A Dynamic Minimal Model of the Listener for Feedback-based Dialogue Coordination

Hendrik Buschmeier and **Stefan Kopp**
Social Cognitive Systems Group — CITEC and Faculty of Technology
Bielefeld University, Bielefeld, Germany
{hbuschme, skopp}@uni-bielefeld.de

Abstract

Although the notion of grounding in dialogue is widely acknowledged, the exact nature of the representations of common ground and its specific role in language processing are topics of ongoing debate. Proposals range from rich, explicit representations of common ground in the minds of speakers (Clark, 1996) to implicit representations, or even none at all (Pickering and Garrod, 2004). We argue that a minimal model of mentalising that tracks the interlocutor's state in terms of general states of perception, understanding, acceptance and agreement, and is continuously updated based on communicative listener feedback, is a viable and practical concept for the purpose of building conversational agents. We present such a model based on a dynamic Bayesian network that takes listener feedback and dialogue context into account, and whose temporal dynamics are modelled with respect to discourse structure. The potential benefit of this approach is discussed with two applications: generation of feedback elicitation cues, and anticipatory adaptation.

1 Introduction

Communicative feedback (*mhm, okay,* nodding, and so on) is a dialogue coordination device used by listeners to express their mental state of listening — e.g., I understand what you say (Allwood et al., 1992) — and by speakers to hypothesise about this mental state and adapt their language production accordingly — e.g., she understood it, I can provide new information (Clark and Krych, 2004). 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 of 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 dialogue in example (1):

(1) KDS-1, U01 (9:46–9:58)[1]

```
1 S1: genau
2     allerdings ist Badminton da=
      =wieder verschoben
3     [weiß nicht] ob das jetzt=
  U1: [mhm       ]
  S1  =dauerhaft ist (.)
4 S1: [aber die zwei] Wochen=
  U1: [okay         ]
5 S1: =hab ich's jetzt so drin
  U1:                        ja
6 S1: das is wieder von=
7     =ehm acht bis zweiundzwanzig
      U[hr]
  U1:  [ok]ay (0.34)
8     ja,
9     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 (1.1) to (1.7), U1 provides four instances of communicative feedback. Firstly, she signals understanding with *mhm*, simultaneously producing a single head nod and looking at S1 (1.3). After that, she signals acceptance of the speaker's ignorance concerning the permanency of the time change with an *okay* that is accompanied by

[1] Excerpt from the calendar assistant domain corpus KDS-1 (http://purl.org/scs/KDS-1). Overlapping talk is marked with aligned square brackets. The transcription follows the GAT 2 system (Couper-Kuhlen et al., 2011).

a head nod (1.4). Thirdly, she signals understanding, producing a short and prosodically flat *ja*, German for 'yeah', (1.5). And finally, with S1 gazing at her, she signals understanding of the new time with an *okay* and a head nod (1.7). After a pause, U1 then takes the turn and continues.

In previous work (Buschmeier and Kopp, 2012), we proposed a Bayesian network approach in which single instances of communicative feedback are interpreted in terms of a few general attributes (contact, perception, understanding, acceptance, and agreement; Allwood et al., 1992). However, when multiple feedback instances occur in sequence, as in the dialogue in example (1), the question arises how their interpretations affect each other, and how they relate to what has been and is being said. In keeping with this 'minimal mentalising' approach to the listener's cognitive state, we take the Bayesian network model and make it dynamic. The dynamics is added by extending the model with a temporal dimension that accounts for the incremental and dynamic nature of dialogue. Thus, in this work, we propose a 'dynamic minimal model' of mentalising which 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.

2 Common ground and feedback

Participating in dialogue involves more than utterance planning, formulation, speaking, listening and understanding. One central task for interlocutors is to track the 'dialogue information state,' a rich representation of the dialogue context. The representation includes which information is grounded and which is still pending to be grounded; which knowledge is private and which is believed to be shared; who said what, how and when; how these utterances are related to each other; which objects have been introduced and are accessible for anaphoric reference; what is the current question under discussion; who is having the turn; and potentially much more (Clark, 1996; Larsson and Traum, 2000; Asher and Lascarides, 2003; Ginzburg, 2012).

