J Multimodal User Interfaces (2014) 8:61–73
DOI 10.1007/s12193-013-0147-7

ORIGINAL PAPER

A model for incremental grounding in spoken dialogue systems

Thomas Visser · David Traum · David DeVault · Rieks op den Akker

Received: 18 April 2013 / Accepted: 30 December 2013 / Published online: 1 March 2014

Abstract We present a computational model of incremental grounding, including state updates and action selection. The model is inspired by corpus-based examples of overlapping utterances of several sorts, including backchannels and completions. The model has also been partially implemented within a virtual human system that includes incremental understanding, and can be used to track grounding and provide overlapping verbal and non-verbal behaviors from a listener, before a speaker has completed her utterance.

Keywords Spoken dialogue systems · Incremental language processing · Grounding

1 Introduction

Effective and fluent conversation requires joint effort from both interlocutors [6], and in spoken human dialogue, this effort is often manifested in real time as speech is happening. While speaking, we monitor the listener's reaction to what we say, and as listeners, we give frequent feedback on what we perceive and understand. Such feedback often overlaps the speaker's ongoing utterance and can take the form of head nods, verbal backchannels, interruptions, and other overlapping responses.

These overlapping responses are important for efficient conversation, and emphasize the incremental nature of human–human communication [20,31]. For a spoken dialogue system to understand and generate such behaviors, it needs to process speech incrementally. This requires that the processing of user input and planning of system responses occurs frequently, not only while the user speaks, but also while the user listens. While traditional systems employ a rigid turn-taking model, in which overlapping speech is not supported, recent research has begun to develop some of these incremental processing and response capabilities in implemented systems (e.g., [3,8,17,18,25,27,28,40,41]).

This work has shown that incremental response capabilities can achieve positive effects on user interactions, including user preference over non-incremental systems and increases in perceived human-likeness and efficiency [29,30], and even increased fluency of user speech [12].

To date, however, implemented systems that model the process of *grounding* in dialogue [7], the process by which interlocutors work to add understood content to their common ground, have not closely linked such incremental response behaviors directly to the grounding model. The system presented in [30] is capable of incremental grounding behavior, but, as pointed out by [4], the domain lacks a notion of utterances and a meaning beyond the surface text. We believe that a grounding model should include the intention and conversational meaning of utterances. In this paper, we take up this project, and present an initial computational grounding model that can connect some of these incremental response behaviors to an incrementally evolving grounding state. We begin in Sect. 2 by looking at examples of incremental grounding behavior in spoken conversations between human interlocutors. In Sect. 3, we review

T. Visser · R. op den Akker
University of Twente, Enschede, The Netherlands
e-mail: thomas.visser@gmail.com

R. op den Akker
e-mail: h.j.a.opdenakker@utwente.nl

D. Traum (✉) · D. DeVault
USC Institute for Creative Technologies, Playa Vista, CA 90092, USA
e-mail: traum@ict.usc.edu

D. DeVault
e-mail: devault@ict.usc.edu

prior work on grounding models and overlapping dialogue behavior. In Sect. 4, we present our model of the incremental grounding process. Section 5 discusses its implementation within a working spoken dialogue system. We conclude, with a discussion of future work, in Sect. 6.

2 Incremental grounding behavior in human dialogue

We present examples of incremental grounding behavior from the the AMI Meeting Corpus [5]. The corpus consists of multi-modal recordings of the meetings of a four person product design team. Over the course of four meetings, they brainstorm, negotiate and decide on the design of a universal TV remote. A simple kind of incremental grounding is shown in (1) ("*" indicates an utterance that overlaps the previous utterance). Here B overlaps A's first utterance with evidence of understanding *okay*, while C gives evidence of understanding *Mm-hmm* of A's second utterance. The key point here is that B and C give evidence of understanding while A is still talking.

(1) A : Maybe even pre-programmed sound modes,
: like um
B*: Okay
A : the user could determine a series
: of sound modes.
C*: Mm-hmm

Dialogue excerpt (2) includes two types of incremental grounding behavior. In the middle of C's sentence, C appears to struggle with how to continue his utterance, uttering a verbal hesitation "um". A then utters "Normal coloured buttons" as a completion of C's partial utterance. The dialogue continues without correction by C, so it is reasonable to assume that this was indeed what C intended to communicate (or was close enough). Meanwhile, D gives a simultaneous backchannel acknowledgement of C's utterance, similar to those in excerpt (1).

(2) C : We could just go with um
D*: Yeah
A : Normal coloured buttons
B : Well do you want colour differentiation here?
C : …

Such examples seem to indicate that participants have an ability to predict the meaning (or perhaps even the surface form). Further, such examples raise a question about the grounding status of the partial utterance and predicted completion. We call the content of the partial utterance *explicit* and the content of the utterance completion *predicted*, under the assumption that the completion is what C intended to say.[1] For grounding purposes, note that A's completion not only demonstrates his understanding of the explicit content of C's utterance, but also of the predicted content, and no further explicit evidence of understanding seems to be required by C. A theory of incremental grounding should make claims about the grounding status of such explicit and predicted content as a dialogue progresses.

Attempted utterance completions do not always exactly match a speaker's intended content or surface form, as in dialogue excerpt (3).

(3) B : That would probably not be in keeping
: with the um the
C*: *laugh* Technology
B : fashion statement and such, yeah.
C*: Yeah.

In this dialogue, B and C are reflecting on the features and design of the remote control they created. When B shows hesitation ("… with the um"), C decides to help and offers "Technology" as a completion of B's utterance. B however continues his utterance by saying "fashion statement and such", revealing perhaps more precisely what he intended to say. C then issues an overlapping acknowledgment of B's continuation with "fashion statement", by saying "Yeah".

In this example, C's predicted content "Technology" apparently does not exactly match B's original intention. However, it does provide some evidence of understanding of the explicit content of B's partial utterance. A theory of incremental grounding should also make claims about the status of explicit and predicted content in such cases where a completion is corrected, and how they are similar and different from the previous one, in which no explicit grounding action was needed. Such a theory should also describe how the grounding state is updated by an overlapping acknowledgment such as C's utterance of "Yeah", here, acknowledging B's continuation with "fashion statement…".

There are also some intermediate cases, such as (4), where B abandons his own continuation and accepts D's completion instead. This shows that grounding is not just a matter of getting the speaker's original intention across to addressee, but that successful contributions are a collaboration, and sometimes multiple "authors" contribute content, with an initiating speaker adopting material from another, and eventually grounding it.

