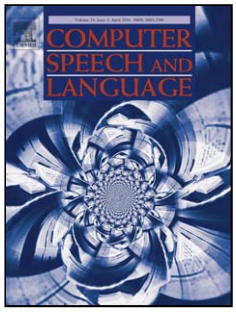
Some background on Dialogue Management and Conversational Speech for dialogue systems

Yorick Wilks, Roberta Catizone, Simon Worgan, Markku Turunen



Title: Some background on Dialogue Management and Conversational Speech for dialogue systems

Authors: Yorick Wilks, Roberta Catizone, Simon Worgan, Markku Turunen

PII: S0885-2308(10)00019-7
DOI: doi:10.1016/j.csl.2010.03.001
Reference: YCSLA 444

To appear in:

Received date: 8-9-2009
Accepted date: 9-3-2010

Please cite this article as: Wilks, Y., Catizone, R., Worgan, S., Turunen, M., Some background on Dialogue Management and Conversational Speech for dialogue systems, *Computer Speech & Language* (2008), doi:10.1016/j.csl.2010.03.001

Some background on Dialogue Management and Conversational Speech for dialogue systems

Yorick Wilks, Roberta Catizone, Simon Worgan and Markku Turunen

1. Introduction

This special issue of the Journal is concerned with speech and language processing issues in the overall environment of end-to-end dialogue systems, and in particular with the sorts of techniques deployed in the COMPANIONS project (www.companions-project.org) which most of the contributors to this issue are associated in one way or another. The aim of the COMPANIONS project was to produce multimodal dialogue agent demonstrators within four years, and the papers in this volume that originate in that project are, in effect, two year prototypes, submitted to evaluations but designed principally as platforms (separately or by a new fusion of components) for further research on the deployment of emotion modelling and of machine learning (ML) techniques of a variety of forms. As will be described, there is already some reportable ML activity in these two-year prototypes.

COMPANIONS was also a much broader concept, embracing both the notion of a new form of conversational interface to the internet, while drawing on some of the traditions of the Embodied Conversational Agent (ECA); this tradition (e.g. Nagao & Takeuchi, 1994 and Traum and Rickel, 2002) has developed rich models going beyond the basis "talking head" of its early days, but is nevertheless not at its heart a form of HLT (Human Language Technology) research and development. It is on this latter strand that the papers in the volume concentrate, along with the assumption that much research on emotion and politeness

is far more dependent on language than its originators realise, and that specifically language and speech phenomena may be the best place to locate emotion and politeness----both crucial to a Companion---- as opposed to say facial expressions and gestures, which are at the core of ECA work.

This initial paper surveys work in two areas: first, Dialogue Management (DM) which is at the core of the language processing system and extends from the understanding of input, in symbolic transcribed form, to decisions based on reasoning as to what to say next, right up to decisions about how to reply. Here we shall concentrate mainly on the core DM itself and its associated knowledge representation and reasoning. Secondly, we shall look very broadly at the speech recognition aspect of conversational speech: this is a very large area and we can only lay out very broad categories of work.

Dialogue systems have been around since the 1960's, the best known are conversation programs such as Eliza (Weizenbaum 1966) and Parry (Colby 1973). The approaches we describe are categorised as follows: finite state/dialogue grammars, plan-based and collaborative; however, this division is not perfect, since any system can in the end be implemented as a finite state system, but the distinction corresponds to design approaches versus implementation approaches, since finite state models can be used to implement a variety of approaches independently of the design choice. Again, collaborative models may or may not be plan-based, so this distinction too, is less than firm.

2. Basic Types of Dialogue Management Systems

2.1 Dialogue Grammars and Frames

Dialogue grammars, are systems that identify and represent local or global surface patterns of dialogue or patterns of speech acts (Searle 1969) and their responses. Dialogue grammars, which have a long history (Polany and Scha, 1984; Reichman, 1981; Sinclair and Coulthard, 1975), use prescriptive grammars for pattern sequences in dialogues. The first grammars described the structure of the complete dialogue, from beginning to end, whereas more recent approaches are based on the observation that there are a number of sequencing regularities in dialogues, which are called **adjacency pairs**. It has been proposed that a dialogue is a collection of such pairs (Jefferson, 1972), which describe facts such as that questions are generally followed by answers, proposals by acceptances, denials etc. Digressions and repairs are dealt with by using embedded sequences.

Dialogue grammars are used to parse the structure of a dialogue, just as syntactic grammar rules are used to parse sentences. Phrase-structure grammar rules and various kinds of state machines have been used to implement dialogue grammars. For example the SUNDIAL system, uses a dialogue grammar to engage in dialogue about travel conversations.

Although dialogue grammars have been successfully implemented (Müller and Runger, 1993; Nielsen and Baekgaard, 1992), they have been criticised on the grounds that they lack flexibility both as to deviations in the dialogue as well as portability to other domains.

A significant extension of dialogue grammars are **frame-based approaches**, which have been developed to overcome the lack of flexibility of dialogue grammars. The entities in the

application domain are hierarchically modelled, and the system can control the dialogue according to the requirements of those entities. Hulstijn et al. (1996), for example, who developed a theatre booking system, arranged frames hierarchically to reflect the dependence of certain topics (like the details of the performance the user wants to see) on others. In Veldhuijzen van Zanten (1996), a train timetable enquiry system, a frame structure relates the entities in the domain to one another, and this structure captures the meaning of all possible queries the user can make. The point of frames is to try to capture a whole topic of dialogue: Lemon and Peters (Lemon 2001) is essentially a frame system, as is the COMIC DM system (Catizone et al 2003) where it is combined with a specific central system to increase flexibility of response.

