*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

KRISTINN R. THÓRISSON

# NATURAL TURN-TAKING NEEDS NO MANUAL: COMPUTATIONAL THEORY AND MODEL, FROM PERCEPTION TO ACTION

## 1. INTRODUCTION

***t-minus-460 msec***

Beth and Alan are sitting at a Fifth Avenue outdoors restaurant in Manhattan. Alan is telling Beth an exciting story about his vacation in Nice. Alan presents the story through gesture and speech. Then Beth's arm starts moving and her neck stiffens.

We, the viewers, know that she's surprised to see an elephant in the middle of Manhattan, and that in 460 milliseconds her arm and hand motion will turn into a well-defined deictic gesture, her eyebrows will rise, and her mouth will open with surprise, at which point Alan will most certainly recognize the signs and look over at the elephant. But right now, at t-minus-460 milliseconds, Beth's gesture is barely recognizable as a communicative action, so Alan doesn't know for sure. And thus, before that all happens, in the next 460 milliseconds, Alan has to decide what to do about Beth's behavior. Should he stop telling his story? Or should he go on, in case Beth is simply adjusting her jacket?

Decisions like these are made by dialogue participants as often as 2-3 times per second. For a 30 minute conversation that's over 5000 decisions. And that's just a fraction of what goes on. How do we do it? Face-to-face dialogue consists of interaction between several complex, dynamic systems — visual and auditory display of information, internal processing, knee-jerk reactions, thought-out rhetoric, learned patterns, social convention, etc. One could postulate that the power of dialogue is a direct result of this fact. However, combining a multitude of systems in one place does not guarantee a coherent outcome such as goal-directed dialogue. For this to happen the systems need to be architected in a way that guides their interaction and ensures that — complex as it may be — the interaction tends towards homeostasis in light of errors and uncertainties, towards the set of goals shared by participants.

Past research into what kinds of architectures might guide such systems has resulted in a broad range of studies that combines linguistics, psychology and artificial intelligence (cf. Maes 1990a, Adler 1989, Grice 1989, Grosz & Sidner 1986, Goodwin 1981). The body of work is impressive. However, much of the work focusing on human behavior and cognition has been descriptive, and not well suited, except in very general ways, for working implementations of artificial systems that can participate in interactive face-to-face dialogue. The divide-and-conquer approach of academic

research has further resulted in neglect of real-time interpretation and generation of multimodal behavior, a critical component to any such system. Building a generative model involves identifying the contributing processes and formalizing the interaction of these in a system capable of taking the role of a simulated human participant. To be certain that the model performs to spec, the best way to test it is in actual interaction with humans; without real-time constraints and the complexities of the real world the system could easily fail to address fundamental constraints in human communication, chief among them the march of a real-world clock. For this the system needs both real-time perception and action.

This chapter presents a computational model of natural turn-taking in goal-oriented, face-to-face dialogue. The model demonstrates fluent psycho-social dialogue skills in real-time interactions with human users, perceiving their multimodal actions — speech, prosody, body language, manual gesture, gaze — and generating multimodal behavior as output, including speech, facial expressions, manual gesture, spatial attention via head and eye movements, as well as manipulations in a topic domain. The first half of this chapter presents the theory of the model, formulated as a series of hypotheses. Here we look at prior research, identify the missing pieces and relate this to our computational perspective, rooted in classical and behavior-based artificial intelligence. The theory is not an analysis of the 'turn-taking rules' observed in human dialogue — which vary between cultures and can be separated out (our implementation leans on research in this area) — the theory and model present a *turn-taking mechanism*. The second half describes a model based on these hypotheses, and its implementation. The implementation has been tested with a wide range of users and shows significant promise as a first step in bridging semantic analysis, situated dialogue, discourse structure, auditory perception, computer vision, and action selection under a unifying framework. Performance examples of the prototype are given at the end of the chapter.

Pragmatically speaking, a generative model of turn-taking has the potential to free users from the "vending machine" symptoms that have plagued many communicative computer systems in the past: Arbitrary pauses, beeps, button pushes, and instruction guidelines. Any decent implementation of a generative, multimodal turn-taking model should allow for interaction with machines in the same way human interaction works, supporting seamless, finely-timed turn-taking, giving invisible support to the task and the situated natural language communication at hand — without the need for a manual.

The model presented here assumes no artificial protocols. It builds on work from psychology (Sacks et al. 1974, Duncan 1972) and artificial intelligence (Maes 1990b, Selfridge 1959), and is based on the Ymir mind-model for communicative creatures and humanoids (Thórisson 1996, 1999). The part of this model concerned with turn-taking is called the *Ymir turn-taking model*, YTTM, and it address the full perception-action loop of real-time turn-taking, from (1) the basics of multimodal perception to (2) knowledge representation, (3) decision making, and (4) action generation for gaze, gesture, facial expressions and speech planning and execution. The model assumes a task-oriented dialogue, and interfaces with knowledge systems via a limited set of primitives. It does not address the specifics of topic knowledge, and is therefore complementary to models such as Grozs & Sidner's (1986) focus space model, and Clark and Schaefer's contribution model (Cahn & Brennan 1999, Clark

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

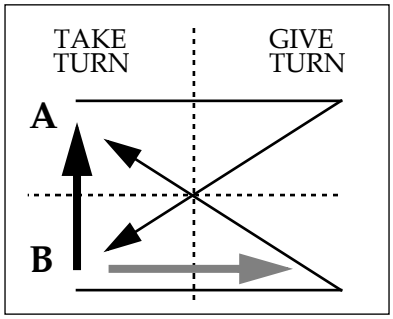


*Figure 1. The task of efficient turn transitions includes detecting acceptable transition points. In this figure, Alan (A) and Beth (B) are engaged in dialogue with each other. Beth is talking, Alan listening. Thin arrows demonstrate smooth turns, lower arrow indicates Beth giving turn to Alan; solid bold arrow constitutes an interruption (of B by A) with the possibility of overlapping speech, gray bold arrow shows a failure of the listener, A, to take the turn when it is given (by B), possibly with an unwanted silence.*

& Schaefer 1989), which model tracking of dialogue topic in general. YTTM has been implemented in two working systems (Bryson & Thórisson in press, Thórisson 1996), and tested in interaction with human users (Cassell & Thórisson 1999, Thórisson 1996). Results show the broad strokes of dialogue behavior it produces to be acceptable, both in perceiving/generating correct and acceptable turn transitions and in producing the necessary and sufficient turn-taking behaviors in real time, resulting in dialogue patterns similar to that observed in human-human conversation. (We present example interactions in Section 5.4.2.) The encouraging results have prompted the decision to summarize the model's background assumptions, as presented in this chapter. The implementation presented here is limited (1) to two participants, and (2) by the assumption of a single topic. Solutions to both limitations are well understood within the framework of the Ymir architecture (Thórisson 1999), and are currently being worked out.

We will now look at the main features of real-time dialogue as presented in prior research — the features which the model needs to address — and then go on to describe how they relate to the YTTM.

## 2. PRIOR RESEARCH

When people communicate in face-to-face interaction they take turns speaking (Duncan 1972). The system's main function is to sequentialize information exchange between two or more communicating parties and ensure efficient transmission (Figure 1). It is the key organizing principle of real-time dialogue. The information exchanged during typical face-to-face interaction is constructed through speech, hand gestures, body language, gaze, facial expressions, and multiple combination thereof (Sacks 1992, McNeill 1992, Goodwin 1981). Turn-taking and back-channel feedback (Yngve 1970) have both been shown to be important for conducting successful dialogue (Sacks et al. 1974, Nespolous & Lecours 1986). Turn-taking is, for example, crucial in both negotiation and clarification (Whittaker et al. 1991, Whittaker & Stenton 1988, Sacks et al. 1974).

Goodwin (1981, p. 2) says about the turn:

> "In the abstract, the phenomenon of turn-taking seems quite easy to define. The talk of one party bounded by the talk of others constitutes a turn, with turn-taking being the process through which the party doing the talk of the moment is changed."

Like many before (and after) him he goes on to say that on closer inspection things are not as simple as they look in the abstract. This is certainly true. However, we

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

argue that the complexity in turn-taking comes from the broad range of contextual influences on the processes, resulting in emergent phenomena that baffle efforts that only look at the surface phenomena. It is only through a thorough analysis of the underlying mechanisms, at multiple levels of detail, that the simplicity of the system becomes apparent.

Sacks et al. (1974) put forth a model of turn taking that characterizes the structure of human conversation as (1) an emergent property of (2) local decisions based on (3) prediction by the participants. In their view, turn taking is locally managed and participant-administrated. Local management means that "all the operations [within the system] are 'local', i.e. directed to 'next turn' and 'next transition' on a turn-by-turn basis" (Sacks et al. 1974, p. 725). In this view, any pattern that arises out of interaction is emergent in the sense that it results from the complex, non-scripted interaction between decisions that are made by each conversant with incomplete knowledge and an independent set of interaction rules. They say further (p. 725-6) that

> "...the turn-taking system is a local management system ... in the sense that it operates in such a way as to allow turn-size and turn-order to vary and be under local management, across variations in other parameters, while still achieving both the aim of all turn-taking systems—the organization of 'n at a time'—and the aim of all turn-taking organizations for speech-exchange systems — 'one at a time while speaker change recurs'".

"Party-administration" refers to the fact that the rules of turn-taking are subject to the conversants' control, i.e. that the rules are designed for being used by each participant individually to manage their communication with others. By hypothesizing the existence of turn-constructional units, Sacks et al. were able to model turn taking with only five — albeit relatively complex — rules. But the most important part of their theory is the set of turn-constructional units they propose, which are *sentential*, *clausal*, *phrasal* and *lexical*. (More unit candidates would clearly have resulted if they had included multiple communication modes in their analysis.) According to their theory, these units are used by speakers to construct a turn (i.e. determine transitions). For example, recognizing that a particular sentence of type *S* is being uttered by a speaker, an interpreter can use her knowledge about sentence type *S* to predict when it ends, making it possible to take turns with no gaps. However, Sacks et al. do not specify what kinds of turn-constructional units distinguish one type of utterance — or multimodal act for that matter — from another. If we assume that a listener is continuously looking for clues to classify each utterance we might conclude that the only features that matter are present in the stream of the audio signal. But this would be a mistake: Anyone who ignored all but the audio signal in a multimodal interaction would be throwing away a wealth of information that can be gleaned from the utterer's behavior pertaining to both the content and the process of the dialogue. We can be pretty certain that the 'evidence' people use to classify turn segments includes a number of sources, all the way from gaze to facial gesture to body stance (Taylor & Cameron 1987, Goodwin 1981).

This leaves us with two problems. From a descriptive point of view, the idea of turn-constructional units may be valid, but it says nothing about the way people actually recognize these units. Furthermore, even when a unit is recognized, its length would not be completely predictable, and the task of prediction becomes clearly also

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

a perceptual task: paying attention to cues that signify the end of the unit.[1] The second problem is that the turn-constructional units that Sacks et al. propose are purely based on the audio stream produced. The mistake is to think only of how utterances relate to the turn, when we really need a theory of how *communicative acts* can be constructed in turns.

What is needed is a mechanism that allows sentential, clausal, phrasal and lexical features — as well as all other types of speaker behaviors indicative of dialogue state — to be recognized in real-time and integrated with a discourse participant's actions. Furthermore, Sacks et al.'s (1974) model does not take into account the internal state of cognitive processing of the participants, which clearly also affects the way they respond to cues in the dialogue. Whether we call this yet another class of turn-constructional units or not is beside the point: We will lump all of these ideas together into a bag called *context*. In the section on the YTTM architecture we will present an approach for doing this based on a version of the blackboard architecture (Nii 1989, Selfridge 1959). In the process we will define operationally what 'context' means in this context.

In what seems to be an incompatible approach to that of Sacks' et al., Duncan (1972) proposed the existence of "cues" for turn signalling. Such cues are generated by interlocutors for the purpose of "signaling" to each other the state of the dialogue, such as whether they want the other to take the turn, whether they want to keep the turn, etc. The claim here is that Duncan's cues are simply the features missing from Sacks et al.'s model — the features that conversants use to identify the turn-constructural units, and their boundaries. These, naturally, vary between cultures and individuals — which is why we find it more difficult to interact smoothly with strangers than with people we know. What, exactly, the set of such 'cues' consist of is not easy to determine, and is bound to vary on an individual basis. The best we might do is to create a collection of what may *look* like 'typical' patterns (cues) for a given group of individuals, families, or cultures. A more important first step though, is to identify which kinds of data and channels carry information relevant to dialogue, and to propose mental mechanisms that might be at work for producing — and especially perceiving — such information. This will be our focus in section 4., "Principles of the YTTM".