(4) B : However I've got a couple of worries about that
The power required , um and the ability to
D*: the cost
B : cost It seems like for an embedded system …

[1] It is sometimes useful to distinguish further between the explicit or predicted surface form, as opposed to the explicit or predicted meaning.

3 Background work

Clark and Schaefer [7] distinguish two phases in the grounding of a contribution: the *presentation* phase and the *acceptance* phase. In the presentation phase, the speaker presents a piece of new content for the listener to consider. The speaker assumes that if the listener provides evidence of at least a certain strength, he can believe that the listener understands what he meant. In the acceptance phase, the listener accepts the new content by giving evidence of understanding, assuming that this evidence will make the speaker believe that he understands. The acceptance itself is also considered a contribution, which in turn needs to be accepted.

The evidence of understanding that a listener can give to show his acceptance of the contribution can, according to Clark and Schaefer, be one of the five types listed in Table 1.

Some of these evidence types can occur simultaneously with the original presentation, as seen in the previous section. Clark mentions a few forms of completions. A variant of background acknowledgment is the *unison completion* in which the listener and speaker complete the speaker's sentence in unison (see the example given in [6], p. 231). In *collaborative completions* the listener shows acceptance of the part so far by completing the speaker's utterance. He thereby presents a new part. These type of completions are often accepted or rejected explicitly. Clark sees two contributions here. "The one contains the other as its part". In a third form, *truncations*, the listener interrupts the speaker by giving the answer to a question only expressed half way. The primary speaker accepts the answer by a short acknowledgement ([6], p. 238–239).

3.1 The Traum '94 grounding model

Traum's computational model of grounding [36], defines seven *grounding acts*: initiate, continue, acknowledge (abbreviated *ack*), request repair (abbreviated *reqrepair*), repair, request acknowledgement (abbreviated *reqack*), and cancel. Every behavior, either verbal or non-verbal, can convey one or more grounding acts relating to one or more Common Ground Units (CGU). A CGU is similar to Clark and Schaefer's 'contribution' [7], but it is more closely related to surface structure of the dialogue and therefore more suitable in on-line systems [19].

Traum's theory uses a finite state model that assigns each CGU to one of seven states at each point of the dialogue. The processing of an utterance's impact on common ground consists of two steps. The system first has to determine the grounding acts that are being conveyed by the utterance and to which CGUs they apply. Then, the corresponding CGUs are updated. Grounding acts may change the information state in two ways: changing the grounding state of a CGU, or changing the content of CGUs (or both).

Traum's model has been used with several different models for CGU and utterance content. In each case, content consists not only of the words that are spoken, but their semantics (representation of the meaning of the words and structures) and pragmatics (references to domain objects, core speech acts, and speech act effects, such as beliefs and obligations). The Trains-95 system used Event-based Temporal Logic (EBTL) [39] to represent utterance contents. Later work represented information state content in the form of discourse representation structures [22], and records in the EDIS system [16]. More recent work has used a representation of semantics frames, as described in [32].

Table 2 contains the CGU transition diagram that indicates how the grounding state is updated when grounding acts are performed by initiator (abbreviated with a superscript "i"), the partner who performs the initiate act, or a responder (abbreviated with a superscript "r"), another dialogue participant. In general, a CGU is placed into the starting state upon being initiated by a speaker; eventually (if all goes well), the CGU moves into a final state signifying that the CGU's content has entered the common ground. In the meantime, various patterns of continue, repair, acknowledgment, and other grounding acts may occur. Throughout this process, speaker and addressee information is used to determine which role, either initiator or responder, the participants have with respect to each CGU. A CGU is said to be in the common ground when it reaches state F.

In the Traum '94 grounding model, some grounding acts (acknowledgment, request acknowledgment and cancel) will only affect the grounding state, while others (initiate, continue and repair) will change its content. We will call the latter category *authorial grounding acts*, as they make the uttering party co-author of the CGU. By becoming an author, the burden of providing evidence of understanding of the CGU content shifts to the other interlocutor. For example, after ini-

Table 1 Types of evidence of understanding, from Clark and Schaefer [7, p. 267]

1.	*Continued attention*. B shows that he is continuing to attend and therefore remains satisfied with A's presentation
2.	*Initiation of the relevant next contribution*. B starts in on the next contribution that would be relevant at a level as high as the current one
3.	*Acknowledgment*. B nods or says "uh huh", "yeah", or the like
4.	*Demonstration*. B demonstrates all or part of what he has understood A to mean
5.	*Display*. B displays verbatim all or parts of A's presentation

tiation, the responder must acknowledge for the CGU to be grounded, however, if the responder repairs a CGU, then it is the initiator who is required to provide evidence of understanding for that CGU to be grounded. If the initiator however decides, in his turn, to repair the CGU again, he becomes the most recent author and the burden of evidence shifts back to the responder. In Traum's original model, these notions are implicitly contained in the four in-progress grounding states (see Table 2). In state 1 and 2, the initiator is the most recent author and the burden of evidence lies with the responder, an acknowledgment act by the responder from those states will move the CGU to the final state. In states 3 and 4, the situation is reversed.

3.2 Verbal and non-verbal feedback behaviors

Listener feedback can be characterized with respect to several dimensions. Most important are: expression and modality used in feedback (e.g. audible speech, visible body movements), types of function/content of the feedback expressions, types of reception preceding giving of feedback, types of appraisal and evaluation occurring in listener to select feedback. Allwood et al. ([2], p. 256, Table 1) show how different types of embodied feedback behavior can be differentiated according to these dimensions. The authors provide a predictive model of embodied feedback based on an empirical corpus study to support simulation in an embodied conversational character [15].

An **acknowledging head nod** conveys an acknowledgment grounding act. It is an alternative to a verbal acknowledgment. A **verbal backchannel** (e.g. "okay", "right", "uh-huh") can also be used to perform an acknowledgment act. During a speaker's utterance, a listener whose understanding is progressing adequately may signal continued attention with an **attentive head nod**, inviting the speaker to proceed with their utterance. A **frown** can be used to realize a request for repair. As discussed in Sect. 2, **completion** can be used to acknowledge understanding of both explicit and predicted content. An example can be found in Dialogue Excerpt (1), where A completes C's unfinished utterance. This behavior conveys an acknowledge act for the full predicted utterance it is completing. Completions will generally occur when understanding confidence is high, although trial completions may be used in cases of lower confidence.

