'How was your day?'

S. G. Pulman, J. Boye
University of Oxford
sgp@clg.ox.ac.uk

M. Cavazza, C. Smith
Teesside University
m.o.cavazza@tees.ac.uk

R. S. de la Cámara
Telefonica I+D
e.rsai@tid.es

Abstract

We describe a 'How was your day?' (HWYD) Companion whose purpose is to establish a comforting and supportive relationship with a user via a conversation on a variety of work-related topics. The system has several fairly novel features aimed at increasing the naturalness of the interaction: a rapid 'short loop' response primed by the results of acoustic emotion analysis, and an 'interruption manager', enabling the user to interrupt lengthy or apparently inappropriate system responses, prompting a replanning of behaviour on the part of the system. The 'long loop' also takes into account the emotional state of the user, but using more conventional dialogue management and planning techniques. We describe the architecture and components of the implemented prototype HWYD system.

1 Introduction

As the existence of this workshop shows, there is a good deal of interest in a type of spoken language dialogue system distinct from the traditional task-based models used for booking airline tickets and the like. The purpose of these 'social agent' systems is to be found in the relationship they can establish with human users, rather than on the assistance the agent can provide in giving information or solving a problem. Designing such agents provides many significant technical challenges, requiring progress in the integration of linguistic communication and non-verbal behaviour for affective dialogue (André et al. 2004). In this paper, we present the implementation of a Companion Embodied Conversational Agent (ECA) which integrates emotion and sentiment detection with more traditional dialogue components.

2 From Dialogue to Conversation

Most spoken language dialogue systems are 'task-based': they aim at getting from the user values for a fixed number of slots in some template. When enough values have been found, the filled template is sent off to some back-end system so that the task in question - ordering a pizza, booking a ticket etc. - can be carried out. However, a social Companion agent assumes a kind of conversation not necessarily connected to any immediate task, and which may not follow the conventions associated with task-driven dialogues, for example, the relatively strict turn-taking of task-based dialogue. In everyday life, many interhuman conversations see one of the participants producing lengthy descriptions of events, without this corresponding to any specific request or overall conversational purpose. Our objective was to support such free conversation, whilst still obtaining meaningful answers from the agent, in the form of advice appropriate both to the affective and informational content of the conversation. In order to balance the constraints of free conversation with those of tractability, we have deliberately opted for a single-domain conversation, in contrast with both small talk (Bickmore and Cassell, 1999) and 'chatterbot' approaches. Our HWYD domain involves typical events and topics of conversation in the workplace, ranging from the relatively mundane - meeting colleagues, getting delayed by traffic, project deadlines - to rather more important - promotions, firings, arguments, office politics - designed to evoke stronger emotions and hence more affective dialogues.

However, our HWYD Companions retains some features of a typical task based system, in that each of these subtopics can be thought of as a task or information extraction template. Unfilled slots will drive the dialogue manager to question the user for possible values. When enough slots

Proceedings of the 2010 Workshop on Companionable Dialogue Systems, ACL 2010, pages 37–42,
Uppsala, Sweden, 15 July 2010. ©2010 Association for Computational Linguistics

are filled, the initiative will be passed to an 'affective strategy' module, which will generate a longer response designed to empathise appropriately with the user over that particular topic.

3 System Overview and Architecture

The HWYD Companion integrates 15 different software components, covering at least to some degree all the necessary aspects of multimodal affective input and output: including speech recognition (ASR, using Dragon Naturally Speaking), emotional speech recognition (AA: the EmoVoice system (Vogt et al. 2008)), turn detection (ATT), Dialogue Act segmentation and tagging (DAT), Emotional modelling (EM), Sentiment Analysis (SA) (Moilanen et al. 2007), Natural Language Understanding (NLU), Dialogue Management (DM), user modelling and a knowledge base (KB/UM), an 'Affective Strategy Module' (ASM) generating complex system replies, Natural Language Generation (NLG), Speech Synthesis (TTS), an avatar (ECA), and Multimodal control of the ECA persona (MFM): gesture and facial expression, supported by the Haptek animation toolkit. Clearly the use of Naturally Speaking imposes on us speaker dependence, since the system needs training: in the scenario we have chosen this is in fact not too unrealistic an assumption, but this is merely a practical decision - we are not doing research on speech recognition as such in this project and so want to get as good a recognition rate as possible.