### *2.1. Back-Channel Feedback*

No treatise on turn-taking is complete without a discussion of back-channel feedback (Yngve 1970). Face-to-face interaction quickly breaks down if communication can only happen at or above the turn level (Nespolous & Lecours 1986) — there needs to be a two-way incremental exchange of information within the turn. Part of the task for a listener is to make sure that the other party knows that she is paying attention, and indicate that she is at the same state in the conversation. This is done mainly in the back channel (Yngve 1970). Back channel feedback is in effect information exchange that supports the interaction itself and helps move it along (McNeill 1992, Goodwin 1981). In English speaking countries it includes using

1. Dead-reckoning — the act of committing to a course of action ahead of time and then blindly executing it — would be another way to solve this problem. More on this in section 3.1, "Achieving Seamlessness Through Perceptual Anticipation".

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

paraverbals such as “m-hm,” “aha,” etc., indicating confusion, expressing feelings at given points (by facial gesture, laughter, etc.), and indicating attentional focus. The absence of such regulatory gestures from a listener may disrupt the discourse (Dahan, as referenced in Nespolous & Lecours 1986).[2]

While it may rightfully be argued that overlapping talk in the main communication channel is counter-productive because it interferes with the flow of a conversation (Sacks 1992), co-occurring speech in the paraverbal channel does not (Yngve 1970), unless it is misclassified (by the speaker) as being part of the class of accepted turn-transition cues. One rule of thumb definition of back-channel feedback then is that it is the ongoing (communicative) behavior of a dialogue participant that does not change who is in control of the dialogue at the moment. So, whether something "is" back-channel feedback is not based on what an act looks like (morphology) or who has the turn, because what may be intended as back-channel feedback may turn into an interruption if the speaker misinterprets it. For the perceiver of such behavior this is therefore an issue of ongoing *functional classification*. Functional classification is executed continuously by all participants during dialogue. We will discuss functional classification in section 4.4.1, "Functional Analysis: Characterizing the Broad Strokes ’First’".

Back-channel feedback is modeled in the YTTM as resulting from two very different sources: The processing of the *content* of dialogue, e.g. smiling when we find funny the content of what the speaker is saying, or from the mental machinery *orchestrating* the dialogue, e.g. when we look at the speaker to show we are paying attention. We will look at this claim in section 4.3, "Separating Interaction Control from Content Generation & Delivery".

### *2.2. Embodiment*

At least two types of spatial constraints are critical to situated conversation: the *location* and *orientation* of conversants to each other and surroundings, referred to here as *positional elements* and *directional elements*, respectively. The position of conversational participants has implications for spatial reference: glances, pointing gestures and direction-giving head nods will be done differently (varying morphology) depending on where the speaker and listener are positioned in space. The display of visual cues such as facial gesture is bound to a specific location, i.e. the participants’ faces. A number of turn-taking signals rely on participant location and facial cues (Duncan 1972), and back-channel feedback is often given through the face (Goodwin 1981). Manual gesture is usually done in the area right in front of the gesturer’s body (McNeill 1992), and a perceiver needs to be able to locate these in space. Gaze is often used to reference this space (Goodwin 1986), and can be indicative of the kind of gesture being made (McNeill 1992, Goodwin 1981); gesturers tend to look at their own iconic gestures.

Directional elements have to do with how the participants are turned relative to each other, how various body parts are oriented, and how this changes over the

2. Nespolous & Lecours (1986, page 61) say: “... Dahan (see ref., op. cit.) convincingly demonstrated that the absence of regulatory gestures in the behavior of the listener could lead the speaker to interrupt his speech or to produce incoherent discourse.”

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

course of the interaction. When talking face-to-face, most people prefer to orient their bodies approximately 90° to each other (Sommer 1959). Turning your head away right after your partner finishes speaking can indicate that you think he's done and that you are now preparing a response (Goodwin 1981, 1986). All these features require spatial computation of both participants. We will look at how some of these are implemented in section 5.3.2, "Multimodal Integrators (Table 4)".

## 3. THE LEAP TO GENERATION

From the discussion so far it is clear that a step-lock "transmitter/receiver" model will not be sufficient when imparting multimodal interaction to the computer. Back-channel feedback, interruptions, real-time construction, unforeseen events all hint at a much more complex, dynamic system in which multiple states and events serve to provide a rich context for the participants' mental processing.

As numerous researchers have shown (Walker & Whittaker 1990, Goodwin 1986, Sacks et al. 1974), turn-taking defines the two main roles of conversants, often referred to as 'speaker' and 'listener'. These terms are too limiting to describe the roles of dialogue participants. We will use the terms *content presenter* and *content interpreter* to refer to these roles, respectively. Firstly, this separates the roles of communicating parties from the modes they use for the communication. Secondly, it avoids confusion between turn mechanisms and the act of speaking (when giving back-channel feedback "listeners" can speak without taking the turn). The relation between content presentation and turns is that, generally speaking, one needs to have the turn to present content. In section 4.2, "Presentation and Interpretation: Role-Based Processing", we will explore how each role calls for its own cognition repertoire.

The model of turn-taking advanced by Sacks et al. (1974) is a good descriptive model of turn-taking. A generative model has to go beyond describing surface phenomena in dialogue, however, it has to re-create the surface events observed through a performance model. Given the amount of sensory data and motor control needed for this to happen, the challenge in turn-taking is how to make context-sensitive mechanisms without having to connect everything to everything else. Modularization of the computational processes must be a significant part of a successful model.

The hypotheses on which the Ymir Turn-Taking Model is based create a necessary bridge between a backdrop of relatively coarse-grain studies of turn-taking and dialogue from the psychological and linguistic literature on the one hand, and, on the other, a computational architecture that dictates mental functioning at much smaller levels of granularity. Although the work described can be classified solely under the rubric of artificial intelligence it is inspired by cognitive models, and the following hypotheses are provided to enable future assessment of the model's psychological plausibility. (This task remains outside of our scope here.) The hypotheses represent a theoretic-complete foundation for the creation of YTTM and stand as a generalization of the computational implementation presented in the second half of the chapter.

### *3.1. Achieving Seamlessness Through Perceptual Anticipation*

People do not particularly notice the mechanisms by which they take turns speaking. They do not have to pay much attention to how they interweave glances, content,

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

gestures, body movements, etc., seamlessly during conversation. How is this possible? It might be argued that after years of participating in dialogue almost every day, people achieve turn-taking using dead-reckoning.[3] Perhaps they committ to taking the turn several hundreds of milliseconds in advance, and then blindly stick to it, from that moment on running ballistic. This is a valid hypothesis that deserves consideration.

In the context of face-to-face dialogue, which can go on for hours, 100 msec is not a very long time to be spending between turns. Yet, as Goodwin (1981) and other have shown, turn transitions of 100 msec or less, even ones with no pauses in the speech channel, happen frequently in spoken dialogue. Given the complexity of turn-taking, dead-reckoning a long time ahead would greatly increase the likelihood of erroneous turn transitions. If it exists, it is therefore likely to span only a short interval. Of course both parties in a dialogue have a choice reaction time of 100 msec — a speaker can decide to continue an utterance on a whim, destroying any conclusions about a valid turn-transition that the other party may have predicted equally many milliseconds ago. So even for a relatively short turn, lasting, say, 3-4 seconds, a valid turn transition predicted by the interpreter 400-500 msec in advance can be destroyed 200 msec later by the presenter's decision to continue speaking at the end of that segment, leading unavoidably to overlapping speech. (In any goal-directed conversation overlapping speech is considered non-cooperative (Grice 1989) and is thus to be avoided.) An interpreter who has predicted a turn-transition by dead-reckoning will thus also have to monitor, during the exact moment of turn transition, whether this prediction was erroneous. Given the speed of simple reaction, and the price of erroneous dead-reckoning resulting in unwanted speech overlaps and pauses, it is unlikely that any dead-reckoning turn-taking scheme would span longer than 100-200 msec into the future. Moreover, any such dead-reckoning behavior would become useless in interaction with a non-native speaker with a different rhythm, syntax structure and intonation. Whichever way we look at it, no matter whether some amount of dead-reckoning is happening or not in native-speaker turn-taking, we still end up with the conclusion that there has to be ongoing perceptual monitoring during transitions. Moreover, because of the unpredictability of turn-taking, the extent of dead-reckoning is likely to be very short, possibly close to being negligible. The assumption here is that the role of open-loop — ballistic — action in turn-taking is likely to be very small, and can for all practical purposes be ignored.

Both Sacks et al.'s (1974) and Duncan's (1972) work provide evidence that the difficulty of modeling turn-taking lies first and foremost in perception, because a participant has to infer what constitutes a valid turn-giving "signal" solely from perceptual information. Moreover, no decision can be made without the proper percep-

3. Dead-reckoning here means committing to future actions before they are to be executed. The shortest human reaction time is approximately 100 msec (Boff et al. 1986). This is so-called *simple reaction time;* a boolean event where a person only has to choose between action or in-action depending on an external, pre-determined event — for example pressing a button when a light comes on. This is an appropriate measure to use here since it represents the lower limits of what can be achieved via the voluntarily controlled perceive-act cycle, and would be expected for a highly practiced skill like turn-taking. To be considered dead-reckoning, the interval from commitment to execution would then be longer than 100 msec.

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

tual data to base it on. There is a type of prediction besides dead-reckoning which may exist in turn-taking.[4] This kind can best be thought of as expectation or *anticipation*. We propose this as the first step towards incorporating prediction into turn-taking. This kind of prediction only affects one of the four elements of mental processing (i.e. perception — the others in our classification are cognition, decision and action), and can therefore be considered the weakest form of prediction. We hypothesize that

**{H1}** *Opportunities for turn-transitions are identified using a mechanism of anticipatory perceptual processing.*

Given a perceived dialogue progression P, participant A will be anticipating turn-transition T. T has associated with it a set of perceivable behavioral features F (some of which may be the traditionally called "turn signals"). Participant A will focus attention towards the occurrence of F. This he does by priming his perceptual system, thus engaging in *anticipatory perceptual processing*. We will discuss the implications of this in section 4.2, "Presentation and Interpretation: Role-Based Processing".

### *3.2. Temporal Constraints in Face-to-Face Dialogue*

Face-to-face interaction is unique because it contains processes that span as much as 5 orders of magnitude of execution time, from about 100 ms (gaze, blinks), to minutes and hours (Thórisson 1999). Another way to say this is that co-temporal, co-spatial discourse contains rapid responses and more reflective ones interwoven in a complex pattern dictated by social convention. The structure of dialogue requires that participants agree on a common speed of exchange (Goodwin 1981). If the rhythm of an interaction is violated, it is expected that the violating participant make this clear to others, at the right moment, so that they can adjust to the change. For example, if a story teller suddenly forgets what comes next in her story and has to pause, she is sure to indicate this to her audience by saying something like "ahhh" or even the more explicit "Hmm, I can't seem to remember what happened next". This common speed sets an upper limit to the amount of time participants can allocate to thinking about the dialogue's form, content, and to forming responses. Newell's (1990) classification of time scales in human mental processing proposes a "cognitive band" which spans three orders of magnitude of time, from 100 ms to tens of seconds. A lot of mental processing during dialogue happens in this band, yet very few have looked at the real-time performance aspects of mental processing.

The issue of real-time is not only about speed but about *proper mental load-balancing:* ensuring that the most important processes are always run. If the story teller fails to explain the pause in her story telling, unwanted interruptions are bound to happen; the processes supporting the *delivery* of the explanation represent a higher priority than the production of the story itself. Based on Dodhiawala's (1989) principles of real-time performance, we can identify the following four aspects of real-time performance:

4. It is not clear whether by 'prediction' Sacks et al. (1974) mean ballistic action, prediction of a turn transition point in the future, or to some kind of anticipation.

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

1. Responsiveness: The system's (in this case dialog participant's) ability to stay alert to, and respond to, incoming information.
2. Timeliness: The system's ability to manage and meet deadlines.
3. Graceful adaptation: The system's ability to (re)set task priorities in light of changes in resources or workload, and to rearrange tasks and replan when problems arise, e.g. in light of missed deadlines.
4. Speed.

The first three are about load-balancing; the fourth requires a reference — speed compared to what? That 'what' is the real world. Following the Model Human Processor model of cognitive processing (Card et al. 1983), we can look at speed of cognition at three stages: (1) Speed of *perceptual analysis*, (2) speed of *decision*, and (3) speed of *action composition*.[5] What really matters, of course, isn't the speed of any one of these stages but that their combined output, e.g. bending down to avoid being hit in the head, are composed and executed fast enough to get the head out of the way. Of equal importance, the system has to know that at that point in time this is the most important thing to compute, both in perception and action. To achieve this feat the system has to be capable of simultaneous production of multimodal action and multimodal perception, in other words, it has to be capable of parallel processing.

