Garrod it would be too computationally expensive if such coordination processes were continuously active. Explicit common ground only needs to be considered, and adaptations designed for the interlocutor only need to be planned when repair is needed because a certain amount of mis-alignment has been detected (ibid., p. 179).

Pickering and Garrod's main motivation for the interactive alignment account is that speech and language processing without a high degree of automaticity would be far to effortful. Although researchers mostly agree that priming plays a role in adaptation phenomena, the interactive alignment model's strong claims about dialogue processing being mechanistic, as well as its aspiration to be a general theory of communication, have raised criticism (see, e.g., the peer commentary in Pickering and Garrod [ibid., pp. 190–211] and the contributions in the collection edited by Wachsmuth et al. [2013]).

[17] Situation models are cognitive representations of speakers and listeners which reconstruct the content of an utterance or a discourse (Zwaan and Radvansky 1998). According to the interactive alignment model (Pickering and Garrod 2004, pp. 172–173) interlocutors in dialogue understand each other when their situation models are aligned, that is, when they have constructed situation models that are sufficiently similar. Reitter and Moore (2014) found evidence for this in a corpus of task-oriented dialogue.

2.3.3 FULL COMMON GROUND

The proclaimed effortlessness of the interactive alignment account of adaptation in dialogue is in strong opposition to models of dialogue in which meaning making and understanding assume a collaborative effort of the interlocutors (e.g., Clark 1996). According to this perspective, utterances are 'designed' for the recipient (or an audience; Clark and Carlson 1982, pp. 10–11), take common ground into account (Clark and Marshall 1981; Clark and Brennan 1991), and are additionally shaped by interlocutor feedback that is received while they are produced.

A metaphor for dialogue processing that Brennan (1990, pp. 30–33) uses is one of 'hypothesis testing'. The assumption is that in order to reach its intended effect, an utterance needs to be tailored to its addressee — for example by taking into account what is in the interlocutor's common ground. As common ground is inherently uncertain (see page 25 above), however, it is not necessarily foreseeable whether an utterance will be understood. Utterances can thus be regarded as hypotheses about the interlocutor's common ground that a speaker tests by producing them. The interlocutor's responses to such utterances then serve as evidence for (or against) the hypotheses. If an interlocutor's response suggests that a hypothesis embodied in an utterance is false, the speaker needs to revise her hypothesis and test anew. This can either be done by producing a new utterance that embodies the revised hypothesis, or — when the interlocutor responds while the utterance is being produced — through marked 'self repair' (Schegloff et al. 1977), or changes to the plan and course of the continuing utterance (which will go unnoticed because what has already been articulated is not affected by the change).

When processing utterances, addressees are engaged in hypothesis testing as well. While an utterance is unfolding they hypothesise what the speaker could mean. If sufficiently certain of their hypothesis, they provide evidence of understanding or even act early, thus making the interaction more efficient. If uncertain, they have to revise their hypothesis or, if they cannot come up with a good hypothesis, can provide evidence of non-understanding or request clarification.

This makes dialogue a process of collaborative hypothesis testing in which interlocutors 'continually seek and provide evidence of mutual understanding, and [...] collaborate in testing and revising their complementary hypotheses until the difference between them is too small to matter' (Brennan 1990, p. 33).

2.3.4 MONITORING AND ADJUSTMENT

The interactive alignment model and the full common ground model of language use and partner-specific adaptation in dialogue lie at opposite ends of the scale of processing effort during language use. In addition to concerns about the efficiency of the use of full common ground in online language processing, concerns about its parsimony were raised and it was proposed that experiments need to take the possibility into account that a model subsumed by the full common ground model would explain the observed adaptation phenomena equally well, or even better (Keysar 1997).

The 'monitoring and adjustment' model for speech production (Horton and Keysar 1996) and the 'perspective adjustment' model for comprehension (Keysar et al. 1998) are more parsimonious and lie in between the two extremes. Both claim that language processing in dialogue is, initially, egocentric and does not take common ground into account. This is unproblematic when the interlocutors' perspectives are already aligned. If not, however, it may lead to production and comprehension errors, which are addressed in a second process. Here common ground becomes relevant and adjustments to the utterance or interpretation are made.

It was shown that under certain conditions — for example time pressure[18] (Horton and Keysar 1996) — speakers do not take common ground into account, whereas they do when these conditions do not hold. The adjustment process does not necessarily need feedback from listeners to take place. Through monitoring common ground violations in their initial utterance plan, speakers can become aware of (potential) problems of not taking common ground into account themselves and adjust before articulation even starts. In certain conditions there may, however, not be enough resources for such adjustments to take place before articulation. When and if this is the case is not just a result of certain conditions but also of the speaker's familiarity with certain types of interactions (Horton and Gerrig 2002).

On the comprehension side listeners initially act egocentrically as well. They process definite references according to their own perspective and only adjust their perspective if a later monitoring-stage detects common ground violations in their interpretation, in which case referent resolution is delayed (Keysar et al. 1998).

[18] Time pressure of varying degree is a natural factor in speech production. When articulation finishes producing currently planned units, speech planning (conceptualisation and formulation) need to have subsequent units ready for articulation. In dialogue settings, when intending to make relevant and coherent contributions to an ongoing discourse, speakers need to be ready to start speaking instantaneously when the preceding turn ends. If not, other participants may take the turn. Even though there are other mechanisms to mitigate time pressure in these cases (e.g., filled pauses such as *uh* and *uhm*, Clark 2002; Clark and Fox Tree 2002) these are only tolerated by interlocutors to some degree.

The difference of these adjustment models to the full common ground model is that common ground reasoning in language processing is not needed by default. Common ground assumptions only come into play if subsequent monitoring detects their violation. As many utterances need not be tailored for addressees, e.g., because they can be fully understood by relying on conventions and shared context, monitoring and adjustment is more efficient than partner specific production and comprehension that fully rely on common ground. The model can also explain grounding violations under conditions such as time pressure. The parsimony of the model can however be disputed as it proposes two-stage processes in production and comprehension (Brennan and Hanna 2009, pp. 284–286).

The adjustment models also differ from the interactive alignment model (which is sometimes called a two-stage model as well, ibid.) in that they acknowledge explicit common ground as a factor in speech production and comprehension even without listener-initiated repair or clarification requests.

2.3.5 MINIMAL PARTNER MODELS

Brennan and Hanna (ibid., pp. 284–286) reject the proposal of two-stage adjustment models because the timings for initial and later stages processing that were measured in various studies differ widely. They also refute the claim that these models are more parsimonious than integrated models. Brennan and colleagues (Brennan and Hanna 2009; Galati and Brennan 2010), however, generally acknowledge the concerns about processing effort that would be needed when taking full common ground into account.

At the same time, Galati and Brennan (2010) find evidence for partner specific adaptation in articulation, which, among many other parts of an utterance, is attenuated in repeated interactions with the same interlocutor, but not in interaction with new interlocutors. This indicates that speakers do not simply attenuate information for their own benefit, but for the addressees'. Their explanation of why this is possible even in highly automatised low-level processes such as articulation is twofold: the addressee's needs are (i) clearly indicated through available 'cues' and 'salient knowledge', and (ii) can be represented in a simple way (ibid., p. 47). Galati and Brennan propose that adaptation of utterances to addressees is based on a simple — almost stereotypical — model of the addressee. Each dimension of such a 'most minimal model' consists of only 'one-bit' of information (ibid., p. 47), such as, for example, the binary knowledge of whether interlocutors talk about something for the first time, or not.

To illustrate their proposal, Brennan et al. (2010, p. 324) list several dimensions of most minimal partner models identified in experimental studies (references to which are omitted here):

> *my partner can see what I'm doing,* or not [...]; *my partner can reach the object she's talking about,* or not [...]; *my partner has a picture of what we're discussing,* or not [...]; *my partner and I have spoken about this before,* or not (Galati and Brennan 2010; [...]); *my partner is currently gazing at this object,* or not [...]; *my partner needs to distinguish this referent from a competitor,* or not [...]; *my partner is a young child,* as opposed to older [...]; *or my partner is a native speaker of English,* or not [...].

It should be noted that this model can be informed by examining the context in which an interaction takes place and/or by making (mostly static) inferences about the addressee. Brennan et al. (2010) suggest that 'a "partner model" need not entail a detailed record of all of the knowledge one partner has about what the other is likely to know [...] as well as what the other does not know' (ibid., p. 324). Instead, speakers using such a minimal model 'represent relevant aspects of common ground in a simple, clear way' (Galati and Brennan 2010, p. 47). Nevertheless, even such minimal models potentially encompass many dimensions[19] and are likely compiled and updated dynamically, depending, *inter alia,* on the dialogue situation, the visual context, the interlocutor, and the interlocutor's feedback. Thus even such most minimal models can be complex and powerful.

The question whether binary representations are sufficient to account for real world adaptation phenomena is left open (ibid., p. 47). It is certainly plausible to assume that some of the dimensions have richer, gradient representations. Brown-Schmidt (2012), for example, finds evidence that speech production is influenced by the strength of groundedness of information (as already suggested by Clark and his colleagues, see section 2.2.1). Similarly, dimensions that are informed from multiple cues and/or sources are likely to contain uncertainty. In both cases the dimensions in the most minimal model could be probabilistic, i.e., represented as values between 0 (impossible) and 1 (certain). A speaker could for example be fairly certain (0.83) that her 'partner is currently gazing at this object' (see above). Even such richer models retain the advantage of being relatively simple in their representation. Each dimensions coarsely models a certain aspect of the interlocutor which is relevant for adaptation — and is derived from various cues and knowledge.

[19] Brennan et al.'s (2010) proposal is underspecified in that they do not provide criteria for when a variable should be part of their model.

2.3.6 INTERMEDIATE SUMMARY

How and when adaptation in language processing takes place is not only relevant for psycholinguistic modelling, but for the design and implementation of artificial conversational agents, too. We think that each of the models discussed in the previous sections — despite being antagonistic from a theoretical point of view — contains ideas that are useful for modelling the processes of interactive and adaptive natural language generation for conversational agents.[20]

The interactive alignment model suggests itself as a lightweight (i.e., priming-based) mechanism for automatic adaptation of an agents utterances' surface form to the human interlocutor (e.g., Buschmeier et al. 2010; Isard et al. 2006; de Jong et al. 2008).

The full common ground model can be used as a principal framework for making decisions in generation that take into account the objects that are assumed to be part of the common ground (e.g., Stone et al. 2003; DeVault 2008) — which are typically rather few in dialogues between artificial conversational agents and human interlocutors.

The monitoring and adjustment model on the other hand suggests a strategy for situations where an agent's resources are limited (e.g., because the system needs to start speaking at a certain point in time in order to be able to take the turn) or when there is too much uncertainty in the representation of common ground (e.g., DeVault 2008).

Finally, minimal models of the partner may be useful when the dimensions (i.e., the bits) that they represent directly map to the adaptation capabilities that a natural language generation component possesses (e.g., Walker et al. 2007; Mairesse and Walker 2010). As will be shown later on (section 8.3.4), they may also be useful for making decisions that shape the course of the interaction — on the level of dialogue management.

Ideas from the alignment account can be seen as a way of coordinating on the level of behaviour (Kopp 2010). In contrast to this, ideas taken from the full common ground and monitoring and adjustment models, seem to address coordination on the levels of belief (and attitude). The ideas taken from the minimal partner model approach seem to be useful for all levels of coordination as well as for shaping the interaction.

[20] As can be seen, some of the ideas are already part of current natural language generation systems.

2.4 REACHING UNDERSTANDING WITH ARTIFICIAL CONVERSATIONAL AGENTS

In the following we will review in how far current approaches to artificial conversational agents take the properties of dialogue that are described in this chapter into account.

Artificial conversational agents are computational artifacts that use natural language processing and artificial intelligence in order to be able to interact with human interlocutors using natural language. Examples for such agents are purely speech-based interfaces such as spoken dialogue systems (McTear 2002), embodied conversational agents (Cassell et al. 2000), which are virtually embodied and can thus produce verbal and non-verbal acts (e.g., gestures, facial displays, gaze), or sociable robots (Fong et al. 2003), which have a physical presence and may even be able to manipulate the physical environment they share with their interlocutors.

Early on, natural language-based communication with artificial conversational agents has been seen as an important problem in artificial intelligence. In proposing that a computer program should be regarded as 'intelligent' if it can convince human interlocutors that they talk to a human and not to an artificial conversational agent,[21] Turing (1950) suggested that conversational interaction is a task that needs human-like abilities to be solved.

Early progress on artificial conversational agents, beginning with the program ELIZA (Weizenbaum 1966), however, has shown that intelligence is not necessary for a computer program to be an engaging and somewhat believable dialogue partner. ELIZA certainly has no chance passing a Turing test, since it simply consists of a set of rules that define its behaviour in reaction to its interlocutors (it can neither understand their communicative acts nor its own ones). Nevertheless, Weizenbaum demonstrated that conversational agents can get a long way without pragmatic inference and interactional intelligence. Especially application-oriented artificial conversational agents (e.g., telephone-based dialogue systems) — but also research-oriented systems in which the research focus does not lie on dialogue (e.g., sociable robots) — benefitted from this.

Yet, some research on conversational agents takes a more interaction-oriented approach to dialogue, resulting in conversations that show some of the characteristics of interactional intelligence. Although even the most basic dialogue management approaches, e.g., traversing a finite state automaton (McTear 2002, § 5.1) or filling a frame (ibid., pp. 5.2) can be used to model simple forms of interactive processes

[21] The original formulation of the task in this, so called, 'Turing test' is somewhat more intricate in that participants do not even suspect to be talking to a computer (Turing 1950, pp. 433–434).

in dialogue, interaction is neither a central nor a general principle for them. In the following we describe approaches in which interaction is central.

Heeman and Hirst (1995) present a plan-based computational model for interactively solving a single, but central, problem in dialogue: reference. Their model of collaborative reference (Clark and Wilkes-Gibbs 1986) is able to account for the generation and understanding of reference and involves proposing an expression, judging and potentially clarifying it, rephrasing it, and, eventually, accepting and adopting it.

Poesio and Traum (1997) present a formal discourse theory (PTT) that uses a unifying representation of context to be able to account for various discourse phenomena (here interactively generating a referring expression is just a special case). Central to this theory is that dialogues are constructed from 'micro conversational events', which allows the model to capture interactivity in discourse even on the sub-utterance level, which is important to model grounding and other phenomena. Simplified versions of PTT have been applied in a number of dialogue systems that use the 'information-state update' model for dialogue management (Larsson and Traum 2000, § 4), as well as for in-depth analysis of real dialogues (Poesio and Rieser 2010), where it is able to account for the coordination phenomena that occur.

Regarding more practical dialogue systems, Skantze (2008), presents a computational discourse model that keeps track of the grounding status of the concepts (instead of utterances) that are present in the discourse, which is then used for error detection and interactive error handling with situation-specific strategies.

Skantze and Schlangen (2009) describe a dialogue system in a micro-domain that can incrementally display and ensure its understanding of telephone numbers — that its users can dictate in a conversational manner (e.g., in prosodically marked 'installments' [Clark 1996, p. 236]) — by producing human-like clarification and grounding acts.

Hough and Schlangen (2016) present an incremental dialogue system that models grounding in a task-oriented human–robot dialogue (where the human instructs the robot) by tracking the dialogue state in two parallel, but interacting, state-machines (based on statecharts [Harel 1987]), one for its own state and one for the estimated state of its human interlocutor. These state machines model an interactive repair process. A goal is considered to be grounded ('publicly manifest') between the dialogue partners when the interlocutors commit themselves to the goal, which they only do when the robot displays enough commitment towards this goal (e.g., through actions). The grounding and repair process is guided by globally set thresholds for the strength of evidence for individual goals. Lowering these thresholds allows for a higher level of incrementality in the system.

Yaghoubzadeh, Pitsch et al. (2015) describe a dialogue system for elderly users and users with cognitive impairments. In order to ensure that users and system mutually understand each other correctly, the system's dialogue management approach employs a flexible grounding strategy. Depending on how well a user can likely process multiple pieces of information in one utterance, the system can either explicitly ensure the understanding of every slot of information that the user provided individually or only do so if it suspects that the information for a slot might be wrong (based on the clarity of the results from automatic speech recognition).

All of these systems try to reach understanding by modelling the grounding process. They can detect problems in understanding (either their own or their interlocutors' understanding) and most are able to adapt their communicative actions, e.g., by engaging in clarification and repair or by choosing specific presentation strategies. Some of these system even account for the incremental nature of interactive grounding.

2.5 SUMMARY AND CONCLUSION

In this chapter we elaborated on three aspects of dialogical interaction that were already mentioned in the introduction. We first debated what it means when interlocutors in dialogue understand each other and came to the conclusion that a strong notion of understanding — which causes the paradox discussed in the introduction (Rapaport 2003) — is practically irrelevant. Similar to the solution that Rapaport offers, we concluded that interlocutors in dialogue can attain a weak form of understanding, which can be reached and improved upon interactively, until understanding is sufficient.

Following this, we described the concept of common ground, a resource that is shared between interlocutors and upon which understanding is build. Somewhat paradoxically, the common ground is itself expanded when interlocutors reach understanding about things that were not yet in their common ground and publicly share this achievement with one another — by providing evidence of understanding. We described this grounding process and reviewed computational models of grounding which deal with the problems of recognising whether an artificial conversational agent can regard an utterance to be grounded as well as how and when an agent should provide evidence of its understanding.

We then turned the discussion to another capability that is important for interactively reaching understanding and establishing common ground, namely being able to make oneself understood by taking the common ground and needs of the interlocutors into account during language production. We illustrated the interactive nature of this process and the effects of adaptation on various levels of the speech production process. We reviewed psycholinguistic models of (adaptive) language production — which vary in their requirements of cognitive resources and their approach to adaptation — and suggested how ideas from these models could be combined to make an artificial conversational agent adaptive, and enable it to coordinate with its interlocutor on different levels of processing (Kopp 2010).

In light of the information presented in this chapter we can draw the conclusion that being a successful participant in conversation and dialogue hinges on the ability to interactively make oneself understood, i.e., it requires 'interactional intelligence' (Levinson 1995). The review of dialogue models for artificial conversational agents illustrates that the relevance of grounding and common ground as well as adaptation has been recognised in these research fields early on (and is still an ongoing research topic). The interactional nature of these processes is already embodied in some of these systems, but usually not as the guiding principle in their design.

CHAPTER 3
COMMUNICATIVE FEEDBACK

In this chapter we introduce the dialogue phenomenon 'communicative listener feedback' as a mechanism for belief and attitude coordination in dialogical interaction. We begin by exploring the origins of the concept of feedback in cybernetics and analyse whether the underlying ideas are of relevance to communicative feedback. We then compare this term to the terminology that is in use for the phenomenon. Following this introduction, we describe the phenomenon itself: we look at the form of feedback (both verbal, prosodic and non-verbal) as well as the functions and meaning of feedback signals. These descriptions are based on reviews of the literature as well as on our own research on German listener feedback in the ALICO-corpus. Following this, we review the literature on timing and placement of feedback, and on how feedback phenomena have been modelled in artificial conversational agents.

3.1 ON THE ORIGINS OF THE CONCEPT OF FEEDBACK IN COMMUNICATION

The term 'feedback' has its roots in 18th century engineering. It describes the principle of automatically regulating mechanical apparatuses to adhere to a certain state.[22] Well known examples are 'centrifugal governors' (Maxwell 1867, see fig. 3.1) as used, for example, in Boulton and Watt's 1788 steam engine to maintain a constant speed under

[⋆] The ALICO-corpus, that we use in this chapter to illustrate various aspects of feedback, was collected and analysed in collaboration with Zofia Malisz, Marcin Włodarczak, and Petra Wagner from Bielefeld University's Phonetics and Phonology Group, with contributions from Joanna Skubisz. Analyses of various aspects of feedback in the ALICO-corpus have been published in Buschmeier et al. (2011), Malisz et al. (2012), Włodarczak et al. (2012), Buschmeier et al. (2014) and Malisz et al. (2016).

[22] Self-regulating mechanical constructions were invented much earlier. In a study of constructions and machines from antiquity to modernity, Mayr (1970, pp. 11–16) identifies a water clock from the third century BCE — constructed by the Hellenistic mechanician Ktesibios from Alexandria — as the first (known) device using a feedback mechanism.

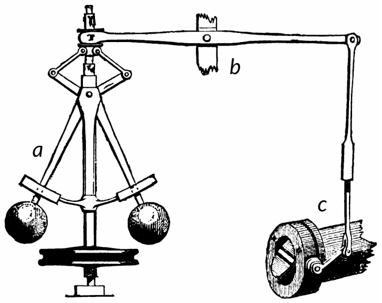


Figure 3.1: Schematic drawing of a centrifugal governor that keeps the rotational speed of a steam engine constant under varying load conditions based on feedback principles. Changes in engine speed cause the centrifugal pendulum (a) to swing in- or outward mechanically moving a lever (b) which opens or closes the inlet valve for steam (c) thus regulating the amount of steam getting into the engine (Mayr 1970, pp. 2–3, 109–113).

varying load conditions. This is achieved by 'feeding back' a signal of the effect of a machine (e.g., rotational speed) to the controller of its action (e.g., a steam valve). The controller regulates the action, which changes the effect (e.g., a change in the amount of steam changes the rotational speed). Feeding back this information (the feedback) creates a closed signalling-loop between two parts of a machine which then affect each other and results in a 'circularity of actions' — the defining criterion for feedback (Ashby 1956, p. 53).

The study of self-regulating machines was later taken up in the field of cybernetics, which promoted feedback as a first principle for the study of general 'systems', i.e., declaring it to be the prominent mechanism at work in 'the animal and the machine' (Wiener 1948/1961). As systems theory is all-encompassing in its aspiration (everything is a system), it was natural to not only think of the mechanisms at work within a single entity — e.g., a human being — in terms of being feedback controlled, but to apply this thinking to the interaction processes taking place between multiple entities, even whole societies (Ashby 1956, p. 5). From there on, it did not take long until the concept of feedback control was used in theories of human communication.

Using a simple communication experiment (in which speakers described geometric patterns to groups of listeners who had to draw them), Leavitt and Mueller

(1951) compared interactions in which listeners were prohibited from giving feedback with interactions in which listeners could speak freely.[23] It was found that allowing listener feedback resulted in drawings that were more accurate and made speakers and listeners more confident in their own performance.[24]

In similar experiments, Maclay and Newman (1960), Stolz and Tannenbaum (1963), and Krauss and Weinheimer (1966) analysed the specific effects that listener feedback has on speakers' speech production and found that the length of phrases (ibid.), total quantity of speech, word and phrase selection (Maclay and Newman 1960), speech encoding time, utterance duration, and speech rate as well as hesitation phenomena such as quantity of filled pauses, false starts, and repetitions (Stolz and Tannenbaum 1963) are affected. The studies of Maclay and Newman (1960) and Stolz and Tannenbaum (1963) also analysed the effect of different types of feedback (positive and negative) in comparison to a no feedback condition. It was found that negative feedback has an effect on speaker behaviour, whereas a difference between no feedback and positive feedback cannot be shown in their data. Stolz and Tannenbaum (ibid.) further suggest that speakers react to negative feedback by actively trying to alter their speech production, i.e., stopping, replanning, and restarting their utterance.

All four studies explicitly use the term feedback and refer to its concept. In contrast to cybernetics, however, they use it in a rather informal way, which leaves its status unclear. Is it really feedback in the cybernetic sense? Krauss and Weinheimer (1966) think that a feedback-controlled system is, at least, a good analogy for

> the model of a speaker as intent upon effecting some end state in his listener, monitoring the listener's behaviour for indications of change, and adjusting his subsequent output on the basis of this information (ibid., pp. 343–344).

Analysing Leavitt and Mueller's (1951) study for the appropriate use of the term, it can be seen that they regard the dyad of dialogue partners, a speaker and a listener, as their system. These two interlocutors then form the two parts of the closed signalling-loop. The quality of the speaker's description affects the listener's ability to reproduce a geometric pattern. The listener provides information about her ability to do so, which then allows the speaker to change his description, which in turn might change the

[23] It should be noted that Leavitt and Mueller's (1951) concept of feedback is a broad one. Listeners in the feedback condition were allowed to ask questions and to interrupt the speaker. In stark contrast to this, the no-feedback condition even prevented visual contact between listeners and speakers.

[24] A different experiment that additionally included two less disparate conditions — visibility; listeners were allowed to respond with 'yes' or 'no' to speaker questions — supports the finding, but did not yield statistically significant results (ibid., experiment 1).

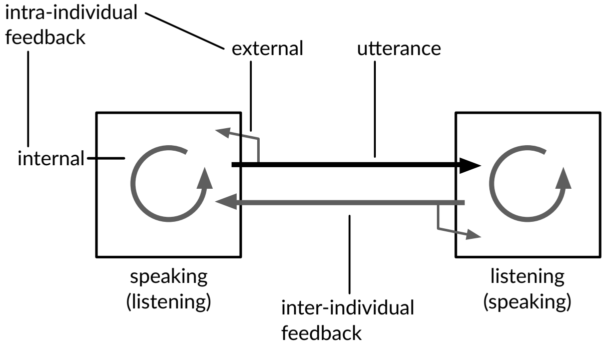


Figure 3.2: Allwood's application of the cybernetic concept of feedback to speech-based communication. Two levels of feedback are suggested. Intra-individual feedback — Levelt's (1989) self-monitoring — comprises an internal (covert) and an external (overt) feedback loop. Inter-individual feedback flows from the listener to the speaker. Redrawn and translated from Allwood (1988, p. 91, fig. 1).

listener's ability to reproduce the pattern, and so on. From this it can be said that a circularity of actions within the system is clearly present so that the use of the term feedback for the information that the listener provides to the speaker is warranted — despite a lack of formal rigour that is present in cybernetics and engineering.

A study that makes an explicit reference to Wiener (1948/1961), even featuring a schema that explicates the 'control flow' in communication in form of a block diagram (see fig. 3.2), is Allwood's (1988) work on the system for Swedish linguistic feedback. Allwood sees multiple feedback loops at work in speaker–listener dyads: intra- and inter-individual feedback. Intra-individual feedback occurs within speakers (and also within listeners) via the two self-monitoring loops (covert and overt) present in the human speech production system (see, e.g., Levelt 1989, fig. 1.1, pp. 13–14). This feedback loop is used to notice and repair deficiencies in planned and not yet articulated (covert), or articulated (overt) speech, respectively. Inter-individual feedback flows from listeners to speakers. This is the concurrent listener feedback that the four studies presented above focus on — although Allwood (1988) limits its scope to small, quickly produced signals (which he calls 'interjections', see below). Speakers use this kind of feedback in order to adapt their speech production process to the listeners' needs.

Sharing Allwood's concept of intra- and inter-individual feedback, Allen and Guy (1974, pp. 25–26), taking a conversation analyst perspective, apply the term feedback more broadly and identified two further feedback loops. Their third loop is feedback that occurs after the turn changes (i.e., speaker and listener switching roles). Turns, in coherent discourse, add to the representation of the interaction that individual participants hold, which is an important basis for subsequent turn formulation. According to Allen and Guy, it is this feedback loop that closes the circle of communication. Their fourth and final loop of feedback comprises the outcomes of the interaction which lead to 'behavioural and conceptual reorientations of [the participants] toward the other' (ibid., p. 26). The four loops run on increasing time scales. The intra-individual loop spans the time of syllables, words up to an utterance, the inter-individual loop the time of one or two utterances, the third loop spans two turns, and the fourth loop extends beyond the interaction.

But there are also critical views on the feedback analogy. Stolz and Tannenbaum (1963), despite using it, believe that

> feedback [in human communication systems] is not a unitary factor with unitary consequences. Its effect on further encoding behaviour is probably a function of several variables — the kind of feedback, its source, its focus, and so forth (ibid., p. 225).

Feedback in human communication is more complex than the feedback in self-regulating apparatuses that are typically studied in engineering. This can be seen as a problem. Ashby (1956) warns that

> the concept of 'feedback', so simple and natural in certain elementary cases, becomes artificial and of little use when the interconnections between the parts become more complex (ibid., p. 54).

This is the case because

> [s]uch complex systems cannot be treated as an interlaced set of more or less independent feedback circuits, but only as a whole (ibid., p. 54),

which makes a formal treatment, in mathematical terms, difficult.

Commenting on the application of the concept of feedback in the social sciences in general, but equally relevant for its applicability to communication as well, Spink and Saracevic (1998) raise the point that the view of feedback as a simple control signal, even though it works well in engineering, is 'devoid of any cognitive and situational references, interpretations, and processes' (ibid., p. 251), which, as we seen have in chapter 4, are central in communication and dialogue. They further note that the feedback analogy puts too much 'emphasis on the loops between intervening variables [and hence] de-emphasizes the study of the variables themselves' (Spink and Saracevic 1998, p. 251), that is, the interacting agents and their cognitive processes.

Both strands of criticism suggest that the cybernetic feedback analogy, though it may certainly be helpful when thinking about the general processes, is much too simplistic for describing or modelling the actual interaction that takes place in human conversation. The feedback analogy may — similar to the view of code models of communication (Shannon 1948; Blackburn 2007) — be helpful in thinking about the phenomena at hand, but falls short of the complex and rich inferential processes actually taking place, or may even be misleading.

3.2 TERMINOLOGY

Despite the above criticism we, nevertheless, think that 'communicative listener feedback' (or simply 'feedback)' is an appropriate general term (and analogy) for thinking about the phenomena related to the inter-individual concurrent information flow from listeners to speakers, their effects on the speakers' speech production and the dynamics and outcomes of interactions that we focus on in this thesis.

The communicative acts that serve as communicative listener feedback in dialogue received a large number of different names in the literature. A handbook chapter on 'listener responses' (Xudong 2009) lists a number of terms that are and were in use:

> […] 'signals of continued attention', 'recognition', 'concurrent feedback', 'accompaniment signals', 'listener responses', 'assent terms', 'back channel' or 'backchannel responses', 'encourager', 'continuers', 'limited feedback', 'responsive listener cues', 'minimal responses', 'reactive tokens', 'acknowledgment tokens', 'receipt tokens', 'response tokens', and 'project markers' (ibid., p. 104; citations omitted)

This list is far from complete. Fujimoto (2007), in a critique of the terminology of the widely used term 'backchannel', lists further terms and we have additionally encountered the names 'non-lexical speech sounds', 'conversational grunts' (Ward 2000), 'listener vocalisations' (Pammi 2011), '(affirmative) cue words' (Lai 2010; Gravano et al. 2012), 'interjections' (Ehlich 1986), and 'discourse particles' (Siegert et al. 2013).

The diversity in terminology has various reasons. It can, for example, be seen as evidence that the phenomenon has received attention from various research fields and research traditions. Communicative listener feedback has been investigated in sociology, psychology, linguistics, computer science and cognitive science. Even within these fields, various sub-fields created their own terminology based on established terms or due to traditional preference for coining terms. Some of the terms also reflect the specific research question and focus that investigators had in mind.

The diversity in terminology could be seen as valuable for mapping the phenomenon. Splitting up the different terms above into their individual components allows us to make a first characterisation of feedback in dialogue. Feedback is produced by 'listeners', in 'response' or as a 'reaction' to utterances by a speaker. It is produced 'concurrently' and in 'accompaniment' to the speaker's actions, and uses a 'back channel' — not the main 'channel' — of communication. Although instances of feedback may actually be words (e.g., exclamations, interjections, discourse markers) researchers do not primarily see them as words, but rather as 'cues', 'markers', 'responses' or 'signals'. This choice in terminology is suggestive of their usage in dialogical interaction, especially of the fact that they are being produced for the interaction partner. In their form they are 'tokens' or 'terms', which suggests a certain compactness. Its extent is characterised as being 'limited' and 'minimal' in contrast to normal utterances in dialogues. Feedback communicates 'continued attention', 'recognition', affirmation and 'assent' of the listener and it 'acknowledges' a speaker's actions. Furthermore, it is used to 'encourage' the speaker to continue. Feedback is often a 'non-lexical' 'sound' or 'vocalisation' that is sometimes even characterised as 'grunt'-like.

Where the terms 'continuer' and 'encourager' suggest that a listener wants the speaker to continue, 'signal of continued attention', 'recognition', 'assent term', 'acknowledgement token', and 'receipt token' imply that listeners communicate their inner state. The terms 'accompaniment signals', 'backchannel', and 'concurrent feedback' highlight the aspect that they do not take centre stage in communication but rather occur in the background. 'Limited feedback' and 'minimal responses' characterise their form, namely that they are short. Finally, 'reactive tokens', 'response tokens', and 'responsive listener cues' highlight the fact that feedback may be given in response to something that occurred.

Not all of these terms operate on the same level and they all capture different aspects of the phenomenon. 'Interjections' or 'discourse markers' are broader linguistic concepts that can be produced by listeners and speakers alike and also play a role in written language. In contrast to this, terms such as 'continuer' or 'affirmative cue word' imply a very specific feedback function and, therefore, exclude feedback that may communicate different functions. 'Concurrent feedback' and 'listener responses' are terms that fall in between these two extremes. They do not imply any specific function, yet they have been created specifically for the phenomenon at hand. The widely used term 'back-channel', which originally focussed on the observation that listeners produce short 'messages' on a second 'channel' — which operates in parallel to the main channel of communication (Yngve 1970) — is now often used as a similarly broad term, but usually limited to the verbal/vocal modalities. Fujimoto (2007) rejects the term backchannel for its broadness — which she thinks makes it meaningless — and especially because it belittles the role that listeners and their feedback play in shaping a conversation. She therefore suggest usage of the neutral term 'listener response', as do Xudong (2009), and de Kok (2013).

We share Fujimoto's (2007) criticism of the term backchannel, but in addition to her points, we also criticise the term for its endorsement of a channel-like model of communication (Shannon 1948), which neglects important aspects of human conversation. Instead we prefer the term 'feedback signal' when referring to communicative feedback produced by listeners in dialogue. In contrast to the neutral term 'listener response', it alludes to its potential effects (causing change in a speaker's behaviour) but is still general enough to encompass the diversity of properties that will be described in the upcoming sections of this chapter. We begin with describing the form and structure of feedback signals.

3.3 FORM AND STRUCTURE OF FEEDBACK SIGNALS

Communicative feedback is an inconspicuous phenomenon that does not take centre-stage but is secluded in the background. Feedback takes place in the 'back channel' (Yngve 1970, p. 568), on 'track 2' (Clark 1996, p. 241). One of its defining features is that it does not adhere to — nor interferes with — the systematics of turn-taking. It does not occupy a turn, but may be placed with relatively few restrictions in parallel to an ongoing turn (see section 3.5). To be unobtrusive, feedback signals are generally (i) short (i.e., consist of minimal verbal/vocal expressions),[25] (ii) locally adapted to their prosodic context (i.e., the speaker's utterance) by being more similar in pitch to their immediate surrounding than regular utterances (Heldner et al. 2010), or (iii) taking place in the visual modality, for example as head gestures or facial expressions (Allwood and Cerrato 2003; Allwood, Kopp et al. 2007).

In the following sections we analyse the form properties of short verbal/vocal feedback expressions and non-verbal embodied feedback in detail.

[25] Longer forms of verbal feedback are possible as well, even normal utterances may be characterised as feedback that listeners produce for their interaction partners. In this thesis, however, we focus solely on short feedback expressions.

3.3.1 SHORT VERBAL/VOCAL FEEDBACK

Following Allwood and Cerrato (2003), we consider feedback 'verbal/vocal', if it is spoken, i.e., produced as a speech sound in the vocal tract of a listener. Examples of such feedback found in the ALICO-corpus (Malisz et al. 2016) are *genau* ('exactly'), *ja* ('yes'), *mhm* ('uh-huh'), and *m*.

Genau and *ja* are regular German words. *Genau* is an adverb that is used in the same way as its English counterpart *exactly*. *Ja* is a particle, that can be used in the same way as the English *yes*, e.g., to affirmatively answer a polar question. Both words are lexical in the classic sense: they can be found in a German dictionary (such as Duden [2013]), and they have conventionalised forms, pronunciations, and meanings.

Mhm and *m* are somewhat different from *genau* and *ja*. *Mhm* is listed in the Duden as well, as a discourse particle.[26] *m*[27], however, is not listed there, although it is more frequent in the ALICO-corpus (there are 346 occurrences of *m* and 191 occurrences of *mhm*). This suggests that *mhm* is a borderline case of a lexical entry and, indeed, verbal feedback expressions such as *mhm*, *m*, and the like are considered to be non-lexical in nature, and sometimes categorised as 'conversational grunts' (Ward 2000; Ward 2006; Neiberg and Gustafson 2010).

The differences in the surface form of these four feedback expressions is that *genau* is verbal, whereas *mhm* and *m* are considered interjections, inter alia because of the variability in their phonetic-phonological structure (Pompino-Marschall 2004) and intonational structure (Ehlich 1986, § 3.3.1). Because of this, Allwood and Cerrato (2003) call them 'vocal'. *Ja* lies somewhere in between. It is lexical, but also carries the property of many non-lexical vocal feedback expressions in that it is simple in its basic structure and can be easily modified prosodically since it is sonorant in form (Stocksmeier et al. 2007). *Ja* can also be used as an interjection.

Allwood (1988), in an analysis of Swedish feedback, makes a similar observation and distinguishes two groups of 'feedback morphemes'[28]: primary and secondary[29]. Primary feedback morphemes mainly express the basic communicative functions of feedback. In contrast to this, secondary feedback morphemes additionally express more specific functions and can often be used predicatively, attributively, and ad-

[26] The two meanings of *mhm* that Duden (2013) provides are (1) 'drückt (zögernde) Zustimmung aus' (expresses [hesitant] affirmation) or (2) 'drückt Nachdenklichkeit aus' (expresses thoughtfulness).

[27] Following the Chicago Manual of Style (2010, § 7.30) 'for [interjections] not found in the dictionary — or where a different emphasis is required — plausible spellings should be sought in literature or invented.'

[28] Allwood does not use term morpheme in its strict sense.

[29] This distinction parallels the one commonly made for interjections: primary interjections are considered to be those words and non-words which exclusively belong to the part-of-speech class interjection. And secondary interjections are words that belong to other part-of-speech classes that can be used as interjections as well (e.g., Ameka 1992, p. 105).

verbially. Allwood acknowledges that the distinction between primary and secondary feedback morphemes is one of degree rather than clear cut. Examples for primary single feedback morphemes of Swedish that Allwood identified are *ja* ('yes'), *nej* ('no'), *okej* ('okay'), *jo* (similar to German *doch*; there is no equivalent word in English), and also sounds like *m*, *n* and more (Allwood 1988, p. 95). As examples for secondary single feedback morphemes Allwood provides *bra* ('good'), *precis* ('exactly'), or *aldrig* ('never') and also exclamations such as *aj* ('ouch') or *usch* ('ugh').

From these single feedback morphemes, a large number of feedback expressions can be build by applying linguistic (prosodic, phonologic, morphologic and syntactic) operations. For Swedish, Allwood (ibid., pp. 96–98) names the following operations: changing intonation and voice quality (prosodic operations, see section 3.3.2); lengthening/shortening, omission of initial consonant, reduplication, repetition, reduplication with glottal stop, reduplication with *h*, omission of reduplicated prefix, addition of vowel, and breathiness (phonologic operations); reduplication, repetition, derivation, and composition (morphologic operations); and creating two word expressions (syntactic operation).

These operations are also constructive for communicative feedback expressions in German. As mentioned above, *ja* can be modified prosodically (Stocksmeier et al. 2007), but is also subject to morphological changes, e.g., through reduplication (*jaja* Golato and Fagyal 2008). Similarly, *m*, *mhm*, and *hm* — and other combinations of these sound — are constructed through these operations (Ehlich 1979; 1986). Analyses of feedback expressions in American English come to similar findings (e.g., Ward 2006). This suggests that modification of feedback expressions are a cross-linguistic phenomenon.[30] Table 3.1 shows examples of communicative feedback expressions that are subject to these operations in these three languages.

3.3.2 PROSODY AND INTONATION OF FEEDBACK

Whereas the phonological, morphological and syntactical operations on short verbal/-vocal feedback signals expand the discrete space of potential feedback signals combinatorially, prosodical operations add further dimensions that are rather continuous in nature.

Prosody encompasses acoustic features such as 'the perceived F_0 pattern' (intonation) as well as 'pauses, relative loudness, voice quality, duration, and segmental phenomena related to varying strengthening of the speech organs' (Vaissière 2005, p. 238). Prosody serves multiple (pragmatic) functions (ibid., tbl. 1), an analysis of what

[30] English, German and Swedish are closely related languages, but modification of feedback expressions occur in non-Germanic languages as well, e.g., in French (Prévot et al. 2016) or Japanese (Den et al. 2012).

Table 3.1: Operations for building feedback expressions from single feedback morphemes (Allwood 1988) in Swedish, American English, and German.

		Examples		
Type	Operation	Swedish	American English	German
prosodical	intonation	●	●	●
	voice quality	●	●	●
phonological	lengthening	ja: – ja::	okay – ookay	ja: – ja::
	shortening	ja: – ja	—	ja: – ja
	omission of initial consonant	ja: – a:	yeah – eah	—
	reduplication	ja: – ja:a:	mm – mmm	ja: – ja:a:
	repetition	ja: – ja: a:	mm – mm-mm	ja: – ja: a:
	reduplication with glottal stop	ja: – ja:'a:	—	m – m'm
	reduplication with h	ja: – ja:ha:	mm – mm-hm	ja: – ja:ha:
	omission of reduplicated prefix	m – mhm – hm	—	m – mhm – hm
	addition of a vowel	ja: – ja:α	oh – oh-eh	ja: – jo:a:
	breathiness	ja: – $_{o}$ja:	mm – hmm	m – hm
morphological	reduplication	ja – jaja	—	jaja
	repetition	ja – ja ja	yeah – yeah yeah	ja – ja ja ja
	derivation	oj – ojsan	—	—
	composition	ja, så – jaså	myeah	achja
syntactical	two word expression	ja säkert	yeah okay	ja genau

Sources: Swedish examples from Allwood (1988), American English examples from Ward (2006), German examples from the ALICO-corpus (Malisz et al. 2016). A '—' indicates that no example was described, or could be found in the corpus.

contributes to the perception of these, however, is difficult since (i) usually multiple acoustic features occur simultaneously (Vaissière 2005, tbl. 239), and (ii) perception of function interacts heavily with discourse context (ibid., p. 242).

Remarkably, the prosody and intonation of feedback signals (and interjections in general) is similar to the prosody and intonation of complete utterances (Ehlich 1986, pp. 36–37). Prosody plays an important role in feedback form, as its modifies feedback functions and meaning (see section 3.4) and may transport further subtle cues of listeners' mental states.

For German, for example, the intonation of the feedback expressions and discourse particles *hm*, *m*, *mhm* (and further variants) have been researched quite extensively. As early as 1913, Hermann described the wide variety in meaning — he lists 17 — which different forms of these vocalisation have (Hermann 1913, pp. 27–29). Ehlich (1986, pp. 36–44) compared the effect that the intonation of *hm* has on its meaning to the tone systems of languages such as Chinese. He identified four intonation contours with different meanings (convergence, divergence, pre-divergence, and complex divergence) which, however, were stable (with subtle variation) across different phonological and morphological forms, such as for example *hm̀* and *hmhm̀* (ibid., p. 54; tbl. IV). Schmidt (2001, p. 25; tbl. 3) defined seven prototypical form-meaning mappings of intonation for *hm*, some of which serve further communication management functions such as turn-taking, closing, and feedback elicitation.

Building on the intonation contours identified by Ehlich (1986), Stocksmeier et al. (2007) synthesised twelve different variants of the German feedback signal *ja* and let them rate on seven semantic differentials: happy — sad, brave — anxious, certain — hesitant, agreeing — rejecting, pushing — not pushing, surprised — bored, and angry — balanced. Three clear clusters emerged — agreeing, boredom, and hesitation — and almost all properties from the differentials received high rating for some of the synthesised signals.

In Malisz et al. (2012), we analysed 24 acoustic features of three German feedback expressions (*ja*, *m*, and *mhm*) in the ALICO-corpus and identified prosodic correlates of listeners' attentiveness (in contrast to distraction), as well as of the level of feedback functions. Attentiveness, for example, could be predicted through higher mean intensity, higher energy variability, and higher pitch variability. For feedback function the results were less clear since prosody often interacted with the segmental structure of the different feedback expressions.

Similar analyses were also carried out for American English. Ward (2004, tbl. 2) identified a range of prosodic features of feedback signals — namely 'syllabification, duration, loudness, pitch height, pitch downslope/upslope, [and] creaky voice' — as meaning bearing. He assigned each feature a vague meaning continuum with which it is supposed to covary (e.g., pitch height with listeners' degree of interest; pitch slope with her degree of understanding — or lack thereof).

More concretely, Gravano et al. (2012) showed that the pragmatic functions of a range of American English affirmative cue words (as either a back-channel or an acknowledgement/agreement signal) can be distinguished by their prosodic features. Beňuš (2012) showed that similar distinctions can be found for the polysemous discourse marker *no* in Slovak, too.

Lai (2008), on the other hand found that simple prosodic features (intonation, duration, and intensity) are not sufficient to distinguish between the plain back-channel function and a back-channel question function of the feedback expression *really*. In later work, when trying to distinguish between a surprise and question function of *really* and *right*, Lai (2009) found that prosodic features interact with the semantics of these expressions. In further investigation of such interactions, Lai also showed this for uncertainty, the presence of which was rather reflected in the choice of a feedback expression and not so much through (rising) intonation of the expression, as was expected. Rising intonation in a feedback signal, however, was indicative of listeners' problems in understanding speakers' utterances (Lai 2010).

In contrast to these complex relationships of prosody, semantics, and feedback function and meaning, van Zyl and Hanekom (2012) found that prosodic cues can be used to identify when a listener expresses a state of reluctance with the feedback expression *okay*, with the simple feature of duration to be the best predictor. Neiberg et al. (2013) came to similar conclusions when investigating the acoustic correlates of Swedish feedback expressions. They found that although feedback expressions have inherent meaning, different prosodic realisations have different functions.

Abstracting from concrete feedback tokens and analysing different realisations of a generic feedback-like vocalisation (*na*), Philippsen et al. (2013) were able to classify, based only on prosodic features, whether positive or negative polarity is expressed. The classifier could also be applied, with reasonable results, on natural German feedback signals.

Recently, Lotz et al. (2016), using the ALICO-corpus could distinguish two of Schmidt's (2001) seven functional meanings of *hm* using pitch contour only.

3.3.3 EMBODIED, NON-VERBAL FEEDBACK

Human feedback is not limited to the verbal/vocal modality. As is generally the case in human face-to-face communication, nonverbal modalities are important, too, and feedback can be expressed, e.g., through head gestures, facial expressions, gaze, or manual gestures. This has the advantage that it happens in the visual modality and interferes even less with the interlocutors' speech signal when produced concurrently to an ongoing utterance.

As in verbal/vocal feedback, non-verbal feedback signals combine discrete and continuous features (Allwood, Kopp et al. 2007, § 1.1.2), which makes them comparable in their expressiveness. Head movements, for example, can be categorised into a small set of discrete gestures — such as nods, shakes, tilts — each describable with simple rotation- and/or translation-based movement patterns (Wagner et al. 2014, § 2.2; fig. 1). Head gesture units can be comprised of such a single head movement, but can also form phrases that combine multiple such movements (Heylen 2008, p. 252) — either of the same or of different movement types (or combinations thereof). In the former case, the head gesture is polycyclic, but 'simple', in the latter case polycyclic and 'complex' (Malisz et al. 2016, § 5). In ALICO, for example, more than 70 % of listeners' head gestures were polycyclic, 23.1 % of which were complex (ibid., tbls. 8, 9).

Compositionality allows for a very large number of different head gesture units. In ALICO, for example, 20 listeners produced head gesture units with 303 different discrete forms — 71 % of which occur only once and 88 % less then five times. In contrast to this, the five most frequent head gestures units, all of them simple nod units with different cyclicity, make up 60 % of all observed head gestures.

In addition to these variations in a head gesture unit's discrete form, the kinematics of the movement is variable as well (Heylen 2008, § 5), making continuous adjustments possible: listeners can for example influence the amount of energy that is put into a head gesture unit by varying the amplitude of the individual movements, or by varying the duration of the complete head gesture unit (or its individual constituents).

Both aspects of listener head gestures have an influence on the meaning and function that can be derived from a feedback signal. Even in isolation, e.g., when not accompanying verbal/vocal feedback, some head gestures units are associated with certain meanings: nods are generally considered to be positive and to signal agreement (e.g., Poggi et al. 2010a; Poggi et al. 2010b), shakes transport negativity (e.g., Kendon 2002, pp. 151–152), and jerks transport surprisal and understanding (e.g., Allwood and Cerrato 2003, § 3.2). Analysing these coarse meanings in greater detail, however reveals that each gesture can actually fulfil different functions. Analysing rather different feedback functions of head nods (confirmation, taking note, and agreement), Poggi et al. (2010b) found differences in form features. The variation is, however, rather subtle such that discourse context needs to be taken into account as well in order to differentiate between functions. They also found that, in combination with other facial displays, nods may actually express disagreement or display ongoing processing.

Similar to how the perceived meaning of short verbal/vocal feedback expressions changes with prosodic realisation, head gestures were found to interact with verbal/vocal feedback as well. When accompanying verbal/vocal feedback, the perceived meaning of head gesture units containing a single head movement was found to be mainly determined by the function of the co-occurring verbal/vocal feedback signals (Allwood and Cerrato 2003, § 3.2). Complex head gestures, on the other hand, were found to modify the function of the verbal/vocal feedback expression (they may, for example reinforce, enforce, weaken, or contradict the function [ibid., § 2.3]).

Similar observations can be made about other non-verbal feedback mechanism, especially facial expressions, which may convey listeners' mental states such as uncertainty (Krahmer and Swerts 2005), emotional stance (Kaukomaa et al. 2015), or the cognitive effort of a listener (van Amelsvoort et al. 2013). Heylen et al. (2007), for example, evaluated the feedback function of 21 non-verbal signals (various head gestures, facial expressions, and combinations thereof) synthesised with an embodied conversational agent and identified prototypical expressions for ten out of twelve functions that were analysed: accept — refuse, agree — disagree, like — dislike, understand — not understand, disbelief, and not interested (ibid., § 2.1).

Gaze, on the other hand, is quite a different signal. Gaze is often merely indicative of listeners' cognitive processes and states (e.g., scanning the environment for referents, looking at the object being referred to [Tanenhaus and Brown-Schmidt 2008; Garoufi et al. 2016]). Gazing at specific points or targets can thus be evidence of understanding — or non-understanding — at least when talking about things that are part of the physical situation in which the interaction takes place. But gaze can also be used by listeners as a feedback signal, e.g., to show that they are in contact or that they are ready to hear more. Nakano et al. (2003, p. 555–556) found that speakers use listener gaze cues of both sorts (indicated and signalled) as feedback and adapt their dialogue behaviour accordingly (e.g., by either elaborating the current topic or continuing with the next one). Clark and Krych (2004, p. 76) found that speakers actively monitor their interaction partners' gaze behaviour, use it as evidence, and, if necessary, adapt their behaviour immediately. Garoufi et al. (2016) actually utilised listener gaze as feedback in an interactive natural language generation system and improved the system's task performance.

Referring to the research program of 'embodied cognition' (Wilson and Foglia 2015), Allwood, Kopp et al. (2007) and Kopp et al. (2008) argue that non-verbal feedback is 'embodied', i.e., directly caused by biological and/or cognitive processes in the listeners' bodies. This, they argue enables the kind of expressiveness in feedback that is difficult to capture with classic approaches to semantics. Speakers — having embodiments similar to listeners — can, however, 'ground perception and understanding of physical expressions of the other in own bodily experiences' (ibid., p. 22), which enables them to interpret their interaction partners' multidimensional and multilayered feedback signals.

3.3.4 INTERMEDIATE SUMMARY: FORM

In summary we can say that communicative listener feedback signals, even though they are seemingly small and produced unobtrusively, are very rich in their form. The set of basic signals — those to which a form of conventionalised meaning could be ascribed (*ja*, *mhm*, etc.; nod and shake head gestures, certain facial expressions) — is rather small in comparison to the total number of words that natural languages have. As it is possible to modify the form of these basic unimodal signals through various operations and by the possibility to form multimodal combinations, the actual space of feedback signals that can be produced is extensive though. This abundance of possible signals is used by listeners when producing feedback as it allows them to express rather subtle differences in meaning, e.g., they cannot just express understanding, but can modulate how certain (or uncertain) they are of their understanding onto the signal.

3.4 MEANING AND FUNCTION OF LISTENER FEEDBACK

The analysis of the origins of the concept of feedback (section 3.1) and of the different terms used to describe the phenomenon (section 3.2) has already shown some of the roles and functions that communicative listener feedback fulfils in dialogical interaction. The discussion of feedback form (section 3.3) has shown that the relation between form and meaning is complex and that subtle differences in meaning, difficult to capture in traditional approaches to semantics, can be expressed.

Feedback signals can be assigned functions on various levels of granularity. The simplest approach is to assign them the dialogue act 'backchannel' or see them as having the interactional function to show continuous attention. The field of conversation analysis distinguishes between feedback signals that are continuers and feedback signals that are assessments (e.g., Goodwin 1986), and/or feedback that signals changes-of-state (e.g., Heritage 1984). Although conversation analysts are aware of the richness of feedback — see for example Heritage (ibid., p. 335) — they analyse the interactional functions of individual feedback signals (e.g., *oh*) one by one.

A different perspective is to distinguish between 'generic' and 'specific' feedback signals (Bavelas et al. 2000, tbl. 1), where generic signals are appropriate in many situations and specific signals are appropriate specifically in the situations in which they are produced — i.e., they relate to the content of the interaction partners' utterances.

A different approach is taken by Allwood et al. (1992), who aim at modelling a broader spectrum of feedback meaning. They concentrate on three semantic and pragmatic dimensions of feedback signals: their communicative function, polarity, and communicative status. These aspects will be discussed in the following:

Allwood and colleagues hold the view that communicative feedback is a 'linguistic [mechanism] which enable[s] the participants of a conversation to exchange information about four basic communicative functions' (ibid., pp. 2–3), namely 'contact', 'perception', 'understanding', and 'attitudinal reactions'. This list has been expanded with the functions 'acceptance' (Allwood, Cerrato et al. 2007) and 'agreement' (Buschmeier and Kopp 2012b).

Allwood et al. derive these functions directly from fundamental properties of communication. For example, communication is impossible when the potential communicators are not in contact. Being in contact requires co-presence in time (delays in transmission are acceptable as long as they are expected) and space (in an abstract sense, possibly mediated), and the ability to transmit and receive information. Furthermore, contact between two communicators is only established when both have the desire to communicate. Feedback fulfilling the function contact thus expresses whether communicators are 'able' and 'willing' to communicate (Allwood et al. 1992).

Similarly, communication requires the ability and willingness of communicators to perceive and understand each other. When co-presence in time and space is established, perception can be impaired for several reasons. One possibility is that the 'communication channel' is 'noisy' (Shannon 1948). This is the case when interfering noise is present in the environment, but also when communication is interrupted for other, e.g., technical, reasons. Perception can also be impaired when the communicators have problems with articulation or the ability to hear. When perception is possible, the next barrier to successful communication are problems in understanding (see section 2.1). These may emerge for several reasons and on different stages of language comprehension. Interlocutors might speak different languages, or be on different levels of proficiency, or use different (regional) dialects of the same language. In general, listeners might not know the meaning of a word that the other interlocutor uses, or they might not be able to infer the speaker's meaning from the utterance, or integrate the content that is expressed into the dialogue context (e.g., they may not be able to resolve an anaphora or a referring expression), or follow a line of arguments (e.g., because they missed a cue word of a rhetorical relation). The reasons for this are manifold, but once a listener has detected a problem, she can express it via communicative feedback.

Once understanding is possible and depending on the content of the utterance communicative listener feedback may communicate a listener's willingness and ability

Table 3.2: Hierarchies of communicative functions according to Allwood et al. (1992) and Allwood (2000), Clark (1996), and Bunt (2011). A level L_i is lower than a level L_{i+1}.

Level	Allwood et al.	Clark	DIT^{++} (Bunt)
L_5			execution
L_4	reaction to evocative intention	consideration/uptake	evaluation
L_3	understanding	understanding	understanding
L_2	perception	identification	perception
L_1	contact	attention	attention

to ‘accept’ the speaker’s utterance (e.g. a claim, or a proposal) or the listener’s willingness and ability to ‘agree’ with the speaker (e.g., to an expressed opinion). Listeners may also express their attitude towards utterances of their interaction partner, e.g., based on the listener’s affective state (e.g., being surprised, liking what was said, etc.).

In general, feedback signals that communicate the basic functions contact, perception and understanding are ‘process related’, whereas feedback communicating acceptance, agreement and attitudinal reactions is ‘content related’ (Allwood, Cerrato et al. 2007, p. 276).

In contrast to the interactional functions that are assigned to feedback in the field of conversation analysis, the communicative functions of Allwood et al. (1992; 2007; 2008) are rooted in the listeners’ cognitive state (Kopp et al. 2008, p. 29).

3.4.1 HIERARCHICAL RELATIONS

The basic communicative functions of feedback are neither independent nor on the same level. They are related to each other, forming a hierarchy with contact at its lowest end, followed by perception, understanding, and ending in acceptance and agreement at the top, see table 3.2.

The hierarchy is based on the assumption of ‘upward completion’ (Clark 1996, p. 147), meaning that, a lower level of processing L_i[31] needs to be successful for a

[31] Let $\mathcal{L} = \{L_\bot, \ldots, L_i, \ldots, L_\top\}$ be a (partially) ordered set of levels of processing, with $L_\bot$ being a minimal and $L_\top$ a maximal level. For two levels L_i and L_j, we write $L_i \prec L_j$ if L_i precedes L_j, $L_i \succ L_j$ if L_i succeeds L_j, and $L_i \simeq L_j$ if they are on the same level of processing.

higher level of processing L_{i+1} to be in reach. There is no perception without contact, no understanding without perception, and no acceptance or agreement without understanding.

Upward completion in combination with the assumption that interlocutors adhere to the cooperative principle (Grice 1975) allows for semantic and pragmatic inferences to be drawn (Bunt 2011, pp. 237–238). Explicit feedback of positive polarity on a level L_i (e.g., understanding),

(1) *entails* positive feedback on all preceding levels $L_{i-1}, \dots, L_\bot$ (e.g., perception and contact), and

(2) *implicates* negative feedback on the succeeding level L_{i+1} (e.g., acceptance or agreement).

Explicit feedback of negative polarity on this level, however,

(3) *entails* negative feedback on all succeeding levels $L_{i+1}, \dots, L_\top$ (e.g., acceptance or agreement), and

(4) *implicates* positive feedback on the preceding level L_{i-1} (e.g., perception).

See fig. 3.3 for a visualisation of these relationships.

Entailments (1) and (3) directly follow from the upward completion assumption. L_i being positive would not be possible if preceding levels were not positive as well. Negative feedback on L_i blocks upward completion and hence subsequent levels need to be negative. The cognitive reality of the assumption is questionable though. Although understanding, in general, requires perception, it is easy to imagine situations in which it does not, e.g., when dialogue context provides sufficient information for understanding to be possible. Similarly for the other levels. Allwood (2000, pp. 72–74) therefore assumes a weaker form of upward completion in which the entailment relations between levels of processing, as proclaimed by Bunt (2011), is a 'default chain of implications', that is, the entailment is defeasible.

Consequences (2) and (4), on the other hand, are 'upper-bounding implicata' (Horn 2004, p. 13) generated by the cooperative principle.[32] A rational, cooperative interlocutor provides an optimal amount of information. An ideal listener would have no reason to provide feedback of positive polarity on a level of processing L_i if she could as well provide feedback of positive polarity on the subsequent level of

[32] Upper-bounding implicata are chiefly due to Grice's (1975, p. 45) first sub-maxim of quantity — 'Make your contribution as informative as required (for the current purposes of the exchange)' — but see the discussion in Horn (2004, pp. 12–14).

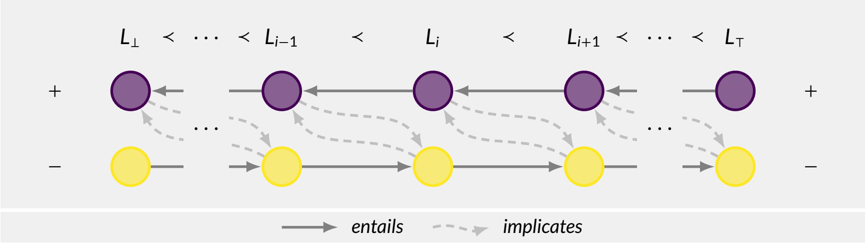

$L_\perp$ ≺ · · · ≺ L_{i-1} ≺ L_i ≺ L_{i+1} ≺ · · · ≺ $L_\top$
+
+
–
–
entails
implicates

Figure 3.3: Pragmatic relations among feedback functions of different levels of processing L_i as described in Bunt (2011, tbl. 5). When receiving feedback of positive polarity (+) on level L_i, positive feedback on all preceding levels L_{i-1} to $L_\perp$ is entailed and negative feedback on the directly succeeding level L_{i+1} is implicated. When receiving feedback of negative polarity (−) on level L_i, negative feedback on all succeeding levels L_{i+1} to $L_\top$ is entailed and positive feedback on the directly preceding level L_{i-1} is implicated.

processing L_{i+1}. Consequentially, the optimal level L_h for providing positive feedback is the highest level of processing that is successful

$$L_h = \max_i \left\{ L_i \in \mathcal{L} \mid L_i = \mathit{success} \right\}.$$

It follows that no positive feedback could be provided on the subsequent level L_{h+1} and thus feedback of positive polarity on level L_h implicates a problem on level L_{h+1}.

Conversely, an ideal listener would have no reason to provide feedback of negative polarity on a level of processing L_i if she could as well provide feedback of negative polarity on the preceding level of processing L_{i-1}. Consequentially, the optimal level L_l for providing negative feedback is the lowest level of processing on which a problem occurred

$$L_l = \min_i \left\{ L_i \in \mathcal{L} \mid L_i = \mathit{problem} \right\}.$$

It follows that no negative feedback could be provided on the previous level L_{l-1} and thus feedback of negative polarity on level L_l implicates successful processing on level L_{l-1}.

It is apparent that L_h directly precedes L_l (and that L_l directly succeeds L_h). As feedback of positive polarity on level L_h implicates a problem on L_l and feedback of negative polarity on level L_l implicates success on level L_h, one implicates the other. It may look as if there is virtually no difference between feedback of negative polarity on level L_l and feedback of positive polarity on level L_h. Actual listeners (providing feedback), however, cannot rely on speakers to infer what they implicated and, conversely, speakers cannot rely on listeners actually meaning what seems to be implicated in their feedback (Bach 2006, p. 23).[33]

3.4.2 POLARITY

By choosing to either provide positive or negative feedback, a listener makes a statement about a specific level of processing.

Borrowing the terminology from cybernetics, feedback can be either negative or positive. Whereas in cybernetics negative and positive feedback are principled ways in which a system is controlled — classically negative feedback leads to stability, positive feedback leads to instability and growth (Krippendorff 1986, pp. 22–23, 30) — , manifestation of polarity in communicative feedback is usually interpreted differently: negative feedback tells the speaker that a problem occurred and that something needs improvement. With positive feedback, on the other hand, listeners communicate that everything is alright and that speakers can continue as is — or even reduce their effort.

It is important to note though that the polarity of a feedback signal is, in general, not directly observable from its basic form. To be able to interpret a feedback signal, it needs to be analysed in its dialogue context. Allwood et al. (1992, pp. 8–10) make the case that — depending on the context in form of the preceding speech/dialogue act — an inherently negative *no* can actually be positive feedback of acceptance and an inherently positive *yes* can serve as a signal of rejection.

3.4.3 AWARENESS AND INTENTIONALITY

Feedback in dialogue can be produced on different levels of awareness and intentionality. For communicated information in general Allwood et al. (ibid.) state that gradual differences in awareness exist but, for simplicity, make three such levels explicit. Communicated information, and therefore feedback, may be 'indicated', 'displayed', or 'signalled'.

Indicated feedback is produced on the lowest level of awareness. Listener's are not necessarily aware of and possibly not in control of it. An example of indicated communicative feedback would be a listener blushing when her interlocutor makes an utterance that embarrasses her. As she is not in control of the blood flow into her face,

[33] One could argue that these implicatures are conventional in nature.

her reaction cannot be considered intentional, but is causal to her internal mental state. She may, however, feel the warmth and can be aware of it.

Displayed feedback is produced intentionally. The listener displays information and intends the speaker to see it — but does not expect, or even want, it to be recognised as being intentional. An example of displayed communicative feedback would be if a listener pretends to be happy about what her interlocutor says and displays this with a smile. If the smile is done right (a 'Duchenne' smile), it looks involuntary and suggests a genuine pleasure to the interlocutor, who is not able to recognise that there is an intention behind it. For him it appears to be indexical.

Signalled feedback is produced intentionally as well, but in contrast to displayed feedback, a listener also intents the speaker to recognise that it is displayed. An example of signalled feedback would be a listener that intends to communicate to her interlocutor that she understood and accepts a suggestion made by him by uttering a straight 'yes, okay'. This expressions of feedback displays understanding and acceptance linguistically, which makes it a signal by convention as 'ordinary linguistic expressions (verbal symbols) [are] signals by convention' (Allwood et al. 1992, p. 6). For the interlocutor, it is thus clear that the listener is intending him to recognise that she displays him understanding and acceptance.

In general, it may be difficult for a speaker who encounters listener feedback to recognise whether it is purely indicative, or displayed, or signalled by the listener. The same behaviour occurring in very similar situations can arise from different levels of intentionality and awareness. Consider the following example, set in a noisy environment such as a crowded bar, where, say, Stanisław addresses Lydia and where she does not react at all to his address (i.e., gives negative contact feedback).

If Lydia is not aware of the fact that she is being addressed by Stanisław (she neither hears nor sees him) she is also not aware and in control of the negative contact feedback she provides to him. Consequently, Lydia's feedback is *indicative* and the situation is a good example of Watzlawick et al.'s (1967, p. 51) insight that 'one cannot *not* communicate'.

Now imagine that Lydia is aware of Stanisław's address (she hears him), but pretends not to have noticed. Her behaviour is now intentional — she *displays* not being in contact — but at the same time she does not want him to recognise her intention. From Stanisław's perspective this situation does not differ from the one above and it is basically impossible for him to see whether Lydia's behaviour is purely indexical or a display.

Finally, imagine a slightly different situation in which it is virtually impossible for Lydia to miss Stanisław's address (he is in her line of sight, and clearly audible) but she is not willing to talk to him and ignores his address in the same way as above (she

Table 3.3: Allwood et al.'s (1992) levels of awareness and intentionality. Listener feedback can be purely indicative, or displayed, or signalled. For speakers it may be difficult to differentiate between displayed and indicated as well as indicated and signalled feedback.

	listener		speaker	
level	aware	intentional	detect intention	confusion
indicate	possibly	○	—	display
display	●	●	○	indicate/signal
signal	●	●	●	—

does not react). In this case she *signals* her non-willingness to interact as she cannot expect him not to detect and recognise the intention behind her feedback.

In the first two situations, if Stanisław, for whatever reasons, suspects that Lydia might ignore him, he might falsely attribute an intention to her (although the feedback is purely indexical) or recognise that she is displaying ignorance. He might even think that her ignorance of his address is a signal for him. In the third situation, Stanisław, not suspecting that Lydia would ignore him, may even falsely interpret her signal as being purely indexical.

Of these three levels of awareness and intentionality of feedback, displayed feedback poses most problems to speakers. Differentiating between indicated and signalled feedback is most probably possible in most situations. Displayed feedback, however, can be used ambiguously and may be confused with indicated or (if acted poorly) signalled feedback. An overview of the properties of the levels of awareness and intentionality for the listener and the speaker is given in table 3.3.

From the speaker's perspective, a more useful distinction is to differentiate between two dimensions of feedback: 'intentionality' and 'veridicality' (Nivre 1995). On the intentional dimension, speakers perceive feedback as being produced on a spectrum from intentionally to involuntary. Linguistic feedback is clearly produced intentionally, while indexical feedback is produced involuntarily. On the veridical dimension, feedback is either a truthful representation of the inner state of the listener or not. In general, non-intentional feedback is most likely veridical (cf. the blushing example above), whereas intentional feedback may be veridical or not.

Non-veridical intentional feedback does not necessarily imply a deceptive intention, nor does it mean that the listener does not want to be cooperative. A listener might, for example, intentionally signal understanding without even knowing that she has no real basis for claiming understanding (see section 2.1). She might truthfully think that she understands what is being explained to her and only later find out that in fact she did not. She might also intentionally non-veridically signal understanding for social reasons, for example, because she is confident that she will be able to understand in a while.

3.5 PLACEMENT AND TIMING

In general, '[n]o location restrictions are placed on the occurrence of back-channel signals' (Duncan and Fiske 1977, p. 202), that is, listeners may provide feedback at any point of time during a speaker's turn. Some points of time, however, seem to be more appropriate for feedback placement than others (e.g., de Kok 2013, pp. 31–34), that is, the probability that a feedback response can be observed at such a point is higher than at other points. These points, or perhaps intervals, of time can be seen as 'feedback opportunities' (ibid., p. 13) or 'feedback relevance spaces', 'intervals where it is relevant for another speaker to produce a backchannel' (Heldner et al. 2013, p. 137) — paralleling Sacks et al.'s (1974) notion of 'transition relevance places' at which turn-changes may take place.

As a surface phenomenon, feedback placement and timing has been researched extensively, often in combination with turn-taking. Based on the assumption that feedback is not timed randomly, but placed systematically, one particularly influential idea (due to Duncan 1974) has been that listeners place feedback (or take the turn) after the speaker displays a signal (a 'speaker within-turn signal' for feedback and a 'speaker turn-signal' for turn-taking) and that such signals are realised as observable 'behavioural cues'. Duncan found that two behavioural cues were good predictors for listener feedback that occurs between units of analysis (phonemic clauses that are multimodally marked; ibid., p. 164), or shortly after the subsequent unit begins: 'completion of a grammatical clause', and 'turning of the [...] head towards the [listener]' (ibid., p. 172). Using both cues simultaneously increased the probability of them being followed by feedback, and both cues preceded verbal/vocal and non-verbal feedback alike. No cues, however, could be identified for feedback that occurred within a unit of analysis (ibid., p. 173).

Duncan's analysis was grounded in only two recorded dialogues. Subsequent research on feedback placement built on his methodology and tried to identify behavioural cues preceding feedback on a larger scale, using more data, automatic feature (i.e., potential elicitation cue) extraction, and more powerful statistical methods. Such an approach was especially appealing to researchers in the field of spoken dialogue systems who hoped to be able to automatically identify cues in the speech of a human user and respond with appropriately placed feedback (seesection 3.6).

Koiso et al. (1998) use syntactic (part-of-speech) and prosodic features immediately preceding a feedback signal to predict feedback placement in a corpus of eight task-oriented dialogues in Japanese. They constructed a decision tree using prosody as a filter on the syntactic features — prosodic features need to be consulted when syntax is a potential cue, but syntactic features are often sufficient to rule out a cue. Their conclusion is that both prosodic as well as syntactic features are important cues for feedback placement. The constructed decision tree further suggests that looking at individual prosodic features is insufficient and feature combinations need to be taken into account (ibid., tbl. 8).

A systematic analysis of intonational cues was carried out by Ward and Tsukahara (2000) on eight conversations in American English and 18 conversations in Japanese. They identified a simple 'low pitch cue' to be a reliable predictor for feedback in both languages, and were able to formulate a precise rule of when feedback is produced:

> Upon detection of (P1) a region of pitch less than the 26th-percentile pitch level and (P2) continuing for at least 110 milliseconds, (P3) coming after at least 700 milliseconds of speech, (P4) providing you have not output back-channel feedback within the preceding 800 milliseconds, (P5) after 700 milliseconds wait, you should produce back-channel feedback (ibid., pp. 1186)

In contrast to Duncan (1974) and Koiso et al. (1998), Ward and Tsukahara did not use a fixed unit of analysis, but examined the signal quasi-continuously (in 10 ms steps). Hence, they did not need to make the assumption that cues are only produced at the end of certain units (Ward and Tsukahara 2000, p. 1202).

Ward and Tsukahara already intended their model to be used in an 'automated listener', a conversational agent that is able to provide feedback in response to the speech of a human user. Also working on the side of spoken dialogue systems, Cathcart et al. (2003) devised a model suitable for online-use based on pause duration in combination with an n-gram part-of-speech model (with $n = 3$), which outperformed Ward and Tsukahara's model by a factor of two (ibid., p. 57).

Morency et al. (2010) used sequential probabilistic models (Hidden Markov Models and Conditional Random Fields) to predict feedback placement in human computer interaction. This required them to use only features that can be derived in real-time from the users' multimodal behaviour (speech and gaze). During the training process of their model, they automatically selected the best features from a set of prosodic features (e.g., the individual parts of Ward and Tsukahara's low pitch cue, pauses, lengthening or emphasising of words), lexical features (e.g., word unigrams in contrast to Cathcart et al.'s part-of-speech trigrams, fillers, incomplete words), as well as one multimodal feature (whether the speaker is gazing at the listener). Each of these features was additionally encoded in multiple ways in order to model different relationships between a feature and feedback behaviour (strength of the trigger, delay of subsequent feedback). As a result of the training and feature selection process on 50 short dialogues, three features were chosen: occurrence of a pause, use of the word *and*, as well as the gaze feature encoded in two ways (Morency et al. 2010, p. 81).