2.2 Plan-based and Collaborative Systems

Plan-based approaches take the view that humans communicate to achieve goals, including changes to the mental state of the listener. Utterances are seen not just as strings but as performing speech acts (Searle, 1969) and are used to achieve these goals. The listener has to identify the speaker's underlying plan and respond accordingly. For example, in response to a customer's question of "Where are the steaks you advertised?", a butcher's reply of "How many do you want?" is appropriate, because the butcher understands the customer's underlying plan to buy the steaks (taken from Cohen (1990)). Plan-based theories of communicative action and dialogue (for example: Allen and Perault, 1980; Appelt, 1985; Cohen and Levesque, 1990) claim that the speaker's speech act is part of a plan and that it is the listener's job to identify and respond appropriately to this plan. Plan-based approaches attempt to model this claim and explicitly represent the (global) goals of the task. Plan-based approaches have been criticised on practical and theoretical grounds. For example, the

processes of plan-recognition and planning are combinatorically intractable in the worst case, and in some cases they are undecidable. Plan-based approaches also lack a sound theoretical basis. There is often no specification of what the system should do, for example, in terms of the kinds of dialogue phenomena and properties the framework can handle or what the various constructs like plans, goals, etc are. Again, a great deal of conversation and dialogue, as the ATT corpora show, are not about planning or tasks **at all**, they are merely conversation and most of this approach is irrelevant.

Conversational Games Theory (Carletta et al., 1995; Kowtko et al., 1991) uses techniques from both discourse grammars and plan-based approaches by including a goal or plan-oriented level in its structural approach. It can be used to model conversations between a human and a computer in a task-oriented dialogue (Williams, 1996).

A (task-oriented) dialogue consists of one or more transactions, each transaction representing a subtask. A transaction comprises a number of conversational games, which in turn consist of an opening move, and (sometimes optional) end move. An example is an INSTRUCTION game which consists of three nested games: an EXPLAINING game, a QUERY-YN game, and a CHECKING game. The CHECKING game, for example, can consist of a QUERY-YN and a REPLY-Y or a REPLY-N.

The approach deals with discourse phenomena such as side sequences, clarifications etc. by allowing games to be have another game embedded within it - a technique which allows for the modelling of the complexity of natural dialogue. This approach also makes clear that

there is no firm distinction between these and frame systems of section 2.1, since plans can be represented as frames since the days of Schank's Planning scripts (Schank 1977).

A variant called collaborative approaches is based on viewing dialogues as a collaborative process. Both partners work together to achieve a mutual understanding of the dialogue. The motivations that this joint activity places on both partners motivates discourse phenomena such as confirmation and clarification - which are also evident in human-to-human conversations, though, of course, all this rhetoric fits equally well into a planning view. Collaborative approaches try to capture the motivations behind a dialogue and the mechanisms of dialogue itself, rather than concentrate on the structure of the task. The beliefs of at least two participants will be explicitly modelled. A proposed goal, which is accepted by the other partner(s), will become part of the shared belief and the partners will work cooperatively to achieve this goal.

In the TRAINS-93 dialogue manager, Traum's (1996) model of conversation agency extended Bratman's et al. (1988) Beliefs Desires Intentions (BDI) agent architecture. In the BDI model, actions in the world affect an agent's beliefs and the agent can reason about its beliefs and thus formulate desires and intentions. Beliefs are how the agent perceives the world, desires are how the agent would like the world to be, and intentions are formulated plans of how to achieve these desires. Traum states two major problems with the BDI model. He argues that an agent's perceptions not only influence its beliefs but also its desires and intentions. Also, the BDI model does not support more than one agent. Traum thus extended the BDI model by incorporating mutual beliefs, i.e. what both agents believe to be true and also let perceptions influence desires and intentions as well as beliefs.

Viewgen (Wilks and Ballim 1991b) is a representational system for modelling agents and their beliefs and goals as part of a dialogue system. It has two types of structures: those for agents that can have views of other agents and entities, and those for entities that have no points of view of their own. It is based on a virtual machine that nests these entities to any depth required for analysis by nesting either type of object inside the first type: i.e. agents can have perspectives of entities and other agents. The important notion is that nested beliefs (about beliefs and goals) are created only at need and not prestored in advance as in the Cohen, Allen, Perrault-type systems above that compute over goals and beliefs. Other related approaches include Novick and Hansen (1995), Novick and Ward (1993), Chu-Carroll (1996), who extends Sidner's (1992, 1994), and Beun (1996).

3. DM Architectures

3.1 SmartKom

SmartKom (Alexandersson and Becker 2001) is a multimodal dialogue system that combines, speech, gesture and mimics input and output within an overall DM architecture of a Blackboard type, called here a “pool” architecture. One of the major scientific goals of SmartKom is to design new computational methods for the seamless integration and mutual disambiguation of multimodal input and output on a semantic and pragmatic level.

The SmartKom Architecture

- Interface modules: on the input side: there is an audio module, on the output side the display manager

- Recognizers and synthesizers: on the input side, there is gesture recognition, prosody and speech recognition modules, on the output side speech synthesis and the display manager.
- Semantic processing modules: this group of modules comprises or transforms meaning representations: gesture and speech analysis, media fusion, intention recognition, discourse and domain modelling, action planning, presentation planning and concept-to-speech generation.
- External services: the function modelling module is the interface to external services, e.g. EPG databases, map services, and information extraction from the Web.

The discourse module receives hypotheses directly from the intention analysis module. The hypotheses are validated and enriched with (consistent) information from the discourse history. During this process a score is computed which mirrors how well the hypothesis fits the history. Depending on the scores by the analysis modules and the score by the discourse modeller, the intention analysis module picks the "best" hypothesis.

3.2 Trindi

The Trindi project (Larsson 2000) proposes an architecture and toolkit for building dialogue managers based on an *information state* and *dialogue move engine.* The Information state of a dialogue represents the information necessary to distinguish it from other dialogues, representing the cumulative additions from previous actions in the dialogue and motivating future action. It can be seen as an attempt to make a finite state system more plausible as a

general architecture for DM, when combined with other components expressing the overall "information-state" of the system.

Trindi offers a platform for the formalization of the notion of an information state which allows specific theories of dialogue to be formalized, implemented, tested, compared and iteratively reformulated. Key to this approach is the notion of UPDATE of the information state with most updates related to the observations and performance of DIALOGUE MOVES.

The Information State Theory of Dialogue Modelling consists of

- A description of the **informational components** of the theory of dialogue modelling including common context and internal motivating factors (common ground, commitments, beliefs, intentions, etc.)
- **Formal representations** of the above components (e.g. as lists, sets, typed feature structures, records, etc)
- A set of **dialogue moves** that will trigger the update of the information state (also correlated with externally performed actions such as particular natural language utterances).
- A set of **update rules** that govern the updating of the information state given various conditions of the current information state and performed dialogue moves including a set of selection rules.
- An **update strategy** for deciding which rule(s) to select at a given point from the set of applicable ones.

There is an important distinction between information state approaches (Cooper and Larsson 1999) and dialogue state approaches. In a dialogue state approach, a dialogue behaves according to some grammar where the states represent the results of a dialogue move in a previous state and each state has a set of allowable moves. The "information" is implicit in the state itself and the relationship it plays to other states.

Here, the **informational components** are not conceived of as monolithic nodes in a transition network (as with dialogue state, but rather as consisting of several interacting components). One could model the mental state of an agent or take a more structural view and model the performance of actions. The **formal representations** for modelling various aspects of the dialogue structure range from simple abstract data types to more complex informational systems such as logics.

3.3 WITAS

This system (Lemon 2001) contains a dialogue interface for multi-modal conversations to the WITAS robot helicopter. The requirements of this dialogue system are:

- Asynchronous
- Mixed-Initiative
- Open-ended
- Involves a dynamic environment

The Dialogue Manager creates and updates an Information State corresponding to a notion of dialogue context. Dialogue moves have the effect of updating information states and moves can be initiated by both the operator and the robot. This system can be seen as a dialogue state/information state hybrid that began with a stack structure like COMIC (see below).

They dropped this arguing it was too restrictive because navigation back and forth between different sub-dialogues and topics was impossible because information was lost when the stack was popped. To compensate for this, they implemented Version II of the dialogue management system which uses a tree structure of dialogue states (*dialogue move tree*), where edges are dialogue moves and branches represent conversational threads. They also wanted to enrich their domain knowledge and inference methods so they implemented a dynamic hierarchical *task tree.* The *task-tree* grows as part of the developing dialogue context and represents tasks and sub-tasks described by the operator and their temporal ordering. This structure allows for reordering and reference to tasks. They also implemented an inference-based model of the robot's changing abilities, which depends on dynamic information about the world and the robot's internal state and location.

3.4 CONVERSE

CONVERSE (Levy et al. 1997) ,a machine dialogue system funded by Intelligent Research of London , won the Loebner prize in 1997. That year was the first in which there was no restriction on the topic of discussion with judges, and CONVERSE covered about 80 topics, which were appropriate to its persona as a young female New York-based journalist. It embodied substantial resources, such as WordNet, the proper names of Collins dictionary etc. It could store the personal information it elicited from a user and build it into the conversation later. Its topic structures were complex ATN scripts that could be left and reentered appropriately and could generate responses using stored/elicited material.

It had no conventional analysis/DM/generation division, though it used a commercial ,statistically based parser to pass input to the ATN's. Its control structure was implemented as

a simple blackboard system in which the ATNs competed to take control of the generation; these decisions were made numerically based on weights assigned by the closeness of fit of the input to the expected input etc. The system had only limited recovery mechanisms if it was not able to find a topic relevant to the input, and relied on seizing control of the conversational initiative as much as it could. Since this system models only plausible conversation, the dialogue had no application goals of any kind.

3.5 COMIC

COMIC was a Framework Five funded IST project (ended in 2004) which applied research in human-human interaction to human-computer interaction. The application of COMIC was bathroom design and it contained speech and gesture input/output with the use of an avatar to generate facial emotion. The DM in COMIC was designed at the University of Sheffield as a general-purpose dialogue management system, designed so that the domain data is separate from the DM control mechanism. The domain data is expressed using Dialogue Action Forms (DAFs) which are augmented transition networks – a series of nodes and their connected arcs containing tests and the corresponding actions. In order to create and modify the DAFs, a GUI editor (DAF editor) was developed. With the DAF editor, it is a straight forward process to create and modify DAFs. This allows for a, relatively self-contained, way to maintain the domain data in a dialogue management system. This method of separating domain data along with the visual aid of editing using a graphical representation is novel. COMIC's information sturctures were modelled on those of the higher functionality CONVERSE system, excluding the blackboard architecture, but with the flexibility of a stack system to allow reaccess to "pushed" structures, arguing that the losses experienced with this method in the WITAS system (above) were fact acceptable.

The general purpose nature of the DM means that it could easily be accommodated to other Dialogue systems with a minimum of application specific reorganisation.

The most important features of the DM in COMIC are:

- It is general purpose;
- It can be re-used in other applications with minimal changes/effort;
- It is able to handle different types of Dialogue Management such as user driven, system driven and mixed initiative dialogues;
- It is able to handle different Dialogue Styles;
- It can deal with topic shift nd topic recovery;
- It includes multi-leveled error handling;

3.6 Agent-based dialogue management

A great deal of work has been done in the field of dialogue management to achieve flexible and robust interaction with compact software agents, and this can be seen as an extension of distributed DM architectures such as Communicator (http://communicator.sourceforge.net/index.shtml) in the US These approaches include the agenda-based dialogue management architecture (Rudnicky et al., 1999) and its RavenClaw extension (Bohus & Rudnicky, 2003), Queen's Communicator (O'Neill et al., 2003), SesaME (Pakucs, 2003) and Jaspis (Turunen & Hakulinen, 2000; Turunen et al., 2005a). In these approaches, dialogue management is often implemented using the object-oriented approach.

Most importantly, inheritance is used to separate generic dialogue management from domain specific actions.

The modular agent-based approach to dialogue management makes it possible to combine the benefits of different dialogue control models, such as state-based dialogue control and frame-based dialogue control. Similarly, the benefits of alternative dialogue management strategies, such as the system-initiative approach and the mixed-initiative approach (Walker et al., 1998), can be used together in an adaptive way. Using multiple agents for the same purpose makes it possible to combine rule-based and machine learning approaches (Turunen, 2004).