The software architecture of the prototype relies on the Inamode Framework developed by Telefnica I+D. Communication between modules follows a blackboard-like paradigm, in which central hubs broadcast any incoming message from any module to all of the other modules that are connected to it. Figure 1 below shows the system architecture, and Figure 2 shows one version of what is on the screen when the system is running.

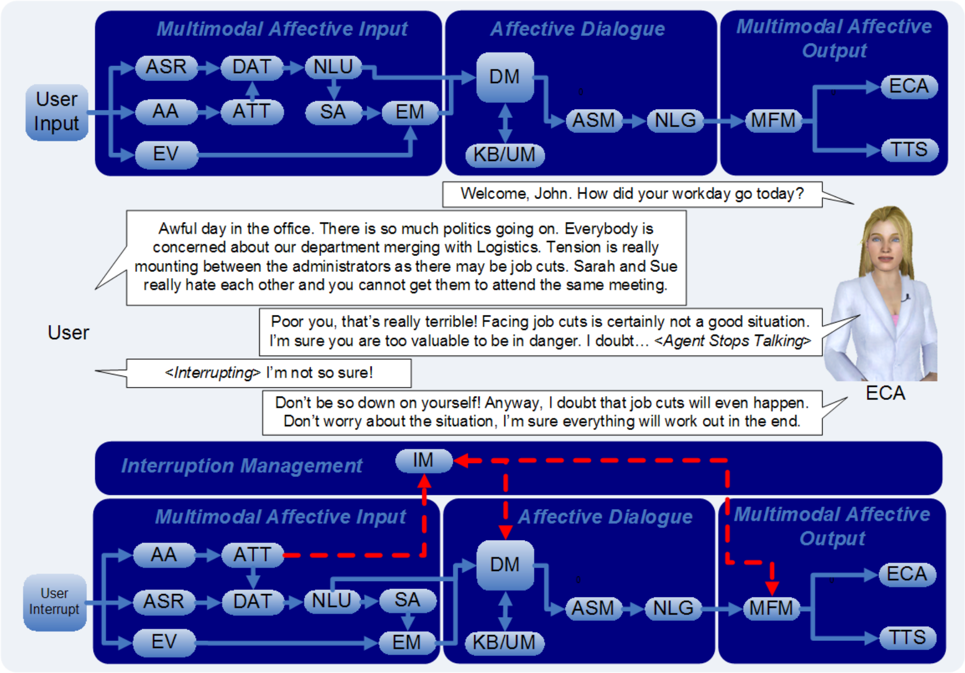
4 Emotional Feedback Loops

Recognising and responding appropriately to different emotions is an important aspect of a social agent. In our HWYD Companion, emotion and sentiment are used in two ways: firstly, to provide immediate feedback to a user utterance (given that there will inevitably be some delay in the response from natural language and dialogue processing modules) and secondly to inform the more extended responses given by the system when it has learned enough about the current sub-topic. There are two feedback loops: the 'short loop' (response time < 700 ms) provides an immediate backchannel, and its main purpose is to maintain contact and keep the communication alive and realistic. This is achieved by matching the non-verbal response (gesture, facial expression) of the avatar to the emotional speech parameters detected by EmoVoice prior to affective fusion (where the emotion detected from speech and the sentiment value detected from the corresponding text are merged: see below), and occasionally including an appropriate verbal acknowledgement, on a random basis to avoid acknowledging all user utterances. The short loop essentially aligns the ECA response to the user's attitude, thus showing empathy. (We should also use SA for this, but currently processing speed is not fast enough).

The 'major loop' (response time < 3000 ms) involves the ECA's full response to the user utterance in terms of both verbal and non-verbal behaviour. There are effectively two sources of system output: the dialogue manager engages with the user to find out what happened during their work day, and will ask questions, or drop into clarificatory sub-dialogues, gradually building up a complex event description along with an assessment of the prevailing emotions of the speaker. When sufficient information has been gathered, control is passed to an 'affective strategy module' which will produce a longer output, typically advice or warning in response to the user's recollection of his daily events.

The system also includes an interruption manager which detects interruption and barge-in by the user, resulting in the immediate suspension of the current system utterance, triggering the processing of any content specific to the interrupting utterance, and consequent replanning on the part of other modules to produce an appropriate response. Such an interruption is illustrated in Figure 1. The design of such an interruption manager in a system with so many separate modules is quite challenging, in fact: the system is described further in Crook et al. (2010).