One problem with all the above approaches is that they only identify those cues and places, where listeners actually responded with feedback. However, feedback elicitation cues do not necessarily impose a strict obligation on listeners to respond. A different person might have responded differently in the same situation. Consequently, if the person that happened to be the interaction partner in a dialogue did not respond to a feedback elicitation cue of the speaker, this behaviour will not be identified as a cue and missed in subsequent analyses.

This problem was identified and addressed by Huang et al. (2010) and de Kok and Heylen (2012), who explored ways to have different people act as listeners in the same dialogue situation. Huang et al. (2010) let multiple persons do a 'parasocial interaction' with a video recording of a speaker taken from a dialogue where the original listener is cut out. The parasocial-listeners are then instructed to put themselves into the dialogue situation and respond to the speaker by pressing a button every time they would give feedback. De Kok and Heylen used a more ecologically valid — but less scalable — approach that made three listeners believe that they are in a one-on-one interaction with a speaker (only one of them was, the other two were maintaining an illusion). Collecting data from multiple listeners has several benefits. First of all, it increases the number of cues that are revealed as individual listeners might respond to cues that the others do not respond to. Secondly, cues that are responded to by multiple listeners might be seen as more prominent or important cues than cues that get fewer responses. Thirdly, cues by multiple listeners that occur in proximity to each other can be used to measure out intervals that constitute response opportunities or feedback relevance spaces (de Kok 2013; Heldner et al. 2013).

As mentioned above, one reason for some feedback elicitation cues to be perceived as more prominent or important than others might be that multiple cues are produced simultaneously (Duncan 1974, p. 172), This was confirmed in a systematic analysis of prosodic and syntactic features at the boundary of units preceding feedback signals in American English, where Gravano and Hirschberg (2011, § 5.2) found a quadratic relationship between the number of simultaneous cues (up to six) and the likelihood that a feedback signal is produced.

The approaches to feedback-timing and placement discussed above focus on elicitation cues. They provide good insight into the form of potential cues and even provide mechanism to detect feedback opportunities in audiovisual signals. The perspective on feedback that these approaches hold is one of feedback as stimulus–response behaviour. Speakers provide elicitation cues (the stimuli) and listeners respond to these cues. Whether missed opportunities result from listeners' non-detection or voluntary ignorance of a cue is a question that these approaches do not address.

A different perspective on feedback placement focusses on the interactional functions it fulfils in dialogue. According to this view, listeners may, of course, respond to speakers' feedback elicitation cues (because they are cooperative dialogue partners and recognise a speaker's need of information), but listeners may as well provide feedback because they feel a need to communicate their state of understanding or their attitude. This may be the case for different reasons. In the case of difficulties in understanding, feedback that expresses such a state signals to the interlocutor that there is a problem. It can therefore be considered an attempt of the listener to initiate repair (i.e., 'other-initiated repair', Schegloff et al. 1977[34]). Similarly, the organisation of dialogue in terms of contributions (Clark and Schaefer 1989) necessitates a form of closure (see section 2.2.1).

Eshghi et al. (2015) model communicative feedback placement specifically in terms of its grounding function. Operating in the incremental syntactic and semantic framework DS-TTR — a combination of Dynamic Syntax (Kempson et al. 2001) and Type Theory with Records (Cooper 2005) — they see feedback as a mechanism to synchronise the grounding state between interlocutors. This state is represented as two pointers, a 'self' and an 'other' pointer, in the parser and generation context graph, which is maintained by each interlocutor. From the speaker's perspective, the self-pointer incrementally proceeds with each word uttered, while the other-pointer remains behind, at the latest position that is considered grounded. As soon as the speaker encounters — positive — feedback ('backchannels'; Eshghi et al. 2015) from the listener, the other-pointer is moved to the frontmost position in the graph. Feedback is furthermore associated with DS-TTR's computational action 'completion', which is only applicable at points of semantic completion (which accounts for appropriate feedback placement). In addition, the point where listener feedback occurs may clarify which of several possible interpretations a listener adopted (those that are complete at that point of time) and thus helps to consolidate the graph of possible parses. Alternatively,

[34] Interestingly, Schegloff (1982, p. 87f), discussing '"uh-huh" and other things that come in between sentences', states that such feedback signals, being of the type continuer, signal to the speaker that no repair of the preceding 'unit of talk' is needed. Specific paralinguistically enriched realisations of these tokens, however, can actually signal the need for repair (see, for example, the analysis in Uematsu [2000]).

it shows that the listener is predicting an upcoming semantic completion (if feedback occurs right before the completing words are uttered). In Eshghi and colleagues' model, whether and when to provide feedback is entirely decided by listeners, based on their processing of the speaker's speech and the conventionalised conversational function of feedback that is assumed in the model.

Ideas from elicitation-cue driven models and listener-intention driven models are combined in the dual-route feedback production model developed by Kopp et al. (2008), see fig. 3.4. This model consists of a 'deliberative' route that analyses the speaker's language and multimodal behaviour, evaluates whether it can perceive and understand what is said and meant, maintains the result of this evaluation in a 'listener state', and plans, generates, and produces feedback signals from the generic–specific continuum (see the discussion of Bavelas et al. [2000] in section 3.4). Generic feedback is produced in response to elicitation cues and does not necessarily take information from the listener state into account. In contrast to this, specific feedback can also be produced without an elicitation cue, e.g., when a problem in understanding arises. Concurrently, the 'reactive' route of the model produces feedback signals that are on lower levels of awareness (e.g., smiling, blushing; see section 3.4.3). The form of these signals results from the listener's emotional state (which is updated by appraisal of events such as her interlocutor's behaviour and the situation).

3.6 COMMUNICATIVE FEEDBACK IN ARTIFICIAL CONVERSATIONAL AGENTS

Communicative listener feedback has been a topic in computational sciences such as computational linguistics, artificial intelligence, and human–computer interaction. Most research has been carried out on the question of how feedback behaviour can be generated, i.e., in applications where the computer, in form of an artificial conversational agent, was supposed to produce communicative listener feedback in response to users speaking to it. Within this broad topic, most work has focussed on timing of feedback signals, as described in the previous section, but some work has focussed on the form of feedback signals that the agent should produce.

Nakano et al. (1999) describe a dialogue system that incrementally translates user utterances into a frame-based representation and produces feedback of different type depending on the state of this representation.

Wang et al. (2013) present a model for the production of multimodal communicative feedback behaviour that is able to generate both generic as well as specific feedback and takes into account the conversational agent's role in the dialogue (e.g., is it the

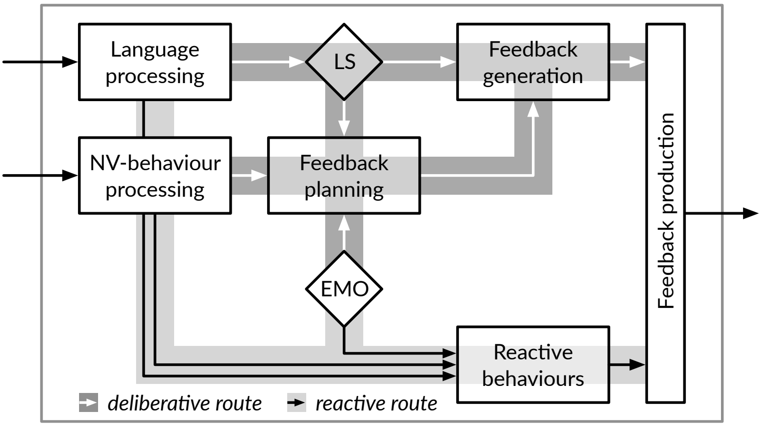


Figure 3.4: Dual-route architecture of the 'feedback system' for embodied conversational agents developed by Kopp et al. (2008, fig. 2, simplified). Using the 'deliberate' route (white arrows on dark grey background), the system plans, generates, and produces generic and specific feedback signals based on feedback elicitation cues as well as its current 'listening-related mental state' (LS). Using the 'reactive' route (black arrows on light grey background) the system concurrently produces feedback on a low level of awareness, mainly based on its current emotional state (EMO), possibly triggered by feedback elicitation cues.

addressee or merely a side participant) and its confidence in being able to understand the ongoing utterance.

Bevacqua (2009) and Schröder et al. (2012) describe a variety of 'sensitive artificial listeners' (SAL) with different personalities, that could provide multimodal feedback in response to the spoken language and multimodal behaviours of their human interlocutors. The SAL-agents were able to detect multimodal feedback elicitation cues of the user, based on which they then chose when to product feedback. The feedback generation model was capable of producing feedback signals on different levels of awareness, either purely reactive and neutral backchannel feedback or responsive feedback with a communicative function (similar to those defined by Allwood et al. [1992] and Kopp et al. [2008]). Which type of feedback got produced depended on whether the agents' model of mental state contained information that it deemed worth communicating to the interlocutor.

Mukai et al. (1999) model feedback production for a virtual conversational agent as a dynamic system that describes a changing 'desire' for inserting feedback. Placement and choice of modality (head gesture or verbal feedback expression) are decided upon based on the agent's internal desire level.

These systems take the artificial conversational agent's cognitive or emotional state into account when choosing feedback form. In contrast to this, Kawahara et al. (2016) describe a model that makes choices based on features derived from the preceding utterance of the interaction partner. They found in a corpus study that boundary type as well as syntactic complexity of an utterance are predictors for specific types of feedback expressions (e.g., complex syntax increases the likelihood of reduplication of an expression) and learn a model that is rated more naturally than a random baseline in a subjective evaluation study.

We now look at work that takes the opposite perspective on feedback processing: artificial conversational agents that recognise and interpret communicative listener feedback of their human interlocutors and react to it.

Dohsaka and Shimazu (1997) present an agent that generates utterances incrementally while simultaneously attending to interlocutor utterances, enabling immediate reaction by re-planning output (changing the content of explanation), if necessary. Nakano et al. (2003) present a simple probabilistic model for estimating groundedness of its utterances based on verbal and non-verbal feedback acts of the interaction partner, and for deciding how to proceed given this information (continue, elaborate, or repeat). Reidsma et al. (2011) explore various aspects of conversational agents that are able to produce feedback elicitation cues from their interlocutors, interpret their interlocutors' feedback, and adapt the timing of their communicative actions based on this feedback.

A slightly different approach is taken by Garoufi et al. (2016), who present a system that interactively takes its interaction partners' gaze behaviour and movements into account when generating referring expressions and instructions for a direction-giving task in a virtual environment. They could show that their adaptive system outperformed a non-adaptive system and was able to avoid confusion of the human interlocutor. The adaptive system was better in generating successful references and could generate adaptations earlier. The type of feedback that this system processes is different than in the agents described above because it is non-communicative (i.e., purely indicative). The focus of this work is on adaptive generation though, the model of which is more complex than in the adaptation mechanisms in the three models described above.

There has, apparently, not been much work on feedback-adaptive artificial computational agents. On the one hand this is surprising as the underlying ideas are not new. Maclay and Newman (1960, p. 226), e.g., write

> The differences in response to feedback are of considerable general interest. Effective communication, in addition to requiring an accurate perception of what the hearer is likely to understand, must involve a sensitivity to feedback. Is the speaker willing and able to alter his approach in the face of failure to communicate?

Similarly, Krauss and Weinheimer (1966, pp. 343–344) state that

> [...] the model of a speaker as intent upon effecting some end state in his listener, monitoring the listener's behavior for indications of change and adjusting his subsequent output on the basis of this information is plausible [...].

On the other hand, even though the models and agents described above are able to establish a feedback loop (they perceive interlocutor feedback and react to it), they are, scientifically, still at an early stage of modelling the problem (which the authors acknowledge). On the interpretation side, the recognisers are not sophisticated enough to be able to deal with the richness of the feedback phenomenon, which, given the complexity of the recognition and classification task, is hardly surprising. More basic research is need here. On the generation side, the adaptation choices of the agents described above are limited to a small set of (usually fixed) alternatives — Garoufi et al. (2016) is the exception. As will be discussed later on (section 6.2), some work on adaptive natural language generation exists (usually quite application-driven), but the general problem is also not yet understood well enough. Hence, more basic research is needed here as well.

3.7 SUMMARY AND CONCLUSION

Communicative feedback is a mechanism that listeners can use to provide 'evidence of understanding' (Clark 1996) in dialogical interaction. In this chapter, we argued that some properties of communicative feedback make it particularly suitable for this task.

As described in section 3.3, listeners do not need to have or take the turn in order to provide feedback. They can provide it any time, as soon as they consider it useful or necessary. Similarly, speakers can quickly elicit feedback from their interaction partners without interrupting their speech or yielding the turn. Because of this, the mechanism of communicative feedback is fast and allows speakers to incrementally adapt their ongoing utterance.

We saw that this is possible because feedback is relatively unobtrusive: (i) it makes extensive use of the non-verbal modalities, namely head gestures, facial expressions, and gaze. It thus does not interfere with speakers' linguistic processing. (ii) Verbal/vocal feedback expressions, that may interfere with speech processing, are often non-lexical, short, and prosodically similar to the speech context in which they occur. Because of this, speakers are less likely to take feedback signals from their interlocutors as turn taking attempts. Feedback does not just happen in the back channel conceptually, it is literally kept away from the main track of communication.

In this chapter we argued that, despite being short and unobtrusive in form, feedback signals are very expressive. This is the case because feedback signals are

- productive (a large number of discrete forms can be produced from a relatively small set of base forms),
- multimodal (feedback signals often consist of coordinated behaviours on different modalities),
- continuous (prosodic variation of feedback signals — verbal and non-verbal alike — allow for the expression of qualitative and gradual differences), and
- contextually embedded (feedback signals interact with the dialogue context).

Because of these properties, the meaning of feedback signals (the speaker intention) is mostly non-conventionalised. A plain *yeah* or even an *uh-huh* — in a typical back-channel position — might be regarded as conventional and can be processed routinely. The 'speaker meaning' that the listener has in mind when producing a complex multimodal feedback signal, one which is qualitatively loaded due to its prosodic realisation, however, can only be understood using pragmatic reasoning and may rely on embodied cognition.

Finally, we argued that communicative feedback is reflective of listeners' mental state with respect to language and dialogue processing. It indicates (or is used to signal) whether listeners are in contact with speakers, whether they are able and willing to perceive or understand what is being or has been said, whether they are able and willing to accept the message, to agree with it, and what their attitude is towards it (Allwood et al. 1992; Allwood, Cerrato et al. 2007). Furthermore, depending on its prosodic realisation, its placement, or its timing, feedback may also be indicative of the listeners' uncertainty about their own mental state, their urgency for providing feedback, the importance of this feedback item, and more such qualifiers to its basic communicative functions (Petukhova and Bunt 2010).

Because of these properties, listener feedback is a viable basis for estimating groundedness and common ground. Since the communicative functions of listener feedback reflect the interlocutor's internal state, a somewhat detailed picture of the interlocutor (and hence the dialogue) can be formed based on it. Especially the latter two properties suggest that feedback facilitates a form of mentalising about the cognitive state of the dialogue partner that goes beyond what is usually considered groundedness.

Work on computational models of feedback interpretation, as well as models of what to do with communicative listener feedback provided by human interlocutors in human–agent dialogue, is sparse. And the approaches that exist are at a relatively early stage of modelling the phenomenon and its role in dialogue.

PART II

A COMPUTATIONAL MODEL OF ATTENTIVE SPEAKING

CHAPTER 4
INTERACTIONAL INTELLIGENCE FOR ATTENTIVE SPEAKING

In the introduction (section 1.2) we declared this thesis' general objective to be an investigation of how 'interactional intelligence' (Levinson 1995) can be modelled computationally for artificial conversational agents and chose communicative listener feedback as the specific subject of this research. Given the two preceding chapters 2 and 3 as background for this investigation, we will argue in this short chapter that feedback is indeed an interesting and feasible choice for investigating interactional intelligence.

The most fundamental insight we gained so far is that interlocutors in conversation make meaning and establish joint understanding through interaction. New meaning cannot be simply encoded in and decoded from an utterance, but needs to be constructed through interaction. Consider the following example:

Starting from the established common ground between them, Susanne can contribute new information by making an attempt to present it in such a way that Lothar is likely able to understand and accept it. Lothar tries to infer what Susanne meant. If he is certain that he understood Susanne, he can accept her presentation by providing evidence of his understanding. Both can then update their common ground with this contribution. If not, he signals that he has problems understanding her presentation, (which itself is a presentation). Susanne then needs to understand Lothar's presentation in order to infer which aspect of her presentation might have caused the problem. She updates her representation of their common ground accordingly and can make a new attempt in communicating the meaning she wants to contribute, making a new presentation that is adapted in such a way Lothar is now likely able to understand and

⭑ This chapter contains material previously published in Buschmeier and Kopp (2011).

accept it. They engage in this collaborative interaction until Lothar can finally accept Susanne's presentation.

Through this process, their individual representations of their common ground become more accurate, making it easier to contribute to the discourse in the future.

Interactional intelligence is needed at various points in such collaborative meaning making. In the speaking role, dialogue participants need to be able to (i) produce utterances tailored to their addressees, (ii) interpret evidence of understanding from their addressees, (iii) update their individual representation of the common ground, and (iv) produce adapted utterances that take the updated common ground and/or addressees' evidence of understanding — or clarifying presentation — into account. In the listening role/as an addressee, dialogue participants need to be able to (v) interpret utterances from the speaker, (vi) update their individual representation of the common ground, and (vii) produce helpful evidence of understanding, or clarification, to the speaker. We believe that the interactional intelligence that an artificial conversational agent would need in order to be able to engage in such conversational interactions is not yet feasible from a scientific and technical point of view, e.g., because some fundamental aspects of the process are not yet understood well enough to be computationally modelled in sufficient detail.

We do belief, however, that we can conceptually and computationally model the processes of interactional intelligence that are needed for an artificial conversational agent that is limited to the role of the speaker, and limited to evidence of understanding in the form of communicative listener feedback (in contrast to unrestricted utterances that addressees can produce as acceptance, or clarifying presentation).

As argued in section 2.2.1, communicative listener feedback is a way of providing evidence of understanding that is often sufficient in strength (Clark and Brennan 1991, p. 224), but quickly produced because it is short and not subject to turn taking. The timeliness with which feedback can be produced can give it an accuracy in pointing out the source of a problem that may otherwise be difficult to articulate for listeners. Feedback thus occupies a middle ground between expressiveness and costs of production and interpretation of evidence of understanding, a property that benefits both listeners and speakers.

Here, we argue that the meaning of feedback, though rich and expressive, is much simpler to interpret than the meaning that unrestricted natural language utterances can bear. It is certainly different from interpreting unrestricted utterances in that it relies less on convention and compositional semantics and more on qualitative properties. Similar to natural language, feedback meaning interacts with context — and understanding feedback likely requires inference that takes context into account. The important difference, however, is that the context for feedback interpretation is only the dialogue itself — perhaps even just the preceding (or ongoing) utterance it refers to — , while the context for interpreting unrestricted natural language utterances is, additionally, the extensive world knowledge of dialogue participants. This difference makes utterance interpretation AI-complete (section 1.1) and, as we would argue, feedback interpretation feasible.

Limiting our model of interactional intelligence to the speaker role is also a useful simplification because the artificial conversational agent will have direct control over the dialogue and can thus define the direction that it takes itself.[35]

Limiting evidence of understanding to communicative listener feedback also makes the processes for updating the conversational agent's representation of common ground, and for adaptive language generation, more feasible from a conceptual and computational point of view.

More concretely, the objective of this thesis is to model processes needed for an ability that we name 'attentive speaking'. In addition to being able to produce natural language utterances, artificial conversational agents endowed with such processes, we call them 'attentive speaker agents',[36] should be able to

(1) interpret communicative listener feedback from their human interlocutors, taking the dialogue context into account, and

(2) adapt their ongoing and/or subsequent natural language utterances, paying heed to their interlocutors' needs — as inferred in (1).

In the introduction to this thesis (section 1.2), we already mentioned an additional property that specifically makes such agents 'attentive', namely a desire to be understood by their interlocutors. This means that attentive speaker agents should be willing to work towards being understood (sufficiently well for current purposes), and to make extra efforts — if necessary — to achieve this. This desire plays a role in (1) and (2), and additionally requires attentive speaker agents to

(3) invite feedback from their interlocutors by providing opportunities, or producing feedback elicitation cues, when needed.

[35] This is probably the main reason why computational models of feedback production are often limited to feedback timing. Producing meaningful feedback — which, in addition to timing, includes an informational intention, and the choice of an appropriate feedback form — in response to utterances of a human interlocutor requires the ability to understand (possibly unrestricted) natural language.

[36] The term was used by Reidsma et al. (2011), too.

This part of the thesis, spells out, conceptually and computationally, three processes, that address the requirements (1)–(3) for attentive speaker agents.

In chapter 5, we develop the process for interpreting evidence of understanding (and other listening-related mental states, namely contact, perception, acceptance, and agreement) in the form of communicative listener feedback provided concurrently by the interlocutors in response to an agent's ongoing behaviour. We frame this as mentalising, or mental state attribution, a process in which the attentive speaker agent theorises — using probabilistic inference — about the dynamic mental states of understanding of its interlocutors. This process is a 'minimal' approach to mentalising in that the resulting representation is simple and lightweight, yet more expressive than the one-bit variables in the 'minimal partner model' theory (Galati and Brennan 2010; Brennan et al. 2010, see section 2.3.5). Since the model takes the agent's dialogue context into account, it embodies a computational probabilistic pragmatics approach (Goodman and Frank 2016; Franke and Jäger 2016) to speaker (or rather listener) meaning.

In chapter 6 we then develop the process for adaptive language generation on different levels of processing (dialogue management, natural language generation, speech synthesis) that takes the attributed mental state — and local common ground — into account when making generation decisions.

Finally, in chapter 7, we develop a model of feedback elicitation based on the attentive speaker agent's needs. Based on this model, the agent can decide when and how to produce behavioural cues that elicit communicative feedback.

How these models are integrated into an actual attentive speaker agent is described and evaluated later, in chapters 8 and 9 in part III of this thesis.

CHAPTER 5
MENTAL STATE ATTRIBUTION BASED ON COMMUNICATIVE LISTENER FEEDBACK

In this chapter we develop a model for interpreting evidence of understanding (and other listening-related mental states, namely contact, perception, acceptance, and agreement) in the form of communicative listener feedback provided concurrently by interlocutors in response to the ongoing behaviour of a speaker. We begin by introducing the concepts of mental states and mentalising and, based on this, develop a conceptual and computational model. We first illustrate the feedback loop in dialogue and describe it in terms of a causal model. We then develop a formal representation of variables that correspond to listeners' listening-related mental states (an 'attributed listener state') and probabilistically model the interactions among these variables in a Bayesian network. Following this, we extend the model, step-by-step, by modelling the influences that properties of the feedback signal, information in the dialogue context, and the temporal dimension of the dialogue exert on the attributed listener state. We also show how the model can be used as a context-aware grounding criterion in a computational theory of grounding. We finish the chapter by discussing the theoretical merits, as well as limitations of our approach to interpreting and modelling communicative listener feedback.

5.1 INTRODUCTION

The central idea pursued in this thesis is that processing of communicative feedback in dialogue can be modelled as a form of 'mentalising'. Mentalising is an important concept in social cognition, defined to be a process for making inferences about the mental states of other agents (Frith and Frith 2006, p. 531). A prerequisite for

[⋆] This chapter contains material previously published in Buschmeier and Kopp (2011; 2012; 2013; 2014).

mentalising is that an agent has a 'theory of mind' (Premack and Woodruff 1978), that is, the agent introspectively perceives that it possesses mental states itself and hence believes that other agents, at least those that are similar, e.g., conspecifics, must possess mental states as well. Based on its empirical insight into the relationship between its own mental states and actions the agent can theorise about the mental states of other agents, and, based on their actions, 'attribute' mental states to them. How exactly the attribution and inference processes work in human cognition is a topic of ongoing research and depends, *inter alia*, on the 'category' of mental state that is to be inferred.

Mental states are generally thought to fall into two types (Boghossian 2009): There are 'propositional attitudes', such as beliefs, knowledge, intentions, fears, doubts, etc., that refer to or are about something (i.e., they have content). And there are 'phenomenal' states, such as pain, thirst, sadness, uncertainty, perception of the colour red, etc., that are identical to the quality of experiencing them, but are not about something. Hybrids of the two types of mental states are possible as well (Pitt 2013, § 3). Propositional attitudes can be about mental states of the phenomenal type (Mary believes that she sees a red rose) and mental states of the phenomenal type can be caused by propositional attitudes (Mary experiences a feeling of certainty because she knows that she knows everything about the colour red).

The most basic — perhaps trivial — assumption underlying the mentalising-based approach to feedback processing that we develop in this chapter is that listeners in dialogue have mental states that are specifically related to their task of listening. Consider the following example. Talking about their afternoon plans, Ludwig *perceives* that Sybille makes an utterance, he *understands* that she wants to visit the botanical gardens and he *agrees* that it will be nice looking at some exemplars of *Cephalanthera damasonium*.

Listening-related mental states are not limited to propositional attitudes, but can be phenomenal (and hybrid) as well. Ludwig may for example have a feeling of *non-understanding* that is the result of him not knowing what a *Cephalanthera damasonium* might be.[37] He may also feel generally *positive* as he usually enjoys afternoon activities with Sybille, or he may be hesitant accepting Sybille's suggestion as he does not like to go outside when it is too warm.

It suggests itself that these listening-related mental states play a role in listeners' feedback behaviours.[38] When Ludwig understands that Sybille is suggesting to visit the

[37] *Cephalanthera damasonium* is the orchid species of the year 2017, chosen by the German 'Arbeitskreise Heimische Orchideen'.

[38] Kopp et al.'s (2008) computational model of feedback generation for virtual agents (see section 3.5 and fig. 3.4) is based on this assumption. An agent's capability to deal with a user utterance (is the agent

botanical gardens, he may signal his understanding with a head nod and his agreement with an enthusiastic *yeah*. In this case Ludwigs' feedback signals have an underlying communicative intention. Being a cooperative interlocutor, he wants to inform Sybille about his listening-related mental states. In other cases his feedback signals might be purely indicative. A lengthened *yeah* could indicate that he is hesitant to accept the suggestion and a puzzled facial expression could unwillingly indicate his non-understanding of the term *Spiranthes aestivalis*. Here Ludwig's listening-related mental states become manifest in his behaviour even though he might not have intended to communicate them.

Given this background, we propose that speakers, when perceiving conversational feedback, reason about the listeners' listening-related mental states that caused them to produce their feedback behaviour. A computational model of this process will be described in the following sections.

5.2 A CAUSAL MODEL OF THE INTERACTION

We set out by illustrating one iteration of the feedback loop between speaker and listener in dialogue, see fig. 5.1, describing the model in the 'causal network' formalism (Jensen and Nielsen 2007, pp. 22–27)[39]. Causal networks allow us to describe the information flow (from cause to effect) in the model and are a subset of the formalism that will later be used for the actual model — Bayesian networks.

Let us revisit the dialogue between Sybille and Ludwig described above. Sybille, the speaker, produces an utterance in the presence of Ludwig, the listener, and is interested in what Ludwig's listening-related mental state towards her utterance turns out to be, e.g., whether Ludwig is in contact, has perceived, understood, and accepts or agrees with her utterance. As it is impossible for Sybille to directly observe Ludwig's mental states, she pays attention to his feedback behaviours and uses them as evidence in her mentalising processes, inferring Ludwig's mental states in a mental representation of her own, which we call 'attributed listener state' (ALS).

When modelling this interaction in a causal network, we have to take, for the moment, an observer perspective (instead of an agent-centric perspective). This per-

familiar with the words used? can it generate a coherent answer?) is mapped onto a simple representation of 'listener state'. This representation then informs the choices that are made in generating the agent's embodied feedback behaviours.

[39] In a causal network, variables represent causes, effects or both, and directed links between variables represent causality (Jensen and Nielsen 2007, pp. 22–27, 32–35). Consider, for example, a simple network $A \rightarrow B \rightarrow C$. The directed link from variable A to variable B models that A is a cause for B, and that B is an effect of A. The second directed link from B to the third variable C, makes B the cause of C. Being intermediate, it is possible that B is both an effect (of A) and a cause (of C).

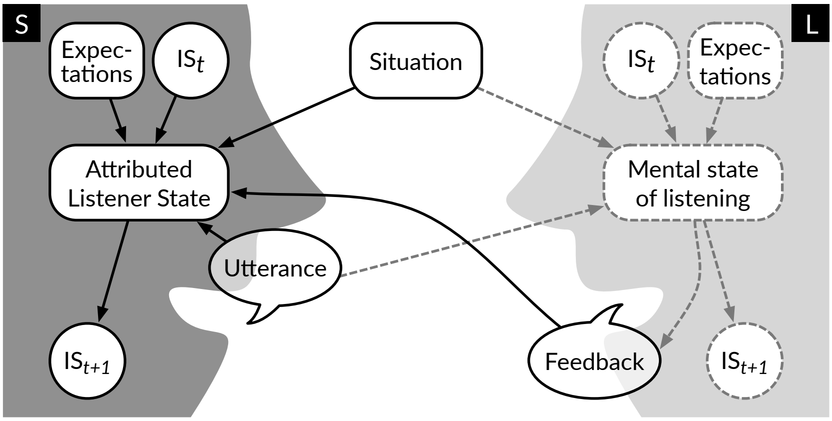


Figure 5.1: Causal network of the feedback loop between speaker *S* and listener *L* in dialogue. *S*'s utterances gives rise to specific listening-related mental states in *L*. These cause *L* to provide feedback signals, which *S* uses as evidence when attributing mental states to *L*.

spective unites both agents in a single network (variables and links unobservable to Sybille are drawn with grey dashed lines in fig. 5.1).

The information flow in the network is as follows[40]: Sybille produces an utterance. Ludwig perceives and processes this utterance, taking his current dialogue information state (IS$_t$, Larsson and Traum 2000), his expectations, and the communicative situation into account. His speech and language understanding processes give rise to specific listening-related mental states, which may subsequently be indicated, displayed, and signalled in his behaviour and actions. Sybille perceives these feedback signals of Ludwig and forms hypotheses about listening-related mental states that might have caused these behaviours. These hypotheses are represented in her attributed listener state, which she uses to update her dialogue information state (IS$_{t+1}$) and to formulate a subsequent utterance that may take Ludwig's feedback into account. This utterance closes the loop and creates the circularity of actions that is needed in a feedback driven system (Ashby 1956), see section 3.1.

[40] The temporal ordering of actions, as presented in the next paragraphs, is greatly simplified. As listener feedback is produced concurrently to a speaker's ongoing utterance the information needs to flow incrementally (see e.g., section 8.4).

This observer model can easily be reduced to an agent-centric model for Sybille which consists of only those influences that she can observe directly (drawn with black solid lines in fig. 5.1). Despite the resulting 'gap' in the causal chain, variables retain their roles as causes and/or effects.

This causal model will provide the scaffolding of a more detailed model to be presented next. Each variable is a mere place-holder for a complex network structure. These sub-networks are constructed according to information that is available and useful to model feedback processing in a speaker.

5.3 THE ATTRIBUTED LISTENER STATE (ALS)

Recall from section 3.4 that Allwood et al. (1992) hold the view that feedback in dialogue is used to 'exchange information about [a limited number of] basic communicative functions' (ibid., pp. 2–3). They specify these to be contact, perception, understanding, and attitudinal reactions. Kopp et al. (2008) extend this set by adding acceptance/agreement (previously considered an attitudinal reaction) and by regarding expressions of emotion as attitudinal reactions.

For our model we make the assumption that the type of listening-related mental states of a listener in dialogue correspond (one-to-one) to the basic communicative functions expressed through feedback. Feedback that communicates information about the basic communicative function 'understanding', for example[41], is grounded in a listener's mental state of being willing and able to 'understand'. Assuming a hybrid mental state that combines the propositional attitude 'belief' with the phenomenal state 'understanding', we can express this in modal epistemic logic[42] as

$$B_L(u),$$

where B is the belief operator, L is the agent, and u an atomic formula, here representing the phenomenal mental state understanding, that the agent believes it is in.

A listener may, however, not just be in either a state of full understanding or complete non-understanding. Understanding is not bipolar, but a gradual quality. It

[41] In development of the model, we will, at first, focus on the basic communicative function and mental state 'understanding' only. The other functions and listening-related mental states (contact, perception, acceptance, and agreement) are modelled analogously, but omitted for clarity of presentation.

[42] The syntax of modal epistemic logic for reasoning about knowledge and belief simply extends the syntax of propositional logic with modal operators for knowledge (K) and belief (B) and adds the syntactic rule that if φ is a formula, then $K\varphi$ and $B\varphi$ are formulae as well (Halpern 2003, pp. 244–245, 291). In systems of multiple agents — such as a listener and a speaker — , a subscript on a modal operator indicates which agent it refers to, e.g., $B_i\varphi$ is read as 'agent i believes that φ'.

can lie in between the two extremes. Feedback is rich enough in its form to express such qualifying information (see section 3.7; Petukhova and Bunt [2010]). In our model, we support this qualification by allowing atomic formulae of the form $U = u$, where U is a variable representing the listener's mental state of understanding and $\mathrm{Val}(U)$ the finite, non-empty set of values U can take (Milch and Koller 2000), e.g., grades of understanding (see section 2.1.4 and fig. 2.1). If, e.g., $\mathrm{Val}(U) = \{low, medium, high\}$ and a listener L believes her grade of understanding to be *high*, we write

$$B_L(U = high).$$

For practical reasons, we have chosen ternary grading (*low* — *medium* — *high*) for our model, but models with more grades (e.g., quinary grading with additional grades *medium-low* and *medium-high*), or continuous grading ($q \in [0\,..\,1]$) would have been possible as well. Importantly, we do not imply that a human listener's phenomenal state of understanding is ternary.

The 'attributed listener state' is a representation of a speaker's reconstruction of a listener's representation of her listening-related mental states. It is therefore identically structured, that is, it represents the same categories of mental states[43]. If the speaker inferred from a listener's feedback behaviour that her understanding is *high*, we can say that the speaker believes that the listener believes that her understanding is *high*, i.e.,

$$B_S(B_L(U = high)). \tag{5.1}$$

The mental state attribution process of the speaker is not a direct mapping from signal to meaning and neither is the feedback production process in the listener. Feedback signals are not symbolic or well defined, and, as described in section 3.7, they are conventionalised to a much lesser degree than other parts of speech, both on the form and function/meaning side. A reason for this is that they are very expressive. The abilities of feedback to convey graduation and uncertainty further contribute to this. It is thus clear that the mental state attribution process of the speaker has to deal with uncertainty.

Under uncertainty, a belief of the speaker concerning the mental state of the listener, such as the belief that the listener believes that her understanding is high (eq. [5.1]), becomes a matter of 'degree'. Such a subjective 'degree of belief' (or 'credence')

[43] We primarily make the assumption that the speaker's ALS and the listener's listening-related mental states have an identical structure for practical reasons. But even from a theoretical perspective this decision does not seem implausible. Presuming that theory of mind develops (ontogenetically) by introspectively discovering the possession of mental states and then forming the belief that other agents must have mental states as well it seems reasonable that agents expect others to have identical — or at least very similar — categories of mental states. The content of these mental states is very likely not identical, though.

can be regarded as the confidence that an agent has in an object of belief being true at a certain point of time (Huber 2009, pp. 1–2).

Degrees of belief are predominantly modelled as subjective probabilities that can be formalised in the mathematical framework of probability theory (ibid., p. 4). If an agent is certain about the truth of an object of belief, it assigns a subjective probability of $\theta = 1$ to it. Conversely, an object of belief that the agent considers to be false is assigned a subjective probability of $\theta = 0$. If an agent is uncertain about the truth of an object of belief, i.e., when it only beliefs in its truth to a certain degree, it assigns a probability of $\theta \in [0 \,..\, 1]$. The specific subjective probability that agents assign to an object of belief[44] is based on the evidence in favour (and/or against) its truth that they have encountered, possibly in combination with *a priori* information.

We adopt a subjective probability-based degree of belief approach for modelling a speaker's uncertain belief in the listener's belief in her level of understanding. In this framework, we denote the speaker's belief that the listener believes that her understanding is *high* (eq. [5.1]) as

$$b_S(B_L(U = \mathit{high})) = \theta$$

with $\theta \in [0 \,..\, 1]$ being the subjective probability assigned to the specific object of belief (note the use of lower case b_S instead of B_S to indicate a belief of a belief).

To have a proper probability theoretic model, we define $B_L(U)$ to be a discrete random variable that returns a listener's belief in having a certain categorial level[45] of understanding ($u \in \mathrm{Val}(B_L(U))$). We can then get the probability θ of a specific level

[44] In theory — assuming rational agents — the degree of belief/subjective probability is objectively measurable with a 'betting analysis' (Hájek 2012, § 3.3.2):

> [An agent's] degree of belief in [an object of belief] E is p iff p units of utility is the price at which [it] would buy or sell a bet that pays 1 unit of utility if E, 0 if not E.

[45] Given a discrete probability space (Ω, Pr), a discrete random variable X is a function from the sample space Ω to a set of values $\mathcal{X}$. Its range $\{X(\omega) \mid \omega \in \Omega\}$ is named $\mathcal{X}_X$ (Krengel 2005, p. 42). In general, $\mathcal{X}$ is often defined to be $\mathbb{R}$ or a countable subset thereof. For the current purpose, however, it is most useful to use a finite set of named categorial states (e.g., $\mathrm{Val}(X) = \{\mathit{low}, \mathit{medium}, \mathit{high}\}$). ¶ When needed, an underlying mapping from the elements of this set to real numbers, such as $(\mathit{low}, \mathit{medium}, \mathit{high})^T \mapsto (-1, 0, 1)^T$, can easily be created.

(e.g., $B_L(U = \mathit{high})$) from its probability distribution[46] $\mathrm{Pr}_{B_L(U)}$.

$$\theta = b_S(B_L(U = \mathit{high})) = \mathrm{Pr}_{B_L(U)}(\mathit{high}) \tag{5.2}$$

Above, on page 86, we argued for a graded representation of the listener's mental state, e.g., that she believes her level of understanding to be either *low*, *medium*, or *high*. We take these three levels to be the values of the random variable $B_L(U)$, i.e., $\mathrm{Val}(B_L(U))$. It is important to note that this model of the listener's mental state implies that her belief to be in a specific level of understanding is certain and crisp. Uncertainty about the listener's belief is only modelled on the side of the speaker. The listener's uncertainty, as perceived by the speaker, is modelled in a separate variable (see section 5.4) and is incorporated into the attribution process.

The speaker's belief state — the representation of his degree of belief in all possible worlds (Russell and Norvig 2010, p. 480) — about the listener's mental state of understanding is therefore the marginal distribution $\mathrm{Pr}_{B_L(U)}(\boldsymbol{u})$, with $\boldsymbol{u} = (\mathit{low}, \mathit{medium}, \mathit{high})^T$ and $\sum_{u \in \boldsymbol{u}} \mathrm{Pr}_{B_L(U)}(u) = 1$. A distribution of, e.g., $\mathrm{Pr}_{B_L(U)}(\boldsymbol{u}) = (0.052, 0.418, 0.53)^T$, where almost all probability mass is distributed across the *medium* and *high* levels of understanding, means that the speaker is fairly certain that the listener's level of understanding is either *high* or *medium*, with a low subjective probability that the listener's level of understanding is *low*.

When a coarser view on the speaker's belief state is sufficient, we can look at the skewness[47] of the distribution. If the probability distribution is skewed towards *high*, that is, its skewness is $\gamma_1 < 0$, we can simply say that the speaker beliefs $B_L(U)$ is *high*. Similarly for a distribution skewed towards *low* ($\gamma_1 > 0$).

As stated in fn. 41, a speaker's belief state of the listener's belief in being in the other phenomenal listening-related mental states (contact, perception, acceptance, and agreement) is modelled analogously to understanding. Thus, the speaker's complete

[46] The distribution of a discrete random variable X specifies the probabilities of specific values $x \in \mathcal{X}_X$. It corresponds to the probability measure function $\mathrm{Pr}_X : \mathcal{X}_X \rightarrow [0 \,..\, 1]$, which is defined in terms of the sum of the probability masses $\mathrm{Pr}(\omega)$, $\omega \in \Omega$, that X maps to x (Krengel 2005, p. 42):

$$\mathrm{Pr}_X(x) = \mathrm{Pr}(X = x) = \mathrm{Pr}(\{\omega \in \Omega : X(\omega) = x\}) = \sum_{\{\omega \in \Omega | X(\omega) = x\}} \mathrm{Pr}(\omega).$$

[47] The skewness γ_1 of a discrete random variable X is defined as

$$\gamma_1 = \frac{\boldsymbol{E}_{\mathrm{Pr}}[(X - \boldsymbol{E}_{\mathrm{Pr}}[X])^3]}{\sqrt{\mathit{Var}_{\mathrm{Pr}}[X]}^3},$$

with expectation $\boldsymbol{E}_{\mathrm{Pr}}[X] = \sum_x x \cdot \mathrm{Pr}_X(x)$ and variance $\mathit{Var}_{\mathrm{Pr}}[X] = \boldsymbol{E}_{\mathrm{Pr}}[X^2] - (\boldsymbol{E}_{\mathrm{Pr}}[X])^2$. See fn. 45 concerning numerical values of x (needed for the computation of expectation) when dealing with discrete categorial random variables.

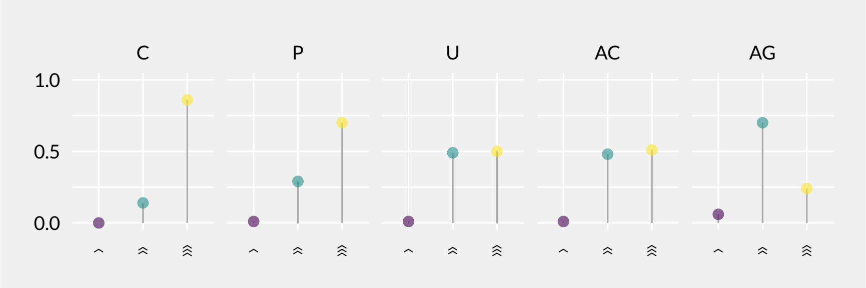


Figure 5.2: Example of the attributed listener state (data from the third worked example presented in Buschmeier and Kopp [2012b, fig. 3]). Each facet shows a comb plot of the belief state of one of the variables $X \in \mathcal{ALS}$, i.e., the distribution of subjective probability $\Pr_{B_L(X)}$ over its values *low* (⌃ / ●), *medium* (⌃⌃ / ●), and *high* (⌃⌃⌃ / ●).

representation of attributed listener state $\mathcal{ALS}$ is modelled as a set of five discrete random variables

$$\mathcal{ALS} = \{C, P, U, AC, AG\} \tag{5.3}$$

representing the graded 'beliefs' of a speaker that a listener beliefs to be in contact ($B_L(C)$), whether a listener is willing and able to perceive ($B_L(P)$), understand ($B_L(U)$), accept ($B_L(AC)$), and agree ($B_L(AG)$). For brevity of notation, we write U instead of $B_L(U)$, P instead of $B_L(P)$, and so forth.

Notably the model implies that speakers simultaneously maintain beliefs about all five of the listener's listening-related mental states at any point in time. Figure 5.2 visualises the distributions over the individual variables of a concrete attributed listener state (taken from Buschmeier and Kopp [2012b]). In this state, the distribution of the speaker's degrees of belief in the listener's understanding are $\Pr_{B_L(U)}(\boldsymbol{u}) = (0.01, 0.49, 0.5)^T$, which means that the speaker is uncertain whether the listener's understanding is *medium* or *high*, but certain that it is not *low*. At the same time, the degrees of belief concerning the listener's mental state of perception — $\Pr_{B_L(P)}(\boldsymbol{p}) = (0.01, 0.29, 0.7)^T$ — clearly peaks on perception being *high*. The other three variables can be interpreted similarly.

This way of modelling a speaker's belief state provides a multi-layered and nuanced view on a listener's listening-related mental states.

The basic communicative functions of feedback are neither independent nor on the same level. They are related to each other, forming a hierarchy with contact at its lowest end, followed by perception, understanding, and ending in acceptance and agreement at the top. From a theoretical perspective, these relationships are governed by two major factors: the property of upward completion (Clark 1996)[48] and the cooperative principle (Grice 1957) — see section 3.4. As illustrated in fig. 3.3 (page 60), positive feedback on a level L_i entails successful processing on lower levels L_{i-1} to $L_\perp$ and implicates a problem in the next higher level L_{i+1}. Conversely, negative feedback on a level L_i entails problems in processing on all higher levels L_{i+1} to $L_\top$ and implicates successful processing on the preceding lower level L_{i-1}.

Based on the argument that feedback functions correspond to underlying mental states, we assume such a relation to be present among the variables of the attributed listener state as well. Accordingly, receiving feedback of *high* understanding should — in addition to influencing ALS-variable U — also have an influence on ALS-variables corresponding to lower levels of processing, namely P and C and on the variables of the next higher level, AC and AG. Assuming strong upward completion, high understanding feedback would result — due to the entailment relationship — in the definite belief of perception and contact being *high* as well, i.e., the ALS-variables P and C would have the degenerate distribution $\mathrm{Pr}_P(\boldsymbol{p}) = \mathrm{Pr}_C(\boldsymbol{c}) = (0,0,1)^T$. The more plausible assumption of weak upward completion, and an uncertain entailment relationship, would result in distributions of degrees of belief of perception and contact that are skewed towards *high*.

As the implicatures of feedback are uncertain per se[49] feedback of high understanding would result in distributions of degrees of belief of acceptance and agreement that are skewed towards *low*.

One way to capture these relationships between the random variables of the ALS would be to define — i.e., specify by hand or learn from data — their joint probability distribution

$$\mathrm{Pr}_{\mathcal{ALS}}(C, P, U, AC, AG),$$

which, in general, is not feasible for more than a few variables.[50] A more compact representation of the relationships between variables makes use of independence

[48] From a cognitive point of view, Clark's (1996) property of upward completion is likely too strong. In section 3.4 we argued for a weaker form of upward completion in which the resulting entailment relationships are merely default implications (Allwood 2000) and therefore uncertain and defeasible.

[49] A speaker cannot be sure that a listener means what is implicated and a listener cannot be sure that a speaker infers what is implicated (Bach 2006, p. 23).

[50] The discrete joint probability distribution for a set $\mathbf{X}$ of random variables consists of $\prod_{X\in\mathbf{X}} |\mathrm{Val}(X)|$ parameters. This number grows exponentially with the number of variables described. ¶ With five variables and three values each $\mathrm{Pr}_{\mathcal{ALS}}$ has $|\mathrm{Val}(C)| \cdot |\mathrm{Val}(P)| \cdot |\mathrm{Val}(U)| \cdot |\mathrm{Val}(AC)| \cdot |\mathrm{Val}(AG)| = 3^5 = 243$

properties, especially conditional independence[51].

We identify such independencies among the ALS-variables on theoretical grounds. According to the property of upward completion — and the entailment relationships that follow from it — a variable representing a level of processing L_i depends on the variables representing its immediate neighbouring levels. Given positive feedback it influences its predecessor L_{i-1}. Given negative feedback it influences its successor L_{i+1}. It also influences its next but one neighbours L_{i-2} and L_{i+2} . This influence, however, is not direct but conditioned on its immediate neighbours. Based on this, we assume that a variable on level L_i is conditionally independent of all other variables given the variables on its immediate neighbour levels L_{i-1} and L_{i+1}:

$$(L_i \perp\!\!\!\perp L_\perp \perp\!\!\!\perp \ldots \perp\!\!\!\perp L_{i-2} \perp\!\!\!\perp L_{i+2} \perp\!\!\!\perp \ldots \perp\!\!\!\perp L_\top \mid L_{i-1}, L_{i+1}). \tag{5.4}$$

This includes that, when multiple variables $L_{i_a}, \ldots, L_{i_k}$ exist on the same level of processing, these are conditionally independent from each other given a variable on the preceding level L_{i-1}[52]

$$(L_{i_j} \perp\!\!\!\perp L_{i_1}, \ldots, L_{i_k} \mid L_{i-1}). \tag{5.5}$$

We assume that the five ALS-variables lie on four levels of processing. C is on the first level, P on the second level, U on the third level, and AC and AG are both on the fourth level. Applying eqs. (5.4) and (5.5) to the ALS-variables yields a set of independence assertions

$$\begin{aligned} \mathcal{I}_{\mathcal{ALS}} = \{ \quad & (C \perp\!\!\!\perp U, AC, AG \mid P), && (5.6a)\\ & (P \perp\!\!\!\perp AC, AG \mid U), && (5.6b)\\ & (U \perp\!\!\!\perp C \mid P), && (5.6c)\\ & (AC \perp\!\!\!\perp C, P, AG \mid U), && (5.6d)\\ & (AG \perp\!\!\!\perp C, P, AC \mid U) \quad \}. && (5.6e) \end{aligned}$$

parameters. Even if defining values for these parameters might still be considered feasible, it becomes virtually impossible once more variables are added to the model (see below, especially section 5.5, and appendix A).

[51] Random variables $X_1, \ldots, X_n$ are 'marginally independent', iff $\Pr(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n \Pr(X_i = x_i)$, for all $(x_1, \ldots, x_n) \in \mathrm{Val}(X_1) \times \ldots \times \mathrm{Val}(X_n)$. We write $(X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n)$. ¶ They are 'conditionally independent' of a random variable Y, iff $\Pr(X_1 = x_1, \ldots, X_n = x_n \mid Y = y) = \prod_{i=1}^n \Pr(X_i = x_i | Y = y)$, for all $(x_1, \ldots, x_n, y) \in \mathrm{Val}(X_1) \times \ldots \times \mathrm{Val}(X_n) \times \mathrm{Val}(Y)$. We write $(X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n \mid Y)$. Note that marginal and conditional independence is a symmetric property, i.e., $(X_1 \perp\!\!\!\perp X_2) \Leftrightarrow (X_2 \perp\!\!\!\perp X_1)$ and $(X_1 \perp\!\!\!\perp X_2 \mid Y) \Leftrightarrow (X_2 \perp\!\!\!\perp X_1 \mid Y)$.

[52] This is the only case relevant here. In general, independence assertions are more complex with multiple variables on a level (see Koller and Friedman 2009, pp. 70–71), especially when there is a subsequent level L_{i+1}, or when multiple variables exist on neighbouring levels, too.

These independence assertions can now be used to find a more compact representation of the joint probability distribution of the attributed listener state model. Using the chain rule of conditional probabilities[53], $\mathrm{Pr}_{\mathcal{ALS}}(C, P, U, AC, AG)$ can be decomposed into a product of conditional probabilities, for example,

$$\begin{aligned}\mathrm{Pr}_{\mathcal{ALS}}(C, P, U, AC, AG) = \mathrm{Pr}(C) \cdot \mathrm{Pr}(P \mid C) \cdot \\ \mathrm{Pr}(U \mid C, P) \cdot \mathrm{Pr}(AC \mid C, P, U) \cdot \mathrm{Pr}(AG \mid C, P, U, AC),\end{aligned} \quad (5.7)$$

which can be further simplified by considering the asserted independencies $\mathcal{I}_{\mathcal{ALS}}$. Applying the independence assertion eq. (5.6c) to the factor $\mathrm{Pr}(U \mid C, P)$ in eq. (5.7) yields $\mathrm{Pr}(U \mid P)$. Similarly, applying eq. (5.6d) to $\mathrm{Pr}(AC \mid C, P, U)$ yields $\mathrm{Pr}(AC \mid U)$, and an application of eq. (5.6e) to $\mathrm{Pr}(AG \mid C, P, U, AC)$ yields $\mathrm{Pr}(AG \mid U)$. The joint probability distribution can thus be expressed as a product of five conditional probability distributions

$$\begin{aligned}\mathrm{Pr}_{\mathcal{ALS}}(C, P, U, AC, AG) = \mathrm{Pr}(C) \cdot \mathrm{Pr}(P \mid C) \cdot \\ \mathrm{Pr}(U \mid P) \cdot \mathrm{Pr}(AC \mid U) \cdot \mathrm{Pr}(AG \mid U),\end{aligned} \quad (5.8)$$

one for each of the random variables in the ALS model. This factorised model is much easier to specify. The probability distribution of each ALS-variable is only conditioned on the variable of its preceding level and it consists of much fewer parameters than the full joint probability distribution[54].

In order to be able to reason computationally with the ALS-model, we can represent it in form of a graphical probabilistic model, a declarative representation which provides data structures and allows for the application of various reasoning algorithms to these data structures. The specific model developed above can be represented as a Bayesian Network[55] (Pearl 1988; Koller and Friedman 2009) with the graph structure shown in fig. 5.3.

[53] Following the 'chain rule' of conditional probabilities, any joint probability distribution $\mathrm{Pr}(X_1, \ldots, X_n)$ can be decomposed into a product of conditional probabilities as follows:

$$\mathrm{Pr}(X_1, \ldots, X_n) = \mathrm{Pr}(X_1) \cdot \mathrm{Pr}(X_2 \mid X_1) \cdot \ldots \cdot \mathrm{Pr}(X_n \mid X_1, \ldots, X_{n-1}).$$

It should be noted that different decompositions — all valid — follow depending on the ordering of the random variables $X_1, \ldots, X_n$.

[54] As joint probability distributions, conditional probability distributions $\mathrm{Pr}(X_1 | X_2 \ldots, X_n)$ for a set $\mathbf{X}$ of random variables consist of $\prod_{X \in \mathbf{X}} |\mathrm{Val}(X)|$ parameters. Since conditional probability distributions are independent, the number of parameters of a factored probability distribution is the sum of parameters of its factors. ¶ Our specific factorisation of the model of the relationships among ALS-variables reduces the number of parameters to $3^1 + 3^2 + 3^2 + 3^2 + 3^2 = 39$ (from 243 for the unfactorised model, see fn. 50).

[55] A Bayesian Network is a directed acyclic graph whose structure encodes the dependencies and in-

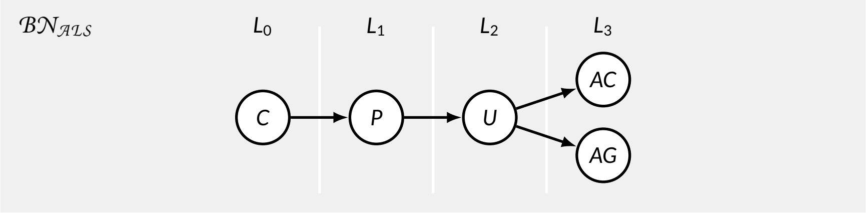


Figure 5.3: Bayesian network representation of the ALS-model ($\mathcal{BN}_{\mathcal{ALS}}$). The network encodes the set of independence assertions $\mathcal{I}_{\mathcal{ALS}}$, eq. (5.6), and the factorised distribution $\Pr_{\mathcal{ALS}}(C, P, U, AC, AG)$, eq. (5.8).

Could the result of the derivation above have been a Bayesian network with a different graph structure, e.g., one where the edges point in the opposite direction, $\mathcal{BN}'_{\mathcal{ALS}} : C \leftarrow P \leftarrow U \leftarrow [AC, AG]$? The set of independence assertions $\mathcal{I}_{\mathcal{ALS}}$, eq. (5.6), is consistent with the structure $C \leftarrow P \leftarrow U$, given a different decomposition of $\Pr_{\mathcal{ALS}}(C, P, U, AC, AG)$ according to the chain rule. AC and AG, however, were connected to U in the 'v-structure' $AC \rightarrow U \leftarrow AG$ (they have U as a common effect), which makes them marginally independent, or conditionally dependent given U (see fn. 52). Because of this v-structure $\mathcal{BN}'_{\mathcal{ALS}}$ lies in a different 'I-equivalence class' (Koller and Friedman 2009, pp. 77–78, def. 3.9, thm. 3.7) than $\mathcal{BN}_{\mathcal{ALS}}$. It is thus not possible, given $\mathcal{I}_{\mathcal{ALS}}$, to factorise $\Pr_{\mathcal{ALS}}(C, P, U, AC, AG)$ in such a way that local probabilistic models for $\Pr(AC)$ and $\Pr(AG)$ result.

The network $\mathcal{BN}''_{\mathcal{ALS}} : C \leftarrow P \leftarrow U \rightarrow [AC, AG]$, however, would be an alternative I-equivalent structure and the question is of course whether this network structure would make a difference for the actual ALS-model. As Bayesian networks with I-equivalent structures can represent the same probability distribution, there is no principal difference between $\mathcal{BN}_{\mathcal{ALS}}$ and $\mathcal{BN}''_{\mathcal{ALS}}$, i.e., the directionality of edges in I-equivalent Bayesian networks does not matter (ibid., p. 77) — at least for their expressiveness. The difference is rather one of consistency — why should the direction of edges differ between L_2 and L_3, and L_3 and L_4? — and of the perspective on the

dependencies of a probability distribution Pr. The nodes of the graph represent the random variables of the distribution. The structure of the graph is such that each node X_i is conditionally independent of its non-descendant nodes Nd_{X_i} given its parent nodes Pa_{X_i}, i.e., $(X_i \perp\!\!\!\perp \mathrm{Nd}_{X_i} \mid \mathrm{Pa}_{X_i})$ (Koller and Friedman 2009, def. 3.1, p. 57). Each node X_i has an associated local probabilistic model $\Pr(X_i \mid \mathrm{Pa}_{X_i})$ which is one of the conditional probability distributions of the factorised distribution.

problem at hand that the model embodies. $\mathcal{BN}_{\mathcal{ALS}}$ takes a 'causal' perspective whereas $\mathcal{BN}''_{\mathcal{ALS}}$ would partially take an 'evidential' perspective (Koller and Friedman 2009, pp. 69–70).

The interpretation of the network's interaction according to the causal view are that the degree of successful processing on higher levels is an effect of the degree of successful processing on lower levels. This view explains upward entailment of negative feedback well (problems in perception cause, i.e., likely result in, problems in understanding). According to the evidential view, on the other hand, the interpretation is that the degree of successful processing on higher levels is evidence for the degree of successful processing on lower levels.

We now analyse an example parametrisation of the attributed listener state network $\mathcal{BN}_{\mathcal{ALS}}$[56] with the following, manually created, local probabilistic models, in form of tabular conditional probability distributions ('conditional probability tables, CPTs')[57]

$$\begin{gathered}\Pr(C) = \begin{bmatrix} ⅓ \\ ⅓ \\ ⅓ \end{bmatrix}, \quad \Pr(P \mid C) = \Pr(U \mid P) = \begin{bmatrix} 0.8 & 0.2 & 0.1 \\ 0.15 & 0.7 & 0.3 \\ 0.05 & 0.1 & 0.6 \end{bmatrix}, \\ \Pr(AC \mid U) = \Pr(AG \mid U) = \begin{bmatrix} 0.8 & 0.2 & ⅓ \\ 0.15 & 0.6 & ⅓ \\ 0.05 & 0.2 & ⅓ \end{bmatrix}. \end{gathered} \tag{5.9}$$

The local probabilistic model for variable C is chosen so that the speaker's belief state of the listener's grade of contact $\Pr(C)$ is uniformly distributed[58]. The model for variable U, $\Pr(U \mid P)$, is chosen so that if P is believed to be *low*, the distribution of U is heavily skewed towards *low*. This is weak upward completion. If P is believed to be *medium*, the distribution of U is centred on *medium* with a slight skew towards *low*. If P is believed to be *high*, the distribution of U is skewed toward *high* but less so than is the case if $P = low$. The motivation for these two choices is that, from a causal perspective, lower grades of success of P have a stronger and more certain effect upward than higher grades of success (whose upward effect is more uncertain). The same model holds for $\Pr(P \mid C)$, but not for $\Pr(AC \mid U)$ and $\Pr(AG \mid U)$ as acceptance and agreement are rather evaluative than processing-related. Given that U is believed to be *low*, the distribution of AC/AG is skewed towards *low* as in the

[56] A machine readable specification of the example model, in 'Bayesian network interchange format' (XBIF, Cozman et al. 1998), is archived and available at DOI: 10.6084/m9.figshare.3827277 .

[57] With $a_{ij} = \Pr(X = \mathrm{Val}(X)_i \mid Y = \mathrm{Val}(Y)_j)$ for each element a_{ij} in the tabular representation of the conditional probability distribution $\Pr(X \mid Y)$. ¶ Given that $\mathrm{Val}(U) = \mathrm{Val}(P) = \{low, medium, high\}$, $\Pr(U \mid P)_{31} = 0.05$ in eq. (5.9) is the conditional probability $\Pr\big(U = high \mid P = low\big)$.

[58] The reason for this choice is that the model is in a stage where no feedback has been presented as evidence to the model yet.

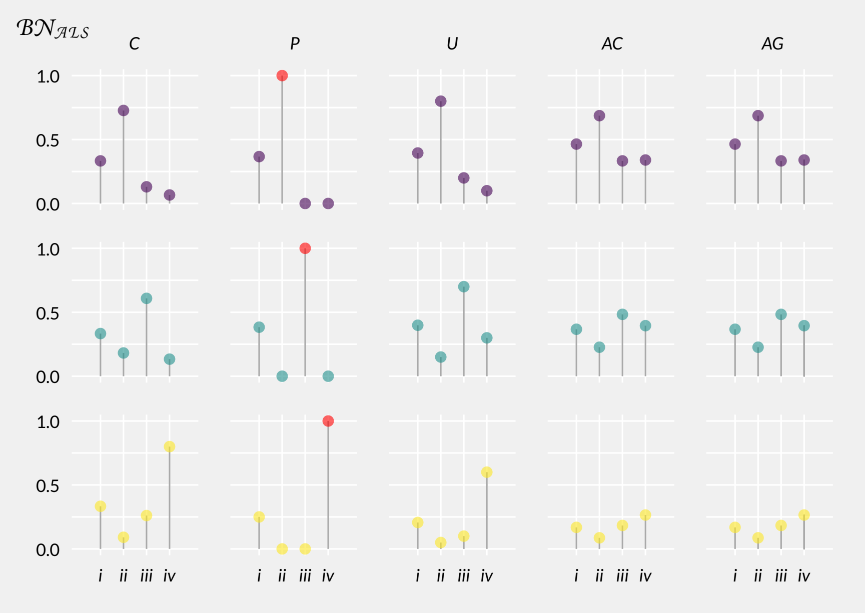


Figure 5.4: Influence of P on the belief states of the other ALS-variables C, U, AC, AG. Given a Bayesian network as in fig. 5.3 with the example conditional probability distributions from eq. (5.9), posterior marginals are computed in four settings: In *(i)* no evidence is specified, in *(ii)*–*(iv)* P is asserted to be *low*, *medium*, or *high*, respectively (marked ●). Facets show the degree of belief in one value — *low* (⌃ / ●), *medium* (⌃⌃ / ●), and *high* (⌃⌃⌃ / ●) — of one ALS-variable.

models above. Acceptance and agreement presuppose understanding and are very likely low given problems in understanding. If U is believed to be *medium*, or *high*, however, it remains more uncertain what AC and AG could be.

Figure 5.4 shows how relations between variables play out in the example model in four settings. Evidence is set and the marginal distributions of all ALS-variables are computed[59]. In setting *(i)*, which serves as a reference, only variable C contains

[59] Bayesian network inference in this thesis is based on the joint-tree algorithm using elimination trees, as described in Darwiche (2009, §§ 7.1–7.6). Inference is done with the PRIMO package for probabilistic

evidence (its initial marginal distribution). Being uniformly distributed, the belief state for C is maximally uncertain. The other variables are skewed towards *low*.

In settings *(ii)*–*(iv)* hard 'evidence' for variable P is specified.[60] An immediate observation is that all variables are correlated with each other. In setting *(ii)*, where P is set to *low*, i.e., $\Pr(P) \leftarrow (1, 0, 0)$, the belief in the other variables being *low* as well is dominant. Similarly in settings *(iii)* and *(iv)*. When P is set to *medium* or *high*, the subjectively most probable beliefs for the other variables are *medium* or *high* as well.[61] $\mathcal{BN}_{\mathcal{ALS}}$ captures weak upward completion based entailment, but overgeneralises into both directions, where, depending on the polarity of perceived feedback, upper-bounding implicata of the feedback signal, which result from the cooperative principle (see section 3.4 and fig. 3.3), should fence off variables from one of the directions.

The relationship among the five ALS-variables is, however, not the right place to model the influence of these potential implicata. The variables represent a speaker's multidimensional attributed mental state and not a specific feedback signal that was perceived and caused the speaker to attribute this mental state in the first place. Once a feedback signal has been received and has informed the attributed mental state, this state might become subject to inferential processes of its own, that is, independent from feedback signals (for example inferences concerning the temporal dynamics of the listeners mental state in the absence of feedback, see section 5.7.3).

Having presented the general ideas and modelling decisions underlying the Bayesian network model of attributed listener state, we will turn to the effect of perceived feedback signals on the five variables in $\mathcal{BN}_{\mathcal{ALS}}$ in the next section.

5.4 FEEDBACK AND ATTRIBUTED LISTENER STATE

Feedback signals carry information about the listener's mental state and are the source of evidence in the attribution process. They constitute the interface between the listener's listening-related mental state and the five variables of the speaker's attributed listener state. This interfacing point is the place where it makes sense to model the upper-bounding implicata of a perceived feedback signal.

inference (available at https://purl.org/scs/PRIMO).

[60] The purpose of this is purely illustrative. The five ALS-variables do not represent directly observable events but are the reconstruction of an interlocutor's hidden state. In the actual model they will not be set to definite values.

[61] The degree values among variables differ though. This is mainly due to the specific local probabilistic models of the variables, but, importantly, also due to their distance to P. The effect of a belief on one level of processing on a variable on a different level of processing gets weaker depending on their distance (in levels).

We design the interface as a random variable *FB* that extends $\mathcal{BN}_{\mathcal{ALS}}$ and represents a perceived feedback signal in terms of its communicative function[62] (contact perception, understanding, acceptance, agreement) and polarity (positive, negative). Each combination of function and polarity is one value *FB* can take[63]:

$$\mathrm{Val}(FB) = \{c^-, c^+, p^-, p^+, u^-, u^+, ac^-, ac^+, ag^-, ag^+\}. \tag{5.10}$$

It is obvious that the five core variables of the attributed listener state are dependent on the type of feedback perceived. This changes the independence structure $\mathcal{I}_{\mathcal{ALS}}$, eq. (5.6), to

$$\mathcal{I}_{\mathcal{ALS}'} = \{ \begin{array}{r} (C \perp\!\!\!\perp U, AC, AG \mid P, FB), \\ (P \perp\!\!\!\perp AC, AG \mid U, FB), \\ (U \perp\!\!\!\perp C \mid P, FB), \\ (AC \perp\!\!\!\perp C, P, AG \mid U, FB), \\ (AG \perp\!\!\!\perp C, P, AC \mid U, FB) \end{array} \}, \tag{5.11}$$

— the only difference is that the variable *FB* is added as a dependency to each local probabilistic model — with the help of which we can derive the factorised probability distribution

$$\begin{aligned} \Pr_{\mathcal{ALS}'}(C, P, U, AC, AG, FB) = \Pr(FB) \cdot \Pr(C \mid FB) \cdot \Pr(P \mid C, FB) \cdot \\ \Pr(U \mid P, FB) \cdot \Pr(AC \mid U, FB) \cdot \Pr(AG \mid U, FB), \end{aligned} \tag{5.12}$$

which translates to the Bayesian network $\mathcal{BN}_{\mathcal{ALS}'}$ with the graph structure shown in fig. 5.5.

The local probabilistic models for the variables, e.g., the model for $\Pr(U \mid P, FB)$ for the variable U, now model both the interaction with the variable P on the preceding level of processing as well as the interaction with the perceived feedback signal.

[62] This modelling decision assumes that raw feedback signals can be reliably classified according to their communicative function even before they enter the attribution process, ideally based only on acoustic and lexical properties (and comparable properties for non-verbal feedback signals) that are objectively measurable. This, however, is currently not possible. Such classification/annotation tasks are quite challenging to do reliably even for human observers, even if trained, and even if communicative context is available to them (Geertzen et al. 2008; Malisz et al. 2016). In a significant effort Neiberg et al. (2013) could identify prosodic features in productive feedback expressions that correlate with some communicative functions of feedback (in Swedish). It is, however, unclear whether their research, based on careful manual analysis, can be easily implemented as a pattern recognition system.

[63] Exactly one of these values is presented as evidence to the model at a time. The hierarchical relations among these values are specified in the local probabilistic models.

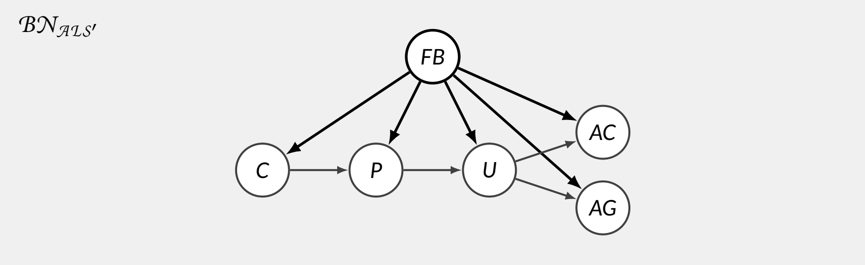


Figure 5.5: Bayesian network representation of the ALS-model with feedback influence ($\mathcal{BN}_{\mathcal{ALS}'}$). This model extends $\mathcal{BN}_{\mathcal{ALS}}$ (fig. 5.3, on page 93) with a random variable *FB* that represents the perceived feedback signal in terms of its function and polarity. The network encodes the set of independence assertions $\mathcal{I}_{\mathcal{ALS}'}$, eq. (5.11), and the factorised distribution $\mathrm{Pr}_{\mathcal{ALS}'}(C, P, U, AC, AG, FB)$, eq. (5.12).

This makes specifying the local probabilistic models of this Bayesian network more difficult than specifying the models of the core ALS network $\mathcal{BN}_{\mathcal{ALS}}$ (as in, e.g., eq. [5.9]). As the graph has a higher density and the added variable *FB* is ten-valued, the resulting factorised conditional probability distribution consists of 400 (instead of 39, see fn. 54) parameters, a more than ten-fold increase.

Unfortunately, the models of $\mathcal{BN}_{\mathcal{ALS}}$ cannot be easily reused and extended for $\mathcal{BN}_{\mathcal{ALS}'}$. For example, each entry in the conditional probability table $\mathrm{Pr}(U \mid P)$ — that is, each individual probability $\mathrm{Pr}(U = \mathrm{Val}(P)_i \mid P = \mathrm{Val}(P)_j)$ — is subdivided into ten probabilities in $\mathrm{Pr}(U \mid P, FB)$, one for each element in $\mathrm{Val}(FB)$. The way they are subdivided differs depending on the assignment of values to U and P. This makes the conditional probability tables-based representations non-modular so that all parameters need to be specified anew. Due to the large number of parameters, it is difficult not to lose track of the big picture and to make consistent choices such that they are replicable and explainable — that is, not arbitrary.

We approach this problem by specifying 'implicit representations' (Koller and Friedman 2009, p. 158) of the local probabilistic models from which the tabular conditional probability distributions can be generated algorithmically[64]. This allows for a

[64] Appendix A, section A.2, describes the approach, representation, and algorithm in detail. Section A.3 contains an example of the implicit representation and generation of the local probabilistic model and conditional probability table $\mathrm{Pr}(U \mid P, FB)$.

Table 5.1: One-dimensional influence vectors $\delta_X(FB_i)$ for each Val(FB) and each local probabilistic model in $\mathcal{BN}_{\mathcal{ALS}'}$, eq. (5.12). The assignment of $\delta_P(FB_{u^-}) = 0.7$ for the local probabilistic model $\Pr(P \mid C, FB)$, for example, means that receiving negative feedback of understanding has a positive influence of strength 0.7 ($\delta \in [-1\,..\,1]$) on the variable P. That is, negative feedback of understanding has the effect that perception is believed to be positive (or skewed towards *high*).

Level		L_1		L_2		L_3		L_4			
	Val(FB) =	c^-	c^+	p^-	p^+	u^-	u^+	ac^-	ac^+	ag^-	ag^+
L_1	$\delta_C(FB_i)$	−1	.9	.8	.95	.8	.95	.8	.95	.8	.95
L_2	$\delta_P(FB_i)$	−.9	−.5	−1	.7	.7	.8	.7	.8	.7	.8
L_3	$\delta_U(FB_i)$	−1	−.8	−.9	−.8	−1	.7	.7	.8	.8	.8
L_4	$\delta_{AC}(FB_i)$	−.9	−.6	−.8	−.9	−.1	−.9	−1	1	−.5	.5
	$\delta_{AG}(FB_i)$	−.9	−.9	−.9	−.8	−.9	−.3	−.5	.3	−1	1

specification of the models with much fewer parameters — and in a modular way.

Table 5.1 shows part of the implicit representation $\mathcal{BN}_{\mathcal{ALS}'}$, namely the specification of the values $\delta_X(FB_i)$ which model the influence that the variable *FB* exerts on the ALS-variables $X \in \Pr_{\mathcal{ALS}'}$.[65] The values in the table are specified such that a variable X on the level of processing L_j is positively influenced by feedback signals of a communicative function that is of either positive polarity and on the same level L_j, or of any polarity and on a higher levels L_k (with $k > j$). X is influenced negatively when the perceived feedback signal is on a lower level L_i (with $i < j$).