In the Jaspis architecture dialogue agents are used for various adaptive features. For example, in the AthosMail application (Turunen et al., 2004) dialogue control is performed using two approaches to make the system robust for different users. The first approach uses agents for pragmatic processing and sense annotation, while the second approach utilizes numerous specialized dialogue agents to make multilingual interaction possible (Salonen et al., 2004). In the timetable domain, agent-based dialogue management approach is used to implement features such as truly mixed-initiative dialogues (Turunen et al., 2005b), and multimodal guidance to help novice users to interact with the system and bring system-initiative features to the user-initiative interface (Hakulinen et al., 2005).

In the area of speech-based pervasive computing systems the agent-based approach has been used to implement distributed, concurrent, open-ended, dynamically constructed dialogues that can involve multiple participants (Turunen & Hakulinen, 2003). For example, agents have been used to distribute multimodal dialogues between the server and mobile devices

(Salonen et al., 2005; Turunen et al., 2005c), and implement pervasive speech-based and auditory dialogues with technology embedded in the environment (Kainulainen et al., 2005).

There have also been applications in DM about the notion of an autonomous agents based on BDI which was originally introduced as an alternative to full planning that could balance reactive and deliberative behaviour (Bratman, Israel & Pollack'88). BDI has been independently developed as a dialog manager around the world (Ardissono'98, Wallis'01) and claims the advantage that it does intentional behaviour and plan failure in a psychologically plausible manner.

4 DM and ASR language modelling

The present situation in dialogue modeling is in some ways a replay, at a lower level, of the titanic struggle in the early 1990's between linguistic models and the data-driven approach to NLP introduced by Jelinek in MT. The introduction into ASR of so called "language models" –which are usually no more than corpus bi-gram statistics to aid recognition of words by their likely neighbours---has caused some, like Young (2002) to suggest that simple extensions to ASR methods could solve all the problems of language dialogue modeling.

Young describes a complete dialogue system seen as what he calls a Partially Observable Markov process, of which subcomponents can be observed in turn with intermediate variables and named:

- Speech understanding
- Semantic decoding
- Dialogue act detection

- Dialogue management and control
- Speech generation

Such titles are close to conventional for an NLP researcher, e.g. when he intends the third module as something that can also recognise what we may call the *function* of an utterance, such as being a command to do something and not a pleasantry. Such terms have been the basis of NLP dialogue pragmatics for some thirty years, and the interesting issue here is whether Young's Partially Observable Markov Decision Process, are a good level at which to describe such phenomena, implying as it does that the classic ASR machine learning methodology can capture the full functionality of a dialogue system, when its internal structures cannot be fully observed, even in the sense that the waves, the phones and written English words can be. The analogy with Jelinek's MT project holds only at its later, revised stage, when it was proposed to take over the classic structures of NLP, but recapitulate them by statistical induction. This is exactly Young's proposal for the classic linguistic structures associated with dialogue parsing and control with the additional assumption, not made earlier by Jelinek, that such modular structures can be learned even when there are no distinctive and observable input-output pairs for the module that would count as data by any classic definition, since they cannot be word strings but symbolic formalisms like those that classic dialogue managers manipulate. Young assumes roughly the same intermediate objects as linguists but in very simplified forms. So, for example, he suggests methods for learning to attach Dialogue Acts to utterances but by methods that make no reference to linguistic methods for this (since Samuel et al. 19w98) and, paradoxically, Young's equations do not make the Dialogue Acts depend on the words in the utterance, as all linguistic methods do. His overall aim is to obtain training data for all of them so the whole process becomes a single throughput Markov model, and Young concedes this model may only be for simple domains, such as, in his example, a pizza ordering system.

All parties in this dispute, if it is one, concede the key role of machine learning, and all are equally aware that structures and formalisms designed at one level can ultimately be represented in virtual machines of less power but more efficiency. In that sense, the primal dispute between Chomsky and Skinner about the nature of the human language machine was quite pointless, since Chomsky's transformational grammars could be represented, in any concrete and finite case, such as a human being, as a finite state machine.

All that being so, researchers have firm predelictions as to the kinds of DM design within which they believe functions and capacities can best be represented, and, in the present case, it is hard to see how the natural clusterings of states that form a topic can be represented in finite state systems, let alone the human ability to return in conversation to a previously suspended topic, all matters that can be represented and processed naturally in well understood virtual machines above the level of finite state matrices.

There is no suggestion that a proper or adequate discussion of Young's views has been given here, only a plea that machine learning must be possible over more linguistically adequate structures than finite state matrices if we are to be able to represent, in a perspicuous manner, the sorts of belief, intention and control structures that complex dialogue modeling will need; it cannot be enough to always limit ourselves to the simples applications on the grounds, as Young puts it, that the typical system S will typically be intractably large and must be approximated.

The future of DM will in part be a reaction to this territorial dispute between ASR and NLP paradigms, but all will agree that the issues remain 1) the extent to which DM data can be learned, and by more sophisticated methods than the reward structures of Walker and Pieraccini (Walker et al., 1998); 2) the ways in which evaluation methods for dialogue systems, and DM in particular, can be evaluated and 3) the extensions to our concept of dialogue that will be needed to deal with distributed dialogues, over time and space, with computers that will come with the spread of small, embedded, "ubiquitous" devices.

5 Conversational Speech

The vast majority of speech is perceived within the context of dialogue, given this context it remains perverse that many approaches within automatic speech recognition (ASR) "derive almost entirely from the study of monologue."(Pickering, 2004) This section will attempt to address this oversight by providing a review of the study of speech within the wider context of dialogue. Motivated by clear indications that the production and perception of the signal is as vital as the text in judging the quality of an interaction (Gregory et. al., 1997) we will consider not only the practical challenges of constructing a conversational system, Section 5.1, but also the role of 'turn-taking', Section 5.2 and paralinguistic cues, Section 5.3. We will then summarize and unify these features by arguing, in Section 5.4, that spoken language processing cannot be viewed as the passive perception of an abstract signal. Rather, speech should be viewed as an intentional act, which is comprehended as such by both speaker and listener. To capture this comprehension we propose that speech should be placed in the context of a wide range of `interaction affordances', expanding upon the direct realist (Gibson, 1986) position, capturing the perception of environmental affordances. As a result of this intentionality these affordances are open to manipulation and as such require the

perceiver to model production in order to allow direct perception to proceed. We then conclude by proposing that this approach has consequences for all aspects of spoken language processing within dialogue, ranging from phonetic perception through to emotional manipulation.

5.1 Speech, data and Artificial Dialogues

When we attempt to construct artificial conversational systems the addition of speech presents a challenge not only for the ASR but also for the system as a whole. As highlighted by Schafer et. al. (2000) recorded training data must strike a balance between the natural rapport of conversational speech while remaining confined to the scenario under consideration. It is not sufficient to simply record unscripted interactions "as recordings of unscripted speech do not readily yield the carefully controlled contrasts required for many research purposes." (Schafer et. al., 2000). Conversely, acted interactions are frequently insufficient as, demonstrated by the emotion in speech community (Burkhardt et. al.,2005), actors fail to capture the prosodic and spectral features present in normal speech. Furthermore, as shown by WOZ approaches (Moore and Morris, 1992), human-machine interactions will result in a different set of behaviors when compared to human-human interactions. Frequently, for example, humans will begin to mimic the acoustic features of the artificial text-to-speech (TTS) system in an attempt to establish rapport. Accordingly, all data collection tasks should be tailored for the conversational scenario under consideration as each scenario can present different properties. For example, "picture description tasks have revealed much about the generation of syntactic and thematic structure (Bock and Loebell, 1990);" while "descriptions of networks of colored nodes have supplied a wealth of data on

aspects of the planning, sequencing, and repair of utterances (Levelt and Cutler, 1983)." (Schafer el al. 2000)

All of these factors present a challenge when training conversational systems, but the recursive nature of dialogue also allows us to exploit a number of beneficial aspects. As highlighted by (Thorisson,1997) we should "keep in mind that the agent can always ask the user a question when the data doesn't make "sense"' and in the context of artificial dialogues an utterance doesn't make sense if it cannot be grounded in the systems dialogue manager (Lauria, 2007). The system only needs to resolve uncertainty to the point where it is able to act and reply in a meaningful fashion. As a result, certain ASR errors can be dismissed as incidental to the flow of conversation and multiple phrase hypotheses can be resolved by selecting that which is most likely to produce a meaningful response. Ultimately, an effective symbiosis can be established between the dialogue manager and ASR system as the previously selected utterance can suggest the nature of the users reply, while the ASR system can focus upon the phrases and keywords required to sustain the conversation.

5.2 Turn taking

At its simplest turn-taking within conversational systems has been sharply delineated with system and user literally 'taking it in turns'. (Field et. al., 2009) By comparison, it is clear that face-to-face communication involves overlapping utterances, seamless transitions (Sacks, 1974), and it is frequently the case that "pauses across turns are sometimes even shorter than pauses within a turn itself" (Cassell et. al., 1999). Clearly then straightforward pause detection will not be sufficient to sustain turn-taking behavior.

To establish rapport we need to both indicate to the user when it is their turn to speak and to detect when the user is willing to give up their turn. As shown by previous multi-modal dialogue system (Casell et. al., 1999; Thorisson,1997) this behavior can usually be achieved by a fusion of gesture, gaze, and intonation. Focusing upon intonation a number of auditory features can be proposed; these include falling pitch at the end of a sentence and lengthening of the final syllable (Duncan, 1972). In combination with the termination of the current gesture and specific gaze behaviors (away from the listener at the start, towards at the end) we can begin to account for the fact that: "The time between the exchange of turns is often too short to be explained as the result of the hearer's waiting for the speaker to finish before the hearer starts to speak" (Cassell et al., 1999). Furthermore, once an artificial system has developed a reasonable approach to turn taking behavior we can make further progress by determining the nature of each turn "by regarding the intonation pattern over a full utterance … determining whether an utterance could be a filler (relatively short and flat pitch pattern), question (final rise) or command (final fall)" (Thorisson, 1996).

5.3 Intonation and paralinguistic context

Beyond turn-taking, intonation within the speech signal can also help to inform the dialogue manager when new information (the rheme) is introduced into the current conversation (the theme). As demonstrated by Lowuerse et. al. (2008) "the average pitch of the rheme in a turn is significantly higher than the average pitch of the phrasal theme of that turn", showing the relation between information and intonation structure. Within the complex interaction between turn, theme and rheme, we can also judge the level of agreement within the interaction as shown by Steedman (2003), Table 5.1.

	Agree	Disagree
Theme	L+H*	L*+H
Rheme	H* or (H*+L)	L* or (H+L*)

Table 5.1: Judging the level of agreement and disagreement within the theme and rheme of a dialogue turn. H/L corresponds to the high/low tone of the utterance while * designates that it is aligned with a stressed syllable.

These features can have powerful implications for any proposed ASR system within a conversational context, as shown by Manusov and Trees (2002) "even when controlling for what a person said, the messages sent by nonverbal cues could all predict subsequent account forms, although not always in the way expected. These results help our argument that nonverbal cues may be an important part of moving through account sequences, both on their own and when combined with verbal utterances."

5.4 The recursive nature of conversational speech

It is clear when considering all of these distinctive acoustic features that speech can be viewed as a recursive process between two or more participants, with each establishing a shared understanding with the whole. Mimicry is the clearest example of this process, as demonstrated by Gregory et. al. (1997) "first, that partners do actively accommodate the fundamental frequency of their voices, and, second, that elimination of the fundamental

frequency from conversation partners' voices profoundly alters perceived positive evaluations by judges overhearing the conversation." Furthermore, as shown by Parrill and Kimbara (2006) "Participants who observed more mimicry reproduced more of the mimicked features in their descriptions—despite the fact that these cases of mimicry were quite subtle—indicating a high degree of sensitivity to mimicry." This recursive process then presents challenges and opportunities for both TTS and ASR. Within TTS there is a need to capture this mimicry in the production of the fundamental frequency, end-of-turn indicators, topic shift, and emotive indications. Within ASR there is a requirement to take into account the role and intentionality of the user, requiring new, conversational, systems to be developed.

For example the PRESENCE ASR system demonstrates this new approach. In previous work, (Moore,2007) the difference between PRESENCE and traditional approaches to ASR has been parodied as two different approaches to heating a room. In the first, a thermostat is installed and the system adjusts the heating according to the deviation from the desired temperature. In the second, a wide range of factors pose a fundamental challenge for our engineer, doors are opened, people enter the room, ambient temperature changes over time. To account for this sensors are fitted to doors and windows and a wide variety of statistical heating models are proposed to account for noise and variance. Why has the speech community taken the second approach? We propose that it is because we have not yet invented the `thermometer'.

Within the framework of conversation this `thermometer' can be thought of as an error signal derived from the intentionality of the listener. The participants of a conversation are seeking to fulfill some purpose, taken from the set of currently available interaction affordances, and

it is the mismatch between the desired perceptual state and the current perceptual state that allows the listener to actively, continuously, refine the process of perception. These refinements are possible because the listener models the speaker's intentionality and conversely the speaker models the listener; far from traditional symbolic conceptions of perception this continuous recursive process establishes an unbroken loop between perception and production within the context of a dialogue.

5.5 Emotion detection and manipulation

One clear consequence of this error signal is that we can begin to take into account the emotional state of the user, emotional deviations from the norm (aggression, joy, etc.) all have implications for any conversational ASR system as these deviations are often expressed within the speech signal. Accordingly, we need to detect the emotional tone of a conversation for two reasons; firstly, emotional deviations need to be accounted for by the ASR system; secondly, the detection and classification of emotion can inform the dialogue manager and so help sustain the flow of conversation. Previous work (Oudeyer, 2003; Vogt et. al., 2008) demonstrates that a reasonably high rate of emotion classification can be achieved by considering the logarithmised pitch, signal energy, Mel-frequency cepstral co-efficients, the short-term frequency spectrum, and the harmonics-to-noise ratio (HNR).

By placing established emotion detection methods within the recursive nature of conversation we can consider discourse as the exploitation of the shared set of interaction affordances (Worgan and Moore, 2009), that is to say what the conversation 'affords' each participant (Gibson, 1986). Emotion can be seen as the manipulation of the range of interaction affordances available to the agent. Within conversation my emotional state affects your

emotional state and your emotional state changes the set of interaction affordances that are available to me. In these terms emotion becomes a strategy for goal fulfillment when coupled with an understanding of the space of possible actions afforded by another. Within artificial dialogue systems this then allows for action selection over a space of possible utterances; selecting the utterance that results in the maximization of the user's emotional state (Publication in this journal, 2009).

6.References

Abelson, R and Schank, R. (1977). "Script Plans Goals and Understanding", Lawrence Erlbaum Associates.

Alexandersson, J. and Becker, T. "Overlay as the Basic Operation for Discourse Processing in a Multimodal Dialogue System", in Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Seattle, August 2001.

Allen, J.F. and Perault, C.R. "Analyzing Intentions in Dialogues", in Artificial Intelligence, 15(3):143-178, 1980

Appelt, D.E. "Planning English Sentences", Cambridge, University Press, 1985.

Ardissono, L and G. Boella "An Agent Architecture for NL dialog modeling" in Artificial Intelligence: Methodology, Systems and Applications, Springer (LNAI 2256) 1998

Beun, R.J. "Speech Act Generation in Cooperative Dialogue", in Luperfoy et al. (1996).

Bock, J.K., & Loebell, H. Framing sentences. Cognition (1990) vol. 35 pp 1-39.

Bohus, D., Rudnicky, A. "RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda", In Proc. of the Eurospech 2003: 597-600, 2003.

Bratman, M.E., D.J. Israel, and M.E. Pollack, "Plans and Resource-Bounded Practical Reasoning", in Computational Intelligence, 4, 1988

Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B., (2005) "A Database of German Emotional Speech.",Proc. Interspeech

Carletta, J.H. Carletta, A. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, "The Coding of Dialogue Structure in a Corpus", in Andernach et al. (1995).

Cassell J., Torres O., Prevost S., (1999) "Turn taking vs. discourse structure: How best to model multimodal conversation", Machine Conversations pp. 143-154.

Catizone, R., Setzer, A. and Wilks, Y. "Multimodal Dialogue Management in the COMIC", Project,Workshop on 'Dialogue Systems: interaction, adaptation and styles of management', European Chapter of the Association for Computational Linguistics(EACL), Budapest, Hungary, April 2003

Chu-Carroll, J. "Response Generation in Collaborative Dialogue Interactions", in Luperfoy et al. (1996).

Cohen, P.R. and Levesque, H. "Rational Interaction as the Basis for Communication", in Cohen et al. (1990).

Cohen, P.R. Morgan, J. and Pollack, M.E. (eds), "Intentions in Communication", MIT Press, Cambridge, Massachusetts.

Colby, K. "Simulations of Belief Systems", In Schank and Colby (1973).

Cooper, R. Larsson, S., Matheson, C., Poesio, M. and Traum, D. "Coding in Structural Dialogue for Information States". Deliverable D1.1. Trindi Project, 1999.

Duncan, S. D., Jr. (1972) "Some signals and rules for taking speaking turns in conversations", Journal of Personality and Social Psychology vol. 23. pp 283-292

Field, D., Catizone, R., Cheng, W., Dingli, A., Worgan, S., Ye, L. and Wilks, Y. (2009) The Senior Companion: a Semantic Web Dialogue System. Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, 10-15 May 2009.

Gibson, J. (1986) "The Ecological Approach to Visual Perception", Lawrence Erlbaum Associates

Gregory, S. W., Dagan, K., Webster, S., (1997) "Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality", Journal of Nonverbal Behavior vol. 21 (1) pp. 23-43

Hakulinen, J., Turunen, M., Salonen, E.-P. "Software Tutors for Dialogue Systems", In Proc. of Text, Speech and Dialogue (TSD 2005): 412-419, 2005.

Huilstijn, J. Streetskamp, R. ter Doest, H., van de Burgt, S. and Nijholt, A. "Topics in SCHISMA Dialogues", in Luperfoy et al. (1996).

Jefferson, G. "Side Sequences", in Sudnow (1972).

Kainulainen, A., Turunen, M., Hakulinen, J., Salonen, E.-P., Prusi, P., Helin, L. "A Speech-based and Auditory Ubiquitous Office Environment", Proc. of 10th International Conference on Speech and Computer (SPECOM 2005): 231-234, 2005.

Kowtko, J.C. Isard, S.D. and Doherty, G.M. "Conversational Games within Dialogue", in Proc. of the ESPRIT Workshop on Discourse Coherence, University of Edinburgh, 1991.
Kruijff-Korbayová , I., Steedman, M. (2003) Discourse and Information Structure, Journal of Logic, Language and Information, vol. 12, pp. 249-259.
Larsson, S. and Traum, D. "Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit", in NLE Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering, 2000.

Lauria, S., (2007) "Talking to Machines: Introducing Robot Perception to Resolve Speech Recognition Uncertainties", Circuits Systems Signal Processing vol. 26 (4) pp. 513-526

Lemon, O., Bracy, A. Gruenstein, A. and Peters, S. "The Witas Multi-Modal Dialogue System I", in Proc. of Eurospeech2001, Aalborg (Denmark), 2001.

Levelt, W. J. M., Cutler A., (1983) "Prosodic Marking in Speech Repair", vol. 2 pp. 205-218.

Levin, E., Narayanan, S., Pieraccini, R., Biatov, K., Bocchieri, E., Di Fabbrizio, G., Eckert, W., Lee, S., Rahim, M., Ruscitti, P. and Walker, M., "The AT&T Darpa

communicator mixed initiative spoken dialog system," ICSLP 2000, Beijing, China, 16-20 Oct. 2000.

Levy, D., Catizone, R., Battacharia, B., Krotov, A. and Wilks, Y. "CONVERSE: A Conversational Companion", in Proc. of the 1st International Workshop on Human-Computer Conversation, Bellagio, Italy, 1997.

Louwerse, M. M., Jeuniaux, P., Zhang, B., (2008) "The Interaction between Information and Intonation Structure: Prosodic Marking of Theme and Rheme", The 30th meeting of Cognitive Science Society, Washington, DC

Manusov, V., Trees, A. R., (2002) "Are You Kidding Me?: The Role of Nonverbal Cues in the Verbal Accounting Process", The Journal of Communication vol. 52 (3) pp. 640-656

Moore, R. K., (2007) "PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction", IEEE Transactions On Computers vol. 56 (9) pp. 1176

Moore, R. K., and Morris, A., (1992) "Experiences collecting genuine spoken enquiries using WOZ techniques", Fifth DARPA Workshop on Speech & Natural Language.

Müller, C. and Runger, F. "Dialogue Design Principles - Key for Usability of Voice Processing", in Proc. of the 3rd European Conference on Speech, Communication, and Technology (EUROSPEECH93), Berlin, Germany, 1993.

Nagao, K. and Takeuchi, A., 1994. "Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation", *ACL 1994*: 102-109

Nielsen A. and Baekgaard, A. "Experience with Dialogue Description Formalism for Realistic Applications", in Proc. of the International Conference on Spoken Language Processing (ICSLP 92), Banff, Canada, 1992.

Novick, D.G. and Hansen, B. "Mutuality Strategies for Reference in Task-Oriented Dialogue", in Andernach et al. (1995).

Novick, D.G. and Ward, K. "Mutual Beliefs of Multiple Conversants: A Computational Model of Collaboration in Air Traffic Control", in Proc. of AAAI'93, 1993.

O'Neill, I., Hanna, P., Liu, X., McTear, M. "The Queen's Communicator: An Object-Oriented Dialogue Manager". In Proc. of the Eurospech 2003: 593-596, 2003.

Oudeyer, P-Y., (2003) "The production and recognition of emotions in speech: features and algorithms", International Journal Of Human-Computer Studies vol. 59 pp. 157-183

Pakucs, B. "Towards Dynamic Multi-Domain Dialogue Processing", In Proc. of Eurospeech 2003: 741-744, 2003.

Parrill, F., Kimbara, I., (2006) "Seeing and Hearing Double: The Influence of Mimicry in Speech and Gesture on Observers", Journal of Nonverbal Behavior vol. 30 (4) pp. 157-166.

Pickering, M., Garrod, S., Barr, D., Keysar, B.,(2004) "Toward a mechanistic psychology

of dialogue", Behavioral and Brain Sciences vol. 27 (2) pp. 169-190

Polany, R. and Scha, R. "A Syntactic Approach to Discourse Semantics", in Proc. of 10th International Conference on Computational Linguistics, Stanford Uni, California, ACL, 1984.

Reichman, R. "Plain-Speaking: a Theory and Grammar of Spontaneous Discourse", PhD thesis, Department of Computer Science, Harvard University, Cambridge, Massachusetts, 1981.

Rudnicky, A. Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu, W., Oh, A. "Creating Natural Dialogs" in the Carnegie Mellon University Communicator System. In Proc. of Eurospeech 1999: 1531-1534, 1999.

Sacks, H., Schegloff, E. A., Jefferson G., (1974) "A simplest systematics for the organization of turn-taking for conversation". Language pp. 696-735

Salonen, E.-P., Hartikainen, M., Turunen, M., Hakulinen, J., Funk, A. "Flexible Dialogue Management Using Distributed and Dynamic Dialogue Control", In Proc. of ICSLP 2004: 197-200, 2004.

Salonen, E.-P., Turunen, M., Hakulinen, J., Helin, L., Prusi, P., Kainulainen, A. "Distributed Dialogue Management for Smart Terminal Devices", In Proc. of Interspeech 2005: 849-852, 2005.

Samuel, K., Carberry, S., Vijay-Shanker, K. "Dialogue Act Tagging with Transformation-Based Learning", Proc. of 17th International Conf. on Computational Linguistics (COLING-ACL '98), 1998.

Schafer, A., Speer, S., Warren, P., White, S., (2000) "Intonational disambiguation in sentence production and comprehension", Journal of Psycholinguistic Research vol. 29 (2) pp. 169-182

Schank, R.C. and Riesbeck, C.K. "Inside Computer Understanding", Hillsdale, NJ: Lawrence Erlbaum

Searle, J.R. "Speech Acts: An Essay in the Philosophy of Language", Cambridge, University Press, 1969.

Sidner, C.L. "Using Discourse to Negotiate in Collaborative Activity: an Artificial Language", in AAAI-92 Workshop "Cooperation Among Heterogeneous Intelligent Systems", 1992.

Sidner, C.L. "An Artificial Discourse Language for Collaborative Negotiation", Proc of AAAI 1994.

Sinclair, J.M. and Coulthard, M. "Towards an analysis of disourse: the English used by teachers and pupils", Oxford University Press, 1975.

Thorisson, K. R. (1996) "Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills", PhD Thesis, MIT, MA

Thorisson, K. R., "Mind Model For Multimodal Communicative Creatures And Humanoid",. International Journal of Applied Artificial Intelligence, vol. 13(4-5), pp 449-486

Traum, D.R. "Conversational Agency: The TRAINS-93 Dialogue Manager", in Luperfoy et al. 1996.

Traum, D.R. and Rickel, J., 2002. "Embodied agents for multi-party dialogue in immersive virtual worlds.",AAMAS 2002: 766-773.

Turunen M., Salonen, E.-P., Hartikainen, M., Hakulinen, J., Black, W., Ramsay, A., Funk, A., Conroy, A., Thompson, P., Stairmand, M., Jokinen, K., Rissanen, J., Kanto, K., Kerminen, A., Gamback, B., Cheadle, M., Olsson, F. and Sahlgren, M. "AthosMail - a Multilingual Adaptive Spoken Dialogue System for E-mail Domain", In Proc. of Workshop on Robust and Adaptive Information Processing for Mobile Speech Interfaces, 2004: 77-86.

Turunen, M. Jaspis – "A Spoken Dialogue Architecture and its Applications", Ph.D. Dissertation, University of Tampere, Department of Computer Sciences A-2004-2, February 2004.

Turunen, M. Salonen, E.-P., Hakulinen, J., Kanner, J., Kainulainen, A. "Mobile Architecture for Distributed Multimodal Dialogues", In Proc. of ASIDE 2005, 2005

Turunen, M., Hakulinen, J. Jaspis, J. "A Framework for Multilingual Adaptive Speech Applications", In Proc. of 6th International Conference of Spoken Language Processing (ICSLP 2000): 719-722, 2000.

Turunen, M., Hakulinen, J. Jaspis, J. "An Architecture For Supporting Distributed Spoken Dialogues", In Proc. of the Eurospeech 2003: 1913-1916.

Turunen, M., Hakulinen, J., Räihä, K.-K., Salonen, E.-P., Kainulainen, A., Prusi, P. "An Architecture and Applications for Accessibility Systems", IBM Systems Journal, Vol 44, (3): 485-504, 2005.

Turunen, M., Hakulinen, J., Salonen, E.-P., Kainulainen, A., Helin, L. "Spoken and Multimodal Bus Timetable Systems: Design, Development and Evaluation", Proc. of 10th International Conference on Speech and Computer (SPECOM 2005): 389-392, 2005.

Veldhuijzen van Zanten, G. "Pragmatic Interpretation and Dialogue Management in Spoken-Dialogue Systems", in Luperfoy et al. (1996).

Vogt, T., Andre, E., Bee, N., "EmoVoice—A Framework for Online Recognition of Emotions from Voice", Perception in Multimodal Dialogue Systems: 4th IEEE Tutorial (2008)

Wahlster, W., Reithinger, N. and Blocher, A. "SmartKom: Multimodal Communication with a Life-Like Character", in Proc. of Eurospeech2001, Aalborg (Denmark), 2001.

Walker, M., Fromer, J., Fabbrizio, G., Mestel, C., Hindle, D. "What can I say? Evaluating

a spoken language interface to Email", In Proc. of ACM CHI 98 Conference on Human Factors in Computing Systems: 582-589, 1998.

Wallis, P., Helen Mitchard, Damian O'Dea and Jyotsna Das, "Dialog Modelling for a Conversational Agent" in AI2001: Advances in Artificial Intelligence, 14th Australian Joint Conference on Artificial Intelligence edited by Markus Stumptner, Dan Corbett and Mike Brooks, Adelaide AU, 2001, Springer (LNAI 2256)

Weizenbaum, J. "ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine", Communications of the Association for Computing Machinery 9, 1966.

Wilks, Y. and Ballim, A. "Beliefs, Stereotypes and Dynamic Agent Modeling", In "User Modeling and User-Adapted Interaction", Vol.1, No. 1, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

Wilks, Y. and Catizone, R. "Human-Computer Conversation", Encyclopedia of Library and Information Science, Vol. 69, Allan Kent (ed.). New York: Dekker, 2001.

Williams, S. "Dialogue Management in a Mixed-Initiative, Cooperative, Spoken Language System", in Luperfoy et al. (1996).

Worgan S. F. and Moore R. K. "Spoken Language Processing as an Aspect of Human Behaviour", Conference on "Grounding language in perception and (inter)action" Wenham MA.

Young, S. "Talking to machines—statistically speaking", Proc. ICSOS02.