In general, maintaining (i.e., representing and constantly updating) an information state is thought to be crucial for being able to successfully participate in dialogue. The necessity of some parts, such as a representation of accessible referents, is agreed upon among researchers. Without this information being maintained, typical dialogues would simply not be possible. Concerning the representation of common ground, however, researchers do not agree on how deep and rich it needs to be and how exactly it is used in language production.

On the one hand, Clark (1996) argues that interlocutors maintain a detailed model of common ground, even to the extent that mutual knowledge (approximated with various heuristics) is necessary to explain certain phenomena in language use (Clark and Marshall, 1981). Pickering and Garrod (2004), on the other hand, believe that dialogue does not involve heavy inference on common ground at all, instead they claim that primed and activated linguistic representations provide sufficient information in themselves.

Use of common ground in language production in dialogue is also a topic of ongoing debate. Clark (1996) and Brennan and Clark (1996) argue that common ground is critical in collaborative discourse. Utterances are designed in such a way that common ground as well as shared knowledge are taken into account. Since this might be cognitively too demanding, Galati and Brennan (2010) propose a lightweight 'one-bit' partner model (e.g., whether the addressee has heard something before or not) that can be used instead of information about full common ground and shared knowledge when producing an utterance. Horton and Keysar (1996) go even further and present evidence that language production is, at its basis, an egocentric process — interlocutors do not take common ground into account when initially planning an utterance unless they identify a possible problem while monitoring utterance execution. Finally, Pickering and Garrod (2004) claim that the only factors guiding language production are priming, activation, and, if necessary, interactive repair.

Speakers infer groundedness and common ground based upon 'evidence of understanding' of the interlocutors (Clark, 1996). One way for listeners to show such evidence is by providing communicative listener feedback as, e.g., short verbal/vocal expressions such as *mhm, okay,* and *oh;* head-gestures such as nods or shakes; facial expressions such as surprise, or frowning; as well as various gaze behaviours. Listener feedback is a particularly interesting kind of evidence of understanding for multiple reasons:

1. When providing feedback, listeners do not need to have or to take the turn, making it

very *fast*. Since it is not constrained by turn-taking, feedback can be given as soon as the need arises, enabling speakers to quickly adapt the ongoing utterance based on this information.

2. At the same time, feedback is *unobtrusive* and does not interrupt speakers during their utterance. It happens in the 'back channel' of communication (Yngve, 1970). Feedback also relies heavily on non-verbal modalities (head, face, gaze) that do not interfere with the speakers' linguistic processing. Verbal/vocal feedback expressions — that have the potential to interfere — are often non-lexical (Ward, 2006), usually short, and even prosodically hidden in the speech context provided by the speaker (Heldner et al., 2010).

3. Despite their shortness, feedback signals are very *expressive*. They are rich in their form (Ward, 2006) — enabling a fine-grained expression of subtle differences in meaning —, multi-functional, and interact heavily with their dialogue context (Allwood et al., 1992). Feedback is only partially conventionalised, relying on iconic properties instead.

4. Finally, communicative feedback is *reflective* of the listener's cognitive 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 and what their attitude is towards it (Allwood et al., 1992). 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

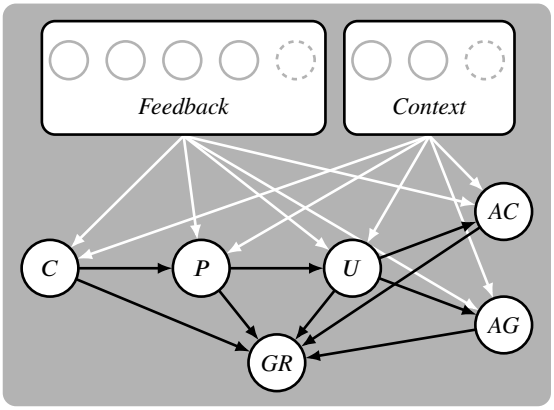

Figure 1: The Bayesian network model of the 'attributed listener state' (ALS; Buschmeier and Kopp, 2012). The random variables C, P, U, AC, and AG model a speaker's degree of belief that a listener is in contact, whether he or she perceives, understands, accepts, and agrees to what is communicated. A speaker's belief in groundedness is informed by all five of these variables.

feedback facilitates a form of mentalising about the cognitive state of the dialogue partner that goes beyond what is usually considered groundedness.

In previous work (Buschmeier and Kopp, 2012), we modelled this capability of speakers as, what we called, an 'attributed listener state' (ALS, cf. Figure 1). The ALS is a Bayesian network-based representation of a speaker's belief of what her listener's cognitive state is in terms of the basic communicative functions underlying feedback in dialogue. Each of the random variables (i.e., the nodes of the network) represent one 'dimension' of the multidimensional cognitive state of the listener: C (is the listener believed to be in contact), P (is the listener believed to perceive), U (is the listener believed to understand), AC (is the listener believed to accept), and AG (is the listener believed to agree). The network captures the dependencies between these variables and models their interactions, e.g., their hierarchical properties (Allwood et al., 1992; Clark, 1996). A belief about the groundedness of the conveyed proposition is formed based on the five ALS-variables, each having a different strength of influence.

The variables consist of the individual elements *low*, *medium*, and *high*, denoting whether the speaker believes the dimension of a listener's cognitive state to be low, medium, or high, respectively. An individual element's probability, e.g., $P(U = \mathit{low}) = 0.6$, is thus interpreted as the speaker's de-

gree of belief in this dimension of the listener's cognitive state to have the specific characteristic, i.e., 'with a probability of 0.6 the listener's understanding is believed to be low'. The probability distribution over all elements of a variable represents the speaker's belief state over the variable.

Buschmeier and Kopp's (2012) model can be considered a *minimal* form of mentalising based on listener feedback. It shares some desirable properties with the lightweight 'one-bit' partner model of Galati and Brennan (2010) — efficient processing in contrast to models of full common ground, a simple variable-based representation — while extending it. In particular, the model is in accordance with gradient representations of common ground (Brown-Schmidt, 2012), as it defines groundedness of a segment on an ordinal, non-binary scale (low < medium < high). Due to its probabilistic nature, each element is associated with a degree of belief from 0 (not believed) to 1 (believed). This information can be used to interactively adapt language production to a listener's need, e.g., by repeating/leaving out parts of an utterance, by giving subsequent parts a lower/higher information density, or by making information pragmatically explicit/implicit (Buschmeier et al., 2012).

3 A dynamic model of the listener

What is missing from the model proposed by Buschmeier and Kopp (2012), however, is a notion of the temporal dynamics that would make the evolution of the ALS coherent and continuous, and enable the model to deal with sequences of feedback such as in the example dialogue (1).

We regard an unfolding dialogue as a sequence of segments $[s_{t_0}, s_{t_1}, \dots s_{t_n}]$, each consisting of a dialogue move of the speaker (Poesio and Traum, 1997), together with any feedback responses of the listener. The static model of Figure 1 (Buschmeier and Kopp, 2012) treats each of these segments s_{t_i} independently and thus only reasons about the listener's cognitive state during one single segment. When doing the listener state attribution 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 cognitive 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 part of a dialogue — in one large network that consists of connected sub-networks ALS_{t_i} — each corresponding to the network in Figure 1 — one for each segment s_{t_i}. The variables in the sub-networks would be uniquely named, and the networks evidence variables would be instantiated from the listener's feedback behaviour as well as the dialogue context of segment s_{t_i}. 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 $s_{i_{t+1}}$, a new sub-network $ALS_{t_{i+1}}$ 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 (Barber, 2012). 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 $X_{t_{i+1}}$ of a sub-network $ALS_{t_{i+1}}$ are only dependent on variables X_{t_i} of the sub-network ALS_{t_i} 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 dynamic Bayesian network approach consists of a maximum of two sub-networks ('time-slices') at any point of time. In such a *two time-slice Bayesian network* (cf. Figure 2), one time slice ALS_{t_i} represents the current dialogue segment s_{t_i} the other time slice the next segment s_{i+1}. As in the growing network approach, temporal influences among dialogue units are modelled by connecting some of the variables between the time-slices. Connection further back are, however, not possible.

In such a network, evolution over time is done by unrolling the network. Bayesian network inference is carried out on time-slice ALS_{t_i} and the resulting marginal posterior probabilities of those variables X_{t_i} that have a connection with variables $X_{t_{i+1}}$ 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_i} that are used as evidence variables to variables $X_{t_{i+1}}$ in the subsequent time slice.

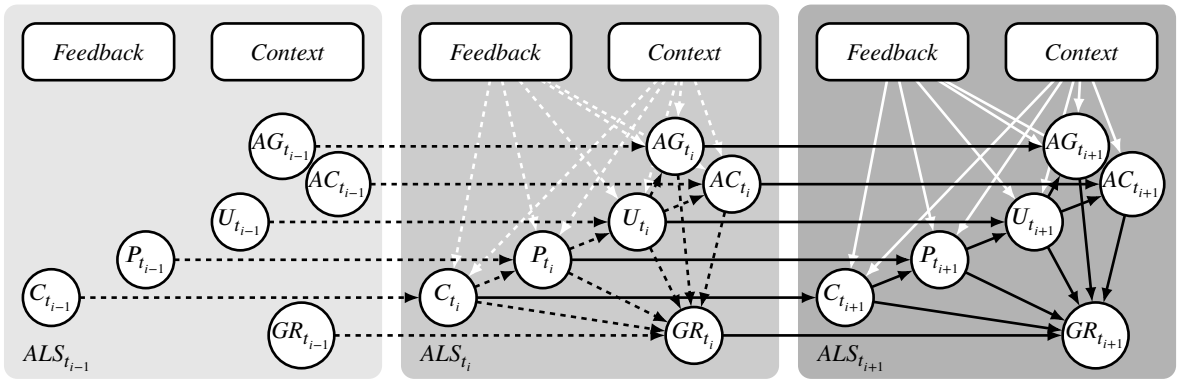

Figure 2: A dynamic two time-slice Bayesian network model unrolling over three steps in time, each corresponding to one dialogue segment. Dashed arrows are disregarded during inference in subsequent time-slices, i.e., variables from time slice $ALS_{t_{i-1}}$ and evidence variable in time slice ALS_{t_i} have no influence on variables in time slice $ALS_{t_{i+1}}$. Posterior distributions of attributed listener state variables in time slice ALS_{t_i} are taken as prior distributions at time t_{i+1} and influence the variables they are connected to in time slice $ALS_{t_{i+1}}$.

Due to the first order Markov assumption, previous time slices ALS_{t_0} to $ALS_{t_{i-1}}$ are not taken into account any more and all connections to them, as well as to all variables X_{t_i} that have no influence into the future, and can be disregarded (dashed lines in Figure 2). The complete history is thus implicitly contained, in accumulated form, in time slice ALS_{t_i}.

In our model, the ALS variables C, P, U, AC, AG, and the groundedness variable GR, are the ones that carry over information between time slices (Figure 2), e.g., understanding at time t_i influences understanding at time t_{i+1} (consequently, variable $U_{t_{i+1}}$ is not only influenced by $P_{t_{i+1}}$, $Feedback_{t_{i+1}}$, and $Context_{t_{i+1}}$, but additionally by U_{t_i}). This is based on the assumption that listener state evolution — and attribution — is usually a gradual process. Indeed, abrupt changes of listener state are often marked by special feedback tokens such as for example *oh* or, in German, *ach* and *ach so*.