As an example, U is influenced positively if feedback of positive understanding or any feedback on higher levels (e.g., positive acceptance, negative agreement) is received and negatively if feedback of negative understanding or lower (e.g., positive perception, negative contact) is received.

Table 5.1 can also be interpreted column-wise. Feedback of a communicative function on a level L_j influences all variables on lower levels of processing L_i (with $i < j$) positively and all variables on higher levels of processing L_k (with $k > j$) negatively. If polarity of the feedback signal is positive, variables on the level L_i are influenced positively, and negatively if it is of negative polarity. As an example, positive understanding feedback influences the variables C, P and U positively and AC and

[65] See section A.2, step G1.2, in the appendix for an explanation of the δ-values.

AG negatively. Negative feedback of understanding influences only the variables C and P positively and U, AC, and AG negatively.

The parametrisation of the local probabilistic models of $\mathcal{BN}_{\mathcal{ALS}'}$, generated from the specification of the implicit representation (partly displayed in table 5.1[66]), thus models both the upper-bounding implicata resulting from Grice's (1975) cooperative principles, as well as the weak upward completion based entailment.

The entailment relationships are now modelled twice in the local probabilistic models of $\mathcal{BN}_{\mathcal{ALS}'}$. Once in the connections $FB \rightarrow (C, P, U, AC, AG)$, and once in the connections $C \rightarrow P \rightarrow U \rightarrow (AC, AG)$ among the ALS-variables. This can be seen as an unnecessary redundancy in the model $\mathcal{BN}_{\mathcal{ALS}'}$ and it could be argued that the connections among the ALS-variables could be dropped. When dialogue context is part of the attribution process (see section 5.5), however, these redundancies are important.

Figure 5.6 shows a comparison of the posterior marginal subjective probability distributions over the ALS-variables (C is omitted for clarity of presentation) given different values for the variable FB. As can be seen, the intended quality of relationships among the variables of the network is reflected in its output.

The representation of feedback as a simple combination of communicative function and polarity, as in the variable FB above, does not meet the characterisation of feedback signals as a multimodal, multidimensional, rich, and nuanced way of expressing a listening-related mental state that we provided in chapter 3. And if it were, a Bayesian network based model of feedback interpretation as in $\mathcal{BN}_{\mathcal{ALS}'}$ would be overly complex to model the phenomenon at hand. But even under the assumption that much of the nuances are used to mark the communicative function and polarity, other properties of feedback signals can be fed into the model as well, using the same modular approach of implicit representation.

As an example we pick the certainty/uncertainty that a listener might express, perhaps unwillingly, in her feedback signals — in their timing, their prosody (Pon-Barry 2008; Skantze et al. 2014), via her facial expression (Krahmer and Swerts 2005), head movement (Heylen 2006), and gaze behaviour (Skantze et al. 2014). The certainty/uncertainty can be seen as a modulating factor for the feedback signal. If a listener expresses positive feedback of understanding, and simultaneously indicates uncertainty, the attribution process should come to a different conclusion than if the feedback signal was produced with certainty. As mentioned in section 5.3 (page 88),

[66] The model $\mathcal{BN}_{\mathcal{ALS}'}$ (and code to generate and query it) is archived and available at DOI: 10.6084/m9.figshare.3851475 .

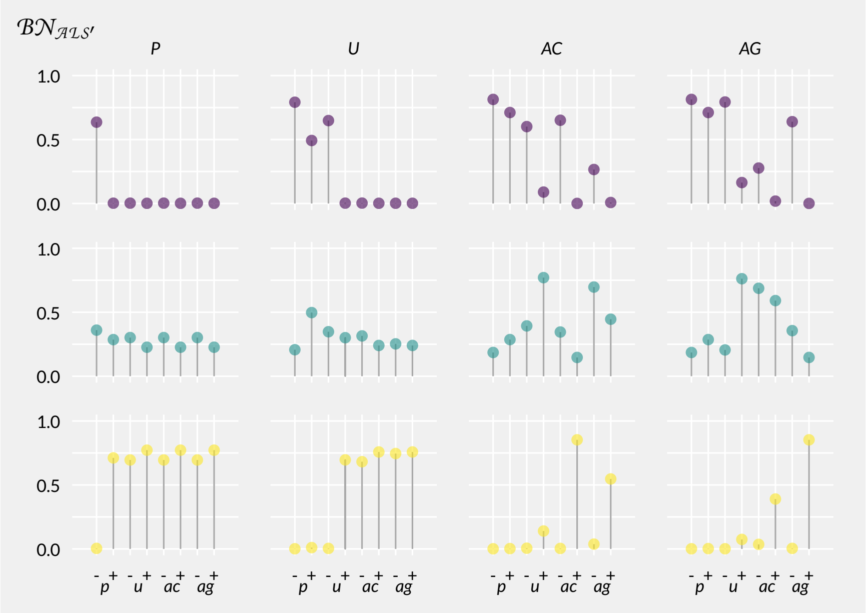


Figure 5.6: Influence of feedback (FB) on the belief states of the other ALS-variables P, U, AC, and AG (C is omitted for clarity of presentation). Posterior marginal distributions are computed for each value in $\mathrm{Val}(FB)$ (see eq. [5.10]). Facets show the degree of belief in one value — *low* (⌃ / ●), *medium* (⌃⌃ / ●), and *high* (⌃⌃⌃ / ●) — of one ALS-variable.

the uncertainty in the five ALS-variables represents the speaker's uncertainty in the attribution of the listener's listening related mental state, not a potential uncertainty on the side of the listener.

The suggestion thus is to model a listener's expressed uncertainty in a Bayesian network $\mathcal{BN}_{\mathcal{ALS}''}$ which has an additional random variable UC with a range $\mathrm{Val}(UC) = \{low, medium, high\}$[67] that influences the ALS-variables P, U, AC and AG (see fig. 5.7). The influence of the perceived uncertainty is difficult to model with the implicit representation based generation of local probabilistic models. The influence of uncertainty

[67] Ternary grading was chosen, again, for practical reasons and implies no claims of cognitive reality.

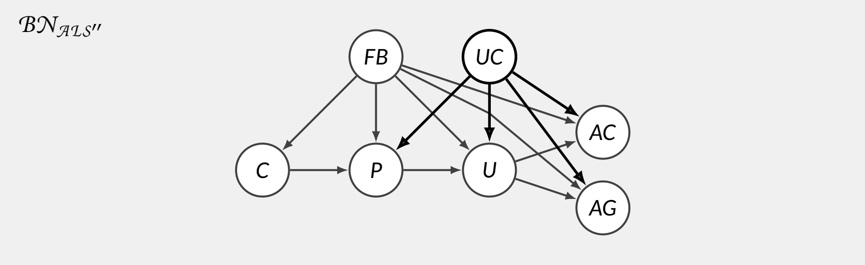


Figure 5.7: Bayesian network representation of the ALS-model $\mathcal{BN}_{\mathcal{ALS}''}$. This model extends $\mathcal{BN}_{\mathcal{ALS}'}$ (fig. 5.5, on page 98) with a random variable *UC* that represents the listener's perceived certainty/uncertainty.

interacts with the polarity of the perceived feedback signal. Feedback of positive polarity with a *high* uncertainty should influence the ALS-variables negatively. It should shift probability mass of the ALS-variables from *high* towards *medium* and *low* (i.e., make the shape of the probability distribution flatter). Feedback of positive polarity with *low* uncertainty should have a positive influence. It shifts further probability mass towards *high* (i.e., skew the probability distribution towards high). For feedback of negative polarity, however, the influence is in the opposite direction. In this case, *high* uncertainty should shift probability mass from *low* towards *medium* and *high*, and *low* uncertainty should shift further probability mass towards *low*.

Due to the modularity of our approach to implicit representation, such interactions between values of variables cannot be directly modelled. Processing the uncertainty accompanying feedback signals of positive polarity would need a setting of, e.g., $\boldsymbol{\delta}_X(UC) = (1.0, 0.0, -1.0)$, whereas for feedback signals of negative polarity a setting of, e.g., $\boldsymbol{\delta}_X(UC) = (-1.0, 0.0, 1.0)$ would be needed. There are several ways this could be addressed, for example by using two uncertainty variables, or by using a six-valued uncertainty variable. In our model, we have chosen a pragmatic solution to this problem. When processing feedback of positive polarity, we let the variable *UC* represent the listener's expressed uncertainty. When processing feedback of negative polarity, we simply inverse the variable's meaning (we write UC^{-1}) and let it represent the listener's expressed certainty. See fig. 5.8 for a comparison of the influences of UC^{-1}/UC on the

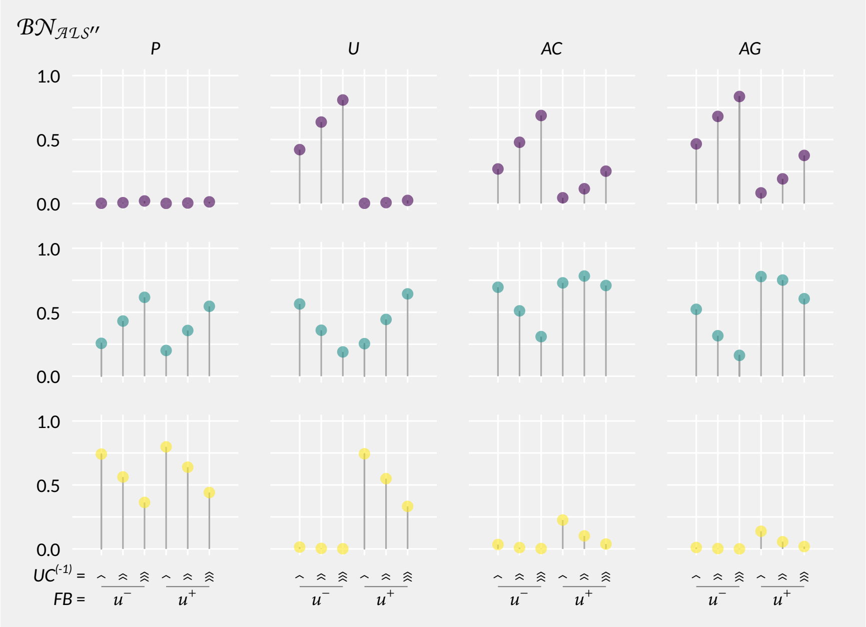


Figure 5.8: Comparison of the influence of the perceived certainty/uncertainty of a feedback signal on the belief states of the ALS-variables P, U, AC, and AG in $\mathcal{BN}_{\mathcal{ALS}''}$ given either negative or positive feedback of understanding (u^-/u^+). The polarity determines whether the variable UC represents the listener's certainty (UC^{-1}, if it is negative) or uncertainty (if it is positive) — see text. Facets show the degree of belief in one value — *low* (⌃ / ●), *medium* (≈ / ●), and *high* (≋ / ●) — of one ALS-variable.

ALS-variables given negative/positive feedback of understanding[68].

Other properties of a listener's feedback behaviour can be integrated into the Bayesian network model for feedback processing in similar ways. The network used in the evaluation study (chapter 9), for example, integrated one variable that represents the listener's gaze behaviour and one that represents the listener's head gestures.

[68] The model $\mathcal{BN}_{\mathcal{ALS}''}$ (and code to generate and query it) is archived and available at DOI: 10.6084/m9.figshare.3971712 .

5.5 INTERPRETING LISTENER FEEDBACK IN CONTEXT

Properties of the perceived feedback signals of a listener are not the only influencing factors on the attributed listener state though, consider the following example:

Meeting, as usual, in a café to discuss the latest development in the coffee world, Sigrid and Linus eventually come to the same old themes, exchanging the arguments both agree with and have already heard numerous times. Sigrid does not even have to unfold an argument in full because Linus already knows what to expect. Listening to Sigrid, Linus nods from time to time but does not produce specific feedback of the communicative function understanding or show his acceptance, and agreement. Given the dialogue context — the usual setting, expectable utterances, a lot of common ground between the interlocutors — Sigrid should nevertheless attribute high understanding, acceptance and agreement to Levi. Would Linus show the exact same behaviour in a different situation, however — e.g., when Sigrid presents a novel and complex argument to him — she should come to different conclusions. That he does not pay attention, that he has problems understanding her argument, or that he does not accept one of its premises.

The same perceived feedback behaviour can originate from different listening related mental states and may thus result in different belief states of the attributed listener state variables. In this example, the factors which determine how the listener's feedback behaviour influences the attributed listener state lie in the wider discourse context: expectable utterances with a low novelty factor due to previous interactions. But contextual factors can also be tied to the speaker's utterance or the immediate situation. Does the speaker explain something complicated, does she use a concept for the first time, does she choose a lexical item or expression with a low frequency, does she use a word that can be misperceived for another word that is plausible in this situation as well, is the environment of the interaction noisy, does the speaker speak loud enough, and so on. This illustrates that the contextual factors that can play a role in feedback interpretation are unique to each dialogue situation and that their number is potentially infinite.

Many contextual factors can only be factored in through deep inferences that often rely on knowledge. They cannot be uncovered with simple, general models such as the one described here. But due to their importance for the interpretation of listener feedback as described in the example above, contextual factors should not be neglected completely during feedback interpretation. We argue that it is possible and useful to model some generally applicable contextual influences in our computational model. Especially speaking related contextual factors are often available in artificial conversational agents, which makes their integration into the model relatively easy.

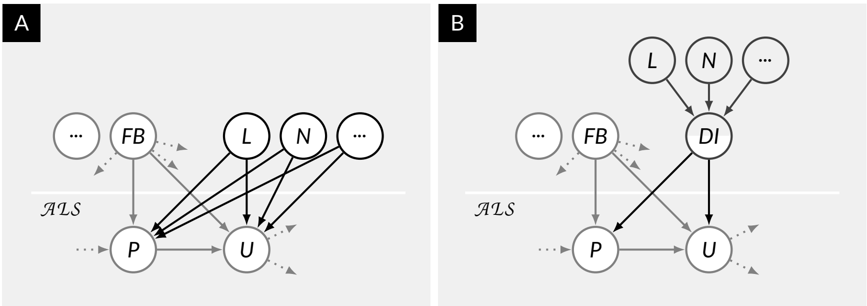


Figure 5.9: Grouping related variables representing contextual factors or feedback properties in a hierarchical Bayesian network. *(A)* variables $L, N, \ldots$ (length and novelty of the speaker's utterance, further related variables) each influence the ALS-variables P and U. *(B)* variables $L, N, \ldots$ form a separate model of utterance difficulty, hidden behind/in the abstract variables DI, which influences the ALS-variables P and U. The complexity of the model is invisible to the ALS-variables, greatly simplifying their local probabilistic models.

Contextual factors can be modelled in the same way as the properties of feedback described in the preceding section. They, and contextual factors in general, are added as a new variable to the network and the parameters of the influence they exert on the ALS-variables are specified in the implicit representation of the local probabilistic model of the ALS-variables.

One contextual factor — modelled as part of the attentive speakger agent described in chapters 8 and 9 — could be the estimated difficulty of the produced utterances. If the system produces an utterance of high difficulty, the feedback its interlocutor produces in response to this utterance has a smaller influence on the subjective probability of the beliefs that understanding is *high*, conversely for utterances of low difficulty. The estimated difficulty of a speaker's utterance is integrated into the model by extending it with a random variable DI that, again, has a ternary grading (i.e., $\mathrm{Val}(DI) = \{\mathit{low}, \mathit{medium}, \mathit{high}\}$) and influences P and U with $\delta_{P/U}(DI) = (1.0, 0.0, -1.0)$.

Contextual factors are often complex concepts. What makes an utterance 'difficult' to process is a research question on its own, but, as stated above, can certainly be reduced to a number of individual components such as the difficulty of the words it is comprised of, its syntactic complexity, its length, the concepts that it describes, its novelty, how much is left implicit, how many presuppositions it assumes, and many things more. Each of these components could be modelled as an individual contextual influence — i.e., a variable in the mental state attribution network — on the ALS-variables (see fig. 5.9*A*). This, however, would increase the complexity and size of the ALS-variables' local probabilistic models substantially.

Alternatively, we can create a separate Bayesian network that models the complex contextual factor in terms of its components and let it interface, via a single node (its 'anchor') with the ALS-variables. The resulting network is a hierarchical Bayesian network (Gyftodimos and Flach 2002) that consists of atomic types (Bayesian network nodes) and composite types (sub-networks that are connected to the overall network through their anchor node).[69] Figure 5.9*B* shows a hierarchical Bayesian network model of the influence of the estimated difficulty of the speaker's utterance on the ALS-variables. The variable *DI* is the anchor node of the sub network that models the complex concept difficulty. It is shared between the networks and shields off — via conditional independence — the ALS-variables from the components of the difficulty model.

Modelling the influence of contextual factors in such a hierarchical way has the advantage of making the overall mental state attribution model simpler (even though an additional variable is introduced) and thus easier to understand. The hierarchical model is also more modular as sub-networks can be expanded or changed without affecting the part of the network it interfaces with (ibid., p. 25). A disadvantage can be that the hierarchical representation may result in influences being less precise since the individual influence of each component might be slightly different from the combined influence of the anchor variable.

All evidence variables in the mental state attribution model can be seen as potential composite types. The variable *FB*, representing the influence of the feedback function, could for example be replaced by a Bayesian network model that estimates the feedback function from low-level features of the feedback signal. The influence that *FB* has on the ALS-variables could remain as it is.

[69] A similar concept are 'object-oriented' Bayesian networks (Koller and Pfeffer 1997; Koller and Friedman 2009, § 5.6), in which sub-networks — that are allowed to have several input and output nodes — define 'encapsulated' conditional probability distributions.

5.6 FEEDBACK AND GROUNDING

After modelling various influences on the ALS-variables, we will now turn to the question of how the inferred beliefs represented in these variables can be used. Given that one important role of communicative listener feedback is to provide 'evidence of understanding' (see section 2.2.1), a relevant question is how the ALS-model is related to grounding and common ground.

Here, we sketch a computational model of grounding in which the speaker's beliefs about her listener's listening-related mental state, as captured in the ALS-variables, is the basis for deciding whether her representation of common ground should be updated. The fundamental idea for this model is that if the evidence of understanding (together with evidence of contact, perception, acceptance, and agreement) provided by the listener causes the speaker to infer a belief state of the ALS-variables that is 'sufficiently' high (Clark and Schaefer 1989), she can regard the communicative acts (and their content) in response to which the listener provided the feedback as grounded.

A way to model this would be the Bayesian network $\mathcal{BN}_{\mathcal{ALS}'''}$ (fig. 5.10), which, in addition to $\mathcal{BN}_{\mathcal{ALS}'}$ (fig. 5.5), has an additional random variable *GR* with a range $\mathrm{Val}(GR) = \{low, medium, high\}$ that is influenced by each of the five ALS-variables C, P, U, AC and AG).[70] *GR* represents the speaker's belief state of the degree of groundedness that, based on the listener's attributed mental state, can be assumed for a communicative act.

Each of the ALS-variables influences the groundedness variable to a different degree. Believing that the listener is in full contact but neither perceives nor understands what the speaker is saying, for example, should lead to a low degree of belief in the groundedness of the object. In contrast, assuming the listener to have at least some understanding might be enough to consider information to be grounded. Figure 5.11 shows how the groundedness variable *GR* varies in a simple model of *GR* given feedback of different functions.

In contrast to the computational models of grounding presented in section 2.2.2 — in which the grounding process advances in a specific way when certain grounding acts are encountered (e.g., Traum 1994; Roque and Traum 2008; Visser et al. 2014) — ALS-based grounding is more flexible.

First of all, instead of a set of specific grounding acts, the ALS-based grounding model can deal with a large variety of feedback-based grounding acts, which may be multimodal and can express subtle differences in meaning and function. This

[70] The model $\mathcal{BN}_{\mathcal{ALS}'''}$ (and code to generate and query it) is archived and available at DOI: 10.6084/m9.figshare.4980743 .

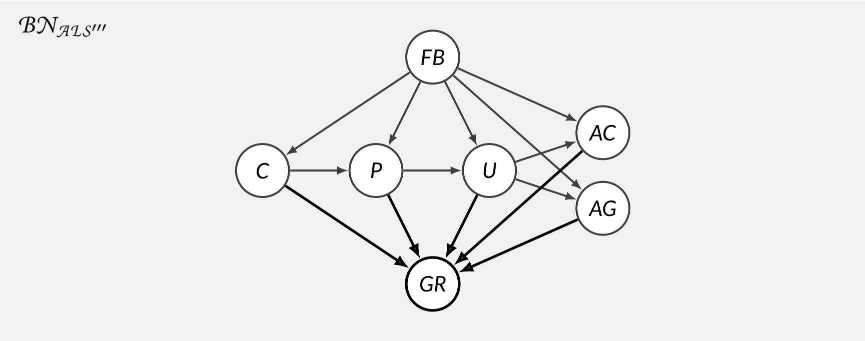


Figure 5.10: Bayesian network representation of the ALS-model $\mathcal{BN}_{\mathcal{ALS}'''}$. This model extends $\mathcal{BN}_{\mathcal{ALS}'}$ (fig. 5.5, on page 98) with a random variable *GR* that represents the speaker's belief state of the degree of groundedness that, based on the listener's attributed mental state, is assumed for a communicative act.

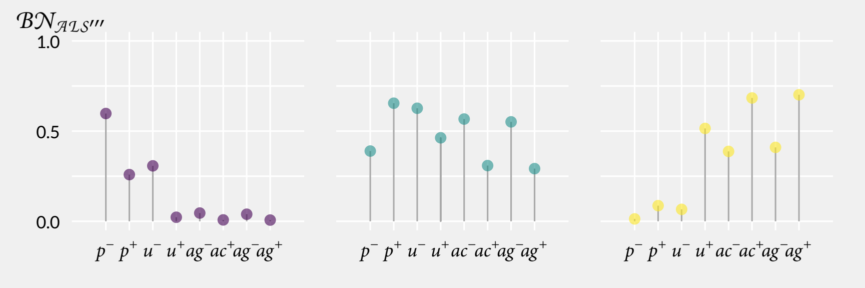


Figure 5.11: Influence of feedback (*FB*) on the belief state of variable *GR*. Posterior marginal distributions are computed for each value in Val(*FB*). Facets show the degree of belief in one value — *low* (⌃ / ●), *medium* (⌃⌃ / ●), and *high* (⌃⌃⌃ / ●) of the variable *GR*.

simplifies the interface between listeners' grounding acts and the operationalisation of groundedness. At the same time, qualifying information of feedback (Petukhova and Bunt 2010), is not lost during inference, but still reflected in the speaker's belief state of the ALS-variables, as well as in the variable *GR*.

Similar to the model of Roque and Traum (2008), the ALS-based model of grounding captures the grounding status with a finite set of 'degrees of groundedness' (here: *low*, *medium*, *high*), but represents them probabilistically in terms of degrees of belief. This adds a further dimension to groundedness, enables the model to deal with uncertainty and removes the need to prematurely commit to a specific degree of grounding.

Another advantage is that a flexible 'grounding criterion' (Clark and Schaefer 1989) is easily operationalised in the ALS-model of grounding because inference about the belief state of groundedness is conditionally independent from the many possible influences. This is a property that reduces the grounding model's complexity significantly. Depending on contextual factors, e.g., the perceived difficulty of the speaker's utterance, the content of a speaker's utterance may be sufficiently grounded even if feedback is absent. In a different context though, even positive feedback of understanding may not be sufficient to consider the content of the utterance to be grounded. The operationalisation of the grounding criterion is context-aware simply because mental state attribution in the ALS model is context-aware. In this regard, it goes beyond Paek and Horvitz's (2000b) decision theoretic formalisation of the grounding criterion.

As the flexibility of the grounding criterion is part of the ALS-based model itself, the decision of whether a belief state is regarded 'sufficient for current purposes' can be modelled as a threshold that should be relatively stable (e.g., within interactions). If the threshold is exceeded, the information can be added to the speaker's representation of common ground. Otherwise, the grounding process needs to continue, e.g., by producing an adapted utterance that provides additional information (see chapter 6).

5.7 INTERPRETING LISTENER FEEDBACK IN AN EVOLVING CONTEXT

The Bayesian network model of attributed listener state developed so far models processing of individual — possibly multimodal — feedback signals in their immediate dialogue context. Up until now, it has no notion of the past and thus disregards the temporal nature of dialogue. Dialogue context, however, is constructed incrementally while the discourse unfolds over time (Poesio and Traum 1997; Asher and Lascarides

2003; DeVault 2008; Ginzburg 2012) and utterances are produced against the common ground established so far (Clark 1996). It is obvious that dialogue participants' listening-related mental states also develop alongside the dialogue. A dialogue partner that believes that her understanding is low may experience, after an additional explanation, a change of her mental state of understanding after which she believes that her understanding is high. She might communicate this development of her listening-related mental state in a sequence of feedback signals — or even by producing a feedback signal that explicitly marks such a 'change-of-state', e.g., *oh* (Heritage 1984) and, in German, also *ach* and *achso* (Golato and Betz 2008; Golato 2012).

One crucial question from the speaker's perspective is how listener feedback signals can be interpreted in the dialogue context, and how they relate to what has been or is being said. Listeners can, in principle, produce feedback signals at any point in time in a dialogue — without having to take the turn. There is also no restriction on the number of feedback signals that can be placed within a dialogue segment, whether it is a turn, an utterance, a pause or a combination of these. Consider the following dialogue fragment:

(5.13) KDS-1, U01 (9:46–9:58)[71,72]

```
(a)  S1:  genau=
(b)       =allerdings ist Badminton da wieder verschoben=
(c)       =[weiß nicht] ob das jetzt dauerhaft ist (.)
     U1:   [mhm       ]
```

[71] Excerpt from the calendar assistant domain corpus KDS-1 (Buschmeier and Yaghoubzadeh 2011). Overlapping talk is marked with aligned square brackets. The transcription follows the 'GAT 2' system (Selting et al. 2011).

[72] English translation of example 5.13:

```
(a)  S1:  precisely=
(b)       =but badminton has been moved again=
(c)       =[don't know] if that's now permanently so (.)
     U1:   [mhm       ]
(d)  S1:  [but for the two] weeks=
     U1:  [okay           ]
(e)  S1:  =it's what I have in
     U1:                      yeah
(f)  S1:  that it's again=
(g)       =um from 8 to 10 [pm]
     U1:                   [ok]ay (0.34)
(h)       yes,=
(i)       =then um I'll go there nevertheless (.) . . .
```

```
(d)  S1:  [aber die zwei] Wochen=
     U1:  [okay         ]
(e)  S1:  =hab ich's jetzt so drin
     U1:                          ja
(f)  S1:  das is wieder von=
(g)       =ehm acht bis zweiundzwanzig U[hr]
     U1:                                [ok]ay (0.34)
(h)       ja,=
(i)       =dann ehm geh ich da trotzdem hin (.) . . .
```

Speaker S1 explains to her interlocutor U1 that the regular badminton training has (again) been moved to a different time, and now takes place from 8 to 10 p.m. She also says that she does not know whether this change is permanent, but that it is scheduled like this for the next two weeks. During S1's nine seconds short turn (5.13*a*) to (5.13*g*), U1 provides four instances of communicative feedback. Firstly, she signals understanding with *mhm*, simultaneously producing a single head nod and looking at S1 (5.13*c*). After that, she signals acceptance of the speaker's ignorance concerning the permanency of the time change with an *okay* that is accompanied by a head nod (5.13*d*). Thirdly, she signals understanding, producing a short and prosodically flat *ja*, 'yeah', (5.13*e*). And finally, with S1 gazing at her, she signals understanding of the new time with an *okay* and a head nod (5.13*g*). After a pause, U1 then takes the turn and continues.

When multiple feedback instances occur in sequence, as in the dialogue fragment in example (5.13), the question arises how their interpretations affect each other, and how they relate to what has been and is being said.

In the series of Bayesian network models for reasoning about the listener's listening-related mental state developed above ($\mathcal{BN}_{\mathcal{ALS}}$ to $\mathcal{BN}_{\mathcal{ALS}'''}$), a concept for the temporal dynamics — which would make the evolution of the ALS coherent and continuous, and enable the models to deal with sequences of feedback, such as in the dialogue fragment in example (5.13) — is absent. In the following, we extend these 'atemporal models' with a temporal dimension that accounts for the incremental and dynamic nature of dialogue. Such a 'dynamic model' of mentalising can naturally deal with multiple instances of feedback by updating its representation — taking the immediate dialogue history into account as well — when the dialogue proceeds and feedback occurs.

5.7.1 DYNAMIC MINIMAL MENTALISING

We regard an unfolding dialogue as a sequence of segments $[s_0, s_1, \dots s_t, \dots s_N]$, each consisting of an 'utterance unit' of the speaker (Traum and Heeman 1997; Poesio and Traum 1997, p. 317), together with any feedback responses of the listener. The atemporal models treat each of these segments s_t independently and thus only reason about the listener's listening-related mental state during one single segment. When attributing the listener state for the next segment, information from the preceding segments is not taken into account at all. To overcome this limitation, i.e., to account for the evolution of the listener's mental state over time, we need to give the model of the listener a temporal dimension.

As Bayesian networks are, in general, not limited in the number of edges and nodes, it would be possible to capture a whole dialogue — or at least a self contained and coherent fragment of a dialogue — in one large network that consists of connected sub-networks $\mathcal{BN}_{\mathcal{ALS}_t}$ — each corresponding to one network such as $\mathcal{BN}_{\mathcal{ALS}'''}$ in fig. 5.11 — one for each 'segment' s_t. The variables in the sub-networks would be uniquely named and the network's evidence variables would be instantiated from the listener's feedback behaviour as well as the dialogue context of segment s_t. Furthermore, the variables between the sub-networks could be arbitrarily connected to model any desirable interaction between feedback and context across segments.

Theoretically, this approach could even work in an incremental framework. With each new dialogue segment s_{t+1}, a new sub-network $\mathcal{BN}_{\mathcal{ALS}_{t+1}}$ would be added and connected to the network and Bayesian network inference would be carried out. However, even though there is, in principle, no limit in the size of a Bayesian network, the computational costs are rising polynomially with the number of nodes, and may even become intractable if the nodes are unfavourably connected (Koller and Friedman 2009, § 9.7). This makes this 'growing network approach' unsuitable for practical applications.

A slightly more constrained approach is to make a first-order Markov assumption, i.e., to assume that variables X_{t+1} of a sub-network $\mathcal{BN}_{\mathcal{ALS}_{t+1}}$ are only dependent on variables X_t of the sub-network $\mathcal{BN}_{\mathcal{ALS}_t}$ that directly precedes it. This can be achieved efficiently in the framework of 'dynamic Bayesian networks'. In contrast to a constantly growing network approach, the flavour of the dynamic Bayesian network approach we choose here consists of a maximum of two sub-networks ('time-slices') at any point of time. In such a 'two time-slice Bayesian network' (ibid., defs. 6.3,6.4), one time slice $\mathcal{BN}_{\mathcal{ALS}_t}$ represents the current dialogue segment s_t, the other time slice $\mathcal{BN}_{\mathcal{ALS}_{t+1}}$ the next segment s_{t+1}. As in the growing network approach, temporal influences among dialogue units are modelled by connecting some of the variables

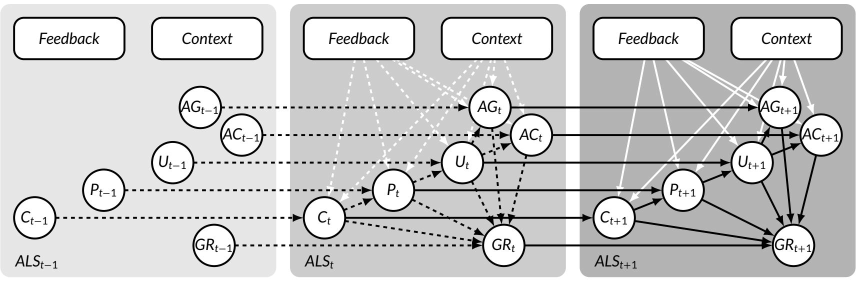


Figure 5.12: Illustration of a dynamic two time-slice Bayesian network model unrolling over three steps in time, each corresponding to one dialogue segment. Dashed arrows are disregarded during inference in subsequent time-slices, i.e., variables from time slice $\mathcal{BN}_{\mathcal{ALS}_{t-1}}$ and evidence variables in time slice $\mathcal{BN}_{\mathcal{ALS}_t}$ have no influence on variables in time slice $\mathcal{BN}_{\mathcal{ALS}_{t+1}}$. The posterior distributions of attributed listener state variables (C, P, U, AC, AG) as well as the groundedness variable GR in time slice $\mathcal{BN}_{\mathcal{ALS}_t}$ are taken as prior distributions at time $t+1$ and influence the variables they are connected to in time slice $\mathcal{BN}_{\mathcal{ALS}_{t+1}}$. For reasons of simplicity, we only we only show an abstract form of the influences that properties of feedback and dialogue context have.

between the time-slices. Connections further back in time are, however, not possible.

In such a network, evolution over time is done by unrolling the network (see fig. 5.12). Bayesian network inference is carried out on time-slice $\mathcal{BN}_{\mathcal{ALS}_t}$ and the resulting marginal posterior probabilities of those variables X_t that have a connection with variables X'_{t+1} in the next time-slice are computed. These posteriors are then used as 'prior feedback' (Robert 1993), i.e., they are interpreted as prior distributions of those variables X_t that are used as evidence variables to variables X'_{t+1} in the subsequent time slice. Due to the first order Markov assumption, previous time slices $\mathcal{BN}_{\mathcal{ALS}_0}$ to $\mathcal{BN}_{\mathcal{ALS}_{t-1}}$ are not taken into account any more and all connections to them, as well as to all variables X_t that have no influence into the future, can be disregarded (dashed lines in fig. 5.12). The complete history is thus implicitly contained, in accumulated form, in time slice $\mathcal{BN}_{\mathcal{ALS}_t}$.

In the model in fig. 5.12, the ALS-variables C, P, U, AC, and AG, as well as the groundedness variable GR, are the ones that carry over information between time slices. Understanding at time t, for example, influences understanding at time $t+1$

(consequently, variable U_{t+1} is not only influenced by P_{t+1}, $\mathit{Feedback}_{t+1}$, and $\mathit{Context}_{t+1}$, but additionally by U_t). This is based on the assumption that listener state evolution — and attribution — is usually a gradual process.

5.7.2 WORKED EXAMPLE

Figure 5.13 shows the results of simulating the dialogue fragment in example (5.13) on page 111 using the (dynamic) Bayesian network $\mathcal{DBN}_{\mathcal{ALS}}$ in two contrasting conditions:[73] (A) Without temporal influences between dialogue segments s_t and s_{t+1}, i.e., modelled with an atemporal network (only the $\mathcal{BN}_{\mathcal{ALS}_0}$ slice of $\mathcal{DBN}_{\mathcal{ALS}}$), and (B) with temporal influences between dialogue segments, i.e., modelled with the full dynamic Bayesian network $\mathcal{DBN}_{\mathcal{ALS}}$. Each set of graphs shows how speaker S1's belief state of the ALS-variables P, U, and AC, as well as the groundedness variable GR evolve over time. Nine time-steps (a, b, …, i) are shown, each corresponding to one segment of the dialogue fragment (5.13*a*, 5.13*b*, …, 5.13*i*).

In time step *c*, positive feedback with the function understanding (u^+), alongside the information that the listener gazed at the speaker and produced a head nod, is provided. In time step *d* positive feedback with the function acceptance (ac^+) is provided together with the information that the listener produced a head nod. In time steps *e* and *g*, positive feedback of understanding (u^+) is provided (in addition to information that the listener produced a head nod, in step *g*).

In fig. 5.13*A*, each feedback event is treated in isolation and independently from the dialogue history. This results in a belief state state that does not change in the beginning, when no feedback is provided by listener U1 (from *a* to *b*). When U1 provides feedback (from *c* to *e* and at *g*), S1's belief state changes abruptly, jumping between rather distant degrees of belief, and returning to the idle state for a brief period of time when no feedback is present (at *f*).

In contrast to this, the dynamic model in fig. 5.13*B*, leads to a gradually evolving attributed listener state. In the beginning, when no feedback is provided by U1 (from *a* to *b*), the belief state shifts towards *low* perception, understanding, acceptance, and groundedness. This changes, cautiously, as soon as feedback is provided at *c* and grows towards *medium* to *high* with each subsequent feedback signal provided by U1 (at *d*, *e*, and *g*). Notably, at *f*, the belief state does not jump to the initial state, but degrades only slightly while U1 does not provide feedback.

[73] The two-time-slice Bayesian network model $\mathcal{DBN}_{\mathcal{ALS}}$ — specifically the two networks $\mathcal{BN}_{\mathcal{ALS}_0}$ and $\mathcal{BN}_{\mathcal{ALS}_\rightarrow}$ — is archived and available at DOI: 10.6084/m9.figshare.4981823 . This is the network that was used in the implemented attentive speaker agent described in chapter 8 and evaluated in chapter 9.

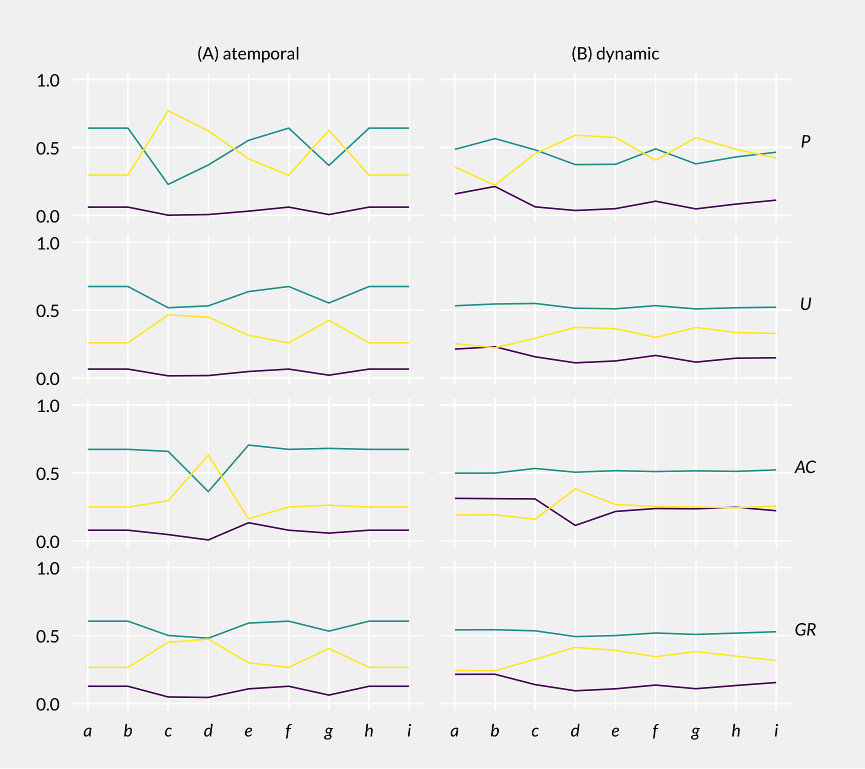


Figure 5.13: Simulated belief state evolution for example dialogue (5.13) on page 111. The graphs show how speaker S1's degrees of belief in the attributed listener state variables P, U and AC, as well as the groundedness variable GR change over time, given the feedback provided by listener U1 at c, d, e, and g. Two conditions are contrasted: (A) without temporal influences between dialogue segments, simulated with an atemporal version of the model ($\mathcal{BN}_{\mathcal{ALS}_0}$); and (B) with temporal influences between dialogue segments, simulated with the two time-slice dynamic Bayesian network model (fig. 5.12). Lines colour encodes the values of the variables as follows: (— *low*, — *medium*, and — *high*.

This example illustrates that the dynamic model of mentalising about attributed listener state produces more coherently evolving belief states. This is a desirable property since it is implausible that a temporary lack of feedback for one utterance unit (e.g., at f) is an immediate sign that an interlocutor's mental state of listening has significantly degraded. The dynamic model can easily bridge such stretches of dialogue in which feedback is absent. In a similar vein, the example also shows that the dynamic model can integrate sequences of feedback signals (or lack thereof). Multiple feedback signals with similar function will reinforce each other. Considering that an attentive speaker agent should quickly adapt its behaviour based on these variables, we can expect that the coherence in evolution will thus lead to more coherent adaptation behaviour of the agent.

A disadvantage of this rather slow evolution of the belief states is that when listener feedback indicates radical changes in an interlocutor's mental state (e.g., change of state feedback such as *oh*), it takes longer for these changes to be reflected in the attributed listener state. This could, however, be mitigated by treating such feedback signals (which are likely marked in some way) differently, e.g., by boosting their influence while simultaneously decreasing the influence of the preceding dialogue context.

5.7.3 DISCOURSE STRUCTURE AND BELIEF STATE EVOLUTION

A question that needs to be addressed is how the attributed listener state in the dynamic model should develop over time, i.e., to what extent and how the belief state $\mathcal{BN}_{\mathcal{ALS}_t}$ influences its successor state $\mathcal{BN}_{\mathcal{ALS}_{t+1}}$. For the example, in fig. 5.13*b*, the transitions were assumed to be fixed, that is, the influence $\Pr(X_{t+1} \mid X_t)$ of each of the variables $X_t \in \{C_t, \dots, GR_t\}$ on its successor $X_{t+1} \in \{C_{t+1}, \dots, GR_{t+1}\}$ was the same for each transition between dialogue segments $s_t \in \{s_a, \dots, s_i\}$.

This assumption is certainly simplified. As Muller and Prévot (2003) argue, feedback is deeply embedded in the discourse and its relation to the discourse structure is one of its pivotal features. As an example, consider a situation in which at time $t+1$ either the topic changes, or the narration simply continues. Intuitively, the influence of the speaker's attributed listener state $\mathcal{BN}_{\mathcal{ALS}_t}$ on the attributed listener state $\mathcal{BN}_{\mathcal{ALS}_{t+1}}$ is different in the two situations.

Given a topic change, there is, e.g., little reason to believe that understanding or acceptance as estimated in $\mathcal{BN}_{\mathcal{ALS}_t}$ has much to contribute — i.e., is a good predictor — to understanding and acceptance in $\mathcal{BN}_{\mathcal{ALS}_{t+1}}$ (arguably this also depends on the relatedness of the two topics). In contrast to this, understanding and acceptance as estimated in $\mathcal{BN}_{\mathcal{ALS}_t}$ seems to be very relevant for $\mathcal{BN}_{\mathcal{ALS}_{t+1}}$ in the case where the narration simply continues.

This simple example indicates that the type of relation between discourse segments — a rhetorical or discourse relation (Asher and Lascarides 2003) — plays a role in the development of attributed listener state over time (a similar argument is made by Stone and Lascarides [2010], see section 5.8).

As a first approach, we propose that the dynamic model of the listener takes the discourse relation between two consecutive discourse segments into account by simply varying the strength of the influence that a variable X_t has on a variable X_{t+1} in the next time-slice. This strength is defined in terms of a weight w that the temporal influence has in relation to the influences of feedback, dialogue context, and other ALS-variables. A weight of $w = 0.5$, for example, results in the influence of X_t on X_{t+1} being the same as the influence that all non-temporal variables combined have on X_{t+1}. A weight of $0 \leq w < 0.5$ results in temporal influence that is smaller than the influences of the non temporal variables and larger for a weight of $0.5 < w \leq 1$.

In practical terms, this approach involves (i) having different dynamic Bayesian network models for each of the discourse relation types, and (ii) switching the networks — carrying over the variable assignments and distributions — when proceeding from dialogue segment to dialogue segment. Concrete weights for individual discourse relations need to be determined empirically.

5.8 RELATED WORK

Bayesian networks have already been used to model grounding and dialogue behaviour in artificial conversational agents. Paek and Horvitz (2000a), for example, use Bayesian networks to manage the uncertainties, among other things, in a model of grounding behaviour in spoken dialogue systems. Stone and Lascarides (2010) propose to combine dynamic Bayesian networks with the logic based Segmented Discourse Representation Theory (SDRT; Asher and Lascarides [2003]) for a theory of grounding in dialogue that is both rational (in the utility theoretic sense) and coherent (by assigning discourse relations a prominent role in making sense of utterances). This approach, which can build upon a detailed model of the discourse structure, is purely theoretical so far.

In contrast to these approaches, but in line with our perspective on attentive speaking and the way we model inference about the mental states of interlocutors, Schäfer et al. (1997) present a dynamic Bayesian network for modelling the cognitive limitations of users in a spoken dialogue system. Their model takes into account properties of the conversational agent's utterance and combines these with properties of the task and the interaction partner, such as time-pressure, or temporary limitations in working memory. The model itself is then used for planning the dialogue and choosing utterances with the interlocutor's resources in mind.

5.9 DISCUSSION

Our approach to interpreting communicative listener feedback is quite different to previous work on the problem, which usually focussed on lower level aspects, using classification-based approaches that map feedback signals to their functions (e.g., Reidsma et al. 2011; Gravano et al. 2012; Neiberg et al. 2013; Philippsen et al. 2013; Lotz et al. 2016, see sections 3.3.2 and 3.6). In contrast to this, we focus on the higher level processes and require a feedback function and a description of other relevant properties of feedback signals (e.g., prosody) as input.

The listener state attribution process can take subtle properties of the signal (e.g., qualifying information such as uncertainty) as well as high-level interactions between feedback signals and dialogue context into account. Realising this model in a probabilistic framework makes it possible (i) that the result of the attribution process — the belief states of the ALS-variables and the groundedness variable — still reflect the uncertainties of the input, thereby preserving information from the signals as long as possible, and (ii) that information from different sources can be integrated in a unified and well understood framework.

Our perspective on feedback interpretation can be characterised as cognitive and inferential. It is cognitive in that it models, on a computational level, a process of mentalising, i.e., reasoning about the mental states of others (the interlocutor). At the same time, the model is kept 'minimal' in that it does not attempt to represent the interlocutor's mental state in all details (Butterfill and Apperly 2013), but limits itself to a small set of variables that capture rather abstract hybrid mental states that combine the propositional attitude 'belief' with a phenomenal state (such as understanding). This minimal form of mentalising is what makes the ALS-model a good match to the 'minimal partner model' theory (Galati and Brennan 2010; Brennan and Hanna 2009) for partner specific adaptation in dialogue (see section 2.3.5). This theory suggests that adaptation in dialogue takes its bearing from a lightweight variable-based representation, instead of making adaptation decisions based on full common ground or based on a monitoring and adjustment approach. We suggest that the ALS-variables can be thought of as further dimensions in such a minimal partner model, but offer a more expressive gradient representation for the variables, which allows for more fine-grained, but still lightweight, adaptation. It should be noted that the ALS-based grounding model (see section 5.6) is nevertheless able to inform the grounding process and can contribute to a representation of full common ground — even to one that is gradual (Brown-Schmidt 2012). This can be used as a fall-back should the minimal model be insufficient or for specific production/generation choices that require such detailed information about common ground.

Furthermore, our perspective on feedback interpretation is inferential in that it does not implement a code model of feedback meaning, but a computational and probabilistic pragmatics approach (Goodman and Frank 2016; Franke and Jäger 2016). This is what enables the dynamic, context-specific inference of the attributed listener state, and has the potential to be able to process unseen feedback signals (once the model interfaces with automatic methods for describing their features) in novel dialogue contexts.

A limitations of the attributed listener state approach to feedback interpretation presented here is that the probabilistic models are manually created, mostly based on theoretical considerations. This limitation is not a general one though. It is, in principle, possible to learn the parameters of Bayesian networks — even given a pre-specified structure — from relatively small data sets. A difficulty for this, however, is the acquisition of data that contains reliable information on human mental states underlying observed feedback behaviour. A possible solution to this problem would be to use an approach that learns the model in interaction with human interlocutors that make their listening-related mental states explicit upon request. The use of hierarchical Bayesian networks further allows for an independent formulation or learning of sub-networks. Whether a manually created model creates problems when using it in an attentive speaker agent is unknown. In general, it is not uncommon to specify Bayesian network models by hand (Koller and Friedman 2009, box 3.C, pp. 64–67,) — even for conversational agents (e.g., Paek and Horvitz 2000a). Koller and Friedman (2009, p. 95) state that 'even fairly significant changes to network parameters cause only small degradations in performance, except when the changes relate to extreme parameters — those very close to 0 and 1'. This finding suggests that our approach, via implicit representations, seems to be suitable, as the rather coarse control it provides may be sufficient. A data driven model would have the advantage of an empirical, rather than theoretical, grounding, though. Learning-based formulation of ALS-models will not be addressed in this thesis but should be explored in future work.

It should be noted that this chapter does not aim to present a definitive model of attributed listener state and its relevant contextual factors. The objective pursued in this chapter is to introduce attributed listener state modelling as a conceptual framework for inferential, context-aware and mentalising-based reasoning about the semantics and pragmatics of feedback signals.

Similarly, the visualisation of the belief states of attributed listener state variables given various input configuration provided throughout this chapter should not be regarded as an evaluation of the model. Their purpose is to illustrate which kinds of interactions can be modelled. The model will be evaluated as part of the implemented attentive speaker agent in chapter 9.

5.10 SUMMARY

In this chapter we developed a framework for computationally modelling feedback interpretation for attentive speaker agents. It frames the process as one of reasoning about the human interlocutors' listening-related mental state. Inference is carried out using a Bayesian network model of the attributed listener state that we extended, step-by-step, with various influencing factors that need to be considered in the attribution process, for example, properties of feedback signals or the dynamic dialogue context in which a feedback signal occurs. Due to the properties of Bayesian networks and in combination with the use of implicit representations, the complex interactions between these factors can be expressed in a compact manner.

The model can be thought of as a dynamic minimal interlocutor model for interactive adaptation in natural language generation using mechanisms based on the the minimal partner model theory. At the same time, the ALS-model can also be used as a way to model a context-aware grounding criterion for reasoning about grounding and common ground, given feedback-based evidence of understanding.

How adaptations can be carried out interactively and incrementally based on the inferred variables from the attributed listener state will be described in the next chapter.

CHAPTER 6
INTERACTIVE ADAPTATION

In this chapter we develop a component for attentive speaker agents that incrementally and interactively generates natural language in a way that takes the interlocutors' needs — based on the attributed listener state — into account. We begin with a review of adaptation mechanisms on various levels of language production. Following this, we describe, in detail, the incremental and adaptive natural language generation system SPUD$_{ia}$. We do this in four steps and start with a general introduction to adaptive natural language generation in which we make a case for the use of incremental processing. This is followed by a description of the SPUD microplanning framework — and its variant SPUD$_{coref}$ — on which the systems developed here are based. In the third step, we describe our incremental variant SPUD$_{inc}$, illustrate how it incrementally generates an utterance, and discuss the overall concept. Following this we describe the adaptive variant SPUD$_{ia}$ and the available adaptation mechanism. Finally, before concluding this chapter, we briefly describe our work on adaptive speech synthesis.

6.1 LEVELS AND MECHANISMS OF ADAPTATION

Based on the the dynamic minimal mentalising model of attributed listener state, speakers may decide if it is necessary or helpful to adapt to the listeners by changing aspects of their language production behaviour. This section describes mechanisms of adaptation based on findings in the literature.

☆ This chapter contains material previously published in Buschmeier et al. (2012) and Buschmeier and Kopp (2012; 2013). The incremental and adaptive natural language generation system 'SPUD$_{ia}$', described in section 6.2, is based on DeVault's (2008) SPUD implementation 'SPUD$_{coref}$'. The initial version of SPUD$_{ia}$ was developed in a Master's thesis (Dosch 2011) supervised by the author of this thesis. ¶ In Buschmeier et al. (2012), the author of this thesis was responsible for the part on (incremental) natural language generation and the subjective evaluation of the integrated system. Timo Baumann was responsible for the part on (incremental) speech synthesis and the objective evaluation. The integration was done jointly.

Table 6.1: Levels of adaptation, from the lowest level 'realisation' to the highest level 'perspective'.

Levels	Mechanisms
Perspective	perspective-change
	provide missing information
Rhetorical structure	elaboration
	explanation
	repetition
	summary
	pragmatic explicitness
Surface form	verbosity
	redundancy
	focus/stress/prominence
	vocabulary
Realisation	hyper- and hypo-articulation
	Lombard speech
	speech rate
	volume

The different needs of a listener need to be addressed on different levels of language production and using different adaptation mechanisms. A problem in perception might, for example, be resolved by simply repeating the utterance or the problematic phrase or word. If a speaker notices, however, that the listener has built up a completely different situation model and is stuck in this incorrect conceptualisation of what the speaker means (detection of a misunderstanding), starting anew from a different perspective might be the right way for the speaker to resolve the problem. Table 6.1 provides an overview of different levels of adaptation along with a choice of mechanisms that operate on each level.

The lowest level of adaptation is the realisation level, i.e., how an utterance is articulated and presented. Adaptation on this level might happen automatically during articulation along the hyper–hypo continuum (Lindblom 1990). A speaker might choose to hyper-articulate when the listener has difficulties perceiving the speaker's speech (e.g., due to noise in the environment, hearing impairment, importance of the message or possible ambiguities). If, on the other hand, the listener perceives well and the message is not overly important, speakers might choose to conserve energy through hypo-articulation. The realisation level is also where speakers may choose to adapt their speech rate, the volume of their voice, and other acoustic features. These low-level adaptations are often automatic (e.g., Lombard adaptation of speech. [Cooke et al. 2014]).

If adapting the realisation is insufficient to accommodate the listener's needs, the utterance's content itself can be adapted. This is possible on all of the higher adaptation levels. The simplest way of adapting utterance content is to change the surface form, keeping the utterance's semantic content fixed. A speaker may choose to be more 'verbose', i.e., use more words to communicate the same semantic content. Although the additional words and phrases might not add semantic content, they can nevertheless serve important communicative functions. Using signpost language and other cue phrases for example helps in drawing the listener's attention to a specific aspect of an utterance. It might also be used to make the speaker's underlying intentions more explicit and to reveal the rhetorical structure of the speaker's argument (Grosz and Sidner 1986). Verbosity also has the simple property of giving listeners more time to process the important meaning-bearing parts of an utterance.

Speakers may also use different degrees of redundancy to adapt surface form. Similarly to verbosity, redundancy usually does not introduce novel semantic objects, but highlights important information and increases the probability of the message being understood (Reiter and Sripada 2002). Redundancy is also a frequent mechanism used to repair misunderstanding (Baker et al. 2008).

Another mechanism that operates on the surface structure is stress and focus. The speaker might put stress (or more general 'prominence') on the important parts of an utterance with the help of prosodic cues as well as by using different syntactic constructions that distribute weight differently (e.g. active vs. passive voice). Furthermore, the speaker can choose a different vocabulary, thereby accommodating the listener's level of expertise (Janarthanam and Lemon 2014).

Adaptation at higher levels requires more than a change of packaging for semantic content, producing instead a different message. 'Rhetorical structure' is the level of adaptation most easily identified and often found in the analysis of our corpus. Speakers often adapt to listener feedback by changing the amount of information they provide. They commonly elaborate on an utterance by providing more information or giving explanations. Another is to repeat the previous utterance or to summarise several utterances. On this level, speakers also adapt by making previously implicit information pragmatically explicit.

Finally, when speakers notice that the listener's conceptualisation of the dialogue topic deviates from their own, they may adapt on the level of 'perspective'. They adjust their own perspective to be closer to that of the listener, or track back to a point in the dialogue where they assume the conceptualisation to have still been consistent. Speakers might also provide further background information that they had previously assumed was already a part of common ground.

It should be noted that adaptation can take place at multiple levels simultaneously. A speaker might very well choose to communicate more clearly by combining several mechanisms. Furthermore, the function of adaptation is not limited to accommodating for the listener's problems in perception, understanding, and so forth. It also serves to modify dialogue when communication is going 'too well'. For example, if a speaker notices that a listener is already ahead in her thinking, he might skip planned parts of his utterance. Similarly, if there are no problems in perception and understanding, the speaker can be more relaxed in his or her articulation.

In the following we focus on the level of surface form, specifically on adaptation in natural language generation. Structural adaptation, on the level of dialogue management, will be described in section 8.3.4.

6.2 ADAPTIVE NATURAL LANGUAGE GENERATION

Traditionally, natural language generation (NLG) is defined as the field 'that focuses on computer systems that can produce understandable *texts* in [...] human languages' (Reiter and Dale 2000, p. 1; emphasis added), typically from data that has a non-linguistic representation. A natural language generation system for a conversational agent, however, does not produce text, but utterances for use in conversation, which is a fundamentally different setting of language use (Clark 1996, pp. 4–11). Where text is produced at one point and read (or perhaps listened to) at a later point, utterances are produced extemporaneously, in real-time, for an interaction partner who listens to them instantaneously, while they are spoken. Whereas the readers of a text are often unknown when it is written, the addressees of an utterance in conversation are co-present. Where text is static and cannot be changed post writing, utterances are dynamic and subject to change and adaptation while unfolding (see section 2.3), and can be stopped at almost any point (Tydgat et al. 2011).

Because of these properties of spoken language, especially instantaneity and extemporaneity, natural language generation systems for text and utterance generation in conversational agents have different requirements. Whereas a text generating system is allowed to take its time to produce a text that will be read eventually, an utterance generating system needs to be able to produce at least some words as soon as the conversational agent gets the turn. Whereas a text generating system knows up front what, in general, constitutes a 'good' text in its domain (often from corpora and user evaluation, Reiter and Dale [2000, §§ 2.3 and 2.4]), an utterance generating system can get direct feedback about the quality of its output from its interaction partner and could — in principle — optimise its output in a timely manner, by adapting it interactively to the interlocutor's needs.

Thus, a fundamental requirement for a natural language generation system that generates utterances in real-time and can adapt its utterances while still producing them, is incremental processing. Guhe (2007, p. 70), extending Levelt (1989, p. 26), defines this to be the case when

> each processing component [is] triggered into activity by a minimal amount of its characteristic input and produces characteristic output as soon as a minimal amount of output is available.

Psycholinguistic research has identified incrementality as an important property of human language production early on and it has been incorporated into several models of human speech production (e.g., Kempen and Hoenkamp 1987; Levelt 1989). Natural language generation systems that support incremental processing are rare, however.

A notable exception is the in-depth analysis of requirements for and properties of incremental natural language generation by Kilger and Finkler (1995), who describe an LTAG-based incremental sentence generator ('VM-GEN'). VM-GEN takes incremental input (lemmata and their semantic relations, basically a 'preverbal message' in Levelt's [1989] terms), processes it and produces output as soon as at least a prefix of the final sentence is syntactically complete. If VM-GEN notices that it committed itself to a prefix too early, it can initiate an overt self-repair.

Guhe (2007) presents a computational model of incremental conceptualisation in natural language generation ('INC'), which incrementally builds conceptual representations from perceived domain objects, selects them for verbalisation, linearises them, and produces a 'preverbal message'. This output of INC can then be used for sentence planning in subsequent generation stages, which, however, are not part of Guhe's model.

Both VM-GEN and INC focus on specific parts of the natural language generation process.[74] In contrast to this, Skantze and Hjalmarsson (2013) present a system that performs incremental generation in the context of a spoken dialogue system and thus needs to cover the whole generation process in interaction with speech input from a user. This system can already start to produce output when the user has not yet finished

[74] Natural language generation is usually seen as a process that involves three consecutive stages: document planning, microplanning/sentence planning, and surface realisation (Reiter and Dale 2000, p. 49, table 3.1). This 'consensus architecture' of NLG-systems (Reiter 1994) is similar to Levelt's (1989) psycholinguistic model of speech production which consists of the three stages conceptualisation, formulation and articulation.

speaking and only a preliminary interpretation exists. By flexibly changing what to say and by being able to make self-repairs the system can recover from situations where it selected and committed on an inadequate speech plan.

None of these incremental natural language generation systems, however, is able to flexibly adapt the language that it generates to changing requirements due to changes in the situation or changing needs on the side of the user.

Natural language generation systems that are able to adapt their output significantly — e.g., stylistically (e.g., Walker et al. 2007; Mairesse and Walker 2010), based on the interlocutors' estimated level of competence (e.g., Janarthanam and Lemon 2014), to the lexical and syntactic structures that the interlocutors use (e.g., Buschmeier et al. 2010; Isard et al. 2006), to the interlocutors' level of politeness (e.g., de Jong et al. 2008), or to the interlocutors' gaze behaviour Garoufi et al. (2016) — do not work incrementally and cannot adapt to their interaction partners on the fly and in real-time. The following sections present an incremental natural language generation system that is able to adapt its utterance while it is still being articulated.

6.2.1 THE SPUD MICROPLANNING FRAMEWORK

The incremental and adaptive natural language generation system described here, 'SPUD$_{ia}$', is based on the non-incremental and non-adaptive microplanning framework 'SPUD' (an acronym for 'sentence planning using descriptions', Stone et al. 2003), more specifically on the variant 'SPUD$_{coref}$' developed and implemented by DeVault (2008, 'SPUD$_{coref}$'), see fig. 6.1 for an overview of SPUD-based NLG-systems.

The SPUD$_{coref}$ generator constitutes one part in a more general conversational agent architecture (ibid., pp. 77, 129). It possesses a set of linguistic resources and maintains a model of its domain knowledge, its interlocutor, and the discourse context (Stone et al. 2003, p. 347). Within its model of discourse context, it represents the information status of its knowledge: it distinguishes knowledge that it considers to be shared between itself and its interlocutor (basically a representation of their common ground) from knowledge that it considers to be private.

SPUD$_{coref}$'s model of language — its linguistic resources — is specified in the feature-based tree-rewriting grammar formalism TAGLET[75] (Stone 2002, app. A). As in construction grammars, each linguistic resource (a TAGLET element) in SPUD

[75] TAGLET is inspired by the lexicalised tree-adjoining grammar formalism LTAG, (Joshi and Schabes 1997), which is especially well suited for natural language generation (Joshi 1987) and used in SPUD. TAGLET retains properties of LTAG — e.g., 'enabling use of grammar in high-level [planning] tasks' (Stone 2002, p. 79) — but is merely context-free in expressive power and therefore computationally more lightweight than LTAG, which is mildly context sensitive.

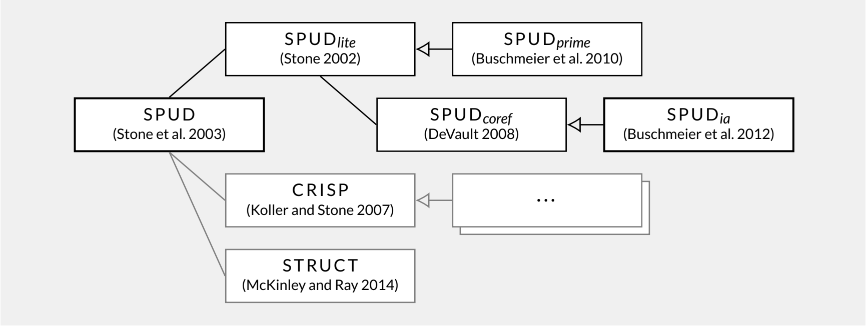

SPUD$_{lite}$ (Stone 2002)
SPUD$_{prime}$ (Buschmeier et al. 2010)
SPUD (Stone et al. 2003)
SPUD$_{coref}$ (DeVault 2008)
SPUD$_{ia}$ (Buschmeier et al. 2012)
CRISP (Koller and Stone 2007)
...
STRUCT (McKinley and Ray 2014)

Figure 6.1: Family tree of natural language generation systems build on the ideas of the SPUD microplanning framework (Stone et al. 2003). Connections indicate inheritance of ideas, arrows indicate inheritance of source code (from the pointed-to system).

combines lexical, syntactic, and semantic[76] information. Surface form, syntax, and semantics of a communicative action are thus constructed simultaneously — through the syntactic operations substitution, pre-, and post-modification, and unification of its features structures — during generation.

In SPUD, a microplanning task is not a specification of the semantics of a sentence (the semantics of the sentence — its interpretation — is constructed during generation), but a specification of the communicative effects (in form of a list of logical formulae, or 'updates') that the final sentence should have. Once the utterance has been verbalised, and assuming understanding on the side of the interlocutor, these updates can then be added to the representation of common ground. If some of the formulae in the specified communicative effects are already part of its representation of common ground, SPUD can presuppose them in the generation process and only needs to assert those formulae that it still considers private. Pragmatic choices — adhering, e.g., to Grice's (1975) maxim of quantity — are thus made automatically.

SPUD frames microplanning as an AI search problem.[77] It carries out deliberate goal-directed actions towards finding an optimal solution to all sub-tasks in sentence

[76] SPUD uses a 'flat' approach to semantics in which the meaning of a derivation (a derived tree in which all leaf nodes are non-terminals) is simply a conjunction of the parametrised meanings of its elements (Stone and Doran 1997).

[77] How the search problem is approached differs in the various implementations of SPUD. It was originally formulated as a constraint satisfaction problem (Stone et al. 2003; Stone 2002), later as an automatic planning

generation — lexical and syntactic choice, referring expression generation, aggregation (Reiter and Dale 2000) — at once. The state space of the search problem is the (potentially) infinite set of sentences that can be generated from SPUD's linguistic resources. The search starts from an initial state: an elementary tree that will become the root node of the derived tree, and the shared and private knowledge relevant in the generation task (Stone et al. 2003, p. 347).

Using an informed search strategy, SPUD attempts to make progress towards a syntactically and semantically complete and valid derived tree whose communicative intent satisfies the specified communicative effects given the representation of discourse context, that is, 'the generator's communicative intent must provide a complete sentence [...] that says what is needed [...] in a way the hearer will understand' (ibid., p. 348–49). In each step, the search algorithm expands the 'provisional' utterance (basically the search state) by adding the linguistic resource (using one of the syntactic operations, unifying its features) that maximally reduces the estimated distance[78] to the final utterance.

6.2.2 INCREMENTAL GENERATION WITH SPUD*INC*

SPUD constructs utterances incrementally in the sense that the communicative intent of the provisional utterance progresses towards the communicative intent of the final utterance with every expansion of the search state. This, however, does not mean that the surface structure of provisional utterances is generated incrementally (i.e., from left to right) as well, which would require special considerations in its formal foundations. Such considerations are addressed in a variant of LTAG that is 'psycholinguistically motivated' (PLTAG Demberg et al. 2013), but has, so far, only been used for incremental parsing, not generation. Full word-by-word incrementality in natural language microplanning is thus not within reach for the SPUD framework as is.

From an empirical perspective, it has long been established that different levels of speech production ('from intention to articulation', Levelt 1989) operate on different increment sizes. What the increments on each level look like, however, is still a topic of debate. Message planning, being relevant to incremental natural language generation — and at the interface to microplanning — has been found to be a continuous process (utterances are not fully planned in advance) that allows for updating of the message, even after articulation begins, while maintaining fluency (Brown-Schmidt

problem (Koller and Stone 2007), and most recently as probabilistic decision making (McKinley and Ray 2014).

[78] The heuristic function for the distance measure varies with the actual implementation of SPUD.

and Konopka 2014), especially in conversational settings. It has also been found that message planning works hierarchically in the sense that 'evidence of downstream planning, several words in advance', was found before speech onset and during articulation (Lee et al. 2013, p. 556). This indicates that although utterances are, trivially, articulated word-by-word, message planning and microplanning work on larger increments. Lee et al. conclude that the scope of pre-planning is likely variable, depending on many factors including, situation, syntactic construction, and the speaker itself.

With this in mind, we take a more coarse-grained approach to incremental language generation. Instead of words, we choose 'utterance units' (roughly the size of 'intonation phrases', see Poesio and Traum [1997, pp. 317–318] and Traum and Heeman [1997]), as our incremental output units. We think that this is a good choice of increment size for multiple reasons. Firstly, the utterance unit corresponds to Lee et al.'s (2013) finding that message planning takes grammatical scope into account, this makes it a plausible choice. Secondly, the utterance unit is similar to the intonation phrase, which is a prosodic unit with coherent intonation (Selting et al. 2011, § 3.3.1) that is often semantically complete. This is relevant for the actual realisation of incrementally produced discourse units in speech, as the prosody of individually synthesised utterance units will be natural — to a certain extent — even if speech synthesis is non-incremental (see, for example, the visualisation of the intonation contour of a discourse unit synthesised with different degrees of incrementality in Baumann and Schlangen [2012b, fig. 3]). Finally, the utterance unit is the typical unit that is grounded in dialogue (Traum and Heeman 1997; Poesio and Traum 1997), that is, a unit we might expect to receive feedback for from dialogue partners.

As in SPUD, the task-specification of SPUD$_{inc}$, the incremental version of SPUD, is a set of communicative effects that the utterance to be generated is supposed to achieve. SPUD$_{inc}$ also has access to the model of discourse context and thus knows which information is considered to be shared and which is considered to be private.

In contrast to SPUD, generation in SPUD$_{inc}$ is a hierarchical process comprising two levels: *micro-content-planning* (MCP), and *microplanning-proper* (MPP). On the MCP-level a high-level specification of a communicative goal that can be verbalised within a discourse unit (Poesio and Traum 1997, pp. 317–318) — a speech act, e.g., to inform about a number of events, or about a scheduling conflict between two events — is used to incrementally plan the structure of individual utterance units (of the discourse unit). On the MPP-level a low-level specification of the communicative effects of an utterance unit — e.g., mention the start time of an event such that it can be understood by the interlocutor — is used to non-incrementally construct the communicative intent and surface structure of each utterance unit.

The communicative goal for the MCP-level is specified in form of an outline of a discourse unit. It consists of the set of desired updates (all communicative effects that the discourse unit should achieve), the set of presupposed/shared knowledge, and the set of private knowledge of the speaker agent. Importantly, the outline describes how the communicative goal can be decomposed into 'incremental microplanning tasks' (IMPT) for the MPP-level. Each IMPT consists of (i) a subset of the communicative goal's desired updates that belong together and fit into one utterance unit, and (ii) the presupposed and private knowledge needed to generate this utterance unit. On the MPP-level, the IMPT is provided to the SPUD-algorithm, which generates the utterance unit's surface form and communicative intent as described in section 6.2.1.

The overall generation process in SPUD$_{inc}$ is centrally controlled. IMPTs are incrementally requested from the MCP-level and passed on to the MCP-level. Based on the representation of communicative intent of the generated utterance units, the achieved communicative effects are then added to the agent's discourse context and removed from its private knowledge.

Both discourse unit planning on the MCP-level and utterance unit generation on the MPP-level have access to the representation of discourse context. Thus the SPUD-algorithm is aware if a desired update of an IMPT has already been communicated in a previous utterance unit and can take this information into account during generation. Despite the individual generation of utterance units, the overall discourse unit can thus adhere to pragmatic principles and be coherent.

Generation example The following example illustrates the incremental generation of a discourse unit with the communicative effect of making two upcoming events later in the same week known to the interlocutor.

The first event is the class 'CS 533 Computational Linguistics' which takes place on 27 March 2002, a Wednesday, from 11:30 to 13:20. The information that describes this event is specified as follows:

$$\begin{aligned} e_1 &: \mathit{subject}(e_1, \text{`}\texttt{CS 533}\text{'}), \mathit{on}(e_1, t_1), \mathit{from}(e_1, t_1), \mathit{to}(e_1, t_2), \mathit{until}(e_1, t_2), \\ t_1 &: \mathit{day}(t_1, \text{`}\texttt{27}\text{'}), \mathit{month}(t_1, \text{`}\texttt{mar}\text{'}), \mathit{time}(t_1, \text{`}\texttt{11}\text{'}, \text{`}\texttt{30}\text{'}), \mathit{dow}(t_1, \text{`}\texttt{wed}\text{'}), \\ t_2 &: \mathit{day}(t_2, \text{`}\texttt{27}\text{'}), \mathit{month}(t_2, \text{`}\texttt{mar}\text{'}), \mathit{time}(t_2, \text{`}\texttt{13}\text{'}, \text{`}\texttt{20}\text{'}), \mathit{dow}(t_1, \text{`}\texttt{wed}\text{'}). \end{aligned} \tag{6.1}$$

The second event, 'Lunch', takes place on the same day, from 13:20 to 14:00:

$$\begin{aligned} e_2 &: \mathit{subject}(e_2, \text{`}\texttt{Lunch}\text{'}), \mathit{on}(e_2, t_3), \mathit{from}(e_1, t_3), \mathit{to}(e_2, t_4), \mathit{until}(e_2, t_4), \\ t_3 &: \mathit{day}(t_3, \text{`}\texttt{27}\text{'}), \mathit{month}(t_3, \text{`}\texttt{mar}\text{'}), \mathit{time}(t_3, \text{`}\texttt{13}\text{'}, \text{`}\texttt{20}\text{'}), \mathit{dow}(t_3, \text{`}\texttt{wed}\text{'}), \\ t_4 &: \mathit{day}(t_4, \text{`}\texttt{27}\text{'}), \mathit{month}(t_4, \text{`}\texttt{mar}\text{'}), \mathit{time}(t_4, \text{`}\texttt{14}\text{'}, \text{`}\texttt{00}\text{'}), \mathit{dow}(t_4, \text{`}\texttt{wed}\text{'}), \end{aligned} \tag{6.2}$$

The information about the two events eqs. (6.1) and (6.2), together with the information that they should be announced, is provided as the set of desired communicative effects to the MCP level of SPUD$_{inc}$. The utterance outline for a discourse unit that makes such announcements consists of a variable number of incremental microplanning tasks (depending on the number of events to announce; in this example seven IMPTS, one for marking the speech act, three to communicate the details of each event).