4 A model for incremental grounding

In this paper, we adapt the Traum '94 model (presented in the previous section) to allow more fine-grained incremental processing. The core of our approach is to allow CGUs to be created and updated incrementally, while an utterance is in progress. These incremental updates can affect both the grounding states and the contents of the CGUs. They can also result in the creation of new CGUs. We first describe the model of incremental interpretation that we assume, which will be the inputs for the grounding model. Next we describe two possible ways of adapting the Traum '94 model to account for partial hypotheses of an ongoing utterance. Then we present the details of our model, specifying recognition conditions for the incremental versions of grounding acts, including four different kinds of incremental acknowledgement, and how else the grounding state is affected.

4.1 Incremental interpretation for grounding

We assume a model of incremental speech understanding that delivers a finite sequence of incremental outputs (that

Table 2 Traum's CGU transition diagram from [36]

Next act	In state						
	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[a]	F	
Ack^{R}		F	F[a]			F	
ReqAck^{I}		1				1	
ReqAck^{R}				3		3	
Cancel^{I}		D	D	D	D	D	
Cancel^{R}			1	1		D	

[a] Repair request is ignored

Table 3 Model input for the utterance "Utah we can give you two hundred dollars"

Partial	ASR transcription	NLU attributes	NLU values
0	UTAH	*s.addressee	utah
1	UTAH	*s.addressee	utah
2	UTAH WHAT	*s.addressee	utah
		*s.sem.speechact.type	no-ack
3	UTAH WHAT WE CAN	*s.addressee	utah
		s.sem.speechact.type	no-ack
4	UTAH WHAT WE CAN GET YOU	*s.addressee	utah
		*s.mood	declarative
		*s.sem.type	event
		*s.sem.speechact.type	statement
		s.sem.agent	you
		s.sem.event	providePublicServices
		s.sem.modal.desire	want
		s.sem.modal.holder	we
		s.sem.theme	sheriff-job
5	(same)	(same)	(same)
6	UTAH WHAT WE CAN GIVE YOU TWO	*s.addressee	utah
		*s.mood	declarative
		*s.sem.type	event
		*s.sem.agent	we
		*s.sem.event	give
		*s.sem.destination	you
		*s.sem.modal.possibility	can
		*s.sem.speechact.type	offer
		*s.sem.theme	twohundred
7	UTAH WE CAN GIVE YOU TWO HUNDRED DOLLARS	(same)	(same)

Frame elements marked with an *asterisk* are part of the explicit sub-frame

we call *partials*) as an utterance progresses. Each partial includes output of the natural language understanding component, including a hypothesis of the sequence of words that have been spoken (a common output of many speech recognizers), an estimate of the predicted [10] and explicit [11] content of the utterance at each point in time, as well as a set of confidence scores Suppose that N partials are delivered during a spoken utterance. We will denote the sequence of partials by $\mathcal{O}$, as shown in (5), where S_i is the surface text, E_i is the explicit content P_i is the predicted content, and C_i is the set of confidence scores for the ith partial. Each E_i is a subset of P_i, we may also refer to F_i as the predicted future completion of the utterance, such that $F_i = P_i - E_i$.

$$\mathcal{O} = \langle (S_1, E_1, P_1, C_1), ..., (S_N, E_N, P_N, C_N) \rangle \quad (5)$$

While many different types of semantic representation could be used (e.g. the types described in the previous section), for concreteness, we will use the representation format from [32], used in [11], and the implementation described in the next section. Here meaning is represented as semantic frames, each of which consists of a set of attributes and values. The representation allows recursion, as attribute values can also be frames. The representation is linearized, so that each attribute is described as a path of attribute names from the root to leaf. An example is shown in Table 3. Here we can see the progression of speech hypotheses on the left. First the interpreter thinks this is just calling the character Utah. Partial (2) is interpreted as a signal of lack of understanding. Partial (3) keeps this view, but is no longer as confident that the repair-request is correct (there are other plausible interpretations). In partial (4), the interpreter thinks that the speaker will offer Utah the sheriff job, but thinks the explicit meaning has just specified a declarative statement about an event. Finally by partial (6), the interpreter has finalized on the interpretation that the speaker will offer Utah 200 dollars. We can see that some partials do not change from the previous hypotheses (e.g. 1, 5, where ("same") is used to mean that the

Table 4 Content-first approach to incremental grounding of utterance in Table 3

t	E_t	New content	Removed content	Grounding act
0	s.addressee utah	s.addressee utah		Initiate
1	s.addressee utah			
2	s.addressee utah	s.sem.speechact.type no-ack		Continue
	s.sem.speechact.type no-ack			
3	s.addressee utah		s.sem.speechact.type no-ack	Repair
4	s.addressee utah			Continue
	s.mood declarative	s.mood declarative		
	s.sem.type event	s.sem.type event		
	s.sem.speechact.type statement	s.sem.speechact.type statement		
5	(same)			
6	s.addressee utah		s.sem.speechact.type statement	Repair
	s.mood declarative			
	s.sem.type event			
	s.sem.agent we	s.sem.agent we		
	s.sem.event give	s.sem.event give		
	s.sem.destination you	s.sem.destination you		
	s.sem.modal.possibility can	s.sem.modal.possibility can		
	s.sem.speechact.type offer	s.sem.speechact.type offer		
	s.sem.theme twohundred	s.sem.theme twohundred		

representation is the same as the previous partial), while others can change by the addition of one (2) or more (3, 4, 6, 7) words, or retraction of one or more words (6,7). Likewise the explicit and predicted meanings can change in multiple ways, e.g. addition of new material (2, 4, 6), change in status from predicted to explicit or vice versa (3), or replacement (6).

The tasks of an incremental grounding model are thus to assign a set of grounding acts to each partial input, and to provide an update procedure for partial grounding acts. Much of the Traum '94 model can be retained, however there are some complications that must be addressed. First, there is the question of whether grounding acts should be calculated at the full utterance level or at the level of individual partials. A related issue is that of revisions to the interpretation of prior material, such as partial 6. Unlike the interpretation of full utterances, the interpretation of partials change very rapidly-especially at the frontier of interpretation. That is, very often $P_i \neq P_{i+1}$ and E_i is not a prefix of E_{i+1}.

4.2 Two approaches to incremental grounding

We have investigated two approaches to modeling grounding in an incremental dialogue system: a *content-first* approach and *function-first* approach. The main distinction is whether we should create a new (set of) grounding acts for each partial that is in any way different from previous partials, even if the only difference is a change in content rather than the grounding function, compared to other utterances (content-first), or whether we should create new grounding acts only when the grounding function is changed. There are merits on both sides, as the content-first approach is closer to the original Traum '94 model in terms of act updates, while the function-first model is less sensitive to the sampling rate and lack of stability of the partial interpretations. We discuss the content-first approach, and then adopt the function-first model which is described in more detail in the following sections.

In the content-first approach, the difference between the explicit content of the partials is used as the main input for the grounding model, and function is computed from the relationships between the contents. Each different partial is seen as a new grounding act. The first partial with explicit content is seen as an *initiate* act, while after that, any partial with removed content is seen as a *repair* act, and any partial with new content but no removed content is seen as a *continue* act. For example, Table 4 shows a sequence of incremental results for the same developing utterance as Table 3, breaking down the changes from the previous partial into new and removed content, and showing the resulting grounding acts, according to the content-first approach.

With this approach, the grounding model does not distinguish between what the user said and what the interpreter understood. An interpreter revision is a repair, and not treated differently than an actual repair, when the user fixes his previ-

ous statement, by retracting meaning.The lack of a distinction between these cases emphasizes their similarity and relieves the system from having to reliably identify either of the two individually in an utterance.

Compared to Traum's model, this approach models a single utterance as a sequence of grounding acts instead of a single grounding act that covers the function of the complete utterance. This works well for the authorial grounding acts, because they affect the content of the CGUs as they are being conveyed, which can now be processed incrementally.The remaining grounding acts, (request repair, acknowledge and cancel) do not progress in any way while the utterance is being uttered.The individual partials do not bring any content, but together make for one of the three aforementioned acts, the first partial no different than the last. It is not clear how to indicate grounding acts for individual partials (all, only the first one), or how to update the function of individual partials in a consistent manner with the authorial grounding acts.

The function-first approach looks at partials as contributing to grounding acts, rather than realizing individual grounding acts. This allows a more uniform treatment of authorial and non-authorial grounding acts, and also simplifies the updates in the case of revised interpretations from one partial to the next. As long as consecutive partials are deemed to be contributing to the same grounding act, they are grouped together, and only the content is modified in the grounding act. The sequence of partials in Table 4 would all be seen as one initiate act. We will adopt this model in the discussion below.

4.3 Grounding act recognition and updates

The function-first model will use the predicted frame and the explicit sub-frame in different ways as it processes the partial interpretations. The predicted frame provides more information about the pragmatics of the utterance and is therefore used to determine the grounding acts that the user performs. The explicit sub-frame is a better representation of the current state of affairs at the time of processing and is therefore used to update the contents of the CGUs. When combined, the predicted frame will help the interpretation of the explicit sub-frame. Each grounding act is detected and processed in a different manner, which will be described below.

Initiate acts generally occur when a speaker begins a new utterance which does not include a continue, request repair, repair or cancel act.[2] Initiate acts create a new open CGU, whose content will be the ungrounded explicit content of the evolving utterance. As the utterance progresses, the explicit content E_i of each new incremental understanding output is generally used to update (i.e. replace) the content of the open CGU. E.g. for the example in Table 4, according to the function-first approach, the initiate act contents would be progressively updated with the new explicit meaning for each partial.

If an overlapping backchannel or request for repair is detected, the initiate act ends. Let the overlapping behavior start at partial t of this utterance U, then content of the CGU will be E^U_{t-1}. In the case of an acknowledging backchannel, the CGU is grounded, and if the speaker decides to continue uttering U, this will be a new initiate act. The initial content of the new CGU will be the ungrounded content of E^U_t, which is the new elements that were not grounded with the previous CGU. For example, if Utah nods right after the first partial in Table 4, then partial (2) would be an initiate act for a new CGU, with only the new content. A second overlapping backchannel in the same utterance is handled analogously. In the case of a request repair, the continuation of U is seen as a repair grounding act.

Continue acts occur when a new speaker utterance serves as a continuation of an ungrounded CGU that was previously initiated. As a rule, when an interlocutor begins to speak, if there is an open CGU with the speaker as most recent author, and the utterance does not convey a repair, the utterance is treated as a continue act. Each new incremental output E_i is used to update (i.e. replace) the content of the developing continue act.

Acknowledgments transition CGUs into the final grounded state, and move the content of the CGUs into the common ground. Of particular interest for incremental grounding is the case of overlapping acknowledgment. We can distinguish several kinds of overlapping acknowledgement, as seen by the examples in Sect. 2. In these cases, there is a speaker performing an utterance V, and acting as current author of a CGU. The other interlocutor performs an utterance U, before all partials of V have been completed. E^U_t, P^U_t are the explicit and predicted content of utterance U at time t. E^V_u, and P^V_u are the explicit and predicted content of V when U is started. We can define the following types of incremental acknowledgement, and their updates as follows:

Backchannel: P^U_t conveys a positive backchannel (e.g. "Yeah", "Uh-huh", etc.). The CGU is grounded with E^V_t. The current authorial act in V is ended and if the speaker continues after/during the backchannel, this is interpreted as the start of a new grounding act.
Completion: V is unfinished and E^U_t contains a syntactical or conceptual continuation of E^V_u. The CGU is grounded with content E^V_u. If the completion actually matches the intentions of the speaker of V, then this is treated as a kind of explicit verification (see below), and some additional material from V is also seen as grounded.

[2] Sometimes, an utterance that includes an acknowledgment will also proceed to initiate a new CGU (as in "okay, so let's talk about the other matter").

Explicit Verification: E_t^U contains parts of P_u^V. If V is unfinished, the predicted content of V at the time t of the start of U that is also in the explicit content of U ($P_t^V \cup E_t^U$) is added to the CGU. For example, in Table 3, if after utterance 4, the addressee has completed with "can get you the sheriff job", and the original speaker did not continue or correct this, then the predicted content of partial 4 would be seen as grounded.
Implicit verification: P_t^U is the next relevant contribution to P_u^V (e.g., the answer to a question). This will ground the CGU of V with the full utterance meaning P_t^V as its content. Utterances that are implicit verifications also present new content—hence the *implicit*—that will initiate a new CGU. Partials of this utterance will therefore also be processed according to the description of *initiate* above.

Correct completions are treated as an acknowledgement of the complete utterance, both the explicit and predicted part, since that is the intention of the completing party. The completion makes the predicted content part of the CGU, which so far only contained the explicit content.

Requests for repair change the state of V's CGU to indicate that a repair from the other party is requested. If the request for repair is overlapping, the CGU's content will be E_t^V and the current authorial act in V is ended. If the speaker continues after/during the request repair, this is a new grounding act (e.g. a repair of the current CGU, or initiate of a new one) .

The < S > .sem.speechact.type no-ack frame element, as seen in Table 3, signals misunderstanding by the speaker. If this frame element appears in the predicted utterance meaning, the speaker is said to execute a request repair act related to the most recent CGU that was not initiated by the speaker or recently repaired by someone else than the speaker, if such a CGU exists.

Repairs will modify the content of a CGU. For each partial of this repair, the explicit content is added to the CGU. Existing content in the CGU that is not compatible with the content of the repair is removed. Material can be incompatible in several ways, following the logic of the semantic representation. Frames roughly represent propositions, so propositions that are contradicted by the current frame would be retracted from a CGU (e.g. if a speaker is initially understood to have said that the town is safe, but later thought to say that that the town is not safe). At a lower level, frames can be one of several types, each with a set of permitted and optional attributes. When new material repairs an existing frame, incompatible frame elements are removed (e.g. a change from a state to an action, or a change in the destination or theme of an action).

Cancel acts move the relevant CGU into a special canceled state. No special logic is needed to handle this in the incremental grounding model.

This model is essentially a model of grounding in dyadic conversation, with two participant roles, *initiator* and *responder*. However, like the Traum'94 model it can be applied to multi-party dialogue. The Traum'94 model was used in multiparty systems [21,34,37], with two virtual agents and one or two human participants. In this case, the responder can be any of the non-initiating participants. An utterance is seen as grounded by all conversational participants if any licensed participant provides an appropriate acknowledgement move (e.g. a responder in state 1 or the initiator in state 3). While this model might not capture all aspects of multiparty grounding adequately, it is at least sufficient for exchanges such as those in Sect. 2.

We can now look at how our incremental grounding model can address the AMI corpus examples from Sect. 2. We can not do a full analysis, because we do not have a semantic content model for this domain, or an NLU that can provide interpretations of the text, or confidence values. However, if we assume a naive approach, where non-filler content words will stand in for content, we can analyze them as follows. In example (1), A's two utterances are initiate grounding acts, and B and C's utterances are acknowledges. In (2), C's utterance is an initiate, while D and A's utterances are both acknowledgements of this CGU. A's utterance is a completion and this could have two different impacts, depending on whether or not this is what C intended. If this is what C intended, then the content of A's completion is added to the first CGU, which is now seen as grounded. If not, then A's utterance is also the initiation of a new CGU. The fact that C does not continue his own utterance or correct A, seems to favor the prior interpretation. However, if the latter interpretation were taken, then B's utterance could be seen as implicitly verifying this content, and grounding the second CGU, so the grounded content ends up the same in either case (the difference being that if B or anyone else had not responded, the status of A's content would then be in doubt). In example (3), we can see that C's utterance "the technology" is clearly not what B intended, so this follows the second interpretation described above—C's first utterance grounds the first CGU with content of B's first line, and then B's second line contains a new initiation which is grounded by C's last line. The CGU initiated by C's first utterance is left ungrounded. Example (4) starts the same way, with a completion/initiation by D, but in this case, B abandons his own continuation and explicitly grounds the CGU initiated by D before continuing on with the initiation of another unit.

5 Implementation

We have implemented the incremental grounding model from Sect. 4 as an extension to the ICT virtual human spoken dia-

logue architecture [35], which has been designed to allow trainees to practice their negotiation skills by engaging in face to face negotiation with virtual humans. It has been tested with data from the SASO4 scenario, which extends the scenario described in [21]. In the SASO4 scenario, two human users play the role of a US Ranger and his deputy, and negotiate with two virtual humans, called Utah and Harmony, to try to convince them that Utah should become the new sheriff of a town.

The virtual human system includes the following set of processing components:

- Automatic speech recognition (ASR), mapping speech to words), and producing incremental results. The rate at which the ASR produces results is also the rate of incrementality for NLU and grounding act recognition (this is currently every 200 ms, which leads to results such as the size and sequence of the partials in Table 3).
- Natural language understanding (NLU), mapping from words to semantic frames and confidence scores. Examples of explicit and predicted frames can be seen in Table 3). The confidence metrics make qualitative distinctions about the system's level of understanding, and can judge the current understanding level to be *low*, *high*, *incorrect*, and *correct*, among others; see [9].
- Dialogue interpretation and management (DM), handling context, dialogue and grounding act recognition, reference resolution, and deciding what content to express. The full contents of CGUs and grounding acts includes contents from the NLU, as seen in Table 3, augmented by the dialogue manager with resolved references and additional hypotheses about speech acts.
- Natural language generation (NLG), mapping frames to words.
- Non-verbal generation (NVBG), mapping function markup language (FML) to behavioral markup language (BML).
- Text to speech synthesis (TTS) and
- Behavior realization (SBM), mapping BML to behaviors of an animated character.

To these, we have added our new component to track incremental grounding and implement a feedback policy.

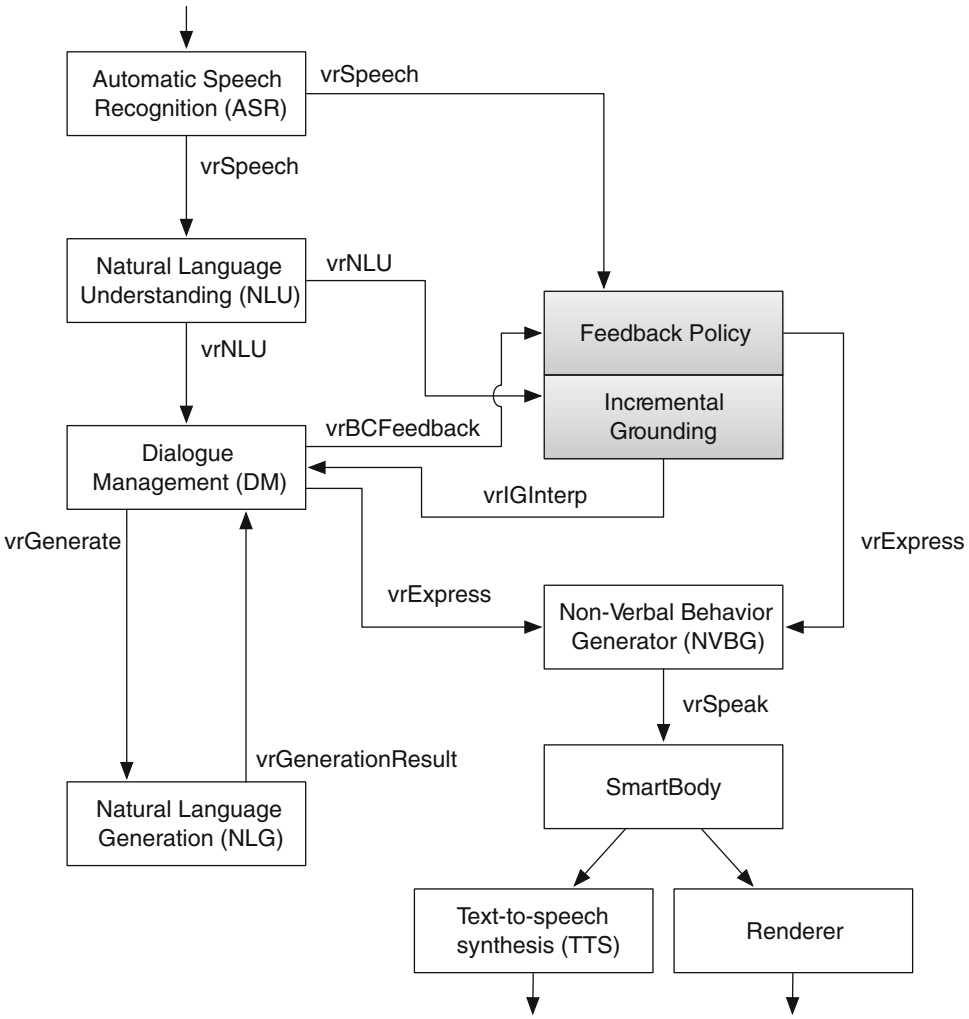
The components communicate via a shared message bus that is provided through the Virtual Human Messaging System (VHMsg). System components can subscribe to certain message types and will be notified when the requested messages are sent. This makes the components loosely coupled, which enables the use of virtually any programming language and gives the freedom to run the system on multiple physical machines.

For our initial incremental grounding implementation, we chose to take advantage of this modular architecture and develop a new proof-of-concept component focused only on incremental grounding and overlapping behavior production, rather than replace the current (non-incremental) Traum '94 grounding implementation that is inside the DM module. This was for simplicity during development, but it also has the advantage of portability, so it can be used with other dialogue managers that adopt the same API. In addition to modeling incremental grounding, the component also executes a simple overlapping behavior policy that showcases up-to-date knowledge of the grounding state. The component selects behaviors according to the policy and instructs the appropriate components to execute those behaviors. Our policy is a rudimentary variation on Wang et al.'s [40] comprehensive listener feedback model.

An overview of the system including our component is displayed in Fig. 1. Most of the messages shown are standard parts of the virtual human toolkit [13]. The vrBCFeedback message was introduced in [33], and gives information about the contextual interpretation of an on-going utterance, including dialogue manager interpretations of reference resolution, participant roles, and references for each partial. We introduce a new message in this work, **vrIGInterp**, to notify the original dialogue manager of updates to the incremental grounding state.

Our new component implements an initial version of the incremental grounding model described in Sect. 4. The implementation initializes and extends CGUs incrementally, as users are speaking, to maintain an incremental grounding state. To generate grounding behaviors in our virtual humans, we have also designed and implemented an overlapping behavior policy for the virtual humans in SASO4, which we summarize in Fig. 2. The behaviors are selected from existing work on feedback models for virtual agents [14,40], describing various types of nods and facial expressions to signal understanding or confusion.

The policy is evaluated with each new partial, which means essentially every 200 ms. The policy approximates "speech activity" as a new partial with different surface form than the previous one—so all partials except (1) and (5) in Table 3 would be seen as having speech activity. A lack of "speech activity" is seen as a pause, which is categorized as "short" if it is only a few partials in duration, but "long" if it goes on for more than three partials (600 ms). The policy allows our virtual humans to provide frequent feedback of their level of understanding. For instance, after a short pause in user speech, when there is ungrounded content in a CGU, three kinds of incremental feedback may be provided. If NLU is fully confident that its predicted understanding is *correct*, a verbal backchannel is generated. If the NLU confidence level is *high* (but the NLU is not confident that its understanding is perfectly correct), an acknowledging nod is generated. If the NLU confidence



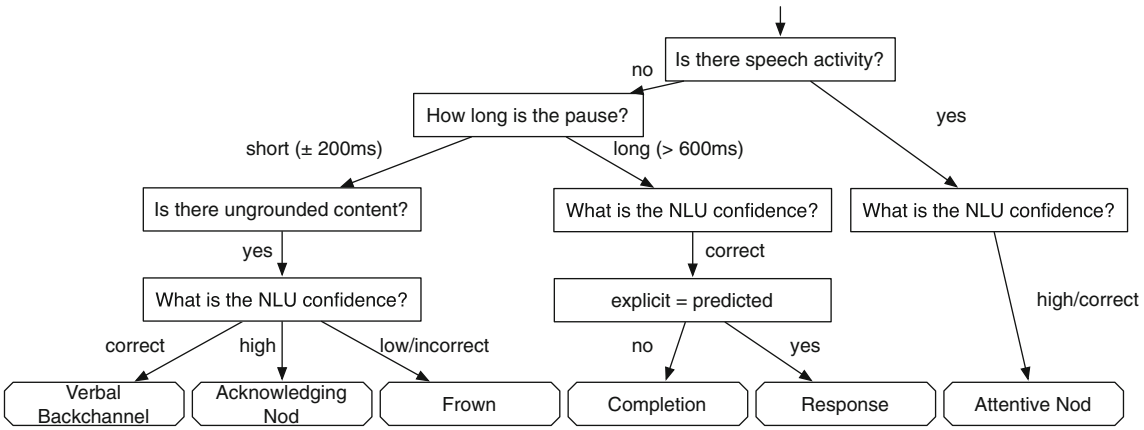


Fig. 1 Overview of the SASO4 dialogue system. Our components are printed in *grey*. The *annotated lines* show the inter-component communications over the VHMsg system, the label is the message type

Fig. 2 An overview of the overlapping behavior policy

level is *low* or *incorrect*, they generate a frown, signaling a request for repair. Similar rules enable the virtual humans to generate utterance completions or to simply respond to the user's utterance during longer pauses in user speech. A response is chosen in cases when the user's utterance is "finished" in the sense that the explicit content is equal to the predicted content. (In such cases, no completion is necessary).

Table 5 Behavior policy and function-first grounding model for utterance in Table 3

Partial	Confidence	Grounding act	CGU state	CGU contents
0	Medium	$initiate_1$	1	s.addressee utah
1	High	$initiate_1$	1	s.addressee utah
utah: acknowledging nod		ack_1	F	s.addressee utah
2	High	$initiate_2$	1	s.sem.speechact.type no-ack
3	Medium	$initiate_2$	1	
4	Low	$initiate_2$	1	s.mood declarative s.sem.type event s.sem.speechact.type statement
5	Low	$initiate_2$	1	(same)
utah: frown		$reqrepair_2$	2	(same)
6	High	$repair_2$	1	s.mood declarative s.sem.type event s.sem.agent we s.sem.event give s.sem.destination you s.sem.modal.possibility can s.sem.speechact.type offer s.sem.theme twohundred
utah: attentive nod			1	(same)
7	correct	$repair_2$	1	(same)

We can illustrate the functioning of this policy by examining how it is applied to the input in Table 3, and how this policy will change the grounding status (and future grounding act recognition, according to the function-first approach), as seen in Table 5. After the first short pause, Utah is confident enough to acknowledge, which completes this CGU and then partial 2 initiates a new CGU. Utah is merely listening until the second pause at partial 5, because the interpreter confidence is not high enough. The frown indicates lack of understanding, and the speaker's continuation is now seen as repairing this lack. At partial 6, utah is now confident of the interpretation and gives an attentive nod, inviting the speaker to continue (without affecting the grounding state). If the speaker were to pause at this point rather than continuing on with more words in partial 7, and the interpreter confidence remained high, Utah would acknowledge, and eventually complete the utterance, e.g. saying "two hundred dollars".

While an initial version of this model is implemented, and the virtual humans' responses often seem appropriate, e.g., nodding, frowning backchanelling and completing as expected, several aspects of the implementation still need to be extended and improved. We also have not yet evaluated the incremental grounding behavior in interactions with users.

6 Future work

There are several strands of work that we would like to engage in. First, the implemented model has not been thoroughly tested. We would like to evaluate the implementation in several ways. First, similarly to [40], we would like to do an observer evaluation, in which people view a video of an interaction between a user and the system, comparing a non-incremental version with an incremental version to gauge whether overlapping behavior is perceived as more natural. Second, we would like viewers to report on their estimates of what is in common ground for such examples to see if these judgements match the model. Finally, we would like to evaluate the full system with users who can interact with incremental and non-incremental versions of the system, and see if there is an impact not just on viewers but on users who are trying to negotiate.

In addition, we would like to consider three types of extensions to the model and the implementation, described below.

6.1 Enhanced behavior policy

The behavior policy in Sect. 5 was sensitive to several important phenomena, including pause length, interpretation con-

fidence, and grounding state, but should also be attentive to a number of other factors, such as prosody (marking both speaker's own expressed confidence in the material, and turn-taking cues), gaze of the speaker, the content of the utterance, and how the agent evaluates the material being spoken. We hope to merge the grounding-related considerations described here with more general work in output behavior planning such as that in [33,40].

6.2 Degrees of grounding

In Traum's model of grounding, a CGU can be in either of three states: ungrounded, in the process of being grounded and grounded. By providing evidence of understanding, the interlocutors ground content, but only if that evidence is strong enough. The type of content, the importance of it being fully understood, shared experiences between the participants, etc. together determine what evidence strength is enough, i.e. the grounding criterion. Evidence that is too weak will not ground the content and evidence that is strong enough will. We took Traum's model as a starting point for our work and therefore follow the same principle. As a result, our model is ignoring the evidence of understanding of the attentive nod from our feedback policy. The evidence is too weak for most cases and therefore we err on the side of not grounding, effectively throwing away the evidence the attentive nod conveys.

In [23,24], Roque presents an extension to Traum's theory that adds degrees of grounding to the model. In the proposed extension, the state of a CGU depends on the type of evidence provided, registering all evidence types, weak and strong. In a continuation of our work, Roque's adjustments to Traum's theory could be merged with our contribution to form a comprehensive grounding model.

6.3 Continuous processing

We have been talking about incremental processing as the early processing of parts of a whole utterance, with clearly defined markers of the beginning and ending of an utterance, such as with a push-to-talk button. At the end of an utterance, the ASR gives its final transcription and the Natural Language Understanding component (NLU) its final meaning representation. That final result is what the components stick with or, in terms of Schlangen and Skantze, *commit* to [26].

In continuous processing, the input is a continuous audio signal, without the artificial source of certainty provided by the user releasing the push-to-talk button. Automatic Speech Recognizers (ASR) evolved from being able to detect individual words to being able to detect a sequence of words. An ASR treats each piece of audio signal as both an additional part of the previous word and the first part of the next word, resulting in many possible transcriptions. This is called the ASR *lattice*, from which the most probable outcome can be selected. This principle can also be applied to the NLU [1], i.e. treating each word as both an addition to the current frame and the first word of the next frame. From these elaborations, the NLU can select the most probable stream of frames for the continuous speech signal instead of a single one per utterance. This is an interesting direction to pursue in the near future.

Certain visual behavior can already be recognized as grounding acts [38], however these are either discrete acts (like head nods) or features that co-occur with utterances (like head direction). More fine-grained, continuous behavior should also be related to the grounding model.

7 Summary

In this paper, we have presented a computational model for incremental grounding and an overlapping behavior policy that leverages the up-to-date grounding state. A theoretical model was presented that extends Traum's grounding acts model, and can account for a number of examples of overlapping grounding. We also implemented this model within an existing virtual human dialogue system, giving it the ability to perform different kinds of grounding related action, based on features like interpretation confidence, grounding state, pause length, and estimation of the completeness of the ongoing utterance.

Acknowledgments Some of the effort described here has been sponsored by the US Army. Any opinions, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. op den Akker H, Schulz C (2008) Exploring features and classifiers for dialogue act segmentation. In: Popescu-Belis A, Stiefelhagen R (eds) Machine learning for multimodal interaction. Lecture notes in computer science, vol 5237. Springer, Heidelberg, pp 196–207
2. Allwood J, Kopp S, Grammer K, Ahlsn E, Oberzaucher E, Koppensteiner M (2007) The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. Lang Res Eval 41(3—-4):255–272. doi:10.1007/s10579-007-9056-2
3. Bohus D, Horvitz E (2009) Learning to predict engagement with a spoken dialog system in open-world settings. In: Proceedings of SIGDIAL 2009. London
4. Buß O, Baumann T, Schlangen D (2010) Collaborating on utterances with a spoken dialogue system using an isu-based approach to incremental dialogue management. In: Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue Association for, Computational Linguistics. pp 233–236

5. Carletta J (2007) Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. Lang Res Eval 41(2):181–190
6. Clark H (1996) Using language. Cambridge University Press, linebreak Cambridge
7. Clark H, Schaefer E (1989) Contributing to discourse. Cogn Sci 13(2):259–294
8. DeVault D, Sagae K, Traum D (2009) Can i finish? Learning when to respond to incremental interpretation results in interactive dialogue. In: 10th SIGdial Workshop on Discourse and Dialogue. London
9. DeVault D, Sagae K, Traum D (2011) Detecting the status of a predictive incremental speech understanding model for real-time decision-making in a spoken dialogue system. In: The 12th Annual Conference of the International Speech Communication Association (InterSpeech 2011)
10. DeVault D, Sagae K, Traum D (2011) Incremental interpretation and prediction of utterance meaning for interactive dialogue. Dialog Discourse 2(1)
11. DeVault D, Traum D (2013) A method for the approximation of incremental understanding of explicit utterance meaning using predictive models in finite domains. NAACL-HLT 2013
12. Gratch J, Okhmatovskaia A, Lamothe F, Marsella S, Morales M, van der Werf R, Morency LP (2006) Virtual rapport. In: Gratch J, Young M, Aylett R, Ballin D, Olivier P (eds) Intelligent virtual agents, vol 2. Springer, Berlin, pp 14–27. doi:10.1007/11821830_2
13. Hartholt A, Traum DR, Marsella SC, Shapiro A, Stratou G, Leuski A, Morency LP, Gratch J (2013) All together now—introducing the virtual human toolkit. In: Aylett R, Krenn B, Pelachaud C, Shimodaira H (eds) IVA, Lecture notes in computer science, vol 8108. Springer, Berlin, pp 368–381
14. Huang L, Morency L, Gratch J (2011) Virtual rapport 2.0. Intelligent virtual agents. Springer, Berlin, pp 68–79
15. Kopp S, Allwood J, Grammer K, Ahlsen E, Stocksmeier T (2008) Modeling embodied feedback with virtual humans. In: Proceedings of the Embodied communication in humans and machines, 2nd ZiF research group international conference on Modeling communication with robots and virtual humans, ZiF'06, Springer-Verlag, Berlin, pp 18–37. http://dl.acm.org/citation.cfm?id=1794517.1794519
16. Matheson C, Poesio M, Traum D (2000) Modelling grounding and discourse obligations using update rules. In: Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics
17. Milward D (1992) Dynamics, dependency grammar and incremental interpretation. In: COLING92, pp 1095–1099
18. Morency LP, Kok I, Gratch J (2010) A probabilistic multimodal approach for predicting listener backchannels. Autonom Agent Multi-Agent Syst 20:70–84. doi:10.1007/s10458-009-9092-y
19. Nakatani C, Traum D (1999) Coding discourse structure in dialogue (version 1.0). Tech. Rep. UMIACS-TR-99-03, University of Maryland
20. Oviatt S, Cohen P (1991) Discourse structure and performance efficiency in interactive and non-interactive spoken modalities. Comp Speech Lang 5(4):297–326
21. Plüss B, DeVault D, Traum D (2011) Toward rapid development of multi-party virtual human negotiation scenarios. In: Proceedings of SemDial
22. Poesio M, Traum DR (1997) Conversational actions and discourse situations. Comput Intell 13(3)
23. Roque A (2009) Dialogue management in spoken dialogue systems with degrees of grounding. Ph.D. thesis, University of Southern California, Los Angeles
24. Roque A, Traum D (2008) Degrees of grounding based on evidence of understanding. In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, Association for, Computational Linguistics. pp 54–63
25. Schlangen D, Baumann T, Buschmeier H, Buß O, Kopp S, Skantze G, Yaghoubzadeh R (2010) Middleware for incremental processing in conversational agents. In: Proceedings of SigDial 2010. Tokyo
26. Schlangen D, Skantze G (2009) A general, abstract model of incremental dialogue processing. In: Proc. of the 12th Conference of the European Chapter of the ACL
27. Schuler W, Wu S, Schwartz L (2009) A framework for fast incremental interpretation during speech decoding. Comput Ling 35(3):313–343
28. Selfridge E, Arizmendi I, Heeman P, Williams J (2011) Stability and accuracy in incremental speech recognition. In: Proceedings of the SIGDIAL 2011 Conference, Association for Computational Linguistics, Portland, pp 110–119. http://www.aclweb.org/anthology/W/W11/W11-2014
29. Skantze G, Hjalmarsson A (2010) Towards incremental speech generation in dialogue systems. In: Proceedings of the SIGDIAL 2010 Conference, Association for Computational Linguistics, Tokyo, pp 1–8. http://www.aclweb.org/anthology/W/W10/W10-4301
30. Skantze G, Schlangen D (2009) Incremental dialogue processing in a micro-domain. In: Proceedings of the 12th Conference of the European Association for Computational Linguistics (EACL)
31. Tanenhaus M, Brown-Schmidt S (2008) Language processing in the natural world. Philos Trans Royal Soc B 363(1493):1105–1122
32. Traum D (2003) Semantics and pragmatics of questions and answers for dialogue agents. In: proceedings of the International Workshop on Computational Semantics, pp 380–394
33. Traum D, DeVault D, Lee J, Wang Z, Marsella S (2012) Incremental dialogue understanding and feedback for multiparty, multimodal conversation. In: Intelligent Virtual Agents. Springer
34. Traum D, Rickel J, Marsella S, Gratch J (2003) Negotiation over tasks in hybrid human-agent teams for simulation-based training. In: Proceedings of AAMAS 2003: Second International Joint Conference on Autonomous Agents and Multi-Agent Systems, pp 441–448
35. Traum D, Swartout W, Gratch J, Marsella S (2008) A virtual human dialogue model for non-team interaction. In: Dybkjaer L, Minker W (eds) Recent trends in discourse and dialogue. Springer, Netherlands
36. Traum DR (1994) A computational theory of grounding in natural language conversation. Ph.D. thesis, University of Rochester, Rochester
37. Traum DR, Marsella S, Gratch J, Lee J, Hartholt A (2008) Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In: Prendinger H, Lester JC, Ishizuka M (eds) IVA, lecture notes in computer science, vol 5208. Springer, Berlin, pp 117–130
38. Traum DR, Morency LP (2010) Integration of visual perception in dialogue understanding for virtual humans in multi-party interaction. In: AAMAS International Workshop on Interacting with ECAs as Virtual Characters
39. Traum DR, Schubert LK, Poesio M, Martin NG, Light M, Hwang CH, Heeman P, Ferguson G, Allen JF (1996) Knowledge representation in the TRAINS-93 conversation system. Intern J Exp Syst 9(1):173–223
40. Wang Z, Lee J, Marsella S (2011) Towards more comprehensive listening behavior: beyond the bobble head. In: Intelligent Virtual Agents, Springer, Berlin, pp 216–227
41. Ward N, Tsukahara W (1999) A responsive dialogue system. In: Wilks Y (eds) Machine conversations. Springer, New York