The ECA listens sympathetically to the user's account of work difficulties, whilst also reacting to apparent discrepancies between perceived mood and the affective content of the recognised events. In the following example from a real conversation,



Welcome, John. How did your workday go today?

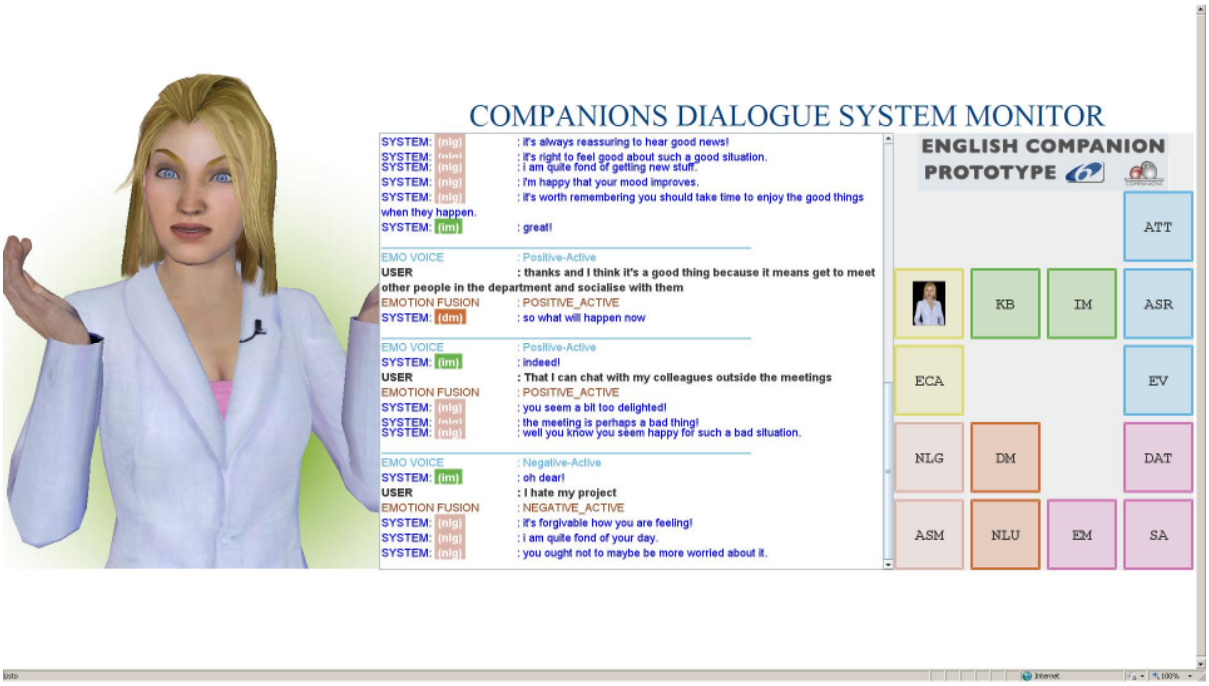
Awful day in the office. There is so much politics going on. Everybody is concerned about our department merging with Logistics. Tension is really mounting between the administrators as there may be job cuts. Sarah and Sue really hate each other and you cannot get them to attend the same meeting.

Poor you, that's really terrible! Facing job cuts is certainly not a good situation. I'm sure you are too valuable to be in danger. I doubt… <*Agent Stops Talking*>

<*Interrupting*> I'm not so sure!

Don't be so down on yourself! Anyway, I doubt that job cuts will even happen. Don't worry about the situation, I'm sure everything will work out in the end.

Figure 1: Overview of the System Architecture illustrating the information flow between modules.