The first IMPT (see table 6.2) desires a single update $intro(announce)$. It is provided to the SPUD-algorithm which, using a single linguistic resource which has no presuppositions and asserts $intro(announce)$, constructs the utterance unit with the surface form *Die Termine sind* ('The events are')[79].

The next three IMPTS (2–4 in table 6.2) communicate the information of event e_1, eq. (6.1): date, time slot, and subject. The desired updates for IMPT$_3$, for example, are start time —$from(e_1, t_1)$ — and end time — $until(e_1, t_2)$ — of the event. The SPUD-algorithm generates the utterance unit *von 11 Uhr 30 bis 13 Uhr 20* ('from 11:30 to 13:20'), see fig. 6.2 for the resulting derived tree.[80] The communicative effects that are actually achieved, go beyond the requested (desired) updates. Details of the timing information (start and end time in hours and minutes) are part of the utterance unit, and, thus, $time(t_1, 11, 30)$ and $time(t_2, 13, 20)$ are part of its communicative intent, and its set of achieved updates.

The second event (e_2) is described in the next three IMPTS (5–7 in table 6.2). The realisation of IMPT$_5$, which should communicate its date, is realised in a different way than the corresponding IMPT$_2$ for event e_1. As both event intervals 'meet' at 13:20 on 27 March 2002 (e_1 ends and e_2 starts at that point of time) and e_1 was communicated right before e_2, the SPUD-algorithm generates the utterance unit *und direkt danach* ('and directly afterwards'), see fig. 6.2 for the resulting derived tree. The meets-relationship between the two events is encoded with a semantic constraint on the linguistic resource s – (*direkt, danach*). It can only be used if (i) a formula $until(e_1, T_i)$ can be presupposed (i.e., is in the common ground), (ii) when the set of desired updates contains the formulae $at(e_2, T_j)$ and $until(e_2, T_j)$, and (iii) if the variables T_i and T_j match. This is the case for t_2 and t_3, see eqs. (6.1) and (6.2). The generated utterance unit implicitly

[79] Generated utterance units are in German.

[80] The linguistic resources used in this example — and in the attentive speaker system described in chapter 8 — differ from linguistically well-founded general LTAG grammars, such as for example XTAG (XTAG Research Group 2001) for English, or a recent corpus-derived PLTAG grammar for German (Kaeshammer and Demberg 2012) that could enable incremental generation. SPUD$_{inc}$'s linguistic resources have been engineered, rather pragmatically, for the generation of domain specific discourse units describing calendar events and operations. This is also due to the fact that available wide coverage LTAG-grammars lack the semantics needed for generation.

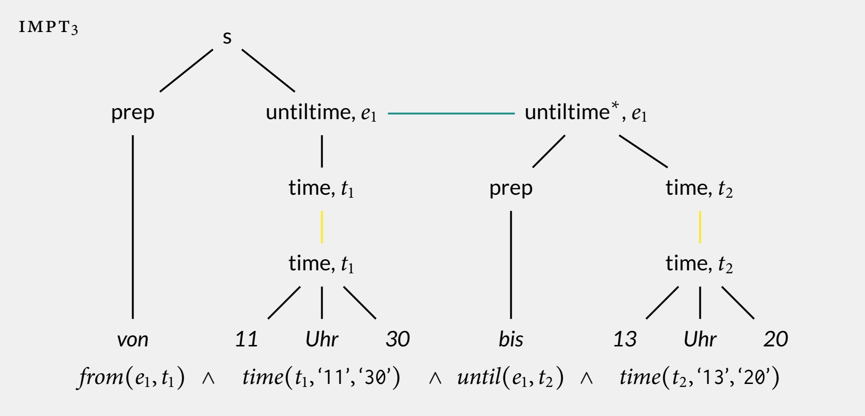

IMPT3
s
prep
untiltime, e_1
untiltime*, e_1
time, t_1
prep
time, t_2
time, t_1
time, t_2
von
11
Uhr
30
bis
13
Uhr
20
$from(e_1, t_1) \wedge time(t_1, \text{'11'}, \text{'30'}) \wedge until(e_1, t_2) \wedge time(t_2, \text{'13'}, \text{'20'})$

Figure 6.2: Derived tree of the utterance unit *von 11 Uhr 30 bis 13 Uhr 20* ('from 11:30 to 13:20') generated from IMPT3 (see table 6.2). Duplicate nodes connected by a coloured line show the construction from linguistic resources, achieved through TAGLET-operations substitution (yellow connection) and sister adjunction (green connection; in this case post-modification), and are unified.

communicates the start time of e_2. In addition to the update $on(e_2, t_3)$, IMPT5 thus achieves the update $from(e_2, t_3)$, which was originally designated as a desired update for the next IMPT.

As $from(e_2, t_3)$ is already part of the discourse context when processing the sixth incremental microplanning task, there is no need for the utterance unit to be generated to explicitly communicate the start time of e_2. The result is *bis 14 Uhr* ('until 14:00'), see table 6.2.

Discussion The outline of a discourse unit that the MCP-level of SPUD$_{inc}$ processes and transforms into incremental microplanning tasks is similar to the modern 'template-based' natural language generation approaches discussed in van Deemter et al. (2005). These use templates which allow for recursive embedding of other templates or chunks of language that are 'properly' generated (e.g., referring expressions) and also make sense from a speech production point of view as speakers often use 'routines' that are nevertheless subject to situation specific adaptations (Pickering and Garrod 2004, § 5,

Table 6.2: SPUD$_{inc}$ example generation of a discourse unit that announces two upcoming events (specified in eqs. [6.1] and [6.2]). Based on the outline for such announcement acts, the MCP-level incrementally generates seven incremental microplanning task (IMPTS) for the discourse unit, each of which specifies a set of updates that are desired to be achieved. On the MPP-level the SPUD-algorithm generates an utterance unit for each of these IMPTS (translated from German). The interpretation of each utterance unit (its communicative intent) achieves a set of communicative effects.

MCP-level		MPP-level	
IMPT	desired updates	utterance unit	achieved updates
1	$intro(announce)$	*The events are:*	$intro(announce)$
2	$on(e_1, t_1)$	*on Wednesday*	$on(e_1, t_1)$, $day(t_1, \text{'27'})$, $dow(t_1, \text{'wed'})$, $month(t_1, \text{'mar'})$
3	$from(e_1, t_1)$, $until(e_1, t_2)$	*from 11:30 to 13:20*	$from(e_1, t_1)$, $time(t_1, \text{'11'}, \text{'30'})$, $until(e_1, t_2)$, $time(t_2, \text{'13'}, \text{'20'})$
4	$subject(e_1, \text{'CS 533'})$	*CS 533*	$subject(e_1, \text{'CS 533'})$
5	$on(e_2, t_3)$ *and*	*and directly afterwards*	$on(e_2, t_3)$, $from(e_2, t_2)$ *and*
6	$from(e_2, t_3)$, $until(e_2, t_4)$	*until 14:00*	$from(e_2, t_3)$, $until(e_2, t_4)$, $time(t_4, \text{'14'}, \text{'00'})$
7	$subject(e_2, \text{'Lunch'})$	*Lunch*	$subject(e_2, \text{'Lunch'})$

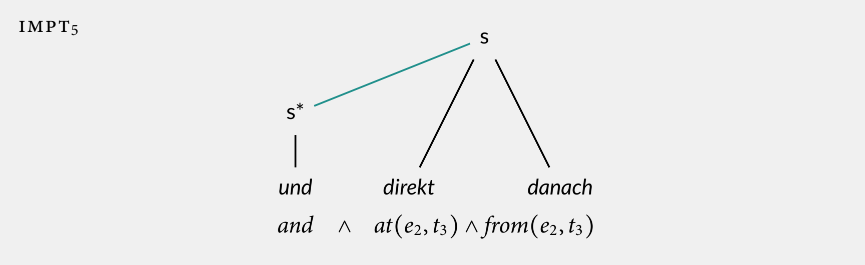


Figure 6.3: Derived tree of the utterance unit *und direkt danach* ('and directly afterwards') generated from IMPT5 (see table 6.2). As in fig. 6.2, the connection by a green line indicates a sister adjunction (in this case pre-modification) of two linguistic resources. The use of the resource s – (*direkt, danach*) is enabled by a semantic constraint modelling the relationship between event e_2 and the preceding event e_1.

especially § 5.2.2). We thus think that it makes sense to use these flexible template-like structures on the planning level.

Natural language generation is a computationally complex problem as the search space is very large and finding a solution to a communicative goal can take a considerable amount of time, especially when generating long utterances.[81] Utterance unit-based incremental generation in SPUD$_{inc}$ does not suffer from this problem as the incremental microplanning tasks on the MPP-level are usually small (in comparison to microplanning tasks for whole discourse units). In addition, usable output is already available after the first utterance unit of a discourse unit. This has the potential to increase the responsiveness of an artificial conversational agents, important for the design of conversational, 'human-like spoken dialogue systems' (Edlund et al. 2008).

Apart from these computational aspects, incremental generation brings a range of interactive phenomena that are fundamental to successful human conversation into reach of artificial conversational agents. Agents can, for example, start speaking before they actually have enough information to form a complete communicative goal (Skantze and Hjalmarsson 2013), produce incremental self repair if needed (Hough 2015), or produce compound contribution with their interlocutor (Howes et al. 2011).

[81] Even for a rather small fragment of German used in the SPUD$_{lite}$-based (Stone 2002) non-incremental, alignment-capable microplanner SPUD$_{prime}$ (Buschmeier et al. 2010), the generation of ten-word-sentences could take several seconds.

In general, the ability to update the message in response to 'interactive demands' (Brown-Schmidt and Konopka 2014, p. 842) — overtly noticeable or not — can enable the interactive adaptations, described in section 2.3.1. In the following, we describe our approach to adaptation.

6.2.3 ADAPTIVE GENERATION IN SPUD*IA*

SPUD$_{inc}$'s capability to generate language in utterance units makes it possible to change increments as long as they have not yet been overtly realised, or, if an utterance unit has already been verbalised, generate a self-repair and re-generate a different version of (only) that utterance unit that is in need of repair. In the following we present SPUD$_{ia}$, a version of SPUD$_{inc}$ that is adaptive on both levels of processing: MCP and MPP. SPUD$_{ia}$ allows changes until the IMPT affected by the change is passed from the MCP-level to the MPP-level. Depending on how generated utterance units are realised in speech (see section 6.3 and especially section 8.3.5), this means that changes are possible almost until the preceding increment finishes. Interactive dialogue phenomena can thus be dealt with in a timely manner.

On the MCP-level, changes to the communicative goal itself can be dynamically incorporated into the generation process. When conceptualising the utterance outline as a queue of IMPTS that need to be communicated, incremental microplanning, from an MCP-perspective, means taking the first element from the IMPT-queue and passing it to the MPP-level and continue doing so until the queue is empty. Given this perspective, adaptation on the MCP-level can be framed in terms of operations that alter the queue. IMPTS may get inserted into, or removed from the queue, or their position in the queue — relative to the other IMPTS — may be changed.

A conversational agent might want to repeat an IMPT that has just been generated and realised, for example, when it becomes aware of a perception problem on the side of the interlocutor. This can be achieved by re-inserting the IMPT that has just been removed from the front of the queue back to same position, optionally in combination with another IMPT that contains information for the incremental generation of a repair marker. A conversational agent might want to skip an IMPT instead of generating it, e.g., when the information it contains is already known, or not relevant given the interlocutor's current level of understanding. In this case the IMPT is simply removed from the queue and will thus never be passed to MPP. Similarly, a conversational agent might want to extend the discourse unit, to dynamically provide more context, for example. One or more IMPTS containing the desired updates can be inserted into the queue at appropriate positions. A conversational agent might also want to postpone or bring forward a planned IMPT. This can be done by shifting the IMPT within the

Table 6.3: Examples of adapted natural language output, subject to variation due to different adaptation mechanisms. Redundancy can either be prohibited (*a*) or permitted (*b*). Verbosity can take different strength, from low to high (*c*–*d*). On a structural level, utterance units can be skipped (*f*), produced as planned (*g*), or postponed for an adapted repetition of the previous chunk (*h*). A '◇' marks utterance unit boundaries, a 'ε', shows where an utterance unit/IMPT has been skipped.

Mechanism		Generated output
Redundancy	*a*	*morgen ◇ von 11 Uhr 30 bis 13 Uhr 20* '(tomorrow ◇ from 11:30 to 13:20)'
	b	*morgen den 27. März ◇ 11 Uhr 30 bis 13 Uhr 20 Uhr* '(tomorrow 27 March ◇ 11:30 to 13:20)'
Verbosity	*c*	*CS 533* 'CS 533'
	d	*Betreff: CS 533* 'subject: CS 533'
	e	*mit dem Betreff CS 533* 'with the subject: CS 533'
Structure	*f*	*ε ◇ 11 Uhr 30 ◇ CS 533* 'ε ◇ 11:30 ◇ CS 533'
	g	*morgen ◇ 11 Uhr 30 ◇ CS 533* 'tomorrow ◇ 11:30 ◇ CS 533'
	h	*morgen ◇ 11 Uhr 30 ◇ ähm ◇ von 11 Uhr 30 bis 13 Uhr 30 ◇ CS 533* '(tomorrow ◇ 13:20 ◇ uhm ◇ 11:30 to 13:20 ◇ CS533'

queue. Table 6.3*f*–*h* shows example surface forms given various adaptation operations.

In addition to such structural changes, more local changes to the communicative goal are also mediated by the MCP-level as the knowledge of how the communicative effects of a discourse unit are distributed across individual IMPTS and utterance units resides there. Communicative effects of an IMPT can thus be updated (e.g., in light of new information), added (e.g., to make a discourse unit more redundant), or removed (to make it simpler).

Adaptations operating on the MPP-level influence the choices (lexical and syntactic) that microplanning makes while transforming IMPTS into communicative intent and surface form. As these choices are ultimately determined by SPUD's heuristic function — which evaluates the candidate next search states during generation and determines which linguistic resource will be integrated into the provisional tree next[82] — , adaptation is achieved through dynamic changes in its parametrisation. If the state of understanding attributed to the interlocutor is low, for example, the heuristic function may rank candidate search states that contain redundancy higher, although redundancy is normally dis-preferred in SPUD.

Adaptation in MCP is controlled top-down, for example by dialogue management. Adaptation in MPP on the other hand depends on the task given and on the status of the knowledge used during generation. The details are then governed by the global parameter settings MPP uses during generation.

If there is, for example, reason for the system to believe that the current increment was not communicated clearly because of noise in the transmission channel, the MCP process might delay future IMPTS and initiate a repair of the current one by re-inserting it at the beginning of the list of upcoming IMPTS of this utterance outline. The MPP process' next task then is to re-generate the same IMPT again. Due to changes in the state information and situation that influence microplanning, the resulting communicative intent and surface form might differ from the previous one.

6.2.4 ADAPTATION MECHANISMS IN SPUD$_{IA}$

A number of adaptation mechanism are integrated into our NLG-microplanning system. The goal of these mechanisms is to respond to a dialogue partner's changing abilities to perceive and/or understand the information the system wants to convey. The mechanisms are implemented either with the knowledge and its conversational status used in generation (i.e., basically relying on what is considered common ground) or by altering the decision structure of SPUD's search algorithm's heuristic function.

Similar to the approach of flexible natural language generation described by Mairesse and Walker (2010), the latter mechanisms are conditioned upon individual flags (see table 6.4), which, in our case, are set based on the value of the attributed listener state' variable U (which represents the level of understanding the system attributes to the user). SPUD$_{ia}$ thus makes its adaptation decisions in a way similar to how decisions in speech production are made according to the minimal partner model theory (Galati and Brennan 2010; Brennan et al. 2010), see section 2.3.5. In the following we describe two adaptation mechanisms: verbosity and redundancy.

[82] As in SPUD$_{coref}$, the evaluation function in SPUD$_{ia}$ creates an order of the candidate next states. This ordering is done according to a number of criteria (ordered descendingly by importance), for example, which candidate search state achieves more desired updates, which is less ambiguous, etc. The order of two candidate next states is determined as soon as a criterion favours one over the other.

Table 6.4: Description of adaptation flags in SPUD$_{ia}$. Flags are set centrally in as soon as an update to the attributed listener state representation becomes available.

Flag	Description
GROUND-EFFECT	Add all asserted information to the local representation of information status so that they can be presupposed in subsequent parts of the utterance unit
PREFER-REDUNDANCY	May use redundant expression
USE-SHORT-DATE	May use a short date format, e.g. *next Wednesday* instead of *next Wednesday, 27 March*
USE-RELATIVE-DAYS	May use relative days like ‘today’ and ‘tomorrow’ to describe dates
USE-RELATIVE-TIME	May leave out date and start time information for a subsequent event, if events are happening one after the other
NO-END-TIME	May refer to events only by their start time.
REFER-ONLY-BY-NAME	May refer to an event only by its title (e.g., when an event is moved)
USE-SIMPLE-SUBJECT	May use the simpler description for an event.

Verbosity The first mechanism aims at influencing the length of an utterance unit by making it either more or less verbose. The idea is that actual language use of human speakers does not adheres to idealised principle such as ‘textual economy’ (Stone et al. 2003). This is not only the case for reasons of cognitive constraints during speech production, but also because words and phrases that do not contribute much to an utterance’s semantics can serve a function, for example by drawing attention to specific aspects of an utterance or by giving the listener time to process.

To be able to vary utterance verbosity, we annotated the linguistic resources in our system with values of their verbosity (these are hand-crafted similar to the rule’s annotation with production costs). During generation in MPP the values of all linguistic resources used in a (provisional) utterance are added up and used as one factor in SPUD’s heuristic function. When comparing two provisional utterances that only deviate in their verbosity value, the one that is nearer to a requested verbosity level is chosen. Depending on this level, more or less verbose constructions are chosen and it is decided whether utterance units are enriched with additional words. Table 6.3*c*–*e* show the differences in surface forms of IMPT$_4$ (see table 6.2), generated with three different levels of verbosity.

Redundancy The second adaptation mechanism is redundancy. Again, redundancy is something that an ideal utterance does not contain and by design SPUD penalises the use of redundancy in its heuristic function. Two provisional utterances being equal, the one exhibiting less redundancy is normally preferred. But similar to verbosity, redundancy serves communicative functions in actual language use. It can highlight important information, it can increase the probability of the message being understood, and is often used to repair misunderstanding.

In incremental microplanning, redundant information can be present both within one utterance unit (e.g., *tomorrow, 27 March, …* vs. *tomorrow …* or across IMPTS. For the former case, we modified SPUD's search heuristic in order to conditionally either prefer an utterance that contains redundant information or an utterance that only contains what is absolutely necessary. Table 6.3*a*,*b* show the differences in surface forms for this. In the latter case, redundancy only becomes an option when later IMPTS enable the choice of repeating information previously conveyed and therefore already established as shared knowledge. This is controlled via the internal structure of an IMPT and thus decided on the level of MCP.

In the following section we briefly show how adaptive incremental generation in SPUD_{ia} works together with speech synthesis.

6.3 ADAPTIVE INCREMENTAL SPEECH SYNTHESIS AND BEHAVIOUR REALISATION

In Buschmeier et al. (2012), we integrated SPUD_{ia} with the incremental speech synthesis component INPRO-ISS (Baumann and Schlangen 2012a)[83], which, by using just-in-time processing, supports changes to 'unspoken' parts of an ongoing utterance. In order to provide some right context (i.e., lookahead), which is important for synthesising coherent inter-unit intonation, SPUD_{ia} generates the second unit directly after the first. INPRO-ISS then synthesises both units. Shortly before the first has been fully spoken the third utterance unit is incrementally generated to serve as right context for the (re-)synthesis of the second unit, and so on. In this approach the increment size can be kept to a single utterance unit.

In an evaluation of the integration of SPUD_{ia} and INPRO-ISS we could show (Buschmeier et al. 2012, § 6.1) that incremental natural language generation and speech synthesis can significantly reduce the response time (time between generation start and speaking start) from an average of 1582 ms if both components operate in non-

[83] INPRO-ISS is based on the MARY text-to-speech system (Schröder and Trouvain 2003).

incremental mode to 271 ms if they operate incrementally and interact in the way described above[84] (Buschmeier et al. 2012, tbl. 2).

We also evaluated the responsiveness and adaptivity of the system in situations where information presentation was randomly interrupted by noise bursts that masked the speech signal (ibid., § 6.2). The evaluation compared system utterances generated and synthesised in three conditions: (A) non-incremental non-adaptive speech production that does not respond to noise at all, (B) non-incremental non-adaptive speech production that pauses upon noise detection and resumes afterwards, and (C) fully incremental adaptive speech production that pauses at the next word boundary when noise is detected and resumes speaking after the interruption by regenerating and re-synthesising the interrupted utterance unit with altered adaptation parameters. Twelve participants rated 27 randomly noise-interrupted system utterances (nine from each condition) for human-likeness on a seven-point Likert scale. The incremental and adaptive behaviour, condition (C), was rated statistically significantly more human-like than the behaviours of the systems in conditions (A) and (B) between which no difference could be found (see ibid., § 6.2, for details of the analysis).

The method to articulate the incrementally generated utterances implemented in the attentive speaker agent, is less sophisticated than the one described in this section. The agent synthesises each utterance unit in isolation, which results in less than optimal utterance intonation. The interplay between incremental generation and behaviour realisation is described in section 8.3.5.

6.4 SUMMARY AND DISCUSSION

In this chapter we developed and described the natural language generation component SPUD$_{ia}$, which is able to generate utterances in increments of the size of utterance units, and can adapt these increments as well as the structure of the utterance in real-time during generation.

This approach is efficient from a computational point of view since (i) generating only the first increment of an utterance uses far fewer resources and is much faster than generating the complete utterance at once, and (ii) re-generation in face of feedback is usually limited to a small part of an utterance. Being able to adapt on this level of granularity often results in utterances that are adapted to the listeners' needs (in real-time) but often do not even show signs of adaptation (overt self-repair). Redundancy, e.g., may be introduced in the next increment even though it was initially unplanned.

[84] With a minor draw back in intonation quality (timing deviation 0.81 ms, pitch deviation 7.08 Hz), response time can be improved further if generation of the next utterance unit is deferred up until the first word of the current utterance unit has been spoken (Baumann and Schlangen 2012b).

Adapted generation is thus similar to adapted human language production, which can also alter yet-unspoken parts of an utterance before articulation.

So far, the repertoire and complexity of the strategies and mechanism for adaptive language generation are quite limited in extent and should only be considered a first step towards the creative adaptation of utterances in speech production that human speakers are capable of. One aspect that should be improved in future work is that adaptation is heuristic, i.e., based on SPUD's heuristic function. Although SPUD evaluates different utterance alternatives, this comparison is done without explicitly reflecting on the likely effects and utilities of a specific adaptation. Ideally, adaptation would also be a coordinated action among different levels.

CHAPTER 7

FEEDBACK ELICITATION

In this chapter we address the third capability that an attentive speaker agent is supposed to have: being able to lead its interlocutors to provide as much communicative feedback as it needs in order to be well informed about their listening-related mental states. Based on the insight that the attributed listener state may, at times, not be informative enough to allow the agent to adapt its language production processes, we devise criteria for detecting such an 'informational need'. These are then used to decide when to generate explicit feedback elicitation cues.

7.1 SEEKING EVIDENCE OF UNDERSTANDING

As described in section 2.3.3, dialogue can be regarded as an ongoing process of collaborative hypothesis testing (Brennan 1990, pp. 30–33): a view on dialogue in which utterances are formulations of speakers' hypotheses about their interlocutors' common ground and responses to utterances are expected to provide evidence for (or against) the hypotheses that are being tested. Hypothesis testing plays a crucial role in dialogue, as it guides the speakers' language production process. If the hypotheses tested in an utterance (e.g., an assumption about something being part of the common ground) turn out to be true, because the listener signals understanding and acts accordingly, the speaker's communicative act can be considered successful and the previously uncertain grounding status of the propositions can be updated to reflect this new information. If, however, the hypotheses turn out to have been (partially) wrong, the speaker can provide extra information and then reformulate her utterance in an adapted way. Brennan (ibid., pp. 77–79) notes that grounding is an incremental (possibly even continuous) process and thus hypothesis testing can be done incrementally as well. If an interlocutor provides evidence of understanding while a speaker's utterance is

⋆ This chapter contains material previously published in Buschmeier and Kopp (2011; 2014).

still ongoing, they might immediately take this updated information into account and adapt.

Speakers are seeking evidence of understanding from their interlocutors, who are usually willing to provide this information and even take the initiative, as — for cooperative interactions — this behaviour is in the interest of both dialogue participants as both are responsible for dialogue success and such cooperation makes the overall dialogue more efficient (Clark and Krych 2004). Being the ones that test their hypothesis overtly, speakers continuously 'monitor[...] addressees for understanding' (ibid.), on multiple levels, and take evidence from different sources into account (e.g., verbal contributions, but also their speech-accompanying gestures and, in task-oriented settings, their task-related actions). Communicative feedback is especially well suited for incremental hypothesis testing, because it allows to provide evidence of understanding (inter alia) in overlap to ongoing utterances.

If the evidence provided by interlocutors, however, is insufficient for verifying or refuting a hypothesis — e.g., because they are not particularly active listeners or only provide feedback of limited informativeness — speakers may pro-actively seek evidence of understanding in the form of communicative listener feedback (ibid., p. 64). This can be achieved by producing feedback elicitation cues (e.g., Gravano and Hirschberg 2011).

Here, we propose that one factor in determining *when* to elicit feedback from an interlocutor are a speaker's 'information needs'. At given points in the dialogue, an attentive speaker agent may be sufficiently certain of a human interlocutor's listening-related mental state. In these cases, additional feedback by the interlocutor might not actually be informative. In other situations, however, the agent's uncertainty about an interlocutor's listening-related mental state may not warrant well-grounded choices in language generation, or may even be completely unknown. Furthermore, when choices for strategies and mechanisms for adaptive generation are limited, the agent needs to know in which listening-related mental state — of a number of the states it knows how to deal with — the interlocutor is most likely to be found. Given that such information needs occur, eliciting feedback from the human interlocutor is one strategy to ensure and achieve an effective dialogue.

In the following, we present a model that enables artificial conversational agents to determine *when* to elicit feedback by assessing their information needs about their human interlocutors' listening-related mental state when processing an utterance. This model informs the agent's decision making for the generation of feedback elicitation cues.

7.2 INFORMATION NEEDS AND FEEDBACK REQUESTS

An assumption commonly made in research on backchannels and communicative feedback is that listeners in dialogue produce feedback, at least partly, in response to behavioural 'elicitation cues' by their interaction partners[85]. These cues have been analysed extensively. It has been found that acoustic features (Gravano and Hirschberg 2011; Koiso et al. 1998; Ward and Tsukahara 2000), syntactic information (Gravano and Hirschberg 2011; Koiso et al. 1998), gaze (Bavelas et al. 2002), as well as head gestures (McClave 2000; Heylen 2006) play a role in eliciting feedback responses from listeners. The mechanism used to identify feedback elicitation cues used in these studies, however, is problematic for two reasons. Firstly, only cues that were actually followed by listener feedback were analysed (i.e., only those cues to which listeners responded). Secondly, speech that preceded listener feedback signals was assumed to contain a cue (i.e., the possibility that the listener produced the feedback signal without being cued by the speaker is not accounted for). Consequently, these types of analyses miss some of the cues that speakers actually produced, while categorising behaviours as a cue that may not have been intended as such.

These problems have been addressed by having multiple listeners respond to the same speaker behaviour in either a 'parasocial interaction' setting (Huang et al. 2010) or by creating the illusion of being in a one-on-one interaction with the speaker for more than one listener simultaneously (de Kok and Heylen 2012). These methods seek to remedy the first problem by increasing the range of available cues (different listeners responding to different cues). Similarly, the second problem may be remedied by clustering feedback (places in the speaker's speech that are followed by feedback signals from multiple listeners are more likely to contain a cue). Nevertheless, the form-features in feedback elicitation cues have proven informative enough to enable automatic detection of feedback elicitation cues in audiovisual data-streams and have been successfully used to model the feedback behaviour of virtual agents (Morency et al. 2010; Schröder et al. 2012).

A different line of research has shown that artificial conversational agents producing synthetic feedback elicitation cues while speaking, actually received feedback responses from their human interaction partners. Elicitation cues were either generated using an HMM-based speech synthesis system trained on a corpus of acted speech containing elicitation cues at interpausal unit (IPU) boundaries (Misu et al.

[85] It should be noted that communicative feedback serves functions for listeners as well, e.g., they can signal comprehension problems early on so that speakers can address them before they get worse. Such cases of feedback may not be a response to an elicitation cue of the currently speaking interaction partner.

2011b; Misu et al. 2011a), or by adding prosodic and non-verbal cues to the behaviour repertoire of a virtual agent (Reidsma et al. 2011).

What is not proposed by either of these two approaches — nor in the general literature on feedback — is a theory of *when* and *why* speakers produce feedback elicitation cues. Empirically, this is due to the problems involved in identifying elicitation cues as described above. From a theoretical point of view, cues are produced at different levels of intentionality. They can be fully intentional, e.g., when the speaker wants to know whether the listener understood what was said. They can also be produced by convention, e.g., by inviting a backchannel at the end of an IPU. Additionally, they can also occur purely coincidentally, e.g., a breathing pause by the speaker might be taken as a response opportunity.

In the following, we will concentrate on intentional feedback elicitation cues which are strategically produced by speakers with the aim of obtaining more — possibly new — information about their interlocutors' listening-related mental states (i.e., cues produced out of 'information needs'), most likely to reduce the uncertainty about the state of the dialogue.

Another common assumption is that communicative feedback and backchannels are one and the same, and that listeners, when giving feedback, merely communicate that speakers can continue speaking. Under this assumption, it would be sufficient for feedback elicitation cue placement to be governed by simple rules. As argued throughout this thesis, backchannels are, however, just one type of feedback. As feedback signals can be much richer in their form and often fulfil specific functions that go beyond the backchannel, strategically placing feedback elicitation cues in a turn can be used as a way of querying information from listeners.

7.3 CRITERIA FOR ELICITING FEEDBACK

Our assumption for modelling *when* speakers elicit feedback is just that. Feedback is elicited in situations where speakers have specific 'information needs' that can be fulfilled by listeners by providing feedback. When seeking to identify these information needs, both the attributed listener state at the current point in time as well as its history — how it developed into this state — are relevant. We propose three criteria that we consider useful for assessing whether an agent has an information need and should elicit feedback from its human interlocutor.

1. When its belief about the human interlocutor's listening-related mental state is not very informative (i.e., when the attributed listener state has high entropy[86]).

[86] The entropy of the probability distribution $\Pr(X)$ of a random variable X with values $x \in \mathrm{Val}(X)$ is

2. When its belief about the human interlocutor's listening-related mental state is static over an extended period of time (i.e., when no feedback was received).
3. When its belief about the human interlocutor's listening-related mental state is different from a desired mental state (e.g., sufficient understanding, high agreement) that is intended as the result of a specific communicative action by the agent or interactive adaptation in a previous utterance (i.e., when the attributed listener state diverges, by a given degree, from a given 'reference state').

A maximal uncertainty about the mental state of an interlocutor would manifest in a uniform probability distribution across the values of (one or more) variables, e.g., when $\Pr(U) = (⅓, ⅓, ⅓)^T$. Conversely, uncertainty would be minimal in a maximally skewed, degenerate distribution such as, e.g., $\Pr(U) = (0, 0, 1)^T$. This way of measuring uncertainty, i.e., in terms of entropy, assumes that the underlying state of the interlocutor is of a discrete nature, rather than fuzzy and with considerable variance persisting over time. We therefore combine the first, entropy-based, criterion with an operationalisation of the third criterion by quantifying the distance between the probability distributions of the current state of a variable and a 'reference state' such as, for example, a state that represents very good or very bad understanding. We measure this difference using the 'Kullback-Leibler divergence'[87].

Figure 7.1 shows a worked example[88] of how the Kullback-Leibler divergence between the current ALS-variables P, U, and AC and positive and negative reference states of these variables[89] changes over time (fig. 7.1*b*), alongside the temporal dynamics of the variables themselves (fig. 7.1*a*). In the example, the dynamic attributed listener state network receives feedback of positive understanding as input at t_1, and

defined as

$$E[X] = -\sum_x \Pr_X(x) \cdot \ln \Pr_X(x).$$

Entropy is a scalar value, which we use to quantify uncertainty, with which it is positively correlated.

[87] The Kullback-Leibler divergence (Kullback and Leibler 1951) of two discrete probability distributions P and Q is defined as

$$D_{KL}(P \,||\, Q) = \sum_i P(i) \cdot \ln \frac{P(i)}{Q(i)}.$$

It is a scalar value greater or equal to zero, with $D_{KL}(P \,||\, Q) = 0$ for $P = Q$, i.e., the more similar the two distributions P and Q are, the smaller their KL-divergence.

[88] The underlying Bayesian network model as well as methods for assessing the criteria for information need used in this example are archived and available at DOI: 10.6084/m9.figshare.4725538 .

[89] For this example we define the positive reference state as $\Pr(X^+) = (0.01, 0.3, 0.69)^T$ and the negative reference state as $\Pr(X^-) = (0.69, 0.3, 0.01)^T$, for $X \in \{P, U, AC\}$. ¶ The implementation of the attentive speaker agent uses slightly less skewed reference states, namely $\Pr(X^+) = (0.01, 0.4, 0.59)^T$ and $\Pr(X^-) = (0.59, 0.4, 0.01)^T$.

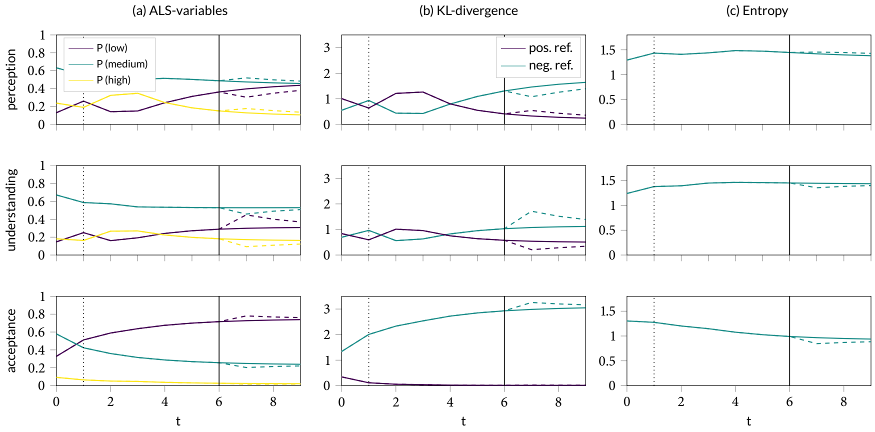


Figure 7.1: (a) Temporal dynamics of the speaker's degrees of belief in the ALS-variables P, U, and AC in a simulated feedback condition where the listener provides understanding feedback of medium certainty at t_1 (visualised by the dotted vertical line), simultaneously gazing near the target object until t_2. (b) Kullback-Leibler divergence between the distribution of the ALS-variables and the positive/negative reference states. (c) Entropy of the ALS-variables. The solid vertical line at t_6 visualises a condition where the speaker can elicit feedback. Dashed lines show how the speaker's degrees of belief would develop when the listener immediately responds with non-understanding feedback of medium certainty while gazing towards the speaker.

the information that the interlocutor gazes near the target object until t_2. No more feedback is received and fed into the network after this. The plots of the KL-divergence show that understanding is believed to be mediocre with a tilt towards low understanding and with some volatility at the beginning when feedback is received. The difference between the distributions of the variable U and the positive and negative reference distributions is not very large, however. In contrast, perception clearly moves toward low, and acceptance is believed to be skewed towards low almost from the beginning. The KL-divergence with the negative reference distributions is almost 0.

Based on this, we can assess the speaker's information needs by looking for points where (1) the KL-divergence to a positive reference distribution (representing an ALS with sufficient certainty and positive listener attributes) has a value higher by a given amount α than what is desired (criterion 3),

$$D_{KL}\big(\Pr(U_t), \Pr(X^+)\big) > \alpha, \quad \alpha = 1.0$$

and (2) where changes in the KL-divergence from one step to the next are smaller than a given value δ, i.e., when the values converge and the belief state becomes almost static (criterion 2):

$$D_{KL}\big(\Pr(U_{t-1}), \Pr(X^+)\big) - D_{KL}\big(\Pr(U_t), \Pr(X^+)\big) < \delta, \quad \delta = 0.1$$

Our model regards these as points where a speaker requires new information in order to know how to deal with the dialogue situation.

Applying these criteria to the example in Figure 7.1 to assess a point in time where feedback should be elicited, we find that they match at time t_6 with $\alpha = 1.03$ and $\delta = 0.077$. Figure 7.1 visualises two contrasting situations in the development of the belief state: (i) a feedback elicitation cue is produced, to which the listener responds with feedback of negative understanding (dashed lines), (ii) or no elicitation cue is produced and no feedback is received (solid lines).

7.4 ELICITING FEEDBACK

In the attentive speaker agent, we use a decision mechanism based on criteria 3 and 2. After an utterance unit (Traum and Heeman 1997; see section 6.2.2) produced by the agent has been communicated we compute the difference of the posterior marginal probability distribution of the ALS-variable U from a positive reference state as well as its dynamics. If the computed values exceed the thresholds for α and δ[90] the behaviour generation components immediately plan and realise an explicit multimodal feedback elicitation cue that consists of a specific gaze behaviour (see section 8.3.6), a pause, and possibly a verbal cue (see section 8.3.4 for details of the implementation of this mechanism). This combination of multiple cues increases the likelihood that an agent's human interlocutor responds with feedback.

In general, planning and realisation of feedback elicitation cues should take into account that different types and forms of cues are likely to elicit different types of feedback. Cues should thus be chosen based on the type of information need, i.e., the feedback type, that the agent wants to elicit, taking into account that not all elicitation cues may be applicable in a certain utterance context or dialogue situation (e.g., some utterances might not be suitable to be followed by *okay?*). This choice mechanism could be realised with the help of a probabilistic mapping from requirements to elicitation cues, as illustrated in fig. 7.2.

[90] The thresholds in the implemented attentive speaker agent are set to $\alpha = 0.85$ and $\delta = 0.2$.

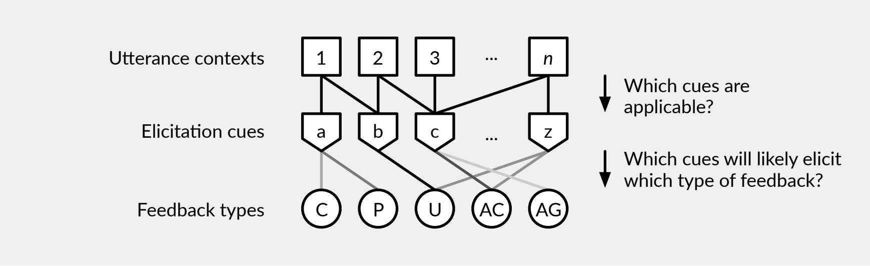

Utterance contexts
1
2
3
...
n
Elicitation cues
a
b
c
...
z
Feedback types
C
P
U
AC
AG
Which cues are applicable?
Which cues will likely elicit which type of feedback?

Figure 7.2: Probabilistic mapping of elicitation cues $\{a, b, \ldots, z\}$ that are applicable in certain utterance contexts or dialogue situations $\{1, 2, \ldots, n\}$ to the feedback types *C*, *P*, *U*, *AC*, and *AG* with which listeners are likely to respond to these cues.

7.5 SUMMARY AND CONCLUSION

In this chapter we addressed an attentive speaker agent's capability to elicit feedback from its human interlocutors. In contrast to most accounts of feedback elicitation, which deal with the form of elicitation cues and how they can be automatically detected in human speech, the model developed here pursues the idea that feedback elicitation has an underlying motivation, namely that speakers need information (e.g., evidence of understanding) from their interaction partners in order to be able to verify or refute the hypotheses that they test with their utterances, and to be able to adapt their language production to their interlocutors' needs. We defined the concept of 'information need' in relation to the representation of listening-related mental states that speakers attribute to interlocutors and specified three criteria that can be used to assess whether an information need is present on any of the levels of processing. This model can serve as the basis for an attentive speaker agent's decision making of when to elicit feedback, i.e., when to generate a behavioural cue.

The criteria for assessing information needs that were presented should not be regarded as definite or even exhaustive. The proposed criteria, based on simple thresholds, should be seen as a first step towards a causal model for feedback elicitation. They are nevertheless sufficient to be applicable in an implemented attentive speaker agent, as will be described in the following chapter. In a more advanced model of feedback elicitation, the criteria should probably differ by level of processing and co-vary with Clark's (1996) grounding criterion.

PART III

EVALUATION

CHAPTER 8
BRINGING IT TOGETHER: AN ATTENTIVE SPEAKER AGENT

In this chapter we describe how we bring together the three models for mental state attribution, interactive adaptation, and feedback elicitation in an artificial conversational agent that should be able to act as an attentive speaker. We first provide a general overview of the agent's information processing architecture, focussing on aspects of behaviour generation, and argue that incremental processing is a necessary requirement for being able to generate adaptive multimodal behaviour in real-time. We present the approach to incrementality that we adopt, and discuss why and how the behaviour generation architecture of the attentive speaker agent deviates from standard approaches to behaviour generation. Following this, we describe how the components in our architecture coordinate on the behaviour planning and realisation task and how the attributed listener state interfaces with these components. We close with a description of and motivation for the scenario in which the agent interacts with the user.

8.1 OVERALL MODEL AND ARCHITECTURE

The overall architecture for the attentive speaker agent includes two information processing streams: behaviour generation and input processing, linked via representations of the dialogue information state. Figure 8.1 shows an overview of the model, its components and architecture.

⋆ This chapter contains material previously published in Buschmeier and Kopp (2011) and Kopp et al. (2014). The incremental processing middleware IPAACA, described in section 8.2.2, was designed and developed in cooperation with Ramin Yaghoubzadeh, with contributions from Herwin van Welbergen, Sebastian Kahl, and others. The dialogue framework, described in section 8.3.4, was also designed and developed in cooperation with Ramin Yaghoubzadeh.

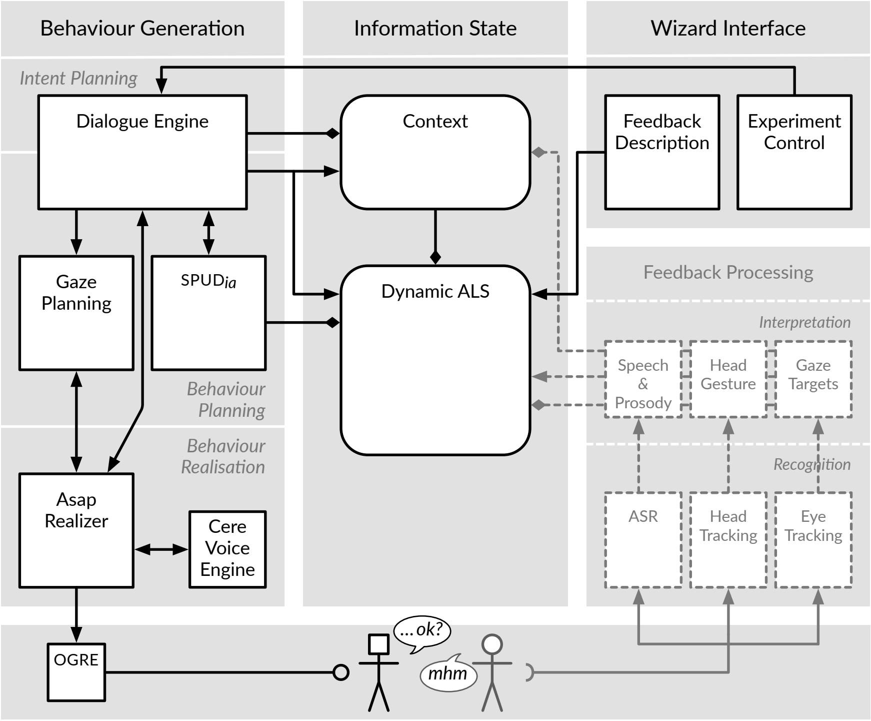


Figure 8.1: Overview of the architecture for the attentive speaker agent consisting of two processing branches (behaviour generation, and feedback processing, respectively wizard interface) and an intermediate representation of information state. Arrows between components visualise directed data flow. Arrows that end in diamonds visualise that data is read from the representation that it attaches to.

The behaviour generation stream of the architecture consists of five components. A 'Dialogue Engine' that manages and processes the agent's agenda (what are the topics to speak about) and decides upon the agent's overall behaviour and timing (when should the agent speak, which message should it convey, when should it elicit feedback, etc.). The dialogue engine is part dialogue manager and part behaviour planner. The behaviour generation stream further consists of a 'Gaze Planning' component (see section 8.3.6 below) which plans the agent's gaze behaviour, and of the natural language generation component 'SPUD$_{ia}$' (see section 6.2.3). Finally, the behaviour generation branch contains the behaviour realisation component 'AsapRealizer' (Reidsma and van Welbergen 2013; van Welbergen et al. 2014), which, in turn, makes use of the 'CereVoice Engine' for speech synthesis (by CereProc Ltd, Edinburgh, UK; Aylett and Pidcock 2007), and drives the 'OGRE' 3D rendering engine, which is used to render the virtual conversational agent's body in real-time computer graphics (OGRE Team 2013).

The feedback processing stream of the architecture, greyed out in fig. 8.1, is not part of the evaluated attentive speaker agent. Recognition and processing of multimodal interlocutor feedback signals (speech and prosody, head gestures, gaze patterns) were explored at the sideline of this thesis (see section 8.4.3), but did not yield results that were reliable enough for actual interaction with the agent. To nevertheless be able to evaluate the attentive speaker agent's capabilities to interpret listener feedback in the dialogue context and adapt ongoing utterances to the listeners' needs, we decided to use the 'Wizard-of-Oz' paradigm[91]. A human (the 'wizard') observes the human interlocutors during the interaction with the agent and enters their feedback signals into the system in real-time, using a graphical user interface (see section 9.3.5). The human interlocutors are not aware of the wizard and are made believe that they interact with a system that is capable of directly processing their behaviour.

The link between the two processing branches in the architecture consists of a representation of the dialogue context (see section 8.3.4) and a component that reasons about and represents the dynamic attributed listener state (ALS; see chapter 5 and section 8.4.2 below).

In the following, the individual components of the attentive speaker agent are described from a technical perspective, focussing on the type of information that is processed as well as the overall flow of information within the architecture. In preparation for this, the next section will motivate the choice of incremental processing

[91] The Wizard-of-Oz paradigm (Kelley 1983) is a common approach in research on human–agent and human–robot interaction, where the wizard is usually used to simulate input processing technology that is either not yet mature enough, not available at all, or not strictly necessary to answer the research question (Riek 2012).

as a principle of processing that underlies most of the components within the agent. This is followed by a discussion of current architectural approaches to behaviour generation and their limitations for incremental adaptive behaviour generation in real-time.

8.2 INCREMENTAL PROCESSING

Human speech production is based on incremental processing (Levelt 1989). For being able to participate in dialogue and conversation, incrementality of speech production alone, however, does not suffice. Many phenomena that occur in dialogue have a prerequisite of incremental processing on the comprehension side as well. There is solid evidence that even low-level comprehension processes in non-interactive language use are incremental and predictive (see, e.g., the line of research based on the 'visual world paradigm', Tanenhaus et al. 1995). This is the case in interactive settings as well (e.g., Tanenhaus and Brown-Schmidt 2008). A clear demonstration that interlocutors' behaviour is processed incrementally is also brought forth by high-level dialogue phenomena such as listeners' ability to produce feedback mid-utterance, or 'compound contributions' (utterances that are seamlessly split across interlocutors, see, e.g., Howes et al. [2011]).

8.2.1 INCREMENTAL PROCESSING FOR ARTIFICIAL CONVERSATIONAL AGENTS

Buß and Schlangen (2010, §§ 2.1, 2.2) review 'sub-utterance phenomena' in dialogue: linguistic feedback, hesitations, interruptions, turn-taking, and relevant non-linguistic actions. They come to the conclusion that — if the goal is to go beyond shallow processing — these phenomena can only be dealt with in an artificial conversational agent (both from an understanding and generation perspective) if it does incremental processing. Analysing prerequisites for fluid, real-time architectures for artificial conversational agents (Kopp et al. 2014, § 2), we came to the same conclusion, and, in addition, argued for tighter linking of input and output processing, as well as for bottom-up and top-down flow of information.

It can be concluded that the cognitive processes for language use work incrementally and that artificial conversational agents need to do incremental processing in order to be able to engage in natural, human-like dialogue with human interlocutors.

8.2.2 THE IU-MODEL

For the attentive speaker agent we adopt the perspective and terminology of the 'general, abstract model of incremental dialogue processing' developed by Schlangen and Skantze (2011) and implement the agent's components using our own implementation of this model ('IPAACA'; briefly described in Schlangen et al. 2010, § 3).

According to Schlangen and Skantze (2011), incremental processing can be modelled in terms of 'processing modules' and 'incremental units' (IUs). Processing modules consist of a left buffer, a right buffer, and a processing proper. The processing proper carries out the actual computations, and consumes and posts IUs from and to its buffers[92]. Processing modules form networks, in which, depending on topology, modules share IUs. If a processing module posts an IU to its right buffer, it is immediately present in the left buffer of processing modules that it is connected to. In IPAACA, the network topology is implicitly and functionally defined through categorisation of incremental units. Posted IUs are of certain categories. Processing modules that are interested in IUs of a category will find all IUs of this category in their input buffers and are notified of any changes in their buffers.

Once posted, an IU can be 'updated' by its owner (the component that posted it to its own output buffer) and also by those components that have the IU in their input buffers. Updates are immediately reflected in all buffers where the IU is present. Figure 8.2 illustrates various actions and operations of IUs in an implicit network of three IPAACA-components.

IUs consist of an identifier, meta-information, relation-information, and the actual payload. The piece of meta information that is specifically relevant here is the binary information of whether an IU is committed, or not (ibid., § 3.3.3). If an IU is set committed, it becomes immutable and thus cannot be updated further. As in Schlangen and Skantze's model, commitment in IPAACA is a technical concept, but it is also used in the attentive speaker agent as a signal that, from the moment of commitment onward, something is an unchangeable fact, e.g., that an utterance unit has been fully articulated (see below).

Relation information is used to build IU-networks from individual IUs. Schlangen and Skantze propose two types of relationships: the hierarchical grounded-in relationship and horizontal same-level-links (ibid., §§ 3.3.1 and 3.3.2). A grounded-in relationship from one IU to another means that the first is grounded in the latter, which can be used to make the flow of information traceable. Horizontal links can reflect various relationships, such as for example successor/predecessor relationships.

[92] In IPAACA, processing modules are called 'components'. They consist of one or more buffers. IUs are consumed via 'input buffers' and posted via 'output buffers'.

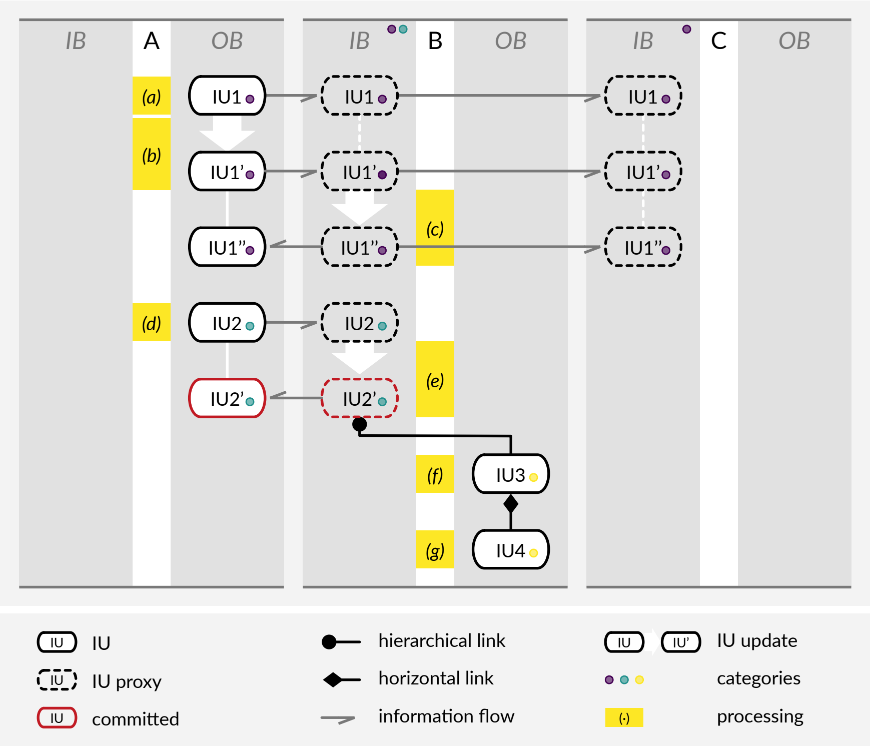


Figure 8.2: Illustration of three IPAACA-components A, B, and C and various IU-based operations over time (from top to bottom). Depending on the interests in IU-categories, IUs posted to the output buffer (*OB*) of a component, appear in the input buffer (*IB*) of other components. (a) IU$_1$, posted by component A, appears in the input buffers of components B and C. (b) IU$_1$ is updated by its owner (component A), and later (c) also by component B that has IU$_1$ in its input buffer. Updates are reflected in all buffers where IU$_1$ is present. (d) IU$_2$, also posted by component A, only appears in the input buffer of component B, as C is not interested in its category. (e) Component B set the meta-information committed, also an update, on IU$_2$. (f) IU$_3$, posted by B, is hierarchically linked to IU$_2$ in a 'grounded-in' relationship, which indicates that it is derived from IU$_2$. (g) IU$_4$, also posted by B, is horizontally linked to IU$_3$, which means that they are on the same level, for example in a successor/predecessor relationship.

In IPAACA both kinds of links are directed and realised with a unified mechanism. Both kinds of links are used in the attentive speaker agent.

The payload of an incremental unit holds the actual information that is transferred between processing modules. In IPAACA this information is encoded in a tree-like data structure that is flexibly composed of dictionary and list objects with numbers, strings, or boolean values as atomic elements (basically JSON-objects [Bray 2014]). Payload objects do not have a type. Their structure is defined based on convention.

All components in the attentive speaker agent are realised as IPAACA components and exchange information via IUs (and construct IU-networks). It should be noted, though, that not every IU passed between components contains an increment of information (in the strict sense), e.g., some contain control signals or commands. Components usually do not keep such information in their buffers once it has been processed.

8.3 BEHAVIOUR PLANNING AND REALISATION

In the following we describe how the attentive speaker agent's multimodal behaviour is incrementally and dynamically planned and realised in real-time. We begin this description with a brief review of the standard approaches to behaviour generation and then discuss the specific requirements that emerge from the need to be able to incrementally adapt behaviour to interlocutor feedback. We then discuss the attentive speaker agent's components for dialogue planning, natural language generation, and gaze behaviour planning and how they interact.

8.3.1 THE SAIBA-ARCHITECTURE FOR BEHAVIOUR GENERATION

The prototypical architecture for multimodal generation of behaviour for artificial conversational agents, 'SAIBA' is envisioned to consist of three broad levels: 'intent planning', 'behaviour planning', and 'behaviour realisation' with standardised interfaces (Kopp et al. [2006, § 3]; see fig. 8.3). Tasks for behaviour realisation are specified as documents in Behavior Markup Language (BML, Kopp et al. 2006; Vilhjálmsson et al. 2007; BML Committee 2011). With multiple realisers implementing BML, it has reached a level of maturity and stability such that it can be considered a widely adopted standard (van Welbergen et al. 2011).[93]

Tasks for behaviour planning are specified to be documents in Functional Markup Language, but standardisation efforts are still ongoing (Heylen et al. 2008; Cafaro et al.

[93] There exist local dialects that extend (or deviate from) the BML-core language, such as, for example, BMLA at Bielefeld University (Kopp et al. 2014; van Welbergen et al. 2014).

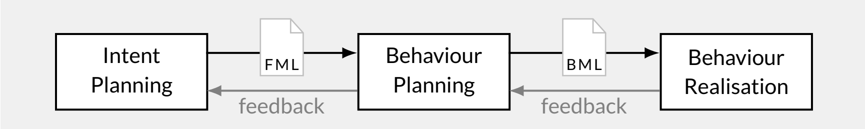


Figure 8.3: Macroscopic schema of the SAIBA architecture for multimodal behaviour generation in artificial conversational agents (redrawn from Kopp et al. 2006, fig. 1). The interface between levels is defined in terms of the markup languages FML and BML, which specify the documents in which information is passed from higher to lower levels (black arrows). Lower levels can pass information (as feedback) to higher levels (grey arrows).

2014). This seems to be the case because, fundamentally, the extent of FML is less clear than the extent of BML. It is not obvious which aspects of a behaviour to be planned result from an intention (i.e., from intent planning) and which are unconscious or merely contextual (Cafaro et al. 2014). Additionally, behaviour planning is potentially much broader in scope than behaviour realisation. A functional markup language would need to comprise specifications of disparate aspects of communication, e.g., mental and emotional states, context, communicative actions and their propositional content, person characteristics, goals on different levels (ibid., p. 82). Some of these aspects are research areas in their own right, with competing, ununified (or not even unifiable) theories that would need their own specification languages. BML allows for an extension with sub-languages (e.g., for the description of facial expressions or manual gestures [BML Committee 2011]) and an FML specification would likely need to rely on extensions to an even greater extent.

In general, standardised interfaces such as FML and BML are useful to make components from different research groups easily exchangeable. For individual endeavours they can, however, be limiting in terms of flexibility.

8.3.2 REAL-TIME GENERATION OF MULTIMODAL BEHAVIOUR

Even though early architectures of embodied conversational agents already specified real-time requirements similar to the ones needed for the attentive speaker agent (incremental processing and real-time reactivity) and also relied on multiple coordinating components that operated on different layers (e.g., the ‘Ymir’ architecture; Thórisson

1996, §§ 7–8), features that can be considered fundamental to real-time behaviour realisation — such as the ability to incrementally add and dynamically change ongoing behaviours — are not part of the BML core standard, but come as an extension that is currently only available in AsapRealizer (BMLA, Kopp et al. 2014; van Welbergen et al. 2014)[94].

But even this extension does not answer the question where incrementally produced behaviours on multiple modalities generated from independent components that operate on different time scales (e.g., different increment sizes) come together. Current non-incremental approaches to multimodal behaviour generation, that are in line with the SAIBA-framework, produce BML behaviour specifications that are fully specified and fused on the behaviour planning level.

The 'Nonverbal Behavior Generator' (Lee and Marsella 2006; Wang et al. 2013), for example, is a single component that plans multimodal non-verbal behaviour based on an FML specification that already includes the surface text to be spoken (the generated behaviour is based on an analysis of this text) and produces a BML behaviour specification to the realiser. The 'GeNetIc' generator for speech-accompanying iconic gestures (Bergmann and Kopp 2009), on the other hand, comprises multiple interacting components on the intent and behaviour planning levels (for natural language generation and for gesture generation), which enables modelling of the complex interactions found in speech–gesture production, e.g., differences in the semantic coordination of these modalities based on linguistic encoding patterns or on cognitive load (Bergmann et al. 2013). Despite being produced in independent components, a single BML behaviour specification is produced within the behaviour planning level and sent off to be realised.

In BML, behaviours are coordinated in relation to synchronisation points which can be located at arbitrary time points. In BMLA, sync-points can be used to coordinate behaviours that are specified across multiple independent BML blocks. Increments of behaviour ('chunks' in BMLA-parlance) can be composed through various operations, even allowing for co-articulation (that is, blended transitions). BMLA also allows for 'pre-planning' of behaviours that are then ready for immediate realisation as soon as they are needed.

[94] AsapRealizer — developed at the Social Cognitive Systems group, Bielefeld University, and at the Human Media Interaction research group, University of Twente — is a successor of the BML-realiser 'Elckerlyc' (van Welbergen et al. 2009; van Welbergen 2011) that, inter alia, integrates principles for continuous adaptation of ongoing plans from the 'articulated communicator engine' (ACE, Kopp and Wachsmuth 2004). AsapRealizer is BML 1.0-compliant (van Welbergen et al. 2011) and goes beyond the core standard with the extensions BMLT (van Welbergen et al. 2009), BMLA (Kopp et al. 2014; van Welbergen et al. 2014), and MURML (Kranstedt et al. 2002).

These mechanisms of BMLA make it possible to start articulating an utterance as soon as an opportunity to take the turn opens up. They also enable the realisation of various other dialogue phenomena. Still, they do not present a general solution for real-time generation and coordination of behaviours that are not pre-plannable, e.g., when they are responsive to user behaviour. Sync-points are not a general solution to such situations either, because they cannot be added on the fly — at least not when the behaviour is already in execution.

8.3.3 BEHAVIOUR GENERATION FOR ATTENTIVE SPEAKING

As the attentive speaker agent requires this kind of flexibility, standard behaviour coordination mechanisms present in BML and BMLA are not sufficient.

Behaviour planning and intent planning in the attentive speaker agent architecture are distributed across multiple components. One of these components, the 'Dialogue Engine', is responsible for multiple tasks, some of which fall on the intent planning, others on the behaviour planning level. Other components, such as natural language generation and gaze planning focus on specific tasks, but may also cut across planning levels. All components exchange information using specific IU-based protocols (instead of framing requests and answers in the FML structures as proposed by Cafaro et al. [2014]).

Behaviour realisation is done in the standard SAIBA way using 'AsapRealizer' (Reidsma and van Welbergen 2013; van Welbergen et al. 2014), which receives incremental units with a payload that contains a BML- or BMLA-specification. During articulation this IU is then incrementally updated with prediction and progress information that would normally be provided as BML-feedback messages (BML Committee 2011). The timing for real-time behaviours, where necessary, however, will not be left to the behaviour realiser alone, but are planned by the behaviour planning components.

The approach we chose here is to frame behaviour planning and realisation as an interactive process. The planning components act autonomously, but coordinate their actions in an interplay with behaviour realisation, the information state, and the interlocutors' actions. Each component is the expert on the planning problem it handles, but can integrate information from other components.

The architecture implements ideas from fluid real-time incremental behaviour generation which we presented in Kopp et al. (2014, fig. 3). Behaviours unfold incrementally on all levels. Higher levels are kept informed about the progress made on lower levels and act accordingly. Lower levels, in turn flexibly adapt their output when plans on higher levels change.

In the following three sections we present the individual components and their role in the attentive speaker agent's architecture.

8.3.4 DIALOGUE ENGINE

The dialogue engine is based on a simple framework for representing dialogue context, actions that can change this context, and rules that trigger these actions. An action and the rules that trigger it form a 'restricted action'. The framework is basically an implementation of the 'Information State Update' approach to dialogue management (ISU; Larsson and Traum 2000). The context corresponds to ISU's informational components and the restricted actions to ISU's update rules (ibid., pp. 324–325).

Restricted actions consist of a function and three conditions (a pre-condition on the context, a pre-condition on potential input, and a post-condition on proposed changes to the context; all three are optional) in form of boolean conditional expressions. The function of a restricted action can be called when both pre-conditions are met (or are unspecified). It has, by convention, no side effects and returns a proposal for a set of changes to the context. These changes are applied to the context if they meet the post-condition (if specified), otherwise they are discarded.

A 'dialogue model' in this framework consists of an ordered set of restricted actions (more specific — in terms of conditions — restricted actions take precedence over less specific ones) and a representation of the context. An active loop continuously evaluates the restricted actions against the context in each iteration (which enables the agent to be very responsive and able to adapt almost immediately upon relevant new information). The result of the first restricted action that meets its pre- and post-conditions (if specified) is applied to the context. The dialogue engine then acts based on the updated state of the context.

In general, the dialogue engine is informed about realisation progress and uses this information to incrementally plan upcoming behaviour on multiple levels. On the highest level it decides which topic to talk about with the interaction partner (topics are maintained in an agenda of things to do).

On an intermediate level it constructs the local structure of the conversation, making decisions that may take the interlocutors' conversational actions (such as their utterances or, via the ALS, their feedback) into account. After the agent finishes producing a presentation, for example, the dialogue engine decides — depending on some of the attributed listener state variables — whether to (i) continue with the next utterance (if understanding is sufficient), (ii) repeat the utterance (if understanding is insufficient), or (iii) explicitly ask which of the two alternatives the interlocutor prefers (if the agent's information about the listener's mental state of understanding is uncertain). When generating these transitions, the dialogue engine chooses, with a simple threshold model, from a set of pre-formulated utterances (see table 8.1) and realises them.

On the lowest level, the dialogue engine decides whether to elicit feedback via a behavioural cue and a pause or whether to generate the next increment of an ongoing utterance to continue speaking without interruption. In the following we focus on these low-level decisions.

Listing 8.1 shows a simplified version of the restricted action `ongoing.elicit` that is used to decide whether and when to produce a feedback elicitation cue in an ongoing utterance. It only consists of a context pre-condition and the action function `elicit_feedback`. The pre-condition is a lambda function that is a conjunction of four boolean conditionals: (i) the current 'state' of the dialogue needs to be `UTTERANCE::Ongoing`, (ii) the current realiser request IU is either being in execution (`IN_EXEC`), or execution is finished (`DONE`), (iii) given the current attributed listener state the agent assesses an information need on the understanding level (see chapter 7), and (iv) realisation of the current request is predicted to be (nearly) finished or finished. All four conditionals make use of the context object (`ctx`), which contains all relevant information on the current state of the dialogue, including, e.g., the attributed listener state.

The `elicit_feedback` action that is called proposes five changes to the context: (i) to produce a verbal feedback elicitation cue, (ii) to make a pause of (up to) 1.5 s, (iii) to change the current high-level gaze behaviour to 'elicit' (see section 8.3.6), (iv) to set the attributed listener state into a mode where it processes incoming evidence immediately (see section 8.4.2), and (v) it changes the state to `UTTERANCE::NeedFB`. The dialogue engine then processes the updated context within the same iteration, for example by posting a behaviour specification containing the chosen verbal feedback elicitation cue to the behaviour realiser.

The following section describes how natural language generation is integrated into the behaviour generation architecture and how the dialogue engine mediates between language generation and behaviour realisation in order to produce continuous and incrementally adaptive verbal behaviour for the agent.

8.3.5 NATURAL LANGUAGE GENERATION WITH SPUD*IA*

The incremental and adaptive natural language generation component SPUD$_{ia}$ (see section 6.2) is implemented as an IPAACA processing component as well. Figure 8.4 visualises the IU-based approach for incremental output generation in the attentive speaker agent that will be presented in this section.

Table 8.1: Pre-formulated utterances from which the dialogue engine chooses when transitioning between information presentation units. Depending on the type of calendar operation (see section 9.3.6) and the attributed listener state variables *U* and *AC*, the dialogue engine chooses one of the transition utterances (or, randomly, a slight variation). The table shows the original formulations (German) and their English translation.

operation	*U*	*AC*	Transition utterance
any	*high*		*Dann machen wir weiter* ‘Then let’s continue’
	low		*Mir scheint, Sie haben mich nicht verstanden. Ich wiederhole es noch einmal* ‘It seems to me that you haven’t understood me. I’ll repeat it again’
	uncertain		*Ich bin mir nicht ganz sicher, ob Sie mich verstanden haben. Soll ich es noch einmal wiederholen, oder weitermachen?* ‘I’m not really sure whether you understood me. Should I repeat it or should we continue’
		uncertain	*Ich kann Ihre Einstellung leider nicht ganz deuten* ‘I am afraid I can’t interpret your attitude’
propose		*high*	*Okay, der Vorschlag scheint Ihnen zu gefallen* ‘Okay you seem to like the suggestion’
		low	*Mir scheint, dass Ihnen dieser Vorschlag nicht so gut gefällt. Auch okay* ‘It seems to me that your do not like this suggestion. That’s okay’
cancel		*high*	*Okay, Sie scheinen einverstanden zu sein, dann streichen wir das* ‘Okay, you seem to accept it, we’ll cancel it’
		low	*Mir scheint, dass Sie diese Änderungen nicht gut finden* ‘It seems to me that you don’t like this change’
move		*high*	*Okay, Sie scheinen einverstanden zu sein, dann verschieben wir das* ‘Okay, you seem to accept it, we’ll move it’
		low	*Mir scheint, dass Sie diese Änderungen nicht gut finden* ‘It seems to me that you don’t like this change’

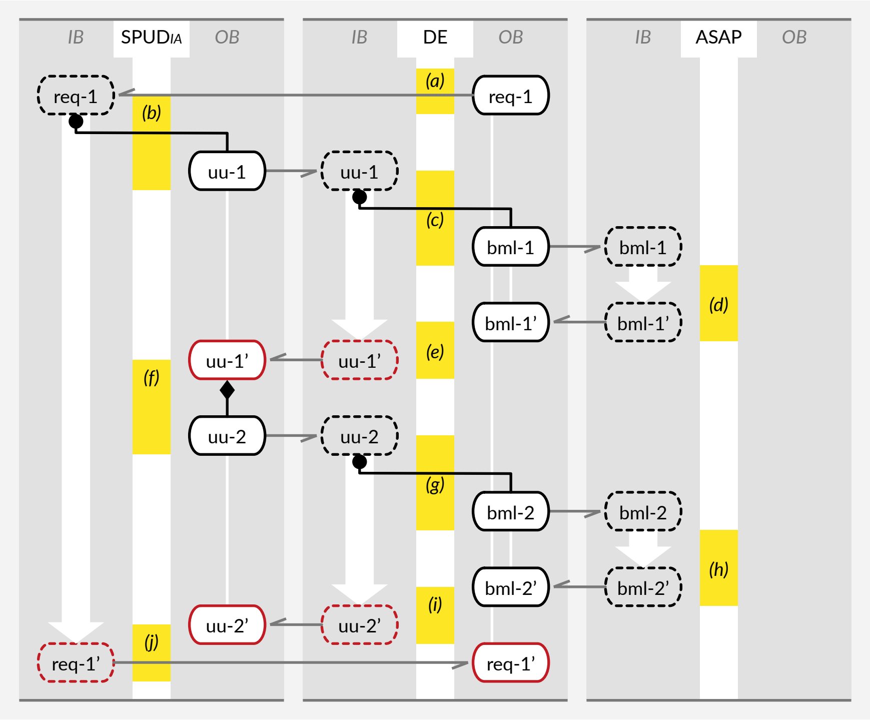


Figure 8.4: Incremental natural language generation (and realisation) architecture for the attentive speaker agent. Three IPAACA-components are involved: SPUD$_{ia}$, the dialogue engine, and AsapRealizer, a BML-realiser. (a) The dialogue engine posts an IU containing an NLG-request. (b) SPUD$_{ia}$ generates the first utterance unit grounded in this request and posts it as an IU. (c) The dialogue engine generates a BML block grounded in this utterance unit and posts it as an IU. (d) AsapRealizer updates this IU with status information and an estimation of when realisation will end. (e) When realisation is nearly finished the dialogue engine sets the utterance unit IU committed. (f) As soon as it is committed, SPUD$_{ia}$ starts generating the succeeding utterance unit. (g)–(i) analogue to (c)–(e). (j) When the last utterance unit of belonging to the request has been realised, the request IU is set committed. Adaptation is not shown as it happens within SPUD$_{ia}$, based on IUs consumed from the attributed listener state component, which is omitted here.

```
RestrictedAction(
  id='ongoing.elicit',
  context_precondition=lambda ctx: ctx['STATE'] == 'UTTERANCE::Ongoing'
    and ctx['REALIZERRQ']['current'].payload['status'] in ['IN_EXEC', 'DONE']
    and information_need(ctx, level='understanding', alpha=0.85, delta=0.2)
    and time.time() >= ctx['REALIZERRQ']['current'].payload['predEndTime'] - 0.3,
  input_precondition=None,
  output_precondition=None,
  action=elicit_feedback)

def elicit_feedback(ctx, inp, outp):
  outp['REALIZERRQ'] = random.choice(verbal_elicitation_cues)
  outp['FB-ELICITATION']['time_to_wait'] = 1.5
  outp['GAZE_STATE'] = 'elicit'
  outp['ALS_MODE'] = 'process-next-fb-immediately'
  outp['STATE'] = 'UTTERANCE::NeedFB'
```

Listing 8.1: Example of the restricted action that plans the production of a feedback elicitation cue, given that the agent has an information need on the level of understanding. Both the context pre-condition as well as the action function are simplified and adapted for presentational purposes.

SPUD$_{ia}$ generates utterances incrementally, utterance unit by utterance unit. In the attentive speaker agent, it becomes active when it consumes an NLG-request-IU (posted by the dialogue engine component, fig. 8.4*a*). Such requests consist of a specification of the type of speech act, as well as the information that is to be verbalised (as, e.g., in eqs. [6.1] and [6.2] on page 130).

The micro-content-planning level (MCP) creates a dynamic utterance plan (a list of incremental microplanning tasks) from the request and triggers microplanning-proper (MPP) into generating the first utterance unit from the first incremental microplanning task (IMPT) and posting it as an utterance-unit-IU. This IU is linked hierarchically to the NLG-request-IU in a grounded-in relationship (fig. 8.4*b*).

After posting an incremental unit containing an utterance unit (say, IU$_i$), SPUD$_{ia}$ becomes inactive until it is notified that IU$_i$ has been set committed (fig. 8.4*e*,*i*). It then provides the next incremental microplanning task IMPT$_{i+1}$ to the MPP-level for generation, the result of which is then posted as IU$_{i+1}$. IU$_{i+1}$ is linked horizontally to IU$_i$ in a successor relationship (fig. 8.4*f*).