To coordinate events at multiple timescales we draw on the modularization approach of behavior-based AI and hypthesize that

**{H2}** *The seamlessness observed in real-time turn-taking comes from the co-operation of multiple processes with different (a) update frequencies, (b) target perception-action loop times, and (c) speed-accuracy tradeoffs.*

The above real-time factors, combined with hypothesis {H2}, have lead to a modularization of YTTM that separates processes according to the urgency of their processing, the *layered feedback loop* model. Processes are load-balanced by different priority levels, and some processes can momentarily suspend processes running at other priorities. Low- and high-priority processes run in parallel, e.g. detection of interrupts, a high-frequency update process, runs in parallel with constructing narrative, a lower-frequency update process, and when the two produce results they may interfere or combine in the output behavior.

But if we have a distributed system where processes with high update rates are — by design — the first to catch a subtle, high-frequency 'interruption cue' from an interpreter, how do they communicate this fact to other processes, e.g. those in charge of telling a story? How does the goal of ignoring a presenter turn off the (modularized) ability to detect and respond to pauses? For this the processes need to have bi-directional control of each other's processing. This leads us to hypothesize that

5. Action composition is not the same as action execution (Thórisson 1997). Here we assume that execution — i.e. movement — characteristics (either simulated in graphics or actual robotics) matches roughly those of the human body.

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

**{H3}** *To support coherent output generation in a modular, distributed system, processes with different perception-action loop times are sensitive to a particular* subset *of the total set of contextual cues available, which includes perceptual data produced from input to the sensors, as well as processing states and partial output of other system elements.*

To take some (simplified) examples, the process of classifying a presenter's silence as a "hesitation", rather than a "turn signal", can rely on the context provided by other parts of the interpreter's mental processing, namely those that classify the presenter's bodily stance as *pensive*, sentence completion and semantic content as *incomplete*, and her gaze as *distracted*, all of which are bottom-up processes which support a "hesitation" theory. Other contextual cues that may influence the interpreter's behavior, given a choice between classifying behavior into a "hesitation" or a "turn signal", could be his complete lack of understanding, or perhaps his lack of something to say, which may simply result in him not taking the turn.

Now, let's look further at the layered feedback loop model.

## 4. PRINCIPLES OF THE YTTM

### *4.1. Multimodal Dialogue as Layered Feedback Loops*

The YTTM follows a layered feedback loop model. The layers in this model are both *descriptive* — they are based on time-scales of actions found in face-to-face dialogue — and *prescriptive* — they specify the prioritization, or load-balancing, of computation. At each level in this model various sensory and action processes are running, primarily providing services to the level below and/or above. The highest priority is concerned with behaviors that have perceive-act cycles shorter than 1 second, typically less than 500 msec. Highly reactive actions, like looking away when you believe it's your turn to speak (Goodwin 1981) or gazing at objects mentioned to you by the presenter (Kahneman 1973), belong in this Reactive Layer. The Process Control Layer includes mental activity that relates to what we would typically categorize as the willful control of the interaction itself: starts and stops, interrupts, recognizing breakdowns, in short, everything that has to do with the *process of the dialogue* (sometimes called 'task level'). The perceive-act cycle of such events typically lie between a half and 2 seconds. Together these two layers contain the mechanisms of dialogue management, or psychosocial dialogue skills.

The lowest-priority layer, the Content Layer, is where the content or "topic" of the conversation is processed, e.g. navigating a rocket ship or cutting grass. Following hypothesis {3} (and {4} — see below), we can treat the topic knowledge residing in this layer as a black box: Its input is provided by perceptual processing in the whole system; its output is speech and multimodal behavior related to the topic of the dialogue, and actions relating to the manipulation of the topic domain.

The set of perception and decision processes actively at work in each of the three levels at any point in time is determined by several factors, one being the role of the participant at that point in time (content presenter or content interpreter). Another factor is the perception-action loop time required for the system to behave correctly. A third factor is incremental processing; multimodal, real-time interpretation is not

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

*Table 1. The table shows which mental processes belong in each of the three priority layers of the YTTM, for content presenter and content interpreter. All tasks run in parallel, but those in the Process Control and Reactive layers have higher priority than those in the Content Layer, both in terms of processing and of execution of actions resulting from the processing. ("Process" in "Process Control" refers to the process of the dialogue, i.e. the* interaction*.) This table links Figure 1, which shows the main states f the turn-taking mechansim, and Figure 2, which shows feedback loops and target loop times for each layer.*

| **DIALOGUE ROLE** / **LAYER** | **CONTENT PRESENTER** ("speaker") **PERCEPTION *(p)* & MOTOR *(m)* PROCESSES** | **CONTENT INTERPRETER** ("listener") **PERCEPTION *(p)* & MOTOR *(m)* PROCESSES** |
|---|---|---|
| CONTENT LAYER (low priority) | Analyze interpreter's content reception *(p)*<br>Present content *(m)* | Interpret content *(p)*<br>Convey status of content interpretation (content-related back channel feedback) *(m)* |
| PROCESS CONTROL LAYER (medium priority) | Analyze interpreter's process control *(p)*<br>Control process *(m)* | Interpret dialogue structure *(p)*<br>Convey status of dialog structure interpretation *(m)*<br>Control process *(m)* |
| REACTIVE LAYER (high priority) | Broad-stroke functional analysis *(p)*<br>Reactive behaviors *(m)* | Broad-stroke functional analysis *(p)*<br>Process-related back-channel feedback *(m)* |

done "batch-style": There are no points in a face-to-face interaction where a full multimodal act or a whole sentence is output by one participant before being received by another and interpreted as a whole. Interpretation of multimodal input happens in parallel with multimodal output generation, continuously produced by processes running in parallel at each level.

### *4.2. Presentation and Interpretation: Role-Based Processing*

Following our discussion about prediction leading up to hypothesis {H1}, we define two different sets of processes, both of which include perceptual, decision and motor tasks, that participants in a dialogue switch between depending on whether they are in the content interpreter or content presenter role.[6] We call this *role-based processing*, and it is really a kind of context sensitivity. Thus, for the period that person *A* takes the role of interpreter, one can expect him to be engaged in a set of mental activities that are different from those he is engaged in when in the role of presenter. To take an example, Goodwin & Goodwin (1986) discuss the activity of searching for a word and how this can be a cooperative activity. A content presenter may indi-

6. This does not mean that there are no processes that run during both states. Indeed, a large portion of typical perceptual, decision and motor tasks, such as glancing at the other party, smiling, etc., may run in both states.

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

cate to her interpreter, using gaze and body language, that she is looking for a word. The interpreter will offer to assist in the search by interjecting plausible words. Although the process is cooperative, it is the presenter who has the turn, and thereby the power to accept or reject the interpreter's suggestions (even in cases where the interpreter knows exactly what the presenter wants to convey). It is not only the relevant behavioral repertoire (visible actions) that is different for each role, but also the demands on the two participant's perceptual and decision-making systems. The roles can be thought of almost as roles in an improvisational play; they are part of the same plot but the rules for each actor's character are very different. The complication is of course that every now and then the actors switch roles according to very complex rules — they take turns.

The roles of content presenter and content interpreter are subjective: For turn-taking to work properly the concept that one participant has the turn has to be represented in the minds of all participants as a mutual belief that this participant has the turn. (They also have to share the goal of achieving efficient and cooperative communication (Grice 1989).) Moreover, their understanding of what represents appropriate moments for taking and giving turns has to also be mutually shared. Cultural differences are the clearest demonstration of how this must be so.

The concept of role-based processing can be taken one step further. According to hypothesis {H1} the process of turn-taking relies on anticipatory perceptual processing; this principle can be extended to perception *during* particular turn states. Hence, back-channel feedback, clarifications during turns, complementary gestures, additional facial expressions etc., can all be generated if needed, based on anticipatory perception. Thus, for a presenter Beth and interpreter Alan, Beth monitors Alans's behaviors (via anticipatory perception) for cues that reveal his understanding of what she is saying; Alan monitors the content of what Beth is presenting, and, via anticipatory perception, identifies places where back-channel feedback, interruptions and the like are appropriate (Table 1).

### *4.3. Separating Interaction Control from Content Generation & Delivery*

One of the questions we need to answer is how the topic of the dialogue relates to the processes that control the timing and style of interaction. The representation of a topic domain is in and of itself a complex matter, and no computer model has succeeded in replicating the detailed knowledge a human has for a given domain of expertise (cf. Lenat 1995). It is unlikely that the rules for how a topic is talked about are replicated for each domain separately; it is more likely that general knowledge about how to convey information is stored once, to be reused for any topic that may be discussed. It can even be argued that this knowledge is a topic in and of itself. YTTM theory splits topic knowledge from dialogue knowledge into separate systems that talk to each other via a well-defined, small set of messages; when you are asked where you were yesterday the knowledge you use to hesitate, look up, roll your eyes, and say "hmm" is controlled by a general mechanism, separate from the processes required to fetch the piece of information required to answer the question proper. So *interaction (process) control* is separate from *content interpretation*. Turn-taking control takes into account the processing status of the domain knowledge in the same way it takes into account any other context. In this case, if it takes longer than typical for the topic knowledge system to process the input the turn-taking mechanism may decide to comment on this fact in one way or another (e.g. look-

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

ing up, saying "Now, let me think..." or by direct semantic information like "Wow, that's a tough question..." — in the latter case the topic knowledge processes would communicate meta-information, i.e. that the question is 'tough', to the turn system.)

Following up on hypothesis {H3}, which claims that various modules in a distributed system are sensitive to various subsets of contextual cues, we present here the *topic-independence* hypothesis of turn-taking:

**{H4}** *Processing related to turn-taking can be separated from processing of content (i.e. topic) via a finite set of interaction primitives.*

This hypothesis has been informally incorporated by others (see e.g. Cahn & Brennan, 1999). In the YTTM prototype the primitives are implemented as a set of messages, exchanged via blackboards, as detailed below (see Figure 3).

### *4.4. Modeling Multimodal Perception*

As a presenter, one's perceptual system is preoccupied with monitoring the progress of one's production of narrative output. But following the proposal of role-based processing, of even higher priority is distinguishing between acts of the interpreter that are insignificant to the dialogue (such as the listener casually adjusting his hair), and those that constitute communicative actions, such as a wish to interrupt. The latter behaviors take priority because they may directly affect turn-taking, and thus the course of the narration. The interpreter's top perceptual priority revolves around interpreting what the presenter is saying and making sure the presenter knows that he is following her story, giving indications of the status of his understanding processes, and interrupting when problems arise.

This emphasis on the presenter-interpreter distinction has the important result of placing the tracking of dialogue state in the driver seat among the sensory activities. It is a process that happens at the decisecond level of granularity and is highly temporally constrained. This is summarized in the following hypothesis:

**{H5}** *Perceptual and decision processes dedicated to tracking dialogue state have the highest priority of all mental activities related to communication.*

In other words, the processes with the highest priority in our system are perceptual processes that produce the data necessary to estimate dialogue state reliably, and the decisions related to these, which change the (mental representation of) dialogue state from one to the next. Why must this be so? Attention is a limited resource and the system has to continuously make trade-offs in processing: The faster it should be responding to a turn-taking cue, the more reliably the cue has to be detected for the interaction quality not to be degraded (increased speed means fewer "sample points" to base the decision on). Most turn-taking cues are multifaceted, involving some combination of many features such as intonation combined with a pause combined with a particular state of content production combined with a particular eye movement. The faster a decision is made in response to any perceptual cue the lower the probability that it actually represents a turn-taking cue, because the total set of events that have some *characteristics* of turn-taking cues far outnumbers that of *actual* cues, and, depending on how fast the needed cues become available, rash decisions will thus often lead to wrong decisions.

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

*4.4.1. Functional Analysis: Characterizing the Broad Strokes 'First'*

Any system that works under time-constraints and uncertainty is forced to always look at the most important data first, since time-pressure may prevent scrutiny of detail. So what constitutes the most general information for a multimodal turn-taking system? How and where do we look for it? We claim that the most significant information in conversation is the *function of discoursal actions*, and people look for it using a system of specialized processes that have a relatively high speed/accuracy ratio. These processes look at the broad strokes of the other participant's behavior 'first' — the word is in quotes because it does not imply sequential processing (all processing is parallel); the principle refers to the *priority* that functional analysis has in multimodal perception. To illustrate further, let's return to the story from the introduction.