COMPANIONS DIALOGUE SYSTEM MONITOR

```
SYSTEM: (nlg)        : it's always reassuring to hear good news!
SYSTEM: (nlg)        : it's right to feel good about such a good situation.
SYSTEM: (nlg)        : i am quite fond of getting new stuff.
SYSTEM: (nlg)        : i'm happy that your mood improves.
SYSTEM: (nlg)        : it's worth remembering you should take time to enjoy the good things
when they happen.
SYSTEM: (im)         : great!

EMO VOICE            : Positive-Active
USER                 : thanks and I think it's a good thing because it means get to meet
other people in the department and socialise with them
EMOTION FUSION       : POSITIVE_ACTIVE
SYSTEM: (dm)         : so what will happen now

EMO VOICE            : Positive-Active
SYSTEM: (im)         : indeed!
USER                 : That I can chat with my colleagues outside the meetings
EMOTION FUSION       : POSITIVE_ACTIVE
SYSTEM: (nlg)        : you seem a bit too delighted!
SYSTEM: (nlg)        : the meeting is perhaps a bad thing!
SYSTEM: (nlg)        : well you know you seem happy for such a bad situation.

EMO VOICE            : Negative-Active
SYSTEM: (im)         : oh dear!
USER                 : I hate my project
EMOTION FUSION       : NEGATIVE_ACTIVE
SYSTEM: (nlg)        : it's forgivable how you are feeling!
SYSTEM: (nlg)        : i am quite fond of your day.
SYSTEM: (nlg)        : you ought not to maybe be more worried about it.
```

Figure 2: The HWYD Companion in operation: the central portion of the screen traces the dialogue, whilst the right-hand one displays which current software module is activated.

the user announces a company restructuring, but in a rather cheerful tone. Notice that some of the English is a little strange: the generator produces a mixture of genuinely generated output along with some canned text.

User: I have just learnt that my department is going to be merged with the Barcelona one (EMOTION FUSION: POSITIVE_ACTIVE)
System: You seem a bit too ecstatic! I'm doubtful that your workload could be maintained.

Every time the Companion has gathered sufficient information it will try to respond more globally to the situation at hand. In the following example (see also Figure 2), after the user has expressed different sorts of concerns, the Companion picks on the latest (positive) information to summarise the few positive elements it picked from the conversation:

User: yes but one good thing anyway we had a new coffee machine arrived today (EMOTION FUSION: POSITIVE_ACTIVE)
System: it's always reassuring to hear good news! It's right to feel good about such a good situation. I am quite fond of getting new stuff. I'm happy that your mood improves. It's worth remembering you should take time to enjoy the good things when they happen.

5 Emotional Modelling and Fusion

Two different modules analyse the emotional content of user utterances: an emotional speech recognition system (EmoVoice (Vogt et al. 2008) which returns information indicating both the arousal and valence of the acoustic properties of the user's speech as negative passive, negative-active, neutral, positive-active or positive-passive, and a text-based Sentiment Analysis module which operates on the utterance transcript after its recognition by the ASR module. The SA module operates in a compositional way and is able to classify linguistic units of any syntactic type: noun phrases, clauses, sentences etc. It is also able to assign a 'strength' of the sentiment expressed. In the current implementation it simply classifies clauses as either negative, neutral or positive. These two emotional inputs are then merged by a fusion procedure, whose purpose is to provide an aggregate emotional category to be attached to the event description template produced by the NLU and DM module. Essentially, the mechanism for affective fusion consists in overriding the valence category of EmoVoice with the one obtained by SA every time the confidence score attached to EmoVoice is below a preset value (depending on the competing valence categories). Fusion is currently an underdeveloped module: for example, detecting mismatches between speech and language emotion and sentiment values could lead to the recognition of irony, sarcasm etc. (Tepperman et al. 2006). Saying an intrinsically negative thing in a positive and cheerful way, or the other way round, suggests that the speaker is trying for some special effect.

6 Natural Language Understanding and Dialogue Management

The task of the NLU module is to recognise a specific set of events reported by the user within utterances which can be of significant length (> 50 words) and which can be difficult to parse due to speech recognition errors. This led us to follow an Information Extraction (IE) approach to dialogue analysis (see Jönsson et al. 2004), using shallow syntactic and semantic processing to find instantiations of event templates. The NLU component of the HWYD Companion demonstration system takes the 1-best output from the speech recogniser (currently: work in progress will take n-best), which has already been segmented into dialogue-act sized utterances (by the DAT module which simultaneously segments and labels the recogniser output: see Figure 1). So, for example, a sequence like 'It was okay there are not many projects at the moment so it is very quiet would be segmented into three separate dialogue acts. The utterances are then part-of-speech tagged and chunked into Noun Phrase (NP) and Verb Group (VG) units. VGs consist of a main verb and any auxiliary verbs or semantically important adverbs. Both of these stages are carried out by a Hidden Markov Model trained on the Penn Treebank, although some customisation has been carried out for this application: relevant vocabulary added and some probabilities re-estimated to reflect properties of the application. NP and VG chunks are then classified into 'Named Entity' classes, some of which are the usual *person*, *organisation*, *time* etc. but others of which are specific to the scenario, as is traditional in IE: e.g. salient work events, expressions of emotion, organisational structure etc. Named Entity classification, in the absence of domain specific training data, is carried out via hand-written pattern matching rules and gazetteers. Each chunk is further annotated with features encoding the head word, stem form, polarity, agreement features, relevant modifiers, etc. for later syntactic and semantic processing. The NPs and VGs are represented as unification grammar categories containing information about the internal structure of the constituents.