SPUD$_{ia}$'s utterance unit IUs are consumed by the dialogue engine and their content is relayed as BML-IUs to the BML-realiser component ('AsapRealizer'; van Welbergen et al. 2014) for synthesis and realisation (fig. 8.4*c,g*). During realisation, the BML-realiser incrementally updates the BML-IUs with status information (feedback; BML Committee 2011), such as the stage of processing and a prediction of when articulation will end (fig. 8.4*d,h*). Towards an utterance unit's end of articulation — timing depends on whether feedback from the interlocutor will be elicited — the corresponding utterance unit IU is set committed by the dialogue engine component,[95] which re-activates SPUD$_{ia}$ and an utterance unit is generated from the next IMPT (see above).

When the last incremental microplanning task of an NLG-request has been committed, the NLG-request-IU, in which the whole utterance is grounded, is set committed as well (fig. 8.4*j*).

Adaptation of utterances is handled SPUD$_{ia}$-internally (see section 6.2.3), based on the attributed listener state. To be informed of the current state, SPUD$_{ia}$ consumes IUs of the attributed listener state component (section 8.4).

8.3.6 GAZE BEHAVIOUR

Finally, we describe how the attentive speaker agent's gaze behaviour is incrementally planned by an autonomous component that acts in coordination with behaviour realisation and other behaviour planning components.

As in many other computational models of gaze behaviour for embodied conversational agents (e.g., Pelachaud and Bilvi 2003; Lee et al. 2007), gaze is planned on two levels. Depending on the interactional context ((i) an interlocutor is present but not in contact, (ii) a conversation is ongoing, (iii) the agent is producing an utterance, (iv) the agent is eliciting feedback) provided by the dialogue engine, the gaze planning component chooses an appropriate high-level gaze strategy. These strategies activate continuously running low-level 'action patterns' (de Kok et al. 2015, § 3.3.2) that dynamically and incrementally generate BML gaze behaviours (effectively shifts of gaze direction using eye and head movement [BML Committee 2011]) that it posts as IUs to be realised incrementally and in real-time by the behaviour realisation component AsapRealizer.[96] Figure 8.5 shows examples of the gaze behaviour for the four high-level strategies.

[95] Utterance unit IUs are always set committed after realisation, regardless of interlocutor feedback. Importantly, commitment is independent of interlocutor feedback, attributed listener state, or estimated groundedness.

[96] AsapRealizer realises specified gaze shifts based on biological models for eye movement, head movement, and saccades (van Welbergen 2011, pp. 62–63).

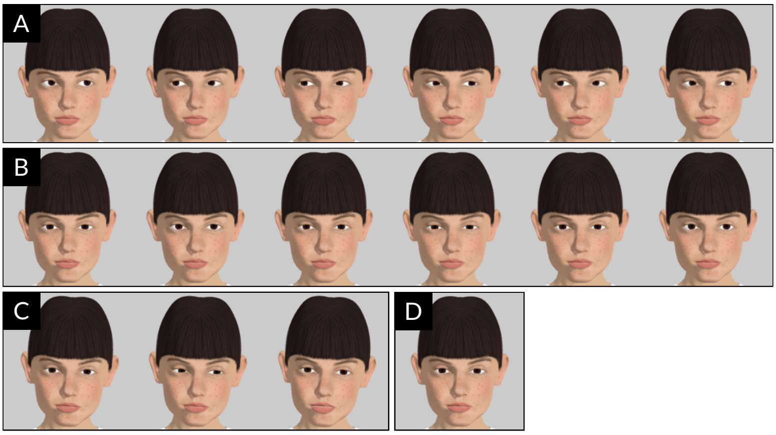


Figure 8.5: Gaze behaviour examples from the four high-level gaze strategies. *(A)* gaze wanders freely in the idle gaze strategy; *(B)* the attentive gaze strategy fixates different points estimated to lie on the interlocutor's face; *(C)* when articulating an utterance, gaze may be averted mid utterance unit and directed back at the interlocutor towards its end; and *(D)* when eliciting feedback, the interlocutor is fixated for a certain amount of time, or until she produces feedback (see listing 8.1).

Generation of low-level gaze behaviour is influenced bottom-up and in real-time by the agent's ongoing behaviour as well as the interlocutor's behaviour.[97] That is, it is not decided on up-front simultaneously with planning of the speech content of the utterance unit.

While the agent is articulating an utterance unit, for example, the gaze component may produce gaze aversion behaviour in order to display 'cognitive effort' (Andrist et al. 2013).[98] It also ensures that gaze is directed back at the interlocutor right before

[97] In the attentive speaker agent evaluated in chapter 9, the interlocutor's feedback behaviour is taken into account via the dialogue engine. In the context of the KOMPASS-project, action patterns in this gaze planning component are also aware of and responsive to the human interlocutor's gaze behaviour (which is tracked in real-time with 'EyeX', a consumer 'eye tracker' from Tobii AB, Danderyd, Sweden).

[98] As information on the internal structure of the utterance relevant for gaze behaviour (e.g., thema/rhema, Cassell et al. 1999; Heylen et al. 2005) is currently not available to the gaze planning component, gaze is averted randomly in ⅓ of all utterance units. When available, information from other behaviour planning components could easily be taken into account during gaze planning though.

articulation finishes in order to display its willingness and ability to receive listener feedback (Lee et al. 2002, § 2.2). Planning of this gaze-while-speaking behaviour is achieved in interaction with the behaviour realisation component that incrementally reports articulation progress as well as a prediction of when articulation will end. Gaze aversion is decided upon between ⅓ to ½ into the articulation of the utterance unit. When more than three quarters of the utterance unit has been articulated, gaze is directed back at the interlocutor. See fig. 8.5*C* for an illustration of the agent's gaze behaviour while articulating an utterance unit.

The description of these three components shows that incremental planning and realisation of multimodal behaviour with multiple behaviour planning components opens up many questions for the SAIBA framework, especially if behaviour planning is required to take into account the actions of the human interlocutor in real-time.

8.4 ATTRIBUTED LISTENER STATE

The theoretical computational model for dynamic listener state attribution, specified in chapter 5, is also implemented as an IPAACA processing component within Schlangen and Skantze's (2011) framework for incremental dialogue processing. The component consumes two types of incremental units: (i) IUs that contain evidence for the network, and (ii) IUs that contain control information for the module. As output the module posts incremental units that contain the current state of the attributed listener state. This results in a loose coupling of the attributed listener state to any of the other components in the attentive speaker agent system. Any other component can provide evidence to be used in the attribution process, which is important for example, for modelling contextual influences (see fig. 5.9). Similarly, information from the attributed listener state is readily available wherever it is needed (in the dialogue engine, in SPUD$_{ia}$, etc.) to make informed adaptation decisions.

8.4.1 COLLECTING EVIDENCE

The fundamental principle for feedback processing in the ALS-module is that evidence is asynchronously collected over a span of time after which the Bayesian network inference is carried out. Evidence is specified in form of pairs of a name of an evidence node and hard evidence for this node, i.e., a value that it can take. Feedback of function understanding, with positive polarity, for example, is specified as $\langle FB, u^{+} \rangle$. Evidence is not just restricted to feedback behaviour of the agent's interlocutor. Contextual factors, such as an utterance's estimated difficulty (see section 5.5) are also specified as evidence. Evidence is thus consumed from multiple processing modules in the attentive speaker agent.

The implication from the collection of evidence over a span of time is that only the latest evidence for each variable is used for inference. Consider the following example where an interlocutor provides feedback to one of the agent's utterance of medium difficulty. While looking at the agent, the interlocutor first provides a feedback signal of understanding with positive polarity but *high* uncertainty. Shortly afterwards, it provides a second feedback signal of understanding with *low* uncertainty. The ALS-module linearly consumes three incremental units containing evidence for the network:

$$\begin{aligned}\textsc{iu}_1 &: \langle DI, medium\rangle \\ \textsc{iu}_2 &: \langle FB, u^+\rangle, \langle GT, agent\rangle, \langle UNC, high\rangle \\ \textsc{iu}_3 &: \langle FB, u^+\rangle, \langle UNC, low\rangle\end{aligned}$$

$\textsc{iu}_1$ is posted by the dialogue engine, which has an estimate of the difficulty of a discourse unit based on the amount of information it contains. This is followed by $\textsc{iu}_2$ and $\textsc{iu}_3$, which are posted by the wizard after the first (respectively second) feedback signal has been perceived. When the collection time span reaches its end, the following hard evidence is set in the network prior to doing the inference:

$$\langle DI, medium\rangle, \langle FB, u^+\rangle, \langle GT, agent\rangle, \langle UNC, low\rangle$$

The fact that the interlocutor, for a certain period of time (at $\textsc{iu}_2$), was uncertain in her understanding is not known to the network when computing the ALS because $\langle UNC, low\rangle$ in $\textsc{iu}_3$ shadows $\langle UNC, high\rangle$ in $\textsc{iu}_2$.

The extent of the time span in which evidence is collected varies. During incremental generation and production of a discourse unit (see section 6.2.2) it encompasses a single increment, that is, one utterance unit. Evidence is collected when an utterance unit is being articulated and collection ends when the next utterance unit is ready to be articulated (see section 8.3.5, especially fig. 8.4), this basically happens exactly at the moment when articulation finishes. If the attentive speaker agent is in need of feedback and produces feedback eliciting behaviours (directing its gaze at the interlocutor, pausing articulation of the next utterance unit, producing a verbal elicitation cue) the evidence collection period extends beyond the end of articulation of the utterance unit. In this case it either ends as soon as feedback is perceived, or, when the interlocutor does not respond to the feedback elicitation cue, after a fixed amount of time (1.5 s).

The dynamic Bayesian network model (structure and parameters) used in the implemented and evaluated attentive speaker agent is the one used for the worked example discussed in section 5.7.2 (see fn. 73 on page 114).

8.4.2 PROCESSING EVIDENCE

Immediately after the evidence collection period ends, Bayesian network inference is carried out using the joint-tree algorithm using elimination trees (see fn. 59 on page 95). A factor tree for the network at time t is created, evidence is set, and the marginal posterior distributions of all ALS-network variables is computed. These are (i) used internally as prior feedback (Robert 1993) for the next time slice $t + 1$ of the dynamic Bayesian network (see section 5.7), and (ii) posted in form of an incremental unit that contains the posterior marginal distributions for each variable at time t, the ALS$_t$. This IU is horizontally double-linked in a successor/predecessor relationship to the IU containing ALS$_{t-1}$, which is set committed at this point.

8.4.3 A NOTE RECOGNISING LISTENER FEEDBACK SIGNALS

Signal processing of multimodal feedback signals is a hard research problem in itself that is not in the focus of this thesis. Recognition of verbal/vocal feedback signals, head gestures, and listener gaze were, however, explored at the sideline of the thesis-work.

Analysing isolated verbal/vocal feedback signals in the ALICO-corpus, we found that prosodic correlates of feedback function interact with the segmental and syllabic structure of feedback expressions (Malisz et al. 2012). Although some relationships could be identified in the data, the overall results are of limited usefulness for practical applications as classification experiments did not yield reliable results. We also attempted to classify short feedback expressions using standard speech recognition technology (Walker et al. 2004) with manually authored pronunciation lexica and grammars, but recognition error rates were too high for non-lexical feedback expressions such as *m*, *mhm*, etc.[99] In general, the research reviewed in section 3.3 shows that current approaches, while in principle being able to find relationships between form and function, are still experimental and not yet ready for use in interactive systems. Recent developments in the field of 'computational paralinguistics' (Schuller and Batliner 2014), however, yield promising results in detecting emotional and other cognitive states of the speaker in larger chunks of speech. This gives hope that, given

[99] Unpublished feedback recognition experiments by Hendrik Buschmeier and Ramin Yaghoubzadeh (2009/2010), as unpublished student work report by Oliver Ast (2014/2015), supervised by Hendrik Buschmeier.

more research, it will be possible to automatically extract features relevant to listening related mental states from short feedback expressions.

Concerning non-verbal embodied feedback, intermediate versions of the attentive speaker agent used a relatively robust and user-independent approach to online head gesture recognition, based on 'ordered means' sequential probabilistic models (Wöhler et al. 2010), that could recognise two types of head gesture movement (nods and shakes), as well as their absence. Extending this approach to a greater variety of head gesture movement types, however, resulted in reduced robustness and recognition of more nuanced differences in head gesture units such as number of cycles, amplitude, as well as a description of complex head-gesture units (see section 3.3.3; Włodarczak et al. [2012]) would have been impossible in that framework. In general, state-of-the-art approaches to conversational head gesture recognition are usually limited to recognising the movement types nod and shake (e.g., Morency et al. 2007; Chen et al. 2015), too.

8.5 PERSONAL CALENDAR ASSISTANT SCENARIO

To apply and test the proposed architecture, we chose a scenario in which the attentive speaker agent is a 'personal assistant' for its human user.[100] The domain for the interaction and topic of interaction is the user's calendar, that is, her appointments, changes to these appointments, recommendations of activities, and so on. The user takes the role of the agent's conversation partner (see fig. 8.6), who — in the evaluation study in chapter 9 — listens to the agent presenting[101] information that is related to her calendar and is able to provide feedback, which the agent can interpret and adapt its ongoing behaviour to.

In addition to being of real-world significance,[102] the calendar domain was chosen for several reasons:

- The basic domain — disregarding the actual content of appointments — is task-oriented, highly structured, and thus relatively well defined (the characteristics

[100] The choice was inspired by the 'Knowledge Navigator' concept developed in 1987 at Apple Computer Inc. (Cupertino, CA, USA). See Scully (2010) for its history.

[101] In the scope of the KOMPASS-project (2015–2018; https://purl.org/scs/kompass; Yaghoubzadeh, Buschmeier et al. [2015]) the calendar assistant agent is — to a certain extent — able to discuss calendar-related topics with its user.

[102] Calendaring is a somewhat natural task domain for personal assistants. Within the large-scale project CALO ('Cognitive Assistant that Learns and Organizes', 2003–2005, led by SRI International [Stanford, CA, USA]) assistance in task and time management was one central research direction (Myers et al. 2007). Apple Inc.'s personal assistant 'Siri' — which, one day, might have functionalities similar to the Knowledge Navigator, see fn. 100 — was a spin-off from the CALO-project.

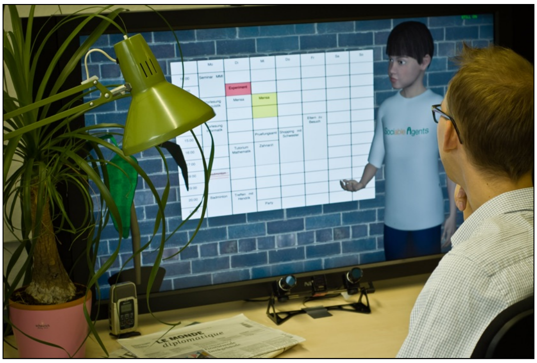


Figure 8.6: Illustration of the personal calendar assistant scenario.

of dates, times, intervals, calendar operations, etc. are almost formal in nature). Because of this, a relatively small and manageable language fragment should be sufficient for the generation of the agent's utterances as well as for the interpretation of the user's utterances.

- Despite the domain's simplicity, misunderstandings are likely to happen, e.g., due to differences in knowledge between agent and user, ambiguous references, etc. Even if misunderstandings seem absent, ensuring understanding between agent and user is important in this domain as there could be undesirable real-world consequences for the user — e.g., missing an important appointment — if understanding is attributed falsely. Being able to attribute other listening-related mental states (e.g., acceptance, agreement) is important in this domain, too. Thus we expect the domain to yield a large amount of listener feedback spanning the rich form–meaning spectrum (sections 3.3 and 3.4).
- Most humans have appointments. Calendars are commonly used tools for recording, planning, reminding, etc. of appointments. Deployed to real-world users, the domain thus has the potential for regular and repeated interactions, potentially over extended periods of time.

  - The agent potentially provides real-world utility to its users. This will likely increase their engagement. This, in turn, will likely increase the chance that users provide natural (i.e., non-acted) and truthful (i.e., corresponding to their mental state of listening) feedback. Similarly for other verbal and non-verbal behaviours that are interesting for research on conversational interaction. Interactions between unacquainted people differ from interactions between people that already know each other, the latter being better coordinated and less cautious in their actions (for example, being less positive; Tickle-Degnen and Rosenthal 1990; Cassell et al. 2007).

- The agent can become familiar with its users. It can get to know a user's idiosyncratic feedback behaviour as well as the effectiveness of its own feedback-based adaptation. As the domain is personal, the agent can also learn its users' preferences regarding appointments, e.g., which appointments are important to a specific user and which ones not. This information can be of interest, e.g., when modelling dynamic grounding criteria (see section 2.2; Clark and Schaefer [1989, p. 291]).

This makes the calendar assistant domain suitable for both real-world applications as well as for smaller, focussed studies of individual aspects of speech based human–agent interaction.

8.6 SUMMARY

In this chapter we described how we integrated the models for attentive speaking — feedback based mental state attribution, adaptive natural language generation, and information-needs based feedback elicitation — into an artificial conversational agent. We discussed the limitations that SAIBA, the standardised architecture for behaviour generation, poses for planning and realising adaptive multimodal behaviour in real-time and presented an alternative architecture which is based on the principle of incremental processing and frames behaviour generation as an interactive process between multiple behaviour planning components, behaviour realisation, information state, and the interlocutor. In the next chapter, we will evaluate whether interactions with this implemented conversational agent show qualities that we expect from an agent that is an attentive speaker, and whether interlocutors are able to perceive this quality.

CHAPTER 9
EVALUATION OF THE ATTENTIVE SPEAKER AGENT

In this chapter we evaluate the implemented attentive speaker agent. We begin by giving a general perspective on the evaluation of artificial conversational agents and, based on this, develop an evaluation strategy and state the general hypotheses that we want to investigate. We describe our metrics and variables, the experimental conditions, the concrete hypotheses that we test, the setup of the study and the procedure. We then discuss the participants that took part in the study and analyse the data that we collected. In particular, we analyse whether participants provided feedback to the attentive speaker agent and whether this feedback is comparable to feedback produced in human–human interaction. We then analyse the objective and subjective quality of the interaction. This is followed by a general discussion of the results.

9.1 EVALUATING ARTIFICIAL CONVERSATIONAL AGENTS

Isbister and Doyle (2004) note that evaluation in the field of embodied conversational agents (and artificial conversational agents in general) is an inherently complex endeavour. The major reason for this is that it is often not clear how evaluation success should be measured (ibid., p. 5) — even to the degree that it is only vaguely defined what constitutes a successful interaction. The difficulty in finding crisp definitions of success is due to the many facets and to the complexity of what is being evaluated.

Embodied conversational agents are 'complete' systems (ibid., p. 5), often consisting of individual components that are integrated to form a whole. The behaviour of an agent arises from the interaction of its components. Thus, a successful evaluation of individual components of an agent does not guarantee a successful evaluation of a combination of them.

When humans become part of an evaluation, the result is a dynamic and continuous interaction of two 'systems' that are already complex in isolation. Humans bring their personal expectations as well as a bulk of (practical) knowledge, acquired during a lifetime, to the interaction. This gives rise to enormous complexity and adds subjectivity to the formula. Each interaction is unique and not easily comparable. What might work well with one person might not work well with another. Individuals might also differ in their subjective judgement of what is important for the interaction to be successful.

The various facets of such systems can be evaluated from different perspectives. A conversational agent can be evaluated in terms of the task it fulfils (e.g., is it useful?), in terms of its implementation (e.g., is it algorithmically efficient?, is it well structured?), and in terms of its appearance and behaviour (is it human-like?, is it believable?), as well as its capabilities (Ruttkay et al. 2004). Multiple scientific disciplines are involved, each having its own perspective on, and approaches to, evaluation (Isbister and Doyle 2004).

Evaluation results may thus differ depending on the perspective taken, or on the specific aspect that is evaluated. A believable agent with good conversational capabilities might fail on the task-level or vice-versa, and neither of the two aspects might be determinant for whether a human interaction partner found the interaction to be enjoyable or not. Thus, it is interesting to note that interactions can be considered a success or a failure depending on the perspective of evaluation. Additionally, a system that performs sub-optimal according to some perspectives might still be considered successful in general.

For the computational models underlying the attentive speaker agent developed in this thesis, we need to take all of this into account.

9.2 EVALUATION STRATEGY AND GENERAL HYPOTHESES

Our research hypothesis in this thesis is that attentive speaker agents endowed with capabilities of interactional intelligence should be able to engage in a simple and effective form of interactive feedback-based coordination on the levels of belief and attitude (see section 1.2, page 8). We break this down into three measurable correlates, that will serve as evaluation criteria, namely

(A) The attentive speaker agent and its interlocutors establish understanding (and evaluate acceptance and agreement) in a loop of feedback and adaptation of the agent's communicative behaviours.

(B) Interlocutors notice that the attentive speaker agent is interested in and able to infer their mental state of listening and responds appropriately.

(C) Interactions with an attentive speaker agent will be better than interactions with non-attentive speakers in that higher understanding will be reached, in a more efficient way.

These criteria will help us guide the process of defining a suitable evaluation strategy as well as concrete measures and hypotheses that can be tested.

Most importantly, all three criteria suggest that the evaluation of the attentive speaker agent has to be based on actual interactions with human interlocutors.[103] It follows from this that the agent needs to be sufficiently robust since evaluation studies with human participants are costly: they take a lot of resources and cannot be easily repeated when problems occur.

A further consequence of using human interlocutors is that the input processing problem needs to be solved, which we address by using the Wizard-of-Oz paradigm (as we have already mentioned in section 8.1). The wizard interprets participants' feedback signals in terms of form and communicative function and feeds this information to the attributed listener state component of the attentive speaker agent. A detailed description of the wizard's tasks are presented in section 9.3.5.

Measuring the correlates of evaluation criterion (A) requires a way to analyse the interaction between the agent and the participant, as well as the processes within the agent. To fulfil this requirement we record the interaction and write log files of the system behaviour (see section 9.3.4).

Measuring the correlates of evaluation criterion (B) requires an assessment of participants' subjective impression of the interaction. To gather this information, participants will have to fill in a questionnaire after the interaction. The questionnaire is presented in section 9.3.2.

Measuring the correlates of evaluation criterion (C) requires an operationalisation of the variables understanding and efficiency. Especially the degree of understanding should be measured shortly after a piece of information has been presented by the agent. The definition of these variables has a large influence on the overall design of the task and interactions and is described in section 9.3.1.

[103] In principle, interaction partners for the attentive speaker agent could also be simulated active listener agents, which would result in greater control and clearly defined behaviours. We opted for human interlocutors for three reasons: (i) adequate models for simulated interlocutors, which produce feedback based on actual understanding, are not readily available, (ii) ad-hoc models of active listening for the purpose of evaluation would carry the risk of only generating listening behaviour that we expect the speaker model to be capable of handling, and (iii) human interlocutors provide greater and more direct empirical validity to the evaluation.

Both (B) and (C) further imply that the evaluation of the interaction is compared to a baseline. We address this issue by defining two different baseline agents to which we can compare the attentive speaker agent. This yields three different experimental conditions, which we evaluate in a between-subjects design.[104] Consequently, several conversational agents — with different behaviours and capabilities — need to be created, and a higher number of participants is needed. The experimental conditions are described in section 9.3.3, the scenario for the evaluation — calendar assistance — was described in section 8.5.

9.3 MATERIALS AND METHODS

In this section, we describe the study design, including the objective and subjective variables that we measured, the experimental conditions we designed, the setup of the study, the task and training procedure of the wizard, and the experimental procedure.

9.3.1 OBJECTIVE METRICS AND VARIABLES

We defined the attentive speaker agent to be effective if interactions will be better than interactions with non-attentive speakers in that higher understanding will be reached in a more efficient way. In the following we develop metrics for the variables understanding and efficiency.

Understanding is problematic to evaluate objectively since it is not directly observable. This could be addressed by letting third parties rate and annotate participants' levels of understanding, e.g., by letting them analyse participants' understanding from their behaviour in a way the model of attributed listener state is not capable of. This would, however, create a new source of uncertainty.

We resort to a more objective — but also less direct — measure of understanding in terms of the ability of a participant to correctly recall information that the agent communicates. Such a 'recall score' is a continuous measure of performance on a ratio scale, which makes it suitable for comparisons. As described in section 8.5, the scenario for the attentive speaker agent is that of a personal calendar assistant. In the evaluation study the agent will present calendar information (it will announce appointments, it will communicate that some event needs to be moved to a different point in time, that an event needs to be cancelled, and it will propose new events to the

[104] Between-subject designs are cleaner in that participants are not influenced in their behaviour (and answers) by noticing differences between conditions. Such designs are also more straightforward to analyse. The disadvantage is that outcomes may be influenced by inter-individual differences and that more participants are needed.

interaction partner). The task of the participants then is to understand the information that the agent presents as well enough to be able to recall as much information as possible later.

Since memory fades over time (and we do not want to assess participants' memorisation skills), recall should be assessed not too long after the information has been presented. At the same time we do not want to assess recall immediately, as it would be interrupting and inhibiting to the unfolding of the dynamics of the interaction.[105] The solution that we choose is to divide the interaction into several blocks, each of which consists of a dialogue phase, in which the agents present the calender information of one week, and a recall phase, in which participants will (try to) recall this information by entering the calender events into a printed empty calendar. These can then be scored later, yielding the recall score.

Efficiency cannot be measured with a single variable. Conceptually speaking, it is a compound that consists of two factors: (i) a measure of performance (here operationalised in terms of recall/understanding), and (ii) a measure of the energy spent on achieving the performance (the costs of making the information understood). The efficiency variable is then defined as the ratio of performance to costs. Costs of the interaction can be defined in (at least) to ways:

First, we can resort to a simple objective metric that is often used to evaluate the quality of interactions with conversational agents: the duration of the interaction (Walker et al. 1998). Whether long or short interaction durations are desirable depends on the type and purpose of the dialogue. For an agent that primarily serves entertainment purposes, longer interactions might be indicators for dialogue success (as used, e.g., in Kopp et al. 2005; Swartout et al. 2010). Conversely, in more task-oriented interactions, where tasks are to be solved efficiently, short interaction durations are often desirable.

With no social talk — aside from a short *hello* in the beginning and a bye at the end of the interaction — the interaction in the calendar scenario of this evaluation study is purely task-oriented. As we are primarily interested in the efficiency of the interaction, low costs, and therefore shorter interactions are, in principle, desirable. Interaction duration can also be extracted automatically from system logfiles written during the interaction.

[105] Another approach to assessing understanding as soon as possible is to let participants answer simple yes-no questions about the information presented. We successfully employed this when evaluating an incremental, adaptive and situation aware in-car dialogue system where participants were distracted by the driving task (Kennington et al. 2014; Kousidis et al. 2014). Here such an approach would likely make recall to easy.

Alternatively, we can also define costs in terms of the amount of information being communicated redundantly. In principle, repeatedly expressing a piece of information that has already been communicated is not desirable, unless it has not yet been understood — in which case producing a repetition may be a useful mechanism to achieve understanding. The number of repetitions produced can also be extracted automatically from system logfiles written during the interaction.

We can thus define two metrics for efficiency of the agent: (i) the ratio of recall score to duration of the interaction, and (ii) the ratio of recall score to the amount of redundant information, in terms of repetitions.

9.3.2 SUBJECTIVE METRICS AND VARIABLES

Having defined these objective measures and variables for criterion (C), we can now turn to the measurement of subjective factors that reflect the participants' perception of the agent (criterion B) as well as some information about participants that may have an influence on the study. We measured these variables with a questionnaire that immediately followed the experiment and consisted of four parts.

The first part asked participants to report their subjective experience of the interaction. Twenty items were presented in random order and had to be rated on seven-point Likert scales.[106] In the following, these 20 items are grouped and presented according to five categories.

There were three items that target whether the agent is perceived to be a competent speaker, the first two (derived from a communicative competence self-report questionnaire [Rubin 1985, p. 177])

(Q1) *Billie drückt sich klar und präzise aus*
('When Billie speaks, his ideas are clearly and concisely presented')

(Q2) *Wenn Billie etwas erklärt ist es oft durcheinander*
('When Billie explains something to someone, it tends to be disorganized')

deal with the agent's ability to speak clearly. This is of interests as self-corrections or repetitions that might occur when adapting to its interlocutor's feedback may result in sub-optimal presentation. Finally, we wanted to know whether participants had problems understanding the agent

[106] Labels for the response anchors were: 1 — *stimme überhaupt nicht zu* ('strongly disagree'), 2 — *stimme nicht zu* ('disagree'), 3 — *stimme eher nicht zu* ('somewhat disagree'), 4 — *teils teils* ('neither agree nor disagree'), 5 — *stimme eher zu* ('somewhat agree'), 6 — *stimme zu* ('agree'), and 7 — *stimme voll zu* ('strongly agree').

(Q3) *Ich konnte Billie gut verstehen*
('I could understand Billie well')

as its speech may contain pronunciation errors due to synthesis artifacts.

Following these questions, six items on the agent's feedback processing capabilities were to be rated. The first two items

(Q4) *Billie hat mir signalisiert, wenn er eine Rückmeldung haben wollte*
('Billie gave me signals when he wanted to have feedback')

(Q5) *Billie war daran interessiert, dass ich ihn verstehe.*
('Billie wanted me to understand him')

target whether participants felt that the agent was interested in their feedback. These were followed by two items that ask participants whether they felt that their feedback was perceived and understood.

(Q6) *Billie hat meine Rückmeldungen wahrgenommen.*
('Billie perceived my feedback')

(Q7) *Billie hat meine Rückmeldungen verstanden.*
('Billie understood my feedback')

These were followed by two items that ask participants whether they felt that the agent made correct attributions of their mental state of listening, both in terms of understanding and attitude towards calendar items.

(Q8) *Billie kann einschätzen, ob ich verstanden habe was er sagt*
('Billie is able to tell whether or not I have understood what he has said')

(Q9) *Billie hat meine Einstellung zu den Terminen wahrgenommen.*
('Billie perceived my attitude towards calendar items')

Finally, participants were directly asked whether the agent was attentive and adaptive.

(Q10) *Billie war rücksichtsvoll und ist auf mich eingegangen.*
('Billie was attentive to me and adapted to my needs')

Participants further rated four items that could be grouped into a category that Ruttkay et al. (2004, p. 58) call 'helpfulness', that is, the items query whether participants perceived the agent to be cooperative. The first three items in this category specifically target whether the agent is attentive and adapted to the participants' needs.

(Q11) *Billie hat mir geholfen Schwierigkeiten beim Verstehen zu beheben*
('Billie helped me resolve difficulties in understanding')

(Q12) *Es war hilfreich, dass sich Billie bei Bedarf wiederholt hat*
('It was helpful that Billie repeated himself, when needed')

(Q13) *Es war hilfreich, dass Billie bei Bedarf weitere Informationen geliefert hat*
('It was helpful that Billie provided further information, when needed')

The fourth item targets interaction duration, with the underlying assumption that a helpful agent does not stretch the interaction unnecessarily.

(Q14) *Billie hat versucht das Experiment nicht länger als nötig dauern zu lassen*
('Billie tried to keep the experiment as short as possible')

Following this, three items on the perceived 'naturalness' of the agent's behaviour were to be rated. Ruttkay et al. (2004, pp. 58–59) define naturalness as being 'in line with the expectations of the user about a living, acting creature with respect to its embodiment and communicative behaviours'. The focus here is on the agent's ability to communicate in a smooth and well coordinated way as would be expected from a human speaker.

(Q15) *Die Interaktion mit Billie verlief reibungslos*
('The interaction with Billie was smooth')

(Q16) *Die Interaktion mit Billie war gut koordiniert*
('The interaction with Billie was well coordinated')

We also wanted to know directly, whether the agent's behaviour was perceived to be similar to a human speaker:

(Q17) *Billies Verhalten ähnelte dem eines menschlichen Sprechers*
('Billie's behaviour was similar to the behaviour of a human speaker')

Finally, three items on the task and the study itself were asked: The first two items target whether the task was difficult but still doable,

(Q18) *Ich empfand die Aufgabe als schwierig*
('I perceived the task to be difficult')

(Q19) *Ich konnte mir die Termine und Terminänderungen merken*
('I could remember calendar events and changes to them').

and the third item asks whether participants would regard their interaction as being successful in the context of the experiment.

(Q20) *Das Experiment war erfolgreich*
('The experiment was successful').

Table 9.13, on page 222 below, provides a structured and concise overview of all questionnaire items.

The second part of the questionnaire asked participants to provide demographic information about themselves. Participants reported their age and gender, their native languages and — if German is not among them — how many years of experience they have in speaking German. We further asked whether participants have normal or corrected-to-normal vision and hearing or not. We finally asked, whether participants had prior experience interacting with virtual agents or humanoid robots, as both experience and non-experience might have an influence on the interaction (ibid., p. 50).

The third part of the questionnaire was a personality test. We wanted to measure personality in case we notice huge differences in feedback behaviour among participants (feedback is likely influenced by personality; see, e.g., [Schröder et al. 2012; Huang and Gratch 2012]). We chose a short 11-item personality test for the 'Big Five' inventory (the BFI-10; Rammstedt and John 2007).[107]

Finally, in part four of the questionnaire, participants were given an opportunity to provide general remarks on the study by filling in a free-form field.

The questionnaire was automatically opened in a web-browser window after the agent closed the interaction and disappeared from the screen. Participants remained alone while answering the questions and left the room to meet with the experimenter once they completed the task.

9.3.3 EXPERIMENTAL CONDITIONS

Having defined the dependent variables that we want to measure, we now turn to the definition of the experimental conditions, which will serve as the independent variable.

The evaluation study consists of three experimental conditions, the main condition (ATTENTIVE SPEAKING or AS), in which participants interact with the attentive speaker agent, and two control conditions that serve as baselines to which condition AS will be compared, one as a lower-bound baseline, the other as an upper-bound baseline. Across all three conditions, the agents' appearance and voice are the same, they present

[107] Although this specific test has less statistical power than the full 44 item test it is derived from (the BFI), we accepted the power versus time-to-complete trade-off since personality is not central to our research questions.

the same calendar operations and items to the participants, and participants receive the same instructions and the same questionnaire.

The attentive speaker agent perceives its participants' feedback in a timely manner (via the wizard), probabilistically attributes a listening-related mental state to them, incrementally adapts its natural language generation process — potentially already tailoring the next utterance unit of an information presentation unit to the participants' needs. If the agent itself has an information need, it can also try to elicit feedback from its users by producing feedback elicitation cues between utterance units and at the end of an utterance. Towards the end of each information presentation unit, the attentive speaker agent evaluates the attributed listener state again and decides how to continue. If the agent attributes sufficiently high understanding to the participant it proceeds to the next information presentation unit. If it attributes low understanding, it will repeat the information unit. If the information in the attributed listener state is not clear, it will explicitly ask the participant, whether it should repeat the information (which will then happen in an adapted manner) or whether it should continue. Depending on the type of calender operation that is communicated (announce, cancel, move, propose) the agent may also want an attitudinal reaction of the listener, on which it may then comment. See section 8.3.4, especially table 8.1), for an overview of these transitions. Whether these models work as intended is subject to this evaluation study.

In contrast to this, the agents in the two control conditions do not have models for any of the above and represent the classic, non attentively speaking artificial conversational agents. The agent in control condition (NO ADAPTATION or NA) is completely ignorant of its interlocutor when presenting information and could, in principle, be replaced with a video. An illusion of interactivity is only created (and maintained) in that the agent replies to the initial greeting of the participants and in that it reacts promptly to participants' requests to continue the 'interaction' after each of the recall phases.

Condition NA is intended to serve as the lower-bound baseline on participants' understanding (how much information can participants recall when it is only presented once) and on interaction duration.

The agent in control condition (EXPLICIT ASKING or EA) has the same basic behaviour as the agent in control condition NA, but differs in that it explicitly asks participants after each presented unit of information, whether it should continue with the next unit or should repeat the current unit again. Participants have to answer this question and the agent then proceeds accordingly. Participants can have an item repeated as often as they want to.

Condition EA is intended to serve as an upper-bound baseline for the level of understanding that participants can achieve in a system that resembles a classical interactive conversational agents that repeats information upon request. It is also (likely) an upper-bound for the duration of interactions.

The overview of the study, fig. 9.4 on page 193 below, schematically illustrates the differences between the three conditions in form of simple flow charts.

The control conditions also serve as baselines for the questionnaire, e.g., whether participants notice that these agents do not attend to and react to their feedback. Moreover, they can be used to assess whether participants provide as much feedback to such an agent as to the attentive speaker agent.

Both control conditions are important for the evaluation of the attentive speaker agent since they allow an assessment of the agent on scales with upper and lower bounds, measured in the same task. Ideally, the attentive speaker agent would approach the upper bound for understanding (EA) and the lower bound for costs of the interaction (NA) — reaching better efficiency than the agents of the control conditions. But such simple predictions are insufficient when evaluating dynamic interactions between complex computational models and human interlocutors. Hence, the arguments on which the actual hypotheses we will test are based will be more complex as well. We will formulated them later, in section 9.5, right before analysing each individual variable.

9.3.4 SETUP

The study was carried out in the lab space of the 'Social Cognitive Systems Group', at CITEC, Bielefeld University. The conversational agent and the recording equipment was set up in the 'Computational Interaction Studies Lab', where instruction of participants as well as the interactions themselves took place.

Participants were seated at a desk at an approximate distance of 50–80 cm to an all-in-one computer with a 27 inch 16:9 screen which ran and displayed the conversational agent during the interaction and, later, the post-interaction questionnaire (see fig. 9.1*AB*). A wireless keyboard/touch-pad-combination was available for doing the questionnaire.

The agent was visible from the chest upwards (see fig. 9.1*C*). The configuration resulted in a slightly smaller than life sized visualisation of the agent with a displayed height of approximately 33.5 cm (head 17.5 cm) and width of 23 cm (head 12 cm). The agent's voice was played from speakers positioned below the desk.

The interactions were filmed with Sony NEX-VG30 HD cameras from two perspectives, see fig. 9.2: (*A*) an ultra wide angle perspective capturing the whole scene from behind — including agent, participant, and experimenter (when present) — , and (*B*) a frontal close-up perspective capturing the participant's face, head, and upper

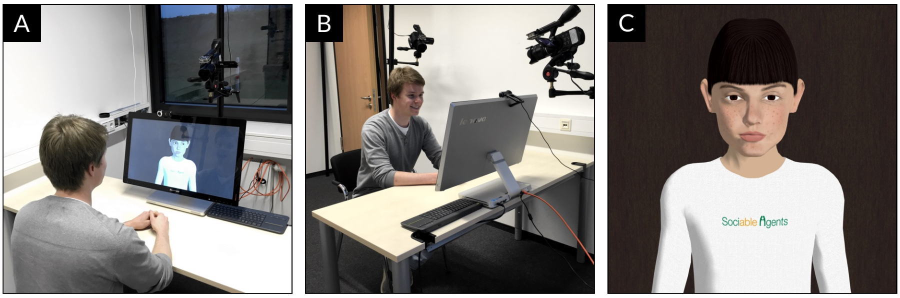


Figure 9.1: Setup of the evaluation study photographed from two perspectives *(A, B)*. Photos are staged with a person that was not a participant. (*C*) shows the conversational agent in the neutral pose.

body. Videos were recorded in 1080p with 50 frames per second. Audio was recorded with the build-in microphones of these cameras. In addition to the behaviour of the participants, log files — logging the processes within the agents as well as all the actions of the wizard — were written.

A third perspective (*C*) — very similar to (*B*) — was filmed with a webcam and displayed in real-time to the wizard. This perspective was streamed and recorded in 720p with 10 frames per second (see fig. 9.2*C*) and also includes an audio stream.

9.3.5 THE WIZARD-OF-OZ

The wizard operated from the 'Interactive Media Lab' next door to the 'Computational Interaction Studies Lab'. The interaction, as well as the general progress of each trial, could be observed and listened to in real-time via the webcam-based video stream (fig. 9.2*C*) and a high-quality audio-link recorded with a room microphone.

The wizard's tasks varied depending on the experimental condition (see table 9.1). In all three conditions the wizard started the interaction as well as the six information presentation blocks as soon as participants signalled that they are ready to begin or continue. In conditions ATTENTIVE SPEAKING and EXPLICIT ASKING the wizard additionally chose which continuation the participants requested when the agent asked them for a preference (continue with next or repeat the current calendar item). Most

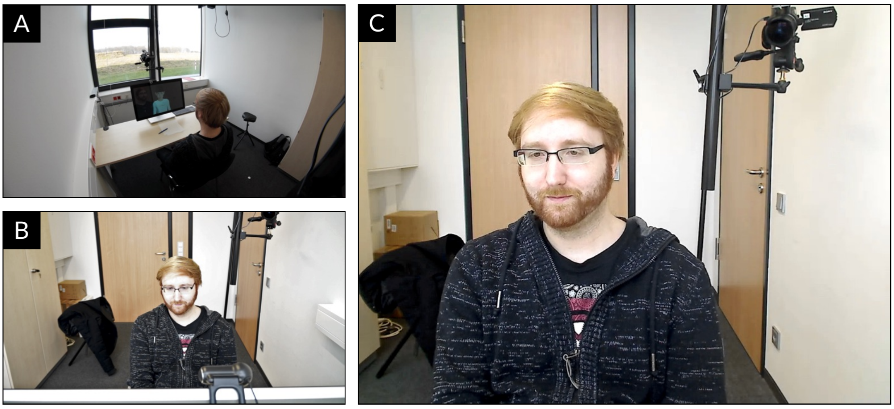


Figure 9.2: Stills from the three camera perspectives. *(A)* ultra wide angle shot capturing the agent, the participants from behind, and the experimenter — during instruction; *(B)* frontal close-up shot capturing participants; and *(C)* webcam-perspective used by the wizard to observe participants during the interaction.

Table 9.1: Wizard-of-Oz tasks depending on experimental condition.

Condition	Start interaction/blocks	Choose continuations	Interpret feedback
AS	●	●	●
EA	●	●	○
NA	●	○	○

importantly, in condition ATTENTIVE SPEAKING, the wizard had to pay attention to the participants' feedback signals — verbal and non-verbal — , categorise their function, polarity, and level of certainty, and feed them as input to the attributed listener state component of the attentive speaker agent system.

The Wizard-of-Oz system consisted of two user interface windows displayed next to each other on a 24 inch screen. One window displayed the streamed webcam perspective (see fig. 9.2*C*), the other the Wizard-of-Oz graphical user interface (see fig. 9.3*A*).

The Wizard-of-Oz interface is controlled via keyboard shortcuts and mouse

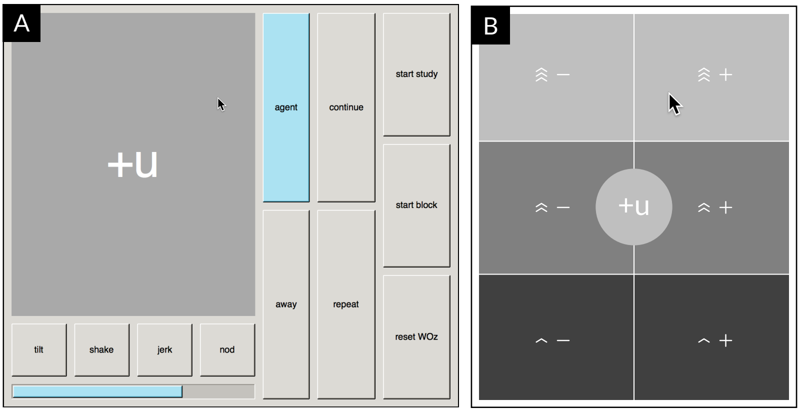


Figure 9.3: The Wizard-of-Oz interface. *(A)* Screenshot of the graphical user interface with a feedback input area; buttons to input a participant's head gestures events (tilt, shake, jerk, nod) and gaze target (agent, away); to choose continuations (continue, repeat); to control the study (start, next block, reset); and utterance progress bar. *(B)* The feedback input area is divided into six invisible regions: negative (−) or positive (+) polarity; and low (⌃), medium (⌃⌃) or high (⌃⌃⌃) level of certainty. The feedback input area shows the active feedback function (chosen by keyboard) and polarity in large type on a background coloured depending on the level of certainty (the former selected by keyboard, the latter two by mouse pointer position). A mouse click inputs a feedback event.

pointer. The basic actions for starting the experiment and experimental blocks as well as for repeating information presentation units or continuing with the next unit are triggered with a click on the respective interface buttons on the right hand side of the window. During the interaction in condition ATTENTIVE SPEAKING the gaze target of the participant (a binary state, either agent or away) can be set by either typing the key `f`, which toggles the gaze target, or by clicking one of the labelled interface buttons. The set gaze target is highlighted in light blue.

Similarly, head gestures (types: nod, jerk, shake, tilt) performed by participants are entered by either typing a key (`v`, `c`, `x`, `z`) or by clicking the labelled interface button.

Participant feedback is specified via the feedback input area, a custom user-interface element (see fig. 9.3*B*) that works in two stages and allows for quick entry of multidimensional features of a feedback signal: function, polarity, and level of certainty. Feedback function is set via the keyboard to either perception/p, understanding/u, or acceptance/ac using the keys (a, s, or d). This function remains set until another function is chosen. When participants produce feedback, the events are entered into the system by simultaneously selecting the polarity (positive/+, negative/−) and level of certainty (low/⌃, medium/⌃⌃ or high/⌃⌃⌃). This is done by clicking into one of six specific regions of the feedback input area. As shown in fig. 9.3*B*, it is divided into two horizontal regions and three vertical regions. The right half corresponds to positive, the left half to negative polarity. The top third corresponds to a high level of certainty, the middle third to a medium level of certainty, and the bottom third to a low level of certainty.

The feedback input area constantly shows the currently selected feedback function in large type. In addition, information on polarity (in front of the function) and level of certainty (different intensity level of the background) that would result from a click into a specific region are shown when the mouse pointer hovers over the area.

As an example, illustrated in fig. 9.3*A*, positive feedback of understanding with a high level of certainty can be entered into the system by first setting the feedback function to u and then clicking somewhere in the upper-right region of the feedback input area. Timing of the click is meant to correspond to the timing the feedback event occurs, disregarding the wizard's processing delay.

Finally, the wizard user-interface shows a progress bar for the utterance that the agent is currently speaking (see bottom left of fig. 9.3*A*). Utterance progress is updated via information on the predicted end time of the utterance (provided through BML prediction feedback [BML Committee 2011]) and the current point of time. This gauge is meant to help the wizard predict when user feedback is most likely to happen.

In all trials that required feedback to be entered into the system (condition ATTENTIVE SPEAKING), the same person — SR — acted as the wizard. In the control conditions EXPLICIT ASKING and NO ADAPTATION, where the wizard's task was straightforward and did not involve subjective decision making, either SR or HB acted as wizards.

SR was instructed and trained starting weeks before the actual study took place. The first step was a familiarisation with the subject matter, especially linguistic feedback. This was done via reading and discussion (with HB) of the theory underlying the model implemented in the attentive speaker agent (e.g., Allwood et al. 1992; Buschmeier and Kopp 2012b; Buschmeier and Kopp 2014a). Following this, the wizard analysed feedback annotations in the KDS-1-corpus (Buschmeier and Yaghoubzadeh [2011]), which follows a scheme of categories very similar to what is fed into the attributed listener state component of the attentive speaker agent system. One dialogue was re-annotated and differences in choice were discussed. As soon as the user interface was available, the wizard tested it on video-data from the KDS-1-corpus. Feedback from these trials was integrated into the wizard interface in several iterations. Finally, a total of 17 pilot trials (see section 9.4) were carried out.

With this amount of preparation, we were very confident that the wizard was well trained and fully understood his task. It should be noted that the wizard was in the known of experimental conditions. We are aware that this is rather undesirable in experimental settings, but it was unavoidable for the study we conducted because the experimental condition can easily be guessed from the agents' behaviour. The option of concealing the agents' actions from the wizard was ruled out since it would have severely impaired his ability to classify feedback in the way needed for the system being evaluated.[108]

By design, the wizard's input influenced the behaviour of the attentive speaker agent. The wizard's choices, however, do not necessarily have a directly predictable effect on the agent's behaviour as it is mediated by the attributed listener state, which brings factors that are not under the control of the wizard (dialogue context, listener state dynamics, processing delays) into play. For this reason, we are confident that the wizard was not making decisions biased towards a desirable outcome of the agent's behaviour. Anecdotally, the wizard was occasionally surprised by the agent's behaviour — as it did not do what he would have expected. Because of this, it was made clear, right at the beginning, that the wizard should not try to optimise their feedback classification ability in responses to unexpected agent behaviour.

9.3.6 PROCEDURE

Having defined all properties of the study, we can now describe the experimental procedure, a schematic overview of which is given in fig. 9.4.

Participants were met in the foyer of the CITEC building, greeted and brought to the lab by the experimenter (HB). Once in the lab, they were asked to sit at the table and to read an information sheet on the study and a consent form. The sheet contained information on the general procedure, a short instruction (109 words, see appendix B.1), information on the data acquired, how this data will be processed and pseudonymised/anonymised, who may access the data, how long the study will take, and how participants will be compensated. Participants were encouraged to ask questions and were given as much time as needed to read the information.

[108] An insight we gained when working on the ALICO-corpus (Malisz et al. 2016) is that dialogue context is needed when classifying feedback functions in an annotation task — and the wizard's task in this study is basically an annotation task that has to be done in real-time.

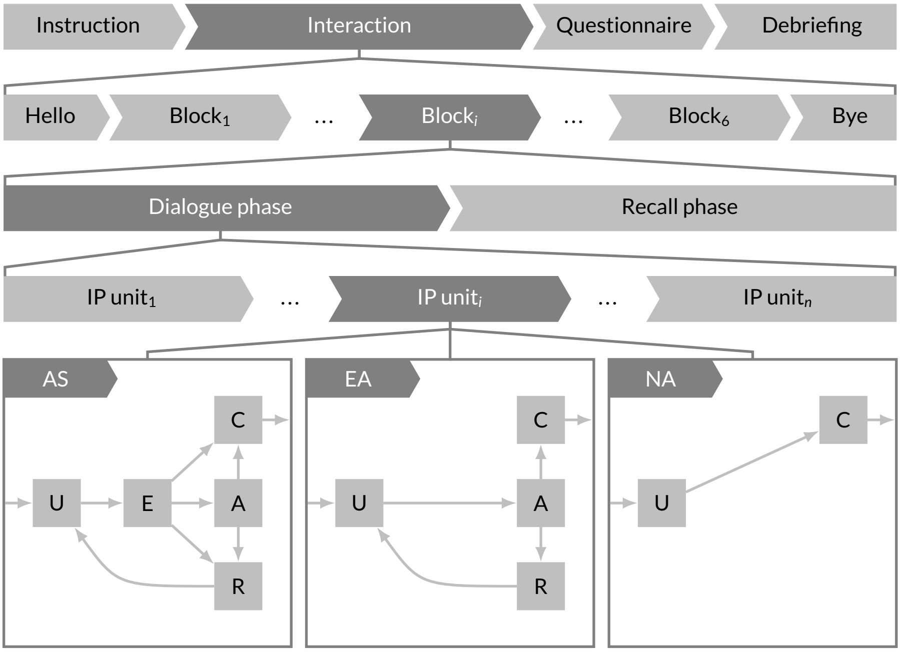


Figure 9.4: Overview of the evaluation study. The interaction basically consists of six experimental blocks, each of which has a dialogue and a recall phase. Dialogue phases consist of two to three information presentation units (IP), the structure of which differs depending on the experimental condition (AS/EA/NA), as the flow charts schematically illustrate. Nodes in these charts represent the following actions: U — present information in an incremental utterance; E — evaluate current attributed listener state, decide what to do next, and describe this to the participant; C — continue with next unit; R — repeat this unit; A — ask interlocutor whether to repeat or continue. Recall phases consist of a pen-and-paper information recall task.

After signing the consent form, camera recordings were started and more detailed oral instructions were provided by the experimenter (see appendix B.2 for the transcript of one exemplary oral instruction). Notably, the instructions contained the information that participants can provide feedback and that the agent may take this information into account in its own behaviour. Participants were again encouraged to ask questions.

After the experimenter left the room, participants started the interaction on their own by gazing towards the screen and calling the agent using a greeting such as *Hallo Billie!* ('Hello Billie!'). After the interaction, the participants had to do the questionnaire (electronically). After that they could leave the room and were awaited by the experimenter. Participants were then asked whether they had suspected something during the experiment and were debriefed (i.e., they were told that, behind the scenes, a Wizard-of-Oz was at work, and that their interaction was subject to a specific experimental condition).

The interaction itself started with a greeting and the agent signposting what will happen next. After that the first of six randomized experimental blocks started. Each block consisted of an information presentation phase followed by a recall phase. During the information presentation phase, two to three information presentation units, each consisting of one calendar operation, were performed. During these units, the agent announced calendar events to the participant, talked about changes to an event (cancel, move) or proposed a new event (see table 9.2). The flow charts in fig. 9.4 schematically illustrate how the information presentation units were presented in each of the experimental conditions.

9.4 PARTICIPANTS

We recruited 59 participants via advertisements posted on bulletin boards at Bielefeld University and Bielefeld University of Applied Sciences as well as in the student-run Facebook groups of the two institutions. The advertisements vaguely described the purpose of the study (to test and rate spoken interaction with a 'virtual assistant'), stated requirements (a high proficiency of spoken German), duration (about 30 minutes) and compensation (5 euro).

Seventeen of the recruited participants served as pilots and six had to be excluded from the analysis for technical problems. We kept recruiting participants until we had successful trials for all twelve slots per condition. Participants were blindly assigned to conditions (i.e., before showing up at the lab).[109]

[109] Assignment of participants to conditions resulted in the following sequence: EEEEENNENENNNEN-

Table 9.2: Structure of the information presentation phase of each of the six experimental blocks. Blocks contain a sequence of two or three information presentation units, each consisting of one calendar operation (announce, cancel, move, or propose calendar items). Announce operations may mention multiple calendar items, resulting in a larger number of calendar items than information presentation units per block.

Block (id)	Calendar operations (type)				IP units			Items (counts)
	announce	cancel	move	propose	U_1	U_2	U_3	
1	●	●	○	●	P	A	C	4
2	●	○	●	○	A	M	A	4
3	●	○	●	○	A	M		4
4	●	●	○	●	A	P	C	4
5	●	○	●	●	P	M	A	3
6	●	○	●	○	A	M	A	5

Table 9.3: Demographics (age and gender) of participants — by condition and overall. *N* is the number of participants.

Condition	Age (years)				Gender (counts)		*N*
	min	max	*M*	*SD*	female	male	
AS	19	31	24.3	3.7	8	4	12
EA	18	40	24.3	6.2	10	2	12
NA	20	28	24.1	2.5	7	5	12
Overall	18	40	24.2	4.4	25	11	36

The 36 participants included in the analysis ranged between 18 and 40 years of age ($M = 24.2$, $SD = 4.4$), with similar age distributions among conditions (see table 9.3). We did not have the objective to balance gender, which resulted in a significant gender-imbalance. 69 % of participants reported to be 'female', and 31 % to be 'male'.

We analysed whether it is likely that gender distribution is independent from condition, as there could be an influence on the outcome of the study, if not. Based on an approximate Pearson's χ^2-test ($\chi^2 = 1.8327, p = 0.543$) the null hypothesis of

NNENEENENAAAAAAAAAAAA. Condition ATTENTIVE SPEAKING was acquired after the acquisition of conditions EXPLICIT ASKING and NO ADAPTATION.

independence cannot be rejected. A contingency table Bayes factor test[110] ($\textsc{bf}_{01}$ = 2.557, prior concentration set to $a = 1$), however, finds only 'anecdotal' evidence[111] for the null hypothesis. Based on this analysis, an influence of gender on the outcome of the study cannot be ruled out.

We asked participants for their native languages and — if German is not among them — how many years of experience they have in speaking German. Three participants (8 %) reported German not to be among their native languages (see table 9.4. One of these participants fell into condition ATTENTIVE SPEAKING (with 2.5 years of experience speaking German), two into condition NO ADAPTATION (with 16 and 23 years of experience speaking German).

We analysed whether it is likely that being a native speaker of German is independent from condition. Based on an approximate Pearson's χ^2-test ($\chi^2 = 2.1818$, $p = 0.765$) the null hypothesis of independence cannot be rejected. Furthermore, a contingency table Bayes factor test ($\textsc{bf}_{01}$ = 6.586, prior concentration set to $a = 1$) finds 'substantial' evidence in favour of independence. As the number of cases is very low and both tests point in the direction of independence, we assume that an influence on the outcome of the study is unlikely.

We also asked participants whether they had prior experience interacting with virtual agents or humanoid robots. Surprisingly, half (50 %) of the participants had — usually from participating in similar studies at Bielefeld University.

[110] Gûnel and Dickey's (1974) contingency table Bayes factor test — as implemented in the 'BayesFactor' R package (Morey and Rouder 2015). We assume Gûnel and Dickey's sampling model (iii), with fixed row marginals and independent multinomially distributed rows, i.e., the null hypotheses states that multinomial probabilities are equal across rows. This model is also used in subsequent contingency table Bayes factor tests.

[111] Interpretation of strength of evidence K (= $\textsc{bf}$) for a hypothesis according to Jeffreys (1961, p. 432):

$1 < K < 3.16$	barely worth mentioning/anecdotal
$3.16 < K < 10$	substantial
$10 < K < 31.62$	strong
$31.62 < K < 100$	very strong
$100 < K$	decisive

Values of K between 0 and 1 are evidence for the 'other' hypothesis, but can be interpreted by calculating $1/K$. ¶ Given two hypotheses A and B, a value of $K > 1$ is evidence for hypothesis A when a Bayes factor $\textsc{bf}_{AB}$ is specified, and for hypothesis B when a Bayes factor $\textsc{bf}_{BA}$ is specified. Similarly, a value of $K < 1$ is evidence against hypothesis A when a $\textsc{bf}_{AB}$ is specified, and against hypothesis B when a $\textsc{bf}_{BA}$ is specified. ¶ Typically A and B are null hypothesis and alternative hypothesis — and represented as 0 and 1, respectively. The Bayes factor is then specified as $\textsc{bf}_{01}$ or $\textsc{bf}_{10}$. Further symbols can be used when comparing more specific hypothesis, such as, for example, > and < for one-sided hypothesis with a specific ordering. The Bayes factor is then specified as $\textsc{bf}_{>0}$, $\textsc{bf}_{0>}$, $\textsc{bf}_{<0}$, $\textsc{bf}_{0<}$, or even $\textsc{bf}_{<>}$ or $\textsc{bf}_{><}$ when comparing one-sided hypothesis against each other.

Table 9.4: Proportions of participants that reported being native speakers of German; having prior experience interacting with virtual agents or humanoid robots; having normal or corrected-to-normal vision and hearing.

Condition	Native speaker	Prior experience	Normal vision	Normal hearing
AS	0.92	0.5	0.92	1
EA	1	0.42	1	1
NA	0.83	0.58	0.75	0.92
Overall	0.92	0.5	0.89	0.97

We analysed whether it is likely that having prior experience interacting with virtual agents or humanoid robots is independent from condition. Based on an approximate Pearson's χ^2-test ($\chi^2 = 0.66667, p = 0.9144$) the null hypothesis of independence cannot be rejected. Furthermore, a contingency table Bayes factor test ($\textsc{bf}_{01} = 3.792$, prior concentration set to $a = 1$), finds 'substantial' evidence in favour of independence. As both tests again point into the direction of independence, we can assume that an influence on the outcome of the study is unlikely.

We also asked participants whether they have normal or corrected-to-normal vision and hearing or not. Four participants (11 %) reported non-normal and non-corrected vision (one in condition ATTENTIVE SPEAKING, three in condition NO ADAPTATION). A different participant reported non-normal and non-corrected hearing (in condition NO ADAPTATION).

With regard to vision, we analysed whether it is likely that distribution of participants reporting non-normal and non-corrected vision is independent from condition. Based on an approximate Pearson's χ^2-test ($\chi^2 = 3.9375, p = 0.2963$) the null hypothesis of independence cannot be rejected, but a contingency table Bayes factor test ($\textsc{bf}_{01} = 2.661$, prior concentration set to $a = 1$), only finds 'anecdotal' evidence in favour of independence. Based on this analysis, an influence of vision on the outcome of the study cannot be ruled out.

The agent's non-verbal (i.e., visible) behaviour, however, only plays a minor role in the evaluation study, with the agent's presence and its gaze behaviour being the only two important features. Considering the study setup (see fig. 9.1), presence is likely perceivable even for participants with lower than normal eyesight. The agent's eye movements, however, are more subtle and are used for feedback elicitation — though only in condition ATTENTIVE SPEAKING. Weighting all these aspects, we consider

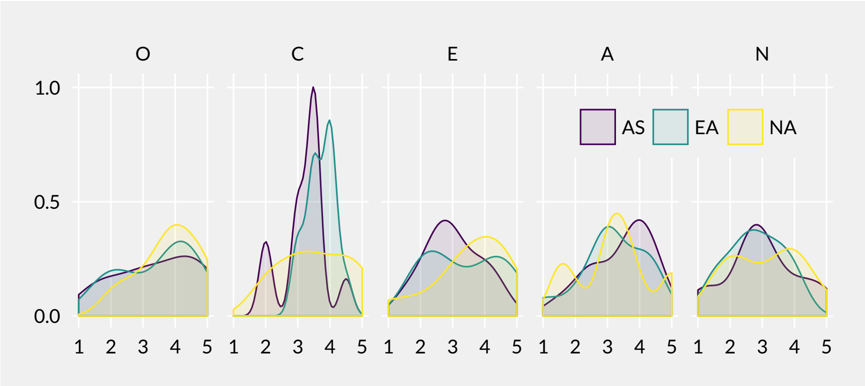


Figure 9.5: Distribution of participants' 'Big five' personality traits by experimental condition. Each facet of the plot shows the distribution of the strength (1–5) of one factor (openness, conscientiousness, extraversion, agreeableness, neuroticism) in all three conditions (purple: ATTENTIVE SPEAKING; green: EXPLICIT ASKING; yellow: NO ADAPTATION).

it unlikely that imperfect vision reported by four participants will have a significant influence on the outcome of this study.

With regard to hearing, it does not make sense to carry out a statistical analysis for the single participant and we do not expect it to have an influence on the outcome of the study.

We also measured the participants' personality (in terms of the 'Big five' inventory using the BFI-10 test, [Rammstedt and John 2007]; see section 9.3.2). Figure 9.5 shows the distribution of the strength (from 1–5) of each factor, split up by factor (openness, conscientiousness, extraversion, agreeableness, neuroticism) and experimental condition. It can be seen that, in general, participants range across the full strength-spectrum on all factors except conscientiousness. Comparing experimental conditions, it can be noticed that the distributions of strength values of each factor are quite similar across conditions (again with an exception in conscientiousness, where values in condition NO ADAPTATION are more uniformly distributed than in conditions ATTENTIVE SPEAKING and EXPLICIT ASKING).

9.5 ANALYSIS AND RESULTS

In this section we analyse the data gathered during the interaction study, directly summarising and discussing intermediate results.

9.5.1 PARTICIPANTS' FEEDBACK BEHAVIOUR

We begin by analysing whether participants actually provided feedback to the agents they interacted with.

Annotation of participant's feedback signal

In order to be able to analyse the feedback behaviour that participants showed during the interactions with the agent, we annotated the recordings using the audio-visual data from the webcam perspective (i.e., annotators had the same information as the wizard; see fig. 9.2*C*).

All utterances by participants were segmented and transcribed from audio data only, using 'Praat' (Boersma and Weenink 2016). Those with feedback characteristics that were not produced in response to questions of the agent were classified as feedback and transcribed on one tier, all other verbal acts were transcribed on a second tier. Transcription of feedback was based on the conventions of the ALICO-corpus (Malisz et al. 2016).

Head gesture feedback was segmented and labelled from video data only, using 'ELAN' (Wittenburg et al. 2006). Head gesture unit labels were limited to Włodarczak et al.'s (2012) head movement types (nod, shake, jerk, tilt, turn, protrusion, and retraction). As no audio was used during annotation, head gestures that were produced while responding to questions of the agent were segmented and annotated as well — and filtered out during the analyses.

In total 33 (of the 36)[112] interactions were segmented, transcribed, annotated,[113]

[⋆] Classical null hypothesis significance testing (NHST) based analyses in this section follow the applied statistics textbooks of Field (Field et al. 2012; Field 2009). Bayesian analyses are based on the work of Rouder, Morey, and colleagues (Rouder et al. 2009; Rouder et al. 2012; Morey 2014; Morey 2015). Both types of analyses were carried out with the statistics software R (version 3.2.3). Non-standard R-packages critical to the analyses are cited when relevant. ¶ The evaluation data and analysis source code are available as a data publication at DOI: 10.6084/m9.figshare.3827277 .

[112] Audio-visual data from the webcam was incomplete or unavailable for three interactions in the EXPLICIT ASKING condition.

[113] Some of the segmentation and transcription was carried out by HB. A second annotator, KS, segmented, transcribed, and annotated the largest part of the data — supervised by HB. Segmentation was done rather coarsely. Transcriptions of verbal feedback signals were checked and corrected through systematic listening.

and analysed[114].

Do participants provide natural feedback in human–agent interaction?

Counting all instances of feedback across conditions yields a total number of 734 signals, 127 (17.3 %) of which are unimodal verbal signals, 296 (40.3 %) are unimodal head gestures, and 311 (42.4 %) are bimodal signals in which a verbal/vocal feedback expression and a head gesture unit are produced in overlap. These numbers show that participants in the evaluation study were willing to provide feedback to an artificial conversational agent, a result that parallels the findings of Reidsma et al.'s (2011) study of an attentive speaker agent.

The question, of course, is whether the feedback signals that participants provided in interaction with the artificial agent can be considered 'natural', i.e., whether they are similar or different to feedback that is provided in human conversation. To this end, we will next analyse form and frequency properties of the feedback signals.

Of the 436 verbal/vocal feedback signals produced, *okay* is the most frequent (41.5 %) feedback expression, followed by *mhm* (18.3 %), *ja* (14.2 %), *m* (6 %), *nein* (3.2 %), *hm* (1.8 %), *ja okay* (1.6 %), and *nee* (1.4 %). These are followed by a 'long tail' of expressions (each < 1 %), 16 of which are single feedback morphemes (e.g., *oh*, *hä?*) and 21 are expressions created through syntactic operation (e.g., *mhm ja*, *hm nein*, *okay ja*; Allwood 1988, see section 3.3.1). This distribution of short feedback expressions is similar to the distribution we found in human–human conversations in the ALICO-corpus (Malisz et al. 2016, tbl. 7), where the four most frequent feedback expressions are *ja*, *m*, *mhm*, and *okay*, too.

Of the 598 head gestures produced as feedback, 81.6 % were labelled nod, 8.9 % tilt, 6.4 % shake, and 3.2 % jerk. This distribution of head gesture types is, again, similar to the one we found in human–human conversation in the ALICO-corpus (ibid., tbl. 4), both are nod-heavy and the four most frequent units are also nod, jerk, tilt, and shake.

The 311 bimodal feedback signals were coherent in that almost all head gestures of the category nod occurred together with verbal feedback expressions that can be considered positive in polarity (*okay*, *mhm*, *ja* and *m*). Head gestures of the type shake, on the other hand, occurred in overlap with the verbal feedback expression *nein* (and variants of it such as *hm nein*) — these have negative polarity. Head gesture types tilt and jerk occurred rarely in overlap with verbal/vocal feedback.

Segmentation and annotation of head gestures were checked for systematic problems. Misspelled labels and transcription were corrected semi-automatically.

[114] Annotations and transcriptions were brought into R via Python and 'TextGridTools' (Buschmeier and Włodarczak 2013).

In summary, it can be said that (i) participants provided multimodal communicative listener feedback to the artificial conversational agents they interacted with, and (ii) the feedback they produced is comparable, in its form and distribution, to feedback found in human–human conversations.

It is important to acknowledge though that participants were told consistently across all three experimental conditions — in the instructions — that they *can* provide multimodal communicative listener feedback to the agent and that the agent can take their feedback into account in its own behaviour (see section 9.3.6 and appendix B). Only the agent in condition ATTENTIVE SPEAKING (the attentive speaker agent), however, perceived participants' feedback signals and adapted its behaviour. Moreover, this agent pro-actively elicited feedback, if participants did not provide it spontaneously (see chapter 7). The agents in the two control conditions EXPLICIT ASKING and NO ADAPTATION did neither of this.

Does the agent's behaviour influence participants' feedback rate?

This raises the question whether the agents' behaviours (i.e., the experimental condition) actually influenced participants in their feedback behaviour — or whether they provided feedback because the instructions mentioned that they could do it. To shed light on this question, we analyse if and how participants' feedback behaviour differed across conditions. Since the agent in condition ATTENTIVE SPEAKING takes feedback into account — and even produces feedback elicitation cues — , our hypothesis is that participants provide more feedback in this condition than in the two control conditions EA and NA, between which we expect no difference (in both conditions the agent ignored listener feedback).

Analysing the absolute number of feedback signals provided by participants[115] is not useful because net durations (i.e., disregarding the time of the recall phases) of the interactions differ (this is further analysed in table 9.7 below). The main reason for this is that information presentation units can be presented repeatedly in conditions AS and EA (see section 9.5.2, page 211 for an analysis of the number of presentations and repetitions). To be comparable across participants and conditions, we therefore base the following analysis of participants' feedback behaviour on the rate at which they provided feedback. We define this 'feedback rate' f_{FB} as the number of feedback signals per presentation (i.e., $f_{\mathrm{FB}} = \#\mathrm{FB}/(\#\mathrm{repetitions} + 17)$; see caption of fig. 9.9).

Across all conditions, participants produced between 0 and 2.4 feedback signals per presentation unit, with a mean feedback rate of $M = 1.1$ ($SD = 0.8$). Analysing participants' feedback behaviour by experimental condition we find differences in

[115] Across conditions, participants produced between 0 and 54 feedback signals ($M = 22.2$, $SD = 17.6$).

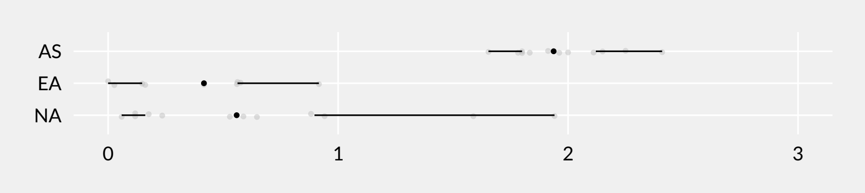


Figure 9.6: Distribution of participants' feedback rate (number of feedback signals per presentation; see text) by experimental condition. Data points are y-jittered in translucent light grey; black dots are medians, black lines are whiskers, mid gaps are quartiles.

feedback rate. Participants in condition ATTENTIVE SPEAKING have a mean feedback rate of $M = 1.97$ ($Mdn = 1.93$, $SD = 0.22$, $\min = 1.65, \max = 2.4$). Participants in condition NO ADAPTATION follow with a mean feedback rate of $M = 0.65$ ($Mdn = 0.56$, $SD = 0.6$, $\min = 0.06, \max = 1.94$). Participants in condition EXPLICIT ASKING only have a mean feedback rate of $M = 0.1$ ($Mdn = 0.41$, $SD = 0.31$, $\min = 0$, $\max = 0.92$). Figure 9.6 shows the distribution of feedback rate by condition. To confirm our hypothesis that participants provide more feedback in the ATTENTIVE SPEAKING condition, we will carry out an inferential analysis of feedback rates.

An independent one-way Welch-approximated ANOVA[116] indicates that feedback frequencies are statistically significantly different between experimental conditions: $F(2, 17.112) = 94.68, p = 5.571\text{e}{-10}$. Post-hoc Welch's one-sided two sample t-tests between conditions (see table 9.5) further reveal statistically significant (to a Bonferroni-corrected alpha-level of $\alpha/3 = 0.0167$) differences of feedback rates between conditions ATTENTIVE SPEAKING and NO ADAPTATION as well as between conditions ATTENTIVE SPEAKING and EXPLICIT ASKING. In both cases the effect size can be considered large ($r > 0.5$; Cohen 1992). However, the null hypothesis that participants' feedback rate in condition EXPLICIT ASKING does not differ from

[116] Shapiro-Wilk tests for normality yield statistically non-significant results ($W_\alpha = 0.859; \alpha = 0.05$) for the distributions of feedback rate in all three conditions ($\text{AS} : W = 0.954, p = 0.7; \text{EA} : W = 0.919, p = 0.38; \text{NA} : W = 0.866, p = 0.06$). Hence, we assume that data was drawn from a normally distributed population and will use parametric tests. ¶ Even though Levene's test does not reject the null hypothesis of equal variance $F(2, 30) = 3.27, p = 0.052$, we assume from the distributions of feedback rate (see fig. 9.6) that the variance between experimental conditions differs and compensate for this by using Welch's approximation method.

Table 9.5: Results of post-hoc NHST and Bayes factor analyses of feedback rate (feedback signals per presentation). Tests AS : EA and AS : NA are one-sided and two sample Welch's t-tests and use 'greater' as the alternative hypothesis, e.g., AS > EA. The test EA : NA is two-sided and two sample. Bayes factor t-tests of conditions AS : EA and AS : NA analyse both alternative hypotheses against the null hypothesis, and against each other (see fn. 118).

Comparison	Welch's t-test				Bayes factor t-test		
	df	t	p	r	$\text{BF}_{>0}$	$\text{BF}_{0<}$	$\text{BF}_{><}$
AS : EA	13.854	13.182	1.575e−9	0.95	3.934e8	225.395	8.868e10
AS : NA	13.901	7.105	2.753e−6	0.83	5.148e4	100	5.016e6
EA : NA	17.141	-1.371	0.1881	0.3	$\text{BF}_{01} = 1.467$		

participants' feedback rate in condition NO ADAPTATION, tested with Welch's two sample t-test, cannot be rejected.

Following this, we analyse feedback rates in a Bayesian framework. A Bayesian ANOVA[117] yields the Bayes factor $\text{BF}_{10} = 1.042\text{e}7$, which is considered 'decisive' evidence for the alternative hypothesis that feedback frequencies differ between experimental conditions against the null hypothesis that only contains the intercept.

Similar to the classical analysis above, we can further analyse this omnibus result with post-hoc tests. Firstly, we analyse our hypothesis that participants in experimental condition ATTENTIVE SPEAKING produced more feedback per presentation than participants in condition EXPLICIT ASKING, i.e, AS > EA. We do this using a BayesFactor two sample t-test.[118,119] For the one-sided alternative hypothesis of a positive effect, i.e., that participants in condition ATTENTIVE SPEAKING produced more feedback

[117] Rouder et al.'s (2012) Bayes factor test for ANOVA designs as implemented in the 'BayesFactor' R package (Morey and Rouder 2015).

[118] Rouder et al.'s (2009) Bayes factor t-test as implemented in the 'BayesFactor' R package (Morey and Rouder 2015). The test calculates how likely the observed effect size is given a prior distribution over the true effect size δ (modelled with a scaled Cauchy distribution). The one-sided two sample version of the test compares both possible alternatives — the observed effect size of one sample is *greater* or *less* than the effect size of the other sample — against the null hypothesis (no effect) and, therefore, yields two Bayes factor values: $\text{BF}_{>0}$ for a positive effect ($\delta \in]0, \infty]$); and its complement $\text{BF}_{<0}$ for a negative effect ($\delta \in [-\infty, 0]$). A Bayes factor that evaluates the more specific hypothesis *greater* vs. *less* (or vice versa) can simply be calculated by division (Morey 2014), e.g., $\text{BF}_{><} = \text{BF}_{>0}/\text{BF}_{<0}$.

[119] Using the default, 'medium'-scaled prior distribution ($r = \sqrt{2}/2$), for each one-sided alternative hypothesis (positive/negative effect) against the null hypothesis (no effect), and then against each other.

signals per presentation than participants in condition EXPLICIT ASKING, this yields the Bayes factor $\text{BF}_{>0}$ = 3.934e8, which is considered 'decisive' evidence against the null hypothesis that there are no differences (see fn. 111). The complementary alternative hypothesis of a negative effect yields the Bayes factor $\text{BF}_{<0}$ = 0.004, which is considered 'decisive' evidence in favour of the null hypothesis. Directly comparing the two alternative hypotheses yields the Bayes factor $\text{BF}_{><}$ = 8.868e10, which is 'decisive' evidence in favour of our hypothesis that participants in condition ATTENTIVE SPEAKING produced more feedback per presentation than participants in condition EXPLICIT ASKING.

Secondly, we analyse the hypothesis that participants in experimental condition ATTENTIVE SPEAKING produce more feedback per presentation than participants in NO ADAPTATION, i.e, AS > NA, in the same way. For the one-sided alternatives of a positive effect this yields the Bayes factor $\text{BF}_{>0}$ = 5.148e4, which is considered 'decisive' evidence against the null hypothesis. Similarly the complementary alternative hypothesis of a negative effect yields $\text{BF}_{<0}$ = 0.01, which is considered 'strong' evidence in favour of the null hypothesis. The direct comparison of the two alternative hypotheses yields the Bayes factor $\text{BF}_{><}$ = 5.016e6, also 'decisive' evidence for our hypothesis that participants in condition ATTENTIVE SPEAKING will produce more feedback per presentation than participants in NO ADAPTATION.

Finally, we analyse the experimental condition EXPLICIT ASKING against condition NO ADAPTATION. Here our hypothesis is that there should be no difference in feedback rate since the agents in both conditions ignored participants' feedback signals. The analysis yields the Bayes factor BF_{01} = 1.467, which can be considered 'anecdotal' evidence for the null hypothesis that there is no difference in feedback rate. Such weak evidence, however, suggests that we do not have enough data to make a definite statement.

Both analyses show that those participants that interacted with the attentive speaker agent (in experimental condition AS) clearly produced more feedback signals per presentation than participants that interacted with the agents that were ignorant of participants' feedback (in experimental condition EA and NA). No difference in feedback rate between these latter two conditions were found. We can conclude from this that the attentive speaker agent's capabilities and behaviour had a decisive effect on the rate of communicative listener feedback being provided.

Does participants' feedback rate change over time?

Participants went into the interaction with the idea that they can provide feedback and that the agent can take their feedback signals into account. The preceding analysis

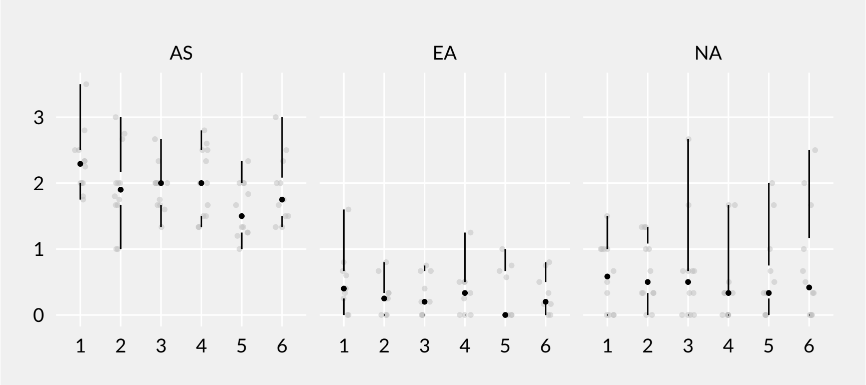


Figure 9.7: Development of participants' feedback rate (feedback signals per presentation) from Block$_1$ to Block$_6$, by condition. Data points are x-jittered in translucent light grey; black dots are medians, black lines are whiskers, mid gaps are quartiles.

leaves open the question whether participants in the control conditions EA and NA had a lower feedback rate throughout the interaction, or whether they noticed at some point that providing feedback has no effect. One way to look at this is to analyse how participants' feedback rate changes over the course of the experiment. As each interaction consists of six consecutive blocks (see fig. 9.4), we can analyse the feedback rate in each of these blocks and look at its development from the beginning of the interaction towards its end.

As can be seen in fig. 9.7, participants' feedback rates do not vary much by block number. In condition ATTENTIVE SPEAKING, the standard deviation of the mean feedback rate across blocks and participants is $SD = 0.23$ and it is even smaller in conditions EXPLICIT ASKING ($SD = 0.08$) and NO ADAPTATION ($SD = 0.11$).

There are two explanation for this. Either participants noticed within the first few information presentation units of the first block whether their feedback behaviour made a noticeable difference or not. Or participants in condition ATTENTIVE SPEAKING simply responded to feedback elicitation cues that the attentive speaker agent produced and, as no such cues were produced in the two control conditions, participants in these conditions did not provide feedback.

Do participants (just) respond to feedback cues?

To investigate this issue, we analyse how effective the feedback elicitation cues of the attentive speaker agent in condition AS are. To this end, we extract the points in time at which the agent produces such cues from the log files of the interactions and jointly analyse elicitation cue timing and the timing of participants' feedback signals. The approach we take is rather simple. Given a feedback elicitation cue that the attentive speaker agent produces, we check whether participants produce a feedback signal within a window of 4 s after the decision to produce a cue is made (a rather generous window size, but it takes some time for the cue to actually be realised). If a listener feedback signal falls within the window, the elicitation cue is considered 'effective'.

Across interactions in condition AS the attentive speaker agent has a mean elicitation cue rate (defined analogously to feedback rate; see above) of M = 1.8 (Mdn = 1.86, SD = 0.26; see table 9.6), that is, on average 1.8 feedback elicitation cues are produced for each presentation. On average, 61 % of the cues were effective (i.e., followed by participant feedback).

Table 9.6 also shows that, on average, 54 % of participants' feedback signals were preceded by an elicitation cue of the agent, which, in turn, means that an average of 46 % of participants' feedback signals were produced 'pro-actively', i.e., not preceded by a feedback elicitation cue (other features in the agent's behaviour may serve as cues as well but are consistent across experimental conditions). Based on this, we can calculate that, on average, participants pro-actively produced 0.91 feedback signals per presentation. This rate is 9.1, respectively 1.4, times as high as the mean feedback rate of participants in conditions EA and NA (see above). Hence, the difference in feedback rate between condition AS and the control conditions cannot be reduced to the factor that the attentive speaker agent deliberately produces feedback elicitation cues. This result suggests that participants indeed noticed that their feedback behaviour has an influence on the agent's behaviour.

The analysis further indicates that the form of an elicitation cue influences its effectiveness. In those cases where the agent simply made a pause and focussed its gaze on the listener (L1-cues), an average of 49 % of the cues was effective, whereas L2-cues, which additionally contained a verbal request (e.g., *okay?*), were, on average, effective 82 % of the time (see table 9.6). As has been found in human–human conversation (Gravano and Hirschberg 2011, pp. 623–625), increasing the number of features in feedback elicitation cues produced by an artificial conversational agent also increases their effectiveness.

Table 9.6: Feedback elicitation cue rate and effectiveness (proportion of cues that are followed by listener feedback in a 4 s window) across interactions in experimental condition AS. The bottom half of the table shows participants' feedback rate as well as the proportions of feedback that can be considered elicited or pro-active (i.e., preceded by a deliberately produced elicitation cue, or not).

	rate			effectiveness (proportion)		
Cue	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*
Total	1.8	1.86	0.26	0.61	0.59	0.19
L1 (pause)	1.29	1.31	0.17	0.49	0.52	0.19
L2 (verbal)	0.51	0.54	0.19	0.82	0.77	0.20
Feedback	1.97	1.93	0.22	proportion of feedback		
elicited	1.06	1.1	0.2	0.54	0.57	0.10
pro-active	0.91	0.87	0.24	0.46	0.43	0.10

Intermediate summary

We can summarise the analysis of participants' feedback behaviour by asserting that (i) in conversation with attentive speaker agents, human interaction partners provide communicative listener feedback that is similar in surface form to feedback that occurs in human–human interaction; (ii) the behaviour of the agent is decisive for its human interaction partner's feedback behaviour, (iii) participants interacting with agents that do not respond to communicative listener feedback quickly notice that providing feedback has no effect, and they immediately stop doing it, and (iv) feedback elicitation cues are effective (depending on their form even highly effective), but the rate of pro-actively produced feedback still exceeds the feedback rate in both control conditions, which suggests that participants in the ATTENTIVE SPEAKING condition noticed that their feedback behaviour has an effect on the agent and the interaction.

9.5.2 OBJECTIVE QUALITY OF THE INTERACTION

In this section, we analyse the objective quality of the interaction, starting with two analyses of costs: interaction duration and repetitions. This is followed by an analysis of participants' performance, i.e., understanding operationalised in terms of their recall of calendar events. Finally, we bring these results together in an analysis of the efficiency of the interactions.

Analysing costs: interaction durations

We begin the investigation of costs by analysing whether the experimental condition (i.e., the agent's behaviour) has an influence on the durations of the interactions. Our hypothesis is that the two control conditions NO ADAPTATION and EXPLICIT ASKING mark two extremes of the duration spectrum, with interactions being shortest in condition NA (where the agent never repeats presentation of information presentation units) and longest in condition EA (where the agent always asks participants if an information presentation unit should be repeated again). We expect experimental condition ATTENTIVE SPEAKING to fall in between these two ends of the spectrum. In this condition, information presentation units are only repeated if (i) the agent attributes perception or understanding below a certain threshold to the participant, or (ii) the agent is unsure about the participant's levels of perception or understanding and lets them explicitly chose to have an information presentation unit repeated.

Interaction durations were automatically extracted from the log file of the interactions. We define duration to be the sum of the durations of the dialogue phases of the six individual blocks (see fig. 9.4). Not included in the measure are therefore the hello and bye sequences (which have the same length across conditions), and the recall phases (which are not considered part of the interaction itself). Measuring begins at the point where the first utterance of a block (e.g., *In diesem Block werde ich über drei Termine sprechen* ('In this block, I will talk about three appointments')) is planned and send to the behaviour realizer[120] and ends when the last utterance of a block is planned and send to the behaviour realiser (e.g., *Dies war der erste Block …* ('This was the first block …')).

Overall, interaction duration varied between 209.5 s and 781.4 s, with a mean of M = 424.5 s and a standard deviation of SD = 177.9. Splitting the data by experimental condition we can observe that there are differences in duration (see table 9.7). Durations in condition NO ADAPTATION are shortest — with a mean of M = 210 s — and almost constant in length (SD = 0.4). In contrast to this, durations in condition EXPLICIT ASKING are longest with a mean of M = 594.9 s, but here the length varies significantly (SD = 103.9). Interaction durations in condition ATTENTIVE SPEAKING fall in between with a mean of M = 468.5 s and considerable variation (SD = 76.7), too. Figure 9.8 shows the distribution of duration by conditions. This descriptive view

[120] Due to variations in processing time, measuring at this point might lead to a negligible error of a few milliseconds per block.

Table 9.7: Duration (in seconds) of interactions by condition and overall. Duration measures the length of an interaction disregarding the recall phases. See fig. 9.8 for the distribution of the data.

Condition	*M*	*SD*	min	max
AS	468.5	76.7	353.8	605.2
EA	594.9	103.9	445.1	781.4
NA	210.0	0.4	209.5	210.6
Overall	424.5	177.9	209.5	781.4

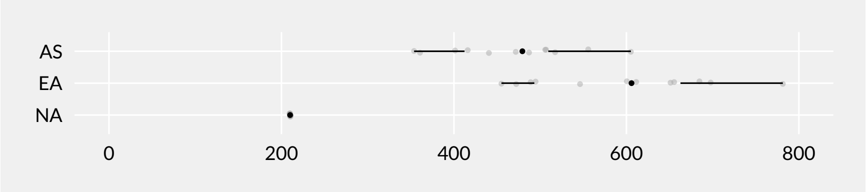


Figure 9.8: Distribution of durations of interactions (in seconds) by condition. Duration measures the length of an interaction disregarding the recall phases. Data points are y-jittered in translucent light grey; black dots are medians, black lines are whiskers, mid gaps are quartiles.

on the data is in line with our hypothesis. To confirm it, we will turn to an inferential analysis of durations.

An independent one-way Welch-approximated ANOVA[121] reveals a statistically significant effect of condition on interaction duration, $F(2, 14.667) = 143.84, p < 0.001, \eta^2 = 0.83$. Post-hoc pairwise Welch's one-sided two sample t-tests between conditions further reveal statistically significant (to a Bonferroni-corrected alpha-level of $\alpha/3 = 0.0167$) mutual differences of interaction duration between conditions

[121] Shapiro-Wilk tests for normality yield statistically non-significant results ($W_\alpha = 0.859; \alpha = 0.05$) for the distributions of interaction duration in all three conditions (NO ADAPTATION: $W = 0.935, p = 0.44$; EXPLICIT ASKING: $W = 0.945, p = 0.57$; ATTENTIVE SPEAKING: $W = 0.97, p = 0.92$). Hence, we assume that data was drawn from a normally distributed population and will use parametric tests in the analysis. ¶ As is apparent from fig. 9.8, the variance in the three groups is not homogeneous. This is confirmed by Levene's test, which rejects the null hypothesis of equal variance $F(2, 33) = 14.05, p < 0.001$. We compensate for this by using Welch's approximation method.

Table 9.8: Results of post-hoc NHST and Bayes factor analyses of interaction durations. All tests are one-sided and two sample. Welch's t-tests use 'less' as the alternative hypothesis, e.g., ATTENTIVE SPEAKING < EXPLICIT ASKING. Bayes factor t-tests analyse both alternative hypotheses against the null hypothesis, and against each other (see fn. 118).

Comparison	Welch's t-test				Bayes factor t-test		
	df	t	p	r	$\mathrm{BF}_{<0}$	$\mathrm{BF}_{0>}$	$\mathrm{BF}_{<>}$
AS : EA	20.246	-3.391	1.431e−3	0.59	28.309	8.695	245.478
NA : AS	11.001	-11.668	7.752e−8	0.93	1.509e8	194.882	2.941e10
NA : EA	11	-12.831	2.914e−8	0.94	8.383e8	219.295	1.838e11

(see table 9.8). With effect sizes of $r > 0.5$, the effects are considered large (Cohen 1992).

A Bayes factor analysis of the interaction durations further confirms these results. A Bayes factor ANOVA yields the Bayes factor BF_{10} = 3.78e10, which is considered 'decisive' evidence for the alternative hypothesis that interaction durations differ between experimental conditions against the null hypothesis that only contains the intercept.

In order to analyse whether our ordering hypothesis is met, we will proceed in two steps. We begin by conducting pairwise one-sided two sample Bayes factor t-tests[122]. As can be seen in table 9.8, the analysis of AS vs. EA yields 'strong' evidence for a negative effect against the null hypothesis, 'substantial' evidence for the null hypothesis against a positive effect, and 'decisive' evidence for a negative against a positive effect. The analyses of NO ADAPTATION against conditions ATTENTIVE SPEAKING and EXPLICIT ASKING both yield 'decisive' evidence for a negative effect against the null hypothesis, for the null hypothesis against a positive effect, and for the evidence of a positive effect against a negative effect. This supports our hypothesis.

As a second step, we conduct a Bayes factor analysis that evaluates the 'specific restricted ordering hypothesis' NA < AS < EA against the more general hypothesis analysed by the omnibus Bayesian ANOVA above (that there is a difference among conditions).[123] For this analysis we drew 10000 samples from the posterior distribution,

[122] Using the default, 'medium'-scaled prior distribution ($r = \sqrt{2}/2$), for each one-sided alternative hypothesis (positive/negative effect) against the null hypothesis (no effect), and then against each other.

[123] Following Morey (2015), this can be done by (i) sampling from the posterior distribution of the data, (ii) computing the proportion of samples that are consistent with the specific ordering hypothesis,

9999 of which matched our specific restricted ordering hypothesis R, resulting in a posterior probability of $p(R|\text{data}) = 0.9999$. Contrasting this with the full model F and considering the riskiness of the model ($1/n!$), yields[124] a Bayes factor $\text{BF}_{\text{RF}} = 5.999$, which can be considered 'substantial' evidence in favour of the model of the specific ordering hypothesis and against the full model. Using BF_{RF} as an 'evidential boost', we can now calculate[125] the Bayes factor $\text{BF}_{\text{R0}} = 2.265\text{e}11$ of our specific order restriction against the null hypothesis (no effect of condition) from the omnibus Bayesian ANOVA above, which is considered 'decisive' evidence for the specific ordering restriction NA < AS < EA of interaction durations in contrast to the null hypothesis of no effect.

Analysing costs: repetitions

We continue the investigation of costs by analysing whether the experimental condition (i.e., the agents' behaviour) has an influence on the amount of information that the agent repeats during an interaction. A simple way to extract this information is by counting the number of information presentation units that the agent repeats (or the total number of presentations).

Repetition of information can only happen in two of the three experimental conditions: in condition AS, the attentive speaker agent repeats an information unit in two cases: (i) when it attributes a level of understanding to an interlocutor that it deems insufficient for current purposes, and (ii) when it is uncertain whether the level of understanding of the interlocutor is sufficient or not. In case (i), the agent tells its interlocutors that it attributes non-understanding and automatically repeats the information (in an adapted way). In case (ii), the agent tells its interlocutor that it is uncertain about their level of understanding and ask them whether they want it to present the information again (see table 8.1 on page 165)

As with the analysis of durations, our hypothesis is that the two control conditions NO ADAPTATION and EXPLICIT ASKING mark two extremes of the amount of information spectrum, with interactions in condition NA — trivially — having the lowest number of repetitions (zero) and condition EA the highest number of repetitions. We expect experimental condition ATTENTIVE SPEAKING to fall in between these two ends of the spectrum.

(iii) computing the posterior odds of a model R of our specific restricted ordering hypothesis and of the full model F of all possible orderings, and (iv) factoring in the riskiness of the specific ordering hypothesis. ¶ Given n factors, there are $n!$ possible orderings, i.e., $3! = 6$ in our case: NA-AS-EA (the specific ordering hypothesis under consideration), NA-EA-AS, EA-NA-AS, EA-AS-NA, AS-EA-NA, and AS-NA-EA.

[124] $\text{BF}_{\text{RF}} = \frac{P(R|\text{samples})}{P(F|\text{samples})} \cdot n! = \frac{0.9999}{1} \cdot 3! = 5.999$.

[125] $\text{BF}_{\text{R0}} = \text{BF}_{\text{RF}} \cdot \text{BF}_{10} = 5.999 \cdot 3.775\text{e}10 = 2.265\text{e}11$.

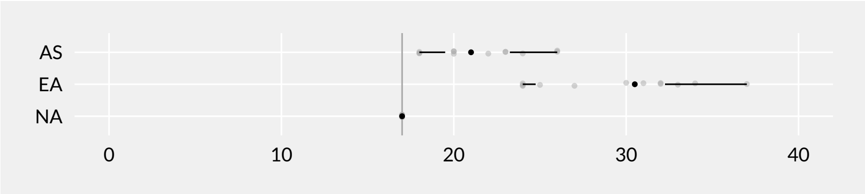


Figure 9.9: Distribution of total number of information presentations by condition. Each condition consisted of the same 17 individual information presentation units (grey vertical line; see table 9.2), each of which could be presented multiple times in conditions ATTENTIVE SPEAKING and EXPLICIT ASKING. The number of repetitions can be easily derived from the number of presentations Data points are y-jittered in translucent light grey; black dots are medians, black lines are whiskers, mid gaps are quartiles.

The mean number of repetitions in condition EA is M = 12.4 (SD = 4.5), whereas it is only M = 4.5 (SD = 2.9) in condition AS, a mean M = 3.2 of which was, initiated automatically by the agent (SD = 2.8). As explained above, there are no repetitions in condition NA. Figure 9.9 shows the distribution of the number of presentations (and therefore repetitions) per condition.

A Welch's one-sided two sample t-test (using 'less' as the alternative hypothesis, i.e., AS < EA) reveals that this difference is statistically significant ($t = -5.1173$, df = 18.984, p = 3.07e−5, r = 0.74) and has an effect size that is considered large ($r > 0.5$; Cohen 1992). Similarly a BayesFactor t-test yields a $\text{BF}_{<0}$ = 892.86 for the negative effect against the null hypothesis, which is considered 'decisive' evidence.

Intermediate summary: Costs

Summarising the analysis of costs, our hypothesis could be confirmed. The experimental condition has an influence on both interaction duration and the amount of redundant information (in terms of repetitions). As hypothesised, interactions were shortest in control condition NA and longest in control conditions EA. The interactions with the attentive speaker agent fell right between these extreme ends of the spectrum.

Having addressed the 'cost' aspect of the promise of computational models of interactional intelligence, we can now turn to the promise of better performance, measured in terms of participants' level of understanding.

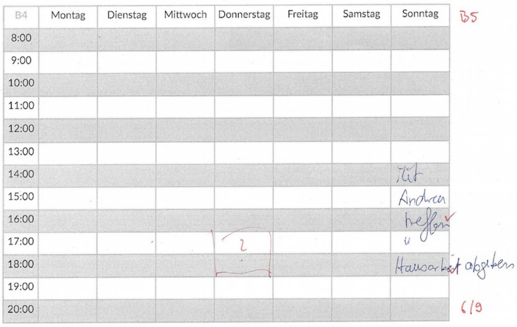
Our hypothesis is that participants will perform worst in experimental condition NO ADAPTATION, since problems in perception or understanding could not be solved through simple repetition or interactive adaptation. It is more challenging to formulate a hypothesis about the relation between recall performance in conditions ATTENTIVE SPEAKING and EXPLICIT ASKING. On the one hand, participants in condition EA can explicitly choose to have an information presentation unit repeated as often as they wish, whereas participants in condition AS can only choose whether they want a repetition when the agent is unsure about their listening-related mental state. As we have just analysed there are less repetitions in condition AS than in condition EA. On the other hand, presentations in condition AS receive presentations that are adapted to their immediate needs. Furthermore, dialogue phases in condition AS were shorter (see above), so events needed to be remembered for shorter periods of time. Weighting these influences, we expect recall scores to be moderately higher in condition EXPLICIT ASKING.

As already mentioned above, interaction duration on its own is hardly a useful criterion for evaluating an attentive speaker agent's performance. The essential idea underlying this thesis is that computational models of interactional intelligence, such as for example the model for interactive feedback-based coordination in an attentive speaker agent, will lead to better understanding among interlocutors, in a cost-efficient manner (see section 1.2).

Analysing performance: understanding in terms of recall

As explained in section 9.3.1, we operationalise participants' understanding by measuring their recall of the calendar events that the agents' present, gathered with a pen-and-paper task in the recall phase following the dialogue phase in each of the six blocks (see fig. 9.4). The scoring of the filled calendars was done manually (by HB), with the help of printed stencil overlays that show the correct positions in the week (weekday, start and end times) and titles of the calendar events. For each calendar event that the agents described three points could be achieved: one for the weekday, one for the start time, and one for the title of the event.[126] The maximum number of points for each block is thus three times the number of calendar events (see column 'Items (counts)' in table 9.2), resulting in a maximum of $24 \cdot 3 = 72$ points.

[126] End times and durations of events were not scored as they were not always mentioned explicitly during the agents' presentation.



B4	Montag	Dienstag	Mittwoch	Donnerstag	Freitag	Samstag	Sonntag
8:00							
9:00							
10:00							
11:00							
12:00							
13:00							
14:00							Mit
15:00							Andrea
16:00							treffen ✓
17:00				?			
18:00							Hausarbeit ✓ abgeben
19:00							
20:00							

B5

6/9

Figure 9.10: Scoring of a participant's recall for block number 5 (the fourth block presented in this case). The participant reached a score of 6 out of 9 points. Two events were recalled correctly, one event is missing completely.

Scoring was done liberally, with rules set up beforehand. Each piece of correct information in a block yielded one point, even when mixed-up among events (e.g., when the titles of two events were interchanged, both points for titles were given nonetheless). Non-queried extra information given by participants (e.g., specifying information from the conflicting event in case of a 'cancel' operation) could make up for wrong/missed information within one block. Figure 9.10 shows the scoring of one of the filled calendars.

Overall, participants reached scores between 29 and 68 points, with a mean of $M = 56.3$ and a standard deviation of $SD = 10.2$ points (see table 9.9). Splitting participants by experimental condition reveals differences in the scores reached, with participants in condition EXPLICIT ASKING reaching highest mean score of $M = 62.8$ ($Mdn = 63.5$, $SD = 5.5$, min = 52, max = 69). Participants in condition ATTENTIVE SPEAKING follow with a mean score of $M = 58.4$ ($Mdn = 59$, $SD = 7.6$, min = 46, max = 68). Participants in condition NO ADAPTATION reached the lowest mean score of $M = 47.5$ ($Mdn = 49.5$, $SD = 10.4$, min = 29, max = 62). Figure 9.11 shows the distribution of scores by conditions. These numbers are in line with our hypothesis. To confirm it, we will turn to an inferential analysis of the recall scores.

Table 9.9: Scores reached in the calendar recall-task by conditions.

Condition	*M*	*SD*	min	max
AS	58.4	7.6	46	68
EA	62.8	5.52	52	69
NA	49.5	10.4	29	62
Overall	56.3	10.2	29	68

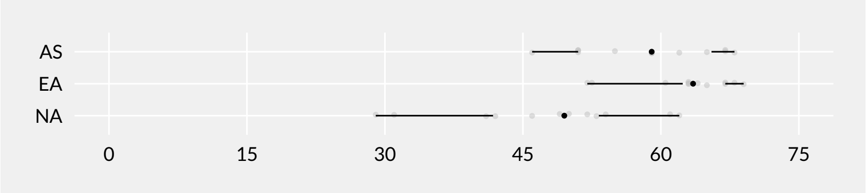


Figure 9.11: Distribution of scores reached in the calendar recall-task by conditions. Data points are y-jittered in translucent light grey; black dots are medians, black lines are whiskers, mid gaps are quartiles.

A Kruskal-Wallis rank sum test[127] indicates that the reached scores are statistically significantly different between experimental conditions $\mathrm{H}(2) = 14.315, p < 7.791\mathrm{e}{-4}$ ($\alpha = 0.05$).

Post-hoc pairwise approximative Wilcoxon-Mann-Whitney tests (10000 Monte Carlo replicas; a Bonferroni-corrected alpha-level of $\alpha/3 = 0.0167$) reveal statistically significant differences of the scores reached by participants between conditions AS and NA, as well as between condition EA and NA (see table 9.10). With effect sizes of $r > 0.5$ the effects of the two comparisons are considered large (Cohen 1992).

The difference between the scores reached by participants between conditions AS and EA, however, are not significantly different.

As before, we also analyse recall scores in a Bayesian framework. A Bayesian ANOVA yields the Bayes factor $\mathrm{BF}_{10} = 160.681$, which is considered 'decisive' evidence

[127] A Shapiro-Wilk test for normality yields statistically significant results ($W_\alpha = 0.859; \alpha = 0.05$) for the distribution of calendar scores in condition EA ($W = 0.846, p = 0.033$). Hence, we cannot assume that data was drawn from a normally distributed population and will use non-parametric tests.

Table 9.10: Results of post-hoc pairwise NHST and Bayes factor analyses of recall scores. Bayes factor *t*-tests analyse both alternative hypotheses against the null hypothesis, and against each other (see fn. 118).

Comparison	Wilcoxon-Mann-Whitney test			Bayes factor *t*-test		
	W	p	r	$\mathrm{BF}_{>0}$	$\mathrm{BF}_{0<}$	$\mathrm{BF}_{><}$
AS : EA	99	0.121	−0.32	0.171	0.571	0.097
EA : NA	133.5	2e−4	−0.79	261.717	9.909	2593.342
AS : NA	115	0.0119	−0.51	12.718	8.065	102.572

for the alternative hypothesis that the recall scores reached differ between experimental conditions against the null hypothesis that only contains the intercept.

In order to analyse whether our ordering hypothesis is met, we will proceed in two steps. We begin by conducting pairwise one-sided two sample Bayes factor *t*-tests[128].

As can be seen in table 9.10, the analysis of AS versus NA yields 'strong' evidence for a positive effect against the null hypothesis (i.e., AS > NA), 'substantial' evidence for the null hypothesis against a negative effect, and 'decisive' evidence for a positive against a negative effect. The analysis of EA versus NA yields 'decisive' evidence for a positive effect against the null hypothesis (i.e., EA > NA), 'substantial' evidence for the null hypothesis against a negative effect, and 'decisive' evidence for a positive against a negative effect. Interestingly, the analysis of AS versus EA[129] yields substantial evidence against a positive effect and for the null hypothesis, basically no evidence for a positive effect against the null, but 'strong' evidence for a negative effect against a positive effect. This is an interesting result, as the NHST-based analysis did not find a difference between these two conditions. The difference in interpretation of the data under the two statistical perspectives shows that the difference in recall scores between the attentive speaker agent and the upper-bound baseline may actually be quite small. Nevertheless, carrying out the additional Bayesian analysis proved important, as it finds strong evidence for a negative effect, i.e., for AS < EA.

This supports our hypotheses and suggests a specific ordering of experimental condition by score, namely R = EA > AS > NA, which we can analyse separately, as in

[128] Using the default, 'medium'-scaled prior distribution ($r = \sqrt{2}/2$), for each one-sided alternative hypothesis (positive/negative effect) against the null hypothesis (no effect), and then against each other.

[129] Note that the test analysed the hypothesis AS > EA (see table 9.10). It can, however, easily be interpreted as its complement (see footnote 118).

the analysis of interaction duration, see fn. 123.

Evaluating the specific ordering hypothesis R, we drew 10000 samples from the posterior distribution. 8809 of these samples matched the specific restricted ordering hypothesis R, resulting in a posterior probability of $p(R|\text{data}) = 0.8809$. Contrasting this with the full model F and considering the riskiness of the model $(1/n!)$, yields a Bayes factor of $\text{BF}_{\text{RF}} = 5.285$, which can be considered 'substantial' evidence in favour of the model of the specific ordering hypothesis and against the full model. The Bayes factor of our specific order restriction against the null hypothesis (no effect of condition) from the omnibus Bayesian ANOVA above is $\text{BF}_{\text{R0}} = 849.26$. Thus, there is 'decisive' evidence for the specific ordering restriction R of recall scores achieved by participants in contrast to the null hypothesis of no effect.

Intermediate summary: Performance

In summary, we can say that the experimental condition has a decisive influence on participants' recall scores. Participants in the control condition NA, who interacted with the agent which only presents each information presentation unit once and does not respond to participants' feedback at all, performed decisively worse than participants in the attentive speaker condition AS and the second control condition EA.

Participants in condition EA, who interacted with the agent which would repeat each information unit as often as participants wished, performed moderately better in the recall task than participants who interacted with the attentive speaker agent, the effect is smaller however.

Given a recall-based operationalisation of understanding, the attentive speaker agent's capabilities could not compete with the simple power of repetition. And, participants in condition EA made extensive use of the opportunity of getting information presented repeatedly (see fig. 9.9 and the analysis of repetitions on page 211). In light of this information, the performance of participants who interacted with the attentive speaker agent does not seems to be too bad after all, even if it is lower in absolute terms.

Revisiting the promise of computational models of interactional intelligence, we can now say that the attentive speaker agent falls in between the two ends of the spectrum on both aspects: Costs (interaction duration, repetitions) and performance (understanding in terms of recall). In the following we will now compute the trade-off between these aspects and analyse the efficiency of the interactions.

Efficiency: Recall score versus duration or repetition

As mentioned above, we model efficiency as the ratio of performance to costs. This yields two measures for efficiency, one based on duration, the other based on the number of repetitions, namely

$$\eta_{\text{dur}} = \frac{\text{recall score}}{\text{duration}}, \quad \text{and} \quad \eta_{\text{rep}} = \frac{\text{recall score}}{\text{\#repetitions}}. \tag{9.1}$$

Our hypothesis for efficiency is that the attentive speaker agent is more efficient than the two agents of the control conditions.

As a first step, we calculate the ratios η_{dur} and η_{rep} for each conditions, using the mean score, mean duration, and mean number of repetitions (see table 9.11). The first point to notice is that the lower bound control condition NA is far more efficient than the two other conditions. Considering η_{dur}, it is 1.8 times more efficient than condition AS and 2.1 time more efficient than condition EA. And it basically does not make sense to consider η_{rep} for condition NA. As no information gets repeated, it is infinitely more efficient.[130] Still, the amount of information that participants in this control condition could recall is (perhaps surprisingly) high. We will discuss this below.

The difference in efficiency between the attentive speaker agent condition AS and the upper-bound control condition EA is rather small when considering η_{dur} (the attentive speaker is 1.18 times more efficient), but quite large in terms of η_{rep} (where it is 2.56 times as efficient). Are these differences, especially between conditions AS and EA, significant? To investigate this, we calculate the efficiency ratios for each participant and test in the usual way.

An independent one-way Welch-approximated ANOVA[131] reveals a statistically significant effect of condition on efficiency in terms of η_{dur}, $F(2, 30.922) = 19.169, p = 1\text{e}{-6}, \eta^2 = 0.73$. Post-hoc pairwise Welch's one-sided two sample t-tests between conditions further reveal statistically significant (to a Bonferroni-corrected alpha-level of $\alpha/3 = 0.0167$) mutual differences of η_{dur} between conditions (see table 9.12). With an effect size of $r > 0.3$ the difference between conditions AS and EA are considered

[130] If we instead replace number of repetitions by number of presentations (i.e., #repetitions+17), condition NA is only marginally more efficient (only 1.02 times more efficient than condition AS and 1.3 times more efficient than condition EA).

[131] Shapiro-Wilk tests for normality yield statistically non-significant results ($W_\alpha = 0.859; \alpha = 0.05$) for the distributions of η_{dur} in all three conditions (ATTENTIVE SPEAKING: $W = 0.937, p = 0.4644$; NO ADAPTATION: $W = 0.943, p = 0.5409$; EXPLICIT ASKING: $W = 0.96, p = 0.7966$). Hence, we assume that data was drawn from a normally distributed population and will use parametric tests in the analysis. ¶ A Levene's test rejects the null hypothesis of equal variance $F(2, 33) = 6.339, p = 0.0047$, for which we compensate by using Welch's approximation method.

Table 9.11: Mean recall scores, mean durations, mean number of repetitions, as well as derived efficiency values η_{dur} and η_{rep} (see eq. [9.1]) by experimental condition.

condition	recall	duration (s)	η_{dur}	# repetitions	η_{rep}
AS	58.4	468.5	0.125	4.5 (+17)	12.98(2.73)
EA	62.8	594.9	0.106	12.43 (+17)	5.06(2.14)
NA	49.5	210.1	0.226	0 (+17)	∞(2.79)

of medium size. The effect sizes of $r > 0.5$ of the two test involving condition NA are considered large (Cohen 1992).

As before, we also analyse η_{dur} in a Bayesian framework. A Bayesian ANOVA yields the Bayes factor BF_{10} = 1.575e7, which is considered 'decisive' evidence for the alternative hypothesis that η_{dur} differs between experimental conditions against the null hypothesis that only contains the intercept.

A one-sided two sample Bayes factor t-test[132] yields a difference between the attentive speaker condition AS and the control condition EA, this is considered 'substantial' evidence for a positive effect against the null-hypothesis. The evidence of a positive effect against a negative effect is even considered to be 'very strong'. Further tests yield 'decisive' evidence for a positive effects against the null hypothesis for comparisons NA > AS and NA > EA.

Turning to the analysis of η_{rep} we only investigate the difference between the attentive speaker condition AS and the upper-bound control condition EA — it does not make sense to investigate it for control condition NA (where $\eta_{\text{rep}} = \infty$). A Welch's one-sided two sample t-test reveals a statistically significant differences of η_{rep} between conditions AS and EA (see table 9.12). The effect size of $r > 0.5$ is considered to be large (ibid.). This is confirmed by a Bayesian analysis. A one-sided two sample Bayes factor t-test yields evidence for a positive effect against the null-hypothesis that is considered 'substantial', similar for the null hypothesis against a negative effect. Comparing the positive effect against the negative effect yields evidence that is even considered to be 'very strong'.

[132] Using the default, 'medium'-scaled prior distribution ($r = \sqrt{2}/2$), for each one-sided alternative hypothesis (positive/negative effect) against the null hypothesis (no effect), and then against each other.

Table 9.12: Results of post-hoc NHST and Bayes factor analyses of efficiency in terms of the ratios of recall to duration (η_{dur}) and recall to repetitions (η_{rep}). All tests are one-sided and two sample. Welch's t-tests use 'greater' as the alternative hypothesis, e.g., AS > EA. Bayes factor t-tests analyse both alternative hypotheses against the null hypothesis, and against each other (see fn. 118).

	Comparison	Welch's t-test				Bayes factor t-test		
		df	t	p	r	$\text{BF}_{>0}$	$\text{BF}_{0<}$	$\text{BF}_{><}$
η_{dur}	AS : EA	2.346	18.295	0.0152	0.45	4.73	7.124	33.696
	NA : AS	6.174	16.199	6.318e−6	0.8	1.509e8	194.881	2.941e10
	NA : EA	7.919	13.077	1.203e−6	0.86	2.491e5	114.85	2.861e7
η_{rep}	AS : EA	2.843	11.18	0.0079	0.52	10.6	7.91	83.834

Intermediate summary: Efficiency

In summary, the cost-efficiency of the interaction — whether operationalised in terms of the ratio of recall score to duration or to number of repetitions — varies with experimental condition.

The results, consistent across both operationalisations, are that (i) the agent in control condition NA (which did not take into account its interlocutor's feedback at all) is more efficient than the attentive speaker agent (condition AS) as well as the agent in control condition EA (which always asked its interlocutors whether it should repeat or continue), and (ii) the attentive speaker agent is more efficient than the agent in control condition EA. That is, the attentive speaker agent's efficiency lies in between the two control conditions.

Our expectation was that, in terms of cost-efficiency, the attentive speaker agent, should perform better than the agents in both control conditions. This raises the question why the agent in condition NA was most efficient. It was clear that this agent would perform best in terms of interaction duration and number of repetitions. It was also expected that it would perform worst in terms of understanding operationalised via recall performance. Despite being low in relation to the two other conditions, participants recall scores in NA were actually quite high in absolute terms. It was not expected, however, that these were high enough for the agent being competitive in terms of efficiency.

One explanation that we have to offer is that the interactions in the two control conditions were more predictable than the interaction with the attentive speaker agent and that this may have influenced participants' recall performance. In the two control conditions, most participants were probably familiar with the structure of the interactions after the first few presentations (see the discussion of changes in participants' feedback rate over time on page 204 and fig. 9.7), which might have enabled them to fully focus on the recall task. In condition EA, participants learned that they could get as many repetitions as needed. In condition NA, however, they learned that they only have a single chance to understand and memorise each event the agent presented. In addition, the presentation phase was short.

In contrast to this, the agent's exact dialogue moves were more difficult to predict in the attentive speaker condition AS. In each presentation it was uncertain whether the agent will offer a repetition, whether it will ask the participants, or whether participants have to make do with a single presentation. These dialogue management decisions were made automatically, based on the agent's attribution of listening-related mental state to its interaction partners. This attribution might not have always reflected participants' actual listening-related mental states and might have thus resulted in decisions that, from a human perspective, were inappropriate and, perhaps, irritating, or even disruptive, for participants. After an initial phase of familiarisation, participants in the control conditions did not have to deal with this.

Having analysed the objective quality of the attentive speaker agent, we can now turn to participants' subjective perception of the agent and the interaction.

9.5.3 SUBJECTIVE QUALITY OF THE INTERACTION

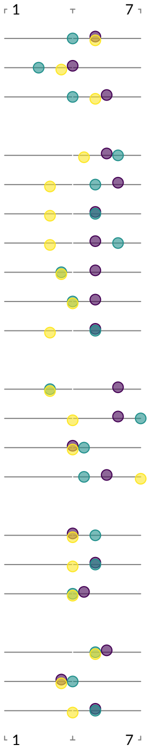
As described in section 9.3.2, participants' subjective perspectives were elicited with the help of 20 items (repeated in table 9.13), each rated on a seven-point Likert scale. Figure 9.12 shows participants' responses to all items grouped by experimental condition.

In contrast to usual analyses of questionnaires of Likert scale items, we will compare ratings of items individually, i.e., not grouped into higher level factors. We do this since the questionnaire was not developed with the intent of items to be grouped.[133] This, however, makes it difficult to do a proper inferential analysis. Carrying out many statistical tests ($20 \times 3 = 60$) would either make type I errors due to multiple testing very likely or lead to a very small Bonferroni-corrected α-level of $\alpha = 0.05/60 = 0.00083$. Given this, classical NHST-based inference would most likely not be very informative. We will therefore carry out an analysis that is descriptive in nature.

[133] The categories according to which the items are discussed in section 9.3.2 and structured in table 9.13 serve presentational purposes.

Table 9.13: Overview of the 20 questionnaire items participants rated on seven-point Likert scales (see section 9.3.2) and their median rating by experimental condition (●AS; ●EA; ●NA).

No.	Item (translated from German)	Rating (*Mdn*)
Speaking competence		
Q1	When Billie speaks, his ideas are clearly and concisely presented	
Q2	When Billie explains something to someone, it tends to be disorganized	
Q3	I could understand Billie well	
Attentive speaking capabilities		
Q4	Billie gave me signals when he wanted to have feedback	
Q5	Billie wanted me to understand him	
Q6	Billie perceived my feedback	
Q7	Billie understood my feedback	
Q8	Billie is able to tell whether or not I have understood what he has said	
Q9	Billie perceived my attitude towards calendar items	
Q10	Billie was attentive to me and adapted to my needs	
Helpfulness		
Q11	Billie helped me resolve difficulties in understanding	
Q12	It was helpful that Billie repeated himself, when needed	
Q13	It was helpful that Billie provided further information, when needed	
Q14	Billie tried to keep the experiment as short as possible	
Naturalness		
Q15	The interaction with Billie was smooth	
Q16	The interaction with Billie was well coordinated	
Q17	Billie's behaviour was similar to the behaviour of a human speaker	
Task and study		
Q18	I perceived the task to be difficult	
Q19	I could remember calendar events and changes to them	
Q20	The experiment was successful	



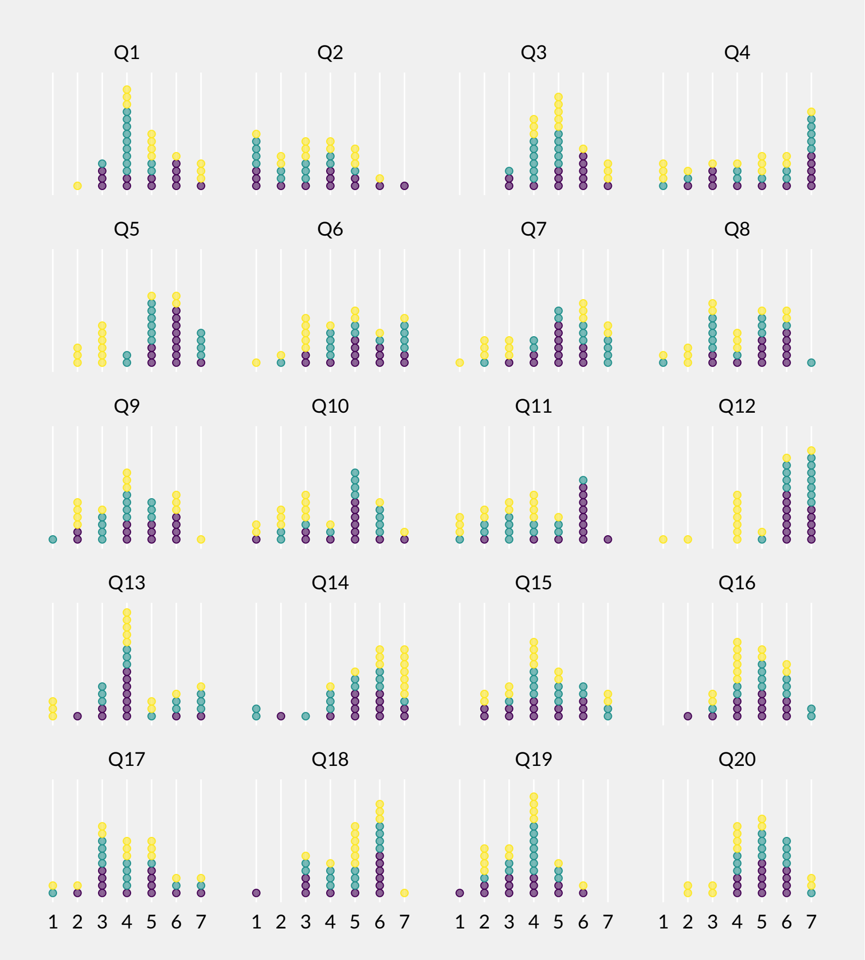


Figure 9.12: Participants' responses to questionnaire items (see table 9.13 and section 9.3.2) by experimental condition. *x*-axes show seven-point Likert scale response anchors (1 [*strongly disagree*] to 7 [*strongly agree*], see fn. 106); stacked dots correspond to number of responses; colours show experimental condition (● AS; ● EA; ● NA).

For each item we will compare the median rating — visualised in table 9.13 and figs. 9.13 to 9.17 — of the experimental conditions. We will substantiate our arguments by making estimations of how likely, given the data, the observed ordering of a rating of an item is in relation to its alternatives. We do this — similar to the procedures used in the Bayesian analysis in previous sections — by first computing the Bayes factor t-test[134], with the default, 'medium'-scaled prior distribution ($r = \sqrt{2}/2$), for each one-sided alternative hypothesis (positive/negative effect) against the null hypothesis (no effect), and then against each other.[135] All computed Bayes factor values are tabulated in table 9.14 and the most relevant ones — where evidence is at least 'substantial' — are visualised in figs. 9.13 to 9.17.

The agent's communicative competence (Q1–Q3)

We will begin the analysis with the first three questionnaire items (Q1–Q3), which deal with the agents' communicative competence. In general, we expected the agent in the attentive speaker condition AS to be a more competent communicator than the agents in the two control conditions EA and NA, i.e., we expected it to present its ideas clearly and concisely, to speak in an organised way, and to be understandable. Figure 9.13 visualises the median ratings and the Bayes factor analyses for each of the three questionnaire items. We start with questionnaire item Q1, the analysis of which we describe in detail in order to illustrate the process we use in analysing all questionnaire items to follow.

Q1 Participants in experimental conditions AS and NA gave questionnaire item Q1 *(When Billie speaks, his ideas are clearly and concisely presented)* a median rating of $Mdn = 5$ ('somewhat agree'), participants of condition EA gave a lower median rating of $Mdn = 4$ ('neither agree nor disagree'). That is, there seems to be no difference between the attentive speaking condition and the lower-bound control condition, but both receive a higher rating than the upper-bound control condition.

[134] We are aware that using t-tests is not recommended for Likert scale data because response anchors are merely ranked and a distance metric between the anchors cannot be assumed. This holds for the Bayes factor t-test, too. We nevertheless use it here because we merely consider it to be a tool for weighing the evidence for specific orderings of experimental conditions.

[135] That is, for each pair of experimental conditions (C_i, C_j), with $C_i, C_j \in \{\text{AS}, \text{EA}, \text{NA}\}, C_i \neq C_j$, we compute a Bayes factor $\text{BF}_{>0}$ (how likely is a positive effect, i.e., that ratings are higher in C_i than in C_j, relative to the null hypothesis), the Bayes factor $\text{BF}_{<0}$ (how likely is a negative effect, i.e., that ratings are lower in C_i than in C_j, relative to the null hypothesis). Based on this we then compute the Bayes factor $\text{BF}_{><}$ (how likely is a positive effect $C_i > C_j$ compared to a negative effect $C_i < C_j$) — see the explanation in fn. 118 for details. These values allows us to evaluate the strength of evidence for the ordering of conditions based on their rating.

Table 9.14: Bayesian analyses of questionnaire items (using Bayes factor *t*-tests). Both one-sided alternative hypotheses (positive/negative effect) are analysed against the null hypothesis (no effect), and against each other. Intensity encodes strength of evidence (see fn. 111) as follows: anecdotal – substantial – strong – very strong – decisive.

Q	AS:EA			AS:NA			EA:NA		
	$\mathrm{BF}_{>0}$	$\mathrm{BF}_{<0}$	$\mathrm{BF}_{><}$	$\mathrm{BF}_{>0}$	$\mathrm{BF}_{<0}$	$\mathrm{BF}_{><}$	$\mathrm{BF}_{>0}$	$\mathrm{BF}_{<0}$	$\mathrm{BF}_{><}$
Q1	1.997	0.165	12.072	0.288	0.508	0.568	0.107	3.449	0.031
Q2	1.544	0.177	8.735	0.373	0.373	1.000	0.118	2.050	0.057
Q3	2.950	0.152	19.386	0.302	0.476	0.636	0.096	7.156	0.013
Q4	0.328	0.429	0.764	1.443	0.180	8.006	1.528	0.126	12.145
Q5	0.765	0.228	3.355	1554.6	0.016	95966	83.150	0.078	1067.6
Q6	0.373	0.373	1.000	4.144	0.143	28.909	2.904	0.110	26.434
Q7	0.206	0.966	0.214	1.214	0.190	6.383	2.300	0.115	20.027
Q8	1.982	0.166	11.959	7.402	0.132	56.119	0.543	0.182	2.980
Q9	1.946	0.166	11.691	0.745	0.231	3.226	0.205	0.440	0.465
Q10	0.342	0.409	0.837	2.992	0.152	19.712	3.526	0.106	33.180
Q11	50.583	0.111	457.68	467.76	0.098	4751.5	0.709	0.161	4.403
Q12	0.302	0.477	0.633	579.31	0.098	5934.4	716.36	0.070	10203
Q13	0.186	1.295	0.144	0.488	0.297	1.642	1.256	0.132	9.482
Q14	1.427	0.181	7.888	0.151	3.064	0.049	0.086	21.416	0.004
Q15	0.180	1.454	0.124	0.373	0.373	1.000	1.112	0.137	8.108
Q16	0.190	1.217	0.156	0.563	0.269	2.092	2.737	0.111	24.642
Q17	0.558	0.271	2.059	0.448	0.317	1.413	0.242	0.348	0.696
Q18	0.313	0.455	0.688	0.217	0.859	0.252	0.166	0.662	0.250
Q19	0.250	0.639	0.392	0.413	0.339	1.219	0.653	0.167	3.920
Q20	0.282	0.525	0.537	2.006	0.165	12.137	2.498	0.113	22.109

The Bayes factor analysis supports this ordering: Comparing the hypothesis AS > EA against its inverse (i.e., AS < EA) yields a Bayes factor $\mathrm{BF}^{\mathrm{AS:EA}}_{><} = 12.072$, which is considered 'strong' evidence in its favour, i.e., that the attentive speaker agent presented its ideas more clearly and concisely than the agent in the upper bound control condition. Similarly, comparing the hypothesis EA > NA against its inverse yields a Bayes factor $\mathrm{BF}^{\mathrm{EA:NA}}_{><} = 0.031$, which is considered 'very strong' evidence in favour of the inverse ordering (i.e., EA < NA). Comparing the hypothesis AS > NA to its inverse, yields a $\mathrm{BF}^{\mathrm{AS:NA}}_{><} = 0.508$, which is considered to be evidence of only 'anecdotal' strength (i.e., we have not enough data to argue for or against either ordering).

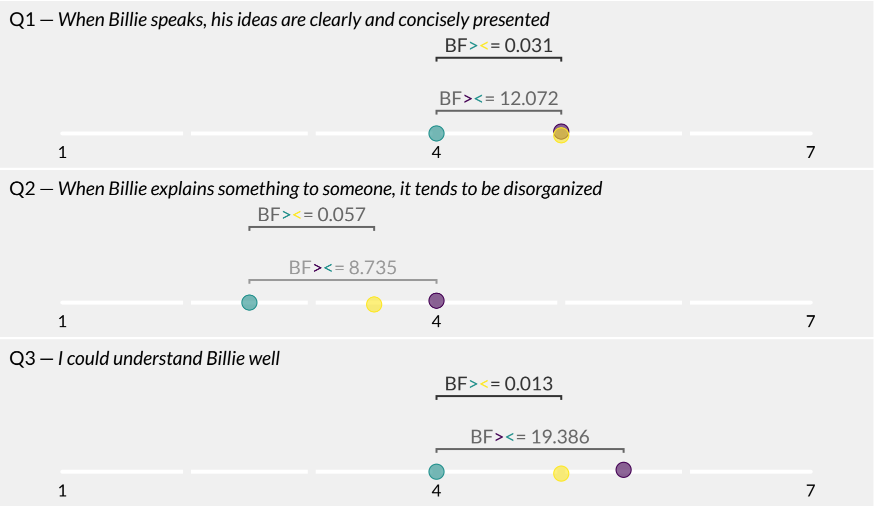


Figure 9.13: Median ratings and Bayes factor based comparison of experimental conditions (●ATTENTIVE SPEAKING; ●EXPLICIT ASKING; ●NO ADAPTATION) of the questionnaire items relating to communicative competence (Q1–Q3). Brackets over two median dots show the Bayes factor t-test value comparing both one-sided alternative hypotheses (positive/negative effect) against each other. Colour-coded angle brackets indicate ordering of conditions (given, e.g., BF><, a value of $K > 0$ is evidence in favour of the ordering AS > EA, a value of $K < 0$ is evidence in favour of the inverse ordering AS < EA). Intensity encodes strength of evidence (see fn. 111) as follows: 'substantial' – 'strong' – 'very strong' – 'decisive'. Brackets for evidence considered merely 'anecdotal' are omitted.

Q2 The analysis of the item *When Billie explains something to someone, it tends to be disorganized* yields 'substantial', respectively 'strong' evidence that participants in AS and NA perceived the agents' speech production to be less organised than participants in EA,. Even though the attentive speaker agent's explanation were rated to be even less organised than that of the agent in NA, there is not enough evidence to claim so.

Q3 The analysis of the item *I could understand Billie well* yields strong, respectively very strong, evidence that participants perceived that they could understand the agents in conditions AS and NA better than the agent in condition EA. Despite receiving a slightly higher median rating, there is not enough evidence to assert that participants interacting with the attentive speaker agent felt they understood better than the agent in the lower-bound control condition NA.

Interestingly, and contrary to our expectations, no differences in perceived basic communicative competence could be asserted between the attentive speaker condition AS and the lower bound control condition NA, whereas there is clear evidence for differences of these two conditions to the upper bound control condition EA.

Participants perceived the attentive speaker agent to be speaking more clearly and concisely (Q1) and also asserted that they could understand it better (Q3) than the agent in control condition EA. It is particularly unexpected that participants of control condition NA rated their own understanding higher than participants in control conditions EA, especially given that their objective understanding — operationalised via recall, see section 9.5.2 — was actually much lower. Given that the control condition in which the agent did not take participants into account (NA) was rated similarly, this might indicate that always asking participants explicitly for their understanding was perceived to be overly verbose and influenced participants' self perception such that they were less able to correctly estimated their own ability to understand the agent.

Explicitly asking for understanding too often, might have raised doubts on their side and diminished their retrospective feeling of understanding while rating questionnaire item Q3. Not questioning participants' understanding, on the other hand, let participants overestimate their level of understanding. The strategy of the attentive speaker agent — informing participants of their estimated level of understanding and only questioning them if necessary seems to have enabled participants to make realistic estimations of their understanding.

The agent's attentive speaking capabilities (Q4–Q10)

Next, we look at the seven questionnaire items Q4–Q10 that deal with the participants' subjective perception of the agent's attentive speaking capabilities. In general we expected that participants in the ATTENTIVE SPEAKING condition will provide higher ratings than participants in both control conditions EXPLICIT ASKING and NO ADAPTATION. We also expect that participants who interacted with the agent that explicitly asked whether it should repeat or continue (EA) is rated higher than the agent that did not adapt at all (NA) — at least for some of the questionnaire items. Figure 9.14 shows the visualisation of the median ratings and the Bayes factor analyses of these six questionnaire items.

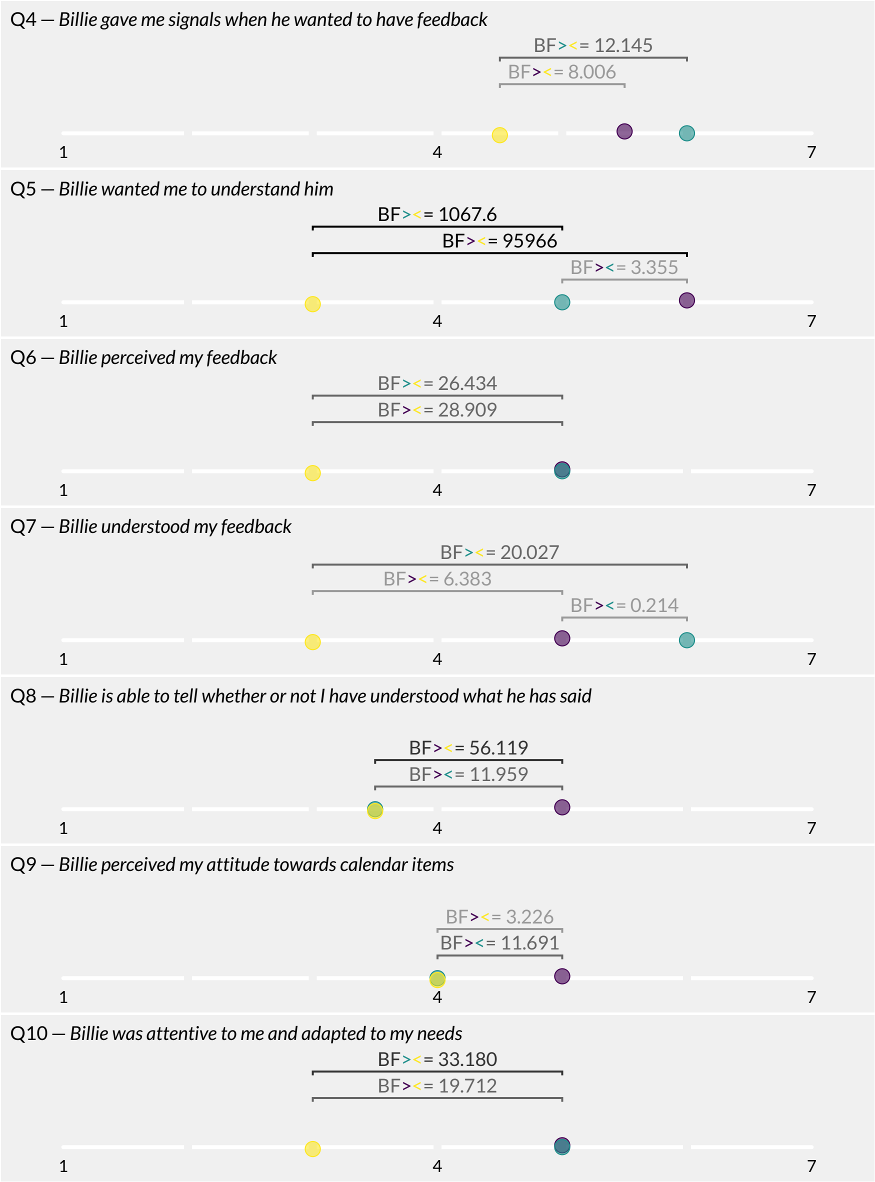


Figure 9.14: Median ratings and Bayes factor based comparison of experimental conditions (●AS; ●EA; ●NA) of the questionnaire items relating to attentive speaking capabilities (Q4–Q10). For an explanation of the plots see the caption of fig. 9.13.

Q4 Examining the answers to the item *Billie gave me signals when he wanted to have feedback*, we see 'substantial', respectively 'strong', evidence that participants in condition AS and EA were more convinced that the agent provided signals in order to elicit feedback than participants in condition NA, who where rather uncertain about this. Even though condition EA received a slightly higher median rating than the attentive speaking condition AS, there is not enough evidence to draw conclusions from it.

Q5 The analysis of the item *Billie wanted me to understand him* finds 'decisive' evidence that participants in conditions AS and EA attributed, to a higher degree, a desire to be understood to the agent that they interacted with than participants in condition NA. Furthermore, there is 'substantial' evidence that condition AS was rated higher than condition EA.

Q6 For the questionnaire item *Billie perceived my feedback* there is 'strong' evidence that participants in conditions AS and EA felt more strongly that the agent they interacted with perceived their feedback than participants in condition NA. The Bayes factor analysis finds no evidence for a difference between the attentive speaking condition AS and the upper bound control condition EA.

Q7 Clear differences in participants' ratings can be observed for the question *Billie understood my feedback*. There is 'substantial', respectively 'strong', evidence that participants in conditions AS and EA were convinced to a higher degree that their feedback is understood by the agent they interacted with than participants in the lower-bound control condition NA in which the agent did not even perceive their feedback. In contrast to our prediction, there is also 'strong' evidence for a higher rating of participants in condition EA than in the attentive speaking condition AS.

Q8 For the questionnaire item *Billie is able to tell whether or not I have understood what he has said* there is 'strong', respectively 'very strong' evidence that participants in the attentive speaking condition AS were convinced to a higher degree that the agent was able to tell whether they understood or not than participants in the two control conditions EA and NA, whereas there is not enough data to claim a difference between conditions EA and NA.

Q9 Similar, only a little less pronounced, were participants' impressions of the item *Billie perceived my attitude towards calendar items*. There is 'strong', respectively 'substantial', evidence that participants in the attentive speaking condition AS rated the agent's ability to perceive their attitude higher than participants in the control conditions EA and NA. Again, the Bayes factor analysis finds no evidence for a difference between the two control conditions.

Q10 For the questionnaire item *Billie was attentive to me and adapted to my needs* there is 'strong', respectively 'very strong', evidence that participants in conditions AS and EA felt more strongly that the agent they interacted with was attentive and adapted to their needs than participants in condition NA. The Bayes factor analysis finds no evidence for a difference between the attentive speaking condition AS and the upper bound control condition EA.

The picture that arises form these question targeting the agents' attentive speaking qualities is as follows. Firstly, we can say that participants rated the agent in condition NO ADAPTATION lowest across all seven items. The agent in this condition was designed to neither elicit feedback from participants, nor to react to their feedback — it basically ignored participants. Given this, it might be surprising that median ratings for these items still ranged between 3 and 4.5, but participants had the expectation that they can provide feedback and that the agent can react to it.

Secondly, we can conclude that participants in the attentive speaking condition AS recognised qualities in the agent that we consider to be constitutive of an attentive speaker agent. Participants felt that the agent wanted them to understand its utterances (Q5), that it perceived and understood their feedback (Q6, Q7), that it is able to reason about their mental state (Q8, Q9), that the agent is interested in their feedback (Q5) and that it was attentive and adaptive (Q10).

Thirdly, we must conclude that the difference of participants' subjective perception of the agents in conditions ATTENTIVE SPEAKING and EXPLICIT ASKING is not clear. Looking at participants' ratings, the agent in condition EXPLICIT ASKING was perceived to have several qualities of an attentive speaker agent, despite not having them by design. This outcome may be due to several reasons. According to our definition, an attentive speaker agent is able to perceive and interpret the feedback of its interaction partners and attribute listening-related mental states to them, it is able to interactively adapt to their interaction partners' needs, and it elicits feedback from its interaction partners — if needed. Firstly, depending on the definition of feedback, the agent in condition EXPLICIT ASKING did neither or all of that. Secondly, it did not perceive and interpret participants' verbal feedback, head gestures, etc. but it perceived their explicit answers to its questions — which can be considered a form of feedback as well. Similarly, the agent in condition EXPLICIT ASKING did not adapt to participants' needs as expressed in their feedback behaviour, but it repeated information presentation units or continued with the next one exactly as participants told it to. Finally, it did not elicit feedback from participants via feedback elicitation cues, but it explicitly asked participants what to do.

The wording of questionnaire items Q4–Q7, and Q10, however, results in vague meanings behind the terms feedback, adaptation to feedback, and feedback elicitation cues. This could be an explanation for the similarity of participants' ratings of the agents in conditions AS and EA.

Questionnaire items Q8 and Q9, on the other hand, address abilities of the agent that are absent in the agent in condition EA. Both items deal with the ability to attribute listener state (understanding and attitude). Although the EA-agent was able to 'perceive/understand [participants'] feedback' (Q6, Q7), it clearly was not 'able to tell whether or not [participants] have understood what it has said' (Q8). It also was not able to perceive participants' attitude towards calendar items (Q9).

The agent's helpfulness (Q11–Q14)

We now turn to the four questionnaire items related to the agents' perceived helpfulness (Q11–Q14). Our prediction was that the agent is, in general, rated more helpful in the ATTENTIVE SPEAKING condition than in the EXPLICIT ASKING and NO ADAPTATION conditions and more helpful in the EXPLICIT ASKING condition than in the NO ADAPTATION condition, i.e, we expect the ordering AS > EA > NA. Figure 9.15 shows the visualisation of the median ratings and the Bayes factor analyses.

Q11 Clear differences in participants' ratings can be observed for the item *Billie helped me resolve difficulties in understanding*. There is 'decisive' evidence that participants in the attentive speaking condition AS felt to a larger degree that the agent helped them resolve difficulties in understanding than participants in the two control conditions. In addition, there is also substantial evidence that the agent in control condition EA was rated higher than the agent in control condition NA in which the agent did not even perceive their feedback.

Q12 Analysing participants' responses to the item *It was helpful that Billie repeated himself, when needed*, we find 'decisive' evidence that the agents in the conditions in which they actually produced repetitions (AS and EA) where rated more helpful than the agent in condition NA. Even though the agent in control condition EA received a higher median rating than the attentive speaker agent, evidence for such a difference can only be considered of 'anecdotal' strength.

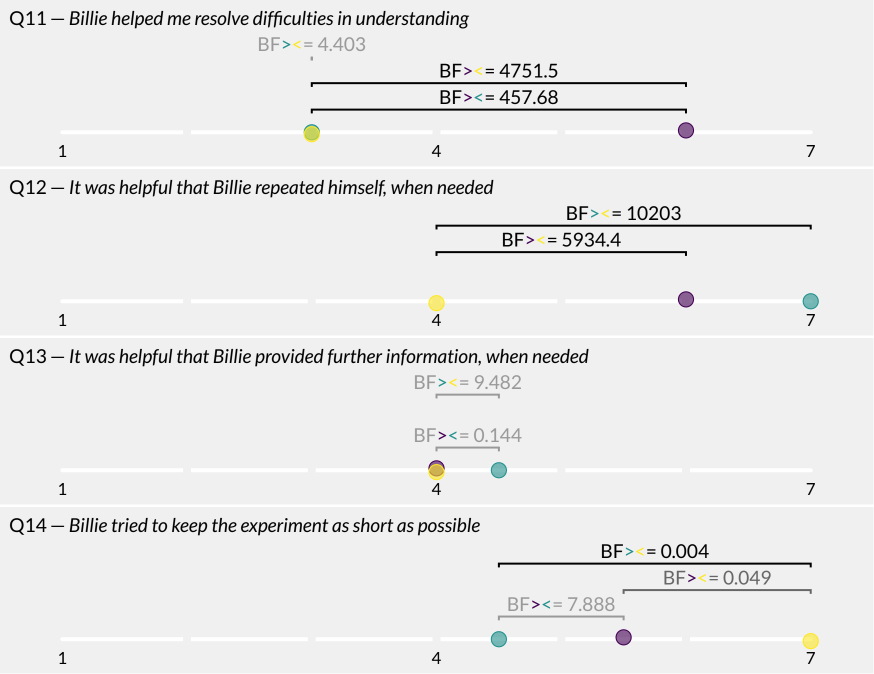


Figure 9.15: Median ratings and Bayes factor based comparison of experimental conditions (●AS; ●EA; ●NA) of the questionnaire items relating to helpfulness (Q10–Q14). For an explanation of the plots see the caption of fig. 9.13.

Q13 This questionnaire item asked whether the provision of additional information by the agent was helpful *(It was helpful that Billie provided further information, when needed)* — an ability that only the agent in the attentive speaking condition AS had. Mostly, participants' neither agreed nor disagreed on this item. Nevertheless, the analysis still finds 'substantial' evidence that participants rated condition EA higher than conditions AS and NA.

Q14 Concerning the question whether *Billie tried to keep the experiment as short as possible* we expect that the lower bound control condition AS will receive the highest and control condition EA the lowest agreement by participants. We expect the attentive speaker condition to fall in between. The analysis of participants' responses yields substantial evidence for a higher rating of AS than EA, strong evidence for a higher rating of NA than AS, and decisive evidence for a higher rating of NA than EA. This subjective characterisation of the agent's behaviour is in line with the objective measures of interaction duration, where the same ordering of conditions EA > AS > NA was observed (see section 9.5.2). Participants accurately perceived how costly interactions were in terms of duration — without having any comparison.

The most important result from these four questions is that the agent in the attentive speaking condition AS was clearly rated — with a large margin — to be most helpful in resolving difficulties in understanding (Q11). This is consistent with their rating of questionnaire items relating to the perception of the agent being an attentive speaker (Q4–Q10), especially with item Q8.

The question, however, is why participants in condition EA — who also felt that the agent wanted them to understand (Q5), that it perceived and understood their feedback (Q6, Q7), and that it was attentive and adaptive to their needs (Q10) — found the agent they interacted with less helpful in resolving difficulties in understanding than participants that interacted with the agent from the attentive speaking condition. Is it because they mis-judged their own level of understanding (Q3), is it because they found the agent to be less able to tell whether they understood or not (Q8), or are the specific adaptation capabilities of the agent in the attentive speaking condition pivotal to its success?

Similar to participants that interacted with the agent in the attentive speaking condition AS, participants in condition EA found it helpful that the agent actually repeated itself upon request (Q12).[136] The agent in the attentive speaking condition could provide additional information as part of its adaptation process — which the agents in the control conditions could not. This capability, however, was not rated to have been particularly helpful in any condition (Q13).[137] Still, the difference in perceived helpfulness in resolving difficulties in understanding may lie in this addi-

[136] There is no differences in perceived helpfulness of repetitions between conditions AS and EA (Q12), despite the possibility that some repetitions in the attentive speaking condition might have been produced due to mis-attribution of listening-related mental states to the participant.

[137] It needs to be taken into account that it might have been difficult for participants in condition AS to notice the introduction of additional information, because (i) the amount of additional information that could have been provided was rather limited, (ii) additional information was not marked as such and only implicitly introduced, and (iii) it is actually possible that — due to participants' state of listening and/or unpredictable communication dynamics — some participants may not have been provided with additional information at all.

tional information that the agent in condition ATTENTIVE SPEAKING could provide. It could also lie in further means of adaptation (i.e., redundancy, see section 6.2.3) that the agent in the attentive speaking condition could use. Another factor might also be that the agent in the attentive speaking condition usually adapted its behaviour 'pro-actively' — instead of requesting whether it should adapt (i.e., repeat information) which the agent in control condition EA did. This question cannot be conclusively answered here.

As before (Q4–Q10), the agent in the lower-bound control condition is rated worst — i.e., least helpful — on all these items. One could argue that it has an edge when equating short interaction duration with helpfulness (Q14). But the argument that duration in itself is not very informative in terms of quality — see section 9.5.2 — applies for the subjective evaluation as well. An agent that wants to keep the interaction short at the cost of understanding cannot be considered helpful. On the contrary a helpful agent can find a trade-off between keeping the interaction short and making the interaction successful, i.e., making itself understood. The agent in the ATTENTIVE SPEAKING condition managed to be perceived as keeping the interaction short (Q14) while still being perceived as helpful in resolving difficulties in understanding (Q11).

The agent's naturalness (Q15–Q17)

Two of the three questionnaire items of the next group deal with the naturalness of the interaction, more specifically whether the interaction is as smooth (Q15) and well coordinated (Q16), as a good interaction with a human speaker would be. An attentive speaker agent should be rated high on both scales. Interaction with them should be smooth because they can — ideally — sense understanding problems of their interaction partners early on and in a non-disruptive way (concurrent feedback instead of explicit other-initiated repair) and can adapt their ongoing utterances before a problem becomes serious. The third item in this group (Q17) asks whether the agent's overall behaviour was similar to that of a human speaker. An attentive speaker agent should be rated high on this scale as well as it can produce behaviour of human speakers which the agents in the two control conditions cannot produce.

Our expectations are thus that participants in the ATTENTIVE SPEAKING condition rate these three items higher than participants in the two other conditions. We also expect that condition EXPLICIT ASKING is rated higher than condition NO ADAPTATION, as the agent in the latter condition basically ignores its interaction partner, which is neither a natural behaviour nor well coordinated and smooth (in cases where participants provide feedback). Figure 9.16 shows the visualisation of the median ratings and the Bayes factor analyses.

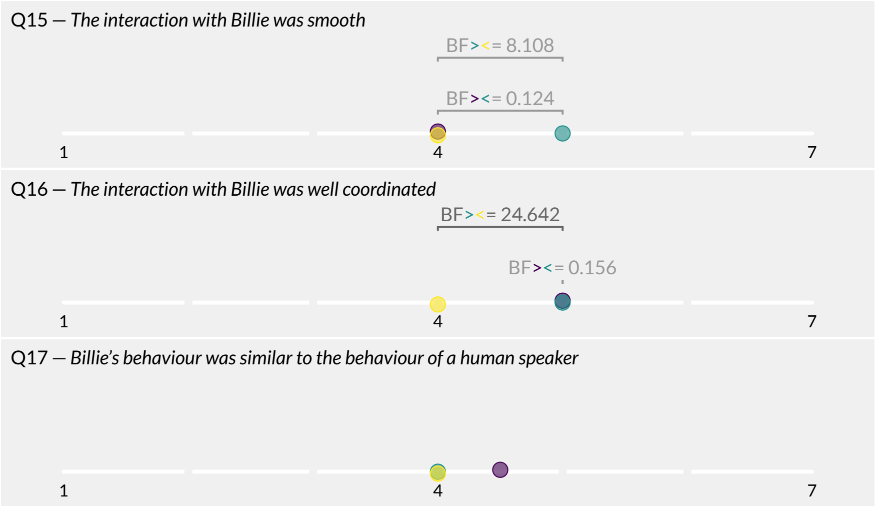


Figure 9.16: Median ratings and Bayes factor based comparison of experimental conditions (●AS; ●EA; ●NA) of the questionnaire items relating to naturalness (Q15–Q17). For an explanation of the plots see the caption of fig. 9.13.

Q15 For the item *The interaction with Billie was smooth*, the analysis finds 'substantial' evidence that participants perceived the interaction with the agent in control condition EA to be smoother (it received a median rating of *Mdn* = 5) than the interactions with the agents in conditions AS and NA (which both received a median rating of *Mdn* = 4).

Q16 The analysis of participants' responses to the questionnaire item *(The interaction with Billie was well coordinated)* yields similar results to the previous item, the attentive speaker condition AS, however, received the same median rating as condition EA. There is, nevertheless, 'substantial' evidence that participants rated the control conditionEA to be better coordinated. It further yields evidence that is considered to be 'strong' that EA is rated higher than condition NA. The data does not allow us to draw conclusions about the ordering of conditions AS and NA.

Q17 The final questionnaire item in this group, *Billie's behaviour was similar to the behaviour of a human speaker*, did not receive pronounced ratings in any of the conditions. Although the median rating of condition AS is slightly higher than the median rating of conditions EA and NA, participants' response data does not contain conclusive evidence for any ordering of conditions (see also table 9.14).

All three questionnaire items were not very informative. Participants' ratings suggest that there are no big differences between the agents in the three experimental conditions. Participants also did not express strong opinions (answer distributions are roughly centred in the middle of the scale, see fig. 9.12), i.e., none of the interactions was perceived as particularly natural, smooth, or well coordinated — but they were also not perceived as unnatural, rough, or uncoordinated.

The results seem plausible, since, in the control conditions, everything proceeded orderly once participants understood the mode of interaction and adapted to it. For the attentive speaking condition we may even read the result as a positive sign: The interaction was not seen as rough or uncoordinated even though the agent might have chosen an unpredictable or wrong communicative action from time to time.

Perception of the task and study

Finally, the questionnaire contained three items on the participants' subjective perception of the task (Q18, Q19) and the study (Q20).

Since we expect the interaction in condition ATTENTIVE SPEAKING to be smoother and better coordinated and the resulting understanding of the participants to be better, we expect that the task would be easier for them in this condition than in the two other. We also expect the task in condition EXPLICIT ASKING to be easier than in condition NO ADAPTATION since participants could listen to information presentation units repeatedly. Figure 9.17 shows the visualisation of the median ratings and the Bayes factor analyses.

Q18 Concerning the questionnaire item *I perceived the task to be difficult*, the analysis yields 'substantial' evidence that participants in condition NA found the task more difficult than participants in AS and EA, for which there is no evidence for a difference in perceived difficulty.

Q19 Concerning the questionnaire item *(I could remember calendar events and changes to them)*, the analysis yields 'substantial' evidence that participants in condition

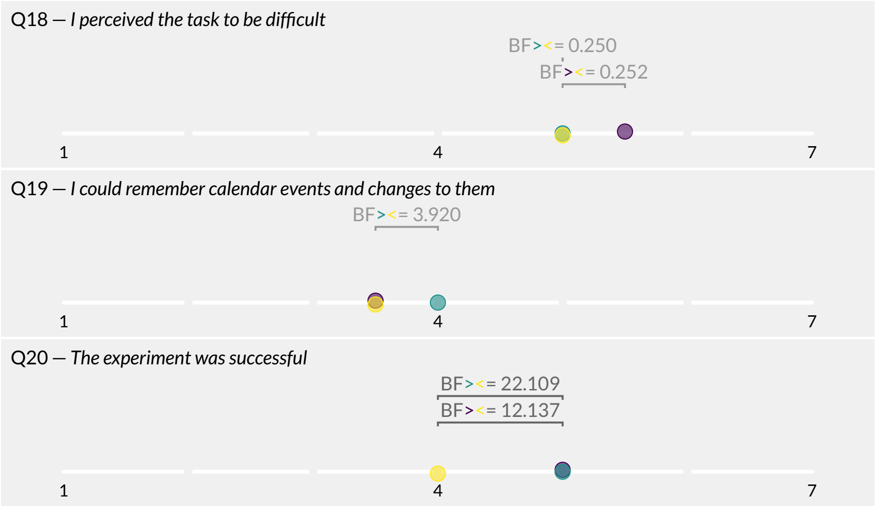


Figure 9.17: Median ratings and Bayes factor based comparison of experimental conditions (●AS; ●EA; ●NA) of the questionnaire items relating to the task and the study (Q18–Q20). For an explanation of the plots see the caption of fig. 9.13.

EA felt that they could remember calender events and changes better than participants in condition NA.

Q20 Finally, the analysis of questionnaire item *The experiment was successful* yields ‘strong’ evidence that participants in conditions AS and EA perceived the experiment to be more successful than participants in condition NA.

Differences in participants' ratings between conditions for questionnaire items Q18 and Q19 are small — but generally correspond to what was found in the analysis of participants' recall scores (see section 9.5.2). For Q20, participants in the lower-bound control condition noticed that something is not the way it should be. They were expecting an agent that listens to their feedback, but instead they were ignored.

Intermediate summary

The results of the subjective evaluation are less clear than those from the objective evaluation. Generally, the attentive speaker agent was rated well. Participants in the ATTENTIVE SPEAKING condition felt that the agent they interacted with speaks clearly and concisely and that they could understand it well. Participants also recognised qualities that are constitutive for attentive speaking, as defined in this thesis. Participants felt that the agent wanted them to understand its utterances, that it perceived and understood their feedback, that it is able to infer their mental state and that the agent is attentive and adaptive. The agent was also rated to be helpful in resolving difficulties in understanding.

Nevertheless, participants' answers were not always consistent across questions, which might be due to ambiguous formulation and unclear terminology. The results, although generally positive for the attentive speaker agent, should thus be taken with a grain of salt.

9.6 GENERAL DISCUSSION

The evaluation study shows that humans are willing and able to provide communicative listener feedback — that is comparable in form to feedback in human communication — in dialogue with artificial conversational agents, both in response to feedback elicitation cues and pro-actively. This confirms previous findings with an attentive speaker agent (Reidsma et al. 2011, § 7.2). Going beyond this result, we find decisive evidence that participants provide more feedback when agents show an actual interest in their interlocutors' feedback behaviour and respond to it by adapting their ongoing communicative actions. This finding is relevant for the design of artificial conversational agents with human-like communicative abilities (Edlund et al. 2008). If a system requires its interlocutors to provide feedback, it should make sure that it displays an interest in it and takes this evidence from interlocutors into account in their behaviour, i.e., the conversational agent needs to be able to speak attentively.

In contrast to this, the results of the analysis of objective quality raise the question whether it is worth modelling a complex attentive speaker agent given that it lies in between the agents from the lower-bound and upper-bound experimental conditions, both in terms of interaction costs and participant understanding (in terms of recall performance). This result was to be expected, as both baselines are strong in their own regard, and are thus hard to beat, but we hypothesised that being an attentive speaker agent is a valuable trade-off between costs and performance in that interactions are more efficient than those with the baseline-agents. This hypothesis could, however, be only partially confirmed. Indeed, we found substantial evidence that the attentive speaker agent is more efficient in its behaviour than an agent that always asks its interlocutors explicitly whether information is understood (the upper-bound condition EA). Participants' understanding in the lower-bound condition (AS), however, was surprisingly high, so that the interactions with this agent were 1.8 times as efficient as interactions with the attentive speaker agent.

When designing artificial conversational agents, the questions thus is how important it is, for a specific interaction scenario or domain, that the human interlocutors will understand the information that the agent presents very well. Even an agent that basically ignores its interaction partners can achieve moderately high performance in its interlocutors' understanding. Improving upon this baseline is costly though. If the scenario requires it, one could argue that the attentive speaker agent's improvement in performance is 'pareto efficient' because the upper-bound baseline (condition EA) is less efficient.

The analysis of the subjective quality of the agents generally indicates that the attentive speaker agent is received positively. Importantly, qualities of attentive speaking were perceived by participants and the attentive speaker agent was rated to be the most helpful agent in resolving understanding difficulties.

Overall conclusion We think that the overall conclusion we can draw from these three perspectives (feedback behaviour, objective and subjective measures) of the interactions in the evaluation study is rather positive. We can confidently make this statement even though the attentive speaker agent did not beat the agents in the baseline conditions in most measures of the objective evaluation. We think that our conclusion is warranted for the following reasons:

- The baselines were designed to be hard to beat on the basic measures of cost (for the lower-bound baseline of condition NA), and performance (for the upper-bound baseline of condition EA). Yet, the attentive speaker agent was able to approach the relevant baselines and even beat the upper-bound baseline on the derived measures of efficiency.

- In contrast to participants in the control conditions, the feedback behaviour of participants that interacted with the attentive speaker agent suggests that they felt that their feedback made a difference in the interaction. The results of the subjective evaluation reflect this as well.

It must be kept in mind that the attentive speaker agent's behaviour is the result of a complex interplay of a number of rather experimental computational models (one of them dynamic probabilistic, the others connected to this dynamic model) in interaction with the human interlocutor (and a wizard). Occasionally, this resulted in generation decisions that were less than optimal (this observation is purely anecdotal). That the interactions with the attentive speaker agent, nevertheless, worked so well and approached the relevant baselines is encouraging for future work in this direction. The occasional problem could not diminish the overall performance and perception of the agent.

The general question whether the attentive speaker agent can be said to have engaged in interactive coordination on the level of belief and attitude (and whether it is interactionally intelligent) can also be answered positively for all three correlates that we expected to see (see section 9.2).

(A) The agent received feedback from its human interlocutors and, in response, adapted its behaviour on multiple levels — thus establishing a feedback loop.

(B) Overall, interlocutors tended to noticed that the attentive speaker agent is interested in and able to infer their mental state of listening and rated it helpful in resolving difficulties in understanding (Q4–Q11).

(C) The attentive speaker agent is more efficient than the upper-bound baseline agent (condition EA) and can thus be said to be pareto-efficient when considering only the three experimental conditions. That the attentive speaker agent could not outperform the lower-bound baseline in terms of efficiency is the result of the extremely short interactions and relatively high performance, which has multiple plausible explanations: e.g., that the behaviour of the agent in these conditions is entirely predictable, or that the interactions are extremely short.

Future work The analysis of data gathered in the evaluation study focussed on comparisons between experimental conditions and neglected a different approach that would have been interesting as well: qualitative analyses of the interactions with the attentive speaker agent. Here we could have studied, inter alia, (i) whether the agent reacted appropriately to its interlocutors' feedback, (ii) the adequacy of the agent's feedback elicitation behaviour in specific dialogue situations, or (iii) the state and the dynamics of the attributed listener state in real interactions. We obtained participants' consent to show individual situations from the interactions to observers and will thus be able to carry out rating studies that evaluate such qualitative aspects in the future.

CHAPTER 10
CONCLUSION

We conclude this thesis by briefly summarising its contributions and results, discussing their implications, as well as their limitations and future research directions.

10.1 SUMMARY

In this thesis we set out to address the broad problem of establishing 'understanding' in dialogue between artificial conversational agents and humans. Based on the insight that conversation in dialogue is, first and foremost, an interactive endeavour, we argued for a perspective shift in conversational agent research: instead of focussing mostly on improvements to conversational agents' natural language processing capabilities, researchers should approach the problem of establishing understanding in human–agent dialogue by endowing agents with 'interactional intelligence'.

In order to investigate how interactional intelligence can be modelled for artificial conversational agents, we have developed and evaluated conceptual and computational models of 'attentive speaking'. The processes that embody these models of attentive speaking enable conversational agents, when holding the turn, to engage in interactive feedback-based dialogue coordination with their human interaction partners. We argued that exploring this restricted form of interactional intelligence is interesting and feasible.

Grounded in theories of conversational interaction in dialogue, and an analysis of the interactional phenomenon of communicative feedback, we defined the concept of attentive speaking and derived three capabilities:

1. to be able to attribute listening-related mental states, based on evidence of understanding in form of communicative feedback signals, to the interlocutors

2. to be able to adapt natural language production such that communicative acts are interactively tailored to the listening-related mental states attributed to the interlocutors
3. to be able to elicit communicative feedback from the interlocutors when not enough evidence of understanding is available to be able to attribute listening-related mental states.

We modelled each of these capabilities conceptually and computationally and described how the individual models are implemented and integrated in an incremental behaviour generation architecture for embodied conversational agents.

Following this, we evaluated the attentive speaker agent in a semi-autonomous Wizard-of-Oz study, comparing it to two control conditions. In the evaluation we could show that, generally, participants perceived the attentive speaker agent to be (i) interested in the feedback that they provided, (ii) able to infer their mental state of listening, and (iii) helpful in resolving their difficulties in understanding. We could further show in the evaluation study, that, from an objective point of view, the attentive speaker agent is more efficient — in terms of cost versus performance — than a control agent that explicitly ensured understanding with its human interlocutors. The attentive speaker agent could, however, not outperform a control agent that did not attempt to ensure its human interlocutors' understanding at all (one possible explanation for this is that this control agent behaved entirely predictable, whereas the dynamic models underlying the attentive speaker agent may have resulted in unpredictable behaviour). Finally, we could show that participants were willing to provide multimodal communicative listener feedback to the attentive speaker agent — both in response to the agent's feedback elicitation cues, but also pro-actively. This was not the case in the control conditions, where participants seemed to notice that the agents did not respond to their feedback behaviour. Feedback signals that human interlocutors produced when interacting with the attentive speaker agent were comparable in frequency as well as form and complexity to feedback in human–human interaction.

In conclusion we can say that the computational models of attentive speaking developed in this thesis are effective and enable artificial conversational agents to interactively coordinate with their human interlocutors on the levels of belief and attitude. We regard this to be an important step towards general interactional intelligence for conversational agents. In the following section we discuss the contributions of this thesis and their implications.

10.2 CONTRIBUTIONS AND IMPLICATIONS

We regard the work presented in this thesis to be relevant to ongoing discourses within multiple research fields.

Communicative listener feedback This thesis contributes to the body of work on communicative feedback in multiple ways.

First of all, this thesis is one of the first few approaches, conceptually as well as computationally, to model the processing of communicative listener feedback of human interaction partners within artificial conversational agents in the speaking role. It is thus complementary to most of the computational work on feedback, which focusses on timing-related aspects of feedback generation.

This thesis builds on theories of the semantics and pragmatics of communicative listener feedback (Allwood et al. 1992; Clark 1996; Bunt 2012) and contributes to them by (i) formally explicating feedback semantics and pragmatics in a state-of-the-art probabilistic, inferential, computational framework, (ii) bringing feedback semantics and pragmatics together with a proposed cognitively motivated theory of feedback production (based on listening-related mental states) and feedback processing (based on minimal mentalising/mental-state attribution), (iii) concretely modelling the hierarchy of feedback functions and their interaction, and (iv) contextualising feedback semantics and pragmatics in dialogue context.

The ALS-approach to the semantics and pragmatics of feedback specifically allows us to embrace the richness in form and meaning as well as the qualitative nature of feedback instead of characterising it in stereotypical ways — such as backchannel, continuer, acknowledgement, assessment, etc. — as traditional accounts do. This is relevant for estimating groundedness (see below), but also for reacting to feedback in general.

The thesis also proposes a novel theory and model of feedback elicitation that — in contrast to previous models of feedback elicitation (Reidsma et al. 2011; Misu et al. 2011b) — does not focus so much on the form of elicitation cues, but on a cognitive motivation for eliciting feedback.

Incremental processing in conversational agents The work presented in this thesis also contributes to the field of incremental dialogue processing, a principle that it applies throughout the implemented artificial conversational agent' architecture (following Schlangen and Skantze's [2011] incremental unit model).

Specifically, the thesis presents a model of incremental multimodal behaviour generation for artificial conversational agents that consists of (i) a real-time architecture that supports the coordinated interplay between multiple behaviour planning components and a component for behaviour realisation that is necessary for incremental multimodal behaviour generation, and (ii) an adaptive natural language generation component that is able to generate the agent's utterances in increments of the size of utterance units, each of which is subject to adaptation at the moment when its generation is requested (when the previously generated utterance unit's articulation is about to finish).

Grounding and interactive adaptation In the context of grounding and interactive adaptation, the thesis make two contributions.

It sketches a computational model of grounding in interaction — modelling grounding as a part of the attributed listener state model — that reflects the gradual nature of common ground (Clark and Schaefer 1989; Clark 1996; Brown-Schmidt 2012), a property that is absent from Traum's (1994) computational model of grounding and less domain-specific than Roque and Traum's (2008) proposal for a model of degrees of grounding. This is possible because the attributed listener state model can process evidence of understanding (in the form of communicative listener feedback) in a way that preserves the non-categorial, qualitative aspects of evidence of understanding that get lost when mapping it to hard categories (such as a set of grounding acts), as previous computational models of common ground do. Additionally, this approach to grounding provides a way to straightforwardly measure 'sufficiency' according to Clark and Schaefer's (1989) 'strength of evidence principle'.

Generally though, the model presented in this thesis does not primarily rely on common ground for adaptive language production and audience design, but adopts the minimal partner models approach (Brennan et al. 2010; Galati and Brennan 2010). Instead of designing communicative actions using full common ground, or engaging in monitoring and adjustment by default, the model of adaptive language production takes variables of the attributed listener state (such as *U* or *AC*) into account during decision making in speech production. Language generation in the attentive speaker agent is adaptive, by default, in that it pays heed to the inferred listening-related mental states of its human interlocutors. These variables of the attributed listener state can be regarded as 'bits', as in Brennan et al. (2010), with the difference that they are continuous (rather than binary). They are still simple, but allow for more specific choices in language generation.

Although the minimal partner model approach to adaptation is also interesting from a computational perspective, we see its adoption in our model of attentive speaking as a programmatic, rather than as a pragmatic, choice. In our opinion minimal partner models — whether based on single 'bits' or on a more expressive continuous variables — are a good characterisation of efficient decisions making in speaking. This makes them a good match for communicative listener feedback, which embodies a similar minimal approach to evidence of understanding and coordination in dialogue.

Intelligent virtual agents The thesis also makes contributions to the field of intelligent virtual agents.

It puts forward a concrete definition of 'attentive speaking' that is similar to the one of Reidsma et al. (2011). The focus of the work carried out in this thesis is different though. While Reidsma and colleagues work on low level processing (detecting and classifying a feedback signal, synthesising a feedback elicitation cue, changing timing of utterances) this thesis models higher level processes. Different directions were taken starting from similar definitions, which suggests that the definition can be spelled out more precisely. Nevertheless, both approaches to attentive speaking seem to be reconcilable, which presents an opportunity for future work.

A second aimed contribution is the thesis' work on incremental multimodal behaviour generation (described in the section on incremental processing above). The approach that is developed here intentionally deviates from the SAIBA model, the de-facto standard in the field of intelligent virtual agents. As incremental processing becomes more common, problems in interactive behaviour generation — especially on the level of behaviour planning — encountered in this thesis will become more widespread, and will likely spawn new discussions on SAIBA. The choices made in this thesis may contribute to these discussions.

Interactional intelligence and dialogue coordination Finally, this thesis make a contribution to the question of interactional intelligence as a basis for language use in dialogue.

In the introduction, we argued for a shift of focus in artificial conversational agent research, from natural language processing to artificial interactional intelligence. Whether such a shift solves the real world problems that artificial conversational agent research faces cannot be answered in this thesis and remains to be shown. Nevertheless, this thesis makes first steps towards computationally modelling interactional intelligence in an actual implemented dialogue system by limiting the model in extent to the speaking role and to evidence of understanding in form of communicative listener feedback. Evaluation results led us to the conclusion that the implemented agent coordinated with its human interlocutors on the levels of belief and attitudes (Kopp 2010).

The thesis shows that the perspective on dialogue as an interactive and iterative process in which speakers and listeners work towards understanding by jointly coordinating their communicative acts through adaptation and feedback is interesting for the fields of artificial conversational agents and formal dialogue modelling. Taking the concept of interactional intelligence (Levinson 1995) as the point of departure, the research described in this thesis raised many interesting research questions and engineering problems in the fields discussed in this sections. Some of these research directions were pursued in this thesis, but many remain unsolved and call for future work.

10.3 LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

The work presented in this thesis is limited in several respects. First of all, some properties that a fully autonomous attentive speaker agent would ideally have could not be researched and/or engineered within the scope of this thesis and are thus absent in the implemented attentive speaker agent.

Notably absent from the system is automatic processing of audio-visual feedback signals of the human interaction partners (as discussed in section 8.4.3). For the purpose of evaluation it was acceptable to rely on a Wizard-of-Oz mediated interaction, where the wizard fed a high-level description of the participants' feedback behaviour into the semi-autonomous system. Future work on fully autonomous attentive speaker agents, however, needs to address the signal processing side of embodied communicative feedback, which is a challenging interdisciplinary research problem — at the intersection of linguistics, phonetics, gesture research, signal processing, pattern recognition, and computer vision — in itself. Besides making an agent autonomous, automatic methods for extraction and classification of relevant features of audio-visual feedback signals have the potential to deliver a more nuanced, accurate, and consistent description of communicative feedback than a human wizard is able to provide in real-time.

Furthermore, the conversational interaction between the implemented attentive speaker agent and the participants of the evaluation study was asymmetric as the agent presented information to its human interlocutors who essentially remained listeners throughout the interactions. An interesting perspective for the models of attentive speaking would be their integration into artificial conversational agents that are able to engage in more symmetrical dialogues with their human interlocutors.

Another limitation is that the probabilistic model of inference-based attribution of listening-related mental states, as presented in this thesis, is not learned from data, but 'expert-modelled' based on theoretical and empirical insights (gained from involvement in empirical research on feedback phenomena, e.g., conducting dialogue studies, annotating corpora). While carefully hand-crafted models have certain advantages — and are a viable approach for Bayesian network construction (Koller and Friedman 2009, box 3.C, pp. 64–67), e.g., when not enough data is available — learning models from data or through interactions with actual human interaction partners can be advantageous for several reasons. Learned models (i) might be more accurate than theoretical models, which are, by definition, idealisations of the world, (ii) might be able to model individual difference in human feedback behaviour, thus allowing the agent to adapt its models to a specific interlocutor, and (iii) could be used as a scientific tool to validate or falsify predictions made by theoretical accounts of feedback in dialogue. Future work utilising the inference-based mental state attribution approach should thus aim to derive the model, at least in parts, from empirical data.

Apart from these practical limitations, there are also some conceptual limitations. First of all, the thesis does not make a statement of how communicative feedback can be brought together with other (multidimensional) communicative acts that provide evidence of understanding (e.g., clarification requests, relevant next utterances). As evidence of understanding may only be a secondary function of such acts, they need to be treated like every other utterance of human interlocutors, e.g., undergo syntactic, semantic, and pragmatic analysis and be integrated into the discourse representation. Future work needs to reconcile these different sources of evidence. Multidimensional communicative acts should inform the model of attributed listener state, and communicative feedback acts should, perhaps, be integrated into the discourse representation.

The repertoire and complexity of the strategies and mechanism for adaptive language production presented in this thesis are rather limited and can only be considered a first step towards the creative adaptation of utterances in speech production that human speakers are capable of. Importantly, the attentive speaker agent is not able to reflect on the likely effects and utilities of the implemented adaptation mechanisms. Hence, the agent cannot evaluate the adapted behaviours against other possibilities, but needs to act according to a fixed configuration of rules. Apart from providing the agent with a larger repertoire of strategies and mechanisms, future work should frame adaptation in language production, on all levels, as planning and decision problems, thus endowing the agent with computationally creative means for producing communicative acts adapted to its interlocutors' needs.

Finally, we need to acknowledge that the overall thesis is rather broad and shallow than narrow and deep. The result of this is that some of the models proposed are not fully spelled out and need to be regarded as mere proofs of concept. We consider, for example, that the aim of this thesis is not to make a claim about an exact configuration of the attributed listener state models. Neither the variables, nor the structure or the parameters of the local probabilistic models should be regarded as definite statements of how feedback needs to be processed in an attributed listener state network. Quite the contrary. We merely argue — and show with our proof of concept — that it is feasible and useful to model feedback interpretation as attribution of listener states using Bayesian networks that model dialogue context and features of the feedback signals.

The ongoing research-project KOMPASS (Yaghoubzadeh, Buschmeier et al. 2015) is beginning to address some of the future research directions described above. It has dedicated personnel working on automatic processing of audio-visual feedback signals, it aims to use computational models based on the ideas developed in this thesis at multiple levels of processing (ensuring contact, understanding, cooperation) of an autonomous socially cooperative artificial conversational agent, and it aims to integrate these models with the language understanding and generation processes of the agent.

APPENDICES

APPENDIX A
MODEL PARAMETRISATION FROM IMPLICIT REPRESENTATION

This appendix contains a description and an example of the approach developed to specify the model parameters for the extended attributed listener state Bayesian network (see chapter 5) via an implicit representation.

A.1 MODEL PARAMETRISATION

An important advantage of Bayesian networks over other probabilistic modelling approaches is that a significant reduction in the number of model parameters that need to be specified (or learned from data) is possible. Instead of specifying the full joint probability distribution over the variables of a model, the structure of a Bayesian network reflects the (conditional and marginal) independencies among the variables and allows for the specification of much more compact 'local probabilistic models' (Koller and Friedman 2009, p. 61).

The joint distribution $\Pr(C, P, U, AC, AG, FB)$ of the five ALS-variables and the observable variable that represents the perceived feedback function, defined in section 5.4, for example, consists of $3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 10 = 2430$ parameters (see fn. 50 on page 90). Given the independence assertions $\mathcal{I}_{\mathcal{ALS}'}$ for this model, eq. (5.11), the local probabilistic models of the factorised distribution, eq. (5.12), require only $10 + 30 + 90 + 90 + 90 + 90 = 400$ parameters.

Specifying these parameters manually may in principle be possible — Koller and Friedman (ibid., pp. 66–67) provide some guidelines and recommend using 'sensitivity analysis' as a supporting tool — but it is tedious, error prone, and makes it difficult to not lose track of the big picture. Given a large number of individual parameters to

[⋆] This appendix contains material previously published in Buschmeier and Kopp (2012b, § 4.2).

specify, it is challenging to choose consistently and to be able to replicate and explain one's choices — this makes them somewhat arbitrary.

The models used in this thesis are therefore manually parametrised with an approach that focusses on fewer parameters which model the essential interactions among variable configurations. It is based on the modelling choice that many variables have a value range that can be considered to lie on an ordinal scale (often *low — medium — high*).

The approach, explained in detail in the next section, generates a conditional probability distribution in tabular format — a conditional probability table (CPT) — via an 'implicit' representation of the local probabilistic model (Koller and Friedman 2009, p. 158). It assumes that the probability density function of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ can be parametrised such that it models the relationship among the values of a variable Y given an assignment of values $(x_1, \ldots, x_n) \in \mathrm{Val}(X_1) \times \ldots \times \mathrm{Val}(X_n)$ of its dependent/parent variables $\mathrm{Pa}_Y = \{X_1, \ldots, X_n\}$. The location of the density function is modelled as a linear combination of the direction and strength of influence that an assignment exerts on Y. The individual conditional probabilities for the values of Y given an assignment of values $(x_1, \ldots x_n)$ can then be derived by mapping the values $y_i \in \mathrm{Val}(Y)$ to the same space and computing their relative likelihood. Repeating this for any assignment of values for $X_1, \ldots, X_n$ and normalising them to probabilities yields the full conditional probability table of the conditional probabilistic distribution $\Pr(Y \mid X_1, \ldots, X_n)$.

With this approach, only $10 + 14 + 18 + 18 + 18 + 18 = 96$ parameters for the implicit representations[138] need to be specified to generate all local probabilistic models of the factorised distribution for $\Pr(C, P, U, AC, AG, FB)$.

A.2 THE CPT GENERATION ALGORITHM

The parameters of a local probabilistic model $\Pr(Y \mid X_1, \ldots, X_n)$ — in form of a conditional probability table — can be generated with the following algorithm:

G1 For each variable $X_i \in \mathrm{Pa}_Y = \{X_1, \ldots, X_n\}$:

G1.1 Specify a value

$$\gamma(X_i) \in [0\,..\,1], \text{ with } \sum_{i=1}^{n} \gamma(X_i) = 1.$$

[138] The implicit representation of each local probabilistic model consists of $|\mathrm{Pa}_Y| + |\mathrm{Val}(Y)| + \sum_{X \in \mathrm{Pa}_Y} |\mathrm{Val}(X)|$ parameters.

This encodes the general strength of influence that X_i has on Y, relative to the other parent variables $\mathrm{Pa}_Y \setminus X_i$.

G1.2 For each value $x_{i,j} \in \mathrm{Val}(X_i) = \{x_{i,1}, \ldots, x_{i,m}\}$ that X_i can take, specify a value

$$\delta(x_{i,j}) \in \mathbb{R},$$

which encodes the direction and strength of influence that X_i has on Y if $X_i = x_{i,j}$. The direction of influence is negative if $\delta(x_{i,j}) < 0$ and positive if $\delta(x_{i,j}) > 0$. Y is not influenced by $x_{i,j}$, iff $\delta(x_{i,j}) = 0$. $|\delta(x_{i,j})|$ quantifies the strength. $\delta(x_{i,j})$ can be thought of as a one-dimensional vector.

G2 For each value $y_i \in \mathrm{Val}(Y) = \{y_1, \ldots, y_l\}$ that Y can take, specify a value

$$\rho(y_i) \in \mathbb{R}.$$

$\rho(y_i)$ relates y_i to the $\delta(x_i)$ — and $\mu(x_1, \ldots, x_n)$, see below — by mapping it onto the same scale. A natural mapping for values *low*, *medium*, and *high* could, e.g., be: $\rho(low) = -1, \quad \rho(medium) = 0, \quad \rho(high) = 1.$

G3 For each assignment of values $(x_1, \ldots, x_n) \in \mathrm{Val}(X_1) \times \ldots \times \mathrm{Val}(X_n)$ of the variables $\mathrm{Pa}_Y = \{X_1, \ldots, X_n\}$[139]:

G3.1 Calculate the value

$$\mu(x_1, \ldots, x_n) = \sum_{i=1}^{n} \gamma(X_i) \cdot \delta(x_i).$$

that models their combined influence on Y as a linear combination of the direction and strength of influence $\delta(x_i)$ of the specific value x_i weighted by each variable X_i's relative strength of influence $\gamma(X_i)$.

G3.2 For each value $y_i \in \mathrm{Val}(Y) = \{y_1, \ldots, y_l\}$:

G3.2.1 Calculate the value

$$\tilde{p}(y_i \mid x_1, \ldots, x_n) = \varphi_{\mu(x_1,\ldots,x_n),\sigma^2}\big[\rho(y_i)\big],$$

where $\varphi_{\mu(x_1,\ldots,x_n),\sigma^2}[\rho(y_i)]$ is the Gaussian probability density function at location $\rho(y_i)$[140].

[139] An assignment of values $(x_1, \ldots, x_n)$ represents a situation in which $X_1 = x_1 \wedge \ldots \wedge X_n = x_n$. Note that the x_i refer to a specific value $x_{i,j} \in \mathrm{Val}(X_i)$ that X_i takes in this situation, that is $X_i = x_{i,j}$.

[140] The probability density function of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$,

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp(-\frac{(x-\mu)^2}{2\sigma^2}),$$

$\tilde{p}(y_i \mid x_1, \ldots, x_n)$ is the relative likelihood of — or the preliminary, not yet normalised degree of belief in — $Y = y_i$ given $X_1 = x_1, \ldots, X_n = x_n$. It brings together a value y_i with an assignment of values $(x_1, \ldots, x_n)$. Figure A.1 illustrates how $\tilde{p}(y_i \mid x_1, \ldots, x_n)$, $\mu(x_1, \ldots, x_n)$, and $\rho(y_i)$ interact.

G3.2.2 Derive a probability

$$\Pr(Y = y_i \mid X_1 = x_1, \ldots, X_n = x_n) \in [0 \,..\, 1]$$

from the preliminary value $\tilde{p}(y_i \mid x_1, \ldots, x_n)$ by normalising it such that

$$\sum_{y_i \in \mathrm{Val}(Y)} \Pr(y_i \mid x_1, \ldots, x_n) = 1.$$

A.3 EXAMPLE — GENERATING A CPT FOR Pr(*U* | *P*, *FB*)

We illustrate the approach and algorithm by generating a conditional probability table for the local probabilistic model $\Pr(U \mid P, FB)$[141], see the network in fig. A.2. For the purpose of this example, we limit *FB* to $\mathrm{Val}(FB) = \{p^-, p^+, u^-, u^+\}$ and choose the parameters for the implicit representation of the conditional probability distribution, see steps G1.1, G1.2, and G2 above, as follows:

$$\gamma_U(P) = 0.3 \qquad \gamma_U(FB) = 0.7 \tag{A.1}$$

$$\boldsymbol{\delta}_U(P) = (-1.0, 0.0, 1.0) \qquad \boldsymbol{\delta}_U(FB) = (-0.9, -0.8, -1.0, 0.7) \tag{A.2}$$

$$\boldsymbol{\rho}(U) = (-1.0, 0.0, 1.0) \tag{A.3}$$

That is, 30 % of the influence on *U* comes from *P* and 70 % from *FB* (eq. [A.1]). When *P* = *low*, *P*'s influence on *U* is negative. *P* does not have an influence when it is *medium*

was chosen because it is unimodal and can easily be positioned (by specifying its mean $\mu \in \mathbb{R}$) and scaled (by specifying its variance $\sigma^2 \in \mathbb{R}_{>0}$). This makes it straightforward to model the relevant relationships between the three-valued random variables commonly used in the ALS-model by choosing an appropriate value for μ, for example — assuming the standard parametrisation of the $\rho(y_i)$ from step G2 — ,

$$\begin{aligned} \mu > 0.5 :& \quad \varphi_{\mu,\sigma^2}[\rho(low)] < \varphi_{\mu,\sigma^2}[\rho(medium)] < \varphi_{\mu,\sigma^2}[\rho(\mathbf{high})], \\ \mu < -0.5 :& \quad \varphi_{\mu,\sigma^2}[\rho(\mathbf{low})] > \varphi_{\mu,\sigma^2}[\rho(medium)] > \varphi_{\mu,\sigma^2}[\rho(high)], \\ -0.5 < \mu < 0.5 :& \quad \varphi_{\mu,\sigma^2}[\rho(low)] < \varphi_{\mu,\sigma^2}[\rho(\mathbf{medium})] > \varphi_{\mu,\sigma^2}[\rho(high)]. \end{aligned}$$

[141] The model (and code to generate it) is archived and available at DOI: 10.6084/m9.figshare.3838047 .

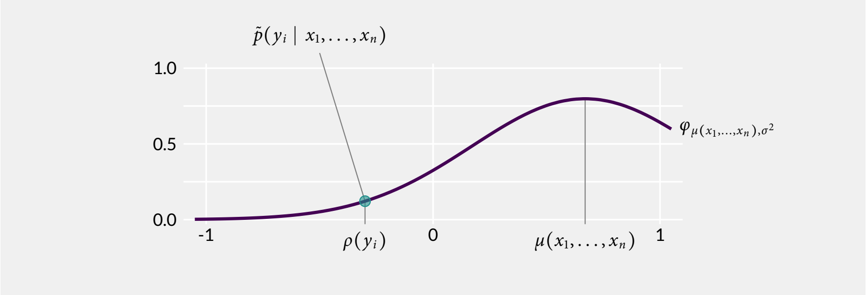


Figure A.1: Illustration of the relationship between $\mu(x_1, \ldots, x_n)$, $\rho(y_i)$, and $\varphi_{\mu(x_1,\ldots,x_n),\sigma^2}$, which determine $\tilde{p}(y_i \mid x_1, \ldots, x_n)$, the relative likelihood of — or not yet normalised degree of belief in — $Y = y_i$ given an assignment of values $X_1 = x_1, \ldots, X_n = x_n$. See step G3.2.1 of the CPT generation algorithm.

and it has a positive influence if its *high*. If feedback of the communicative function perception is received, i.e, if $FB = p^+ \vee p^-$, U is influenced negatively. Likewise if $FB = u^-$. U is only influenced positively if positive feedback of understanding is received, i.e., $FB = u^+$ (eq. [A.2]). Concerning U, as proposed in step G2 of the CPT generation algorithm, a value of *low* is considered to be negative, *medium* is considered to be neutral, and *high* is considered to be positive (eq. A.3).

For this example, twelve parameters $\mu(P = p_i, FB = fb_j)$ for positioning the Gaussian probability density function can be calculated (see table A.1), one for each assignment of values of P and FB (step G3.1). The Gaussian density functions $\mathcal{N}[\mu(P = p_i, FB = fb_j), \sigma^2]$ are plotted in fig. A.2, six of them — those for which $x_P \in \{\wedge, \overset{\wedge}{\wedge}, \overset{\wedge}{\overset{\wedge}{\wedge}}\}$, $x_{FB} \in \{u^-, u^+\}$ — annotated and two highlighted: $\mu(P = \overset{\wedge}{\overset{\wedge}{\wedge}}, FB = u^+)$ and $\mu(P = \wedge, FB = u^-)$.

With each of these twelve Gaussian density functions the three relative likelihood values $\tilde{p}(U = u_k \mid P = p_i, FB = fb_j)$ with $u_k \in \mathrm{Val}(U)$ can be computed at $\rho(u_k)$ (step G3.2.1). Converting these 36 values to probabilities (step G3.2.2) yields the conditional probability table/distribution for $\Pr(U \mid P, FB)$, see table A.2.

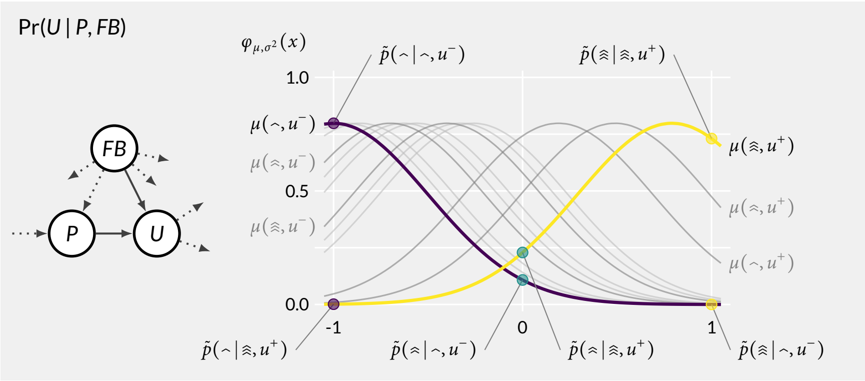


Figure A.2: Example derivation of a local probabilistic model $\Pr(U \mid P, FB)$ from the implicit representation specified in eqs. (A.1) to (A.3). The graph shows plots of the twelve Gaussian density functions $\mathcal{N}(\mu(P = x_P, FB = x_{FB}), \sigma^2)$ with $x_P \in \{\frown, \overset{\frown}{\frown}, \overset{\frown}{\overset{\frown}{\frown}}\}$, $x_{FB} \in \{u^-, u^+\}$ and $\sigma^2 = 0.5$. As an illustration, six (of 36) preliminary — that is, not yet normalised — entries $\tilde{p}$ for the conditional probability table $\Pr(U \mid P, FB)$ are singled out, see table A.2 for their values.

Table A.1: The twelve parameters $\mu(P = p_i, FB = fb_j) = \gamma(P) \cdot \delta(p_i) + \gamma(FB) \cdot \delta(fb_j)$ for positioning the Gaussian probability density functions $\mathcal{N}[\mu(P = p_i, FB = fb_j), \sigma^2]$.

FB =		p^- (-0.9)	p^+ (-0.8)	u^- (-1.0)	u^+ (0.7)
P =	*low*/$\frown$ (-1)	-0.93	-0.86	-1.0	0.19
	medium/$\overset{\frown}{\frown}$ (0)	-0.63	-0.56	-0.7	0.49
	high/$\overset{\frown}{\overset{\frown}{\frown}}$ (1)	-0.33	-0.26	-0.4	0.79

Table A.2: Conditional probability table of the local probabilistic model $\Pr(U \mid P, FB)$ generated via its implicit representation eqs. (A.1) to (A.3). Probabilities annotated with a coloured dot (e.g., ●) are those singled out in fig. A.2.

P =		*low*/$\frown$				*medium*/$\overset{\frown}{\frown}$				*high*/$\overset{\frown}{\overset{\frown}{\frown}}$			
FB =		p^-	p^+	u^-	u^+	p^-	p^+	u^-	u^+	p^-	p^+	u^-	u^+
U =	$\frown$	.848	.808	.880 ●	.047	.625	.556	.688	.010	.328	.268	.395	.000 ●
	$\overset{\frown}{\frown}$	.152	.191	.119 ●	.739	.371	.438	.309	.505	.648	.699	.589	.238 ●
	$\overset{\frown}{\overset{\frown}{\frown}}$	.001	.001	.001 ●	.214	.004	.006	.003	.485	.023	.033	.016	.760 ●

APPENDIX B
STUDY MATERIALS

This appendix contains the instructions given to participants of the attentive speaker agent evaluation study (chapter 9). Participants received the same instructions across experimental conditions.

B.1 SHORT WRITTEN INSTRUCTIONS

Participants read the following short instruction (translated from German) as part of the information sheet containing information on the study (see section 9.3.6).

> In this evaluation study you will interact with the virtual assistant Billie. Billie will present six independent blocks of appointments, changes to appointments, and proposals for appointments for an imaginary week from Monday to Sunday to you. While Billie is presenting these appointments to you, you can provide natural, verbal or non-verbal feedback to him. You can, for example, use feedback signals such as *mhm*, *ja* ('yeah'), *hä?* ('huh?'), *okay*, …; nod or shake with your head; or produce facial expressions. Billie can perceive your feedback and can take it into account in his behaviour. After each block, Billie will prompt you to write down the appointments into an empty calendar. For this reason it is important that you understand Billie as well as possible and that you memorise the presented appointments.

B.2 DETAILED ORAL INSTRUCTIONS

Participants also received more detailed oral instruction from the experimenter once the camera recordings were started. Oral instructions were not fully formulated in advance. The experimenter (HB), however, always presented a fixed set of details and aimed at a consistent verbalisation of key aspects. Participants were allowed to ask questions, possibly causing detours from the intended order of presentation.

The following transcript of the oral instructions, provided to participant X58, is a typical example (translated and lightly edited from German; the participant's utterances are typeset in italics and grey).

> All right. The virtual agent Billie will appear on this screen *mhm* and will talk to you about appointments. He announces appointments or says that appointments need to be moved or cancelled or he makes proposals for announcements. *mhm* He does this in six blocks and in each block he will present between three and five appointments.
>
> After each block he will prompt you to write the appointments into the paper calender. These empty calenders over here. *mhm* There are six of it. I would ask you to have them face down while Billie talks and when he prompts you to write down the appointments you turn them around write them down and turn them around again. After that you can look at him and tell him 'okay let's proceed' or 'you can proceed now' or something like that *all right* and then he will continue *mhm* with the next block.
>
> It is not a real conversation between the two of you, because there is no speech recogniser or something like that besides for the few phrases, for example 'you can continue' or 'repeat it please', when he specifically asks if he should repeat something for example. He presents the appointments and proposals and so on to you and you can signal him understanding — whether you understood mhm what he said or not *yeah* — or you can signal your attitude — do you like what he suggests or not, do you accept an appointment proposal or the changes to an appointment or not.
>
> And you can do that by providing feedback — verbally or non-verbally. Here is the camera through which he can see you and see whether you nod with your head, for example, or whether you look '*huh?*', and with this microphone here he can also hear when you say something like '*mhm*', '*uh-huh*' or '*okay*'. These are the small signals you can use to communicate with him *okay* such small feedback signals. *Can I say something like 'I did not understand this'?* Well, you should not do that. *I should not do that, okay.* It is really about the small signals. *okay*
>
> Billie is able to perceive these small signals and he can, well, I don't want to say he reacts to them, perhaps he can do that as well, but he can incorporate them into his own behaviour. *mhm okay* Okay.

After these six blocks he will say goodbye to you and he will prompt you to fill in a questionnaire which simply appears on the screen. You can move this keyboard here to the front and with it you can also control the mouse cursor.

After you filled in the questionnaire, just go to the door, open it and I will most probably already be there. *okay mhm* And then we will go back into the room *mhm* — you can leave your things here — and then I will pay you and ask you one or two further questions. *okay* Okay? *We will do that.*

All right. *right* As soon as I leave the room, you can start the experiment by looking at the screen and saying 'Hallo Billie'. *okay*. Okay? *mhm all right* Okay, well, then have fun. *Thanks.*

BIBLIOGRAPHY

Allen, Donald E. and Rebecca F. Guy (1974). *Conversation Analysis. The Sociology of Talk*. Den Haag, The Netherlands: Mouton ⊸ p. 45.

Allwood, Jens (1988). 'Om det svenska systemet för språklig återkoppling [On the Swedish system of linguistic feedback]'. In: *Svenskans beskrivning, 16*. Ed. by Per Linell, Viveka Adelswärd, Torbjörn Nilsson and Per A. Pettersson. Linköping, Sweden: Linköping University, Tema Kommunikation, pp. 89–106 ⊸ pp. 44–45, 49–51, 200.

Allwood, Jens (2000). 'An activity-based approach to pragmatics'. In: *Abduction, Belief and Context in Dialogue: Studies in computational pragmatics*. Ed. by Harry Bunt and William Black. Amsterdam, The Netherlands: John Benjamins, pp. 47–80. DOI: 10.1075/nlp.1.02all ⊸ pp. 58–59, 90.

Allwood, Jens and Loredana Cerrato (2003). 'A study of gestural feedback expressions'. In: *Proceedings of the 1st Nordic Symposium on Multimodal Communication*. København, Denmark, pp. 7–22 ⊸ pp. 48–49, 54–55.

Allwood, Jens, Loredana Cerrato, Kristiina Jokinen, Constanza Navarretta and Patrizia Paggio (2007). 'The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena'. In: *Language Resources and Evaluation* 41, pp. 273–287. DOI: 10.1007/s10579-007-9061-5 ⊸ pp. 57–58, 72.

Allwood, Jens, Stefan Kopp, Karl Grammer, Elisabeh Ahlsén, Elisabeth Oberzaucher and Markus Koppensteiner (2007). 'The analysis of embodied communicative feedback in multimodal corpora: A prerequisite for behaviour simulation'. In: *Language Resources and Evaluation* 41, pp. 255–272. DOI: 10.1007/s10579-007-9056-2 ⊸ pp. 48, 54–55.

Allwood, Jens, Joakim Nivre and Elisabeth Ahlsén (1992). 'On the semantics and pragmatics of linguistic feedback'. In: *Journal of Semantics* 9, pp. 1–26. DOI: 10.1093/jos/9.1.1 ⊸ pp. 4, 6, 18, 57–58, 61–63, 69, 72, 85, 191, 245.

Ameka, Felix (1992). 'Interjections: The universal yet neglected part of speech'. In: *Journal of Pragmatics* 18, pp. 101–119. DOI: 10.1016/0378-2166(92)90048-G ⊸ p. 49.

Amelsvoort, Marije van, Bart Joosten, Emiel Krahmer and Eric Postma (2013). 'Using non-verbal cues to (automatically) assess children's performance difficulties with arithmetic problems'. In: *Computers in Human Behavior* 29, pp. 654–664. DOI: 10.1016/j.chb.2012.10.016 ⊸ p. 55.

Andrist, Sean, Bilge Mutlu and Michael Gleicher (2013). 'Conversational gaze aversion for virtual agents'. In: *Proceedings of the 13th International Conference on Intelligent Virtual Agents (IVA)*. Edinburgh, UK, pp. 249–262. DOI: 10.1007/978-3-642-40415-3_22 ⊸ p. 169.

Ashby, W. Ross (1956). *An Introduction to Cybernetics*. London, UK: Chapman & Hall ⊸ pp. 42, 45, 84.

Asher, Nicolas and Alex Lascarides (2003). *Logics of Conversation*. Cambridge, UK: Cambridge University Press ⊸ pp. 109, 117.

Aylett, Matthew P. and Christopher J. Pidcock (2007). 'The CereVoice characterful speech synthesiser SDK'. In: *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA)*. Paris, France, pp. 413–414. DOI: 10.1007/978-3-540-74997-4_65 ⊸ p. 155.

Bach, Kent (2006). 'The top 10 misconceptions about implicature'. In: *Drawing the Boundaries of Meaning. Neo-Gricean Studies in Pragmatics and Semantics in Honor of Laurence R.*

⋆ Pages listed following the symbol '⊸' are back references to where a work is cited in this thesis. ¶ The digital object identifiers (DOIs) provided can be resolved through resolution services that map from a DOI's name (e.g., '10.1007/s13218-013-0241-8') to the associated resources. The resolver of the International DOI Foundation, for example, can be used by appending a DOI's name to the resolver's base URL https://doi.org/, e.g., https://doi.org/10.1007/s13218-013-0241-8. ¶ In electronic copies of this thesis, all DOIs are hyperlinks that resolve in this way.

Horn. Ed. by Betty J. Birner and Gregory Ward. Amsterdam, The Netherlands: John Benjamins, pp. 21–30. DOI: 10.1075/slcs.80.03bac ⊸ pp. 61, 90.

Baker, Rachel, Alastair J. Gill and Justine Cassell (2008). 'Reactive redundancy and listener comprehension in direction-giving'. In: *Proceedings of the 9th Workshop on Discourse and Dialogue (SIGdial)*. Columbus, OH, USA, pp. 37–45 ⊸ p. 123.

Barsalou, Lawrence W. (1999). 'Perceptual symbol systems'. In: *Behavioral and Brain Sciences* 22, pp. 577–660. DOI: 10.1017/S0140525X99002149 ⊸ p. 16.

Baumann, Timo and David Schlangen (2012a). 'INPROISS: A Component for just-in-time incremental speech synthesis'. In: *Proceedings of the ACL 2012 System Demonstrations*. Jeju Island, South Korea, pp. 103–108 ⊸ p. 139.

Baumann, Timo and David Schlangen (2012b). 'Evaluating prosodic processing for incremental speech synthesis'. In: *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, OR, USA, pp. 438–441 ⊸ pp. 129, 140.

Bavelas, Janet B., Linda Coates and Trudy Johnson (2000). 'Listeners as co-narrators'. In: *Journal of Personality and Social Psychology* 79, pp. 941–952. DOI: 10.1037/0022-3514.79.6.941 ⊸ pp. 6, 30, 56, 68.

Bavelas, Janet B., Linda Coates and Trudy Johnson (2002). 'Listener responses as a collaborative process: the role of gaze'. In: *Journal of Communication* 52, pp. 566–580. DOI: 10.1111/j.1460-2466.2002.tb02562.x ⊸ p. 145.

Bazzanella, Carla and Rossana Damiano (1999). 'The interactional handling of misunderstanding in everyday conversations'. In: *Journal of Pragmatics* 31, pp. 817–836. DOI: 10.1016/s0378-2166(98)00058-7 ⊸ p. 19.

Beňuš, Štefan (2012). 'Prosodic forms and pragmatic meanings: The case of the discourse marker *'no'* in Slovak'. In: *Proceedings of the 3rd IEEE International Conference on Cognitive Infocommunications*. Košice, Slovakia, pp. 77–81. DOI: 10.1109/CogInfoCom.2012.6421961 ⊸ p. 53.

Bergmann, Kirsten, Sebastian Kahl and Stefan Kopp (2013). 'Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints'. In: *Proceedings of the 13th International Conference on Intelligent Virtual Agents (IVA)*. Edinburgh, UK, pp. 203–216. DOI: 10.1007/978-3-642-40415-3_18 ⊸ p. 161.

Bergmann, Kirsten and Stefan Kopp (2009). 'Increasing the expressiveness of virtual agents — Autonomous generation of speech and gesture for spatial description tasks'. In: *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Budapest, Hungary, pp. 361–368 ⊸ p. 161.

Bergmann, Kirsten and Stefan Kopp (2012). 'Gestural alignment in natural dialogue'. In: *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci)*. Sapporo, Japan, pp. 1326–1331 ⊸ p. 29.

Bevacqua, Elisabetta (2009). 'Computational Model of Listener Behavior for Embodied Conversational Agents'. PhD thesis. Paris, France: Université Paris VIII ⊸ pp. 6, 69.

Blackburn, Perry L. (2007). *The Code Model of Communication. A Powerful Metaphor in Linguistic Metatheory*. Dallas, TX, USA: SIL International ⊸ pp. 14, 46.

BML Committee (2011). *The BML 1.0 Standard*. URL: http://www.mindmakers.org/projects/bml-1-0/ (visited on 15/06/2016) ⊸ pp. 159–160, 162, 168, 191.

Boersma, Paul and David Weenink (2016). *Praat, a system for doing phonetics by computer.* Version 6.0.14. URL: http://www.praat.org/ ⊸ p. 199.

Boghossian, Paul A. (2009). 'Content'. In: *A Companion to Metaphysics*. Ed. by Jaegwon Kim, Ernest Sosa and Gary S. Rosenkrantz. 2nd. Chichester, UK: Wiley-Blackwell, pp. 188–191 ⊸ p. 82.

Branigan, Holly P., Martin J. Pickering, Andrew J. Steward and Janet F. McLean (2000). 'Syntactic priming in spoken production: linguistic and temporal inference'. In: *Memory & Cognition* 28, pp. 1279–1302. DOI: 10.3758/BF03211830 ⊸ p. 29.

Bray, Tim (2014). *The JavaScript Object Notation (JSON) data interchange format*. RFC 7159. URL: https://tools.ietf.org/html/rfc7159 ⊸ p. 159.

Brennan, Susan E. (1990). 'Seeking and Providing Evidence for Mutual Understanding'. PhD thesis. Stanford, CA, USA: Stanford University ⊸ pp. 3, 32, 143.

Brennan, Susan E. and Herbert H. Clark (1996). 'Conceptual pacts and lexical choice in conversation'. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, pp. 1482–1493. DOI: 10.1037/0278-7393.22.6.1482 ⊸ p. 29.

Brennan, Susan E., Alexia Galati and Anna K. Kuhlen (2010). 'Two minds, one dialog: Coordinating speaking and understanding'. In: *The Psychology of Learning and Motivation: Advances in Research and Theory*. Ed. by Brian H. Ross. Vol. 53. Psychology of Learning and Motivation. Burlington, MA, USA: Academic Press, pp. 301–344. DOI: 10.1016/S0079-7421(10)53008-1 ⊸ pp. 34–35, 80, 137, 246.

Brennan, Susan E. and Joy E. Hanna (2009). 'Partner-specific adaptation in dialogue'. In: *Topics in Cognitive Science* 1, pp. 274–291. DOI: 10.1111/j.1756-8765.2009.01019.x ⊸ pp. 34, 118.

Brown-Schmidt, Sarah (2012). 'Beyond common and privileged: Gradient representations of common ground in real-time language use'. In: *Language and Cognitive Processes* 27, pp. 62–89. DOI: 10.1080/01690965.2010.543363 ⊸ pp. 25, 35, 119, 246.

Brown-Schmidt, Sarah and Agnieszka E. Konopka (2014). 'Processes of incremental message planning during conversation'. In: *Psychonomic Bulletin & Review* 22, pp. 833–843. DOI: 10.3758/s13423-014-0714-2 ⊸ pp. 128, 135.

Bunt, Harry (2011). 'Multifunctionality in dialogue'. In: *Computer Speech and Language* 25, pp. 222–245. DOI: 10.1016/j.csl.2010.04.006 ⊸ pp. 58–60.

Bunt, Harry (2012). 'The semantics of feedback'. In: *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*. Paris, France, pp. 118–127 ⊸ pp. 6, 58, 245.

Burgoon, Judee K., Lesa A. Stern and Leesa Dillman (1995). *Interpersonal Adaptation. Dyadic Interaction Patterns*. New York, NY, USA: Cambridge University Press. DOI: 10.1017/CBO9780511720314 ⊸ p. 28.

Buschmeier, Hendrik, Timo Baumann, Benjamin Dosch, Stefan Kopp and David Schlangen (2012). 'Combining incremental language generation and incremental speech synthesis for adaptive information presentation'. In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*. Seoul, South Korea, pp. 295–303 ⊸ pp. 121, 127, 139–140.

Buschmeier, Hendrik, Kirsten Bergmann and Stefan Kopp (2010). 'Modelling and evaluation of lexical and syntactic alignment with a priming-based microplanner'. In: *Empirical Methods in Natural Language Generation*. Ed. by Emiel Krahmer and Mariët Theune. Vol. 5980. Berlin, Germany: Springer, pp. 85–104. DOI: 10.1007/978-3-642-15573-4_5 ⊸ pp. 36, 126–127, 134.

Buschmeier, Hendrik and Stefan Kopp (2011). 'Towards conversational agents that attend to and adapt to communicative user feedback'. In: *Proceedings of the 11th International Conference on Intelligent Virtual Agents (IVA)*. Reykjavík, Iceland, pp. 169–182. DOI: 10.1007/978-3-642-23974-8_19 ⊸ pp. 77, 81, 143, 153.

Buschmeier, Hendrik and Stefan Kopp (2012a). 'Adapting language production to listener feedback behaviour'. In: *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. Stevenson, WA, USA, pp. 7–10 ⊸ p. 121.

Buschmeier, Hendrik and Stefan Kopp (2012b). 'Using a Bayesian model of the listener to unveil the dialogue information state'. In: *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*. Paris, France, pp. 12–20 ⊸ pp. 57, 81, 89, 191, 253.

Buschmeier, Hendrik and Stefan Kopp (2013). 'Co-constructing grounded symbols—Feedback and incremental adaptation in human–agent dialogue'. In: *Künstliche Intelligenz* 27, pp. 137–143. DOI: 10.1007/s13218-013-0241-8 ⊸ pp. 81, 121.

Buschmeier, Hendrik and Stefan Kopp (2014a). 'A dynamic minimal model of the listener for feedback-based dialogue coordination'. In: *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*. Edinburgh, UK, pp. 17–25 ⊸ pp. 81, 191.

Buschmeier, Hendrik and Stefan Kopp (2014b). 'When to elicit feedback in dialogue: towards a model based on the information needs of speakers'. In: *Proceedings of the 14th International Conference on Intelligent Virtual Agents (IVA)*. Boston, MA, USA, pp. 71–80. DOI: 10.1007/978-3-319-09767-1_10 ⊸ p. 143.

Buschmeier, Hendrik, Zofia Malisz, Joanna Skubisz, Marcin Włodarczak, Ipke Wachsmuth, Stefan Kopp and Petra Wagner (2014). 'ALICO: A multimodal corpus for the study of active listening'. In: *Proceedings of the 9th Conference on Language Resources and Evaluation (LREC)*. Reykjavík, Iceland, pp. 3638–3643 ⊸ p. 41.

Buschmeier, Hendrik, Zofia Malisz, Marcin Włodarczak, Stefan Kopp and Petra Wagner (2011). ''Are you sure you're paying attention?' — 'Uh-huh'. Communicating understanding as a marker of attentiveness'. In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Firenze, Italy, pp. 2057–2060 ⊸ p. 41.

Buschmeier, Hendrik and Marcin Włodarczak (2013). 'TextGridTools: A TextGrid processing and analysis toolkit for Python'. In: *Proceedings der 24. Konferenz zur elektronischen Sprachsignalverarbeitung (ESSV)*. Bielefeld, Germany, pp. 152–157. URL: https://purl.org/net/tgt ⊸ p. 200.

Buschmeier, Hendrik and Ramin Yaghoubzadeh (2011). *KDS-1-Corpus*. Social Cognitive Systems Group, Bielefeld University. Bielefeld, Germany. URL: https://purl.org/scs/KDS-1 ⊸ pp. 110, 191.

Buß, Okko and David Schlangen (2010). 'Modelling sub-utterance phenomena in spoken dialogue systems'. In: *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*. Poznań, Poland, pp. 33–42 ⊸ p. 156.

Butterfill, Stephen A. and Ian A. Apperly (2013). 'How to construct a minimal Theory of Mind'. In: *Mind & Language* 28, pp. 606–637. DOI: 10.1111/mila.12036 ⊸ p. 118.

Cafaro, Angelo, Hannes Högni Vilhjálmsson, Timothy Bickmore, Dirk Heylen and Catherine Pelachaud (2014). 'Representing communicative functions in SAIBA with a unified Function Markup Language'. In: *Proceedings of the 14th International Conference on Intelligent Virtual Agents (IVA)*. Boston, MA, USA, pp. 81–94. DOI: 10.1007/978-3-319-09767-1_11 ⊸ pp. 159–160, 162.

Cassell, Justine, Alastair J. Gill and Paul A. Tepper (2007). 'Coordination in conversation and rapport'. In: *Proceedings of the ACL Workshop on Embodied Language Processing*. Praha, Czech Republic, pp. 41–50 ⊸ p. 175.

Cassell, Justine, Joseph Sullivan, Scott Prevost and Elizabeth Churchill, eds. (2000). *Embodied Conversational Agents*. Cambridge, MA, USA: MIT Press ⊸ pp. 1, 37.

Cassell, Justine and Kristinn Rúnar Thórisson (1999). 'The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents'. In: *Applied Artificial Intelligence* 13, pp. 519–538. DOI: 10.1080/088395199117360 ⊸ p. 6.

Cassell, Justine, Obed E. Torres and Scott Prevost (1999). 'Turn taking versus discourse structure'. In: *Machine Conversations*. Ed. by Yorick Wilks. Boston, MA, USA: Springer, pp. 143–153. DOI: 10.1007/978-1-4757-5687-6_12 ⊸ p. 169.

Cathcart, Nicola, Jean Carletta and Ewan Klein (2003). 'A shallow model of backchannel continuers in spoken dialogue'. In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Budapest, Hungary, pp. 51–58 ⊸ pp. 6, 65–66.

Chartrand, Tanya L. and John A. Bargh (1999). 'The chameleon effect: the perception-behavior link and social interaction'. In: *Journal of Personality and Social Psychology* 76, pp. 893–910. DOI: 10.1037/0022-3514.76.6.893 ⊸ p. 29.

Chen, Yiqiang, Yu Yu and Jean-Marc Odobez (2015). 'Head nod detection from a full 3D model'. In: *Proceedings of the IEEE International Conference on Computer Vision Workshop*. Santiago, Chile, pp. 528–536. DOI: 10.1109/ICCVW.2015.75 ⊸ p. 173.

Clark, Herbert H. (1992). *Arenas of Language Use*. Chicago, IL, USA: University of Chicago Press ⊸ p. 25.

Clark, Herbert H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511620539 ⊸ pp. 3–4, 6, 20–22, 25, 32, 35, 38, 48, 58, 71, 90, 110, 124, 150, 245–246.

Clark, Herbert H. (2002). 'Speaking in time'. In: *Speech Communication* 36, pp. 5–13. DOI: 10.1016/S0167-6393(01)00022-X ⊸ p. 33.

Clark, Herbert H. and Susan E. Brennan (1991). 'Grounding in communication'. In: *Perspectives on Socially Shared Cognition*. Ed. by Lauren B. Resnick, John M. Levine and Stephanie D. Teasley. Washington, DC, USA: American Psychological Association, pp. 222–233 ⊸ pp. 4, 21–22, 32, 78.

Clark, Herbert H. and Thomas B. Carlson (1982). 'Hearers and speech acts'. In: *Language* 58, pp. 332–373 ⊸ pp. 29, 32.

Clark, Herbert H. and Jean E. Fox Tree (2002). 'Using *uh* and *um* in spontaneous speaking'. In: *Cognition* 84, pp. 73–111. DOI: 10.1016/S0010-0277(02)00017-3 ⊸ p. 33.

Clark, Herbert H. and Meredyth A. Krych (2004). 'Speaking while monitoring addressees for understanding'. In: *Journal of Memory and Language* 50, pp. 62–81. DOI: 10.1016/j.jml.2003.08.004 ⊸ pp. 30, 55, 144.

Clark, Herbert H. and Catherine R. Marshall (1981). 'Definite reference and mutual knowledge'. In: *Elements of Discourse Understanding*. Ed. by Aravind K. Joshi, Bonnie L. Webber and Ivan A. Sag. Cambridge, UK: Cambridge University Press, pp. 10–63 ⊸ pp. 4, 25, 32.

Clark, Herbert H. and Edward F. Schaefer (1989). 'Contributing to discourse'. In: *Cognitive Science* 13, pp. 259–294. DOI: 10.1207/s15516709cog1302_7 ⊸ pp. 4, 21–23, 25–26, 28, 67, 107, 109, 175, 246.

Clark, Herbert H. and Deanna Wilkes-Gibbs (1986). 'Referring as a collaborative process'. In: *Cognition* 22, pp. 1–39. DOI: 10.1016/0010-0277(86)90010-7 ⊸ pp. 22, 38.

Cohen, Jacob (1992). 'A power primer'. In: *Psychological Bulletin* 112, pp. 155–159. DOI: 10.1037/0033-2909.112.1.155 ⊸ pp. 202, 210, 212, 215, 219.

Cohen, Philip R., Jerry Morgan and Martha E. Pollack, eds. (1990). *Intentions in Communication*. Cambridge, MA, USA: The MIT Press ⊸ p. 5.

Cooke, Martin, Simon King, Maeva Garnier and Vincent Aubanel (2014). 'The listening talker: A review of human and algorithmic context-induced modifications of speech'. In: *Computer Speech & Language* 28, pp. 543–571. DOI: 10.1016/j.csl.2013.08.003 ⊸ pp. 29–30, 123.

Cooper, Robin (2005). 'Records and record types in semantic theory'. In: *Journal of Logic and Computation* 15, pp. 99–112. DOI: 10.1093/logcom/exi004 ⊸ p. 67.

Cozman, Fabio Gagliardi, Marek J. Druzdzel, Daniel Garcia-Sanchez et al. (1998). *The interchange format for Bayesian networks*. URL: http://www.cs.cmu.edu/~fgcozman/Research/InterchangeFormat/ (visited on 22/09/2016) ⊸ pp. 94, 281.

Darwiche, Adnan (2009). *Modeling and Reasoning with Bayesian Networks*. New York, NY, USA: Cambridge University Press. DOI: 10.1017/CBO9780511811357 ⊸ p. 95.

Dascal, Marcelo and Isidoro Berenstein (1987). 'Two modes of understanding: Comprehending and grasping'. In: *Language & Communication* 7, pp. 139–151. DOI: 10.1016/0271-5309(87)90004-8 ⊸ p. 13.

Deemter, Kees van, Emiel Krahmer and Mariët Theune (2005). 'Real versus template-based natural language generation: A false opposition?' In: *Computational Linguistics* 31, pp. 15–23. DOI: 10.1162/0891201053630291 ⊸ p. 132.

Demberg, Vera, Frank Keller and Alexander Koller (2013). 'Incremental, predictive parsing with psycholinguistically motivated Tree-Adjoining Grammar'. In: *Computational Linguistics* 39, pp. 1025–1066. DOI: 10.1162/COLI_a_00160 ⊸ p. 128.

Den, Yasuharu, Hanae Koiso, Katsuya Takanashi and Nao Yoshida (2012). 'Annotation of response tokens and their triggering expressions in Japanese multi-party conversations'. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey, pp. 1332–1337 ⊸ p. 50.

DeVault, David (2008). 'Contribution Tracking: Participating in Task-oriented Dialogue Under Uncertainty'. PhD thesis. New Brunswick, NJ, USA: Rutgers, The State University of New Jersey ⊸ pp. 18, 27, 36, 110, 121, 126–127.

DeVault, David and Matthew Stone (2006). 'Scorekeeping in an uncertain language game'. In: *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*. Potsdam, Germany, pp. 139–146 ⊸ pp. 20, 27.

Dohsaka, Kohji and Akira Shimazu (1997). 'A system architecture for spoken utterance production in collaborative dialogue'. In: *Working Notes of the IJCAI-97 Workshop on Collaboration, Cooperation and Conflict in Dialogue Systems*. Nagoya, Japan ⊸ p. 70.

Donne, John (1624/1923). *Devotions upon Emergent Occasions*. Ed. by John Sparrow. Cambridge, UK: Cambridge University Press ⊸ p. xiii.

Dosch, Benjamin (2011). 'Adaptive Sprachgenerierung für interaktive virtuelle Agenten [Adaptive language generation for interactive virtual agents]'. MA thesis. Bielefeld, Germany: Faculty of Technology, Bielefeld University ⊸ p. 121.

Dudenredaktion (2013). *Duden – Wissensnetz deutsche Sprache*. Berlin, Germany: Bibliographisches Institut ⊸ p. 49.

Duncan, Starkey, Jr. (1974). 'On the structure of speaker-auditor interaction during speaking turns'. In: *Language in Society* 3, pp. 161–180. DOI: 10.1017/S0047404500004322 ⊸ pp. 64–66.

Duncan, Starkey, Jr. and Donald W. Fiske (1977). *Face-to-Face Interaction: Research, Methods, and Theory*. Hillsdale, NJ, USA: Lawrence Erlbaum ⊸ p. 64.

Edlund, Jens, Joakim Gustafson, Mattias Heldner and Anna Hjalmarsson (2008). 'Towards human-like spoken dialogue systems'. In: *Speech Communication* 50, pp. 630–645. DOI: 10.1016/j.specom.2008.04.002 ⊸ pp. 6, 134, 238.

Ehlich, Konrad (1979). 'Formen und Funktionen von „HM". Eine phonologisch-pragmatische Analyse [Forms and functions of "hm". A phonological-pragmatic analysis]'. In: *Die Partikeln der deutschen Sprache*. Ed. by Harald Weydt. Berlin, Germany: Walter de Gruyter, pp. 503–517. DOI: 10.1515/9783110863574.503 ⊸ p. 50.

Ehlich, Konrad (1986). *Interjektionen [Interjections]*. Tübingen, Germany: Max Niemeyer. DOI: 10.1515/9783111357133 ⊸ pp. 6, 46, 49–50, 52.

Eshghi, Arash, Christine Howes, Eleni Gregoromichelaki, Julian Hough and Matthew Purver (2015). 'Feedback in conversation as incremental semantic update'. In: *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*. London, UK, pp. 261–271 ⊸ p. 67.

Field, Andy (2009). *Discovering Statistics Using SPSS*. 3rd. Los Angeles, CA, USA: Sage ⊸ p. 199.

Field, Andy, Jeremy Miles and Zoë Field (2012). *Discovering Statistics Using R*. Los Angeles, CA, USA: Sage ⊸ p. 199.

Fong, Terrence, Illah Nourbakhsh and Kerstin Dautenhahn (2003). 'A survey of socially interactive robots'. In: *Robotics and Autonomous Systems* 42, pp. 143–166. DOI: 10.1016/s0921-8890(02)00372-x ⊸ pp. 1, 37.

Franke, Michael and Gerhard Jäger (2016). 'Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics'. In: *Zeitschrift für Sprachwissenschaft* 35. DOI: 10.1515/zfs-2016-0002 ⊸ pp. 7, 80, 119.

Frith, Chris D. and Uta Frith (2006). 'The neural basis of mentalizing'. In: *Neuron* 50, pp. 531–534. DOI: 10.1016/j.neuron.2006.05.001 ⊸ pp. 4, 81.

Fujimoto, Donna T. (2007). 'Listener responses in interaction: a case for abandoning the term backchannel'. In: 大阪女学院短期大学紀要 *[Bulletin of the Osaka Jogakuin College]* 37, pp. 35–54. DOI: 10775/48 ⊸ pp. 46, 48.

Galati, Alexia and Susan E. Brennan (2010). 'Attenuating information in spoken communication: For the speaker, or for the addressee?' In: *Journal of Memory and Language* 62, pp. 35–51. DOI: 10.1016/j.jml.2009.09.002 ⊸ pp. 34–35, 80, 118, 137, 246.

Garoufi, Konstantina, Maria Staudte, Alexander Koller and Matthew W. Crocker (2016). 'Exploiting listener gaze to improve situated communication in dynamic virtual environments'. In: *Cognitive Science* 40, pp. 1671–1703. DOI: 10.1111/cogs.12298 ⊸ pp. 55, 70–71, 126.

Garrod, Simon and Anthony Anderson (1987). 'Saying what you mean in dialogue: A study in conceptual and semantic co-ordination'. In: *Cognition* 27, pp. 181–218. DOI: 10.1016/0010-0277(87)90018-7 ⊸ p. 29.

Geertzen, Jeoren, Volha Petukhova and Harry Bunt (2008). 'Evaluating dialogue act tagging with naive and expert annotators'. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakesh, Morocco, pp. 1076–1082 ⊸ p. 97.

Ginzburg, Jonathan (2012). *The Interactive Stance*. Oxford, UK: Oxford University Press. DOI: 10.1093/acprof:oso/9780199697922.001.0001 ⊸ p. 110.

Golato, Andrea (2012). 'German *oh*: Marking an emotional change of state'. In: *Research on Language and Social Interaction* 45, pp. 245–258. DOI: 10.1080/08351813.2012.699253 ⊸ p. 110.

Golato, Andrea and Emma Betz (2008). 'German *ach* and *achso* in repair uptake: Resources to sustain or remove epistemic asymmetry'. In: *Zeitschrift für Sprachwissenschaft* 27, pp. 7–37. DOI: 10.1515/ZFSW.2008.002 ⊸ p. 110.

Golato, Andrea and Zsuzsanna Fagyal (2008). 'Comparing single and double sayings of the german response token *ja* and the role of prosody: A conversation analytic perspective'. In: *Research on Language & Social Interaction* 41, pp. 241–270. DOI: 10.1080/08351810802237834 ⊸ p. 50.

Goodman, Noah D. and Michael C. Frank (2016). 'Pragmatic language interpretation as probabilistic inference'. In: *Trends in Cognitive Sciences* 20, pp. 818–829. DOI: 10.1016/j.tics.2016.08.005 ⊸ pp. 7, 80, 119.

Goodwin, Charles (1986). 'Between and within: Alternative sequential treatments of continuers and assessments'. In: *Human Studies* 9, pp. 205–217. DOI: 10.1007/BF00148127 ⊸ pp. 6, 56.

Gratch, Jonathan, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, R. J. van der Werf and Louis-Philippe Morency (2006). 'Virtual rapport'. In: *Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA)*. Marina del Rey, CA, USA, pp. 14–27. DOI: 10.1007/11821830_2 ⊸ p. 6.

Gravano, Agustín and Julia Hirschberg (2011). 'Turn-taking cues in task-oriented dialogue'. In: *Computer Speech and Language* 25, pp. 601–634. DOI: 10.1016/j.csl.2010.10.003 ⊸ pp. 6, 66, 144–145, 206.

Gravano, Agustín, Julia Hirschberg and Štefan Beňuš (2012). 'Affirmative cue words in task-oriented dialogue'. In: *Computational Linguistics* 38, pp. 1–39. DOI: 10.1162/COLI_a_00083 ⊸ pp. 46, 53, 118.

Grice, Herbert Paul (1957). 'Meaning'. In: *The Philosophical Review* 66, pp. 377–388 ⊸ pp. 2, 13, 90.

Grice, Herbert Paul (1975). 'Logic and conversation'. In: *Syntax and Semantics 3: Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. New York, NY, USA: Academic Press, pp. 41–58 ⊸ pp. 2, 13, 59, 100, 127.

Grosz, Barbara J. and Candace L. Sidner (1986). 'Attention, intentions and the structure of discourse'. In: *Computational Linguistics* 12, pp. 175–204 ⊸ p. 123.

Guhe, Markus (2007). *Incremental Conceptualization for Language Production*. Mahwah, NJ, USA: Lawrence Erlbaum Associates ⊸ p. 125.

Gûnel, Erdogan and James Dickey (1974). 'Bayes factors for independence in contingency tables'. In: *Biometrika* 61, pp. 545–557. DOI: 10.1093/biomet/61.3.545 ⊸ p. 196.

Gureckis, Todd M. and Robert L. Goldstone (2006). 'Thinking in groups'. In: *Pragmatics & Cognition* 14, pp. 293–311. DOI: 10.1075/pc.14.2.10gur ⊸ p. xvi.

Gyftodimos, Elias and Peter Flach (2002). 'Hierarchical Bayesian networks: a probabilistic reasoning model for structured domains'. In: *Proceedings of the ICML-2002 Workshop on Development of Representations*. Sydney, Australia, pp. 23–30 ⊸ p. 106.

Hájek, Alan (2012). 'Interpretations of probability'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2012. URL: http://plato.stanford.edu/archives/win2012/entries/probability-interpret/ ⊸ p. 87.

Halpern, Joseph Y. (2003). *Reasoning about uncertainty*. Cambridge, MA, USA: MIT Press ⊸ p. 85.

Hampton, James A. (1999). 'Concepts'. In: *The MIT Encyclopedia of the Cognitive Sciences*. Ed. by Robert A. Wilson and Frank C. Keil. Cambridge, MA, USA: MIT Press, pp. 176–179 ⊸ p. 16.

Harel, David (1987). 'Statecharts: A visual formalism for complex systems'. In: *Science of Computer Programming* 8, pp. 231–274. DOI: 10.1016/0167-6423(87)90035-9 ⊸ p. 38.

Healey, Patrick G. T., Matthew Purver and Christine Howes (2014). 'Divergence in dialogue'. In: *PLoS ONE* 9, e98598. DOI: 10.1371/journal.pone.0098598 ⊸ p. 29.

Heeman, Peter A. and Graeme Hirst (1995). 'Collaborating on referring expressions'. In: *Computational Linguistics* 21, pp. 351–382 ⊸ p. 38.

Heldner, Mattias, Jens Edlund and Julia Hirschberg (2010). 'Pitch similarity in the vicinity of backchannels'. In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan, pp. 3054–3057 ⊸ p. 48.

Heldner, Mattias, Anna Hjalmarsson and Jens Edlund (2013). 'Backchannel relevance spaces'. In: *Proceedings of Nordic Prosody XI*. Tartu, Estonia, pp. 137–146 ⊸ pp. 64, 66.

Heritage, John (1984). 'A change-of-state token and aspects of its sequential placement'. In: *Structures of Social Action*. Ed. by J. Maxwell Atkinson and John Heritage. Cambridge, UK: Cambridge University Press, pp. 299–345. DOI: 10.1017/CBO9780511665868.020 ⊸ pp. 6, 56, 110.

Hermann, Eduard (1913). 'Über die primären Interjektionen'. In: *Indoeuropäische Forschungen* 31, pp. 24–34. DOI: 10.1515/9783110242713.24 ⊸ p. 52.

Heylen, Dirk (2006). 'Head gestures, gaze and the principle of conversational structure'. In: *International Journal of Humanoid Robotics* 3, pp. 241–267. DOI: 10.1142/S0219843606000746 ⊸ pp. 100, 145.

Heylen, Dirk (2008). 'Listening heads'. In: *Modeling Communication with Robots and Virtual Humans*. Ed. by Ipke Wachsmuth and Günther Knoblich. Berlin, Germany: Springer, pp. 241–259. DOI: 10.1007/978-3-540-79037-2_13 ⊸ p. 54.

Heylen, Dirk, Elisabetta Bevacqua, Marion Tellier and Catherine Pelachaud (2007). 'Searching for prototypical facial feedback signals'. In: *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA)*. Paris, France, pp. 147–153. DOI: 10.1007/978-3-540-74997-4_14 ⊸ p. 55.

Heylen, Dirk, Ivo van Es, Anton Nijholt and Betsy van Dijk (2005). 'Controlling the gaze of conversational agents'. In: *Advances in Natural Multimodal Dialogue Systems*. Ed. by Jan C. J. van Kuppevelt, Laila Dybkjær and Niels Ole Bernsen. Dordrecht, The Netherlands: Springer, pp. 245–262. DOI: 10.1007/1-4020-3933-6_11 ⊸ p. 169.

Heylen, Dirk, Stefan Kopp, Stacy C. Marsella, Catherine Pelachaud and Hannes Högni Vilhjálmsson (2008). 'The Next Step Towards a Function Markup Language'. In: *Proceedings of the 8th International Conference on Intelligent Virtual Agents (IVA)*. Tokyo, Japan, pp. 270–280. DOI: 10.1007/978-3-540-85483-8_28 ⊸ p. 159.

Heylen, Dirk, Maarten Vissers, Rieks op den Akker and Anton Nijholt (2004). 'Affective Feedback in a Tutoring System for Procedural Tasks'. In: *Proceedings of the International Tutorial and Research Workshop on Affective Dialogue Systems*. Kloster Irsee, Germany, pp. 244–253 ⊸ p. 6.

Horn, Laurence R. (2004). 'Implicature'. In: *Handbook of Pragmatics*. Ed. by Laurence R. Horn and Gregory Ward. Malden, MA, USA: Blackwell, pp. 3–28 ⊸ p. 59.

Horton, William S. and Richard J. Gerrig (2002). 'Speakers' experiences and audience design: knowing when and knowing how to adjust utterances to addressees'. In: *Journal of Memory and Language* 47, pp. 589–606. DOI: 10.1016/s0749-596x(02)00019-0 ⊸ p. 33.

Horton, William S. and Boaz Keysar (1996). 'When do speakers take into account common ground?' In: *Cognition* 59, pp. 91–117. DOI: 10.1016/0010-0277(96)81418-1 ⊸ pp. 4, 33.

Hough, Julian (2015). 'Modelling Incremental Self-Repair Processing in Dialogue'. PhD thesis. London, UK: Queen Mary University of London ⊸ p. 134.

Hough, Julian and David Schlangen (2016). 'Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies'. In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*. Los Angeles, CA, USA, pp. 288–298 ⊸ p. 38.

Howes, Christine, Matthew Purver, Patrick G. T. Healey, Gregory Mills and Eleni Gregoromichelaki (2011). 'On incrementality in dialogue: evidence from compound contributions'. In: *Dialogue & Discourse* 2, pp. 279–311. DOI: 10.5087/dad.2011.111 ⊸ pp. 134, 156.

Huang, Lixing and Jonathan Gratch (2012). 'Crowdsourcing backchannel feedback: Understanding the individual variability from the crowds'. In: *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. Stevenson, WA, USA, pp. 31–34 ⊸ p. 185.

Huang, Lixing, Louis-Philippe Morency and Jonathan Gratch (2010). 'Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior'. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Toronto, ON, Canada, pp. 1265–1272 ⊸ pp. 66, 145.

Huber, Franz (2009). 'Belief and degrees of belief'. In: *Degrees of Belief*. Ed. by Franz Huber and Christoph Schmidt-Petri. Berlin, Germany: Springer, pp. 1–33. DOI: 10.1007/978-1-4020-9198-8_1 ⊸ p. 87.

Inden, Benjamin, Zofia Malisz, Petra Wagner and Ipke Wachsmuth (2013). 'Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent'. In: *Proceedings of the 15th International Conference on Multimodal Interaction (ICMI)*. Sydney, Australia, pp. 181–188. DOI: 10.1145/2522848.2522890 ⊸ p. 6.

Isard, Amy, Carsten Brockmann and Jon Oberlander (2006). 'Individuality and alignment in generated dialogues'. In: *Proceedings of the 4th International Natural Language Generation Conference (INLG)*. Sydney, Australia, pp. 25–32 ⊸ pp. 36, 126.

Isbister, Katherine and Patrick Doyle (2004). 'The blind men and the elephant revisited'. In: *From Brows to Trust. Evaluating Embodied Conversational Agents*. Ed. by Zsófia M. Ruttkay and Catherine Pelachaud. Dordrecht, The Netherlands: Kluwer Academic, pp. 3–26. DOI: 10.1007/1-4020-2730-3_1 ⊸ pp. 8, 177–178.

Janarthanam, Srinivasan and Oliver Lemon (2014). 'Adaptive generation in dialogue systems using dynamic user modeling'. In: *Computational Linguistics* 40, pp. 883–920. DOI: 10.1162/COLI_a_00203 ⊸ pp. 123, 126.

Jeffreys, Harold (1961). *Theory of Probability*. 3rd. Oxford, UK: Clarendon Press ⊸ p. 196.

Jensen, Finn Verner and Thomas Dyhre Nielsen (2007). *Bayesian Networks and Decision Graphs*. 2nd. New York, NY, USA: Springer. DOI: 10.1007/978-0-387-68282-2 ⊸ p. 83.

Jong, Markus A. de, Mariët Theune and Dennis Hofs (2008). 'Politeness and alignment in dialogues with a virtual guide'. In: *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Estoril, Portugal, pp. 207–214 ⊸ pp. 36, 126.

Joshi, Aravind K. (1987). 'The relevance of Tree Adjoining Grammar to generation'. In: *Natural Language Generation. New Results in Artificial Intelligence, Psychology and Linguistics*. Ed. by Gerard Kempen. Leiden, The Netherlands: Martinus Nijhoff Publishers, pp. 233–252. DOI: 10.1007/978-94-009-3645-4_16 ⊸ p. 126.

Joshi, Aravind K. and Yves Schabes (1997). 'Tree-Adjoining Grammars'. In: *Handbook of Formal Languages*. Ed. by Grzegorz Rozenberg and Arto Salomaa. Vol. 3. Berlin, Germany: Springer, pp. 69–123. DOI: 10.1007/978-3-642-59126-6_2 ⊸ p. 126.

Jurafsky, Dan and James H. Martin (2000). *Speech and Language Processing*. Upper Saddle River, NJ, USA: Prentice Hall ⊸ p. 2.

Kaeshammer, Miriam and Vera Demberg (2012). 'German and English treebanks and lexica for Tree-Adjoining Grammars'. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey, pp. 1880–1887 ⊸ p. 131.

Kaukomaa, Timo, Anssi Peräkylä and Johanna Ruusuvuori (2015). 'How listeners use facial expression to shift the emotional stance of the speaker's utterance'. In: *Research on Language and Social Interaction* 48, pp. 319–341. DOI: 10.1080/08351813.2015.1058607 ⊸ p. 55.

Kawahara, Tatsuya, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi and Nigel Ward (2016). 'Prediction and generation of backchannel form for attentive listening systems'. In: *Proceedings of Interspeech 2016*. San Francisco, CA, USA, pp. 2890–2894. DOI: 10.21437/Interspeech.2016-118 ⊸ p. 70.

Kelley, John F. (1983). 'An empirical methodology for writing user-friendly natural language computer applications'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Boston, MA, USA, pp. 193–196. DOI: 10.1145/800045.801609 ⊸ p. 155.

Kempen, Gerard and Edward Hoenkamp (1987). 'An incremental procedural grammar for sentence formulation'. In: *Cognitive Science* 11, pp. 201–258. DOI: 10.1207/s15516709cog1102_5 ⊸ pp. 29, 125.

Kempson, Ruth, Wilfried Meyer-Viol and Dov M. Gabbay (2001). *Dynamic Syntax: The Flow of Language Understanding*. Oxford, UK: Blackwell ⊸ p. 67.

Kendon, Adam (2002). 'Some uses of the head shake'. In: *Gesture* 2, pp. 147–182. DOI: 10.1075/gest.2.2.03ken ⊸ p. 54.

Kennington, Casey, Spyros Kousidis, Timo Baumann, Hendrik Buschmeier, Stefan Kopp and David Schlangen (2014). 'Better driving and recall when in-car information presentation uses situationally-aware incremental speech output generation'. In: *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutoUI)*. Seattle, WA, USA, 7:1–7:7. DOI: 10.1145/2667317.2667332 ⊸ p. 181.

Keysar, Boaz (1997). 'Unconfounding common ground'. In: *Discourse Processes* 24, pp. 253–270. DOI: 10.1080/01638539709545015 ⊸ p. 33.

Keysar, Boaz, Dale J. Barr, Jennifer A. Balin and Timothy S. Paek (1998). 'Definite reference and mutual knowledge: process models of common ground in comprehension'. In: *Journal of Memory and Language* 39, pp. 1–20. DOI: 10.1006/jmla.1998.2563 ⊸ p. 33.

Kilger, Anne and Wolfgang Finkler (1995). *Incremental Generation for Real-Time Applications*. Tech. rep. RR-95-11. Saarbrücken, Germany: Deutsches Forschungszentrum für Künstliche Intelligenz ⊸ p. 125.

Koiso, Hanae, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa and Yasuharu Den (1998). 'An analysis of turn-taking and backchannels on prosodic and syntactic features in Japanese Map Task dialogs'. In: *Language and Speech* 41, pp. 295–321. DOI: 10.1177/002383099804100404 ⊸ pp. 65, 145.

Kok, Iwan de (2013). 'Listening Heads'. PhD thesis. Enschede, The Netherlands: University of Twente. DOI: 10.3990/1.9789036506489 ⊸ pp. 6, 48, 64, 66.

Kok, Iwan de and Dirk Heylen (2012). 'Analyzing nonverbal listener responses using parallel recordings of multiple listeners'. In: *Cognitive Processing* 13, pp. 499–506. DOI: 10.1007/s10339-012-0434-3 ⊸ pp. 66, 145.

Kok, Iwan de, Julian Hough, Felix Hülsmann, Mario Botsch, David Schlangen and Stefan Kopp (2015). 'A multimodal system for real-time action instruction in motor skill learning'. In: *Proceedings of the International Conference on Multimodal Interaction (ICMI)*. Seattle, WA, USA, pp. 355–362. DOI: 10.1145/2818346.2820746 ⊸ p. 168.

Koller, Alexander and Matthew Stone (2007). 'Sentence generation as a planning problem'. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*. Praha, Czech Republic, pp. 336–343 ⊸ pp. 127–128.

Koller, Daphne and Nir Friedman (2009). *Probabilistic Graphical Models. Principles and Techniques*. Cambridge, MA, USA: MIT Press ⊸ pp. 91–94, 98, 106, 112, 119, 249, 253–254.

Koller, Daphne and Avi Pfeffer (1997). 'Object-oriented Bayesian networks'. In: *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. Providence, RI, USA, pp. 302–313 ⊸ p. 106.

Kopp, Stefan (2010). 'Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors'. In: *Speech Communication* 52, pp. 587–597. DOI: 10.1016/j.specom.2010.02.007 ⊸ pp. 3, 29, 36, 40, 247.

Kopp, Stefan, Jens Allwood, Karl Grammar, Elisabeth Ahlsén and Thorsten Stocksmeier (2008). 'Modeling embodied feedback with virtual humans'. In: *Modeling Communication with Robots and Virtual Humans*. Ed. by Ipke Wachsmuth and Günther Knoblich. Berlin, Germany: Springer, pp. 18–37. DOI: 10.1007/978-3-540-79037-2_2 ⊸ pp. 6, 55, 58, 68–69, 82, 85.

Kopp, Stefan, Lars Gesellensetter, Nicole C. Krämer and Ipke Wachsmuth (2005). 'A conversational agent as museum guide — Design and evaluation of a real-world application'. In: *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents (IVA)*. Kos, Greece, pp. 329–343. DOI: 10.1007/11550617_28 ⊸ p. 181.

Kopp, Stefan, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn Rúnar Thórisson et al. (2006). 'Towards a common framework for multimodal generation: the behavior markup language'. In: *Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA)*. Marina del Rey, CA, USA, pp. 205–217. DOI: 10.1007/11821830_17 ⊸ pp. 159–160.

Kopp, Stefan and Ipke Wachsmuth (2004). 'Synthesizing Multimodal Utterances for Conversational Agents'. In: *Computer Animation and Virtual Worlds* 15, pp. 39–52. DOI: 10.1002/cav.6 ⊸ p. 161.

Kopp, Stefan, Herwin van Welbergen, Ramin Yaghoubzadeh and Hendrik Buschmeier (2014). 'An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing'. In: *Journal on Multimodal User Interfaces* 8, pp. 97–108. DOI: 10.1007/s12193-013-0130-3 ⊸ pp. 153, 156, 159, 161–162.

Kousidis, Spyros, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp and David Schlangen (2014). 'A multimodal in-car dialogue system that tracks the driver's attention'. In: *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI)*. Istanbul, Turkey, pp. 26–33. DOI: 10.1145/2663204.2663244 ⊸ p. 181.

Krahmer, Emiel and Marc Swerts (2005). 'How children and adults produce and perceive uncertainty in audiovisual speech'. In: *Language and Speech* 48, pp. 29–53. DOI: 10.1177/00238309050480010201 ⊸ pp. 55, 100.

Kranstedt, Alfred, Stefan Kopp and Ipke Wachsmuth (2002). 'MURML: A multimodal utterance representation markup language for conversational agents'. In: *Proceedings of the AAMAS 2002 Workshop on Embodied Conversational Agents — Let's specify and evaluate them!* Bologna, Italy ⊸ p. 161.

Krauss, Robert M. and Sidney Weinheimer (1966). 'Concurrent feedback, confirmation, and the encoding of referents in verbal communication'. In: *Journal of Personality and Social Psychology* 4, pp. 343–346. DOI: 10.1037/h0023705 ⊸ pp. 43, 71.

Kraut, Robert E., Steven H. Lewis and Lawrence W. Swezey (1982). 'Listener responsiveness and the coordination of conversation'. In: *Journal of Personality and Social Psychology* 43, pp. 718–731. DOI: 10.1037/0022-3514.43.4.718 ⊸ p. 6.

Krengel, Ulrich (2005). *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. 8th. Wiesbaden, Germany: Vieweg ⊸ pp. 87–88.

Krippendorff, Klaus (1986). *A Dictionary of Cybernetics*. Norfolk, VA, USA: The American Society for Cybernetics ⊸ p. 61.

Kullback, Solomon and Richard Leibler (1951). 'On information and sufficiency'. In: *The Annals of Mathematical Statistics* 22, pp. 79–86. DOI: 10.1214/aoms/1177729694 ⊸ p. 147.

Lai, Catherine (2008). 'Prosodic cues for backchannels and short questions: Really?' In: *Proceedings of the 4th Conference on Speech Prosody*. Campinas, Brazil, pp. 413–416 ⊸ p. 53.

Lai, Catherine (2009). 'Perceiving surprise on cue words: prosody and semantics interact on right and really'. In: *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brighton, UK, pp. 1963–1966 ⊸ p. 53.

Lai, Catherine (2010). 'What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue'. In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan, pp. 1413–1416 ⊸ pp. 46, 53.

Larsson, Staffan and David R. Traum (2000). 'Information state and dialogue management in the TRINDI dialogue move engine toolkit'. In: *Natural Language Engineering* 6, pp. 323–340. DOI: 10.1017/S1351324900002539 ⊸ pp. 38, 84, 163.

Leavitt, Harold J. and Ronald A. H. Mueller (1951). 'Some effects of feedback on communication'. In: *Human Relations* 4, pp. 401–410. DOI: 10.1177/001872675100400406 ⊸ pp. 42–43.

Lee, Eun-Kyung, Sarah Brown-Schmidt and Duane G. Watson (2013). 'Ways of looking ahead: Hierarchical planning in language production'. In: *Cognition* 129, pp. 544–562. DOI: 10.1016/j.cognition.2013.08.007 ⊸ p. 129.

Lee, Jina and Stacy Marsella (2006). 'Nonverbal behavior generator for embodied conversational agents'. In: *Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA)*. Marina del Rey, CA, USA, pp. 243–255. DOI: 10.1007/11821830_20 ⊸ p. 161.

Lee, Jina, Stacy Marsella, David R. Traum, Jonathan Gratch and Brent Lance (2007). 'The Rickel gaze model: A window on the mind of a virtual human'. In: *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA)*. Paris, France, pp. 296–303. DOI: 10.1007/978-3-540-74997-4_27 ⊸ p. 168.

Lee, Sooha Park, Jeremy B. Badler and Norman I. Badler (2002). 'Eyes alive'. In: *ACM Transactions on Graphics* 21, pp. 637–644. DOI: 10.1145/566654.566629 ⊸ p. 170.

Levelt, Willem J. M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA, USA: MIT Press ⊸ pp. 29, 44, 125, 128, 156.

Levinson, Stephen C. (1995). 'Interactional biases in human thinking'. In: *Social Intelligence and Interaction. Expressions and Implications of the Social Bias in Human Intelligence*. Ed. by Esther N. Goody. Cambridge, UK: Cambridge University Press, pp. 221–260. DOI: 10.1017/CBO9780511621710.014 ⊸ pp. 5, 8, 40, 77, 248.

Levinson, Stephen C. (2006). 'On the human "Interaction Engine"'. In: *Roots of Human Sociality: Culture, Cognition and Interaction*. Ed. by Nicholas J. Enfield and Stephen C. Levinson. Oxford, UK: Berg, pp. 39–69 ⊸ pp. 4–5, 20.

Lewis, David (1969/2002). *Convention. A Philosophical Study*. Oxford, UK: Blackwell ⊸ pp. 17, 20.

Li, Shuyin, Britta Wrede and Gerhard Sagerer (2006). 'A computational model of multi-modal grounding for human robot interaction'. In: *Proceedings of the 7th Workshop on Discourse and Dialogue (SIGdial)*. Sydney, Australia, pp. 153–160 ⊸ p. 27.

Lindblom, Björn (1990). 'Explaning phonetic variation: A sketch of the H&H theory'. In: *Speech Production and Speech Modelling*. Ed. by William J. Hardcastle and Alain Marchal. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 403–439. DOI: 10.1007/978-94-009-2037-8_16 ⊸ p. 122.

Linell, Per (2009). *Rethinking Language, Mind, and World Dialogically. Interactional and Contextual Theories of Human Sense-Making*. Charlotte, NC, USA: Information Age ⊸ p. 13.

Locke, John (1690/1979). *An Essay Concerning Human Understanding*. Ed. by Peter H. Nidditch. Oxford, UK: Clarendon ⊸ pp. 15–16, 19.

Lotz, Alicia Flores, Ingo Siegert and Andreas Wendemuth (2016). 'Classification of functional-meanings of non-isolated discourse particles in human–human-interaction'. In: *Proceedings of the 18th HCI International Conference*. Toronto, ON, Canada, pp. 53–64. DOI: 10.1007/978-3-319-39510-4_6 ⊸ pp. 53, 118.

Maclay, Howard and Stanlay Newman (1960). 'Two variables affecting the message in communication'. In: *Decision, Values and Groups. Reports from the 1st Interdiciplinary Conference in the Behavioral Science Division, held at the University of New Mexico*. Ed. by Dorothy K. Willner. New York, NY, USA: Pergamon, pp. 218–228 ⊸ pp. 43, 71.

Mairesse, François and Marylin A. Walker (2010). 'Towards personality-based user adaptation: Psychologically informed stylistic language generation'. In: *User Modeling and User-Adapted Interaction* 20, pp. 227–278. DOI: 10.1007/s11257-010-9076-2 ⊸ pp. 36, 126, 137.

Malisz, Zofia, Marcin Włodarczak, Hendrik Buschmeier, Stefan Kopp and Petra Wagner (2012). 'Prosodic characteristics of feedback expressions in distracted and non-distracted listeners'. In: *Proceedings of The Listening Talker. An Interdisciplinary Workshop on Natural and Synthetic Modification of Speech in Response to Listening Conditions*. Edinburgh, UK, pp. 36–39 ⊸ pp. 41, 52, 172.

Malisz, Zofia, Marcin Włodarczak, Hendrik Buschmeier, Joanna Skubisz, Stefan Kopp and Petra Wagner (2016). 'The ALICO corpus: Analysing the active listener'. In: *Language Resources and Evaluation* 50, pp. 411–442. DOI: 10.1007/s10579-016-9355-6 ⊸ pp. 41, 49, 51, 54, 97, 192, 199–200.

Maxwell, J. Clerk (1867). 'On Governors'. In: *Proceedings of the Royal Society of London* 16, pp. 270–283. DOI: 10.1098/rspl.1867.0055 ⊸ p. 41.

Mayr, Otto (1970). *The Origins of Feedback Control*. Cambridge, MA, USA: MIT Press ⊸ pp. 41–42.

McClave, Evelyn Z. (2000). 'Linguistic functions of head movement in the context of speech'. In: *Journal of Pragmatics* 32, pp. 855–878. DOI: 10.1016/S0378-2166(99)00079-X ⊸ p. 145.

McKinley, Nathan and Soumya Ray (2014). 'A decision-theoretic approach to natural language generation'. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD, USA, pp. 552–561 ⊸ pp. 127–128.

McTear, Michael F. (2002). 'Spoken dialogue technology: Enabling the conversational user interface'. In: *ACM Computing Surveys* 34, pp. 90–169. DOI: 10.1145/505282.505285 ⊸ pp. 1, 37.

Milch, Brian and Daphne Koller (2000). 'Probabilistic models for agents' beliefs and decisions'. In: *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*. Stanford, CA, USA, pp. 389–396 ⊸ p. 86.

Misu, Teruhisa, Etsuo Mizukami, Yoshinori Shiga, Shinichi Kawamoto, Hisashi Kawai and Satoshi Nakamura (2011a). 'Analysis on effects of text-to-speech and avatar agent in evoking users' spontaneous listener's reactions'. In: *Proceedings of the Workshop on Paralinguistic Information and its Integration in Spoken Dialogue Systems*. Granada, Spain, pp. 77–89. DOI: 10.1007/978-1-4614-1335-6_10 ⊸ p. 146.

Misu, Teruhisa, Etsuo Mizukami, Yoshinori Shiga, Shinichi Kawamoto, Hisashi Kawai and Satoshi Nakamura (2011b). 'Toward construction of spoken dialogue system that evokes users' spontaneous backchannels'. In: *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*. Portland, OR, USA, pp. 259–265 ⊸ pp. 145, 245.

Mol, Lisette, Emiel Krahmer, Alfons Maes and Marc Swerts (2012). 'Adaptation in gesture: Converging hands or converging minds?' In: *Journal of Memory and Language* 66, pp. 249–264. DOI: 10.1016/j.jml.2011.07.004 ⊸ p. 29.

Morency, Louis-Philippe, Iwan de Kok and Jonathan Gratch (2010). 'A probabilistic multimodal approach for predicting listener backchannels'. In: *Autonomous Agents and Multiagent Systems* 20, pp. 70–84. DOI: 10.1007/s10458-009-9092-y ⊸ pp. 6, 65–66, 145.

Morency, Louis-Philippe, Candace Sidner, Christopher Lee and Trevor Darrell (2007). 'Head gestures for perceptual interfaces: The role of context in improving recognition'. In: *Artificial Intelligence* 171, pp. 568–585. DOI: 10.1016/j.artint.2007.04.003 ⊸ p. 173.

Morey, Richard D. (2014). *Bayes factor t tests, part 2: Two-sample tests*. URL: http://bayesfactor.blogspot.com/2014/02/bayes-factor-t-tests-part-2-two-sample.html (visited on 13/05/2016) ⊸ pp. 199, 203.

Morey, Richard D. (2015). *Multiple comparisons with BayesFactor, part 2: Order restrictions*. URL: http://bayesfactor.blogspot.de/2015/01/multiple-comparisons-with-bayesfactor-2.html (visited on 13/05/2016) ⊸ pp. 199, 210.

Morey, Richard D. and Jeffrey N. Rouder (2015). *BayesFactor: Computation of Bayes factors for common designs*. Version 0.9.12-2. URL: https://CRAN.R-project.org/package=BayesFactor ⊸ pp. 196, 203.

Mukai, Toshiro, Susumu Seki, Masayuki Nakazawa, Keiko Watanuki and Hideo Miyoshi (1999). 'Multimodal agent interface based on dynamical dialogue model'. In: *Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology*. Asheville, NC, USA, pp. 69–70. DOI: 10.1145/320719.322586 ⊸ p. 70.

Muller, Philippe and Laurent Prévot (2003). 'An empirical study of acknowledgement structures'. In: *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*. Saarbrücken, Germany ⊸ p. 116.

Myers, Karen, Pauline Berry, Jim Blythe, Ken Conley, Melinda Gervasio, Deborah McGuinness, David Morley et al. (2007). 'An intelligent personal assistant for task and time management'. In: *AI Magazine* 28, pp. 47–61 ⊸ p. 173.

Nakano, Mikio, Kohji Dohsaka, Noboru Miyazaki, J. Hirasawa, Masafumi Tamoto, Masahito Kawamori, Akira Sugiyama et al. (1999). 'Handling rich turn-taking in spoken dialogue systems'. In: *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*. Budapest, Hungary, pp. 1167–1170 ⊸ p. 68.

Nakano, Yukiko I., Gabe Reinstein, Tom Stocky and Justine Cassell (2003). 'Towards a model of face-to-face grounding'. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*. Sapporo, Japan, pp. 553–561. DOI: 10.3115/1075096.1075166 ⊸ pp. 55, 70.

Neiberg, Daniel and Joakim Gustafson (2010). 'The prosody of Swedish conversational grunts'. In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan, pp. 2562–2565 ⊸ p. 49.

Neiberg, Daniel, Giampiero Salvi and Joakim Gustafson (2013). 'Semi-supervised methods for exploring the acoustics of simple productive feedback'. In: *Speech Communication* 55, pp. 451–469. DOI: 10.1016/j.specom.2012.12.007 ⊸ pp. 6, 53, 97, 118.

Nivre, Joakim (1995). 'Communicative action and feedback'. In: *Dialogue and Instruction. Modelling Interaction in Intelligent Tutoring Systems*. Ed. by Robbert-Jan Beun, Michael Baker and Miriam Reiner. Berlin, Germany: Springer, pp. 231–240 ⊸ p. 63.

Novick, David G. and Stephen Sutton (1994). 'An empirical model of acknowledgement for spoken-language systems'. In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Las Cruces, NM, USA, pp. 96–101. DOI: 10.3115/981732.981746 ⊸ p. 6.

Oertel, Catharine, José Lopes, Yu Yu, Kenneth A. Funes Mora, Joakim Gustafson, Alan W. Black and Jean-Marc Odobez (2016). 'Towards building an attentive artificial listener: On the perception of attentiveness in audio-visual feedback tokens'. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*. Tokyo, Japan, pp. 21–28. DOI: 10.1145/2993148.2993188 ⊸ p. 6.

OGRE Team (2013). *OGRE: Object-oriented graphics rendering engine*. Version 1.9. URL: http://www.ogre3d.org/ ⊸ p. 155.

Oviatt, Sharon, Courtney Darves and Rachel Coulston (2004). 'Toward adaptive conversational interfaces: Modeling speech convergence with animated personas'. In: *ACM Transactions on Computer–Human Interaction* 11, pp. 300–328. DOI: 10.1145/1017494.1017498 ⊸ p. 29.

Paek, Timothy S. and Eric Horvitz (2000a). 'Conversation as action under uncertainty'. In: *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*. Stanford, CA, USA, pp. 455–464 ⊸ pp. 28, 117, 119.

Paek, Timothy S. and Eric Horvitz (2000b). *Grounding criterion: Toward a formal theory of grounding*. Tech. rep. MSR-TR-2000-40. Redmond, WA, USA: Microsoft Research ⊸ pp. 28, 109.

Pammi, Sathish (2011). 'Synthesis of Listener Vocalizations. Towards Interactive Speech Synthesis'. PhD thesis. Saarbrücken, Germany: Naturwissenschaftlich-Technische Fakultät I, Universität des Saarlandes ⊸ p. 46.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann ⊸ p. 92.

Pelachaud, Catherine and Massimo Bilvi (2003). 'Modelling gaze behavior for conversational agents'. In: *Proceedings of the 4th International Workshop on Intelligent Vir-*

tual Agents (IVA). Kloster Irsee, Germany, pp. 93–100. DOI: 10.1007/978-3-540-39396-2_16 ⊸ p. 168.

Petukhova, Volha and Harry Bunt (2010). 'Introducing communicative function qualifiers'. In: *Proceedings of the Second International Conference on Global Interoperability for Language Resources*. Hong Kong, China, pp. 123–131 ⊸ pp. 73, 86, 109.

Philippsen, Anja K., Kai. A. Mismahl, Britta Wrede and Yukie Nagai (2013). 'Cross-cultural recognition of auditive feedback using echo-state networks'. In: *Proceedings der 24. Konferenz zur elektronischen Sprachsignalverarbeitung (ESSV)*. Bielefeld, Germany, pp. 173–180 ⊸ pp. 53, 118.

Pickering, Martin J. and Simon Garrod (2004). 'Toward a mechanistic psychology of dialogue'. In: *Behavioral and Brain Sciences* 27, pp. 169–226. DOI: 10.1017/S0140525X04000056 ⊸ pp. 3, 29, 31, 132.

Pitt, David (2013). 'Mental representation'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2013. URL: http://plato.stanford.edu/archives/fall2013/entries/mental-representation/ ⊸ p. 82.

Poesio, Massimo and Hannes Rieser (2010). 'Completions, coordination, and alignment in dialogue'. In: *Dialogue and Discourse* 1.1, pp. 1–89. DOI: 10.5087/dad.2010.001 ⊸ p. 38.

Poesio, Massimo and David R. Traum (1997). 'Conversational actions and discourse situations'. In: *Computational Intelligence* 13, pp. 309–347. DOI: 10.1111/0824-7935.00042 ⊸ pp. 25, 38, 109, 112, 129.

Poggi, Isabella, Francesca D'Errico and Laura Vincze (2010a). 'Agreement and its multimodal communication in debates: A qualitative analysis'. In: *Cognitive Computation* 3, pp. 466–479. DOI: 10.1007/s12559-010-9068-x ⊸ p. 54.

Poggi, Isabella, Francesca D'Errico and Laura Vincze (2010b). 'Types of nods. The polysemy of a social signal'. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*. Valletta, Malta, pp. 2570–2576 ⊸ p. 54.

Pompino-Marschall, Bernd (2004). 'Zwischen Tierlaut und sprachlicher Artikulation: Zur Phonetik der Interjektionen [Between animal vocalisation and spoken articulation: On the phonetics of interjections]'. In: *Zeitschrift für Semiotik* 26, pp. 71–84 ⊸ p. 49.

Pon-Barry, Heather (2008). 'Prosodic manifestations of confidence and uncertainty in spoken language'. In: *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia, pp. 74–77 ⊸ p. 100.

Poppe, Ronald, Khiet P. Truong and Dirk Heylen (2011). 'Backchannels: Quantity, type and timing matters'. In: *Proceedings of the 11th International Conference on Intelligent Virtual Agents (IVA)*. Reykjavík, Iceland, pp. 228–239. DOI: 10.1007/978-3-642-23974-8_25 ⊸ p. 6.

Premack, David and Guy Woodruff (1978). 'Does the chimpanzee have a theory of mind?' In: *Behavioral and Brain Sciences* 1, pp. 515–526. DOI: 10.1017/s0140525x00076512 ⊸ p. 82.

Prévot, Laurent, Jan Gorisch and Roxane Bertrand (2016). 'A CUP of CoFee: A large collection of feedback utterances provided with communicative function annotations'. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portorož, Slovenia, pp. 3810–3815 ⊸ p. 50.

Purver, Matthew (2004). 'The Theory and Use of Clarification Requests in Dialogue'. PhD thesis. London, UK: King's College, University of London ⊸ pp. 4, 18.

Rammstedt, Beatrice and Oliver P. John (2007). 'Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German'. In: *Journal of Research in Personality* 41, pp. 203–212. DOI: 10.1016/j.jrp.2006.02.001 ⊸ pp. 185, 198.

Rapaport, William J. (2003). 'What did you mean by that? Misunderstanding, negotiation, and syntactic semantics'. In: *Minds and Machines* 13, pp. 397–427. DOI: 10.1023/A:1024145126190 ⊸ pp. iii, 3, 8, 39.

Raymond, Eric S., ed. (2003). *Jargon File 4.4.7*. URL: http://jargon-file.org/archive/jargon-4.4.7.dos.txt (visited on 05/01/2017) ⊸ p. 1.

Reidsma, Dennis, Iwan de Kok, Daniel Neiberg, Sathish Pammi, Bart van Straalen, Khiet Truong and Herwin van Welbergen (2011). 'Continuous interaction with a virtual human'. In: *Journal on Multimodal User Interfaces* 4, pp. 97–118. DOI: 10.1007/s12193-011-0060-x ⊸ pp. 6, 70, 79, 118, 146, 200, 238, 245, 247.

Reidsma, Dennis and Herwin van Welbergen (2013). 'AsapRealizer in practice — A modular and extensible architecture for a BML realizer'. In: *Entertainment Computing* 4, pp. 157–169. DOI: 10.1016/j.entcom.2013.05.001 ⊸ pp. 155, 162.

Reiter, Ehud (1994). 'Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible?' In: *Proceedings of the 7th International Workshop on Natural Language Generation (INLG)*. Kennebunkport, ME, USA, pp. 163–170 ⊸ p. 125.

Reiter, Ehud and Robert Dale (2000). *Building Natural Language Generation Systems*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511519857 ⊸ pp. 124–125, 128.

Reiter, Ehud and Somayajulu Sripada (2002). 'Human variation and lexical choice'. In: *Computational Linguistics* 28, pp. 545–553. DOI: 10.1162/089120102762671981 ⊸ p. 123.

Reitter, David and Johanna D. Moore (2014). ‘Alignment and task success in spoken dialogue’. In: *Journal of Memory and Language* 76, pp. 29–46. DOI: 10.1016/j.jml.2014.05.008 ⊸ pp. 29, 31.

Riek, Laurel D. (2012). ‘Wizard of Oz studies in HRI: A systematic review and new reporting guidelines’. In: *Journal of Human-Robot Interaction* 1, pp. 119–136. DOI: 10.5898/jhri.1.1.riek ⊸ p. 155.

Robert, Christian P. (1993). ‘Prior feedback: A Bayesian approach to maximum likelihood estimation’. In: *Computational Statistics* 8, pp. 279–294 ⊸ pp. 113, 172.

Roque, Antonio and David R. Traum (2008). ‘Degrees of grounding based on evidence of understanding’. In: *Proceedings of the 9th Workshop on Discourse and Dialogue (SIGdial)*. Columbus, OH, USA, pp. 54–63 ⊸ pp. 26, 107, 109, 246.

Rosch, Eleanor (1978). ‘Principles of categorization’. In: *Cognition and Categorization*. Ed. by Eleanor Rosch and B. Lloyd. Hillsdale, NJ, USA: Lawrence Erlbaum, pp. 189–206 ⊸ p. 16.

Rouder, Jeffrey N., Richard D. Morey, Paul L. Speckman and Jordan M. Province (2012). ‘Default Bayes factors for ANOVA designs’. In: *Journal of Mathematical Psychology* 56, pp. 356–374. DOI: 10.1016/j.jmp.2012.08.001 ⊸ pp. 199, 203.

Rouder, Jeffrey N., Paul L. Speckman, Dongchu Sun, Richard D. Morey and Geoffrey Iverson (2009). ‘Bayesian *t* tests for accepting and rejecting the null hypothesis’. In: *Psychonomic Bulletin & Review* 16, pp. 225–237. DOI: 10.3758/pbr.16.2.225 ⊸ pp. 199, 203.

Rubin, Rebecca B. (1985). ‘The validity of the communication competency assessment instrument’. In: *Communication Monographs* 52, pp. 173–185. DOI: 10.1080/03637758509376103 ⊸ p. 182.

Russell, Stuart and Peter Norvig (2010). *Artificial Intelligence. A Modern Approach*. 3rd. Upper Saddle River, NJ, USA: Prentice Hall ⊸ pp. 1, 15, 88.

Ruttkay, Zsófia M., Claire Dormann and Han Noot (2004). ‘Embodied conversational agents on a common ground. A framework for design and evaluation’. In: *From Brows to Trust. Evaluating Embodied Conversational Agents*. Ed. by Zsófia M. Ruttkay and Catherine Pelachaud. Dordrecht, The Netherlands: Kluwer Academic, pp. 27–66. DOI: 10.1007/1-4020-2730-3_2 ⊸ pp. 178, 183–185.

Ruttkay, Zsófia M. and Catherine Pelachaud, eds. (2004). *From Brows to Trust. Evaluating Embodied Conversational Agents*. Dordrecht, The Netherlands: Kluwer Academic. DOI: 10.1007/1-4020-2730-3.

Sacks, Harvey, Emanuel A. Schegloff and Gail Jefferson (1974). ‘A simplest systematics for the organization of turn-taking for conversation’. In: *Language* 50, pp. 696–735 ⊸ p. 64.

Schäfer, Ralph, Thomas Weis, Thomas Weyrath and Anthony Jameson (1997). ‘Wie können Ressourcenbeschränkungen eines Dialogpartners erkannt und berücksichtigt werden [How can resource limitations of a dialogue partner be recognised and taken into account]’. In: *Kognitionswissenschaft* 6, pp. 151–164. DOI: 10.1007/BF03354918 ⊸ p. 117.

Schegloff, Emanuel A. (1982). ‘Discourse as an interactional achievement: Some uses of ‘uh-huh’ and other things that come between sentences’. In: *Analyzing Discourse: Text and Talk. 32nd Georgetown Round Table on Languages and Linguistics 1981*. Ed. by Deborah Tannen. Washington, DC, USA: Georgetown University Press, pp. 71–93 ⊸ pp. 6, 67.

Schegloff, Emanuel A., Gail Jefferson and Harvey Sacks (1977). ‘The preference for self-correction in the organization of repair in conversation’. In: *Language* 53, pp. 361–382 ⊸ pp. 18, 32, 67.

Schlangen, David (2004). ‘Causes and strategies for requesting clarification in dialogue’. In: *Proceedings of the 5th Workshop on Discourse and Dialogue (SIGdial)*. Boston, MA, USA, pp. 136–143 ⊸ p. 18.

Schlangen, David, Timo Baumann, Hendrik Buschmeier, Okko Buß, Stefan Kopp, Gabriel Skantze and Ramin Yaghoubzadeh (2010). ‘Middleware for incremental processing in conversational agents’. In: *Proceedings of the 11th Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGdial)*. Tokyo, Japan, pp. 51–54 ⊸ pp. 157, 282.

Schlangen, David and Gabriel Skantze (2011). ‘A general, abstract model of incremental dialogue processing’. In: *Dialogue and Discourse* 2, pp. 83–111. DOI: 10.5087/dad.2011.105 ⊸ pp. 157, 170, 245, 282.

Schlesinger, Izchak M. and Sharon Hurvitz (2008). ‘The Structure of Misunderstandings’. In: *Pragmatics & Cognition* 16, pp. 568–585. DOI: 10.1075/pc.16.3.07sch ⊸ pp. 14–15, 17–18.

Schmidt, Jürgen Erich (2001). ‘Bausteine der Intonation?’ In: *Neue Wege der Intonationsforschung*. Ed. by Jürgen Erich Schmidt. Hildesheim, Germany: Georg Olms Verlag, pp. 9–32 ⊸ pp. 52–53.

Schröder, Marc, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat et al. (2012). ‘Building autonomous sensitive artificial listeners’. In: *IEEE Transactions on Affective Computing* 3, pp. 165–183. DOI: 10.1109/T-AFFC.2011.34 ⊸ pp. 6, 69, 145, 185.

Schröder, Marc and Jürgen Trouvain (2003). ‘The German text-to-speech synthesis system MARY: A tool for research, development and teaching’. In: *Interna-*

tional Journal of Speech Technology 6, pp. 365–377. DOI: 10.1023/A:1025708916924 ⊸ p. 139.

Schuller, Björn and Anton Batliner (2014). *Computational Paralinguistics. Emotion, Affect and Personality in Speech and Language Processing*. Chichester, UK: Wiley. DOI: 10.1002/9781118706664 ⊸ p. 172.

Scott-Phillips, Thom (2015). *Speaking Our Minds. Why Human Communication is Different, and How Language Evolved to Make it Special*. London, UK: Palgrave Macmillan ⊸ p. 5.

Scully, John (2010). 'Genius is seeing the obvious twenty years ahead of everyone else'. In: *Points of View: A Tribute to Alan Kay*. Ed. by Ian Piumarta and Kimberly Rose. Glendale, CA, USA: Viewpoints Research Institute, pp. 49–54 ⊸ p. 173.

Selting, Margret, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen et al. (2011). 'A system for transcribing talk-in-interaction: GAT 2'. Trans. by Elizabeth Couper-Kuhlen and Dagmar Barth-Weingarten. In: *Gesprächsforschung — Online-Zeitschrift zur verbalen Interaktion* 12, pp. 1–51 ⊸ pp. 110, 129.

Shannon, Claude E. (1948). 'A mathematical theory of communication'. In: *The Bell Systems Technical Journal* 27, pp. 379–423 ⊸ pp. 2, 46, 48, 57.

Siegert, Ingo, Kim Hartmann, David Philippou-Hübner and Andreas Wendemuth (2013). 'Human behaviour in HCI: Complex emotion detection through sparse speech features'. In: *Proceedings of the 4th International Workshop on Human Behaviour Understanding*. Barcelona, Spain, pp. 246–257. DOI: 10.1007/978-3-319-02714-2_21 ⊸ p. 46.

Skantze, Gabriel (2007). 'Making grounding decisions: Data-driven estimation of dialogue costs and confidence thresholds'. In: *Proceedings of the 8th Workshop on Discourse and Dialogue (SIGdial)*. Antwerpen, Belgium, pp. 206–210 ⊸ p. 28.

Skantze, Gabriel (2008). 'Galatea: A discourse modeller supporting concept-level error handling in spoken dialogue'. In: *Recent Trends in Discourse and Dialogue*. Ed. by Laila Dybkjær and Wolfgang Minker. Dordrecht, The Netherlands: Springer, pp. 155–189. DOI: 10.1007/978-1-4020-6821-8_7 ⊸ p. 38.

Skantze, Gabriel and Anna Hjalmarsson (2013). 'Towards incremental speech generation in conversational systems'. In: *Computer Speech and Language* 27, pp. 243–262. DOI: 10.1016/j.csl.2012.05.004 ⊸ pp. 125, 134.

Skantze, Gabriel, Anna Hjalmarsson and Catharine Oertel (2014). 'Turn-taking, feedback and joint attention in situated human–robot interaction'. In: *Speech Communication* 65, pp. 50–66. DOI: 10.1016/j.specom.2014.05.005 ⊸ pp. 6, 100.

Skantze, Gabriel and David Schlangen (2009). 'Incremental dialogue processing in a micro-domain'. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Athens, Greece, pp. 745–753 ⊸ p. 38.

Sperber, Dan and Deirdre Wilson (1986/1995). *Relevance. Communication and Cognition*. 2nd. Malden, MA, USA: Blackwell Publishing ⊸ p. 2.

Spink, Amanda and Tefko Saracevic (1998). 'Human–computer interaction in information retrieval: nature and manifestations of feedback'. In: *Interacting with Computers* 10, pp. 249–267. DOI: 10.1016/s0953-5438(98)00009-5 ⊸ pp. 45–46.

Stalnaker, Robert C. (2002). 'Common ground'. In: *Linguistics and Philosophy* 25, pp. 701–721. DOI: 10.1023/A:1020867916902 ⊸ p. 20.

Stocksmeier, Thorsten, Stefan Kopp and Dafydd Gibbon (2007). 'Synthesis of prosodic attitudinal variants in German backchannel "ja"'. In: *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Antwerpen, Belgium, pp. 1290–1293 ⊸ pp. 49–50, 52.

Stolz, Walter S. and Percy H. Tannenbaum (1963). 'Effects of feedback on oral encoding behaviour'. In: *Language and Speech* 6, pp. 218–228. DOI: 10.1177/002383096300600404 ⊸ pp. 6, 43, 45.

Stone, Matthew (2002). 'Lexicalized grammar 101'. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Philadelphia, PA, USA, pp. 77–84 ⊸ pp. 126–127, 134.

Stone, Matthew and Christine Doran (1997). 'Sentence planning as description using tree adjoining grammar'. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*. Madrid, Spain, pp. 198–205. DOI: 10.3115/976909.979643 ⊸ p. 127.

Stone, Matthew, Christine Doran, Bonnie L. Webber, Tonia Bleam and Martha Palmer (2003). 'Microplanning with communicative intentions: The SPUD system'. In: *Computational Intelligence* 19, pp. 311–381. DOI: 10.1046/j.0824-7935.2003.00221.x ⊸ pp. 36, 126–128, 138.

Stone, Matthew and Alex Lascarides (2010). 'Coherence and rationality in grounding'. In: *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*. Poznań, Poland, pp. 51–58 ⊸ p. 117.

Swartout, William, David R. Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams et al. (2010). 'Ada and Grace: toward realistic and engaging virtual museum guides'. In: *Proceedings of the 10th International Conference on Intelligent Virtual Agents (IVA)*. Philadelphia, PA, USA, pp. 286–300. DOI: 10.1007/978-3-642-15892-6_30 ⊸ p. 181.

Tanenhaus, Michael K. and Sarah Brown-Schmidt (2008). 'Language processing in the natural world'. In: *Philosophical Transactions of the Royal Society B* 363, pp. 1105–1122. DOI: 10.1098/rstb.2007.2162 ⊸ pp. 55, 156.

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard and Julie C. Sedivy (1995). 'Integration of visual and linguistic information in spoken language comprehension'. In: *Science* 268, pp. 1632–1634. DOI: 10.1126/science.7777863 ⊸ p. 156.

Taylor, Talbot J. (1986). 'Do you understand? Criteria of understanding in verbal interaction'. In: *Language & Communication* 6, pp. 171–180. DOI: 10.1016/0271-5309(86)90020-0 ⊸ pp. 14–15.

Taylor, Talbot J. (1992). *Mutual Misunderstanding. Scepticism and the Theorizing of Language and Interpretation*. London, UK: Routledge ⊸ pp. 15–16.

Thórisson, Kristinn Rúnar (1996). 'Communicative Humanoids. A Computational Model of Psychosocial Dialogue Skills'. PhD thesis. Cambridge, MA, USA: Massachusetts Institute of Technology ⊸ p. 160.

Tickle-Degnen, Linda and Robert Rosenthal (1990). 'The nature of rapport and its nonverbal correlates'. In: *Psychological Inquiry* 1, pp. 285–293. DOI: 10.1207/s15327965pli0104_1 ⊸ p. 175.

Tomasello, Michael (2003). *Constructing a Language. A Usage-based Theory of Language Acquisition*. Cambridge, MA, USA: Harvard University Press ⊸ pp. 16–17.

Traum, David R. (1994). 'A Computational Theory of Grounding in Natural Language Conversation'. PhD thesis. Rochester, NY, USA: University of Rochester ⊸ pp. 23–27, 107, 246.

Traum, David R. and Peter A. Heeman (1997). 'Utterance units in spoken dialogue'. In: *Dialogue Processing in Spoken Language Systems*. Ed. by Elisabeth Maier, Marion Mast and Susann LuperFoy. Berlin, Germany: Springer, pp. 125–140. DOI: 10.1007/3-540-63175-5_42 ⊸ pp. 112, 129, 149.

Turing, Alan M. (1950). 'Computing machinery and intelligence'. In: *Mind* 59, pp. 433–460. DOI: 10.1093/mind/LIX.236.433 ⊸ p. 37.

Tydgat, Ilse, Michael Stevens, Robert J. Hartsuiker and Martin J. Pickering (2011). 'Deciding where to stop speaking'. In: *Journal of Memory and Language* 64, pp. 359–380. DOI: 10.1016/j.jml.2011.02.002 ⊸ p. 124.

Uematsu, Shigeo (2000). 'The use of back channels between native and non-native speakers in English and Japanese'. In: *Intercultural Communication Studies* 10, pp. 85–98 ⊸ p. 67.

University of Chicago Press Staff, ed. (2010). *The Chicago Manual of Style. The Essential Guide for Writers, Editors, and Publishers*. 16th. Chicago, IL, USA: University of Chicago Press ⊸ p. 49.

Vaissière, Jacqueline (2005). 'Perception of intonation'. In: *The Handbook of Speech Perception*. Ed. by David B. Pisoni and Robert E. Remez. Malden, MA, USA: Blackwell, pp. 236–263. DOI: 10.1002/9780470757024.ch10 ⊸ pp. 50, 52.

Vilhjálmsson, Hannes Högni, Nathan Cantelmo, Justine Cassell, Nicolas E. Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini et al. (2007). 'The Behavior Markup Language: Recent developments and challenges'. In: *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA)*. Paris, France, pp. 99–111. DOI: 10.1007/978-3-540-74997-4_10 ⊸ p. 159.

Visser, Thomas, David R. Traum, David DeVault and Rieks op den Akker (2014). 'A model for incremental grounding in spoken dialogue systems'. In: *Journal on Multimodal User Interfaces* 8, pp. 61–73. DOI: 10.1007/s12193-013-0147-7 ⊸ pp. 27, 107.

Wachsmuth, Ipke, Jan Peter de Ruiter, Petra Jaecks and Stefan Kopp, eds. (2013). *Alignment in Communication. Towards a New Theory of Communication*. Amsterdam, The Netherlands: John Benjamins. DOI: 10.1075/ais.6 ⊸ p. 31.

Wagner, Petra, Zofia Malisz, Benjamin Inden and Ipke Wachsmuth (2013). 'Interaction phonology — A temporal co-ordination component enabling representational alignment within a model of communication'. In: *Alignment in Communication. Towards a New Theory of Communication*. Ed. by Ipke Wachsmuth, Jan Peter de Ruiter, Petra Jaecks and Stefan Kopp. Amsterdam, The Netherlands: John Benjamins, pp. 109–132. DOI: 10.1075/ais.6.06wag ⊸ p. 29.

Wagner, Petra, Zofia Malisz and Stefan Kopp (2014). 'Gesture and speech in interaction: An overview'. In: *Speech Communication* 57, pp. 209–232. DOI: 10.1016/j.specom.2013.09.008 ⊸ p. 54.

Walker, Marylin A., Diane J. Litman, Candace. A. Kamm and Alicia Abella (1998). 'Evaluating spoken dialogue agents with PARADISE: Two case studies'. In: *Computer Speech and Language* 12, pp. 317–347. DOI: 10.1006/csla.1998.0110 ⊸ p. 181.

Walker, Marylin A., Amanda Stent J., François Mairesse and Rashmi Prasad (2007). 'Individual and domain adaptation in sentence planning for dialogue'. In: *Journal of Artificial Intelligence Research* 30, pp. 413–456. DOI: 10.1613/jair.2329 ⊸ pp. 36, 126.

Walker, Willie, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf et al. (2004). *Sphinx-4: A Flexible Open Source Framework for Speech Recognition*. Tech. rep. SMLI TR2004-0811. Sun Microsystems Inc. ⊸ p. 172.

Wang, Zhiyang, Jina Lee and Stacy Marsella (2013). 'Multiparty, multi-role comprehensive listening behaviour'. In: *Autonomous Agents and Multiagent Systems* 27, pp. 218–234. DOI: 10.1007/s10458-012-9215-8 ⊸ pp. 68, 161.

Ward, Nigel (2000). 'The challenge of non-lexical speech sounds'. In: *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*. Bejing, China, pp. 571–574 ⊸ pp. 46, 49.

Ward, Nigel (2004). 'Pragmatic functions of prosodic features in non-lexical utterances'. In: *Proceedings of the International Conference on Speech Prosody*. Nara, Japan, pp. 325–328 ⊸ p. 52.

Ward, Nigel (2006). 'Non-lexical conversational sounds in American English'. In: *Pragmatics & Cognition* 14, pp. 129–182. DOI: 10.1075/pc.14.1.08war ⊸ pp. 49–51.

Ward, Nigel and Wataru Tsukahara (2000). 'Prosodic features which cue back-channel responses in English and Japanese'. In: *Journal of Pragmatics* 38, pp. 1177–1207. DOI: 10.1016/S0378-2166(99)00109-5 ⊸ pp. 6, 65, 145.

Watzlawick, Paul, Janet Helmick Beavin and Don D. Jackson (1967). *Pragmatics of Human Communication. A Study of Interactional Patterns, Pathologies, and Paradoxes*. New York, NY, USA: Norton ⊸ p. 62.

Weigand, Edda (1999). 'Misunderstanding: the standard case'. In: *Journal of Pragmatics* 31, pp. 763–785. DOI: 10.1016/s0378-2166(98)00068-x ⊸ pp. 17–19.

Weizenbaum, Joseph (1966). 'ELIZA — A computer program for the study of natural language communication between man and machine'. In: *Communications of the ACM* 9, pp. 36–45. DOI: 10.1145/365153.365168 ⊸ p. 37.

Welbergen, Herwin van (2011). 'Behavior Generation for Interpersonal Coordination with Virtual Humans. On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior'. PhD thesis. Enschede, The Netherlands: University of Twente. DOI: 10.3990/1.9789036532334 ⊸ pp. 161, 168.

Welbergen, Herwin van, Dennis Reidsma, Zsófia M. Ruttkay and Job Zwiers (2009). 'Elckerlyc — A BML realizer for continuous, multimodal interaction with a virtual human'. In: *Journal on Multimodal User Interfaces* 3, pp. 271–284. DOI: 10.1007/s12193-010-0051-3 ⊸ p. 161.

Welbergen, Herwin van, Yuyu Xu, Marcus Thiebaux, Wei-Wen Feng, Jingqiao Fu, Dennis Reidsma and Ari Shapiro (2011). 'Demonstrating and testing the BML compliance of BML realizers'. In: *Proceedings of the 11th International Conference on Intelligent Virtual Agents (IVA)*. Reykjavík, Iceland, pp. 269–281. DOI: 10.1007/978-3-642-23974-8_30 ⊸ pp. 159, 161.

Welbergen, Herwin van, Ramin Yaghoubzadeh and Stefan Kopp (2014). 'AsapRealizer 2.0: The next steps in fluent behavior realization for ECAs'. In: *Proceedings of the 14th International Conference on Intelligent Virtual Agents (IVA)*. Boston, MA, USA, pp. 449–462. DOI: 10.1007/978-3-319-09767-1_56 ⊸ pp. 155, 159, 161–162, 168.

Wiener, Norbert (1948/1961). *Cybernetics: or Control and Communication in the Animal and the Machine*. 2nd. Cambridge, MA, USA: MIT Press ⊸ pp. 42, 44.

Wienke, Johannes and Sebastian Wrede (2011). 'A middleware for collaborative research in experimental robotics'. In: *Proceedings of the IEEE/SICE International Symposium on System Integration*. Kyoto, Japan, pp. 1183–1190. DOI: 10.1109/SII.2011.6147617 ⊸ p. 282.

Wilson, Robert A. and Lucia Foglia (2015). 'Embodied Cognition'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2015 ⊸ p. 55.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes (2006). 'ELAN: A professional framework for multimodality research'. In: *Proceedings the 5th International Conference on Language Resources and Evaluation (LREC)*. Genova, Italy, pp. 1556–1559 ⊸ p. 199.

Włodarczak, Marcin, Hendrik Buschmeier, Zofia Malisz, Stefan Kopp and Petra Wagner (2012). 'Listener head gestures and verbal feedback expressions in a distraction task'. In: *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. Stevenson, WA, USA, pp. 93–96 ⊸ pp. 41, 173, 199.

Wöhler, Nils-Christian, Ulf Großekathöfer, Angelika Dierker, Marc Hanheide, Stefan Kopp and Thomas Hermann (2010). 'A calibration-free head gesture recognition system with online capability'. In: *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*. Istanbul, Turkey, pp. 3814–3817 ⊸ p. 173.

Wrede, Britta, Stefan Kopp, Katharina Rohlfing, Manja Lohse and Claudia Muhl (2010). 'Appropriate feedback in asymmetric interactions'. In: *Journal of Pragmatics* 41, pp. 2369–2384. DOI: 10.1016/j.pragma.2010.01.003 ⊸ p. 6.

Xiong, Wayne, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu et al. (2017). *Achieving human parity in conversational speech recognition*. Tech. rep. MSR-TR-2016-71 (revised). Redmond, WA, USA: Microsoft Research. arXiv: 1610.05256v2 [cs.CL] ⊸ p. 1.

XTAG Research Group (2001). *A Lexicalized Tree Adjoining Grammar for English*. Tech. rep. IRCS-01-03. Philadelphia, PA, USA: Institute for Research in Cognitive Science, University of Pennsylvania ⊸ p. 131.

Xudong, Deng (2009). 'Listener Response'. In: *The Pragmatics of Interaction*. Ed. by Sigurd D'hondt, Jan-Ola Östman and Jef Verschueren. Amsterdam, The Netherlands: John Benjamins, pp. 104–124. DOI: 10.1075/hoph.4.07xud ⊸ pp. 46, 48.

Yaghoubzadeh, Ramin, Hendrik Buschmeier and Stefan Kopp (2015). 'Socially cooperative behavior for artificial companions for elderly and cognitively impaired people'. In: *Proceedings of the 1st International Symposium on Companion-Technology*. Ulm, Germany, pp. 15–19 ⊸ pp. 173, 250.

Yaghoubzadeh, Ramin, Karola Pitsch and Stefan Kopp (2015). 'Adaptive grounding and dialogue management for autonomous conversational assistants for elderly users'. In: *Proceedings of the 15th International Conference on Intelligent Virtual Agents (IVA)*. Delft, The Netherlands, pp. 28–38. DOI: 10.1007/978-3-319-21996-7_3 ⊸ p. 39.

Yngve, Victor H. (1970). 'On getting a word in edgewise'. In: *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*. Ed. by Mary Ann Campbell, James Lindholm, Alice Davison, William Fisher, Louanna Furbee, Julie Lovins, Edward Maxwell et al. Chicago, IL, USA: Chicago Linguistic Society, pp. 567–577 ⊸ pp. 6, 48.

Zwaan, Rolf A. and Gabriel A. Radvansky (1998). 'Situation models in language comprehension and memory'. In: *Psychological Bulletin* 123, pp. 162–185. DOI: 10.1037/0033-2909.123.2.162 ⊸ p. 31.

Zyl, Marianne van and Johan J. Hanekom (2012). 'When "okay" is not okay: Acoustic characteristics of single-word prosody conveying reluctance'. In: *Journal of the Acoustical Society of America* 133, EL13–EL19. DOI: 10.1121/1.4769399 ⊸ p. 53.

ACCOMPANYING RESOURCES

Example Parametrisation of $\mathcal{BN}_{\mathcal{ALS}}$ A machine readable specification of the example model $\mathcal{BN}_{\mathcal{ALS}}$ (eqs. [5.8] and [5.9] in Bayesian network interchange format [XBIF, Cozman et al. 1998]), and Python source code to compute the (posterior) marginal probability distributions visualised in fig. 5.4.

DOI: 10.6084/m9.figshare.3827277 ⊸ p. 94.

Example Parametrisation of $\mathcal{BN}_{\mathcal{ALS}'}$ Bayesian network model parameters for $\mathcal{BN}_{\mathcal{ALS}'}$ in XBIF-format, and Python source code to generate the local probabilistic models from implicit representation.

DOI: 10.6084/m9.figshare.3851475 ⊸ p. 100.

Example Parametrisation of $\mathcal{BN}_{\mathcal{ALS}''}$ Bayesian network model parameters for $\mathcal{BN}_{\mathcal{ALS}''}$ in XBIF-format, and Python source code to generate the local probabilistic models from implicit representation.

DOI: 10.6084/m9.figshare.3971712 ⊸ p. 103.

Example Parametrisation of $\mathcal{BN}_{\mathcal{ALS}'''}$ Bayesian network model parameters for $\mathcal{BN}_{\mathcal{ALS}'''}$ in XBIF-format, and Python source code to generate the local probabilistic models from implicit representation.

DOI: 10.6084/m9.figshare.4980743 ⊸ p. 107.

Example Parametrisation of $\mathcal{DBN}_{\mathcal{ALS}}$ Dynamic Bayesian network model parameters for the two-time-slice networks $\mathcal{DBN}_{\mathcal{ALS}}$ in XBIF-format, Python source code to generate the local probabilistic models from implicit representation, as well as Jupyter Notebooks to query the model (Python) and generate plot fig. 5.13 (R). This model is also used in the implemented attentive speaker agent described in chapter 8 and evaluated in chapter 9.

DOI: 10.6084/m9.figshare.4981823 ⊸ pp. 114, 172.

Example CPT-Generation for Pr(*U* | *P*, *FB*) Python source code of implicit representation for the local probabilistic model $\Pr(U \mid P, FB)$ developed as part of the illustrative example (section A.3) of the CPT generation algorithm (section A.2). Also contains the Bayesian network model parameters (in XBIF-format) for a model $FB \rightarrow (P \rightarrow U)$, see fig. A.2.

DOI: 10.6084/m9.figshare.3838047 ⊸ p. 256.

Example for assessing information needs The underlying Bayesian network model as well as methods for assessing the criteria for information need used in the worked example that is discussed in section 7.3, (see fig. 7.1).

DOI: 10.6084/m9.figshare.4725538 ⊸ p. 147.

Supplementary material for chapter 9, 'Evaluation of the Attentive Speaker Agent' Evaluation data and analysis source code for reproducing the results of the evaluation study described in chapter 9.

DOI: 10.4119/unibi/2918228 ⊸ p. 199.

IPAACA — Incremental Processing Architecture for Artificial Conversational Agents IPAACA is a middleware for incremental inter- and intra-process communication that embodies the ideas of Schlangen and Skantze's (2011) 'general abstract model of incremental dialogue processing' and is implemented as a layer above the Robotics Service Bus (RSB, Wienke and Wrede 2011). IPAACA is implemented in Python, C++, and Java, can be used on Linux, macOS, and Windows, and is freely available under the GNU Lesser General Public License (version 3).

IPAACA was designed and developed by Ramin Yaghoubzadeh and Hendrik Buschmeier with contributions from Herwin van Welbergen, Sebastian Kahl, and others. In Schlangen et al. (2010, § 3), we briefly describe a preliminary version of IPAACA.

https://purl.org/scs/IPAACA ⊸ p. 157.

PRIMO — Probabilistic Inference Modules A Python package for probabilistic inference in Bayesian Networks.

The first version of the package, `primo`, was developed by Manuel Baum, Dennis John, Lukas Kettenbach and Max Koch (with contributions from Hendrik Buschmeier) as part of the seminar 'Probabilistic Reasoning in Practice' taught by Hendrik Buschmeier and Stefan Kopp (Bielefeld University, winter term 2012/13). An updated version, `primo2`, is being written by Jan Pöppel and Hendrik Buschmeier. `primo` and `primo2` are freely available under the GNU Lesser General Public License (version 3).

https://purl.org/scs/PRIMO (commit c3d90f1) ⊸ pp. 95, 172.

COPYRIGHT PERMISSIONS

Figure 3.1 (p. 42) Adapted from fig. 4, p. 6 of: Routledge, Robert (1883). *Discoveries and Inventions of the Nineteenth Century*. 10th. London, UK: George Routledge and Sons. — Public domain.

AFFIDAVIT

Hiermit erkläre ich, dass ich diese Dissertation konform zu § 8 Abs. 1 lit g der Rahmenpromotionsordnung der Universität Bielefeld vom 15. Juni 2010[*] angefertigt habe, d. h.

- mir ist die geltende Promotionsordnung der Technischen Fakultät der Universität Bielefeld vom 1. März 2011[†] bekannt;
- ich habe die Dissertation selbst angefertigt, keine Textabschnitte von Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel und Quellen in meiner Arbeit angegeben;
- Dritte haben weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Vermittlungstätigkeiten oder für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;
- diese Dissertation wurde noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht; und
- die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung wurde von mir bei keiner anderen Hochschule als Dissertation eingereicht.

Bielefeld, 9. Mai 2017

Hendrik Buschmeier

[*] In: *Verkündungsblatt — Amtliche Bekanntmachungen.* Ed. by Rektorat der Universität Bielefeld. Vol. 39. Bielefeld, Germany: Universität Bielefeld, pp. 98–105

[†] Idem. Vol. 40, pp. 56–59