> Alan is telling Beth an exciting story about his vacation in Nice. He presents his story through gesture and speech. Then Beth's arm starts moving and her neck stiffens. Beth's gesture is not yet recognizable as a communicative action. The movement grabs Alan's attention and keys his perceptual system in to classify the motion further, because in the next half second Alan has to make some decisions about Beth's arm movement that may affect his own behavior. Let's follow Alan's perceptual anticipation for the next 460 milliseconds . . .
>
> ***t-minus-460 msec***
>
> Beth's arm moves. Alan has to decide whether:
>
> 1: Beth's arm movement constitutes a communicative gesture, and if so,
>
> 2: what kind of gesture.
>
> Because Alan is presenting, and thus has the turn, he's reluctant to let himself be interrupted.
>
> ***t-minus-350 msec***
>
> 3: Based on Beth's expression so far, he's persuaded to pause his presentation at t-minus 350 ms (human choice reaction time is ~100 ms (Boff et al. 1986]).
>
> ***t-minus-250 msec***
>
> 4: Using Beth's gaze and the state of the dialogue, Alan decides that he will try to figure out what Beth's multimodal actions mean (i.e. what kinds of phenomena in Beth's mind does her current behavior correlate with — or serve as index of), and thus delay his presentation further.
>
> 5: Alan figures out that Beth has started making a deictic gesture (he's not sure, but "it's worth a glance") so, based on the direction of Beth's gaze, at
>
> ***t-minus-150 msec***
>
> 6: Alan looks over in the direction in which Beth is roughly pointing (where he'll see an elephant).
>
> 7: Beth's gesture becomes fully-fledged, easily recognizable deictic gesture.
>
> 8: Alan should have delayed looking. He had just reached out for his glass of beer, and now, at ***t-minus-0 milliseconds***, he sees the elephant Beth is pointing at... and knocks his beer over.

In order to conduct efficient turn-taking, Alan decided, based on the potential communicative function of Beth's actions, to pause his production of content and succumb to the turn-taking rule which states that generally a wish to interrupt should be acknowledged. Notice that had Alan continued to speak, it would either have been because he chose to do so, or that he had failed to see what Beth was doing. In other words, Alan's acknowledgment of Beth's interruption (by stopping to speak) was not delayed because he was speaking; the only way it could have been delayed was by

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

Alan wilfully pausing. The example illustrates that the highest-priority interpretation of a dialogue participant's behavior should not — in fact *could* not — primarily be concerned with content, for example which lexical elements can be best mapped onto a presenter's utterance, or whether an utterance at any point in time is grammatically correct, it has to be concerned with distinctions that determine broad functional strokes of behavior, i.e. extracting the features that make the major distinctions of the dialogue, *communicative* versus *non-communicative*. Computing the function of a person's behavior to mean that the person is addressing you is a necessary precursor for you to start listening; interpreting a movement's function to be a deictic one will have to happen before you can look in the direction of the pointing arm/hand/finger (or gaze) to find the referent of the action. Depending on the state of the dialogue this could either be a wish to interrupt, as in the example above, or part of conveying content. Thus, a gesture might reveal the meaning of a seemingly meaningless utterance; a nod might indicate the direction the interpreter should look for grasping the meaning of the presentation; intonation might indicate sarcasm, etc. These examples constitute broad strokes — high-level function — of behavior. The *broad-strokes-first hypothesis* postulates that

**{H6}** *In real-time communication, analysis and interpretation of broad-stroke communicative function takes higher priority than content analysis and interpretation.*

Analysis of broad-stroke function is not the same as top-down analysis; our mental processes can use evidence from bottom-up *and* top-down to find broad stroke functions. Broad-stroke functions have to be higher priority because they provide the context for the communication itself, and by extension the presentation. On the feedback generation side, a listener's behavior of looking in the pointed direction is a sign to the presenter that he knows that her gesture is a deictic one, and that he has correctly extracted the relevant direction from the way her arm/hand/finger are spatially arranged. The gaze behavior resulting from correct functional analysis serves double duty as direct feedback, and constitutes therefore efficient process control. Thus, analysis of the contextual function of a presenter's actions and control of the process of dialogue are intimately linked through functional analysis. Furthermore, the information necessary for correct and efficient content analysis is often the necessary information for providing correct and efficient multimodal feedback behavior (Table 1).

### *4.4.2. Combining Multimodal Perceptual Information*

Given the multiple sources of information in multimodal conversation, we are led to the following line of reasoning: A large set of aggregated cues from multiple sources of information must be more reliable than a smaller set of cues from a single mode. Thus, to achieve the most efficient trade-off between speed and accuracy of turn taking, perception related to turn-taking can be expected to draw on cues from any number of modes and sources, as long as they are informative. Therefore:

**{H7}** *Reactive behaviors are based on data produced by highly opportunistic processing.*

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

How would such an opportunistic perceptual system combine 'evidence' from multiple sources and modes? The hypothesis we build on here is:

**{H8}** *Separate features and cues extracted, by perceptual processes of a dialogue participant A, from a particular multimodal action by dialogue partner B, are logically combined (in the mathematical sense of the word) by other perception processes in the mind of participant A, to support generation of appropriate behavior during the interaction.*

Thus, our first approximation to this question is that the process of multimodal integration is based on boolean logic gates (cf. Duncan 1972). This has certain advantages, namely, it is easier to compute and to track the interaction of multiple boolean variables than interaction among equally many scalars, making this probably the simplest possible choice for how to combine data from multiple modes.

To relate this back to the issue of the speed/accuracy trade-off in perception and action, according to these hypotheses, the more features and modes available to someone who is assessing the dialogue (in a single perceptual analysis of turn-transitions and turn-state) the *higher the accuracy* of that perceptual assessment. In other words, estimations on part of the dialogue participants that the dialogue is in a given state, or should change to a new state, will be more accurate with an increased number of modes. This may in fact be one of the reasons why we often prefer to meet face-to-face, rather than simply talking over the phone (or sending e-mail). Paradoxically, the speed of the analysis will not be affected by the presence of more data because it is already a massively parallel process. However, increased perceptual reliability may affect the speed at which the perceiver will *act on* the extracted features. Thus, upon interpreting the multimodal act "He went [deictic manual gesture & gaze] that way," an interpreter may look sooner in the relevant direction if the manual pointing gesture is present, than if the only indication of direction is the presenter's gaze, since a manual deictic gesture is a more reliable indicator of direction than gaze alone. More efficient, speedier turn-taking can thus happen in a face-to-face meeting than any other kind. We give a working example of this in Section 5.3.1.

### *4.5. Decisions*

"Decision" is the event where we turn a perception into a potential action — it's the switch, so to speak, for moving the body of the conversant. If you make a decision half-way (or 3/4th way or 12/27th way) you are not making a decision, you are in fact the very definition of someone who *can't* make a decision. A decision is either made or not made, a crisp event. Turn-taking decisions in YTTM are mainly made about *turn-transitions, provision of back-channel feedback*, and the *timing of the production of content* (i.e. when we start to say what we want to say about the content of the dialogue). Decisions in YTTM aligns with hypothesis {H8}:

**{H9}** *A decision is based on the boolean combination of perceptual features.*

To meet our real-time demands, any *motor events* generated as a result of a decision made in the Reactive Layer has the highest priority for execution; the Process Control Layer has second priority, and the Content Layer the lowest priority. Now,

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

before turning to the system's implementation, let's look briefly at movement generation.

### *4.6. Production of Motor Events*

A significant part of perceptual data related to the turn-taking process lead to decisions that result in actions. A decision of a listener to interrupt the presenter may result in a series of motor events that, in the dialogue participants' culture, is a well-recognized cue for wanting to communicate something. A decision is always discrete, but in the Ymir Turn-Taking Model a decision to move e.g. a body part may or may not result in the movement actually happening: For all decisions the last stop before they become movement is controlled by a relatively monolithic action scheduler (Thórisson 1996, 1997). The action scheduler allows a decision to be cancelled up until 100 ms before its execution by committing to it only at execution time (rather than at decision time), giving YTTM the same choice reaction time found in human behavior (Boff et al. 1986).

The process of turning a (relatively) high-level decision like "interrupt the content presenter" into an acceptable series of motor events is a complex one. Canned responses are a simplification that will not work if our goal is to create a complete generative model of turn-taking; for that the model will have to be able to produce overlapping and interwoven multimodal behaviors. The YTTM uses a multimodal Motor Lexicon for turning a decision (which is a simple kind of a goal) into a motor sequence. The Motor Lexicon is a tree where the nodes are decision names such as "interruptContentPresenter" and the leaves are particular motor sequences that can signify an interrupt in the interpreter's culture. Between a node and the leaves we may have multiple branching; each branching is a named decision/goal node. At each node the motor system can choose between alternative options to achieve that decision/goal. This is where the power of the motor system comes from: Each choice can be compared to the current state of the system and agent's body, and chosen based on "goodness of fit" for that particular moment in the dialogue. For example, the decision/goal "interruptContentPresenter" may branch into the three options:

1. [raise-arm, raise-index-finger, look-at-presenter]
2. [raise-eyebrows, raise-arm, open-mouth]
3. [say-ahhh, look-at-presenter]

Each of the constituents in these three options may in turn have one or more options. If both hands of the interpreter are busy, the last option of producing speech and gaze for interrupting the presenter will be chosen by the system, since both of the other options require the arms to move. Turning the first element of the last option ("say-ahh") into motor events would result in a structure of the form:

[First: Open-mouth[$x$,$d$], Second: Produce-sound [$p$, $v$, $i$, $d$]]

Variable $x$ contains a number signifying how much to open the mouth, variable $d$ tells the system how fast to move, variable $p$ contains the phoneme sequence "ahhh", variable $v$ the volume of the utterance, and $i$ the intonation pattern to use. Further details on this action scheduling scheme can be found in (Thórisson 1997).

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

*Table 2. Prototype YTTM Covert State Deciders change turn-taking state in a real-time human-humanoid interaction. (All are ACTIVE-DURING-STATE: Dialog-On.) Rules are listed in a LISP-like syntax. For items [a]-[g] and a discussion of the rules, see Section 5.3.1.*

| | |
|---|---|
| **STATE:** *I-Have-Turn*<br>TRANSITION TO STATE *I-Give-Turn* IFF:<br>(OR (AND<br>(I-have-something-to-say = F)<br>(I-have-something-to-do = F))<br>[a](AND (Im-executing-topic-realworld-task = F)[b]<br>(Im-executing-communicative-act = F)[b]<br>(Other-wants-turn = T ))) | **1** |
| **STATE:** *I-Give-Turn*<br>TRANSITION TO STATE *Other-Has-Turn* IFF:<br>(OR (Other-accepts-turn = T)[c]<br>(Other-wants-turn = T)<br>(AND (Other-is-paying-general-attention = T)<br>(Im-executing-topic-realworld-task = F))[b]<br>(Im-executing-communicative-act = F)))[b] | **2** |
| **STATE:** *I-Give-Turn*<br>TRANSITION TO STATE *I-Have-Turn* IFF:<br>(Other-accepts-turn = F) | **3** |
| **STATE:** *Other-Has-Turn*<br>TRANSITION TO STATE *I-Take-Turn* IFF:<br>(AND (***Time-since*** 'Other-is-presenting > 50 msec)[d]<br>(Other-produced-complete-utterance = T)<br>(Other-is-giving-turn = T)<br>(Other-is-taking-turn = F))[e] | **4** |
| **STATE:** *Other-Has-Turn*<br>TRANSITION TO STATE *I-Take-Turn* IFF:<br>(AND (***Time-since*** 'Other-is-presenting > 70 msec)[f]<br>(Other-is-giving-turn = T)<br>(Other-is-taking-turn = F)<br>[g](OR (Others-intonation-going-up = T)<br>(Others-intonation-going-down =T))) | **5** |
| **STATE:** *Other-Has-Turn*<br>TRANSITION TO STATE *I-Take-Turn* IFF:<br>(AND (***Time-since*** 'Other-is-presenting > 120 msec)[h]<br>(Other-is-giving-turn = T)<br>(Other-is-taking-turn = F)) | **6** |
| **STATE:** *I-Take-Turn*<br>TRANSITION TO STATE *I-Have-Turn* IFF:<br>(AND (Other-is-paying-general-attention = T)<br>(Other-is-presenting = F)<br>(Other-wants-turn = F)) | **7** |
| **STATE:** *I-Take-Turn*<br>TRANSITION TO STATE *Other-Has-Turn* IFF:<br>(AND (***Time-since*** 'I-Take-Turn > 120msec)[i]<br>(OR<br>(Other-is-speaking = T)<br>(Other-wants-turn = T)<br>(Other-is-presenting = T))) | **8** |

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

*Table 3. Top-level dialogue State Deciders used in the YTTM prototype for Gandalf. States that determine whether the agent is engaged in dialogue, and therefore subsume the turn states. The dialogue states are used to set the stage for the turn-taking; dialogue starts (in this implementation) when a human being addresses the agent by name or by greeting.*

| | |
|---|---|
| **STATE:** *Dialog-on*<br>TRANSITION TO STATE *Dialog-off* IFF:<br>(AND (Other-is-saying-goodbye =T)[a]<br>(Dialog-off = F)) | **9** |
| **STATE:** *Dialog-off*<br>TRANSITION TO STATES (AND *Dialog-on I-Take-Turn)* IFF:<br>(AND<br>(Dialog-off = F)<br>(OR<br>(Other-saying-my-name =T)<br>(AND (Other-is-greeting = T) (Other-is-addressing-me = T))<br>(Other-is-paying-general-attention = T))) | **10** |

## 5. IMPLEMENTATION OF YTTM

The YTTM has been implemented in two working systems, *Puff the Magic LEGO Dragon*, and *Gandalf, the Interactive Guide to the Solar System*. The former was developed as part of a virtual world research project at LEGO, demonstrating the easy integration of YTTM with traditional planning systems, giving Puff high-level goals such as *stay-alive* and *act-playfully*, and the ability to plan actions on larger time scales than the turn (Bryson & Thórisson in press). The focus here will stay on the Gandalf prototype, with its high-fidelity multimodal perception system.

### *5.1. Elements of YTTM*

The YTTM has been implemented by combining features from three A.I. approaches and cognitive modeling: Behavior-based A.I. (cf. Maes 1990b), the (classical A.I.) idea of blackboards and distributed processing (Adler 1989, Selfridge 1959), and the Model Human Processor (Card et al. 1983). Perception is done via a collection of *Perceptor* modules which take in sensory data, or partially processed data from other Perceptors, and compute further results. They are divided into *Unimodal Perceptors* (UPs) and *Multimodal Integrators* (MIs). UPs turn raw sensory data into more meaningful information by segmenting and processing subsets of data from a single mode (e.g. hearing or vision). This data is at or slightly above digital signal processing. MIs take processed output from UPs and other MIs and thus combine information from multiple modes when computing perception. *Decider* modules are divided into *Covert State Deciders*, which keep track of dialogue state and turns, and *Overt Deciders*, which make decisions about an agent's visible behavior. All Deciders read output from the UPs and MIs via shared blackboards. When Overt Deciders fire, they generate a Behavior Request for a particular visible behavior to happen, which are handled by an action scheduler, as explained in Section 4.6.

The Perceptors and Deciders form together the foundation of the YTTM turn-taking system. The rules and processes for the Perceptors and Deciders in the Gandalf

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

prototype agent described here are a distillation of psychological research on the behavior found in human face-to-face dialogue (cf. McNeill 1992, Rimé & Schiaratura 1991, Pierrehumbert & Hirschberg 1990, Goodwin 1986, Kleinke 1986, Nespolous & Lecours 1986, Goodwin 1981, Kahneman 1973, Duncan 1972, Yngve 1970, Ekman & Friesen 1969, Effron 1941), and constitute a blueprint for how a turn-taking system for a particular culture (Western) can be implemented in this architecture.

### *5.2. Turning Hypotheses Into Code*

Before we give a short description of the modules themselves, we will first look at how the nine preceding hypotheses relate to this YTTM implementation.

- {H1} Opportunities for turn-transitions are identified using a mechanism of anticipatory perceptual processing.

This first hypothesis is implemented by tying each Multimodal Integrator (Table 4) to a particular mental (usually turn) state, indicated in the ACTIVE-DURING-STATE slot. Turn- and dialogue states are maintained by a system of Deciders (Table 2) and their supporting Perceptors. To change a state, a Covert State Decider posts a message to a blackboard about the new state.

Multimodal Integrators (Table 4) integrate and use data from Unimodal Perceptors (Table 5). The UPs are in the Reactive Layer and take priority over all other processes. They are not linked to particular states and process continuously when data is available. A higher-level set of modules (Table 3) monitor whether the agent is actually engaged in a dialogue or not, subsuming all turn states. (None of the Deciders in

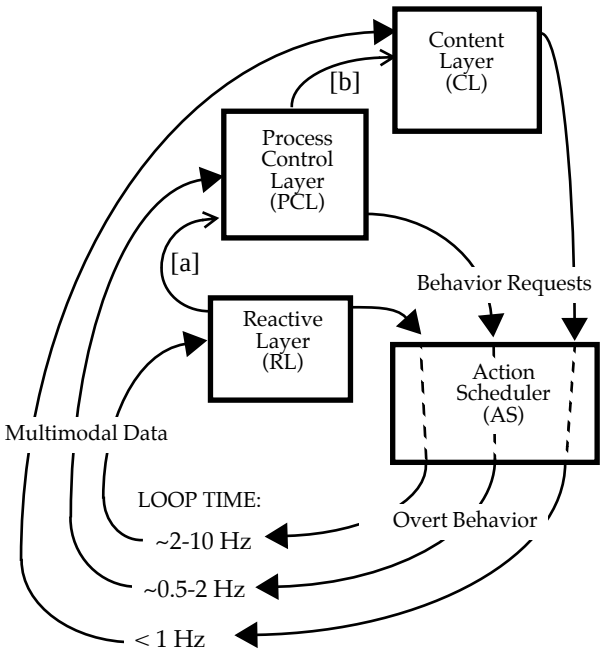


*Figure 2. Multimodal input flows into all three priority layers (from Thorisson [1998]). Decision modules operate on these results and decide when to send Action Requests to an Action Scheduler, which then produces visible behavior. Target loop times for each layer is shown in Hz (compare to Table 1). It is important to note here that the frequency refers not to the layers' internal update time or sampling rate, nor to the speed of decision making, but to a full perception-action loop. [a] and [b] are partially processed multimodal data.*

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

*Table 4. Multimodal Integrators used in the Gandalf prototype using YTTM. For items [a]-[c] see Section 5.3.2.*

| | |
|---|---|
| **Other-is-giving-turn** **11**<br>ACTIVE-DURING-STATE: *Other-Has-Turn*<br>CONDITIONS:<br>(AND<br>(Other-is-speaking = F)<br>(OR<br>(AND<br>(Other-is-looking-at-me = T )<br>(Other-is-facing-me = T))<br>(AND<br>(Other-is-looking-at-me = T)<br>(Other-is-gesturing = F))<br>(AND<br>(Other-is-gesturing = F)<br>(Other-is-facing-me = T)))) | **Other-accepts-turn** **16**<br>ACTIVE-DURING-STATE: *I-Give-Turn*<br>CONDITIONS:<br>(AND<br>(Other-is-looking-at-me = F)[c]<br>(Other-is-presenting = T))<br><br>**Other-is-addressing-me** **17**<br>CONDITIONS:<br>(AND<br>(Other-is-turned-to-me = T)<br>(Other-is-facing-me = T)<br>(Other-is-looking-at-me = T)) |
| **Other-wants-turn** **12**<br>ACTIVE-DURING-STATE: *I-Have-Turn*<br>CONDITIONS:<br>(OR<br>(Other-is-speaking = T )<br>(Others-hand-in-gesture-space = T )) | **Other-wants-my-back-channel-feedback** **18**<br>ACTIVE-DURING-STATE: *Other-Has-Turn*<br>CONDITIONS:<br>(AND<br>(Other-is-looking-at-me = T)<br>(Other-is-speaking = T)) |
| **Other-is-looking-at-own-hand** **13**<br>PERCEPTOR-TYPE: Multimodal-Integrator<br>ACTIVE-DURING-STATE: *Dialog-On*<br>CONDITIONS:<br>(OR<br>(Other-is-looking-at-own-right-hand = T)<br>(Other-is-looking-at-own-left-hand = T)) | **Others-either-hand-in-gest-space** **19**<br>PERCEPTOR-TYPE: Multimodal-Integrator<br>ACTIVE-DURING-STATE: *Dialog-On*<br>CONDITIONS:<br>(OR<br>(Others-left-hand-in-gesture-space = T)<br>(Others-right-hand-in-gesture-space = T)) |
| **Other-is-presenting** **14**<br>ACTIVE-DURING-STATE: *Dialog-On*<br>CONDITIONS:<br>(OR<br>(Others-either-hand-in-gest-space = T)<br>(Other-is-speaking = T)) | **Other-is-gesturing** **20**<br>ACTIVE-DURING-STATE: *Dialog-On*<br>CONDITIONS:<br>(AND<br>(Others-either-hand-in-gest-space = T)<br>(Other-is-speaking = T)) |
| **Other-produced-complete-utterance** **15**<br>ACTIVE-DURING-STATE: *Dialog-On*<br>CONDITIONS:<br>(AND<br>(Others-utterance-is-semantically-correct = T)[a]<br>(Others-utterance-is-syntactically-correct = T))[b] | **Other-is-paying-general-attention** **21**<br>ACTIVE-DURING-STATE: *Dialog-Off*<br>CONDITIONS:<br>(OR<br>(Other-is-turned-to-me = T)<br>(Other-is-facing-me = T)<br>(Other-is-facing-workspace = T)) |

Table 2 will be running unless dialogue is 'on', as controlled by the Deciders in Table 3.)

- {H2} The seamlessness observed in real-time turn-taking comes from the co-operation of multiple processes with different (a) update frequencies, (b) target perception-action loop times, and (c) speed-accuracy tradeoffs.

Processes in a single layer of YTTM are given a single target update frequency, calculated to achieve a desired perception-action loop time. For the Reactive Layer this loop time is 2 - 10 Hz (typical), for the Process Control Layer it is 0.5 - 2 Hz, and for

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

*Table 5. Unimodal Perceptors used in the Gandalf prototype. Some features of the user's behavior are computed continuously, these are referred to with variables filling the INDEX and DATA slots; other data are only computed when a particular module runs (i.e. when needed), necessitating a function call (bold italic). For discussion see Section 5.3.3.*

| | |
|---|---|
| **Other-is-facing-workspace** **22**<br>TYPE: Unimodal-RL-body-perceptor<br>DATA-1: workspace<br>DATA-2: **get-head-direction**<br>FUNC: ***x-facing-y*** | **Other-is-looking-at-me** **30**<br>TYPE: Unimodal-RL- body-perceptor<br>DATA-1: my-own-face<br>DATA-2: **get-gaze-direction**<br>FUNC: ***x-looking-at-y*** |
| **Other-is-looking-at-own-right-hand** **23**<br>TYPE: Unimodal-RL-body-perceptor<br>DATA-1: **get-gaze-direction**<br>DATA-2: **get-r-wrist-position**<br>FUNC: ***u-looking-at-hand?*** | **Others-right-hand-is-in-gesture-space** **31**<br>TYPE: Unimodal-RL-body-perceptor<br>DATA-1: **get-r-wrist-position**<br>DATA-2: **get-trunk-direction**<br>FUNC: ***hand-in-gest-space?*** |
| **Other-is-looking-at-own-left-hand** **24**<br>TYPE: Unimodal-RL-body-perceptor<br>DATA-1: **get-gaze-direction**<br>DATA-2: **get-l-wrist-position**<br>FUNC: ***u-looking-at-hand?*** | **Others-left-hand-is-in-gesture-space** **32**<br>TYPE: Unimodal-RL-body-perceptor<br>DATA-1: **get-l-wrist-position**<br>DATA-2: **get-trunk-direction**<br>FUNC: ***hand-in-gest-space?*** |
| **Others-syntax-is-complete** **25**<br>TYPE: Unimodal-RL-speech-perceptor<br>INDEX: last-utterance<br>DATA: others-word-stream<br>FUNC: ***syntax-complete?*** | **Other-is-speaking** **33**<br>TYPE: Unimodal-RL-prosody-perceptor<br>INDEX: 40-msec-chunk<br>DATA: others-audio-stream<br>FUNC: ***x-speaking?*** |
| **Other-is-turned-to-me** **26**<br>TYPE: Unimodal-RL-body-perceptor<br>DATA-1: my-own-face<br>DATA-2: **get-trunk-direction**<br>FUNC: ***turned-to?*** | **I-see-other** **34**<br>TYPE: Unimodal-RL-vision-perceptor<br>INDEX: body-socket-connection<br>DATA: *socket-object1*<br>FUNC: ***visual-connection-alive?*** |
| **Other-is-turned-to-workspace** **27**<br>TYPE: Unimodal-RL-body-perceptor<br>DATA-1: workspace<br>DATA-2: **get-body-direction**<br>FUNC: ***x-facing-y*** | **Other-is-facing-me** **35**<br>TYPE: Unimodal-RL-body-perceptor<br>DATA-1: my-own-face<br>DATA-2: **get-head-direction**<br>FUNC: ***x-facing-y*** |
| **Others-utterance-is-grammatically-correct** **28**<br>TYPE: Unimodal-RL-speech-content-perceptor<br>INDEX: last-utterance<br>DATA: others-word-stream<br>FUNC: ***grammar-complete?*** | **Others-utterance-is-semantically-complete** **36**<br>TYPE: Unimodal-RL-speech-con tent-perceptor<br>INDEX: last-utterance<br>DATA: others-word-stream<br>FUNC: ***semantics-complete?*** |
| **Others- intonation-is-going-up** **29**<br>TYPE: Unimodal-RL-prosody-perceptor<br>INDEX: 300-msec-chunk<br>DATA: others-audio-stream<br>FUNC: ***inton-direction*** | **Others-intonation-is-going-down** **37**<br>TYPE: Unimodal-RL-prosody-perceptor<br>INDEX: 300-msec-chunk<br>DATA: others-audio-stream<br>FUNC: ***inton-direction*** |

the Content Layer it is 1 Hz and slower (Figure 2). As mentioned earlier, processes in the Reactive Layer have the largest speed/accuracy ratio, those in the Content Layer have the highest accuracy and lowest relative speed, with those in the Process Control Layer in the middle. All UPs are in the RL in our implementation, while other types of modules are found in all layers.