The next stage applies unification based syntax rules which combine NP and VG chunks into larger constituents. These rules are of two types: most are syntactically motivated and are attempting to build a parse tree from which main grammatical relations (subject, object, etc.) can be recognised. These have coverage of the main syntactic constructs of English. But within the same formalism we add domain specific Information Extraction type patterns, looking out for particular constellations of entities and events relevant to the HWYD scenario, for example 'argument at work between X and Y', or 'meeting with X about Y'. Processing is non-deterministic and so sentences will get many analyses. We use a 'shortest path through the chart heuristic to select an interpretation. This is far from perfect, and we are currently working on a separate more motivated disambiguation module.

The final stage of processing before the Dialogue Manager takes over is to perform reference resolution for pronouns and definite NPs. This module is based partly on the system described by Kennedy and Boguraev 1996, with the various weighting factors based on theirs, but designed so that the weights can be trained given appropriate data. Currently we are collecting such data and the present set of weights are taken from Kennedy and Boguraev but with additional salience given to the domain-specific named entity classes. Each referring NP gives rise to a discourse referent, and these are grouped into coreference classes based on grammatical, semantic, and salience properties.

The DM maintains an information state containing all objects mentioned during the conversation, and uses this information to decide whether the objects referred to in the utterance are salient or not. The DM also uses type information to interpret elliptical answers to questions (System: 'Who was at the meeting?' User: 'Nigel.'). After the user's utterance has been interpreted in its dialogue context and the information state has been updated, the dialogue manager decides on the appropriate response. If a new object has been introduced by the user, the DM adds a goal to its agenda to talk about that object. For instance, if a new person is mentioned, the DM will ask questions about the user's relation to that person, etc.

For each turn of the dialogue, the DM chooses which topic to pursue next by considering all the currently un-satisfied goals on the agenda and heuristically rating them for importance. The heuristics employed use factors such as recency in the dialogue history, general importance, and emotional value associated with the goal. We are currently exploring the use of reinforcement learning with a reward function based on the results of SA on the users input to choose goals in a more natural way. The DM also has the option of invoking the ASM (described below) to generate an appropriate answer, in the cases where the user says something highly emotive. Again, this is a decision that could involve reinforcement learning, and we are exploring this in our current work.

The joint operation of the NLU and the DM hence supports a kind of IE or task-specific template-filling: the content of the user's utterances, prompted by questions from the DM, provides the information necessary to fill a template to the point where the ASM can take over. The number of templates for domain events is significantly higher than in traditional IE or task-based dialogue systems, however, since the HWYD Companion currently instantiates more than 30 templates, and will eventually cover around 50.

7 Affective Dialogue Strategies

Once the NLU and DM have a sufficiently instantiated template, which also records emotional value, it is passed to the ASM. This controls the generation of longer ECA narrative replies which aim at influencing the user by providing advice or reassurance. Our overall framework for influence is inspired by the work of Bremond 1973. The narrative is constituted by a set of argumentative statements which can be based on emotional operators (e.g. **show-empathy**) or specific communicative operators. The ASM is based on a Hierarchical Task Network (HTN) planner (Nau et al 2004), which works through recursive decomposition of a high level task into sub-tasks until we reach a plan of sub-tasks that can be directly executed. The operators constituting the plan generated by the HTN implement Bremond's theory of influence by emphasising the determinants of the event reported by the user. For instance, various operators can emphasise or play down the event consequences (**emphasise-outcome-importance, emphasise-outcome-justification, emphasise-outcome-warning**) or comment on additional factors that may affect the course of events (**commend-enabler, reassure-helper**). The planner uses a set of 25 operators, each of which can be in addition instantiated to incorporate specific elements of the event. Overall this supports the generation of hundreds of significantly different influencing strategies.