Figure 3 simulates the dialogue from example (1) in two contrasting conditions. Once without temporal influences between dialogue segments s_{t_i} and $s_{t_{i+1}}$, based on Buschmeier and Kopp's (2012) static model (Figure 3a); and once with modelled temporal dynamics based on the dynamic model presented above (Figure 3b). Each graph shows how speaker S1's belief state of a specific variable — i.e., the probabilities for each of its elements — changes over time (magenta coloured lines show $P(X = low)$, yellow lines $P(X = medium)$ and cyan coloured lines $P(X = high)$ for $X \in \{P, U, AC, GR\}$). Nine time-steps are shown, each corresponding to one dialogue segment.

In Figure 3a, 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 t_0 to t_2). When U1 provides feedback (from t_3 to t_5 and at t_7), 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 t_6).

In contrast to this, the dynamic model in Figure 3b, leads to a gradually evolving attributed listener state. In the beginning, when no feedback is provided by U1 (from t_0 to t_2), the belief state shifts towards *low* perception, understanding, acceptance, and groundedness. This changes, cautiously, as soon as feedback is provided at t_3 and grows towards *medium* to *high* with each subsequent feedback signal provided by U1 (at $t_4, t_5,$ and t_7). Notably, at t_6, the belief state does not jump to the initial state, but degrades only slightly while U1 does not provide feedback.

4 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 ALS_{t_i} influences its successor state $ALS_{t_{i+1}}$. For the example, in Figure 3b, the transitions were assumed to be fixed, that is, the influence $P(X_{t_{i+1}}|X_{t_i})$ of each of the vari-

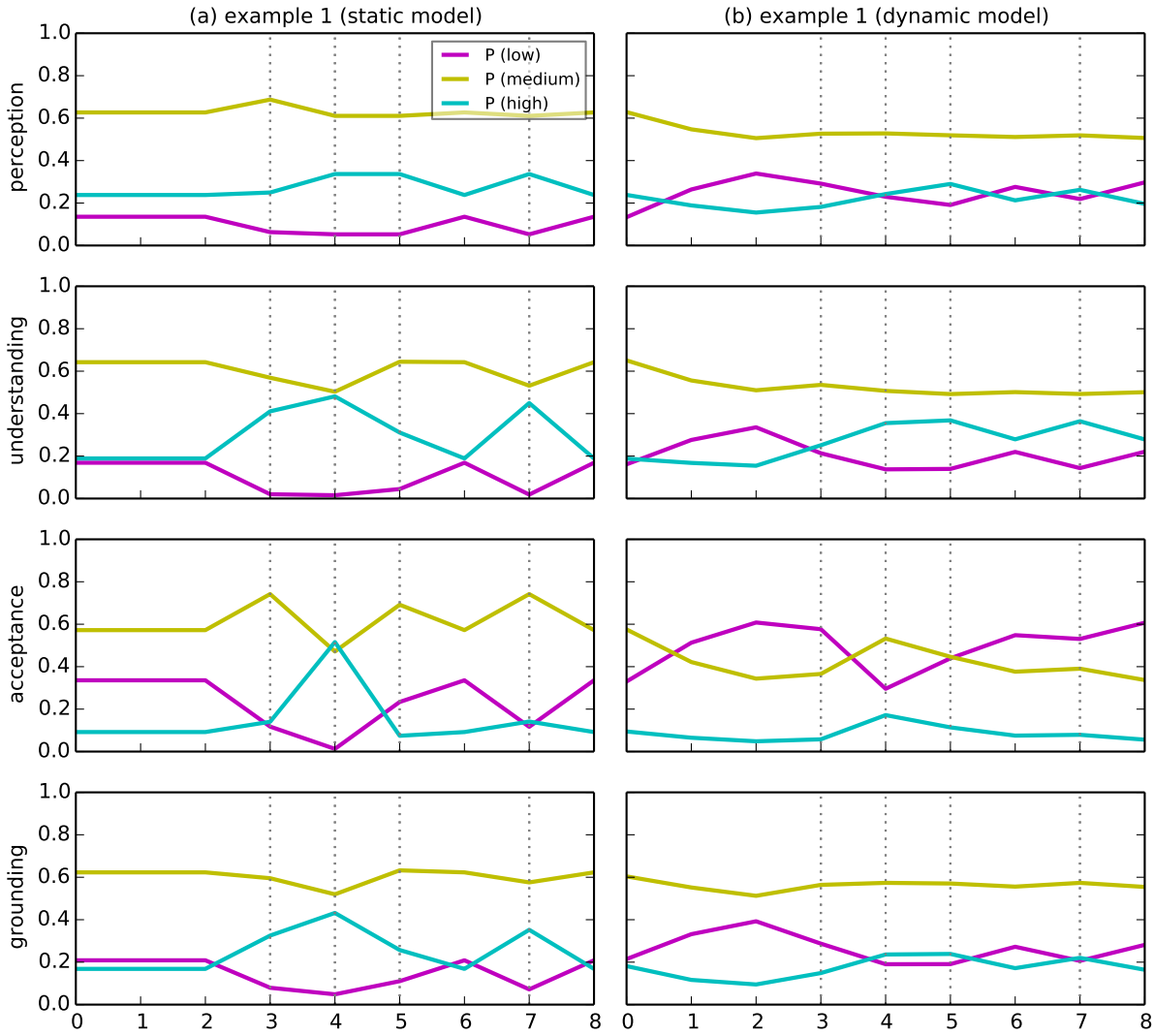

Figure 3: Simulated belief state evolution for example dialogue (1). The graphs show speaker S1's graded belief for the attributed listener state variables P, U, AC, and GR given the feedback provided by listener U1 (dashed vertical lines indicate the exact points in time when feedback occurred). Two conditions are contrasted: (a) without temporal influences between dialogue segments, simulated with Buschmeier and Kopp's (2012) static model; and (b) with temporal influences between dialogue segments, simulated with the two time-slice dynamic Bayesian network model (Figure 2).

ables $X_{t_i} \in \{C_{t_i}, \ldots, GR_{t_i}\}$ on its successor $X_{t_{i+1}} \in \{C_{t_{i+1}}, \ldots, GR_{t_{i+1}}\}$ was fixed for each point in time $t_i \in [t_0, \ldots, t_8]$ (influences among variables varied, i.e., $P(X_{t_{i+1}}|X_{t_i}) \neq P(Y_{t_{i+1}}|Y_{t_i})$ for $X \neq Y$.

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_{i+1} either the topic changes, or the narration simply continues. Intuitively, the influence of the speaker's attributed listener state ALS_{t_i} on the attributed listener state $ALS_{t_{i+1}}$ 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 ALS_{t_i} has much to contribute — i.e., is a good predictor — to understanding and acceptance in $ALS_{t_{i+1}}$ (arguably this also depends on the relatedness of the two topics). In contrast to this, understanding and acceptance as estimated in ALS_{t_i} seems to be very relevant for $ALS_{t_{i+1}}$ in the case where the narration simply continues.

The 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. This is in line with the proposal of Stone and Lascarides (2010), who propose a similar influence of discourse relations on grounding, also within an — albeit so far purely theoretical — dynamic Bayesian network model.

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_i} has on a variable $X_{t_{i+1}}$ 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_i} on $X_{t_{i+1}}$ being the the same as the influence that all non-temporal variables have on $X_{t_{i+1}}$. 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 $0.5 < w \leq 1$. Concrete weights for individual discourse relations need to be determined empirically.

In practical terms, this approach involves (1) having different dynamic Bayesian network models for each of the discourse relation types, and (2) switching the networks — carrying over the variable assignments and distributions — when proceeding form dialogue segment to dialogue segment.

5 Example applications

In addition to being able to better track the attributed listener state and groundedness, the dynamic minimal model of the listener enables novel applications in artificial conversational agents that were not possible with Buschmeier and Kopp's (2012) static model. Two of these will be sketched in the following.

5.1 Eliciting listener feedback

Listeners do not only produce communicative feedback when they feel the need to inform speakers about their cognitive state of dialogue processing, e.g., if they want to give evidence of understanding or if they do not understand what is said. Often feedback is provided cooperatively in response to 'feedback elicitation cues' of a speaker (Ward and Tsukahara, 2000; Gravano and Hirschberg, 2011). Speakers produce these cues since they have an active interest in how their ongoing utterance is perceived, understood, etc., by their interlocutors, and because it helps them in language production and story telling (Bavelas et al., 2000). This is especially the case in situations where they are uncertain about the listener's cognitive state, even to the extent that they cannot make well-grounded choices in language production. In cases of such an 'information need' (Buschmeier and Kopp, 2014b), elicitation of feedback from the listener is a viable strategy to ensure and achieve an effective dialogue. We propose that the following three criteria — in terms of our model — are indicative of a speaker's information needs (Buschmeier and Kopp, 2014b):

1. The entropy of a variable of interest rises (i.e., the probability distribution across the elements of a variables become more uniform, e.g., when $P(U = low) = 0.33, P(U = medium) = 0.33, P(U = high) = 0.33$) so that the belief state becomes less and less informative.

2. A variable of interest remains static for an extended period of time (e.g., when the listener does not provide feedback).

3. The distance (measured with the Kullback-Leibler divergence) between the probability

distributions of the current state of a variable and a desirable 'reference state' — such as, for example, a state that represents very good understanding — grows beyond a certain acceptable value.

These criteria could in principle be used with the static model of attributed listener state. However, the continuous temporal progression of the belief state makes it possible to identify reliable trends which enable informational needs to be detected early on and with high precision.

5.2 Anticipatory adaptation

A second ability that also builds on the mechanism of identifying trends in the development of the attributed listener state is to adapt language production to anticipate needs of the listener, a mechanism that human speakers use all the time. For this, an artificial agent could simulate the most likely evolution of the dynamic ALS and use this projected next listener state in order to make adaptations in natural language generation that serve as a pre-emptive countermeasure against an expected undesirable cognitive state of the user.

As an example, consider a situation where the agent believes that with every discourse segment the user understood less and less. A simulation that is run for the upcoming segment results in a belief state which shows that this trend is likely to continue. Expecting this state in the dynamic model, now allows the agent to change its original plan — say, to present an additional detail — and instead repeat what has already been said in a different way thus giving the subject matter a different perspective which might help the user understand.

6 Conclusion

In this paper we propose a dynamic Bayesian network-based model for minimal mentalising that tracks the interlocutors' cognitive state with respect to their willingness and ability to perceive, understand, accept, and agree by means of their communicative feedback behaviour. We argued that feedback is a particularly suitable way for listeners to provide evidence of understanding at almost any point in the dialogue, and for speakers to reason about the the listener's cognitive state, as well as to make statements about groundedness. The model can serve as a middle ground between theories that assume representations of full common ground (Clark, 1996) and theories that assume no common ground at all (Pickering and Garrod, 2004).

We extended a previous model of attributed listener state (Buschmeier and Kopp, 2012) with a temporal dimension, showed how the attributed listener state develops while a dialogue unfolds, and illustrated how its progression can be influenced by the structure of the discourse. Finally, we briefly described two relevant and novel applications of the presented model for artificial conversational agents that rely specifically on the model's temporal dynamics and its ability to continuously track the development of the attributed listener state in order to identify trends and project its future development.

Future work will involve an investigation of directionality of the influence of the discourse relations in the dynamic model. A result might be that the flow of information will be reversed given certain discourse relations so that recent evidence of understanding can influences variables in the previous time-slice. We will also implement the mechanisms for feedback cue elicitation and anticipatory adaptation sketched out as applications in an artificial conversational agent and evaluate them in interaction with human users.

A Supplementary material

A data publication containing the model parameters supplements this paper (Buschmeier and Kopp, 2014a). Additionally, the dynamic Bayesian network implementation is publicly available under the GPL 3 license at `http://purl.org/scs/PRIMO`.

Acknowledgements This research is supported by the German Research Foundation (DFG) at the Center of Excellence EXC 277 'Cognitive Interaction Technology' (CITEC).

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26. doi: 10.1093/jos/9.1.1

Nicolas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

David Barber. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge, UK.

Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Per-*

sonality and Social Psychology, 79:941–952. doi: 10.1037/0022-3514.79.6.941

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493. doi:10. 1037/0278-7393.22.6.1482

Sarah Brown-Schmidt. 2012. Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, 27:62–89. doi:10.1080/01690965.2010.543363

Hendrik Buschmeier and Stefan Kopp. 2012. 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*, pp. 12–20, Paris, France.

Hendrik Buschmeier and Stefan Kopp. 2014a. Dynamic Bayesian model of the listener. Data publication, Bielefeld University, Bielefeld, Germany. doi:10.4119/unibi/2687517

Hendrik Buschmeier 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*, pp. 71–80, Boston, MA, USA.

Hendrik Buschmeier, 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*, pp. 295–303, Seoul, South Korea.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50:62–81. doi:10.1016/j.jml.2003.08.004

Herbert H. Clark and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. In Aravind K. Joshi et al., (Eds.), *Elements of Discourse Understanding*, pp. 10–63. Cambridge University Press, Cambridge, UK.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK. doi:10.1017/CBO9780511620539

Elizabeth Couper-Kuhlen, Dagmar Barth-Weingarten, et al. 2011. A system for transcribing talk-in-interaction: GAT 2. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, 12:1–51.

Alexia Galati and Susan E. Brennan. 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62:35–51. doi:10.1016/j.jml.2009.09.002

Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford University Press, Oxford, UK.

Augustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25:601–634. doi:10.1016/j.csl.2010.10.003

Mattias Heldner, Jens Edlund, and Julia Hirschberg. 2010. Pitch similarity in the vicinity of backchannels. In *Proceedings of Interspeech 2010*, pp. 3054–3057, Makuhari, Japan.

William S. Horton and Boaz Keysar. 1996. When do speakers take into account common ground? *Cognition*, 59:91–117. doi:10.1016/0010-0277(96)81418-1

Staffan Larsson and David R. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6:323–340. doi:10.1017/S1351324900002539

Philippe Muller and Laurent Prévot. 2003. An empirical study of acknowledgement structures. In *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken, Germany.

Volha Petukhova and Harry Bunt. 2010. Introducing communicative function qualifiers. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pp. 123–131, Hong Kong, China.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226. doi:10.1017/S0140525X04000056

Massimo Poesio and David R. Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13:309–347. doi:10.1111/0824-7935.00042

Christian P. Robert. 1993. Prior feedback: A Bayesian approach to maximum likelihood estimation. *Computational Statistics*, 8:279–294.

Matthew Stone and Alex Lascarides. 2010. Coherence and rationality in grounding. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 51–58, Poznan, Poland.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 38:1177–1207. doi:10.1016/S0378-2166(99)00109-5

Nigel Ward. 2006. Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, 14:129–182. doi:10.1075/pc.14.1.08war

Victor H. Yngve. 1970. On getting a word in edgewise. In Mary Ann Campbell et al., (Eds.), *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–577. Chicago Linguistic Society, Chicago, IL, USA.