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

- {H5} Perceptual and decision processes dedicated to tracking dialogue state have the highest priority of all mental activities related to communication.
- {H6} In real-time communication, analysis and interpretation of broad-stroke communicative function takes higher priority than content analysis and interpretation.

By placing all low-level processes in the Reactive Layer, highest priority is given to processing raw data that can help determine the dialogue state (Table 5). Topic interpretation happens entirely in processes situated in the Content Layer, the lowest-priority layer in the system. A scheduling system ensures a guaranteed processing time for the Reactive and Process Control layers. This varies between deployment platforms, and is tuned based on the target loop-times.

- {H3} To support coherent output generation in a modular, distributed system, processes with different perception-action loop times are sensitive to a particular subset of the total set of contextual cues available, which includes perceptual data produced from input to the sensors, as well as processing states and partial output of other system elements.

As we established in sections 2. and 4., the human mind interprets the world incrementally. For example, the movement of an object becomes available in a perceiver's mind sooner than its color (Kosslyn & Koenig 1992). This means that at any point in time some mental processes contain partial information about the world. To make use of this partial information the intermediate stages of data should be made available to other processes, in case they need it or are able to use it. This is an ideal problem to solve with a blackboard architecture. There are two blackboards used for perception in YTTM; they sit between the three layers. Bottom-up processing pushes incrementally more detailed sensory data upwards, from UPs to MIs to more knowledge-intensive processes; decisions and planning from the deliberative processes in the Content Layer push expectations and anticipatory commands downward, all via the blackboards. Any module that needs a particular set of data to produce its output looks at one of the two blackboards. If it finds the data it needs, it processes it and places the results on one of the two blackboards, making them available to other modules.

```
Topic-Knowledge-System-Received-Speech-Data
Speech-Data-Available-For-Analysis
Topic-Knowledge-System-Parsing-Speech-Data
Topic-Knowledge-Ssystem-Successful-Parse
Content-Layer-Action-Available
I-Have-Reply-Ready
Topic-Knowledge-System-Real-World-Action-Available
Im-Executing-Topic-Speech-Task
Im-Executing-Topic-Realworld-Task
Im-Executing-Topic-Multimodal-Act
Im-Executing-Topic-Communicative-Act
Im-Executing-Communicative-Act
```

*Figure 3. A basic set of communication primitives from processes in the Process Control Layer to a Topic Knowledge Base in the Content Layer, posted on the blackboard shared by CL and PCL. When these are posted they are timestamped, and provided with a pointer that allows other modules to access the data that the message refers to. The primitives form part of the turn-system's contextual cues. The list must be extended for domains more complex than the one explored here.*

- {H4} Processing related to turn-taking can be separated from processing of content (i.e. topic) via a finite set of interaction primitives.

Dialogue and turn states are tracked in the Process Control Layer; knowledge systems related to the topic are placed in the Content Layer and handle everything

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

*Table 6. Overt Decision Modules used in Gandalf's* Reactive Layer. *These modules control Gandalf's reactive behavior. For discussion see Section 5.3.4.*

| | |
|---|---|
| **Show-Im-taking-turn** **38**<br>EL: 5000 msec<br>BehaviorRequest: Show-im-taking-turn<br>FIRE-CONDS: (I-take-turn = T)<br>RESTORE-CONDS: (I-take-turn = F) | **Show-Im-giving-turn** **43**<br>EL: 2000 msec<br>BehaviorRequest: Show-Im-giving-turn<br>FIRE-CONDS: (I-give-turn = T)<br>RESTORE-CONDS: (I-have-turn = F) |
| **Show-I-know-other-is-addressing-me-1** **39**<br>EL: 200 msec<br>BehaviorRequest: Smile<br>POS-CONDS: (Im-executing-speech-act = T)<br>NEG-RESTR-CONDS: (Other-is-turned-to-me = F) | **Show-Im-giving-turn-2** **44**<br>EL: 2000 msec<br>BehaviorRequest: Show-Im-giving-turn<br>FIRE-CONDS: (I-give-turn = T)<br>RESTORE-CONDS: (I-give-turn = F) |
| **Show-I-know-other-is-addressing-me-2** **40**<br>EL: 200 msec<br>BehaviorRequest: Eyebrow-greet<br>POS-CONDS: (AND (Other-is-saying-my-name = T) (Other-is-turned-to-me = T) (Other-is-facing-me = T))<br>RESTORE-CONDS: (Other-is-turned-to-me = F) | **Show-Im-listening** **45**<br>EL: 200 msec<br>BehaviorRequest: Brows-in-pensive-shape<br>FIRE-CONDS: (AND (Other-is-saying-my-name = T) (Other-is-turned-to-me = T) (Other-is-facing-me = T))<br>RESTORE-CONDS: (Other-is-turned-to-me = F) |
| **Initialize-dialogue** **41**<br>EL: 200 msec<br>BehaviorRequest: Face-neutral<br>FIRE-CONDS: (Dialog-On = F)<br>RESTORE-CONDS: (Dialog-On = T) | **Show-I-know-other-is-not-addressing-me** **46**<br>EL: 1000 msec<br>BehaviorRequest: (***Turn-to*** 'Work-space)<br>FIRE-CONDS: (AND (Dialog-On = F) (Other-is-turned-to-me = F))<br>RESTORE-CONDS: (Other-is-taking-turn = T) |
| **Look-puzzled-during-awkward-pause** **42**<br>EL: 1000 msec<br>BehaviorRequest: Look-puzzled<br>FIRE-CONDS: (AND (other-is-turned-to-me = T) (other-is-facing-me = T) (***Time-since*** 'Other-is-facing-me > 400))<br>RESTORE-CONDS: (Other-is-turned-to-me = F) | **Look-aloof** **47**<br>EL: 1000 msec<br>BehaviorRequest: Look-aloof<br>FIRE-CONDS: (AND (Other-is-turned-to-me = T) (Other-is-facing-me = T) (***Time-since*** 'Other-is-facing-me > 800) (Dialog-On = T) (Other-is-speaking = F) (Topic-Knowledge-System-Parsing-Speech-Data = F)<br>RESTORE-CONDS: (Other-is-turned-to-me = F) |

related to the topic. Topic and dialogue processes talk to each other via a limited set of messages (Figure 3).

- {H7} Reactive behaviors are based on data produced by highly opportunistic processing.

To meet this objective, the processing necessary for producing coherent system behavior — including content and gesture analysis — is sliced into small units, each responsible for only a fraction of the overall interpretation. These processes are distributed throughout the three priority layers.[7] Functions for Unimodal Perceptors and Multimodal Integrators (values in slot 'FUNC' in Table 5, bold italic values in tables 2, 6 and 7) provide services that describe various parts of the dialogue state and outside world at any moment in time. Since it is not possible to predict which pieces of

7. Processing related to constructing a morphology (motor program) based on decisions and goals (such as 'greet-other') is done by the action scheduler (Section 4.6) running in parallel with processes in each of the layers, at the same priority as processes in the Reactive Layer.

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

the data will be available at any moment in time, the small units make opportunism the default method by which the system does interpretation and produces behavior.

All modules contain a list of conditions, in their FIRE-CONDS slot, whose boolean combination determines their output to the blackboards, meeting the last two hypotheses:

- {H8} Separate features and cues extracted, by perceptual processes of a dialogue participant A, from a particular multimodal action by dialogue partner B, are logically combined (in the mathematical sense of the word) by other perception processes in the mind of participant A, to support generation of appropriate behavior during the interaction.
- {H9} A decision is based on the boolean combination of perceptual features.

### *5.3. Perceptor & Decider Rules*

This section explains selected modules, where their rules come from and how they interact. As mentioned before, this information is specific to a culture, and is included here as a reference for how rules are encoded in this implementation of the YTTM for supporting the interactive Gandalf character. We start with the high-level turn-rules and trace our steps backwards, ending with the Deciders that turn perceptual and cognitive data into situated behavior.

#### *5.3.1. Turn States (Table 2)*

The turn states in Table 2 form the crux of the turn-taking system in this implementation of the YTTM. These relatively reactive modules help determine the agent's 'interactive personality'. For example, by removing the second condition in Decider 1 ([a] in Table 2) an agent will not give the turn if it is engaged in real-world tasks, even if the other wants the turn. As seen in Transition Rule 4, the transitions are modeled explicitly as states; *I-Take-Turn* leading into *I-Have-Turn*. The conditions marked [b] are determined by the topic knowledge system and form part of the system's contextual cues. Condition [c], Other-accepts-turn, is a perception, not a state; the state *Other-Has-Turn* is thus driven off the perceptual system.

The function ***Time-since*** takes two variables, a status message (e.g. Other-is-speaking) and a time in milliseconds. In case [h] (module 6), for example, it will return *true* if the time since the user stopped presenting (according to the agent's perceptual processing) is greater than 120 msec. The upper bound on the acceptable pause between one partner giving the turn and the other taking it (again, in the Western world) is 250 msec or less (Goodwin 1981). This observation has been implemented as three transition rules 4, 5, and 6, gradually less strict in their necessary conditions. The first rule requires a complete utterance to have been produced (measured by whether the utterance is syntactically correct and if it makes sense[8]) for the agent to take turn. However, given such evidence, the probability of a valid turn transition is so high that a mere 50 msec [d] wait is sufficient. The second rule does not require a complete utterance, but uses intonation direction as an indication [g], which will be

8. The semantic completeness of a multimodal event is computed in various ways from context. For example, should a user utter the words "what is that?" with no accompanying body movements the semantic completeness is given a lower score than if that utterance had been complemented by a glance and/or a deictic gesture that singles out a relevant object.

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

*Table 7. Overt Decision Modules used in Gandalf's* Process Control Layer. *For more details see Section 5.3.4.*

| | |
|---|---|
| **Acknowledge-others-attention-during-presentation** **48**<br>EL: 20000 msec<br>BehaviorRequest: (***Gaze-At*** 'Other)<br>FIRE-CONDS: (AND (Im-executing-act = T) (Other-is-looking-at-me = T))<br>RESTORE-CONDS: (I-take-turn = T) | **Show-Im-done-with-task-by-looking-at-other** **54**<br>EL: 20000 msec<br>BehaviorRequest: (AND (***Turn-head-To*** 'Other) (***Gaze-At*** 'Other))<br>FIRE-CONDS: (Im-executing-topic-realworld-task = F)<br>RESTORE-CONDS: (Im-executing-topic-realworld-task = T) |
| **Turn-to-other-when-I-speak** **49**<br>EL: 20000 msec<br>BehaviorRequest: (***Turn-head-To*** 'Other)<br>FIRE-CONDS: (Im-executing-speech-act = T)<br>RESTORE-CONDS: (Im-executing-communicative-act = T) | **Look-at-domain-with-other** **55**<br>EL: 20000 msec<br>BehaviorRequest: (***Turn-head-To*** 'workspace)<br>FIRE-CONDS: (AND (Other-is-facing-domain = T) (I-have-turn = T) (Other-is-speaking = F))<br>RESTORE-CONDS: (Other-has-turn = T) |
| **Hesitate-during-delay-in-reply-formulation** **50**<br>EL: 500 msec<br>BehaviorRequest: Show-Hesitation<br>FIRE-CONDS: (AND (Dialog-On = T) (I-have-turn = T) (Speech-data-available-from-other = T) (***Time-since*** 'Other-is-speaking > 70 msec) (I-have-reply-ready = F) (other-is-speaking = F))<br>RESTORE-CONDS: (I-give-turn = T) | **Allow-other-to-interrupt-me-during-task** **56**<br>EL: 20000 msec<br>BehaviorRequest: (***Turn-head-To*** 'Other)<br>FIRE-CONDS: (AND (Im-executing-topic-realworld-act = T) (Other-is-looking-at-me = T) (other-is-facing-me = T) (Other-is-speaking = T))<br>RESTORE-CONDS: (Im-executing-topic-realworld-task = F) |
| **Pay-attention-to-my-own-action** **51**<br>EL: 20000 msec<br>BehaviorRequest: (***Turn-head-To*** 'Work-space)<br>FIRE-CONDS: (Im-executing-topic-realworld-task = T)<br>RESTORE-CONDS: (Im-executing-topic-realworld-act = F) | **Show-Im-idle** **57**<br>EL: 20000 msec<br>BehaviorRequest: Restless<br>FIRE-CONDS: (Other-is-facing-me = F)<br>RESTORE-CONDS: (Other-is-facing-me = T) |
| **Show-Im-listening-to-other** **52**<br>EL: 20000 msec<br>BehaviorRequest: (AND (***Turn-head-To*** 'Other) (***Gaze-At*** 'Other))<br>FIRE-CONDS: (AND (Other-is-speaking = T) (Other-is-paying-general-attention = T))<br>RESTORE-CONDS: (I-have-turn = T) | **Turn-to-other-when-I-present** **58**<br>EL: 2000 msec<br>BehaviorRequest: (***Turn-Head-To*** 'Other)<br>FIRE-CONDS: (Im-executing-topic-communicative-act = T)<br>RESTORE-CONDS: (Im-executing-topic-communicative-act = F) |
| **Acknowledge-other-is-addressing-me** **53**<br>EL: 2000 msec<br>BehaviorRequest: (***Turn-To*** 'Other)<br>FIRE-CONDS: (AND (Im-executing-topic-realworld-task = T) (Other-is-looking-at-me = T) (Other-is-facing-me = T) (Other-is-speaking = T))<br>RESTORE-CONDS: (Im-executing-topic-realworld-task = F) | |

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

going down ("final fall") if the user is stating a command and up ("final rise") if the user asked a question (Thórisson, in press, Pierrehumbert & Hirschberg 1990). In this rule the agent waits 70 msec before acting [f] (20 msec longer than in module 4). The third rule catches the condition when the user's utterance is *not* complete (or takes longer than 70 msec to be computed) and intonation is not determinate or is not computed. In this case taking the turn is delayed for an additional 70 msec, bringing the wait up to 120 msec [h]. This breakdown exemplifies how *cascaded decision modules* can be used to track state, and use real-time as part of the processing. It's also an example of our discussion in Section 4.4.2 about the reliability of perceptual data and how soon it can be acted on. Notice that the conditions in all but the last rule work as "evidence" of a certain state of the world being true. Even if the first two rules fail, the third will default to true, unless outside events cause other states in the mean time. If module 6 fails other rules will fire to stabilize the system (see examples below).

Since the conditions *Other-is-taking-turn* and *Other-is-giving-turn* are both perceptual & transitional states, and thus measured by separate, independent perceptual processes, the condition can arise that they are both true ([e] in Table 2). This might happen if for example a non-native speaker uses different rules of conduct when taking turns. Since they are not mutually exclusive, both have to be listed here for higher certainty that the perceived state is correct.

Two rules deal with collaborative mistakes: Transition 3 happens if the agent gives turn but the partner shows no signs of accepting it (time delays may be needed to prevent premature firing of this module). Transition 8 only happens if the agent mistakenly took the turn. If the time since the agent's decision to take turn is greater than 120 msec and the user is still talking or seems to be wanting the turn, the turn transition was possibly made erroneously by the agent ([i]). This might be caused by a failure in the agent's perceptual or decision mechanisms, or because the presenter reversed a decision to give turn.

### *5.3.2. Multimodal Integrators (Table 4)*

In Multimodal Integrator 21, any of the conditions will trigger an *Other-is-paying-general-attention* message to get posted to a blackboard (all OR states are inclusive). Integrator number 13 is used to flag the potential presence of iconic gestures (manual gestures where the hand plays the role of another object (Rimé & Schiaratura 1991, Effron 1941)) and thus prime the knowledge system to analyze its meaning.

Research shows that when interpreters intend to take turn (again, in the Western world) when given by the presenter, they pull away their gaze, which typically was focused on the presenter's face up until that point (Kahneman 1973, Duncan 1972). We have captured this in its simplest form in module 16, where the rule Other-is-looking-at-me = F ([c] in Table 4) requires the other's gaze to fall elsewhere than on the agent's face.

### *5.3.3. Unimodal Perceptors (Table 5)*

The UPs provide all medium and low-level perception necessary, and thus form the foundation for all decisions about turn-taking and covert behavior. The function of Unimodal Perceptors such as module 25 uses complex algorithms and may not always compute the answer fast enough for the dialogue to proceed correctly. Proper

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

load-balancing via the layers, along with flexible rules, is critical to achieve the required realtime performance of the whole system: Should the UPs not get sufficient processing cycles, interaction will suffer.

*5.3.4. Overt Decision Modules (Tables 6, 7)*

The Overt Decider modules have the role of generating visible behavior in response to dialogue events and the agent's own mental events. When their conditions are met they fire a Behavior Request to the action scheduler. After firing they wait for the conditions in RESTORE-CONDS to become true; until then they are unable to fire.

The agent's behavioral system needs to know the time-dependency of every decision and plan made, because one decision may interfere with another, and have to be put on hold. One way this is handled is via scheduling prioritization based on the three layers, a second way is via time-dependency: Each decision made by the system has a time-out associated with it that determines how long it may be buffered in the system, waiting to be executed. In our implementation this is done via the Expected Lifetime (EL) variable, whose value determines how long a Behavior Request produced by an Overt Decider can live in the system without being turned into motor movements. If the EL time is reached before the act can be executed (for one reason or another) the Behavior Request is cancelled. EL values are selected based on psychological studies, but tuned empirically.

Research has shown that greetings are often accompanied by widening of eyes and/or a brief lifting of eyebrows (Schegloff & Sacks 1973). This has been implemented in Decider 40. Goodwin (1981), Duncan (1972) and others have shown that when taking the turn people glance away from their partner, which has led to module 38: The behavior request Show-Im-taking-turn can be realized in one of two ways by the action scheduler: (1) Raising the eyebrows and quickly glancing to the side and back, or (2) turning to face the other, quickly glancing to the side and back, and opening the mouth slightly. In Western cultures the possibilities are of course not limited to these two, and one can imagine a system where, for each such behavior, several variations exist. (Variations can continue to be added to this system, potentially into the hundreds.) The accompanying decision to look at the content presenter to show attention is encapsulated in module 52.

When executing a domain act, Gandalf will look at the events it causes, module 51. In Western turn-taking a presenter tends to look back at the interpreter when he's done presenting, signalling that the turn is available (Goodwin 1981). This is captured in module 54.

To indicate problems or delays in topic processing, module 50 will jump in and display a hesitation. In the prototype a hesitation consisted of three realizations: (1) Saying "ahhh...", (2) gazing upwards, or (3) putting on a pensive facial expression. The EL for this module is 500 msec. This means that if Gandalf decides to hesitate, but then 500 msec pass and this decision is not realized as movement (for whatever reason), the decision will not get realized at all. When this happens, instead of monitoring this event internally, we let the real-world loop catch the result: If Gandalf failed to hesitate and the user took back the turn as a result, Gandalf's turn-taking mechanism would sense this condition via perceptual mechanisms and give back the turn, and thus pick up the slack.

For spatial behaviors that are made relative to real-world objects (most movements besides facial expression) we needed special methods. For example, since the

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

position of the other participant changes, as does the direction of the agent's head, the relative difference needs to be measured each time the agent wants to look at the other. For this we implemented a function that returns the difference between two objects and uses it to calculate where to move the agent's motors. An example of this is module 48 (the condition lm-executing-act refers to any act, whether it originated in the topic knowledge system or not). Module 52 shows the use of two functions, one for turning the head, one for changing the gaze; this module contains a reference to realworld-task which is true whenever Gandalf travels around the virtual solar system.

### *5.4. System Setup & Performance*

#### *5.4.1. Sensory & Display Hardware & Software*

The prototype system for Gandalf consists of eight computers: Four computers are dedicated to sensation; one each for prosody analysis, gaze calculation, geometric body modeling, and speech recognition (Thórisson, in press). Two computers are used to run the Multimodal Integrators, Deciders, topic knowledge system, and motor scheduling. Two computers manage the animation for the face and for the three-dimensional model of the solar system (Figure 4). The display showing Gandalf's head and hand ("Agent screen") is angled in such a way as to allow him to look at the solar system (to his left) and see the human user right in front. The human sees the workspace display right in front of her and Gandalf to her left.

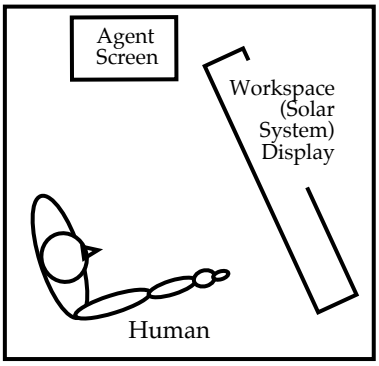


*Figure 4. Top view of display layout in the Gandalf prototype setup.*

#### *5.4.2. Performance*

The Ymir Turn-Taking Model as implemented shows a remarkable flexibility and adaptability considering its relatively small rule base. A large part of that flexibility we believe comes from the layered approach, as well as the explicit handling of time in the system. The number of people who have interacted with Gandalf, and it's cousin Puff, is in the hundreds; most of them have only been given general instructions such as "act as if you are interacting with another person". When Gandalf senses that the human is present, a greeting, such as "I am Gandalf, your guide to the Solar System; I can fly to the planets and tell you about them" sets up the right expectations with regard to the task and guides users to ask the questions that Gandalf understands. Gandalf's sensory and turn-taking mechanisms presented above make sure that the greeting is uttered at the right time — rarely does it fail. Within less than a minute people are communicating naturally and taking turns efficiently, flying to the planets and listening to and watching Gandalf talk about them and their moons. Interacting with Gandalf surely requires no manual. In questionnaires collected from users after talking to Gandalf, users give the system very high grades on interactivity, speech understanding, speech generation, and intelligence (this is prob-

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

ably not as much a reflection of the system as it is of people's perception of it). Scales are grounded in each end by asking people to compare the system to interactions with animals, such as fish, dogs, and cats on the one hand, and humans on the other. Grades on these scales for Gandalf typically fall somewhere betwen the interactivity of a dog and a real human. Interestingly, when the mechanisms presented in this chapter are turned *off*, compared to an idential version with only verbal responses to user's questions, a (statistically) significant decrease in Gandalf's scores for language understanding and language expression is observed, moving closer to the score for these parameters given for human-dog interaction. When all of the dialogue intelligence is turned on people score Gandalf's language abilities significantly closer to human-human interaction.

Figure 5 is a randomly selected five-second segment from a corpus of hours of interaction recordings, showing how the turn-taking system typically performs in interaction with real users. The user was looking and facing Gandalf during the segment. A subset of the modules in the full system are plotted. In this example a Multimodal Integrator for sensing that the other is giving turn (Other-is-giving-turn, module 11 in Table 4) failed on the second turn transition [a] (it worked correctly in case [b]). Nontheless, the system performed correctly because these two perceptual sensors are not mutually exclusive, and in this case the output from the sensor for user taking turn resulted in acceptable behavior. In all cases in this segment the agent gives and takes turn under 70 msec; they turn out to be correctly estimated transitions as well. This is an example of the architecture allowing the system to tend towards homeostasis in the presence of error, and how a behavior-based architecture results in real-time performance while preserving structural simplicity.

Figure 6 shows another example of a typical event sequence, chosen to demonstrate further the internal events that form part of the context of the dialogue, along with external states and events. The system recognizes that the user is taking the turn [a] and gives turn [b], and shows that it's giving turn [c]. The user starts speaking about 50 msec later. His request is "Tell me about that planet" (pointing at the screen). The system is relatively slow to take turn again [d] (approx. 500 msec), and shows that its taking it [e]. The duration since the user stopped speaking and gave the turn is now long (relatively speaking) and the system decides that it should show

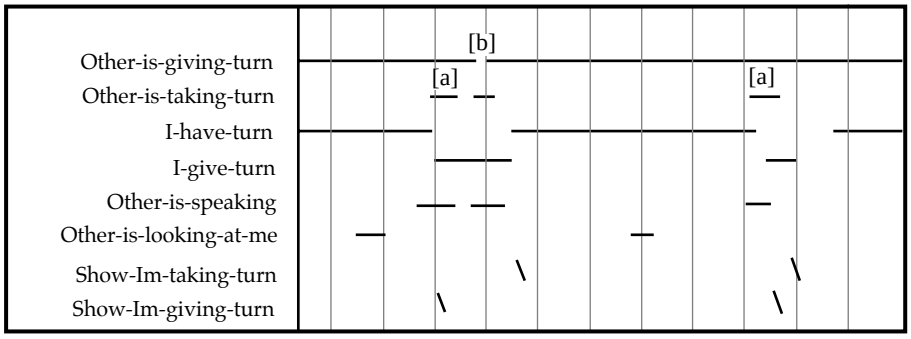


*Figure 5. Example of turn-taking performance by YTTM. Each vertical line marks one second. Horizontal line means condition on left hand side was* True *during that period. Decisions for showing that the agent is taking and giving turn (resulting from state changes) are plotted in the bottom two lines. For further explanation, see Section 5.4.2.*

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

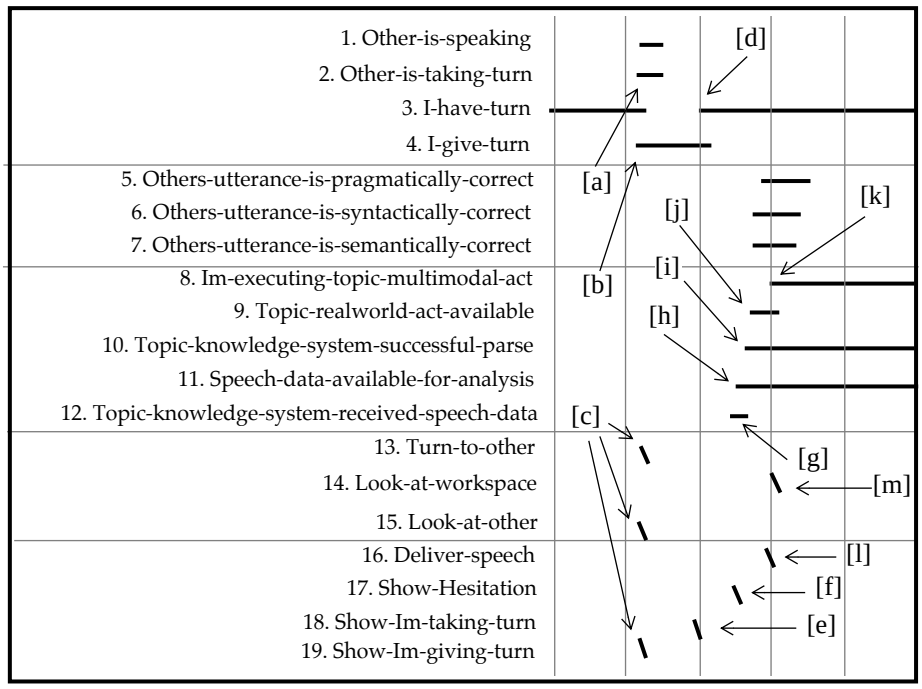


*Figure 6. Example of turn-taking events showing more features of the internal contextual messages, in lines 5-11. The figure shows partial representation of internal variables during run-time. Horizontal lines mark seconds. See text for discussion.*

hesitation [f], since it has accepted the turn but not yet started to respond to the content of what the user said (in fact, as of this moment, the system has no idea *what* the user said, just *that* the user said ... something). A few milliseconds later the topic knowledge system receives a report from the speech recognition subsystem [g]. In this implementation the speech recognition usually delivered words extracted from the speech stream 1-2 seconds after the presenter became silent (real-time information about the speech is generated by the prosody system, which has only a 30 msec time lag). This is identified as a word collection, and a parse is initiated [h]. Roughly 10 msec later the words have been parsed successfully [i] and about 5 msec after that a response has been generated [j]. Post-processing finds the user's input has valid syntax, is meaningful and makes sense in the current task, and hence the system decides (an overt decision) to execute the domain action [l] which was generated in response to the input, and this is subsequently begun [k]. The system's response to the user's request is the speech output "That is Saturn. It has three rings.", and a manual gesture pointing at the planet. As the system explains this the user looks at the screen; to mimic the user's action the system also looks at the screen while telling the user about the planet [m].

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

### *5.5. Discussion & Summary*

The hypotheses presented in this chapter formalize elements needed for a complete generative model of real-time, face-to-face turn-taking. They provide a foundation for the construction of a turn-taking model, YTTM, that has been implemented using completely automated perception of a dialogue participant's behavior, including speech, prosody, gesture and body language, and generating real-time animation and speech output. In goal-directed dialogue with humans the implemented model, in the form of a humanoid agent, produces turn- and dialogue behavior very similar to that seen in human-human dialogue.

The model described combines classical AI and behavior-based AI by proposing a structure for the two to interact to achieve both real-time behavior and long-term planning. It has done so using a particular modularization based on perception-action loop times, along with a set of message types for coordinating domain knowledge with interaction knowledge. YTTM covers the complete loop from perception to action, and, as such, bridges semantic analysis, situated dialogue, discourse structure (Clark 1992), auditory perception, computer vision (Thórisson, in press), and action selection (Thórisson 1997).

The prototype described implements a number of rules based on selected psychological research spanning the last 60 years. More of these rules could be added to create a larger repertoire of perceptual states and response types in this prototype. Undoubtedly this would make the agent capable of dealing better with various boundary conditions — it currently has few error recovery mechanisms — and with conditions such as interruptions, hesitations, restarts, reformulations, and an interaction with a higher degree of mixed-initiative dialogue. We also believe that giving the turn system the ability to do strong prediction, 1-2 seconds into the future, would greatly enhance its interactive intelligence. As the examples of performance show however, the model performs reasonably well in typical situations.

Needless to say, given the tall order of creating a natural ("manual free") interactive dialogue system, significant testing remains to be done to map out the boundaries and limitations of the YTTM. This is not straightforward since any such model postulates dependencies on context: Internal knowledge states, external real-time events, pending plans and current states of the bodies of the participants, as well as their overarching goals. The hypotheses proposed are well suited for empirical testing of how the model relates to human cognitive mechanisms. Future work includes extensions to the agent's knowledge and domain action capabilities, as well as giving it a full body. To explore the practical applications of YTTM it can be employed in settings with greatly varying sensory capabilities, such as keyboard and mouse, speech-only, and speech, mouse and keyboard. These, and other variations on the model, are currently being explored with promising initial results.

### *5.6. Acknowledgments*

I did the bulk of this research while at the M.I.T. Media Lab and LEGO Digital. I am grateful for their support, and the other sponsors of this work: HUMANOID sf., Thomson-CSF and TSG Magic. The work owes a lot to interactions with Richard A. Bolt, Steve Whittaker, Tom Malone, Lynn Walker, Justine Cassell, and last but not least, Pattie Maes. To them I am thankful. I would also like to thank my colleagues

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

Chris Johnson, Rich Cullingford, Inger Karlsson and John DiPirro for helpful comments on this paper.

## 6. REFERENCES

Adler, R. (1989). Blackboard Systems. In S. C. Shapiro (ed.), *The Encyclopedia of Artificial Intelligence*, 2nd ed., 110-116. New York, NY: Wiley Interscience.

Boff, K. R., L. Kaufman, & J. P. Thomas (eds.) (1986). *Handbook of Human Perception.* New York, New York: John Wiley and Sons.

Bryson, J. & K. R. Thórisson (in press). Dragons, Bats and Evil Knights: A Chatacter-Based Approach to Constructive Play. Submitted to *Virtual Reality, Special Issue on Intelligent Agents*. London: Springer.

Cahn, J. E. & S. E. Brennan (1999). A Psychological Model of Grounding and Repair in Dialog. *Proceedings of the Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*, Sea Cliff, Massachusetts, November 5-7, 25-33.

Card, S. K., T. P. Moran, & A. Newell (1983). *The Psychology of Human-Computer Interaction.* Hillsdale, New Jersey: Lawrence Earlbaum Associates.

Cassell, J. & K. R. Thórisson (1999). The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence*, 13 (4-5), 519-538.

Clark, H. H. (1992). *Arenas of Language Use*. Chicago, Illinoi: University of Chicago Press.

Clark, H.H. & E. F. Schaefer (1989). Contributing to Discourse. *Cognitive Science*, 13:259-294.

Dodhiawala, R. T. (1989). Blackboard Systems in Real-Time Problem Solving. In Jagannathan, V., Dodhiawala, R. & Baum, L. S. (eds.), *Blackboard Architectures and Applications*, 181-191. Boston: Academic Press, Inc.

Duncan, S. Jr. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, 23(2), 283-292.

Effron, D. (1941/1972). *Gesture, Race and Culture*. The Hague: Mouton.

Ekman, P. & W. Friesen (1969). The Repertoire of Non-Verbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica*, 1, 49-98.

Goodwin, M. H. & C. Goodwin (1986). Gesture and Coparticipation in the Activity of Searching for a Word. *Semiotica*, 62(1/2), 51-75.

Goodwin, C. (1981). *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY: Academic Press.

Goodwin, C. (1986). Gestures as a Resource for the Organization of Mutual Orientation. *Semiotica*, 62(1/2), 29-49.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, Massachusetts: Harvard University Press.

Grosz, B. J. & C. L. Sidner (1986). Attention, Intentions, and the Strucutre of Discourse. *Computational Linguistics*, 12(3), 175-204.

Kahneman, D. (1973). *Attention and Effort*. New Jersey: Prentice-Hall, Inc.

Kleinke, C. (1986). Gaze and Eye Contact: A Research Review. *Psychological Bulletin*, 100(1), 78-100.

Kosslyn, S. M. & O. Koenig (1992). *Wet Mind: The New Cognitive Neuroscience*. New York, New York: The Free Press.

Lenat, D. B. (1995). Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38(11).

Maes, P. (ed.) (1990a). *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*. Cambridge, MA: MIT Press/Elsevier.

Maes, P. (1990b). Situated Agents can have Goals. In P. Maes (ed.), *Designing Autonomous Agents*, 49-70. Cambridge, MA: MIT Press.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: University of Chicago Press.

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.

Nespolous, J-L & Lecours, A. R. (1986). Gestures: Nature and Function. In J-L Nespolous, P. Perron & A. R. Lecours (eds.), *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, 49-62. Hillsdale, NJ: Lawrence Earlbaum Associates.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Nii, P. (1989). Blackboard Systems. In A. Barr, P. R. Cohen & E. A. Feigenbaum (eds.), *The Handbook of Artificial Intelligence*, Vol. IV, 1-74. Reading, MA: Addison-Wesley Publishing Co.

Pierrehumbert, J. & J. Hirschberg (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. In P. R. Cohen, J. Morgan & M. E. Pollack (eds.), *Intentions in Communication*. Cambridge: MIT Press.

Rimé, B. & Schiaratura, L. (1991). Gesture and Speech. In R. S. Feldman & B. Rimé, *Fundamentals of Nonverbal Behavior*, 239-281. New York: Press Syndicate of the University of Cambridge.

Sacks, H., Schegloff, E. A.. & Jefferson, G. A. (1974). A Simplest Systematics for the Organization of Turn-Taking in Conversation. *Language*, 50, 696-735.

Sacks, H. (1992). *Lectures on Conversation, vol II*. Cambridge, MA: Blackwell.

Schegloff, E. A. & H. Sacks (1973). Opening up Closings. *Semiotica*, 7, 289-327.

Selfridge, O. (1959). Pandemonium: A Paradigm for Learning. *Proceedings of Symposium on the Mechanization of Thought Processes, 1959*, 511-29.

Sommer, R. (1959). Studies in Personal Space. *Sociometry*, 23, 247-260.

Taylor, T. J. & D. Cameron (1987). *Analysing Conversation: Rules and Units in the Structure of Talk*. Oxford, England: Pergamon Press.

Thórisson, K. R. (in press). Machine Perception of Embodied, Real-Time, Multimodal Dialogue. To be published in P. McKevitt (ed.), *Language, Vision and Music*.

Thórisson, K. R. (1999). A Mind Model for Multimodal Communicative Creatures & Humanoids. *International Journal of Applied Artificial Intelligence*, 1999, Vol. 13 (4-5), 449-486.

Thórisson, K. R. (1998). Decision Making in Real-Time Face-to-Face Multimodal Communication. *Second ACM International Conference on Autonomous Agents '98*, Minneapolis, Minnesota, May 12-15.

Thórisson, K. R. (1997). Layered, Modular Action Control in Communicative Humanoids. *Proceedings of Computer Graphics Europe '97*, June 5-7, Genieva, 134-143.

Thórisson, K. R. (1996). Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. Ph.D. Thesis, Massachusetts Institute of Technology, U.S.A.

Walker, M. & Whittaker, S. (1990). Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*.

Whittaker, S., S. E. Brennan & H. H. Clark (1991). Co-ordinated Activity: An Analysis of Interaction in Computer-Supported Co-operative Work. *Proceedings of Conference on Computer Human Interaction*, 361-367.

Whittaker, S. & Stenton, P. (1988). Cues and Control in Expert-Client Dialogues. *Proc. 26th Annual Meeting of the Association of Computational Linguistics*, 123-130.

Yngve, V. H. (1970). On Getting a Word in Edgewise. *Papers from the Sixth Regional Meeting*., Chicago Linguistics Society, 567-78.

## 7. AFFILIATION

*Kris R. Thórisson*
*Communicative Machines Inc.*
*131 E 23rd St., suite 2C*
*New York, NY 10010*
*http://www.cmlabs.com*

*Multimodality in Language and Speech Systems*, 173-207.
Björn Granström, David House & Inger Karlsson (Editors).
The Netherlands: Kluwer Academic Publishers, Dordrecht, 2002.