8 Results and Conclusions

We have described an initial, fully-implemented prototype of a Companion ECA supporting free conversation, including affective aspects, over a variety of everyday work-related topics. The system has been demonstrated extensively outside of its development group and was regularly able to sustain consistent dialogues with an average duration exceeding 20 minutes. The Companion ECA recently won the best demonstration prize at AAMAS 2010,the 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, which is some subjective indication at least that its behaviour is of some interest outside of the project which developed it.

However, we have not yet systematically evaluated the ECA, although this task has begun (Webb et al. 2010). The question of evaluation for systems like this is in fact a rather difficult one, since unlike task-based systems there is no simple measure of success. In our current work we aim to conduct extensive trials with real users and via interview and questionnaires to get some useful measure of how natural and 'companionable' the system is perceived to be.

In other current work we are, as mentioned above, experimenting with reinforcement learning where the reward function is based on the emotion and sentiment detected in the user's input. We are collecting data via Amazon's Mechanical Turk and hope to be able to show how the ECA can develop different 'personalities' depending on how this reward function is defined. For example, we could imagine using simulated dialogues to produce a Companion that was relentlessly cheerful, producing positive outputs whatever the input. Alternatively, we could produce a 'mirror' Companion which simply reflected the mood of the user. We could even produce a 'misery loves company' Companion which, instead of trying to cheer the user up when recognising negative sentiment or emotion, could reply in an equally negative manner.

Acknowledgements

This work was funded by the Companions project (http://www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434. The EmoVoice system has been used courtesy of the Multimedia Concepts and Applications Group of the University of Augsburg. Other contributors to the prototype described in this paper are Karo Moilanen, and from the COMPANIONS consortium: David Benyon, Jay Bradley, Daniel Charlton, WeiWei Cheng, Morena Danieli, Simon Dobnik, Carlos Sanchez Fernandez, Debora Field, Mari Carmen Rodriguez Gancedo, Jose Relano Gil, Ramon Granell, Jaakko Hakulinen, Preben Hansen, Sue Harding, Topi Hurtig, Oli Mival, Roger Moore, Olov Stahl, Markku Turunen, Enrico Zovato.

References

André, E., Dybkjr, L., Minker, W., and Heisterkamp, P. (Eds.), 2004, *Affective Dialogue Systems* Lecture Notes in Computer Science 3068, Springer.

Bickmore, T., and Cassell, J., 1999. *Small Talk and Conversational Storytelling in Embodied Interface Agents.* Proceedings of the AAAI Fall Symposium on Narrative Intelligence, pp. 87-92. November 5-7, Cape Cod, MA.

Bremond, C., 1973, *Logique du Récit*, Paris: Editions du Seuil.

Cavazza, M., Pizzi, D., Charles, F., Vogt, T. And André, E. 2009, *Emotional input for character-based interactive storytelling.* International Joint Conference on Autonomous Agents and Multi-Agents Systems 2009, pp. 313-320.

Nigel Crook, Cameron Smith, Marc Cavazza, Stephen Pulman, Roger Moore, Johan Boye, 2010, *Handling User Interruptions in an Embodied Conversational Agent* Proceedings of International Workshop on Interacting with ECAs as Virtual Characters, AAMAS 2010.

Jönsson, A., Andén, F., Degerstedt, L., Flycht-Eriksson, A., Merkel, M., and Norberg, S., 2004, *Experiences from combining dialogue system development with information extraction techniques*, in: Mark T. Maybury (Ed), New Directions in Question Answering, AAAI/MIT Press.

Kennedy and B. Boguraev, 1996, *Anaphora for everyone: Pronominal anaphora resolution without a parser.* Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, ACL, pp 113-118.

Moilanen, K. and Pulman, S. G. , 2007, *Sentiment Composition*, Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP-2007), pp 378–382.

Nau, D., Ghallab, M., Traverso, P., 2004,*Automated Planning: Theory and Practice*, Morgan Kaufmann Publishers Inc., San Francisco, CA.

J Tepperman, D Traum, and S Narayanan, 2006, *'Yeah right': Sarcasm recognition for spoken dialogue systems*, Interspeech 2006, Pittsburgh, PA, 2006.

Vogt, T., André, E. and Bee, N., 2008 *EmoVoice - A framework for online recognition of emotions from voice.* In: Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems, Springer, Kloster Irsee, Germany, (June 2008).

Webb, N., D. Benyon, P. Hansen and O. Mival, 2010, *Evaluating Human-Machine Conversation for Appropriateness*, in proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta.