Cameron Smith
School of Computing
Teesside University
Middlesbrough, TS1 3BA, UK

Nigel Crook
Simon Dobnik
Oxford University Computing Laboratory
Oxford, OX1 3QD, UK

Daniel Charlton
School of Computing
Teesside University
Middlesbrough, TS1 3BA, UK

Johan Boye
Stephen Pulman
Oxford University Computing Laboratory
Oxford, OX1 3QD, UK

Raul Santos de la Camara
Telefonica I+D
28043 Madrid, Spain

Markku Turunen
Department of Computer Sciences
University of Tampere
Finland

David Benyon
Jay Bradley
School of Computing
Edinburgh Napier University
Edinburgh, EH10 5DT, UK

Björn Gambäck
Preben Hansen
SICS
SE-164 29 Kista, Sweden

Oli Mival
School of Computing
Edinburgh Napier University
Edinburgh, EH10 5DT, UK

Nick Webb
ILS Institute
SUNY Albany
Albany, NY 12222, USA

Marc Cavazza*
School of Computing
Teesside University
Middlesbrough, TS1 3BA, UK

Presence, *Vol. 20, No. 5, October 2011*, 000–000

Interaction Strategies for an Affective Conversational Agent

Abstract

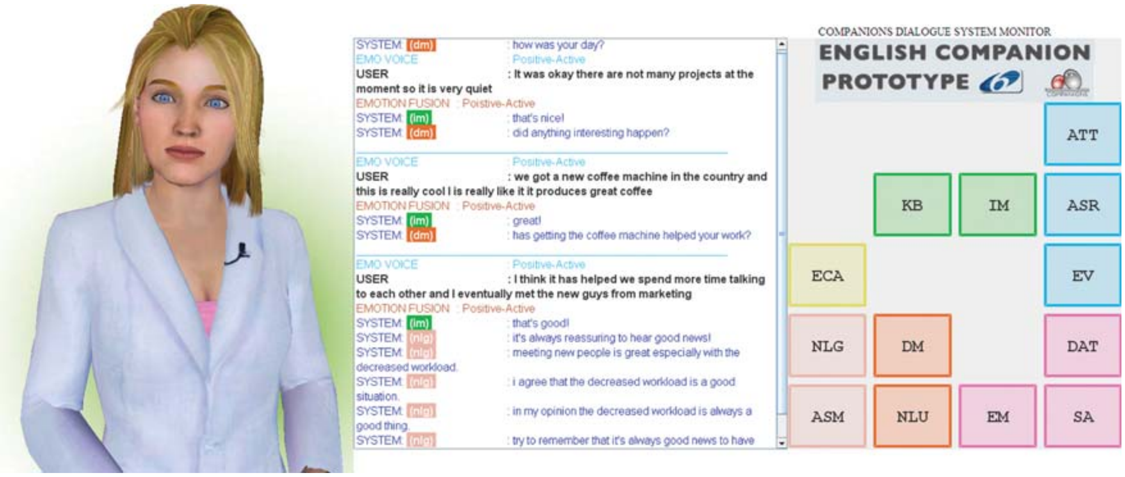
The development of embodied conversational agents (ECA) as companions brings several challenges for both affective and conversational dialogue. These include challenges in generating appropriate affective responses, selecting the overall shape of the dialogue, providing prompt system response times, and handling interruptions. We present an implementation of such a companion showing the development of individual modules that attempt to address these challenges. Further, to resolve resulting conflicts, we present encompassing interaction strategies that attempt to balance the competing requirements along with dialogues from our working prototype to illustrate these interaction strategies in operation. Finally, we provide the results of an evaluation of the companion using an evaluation methodology created for conversational dialogue and including analysis using appropriateness annotation.

1 Introduction

An emerging concept in recent years has been that of a social agent which focuses more on the relationship it can establish with a human user than on the assistance or information it can provide for a practical task. This concept of a companion is particularly significant for embodied conversational agent (ECA) research where the notion of companionship emerges from the overall communicative abilities of the ECA (i.e., embodied and conversational aspects feeding into affective dialogue). Yet there are also significant technical challenges encountered here in the integration of linguistic communication and nonverbal behavior for affective dialogue (André, Dybkjær, Minker, & Heisterkamp, 2004).

In this paper, we present the implementation of a companion ECA integrating all of the above aspects into a single prototype, in a way which supports conversational phenomena one would expect from affective dialogue, namely, lengthy utterances on both sides and interruptions. This presentation mainly focuses on the interaction strategies supported by the agent, which support the principled integration of the large number of software components required to analyze user input, reason upon the situation, control the flow of dialogue, and generate appropriate ECA responses and multimodal behaviors. Our main objective is to give insight into these interaction strategies and to illustrate the companion's performance with detailed examples from a fully-implemented prototype.

*Correspondence to M.O.Cavazza@tees.ac.uk.


COMPANIONS DIALOGUE SYSTEM MONITOR
ENGLISH COMPANION PROTOTYPE
ATT
KB
IM
ASR
ECA
EV
NLG
DM
DAT
ASM
NLU
EM
SA

```
SYSTEM (dm)                 : how was your day?
EMO VOICE                   : Positive-Active
USER                        : It was okay there are not many projects at the
moment so it is very quiet
EMOTION FUSION  : Poistive-Active
SYSTEM (im)                 : that's nice!
SYSTEM (dm)                 : did anything interesting happen?

EMO VOICE                   : Positive-Active
USER                        : we got a new coffee machine in the country and
this is really cool I is really like it it produces great coffee
EMOTION FUSION  : Positive-Active
SYSTEM (im)                 : great!
SYSTEM (dm)                 : has getting the coffee machine helped your work?

EMO VOICE                   : Positive-Active
USER                        : I think it has helped we spend more time talking
to each other and I eventually met the new guys from marketing
EMOTION FUSION  : Positive-Active
SYSTEM (im)                 : that's good!
SYSTEM (nlg)                : it's always reassuring to hear good news!
SYSTEM (nlg)                : meeting new people is great especially with the
decreased workload.
SYSTEM (nlg)                : i agree that the decreased workload is a good
situation.
SYSTEM (nlg)                : in my opinion the decreased workload is always a
good thing.
SYSTEM (nlg)                : try to remember that it's always good news to have
```

Figure 1. *The companion during a typical dialogue.*

2 System Overview and Application

F1 The companion (as shown in Figure 1) presents itself as an ECA with which the user can engage in free conversation, albeit on a select set of topics. As an application scenario, we wanted an everyday life domain that would support conversation with some affective content. We opted for a scenario in which the user, a typical office worker, returns home and talks about the day's events. We refer to this as the "How was your day?" (HWYD) scenario. The system currently supports over 40 work-based conversational topics, with further discussion of a range of influencing factors and event outcomes, across a range of emotional situations. By definition, the conversation is not task-oriented (unless one considers a very high level task of supporting the user through positively influencing their attitudes) and follows a mixed-initiative paradigm. User initiative, as expected, takes a central role, but without reducing the companion to a passive, although sympathetic, listener. As evidenced by the example dialogues of Figures 5, 6, and 7, discussed later, the companion will attempt to offer appropriate advice as soon as it has assessed the user situation and considers such advice as appropriate.

Our system integrates no less than 15 different software components covering aspects of multimodal affective input, affective dialogue processing, interruption management, and multimodal affective output. The software architecture integrating these components follows a blackboard philosophy (Englemore & Morgan, 1988), which provides the control flexibility required to implement various interaction strategies (see below). The system (Figure 2) is composed of speech, language, reasoning, and animation modules. F2 Automatic speech recognition (ASR) is provided by Nuance's Dragon NaturallySpeaking, while text-to-speech (TTS) is an extension of Loquendo's commercial system developed as part of this project. The ECA appearance and animation are based on the Haptek^{TM} toolkit. As expected, all dialogue and natural language understanding (NLU) modules are proprietary. Emotional aspects are pervasive in these modules but their inclusion depends on the module itself: The animation module for the ECA naturally supports nonverbal behavior and the expression of emotions, while our TTS system has been specifically extended to support emotional markers. Finally, some modules are entirely dedicated to affective processing: the recognition of emotional categories from speech is based on the EmoVoice (Vogt, André, & Bee, 2008) system, and the affective content of utterances' transcripts is uncovered using a sentiment analysis module (Moilanen & Pulman, 2007). Depending on the interaction strategy considered, these modules will be used separately or their output will be merged using an

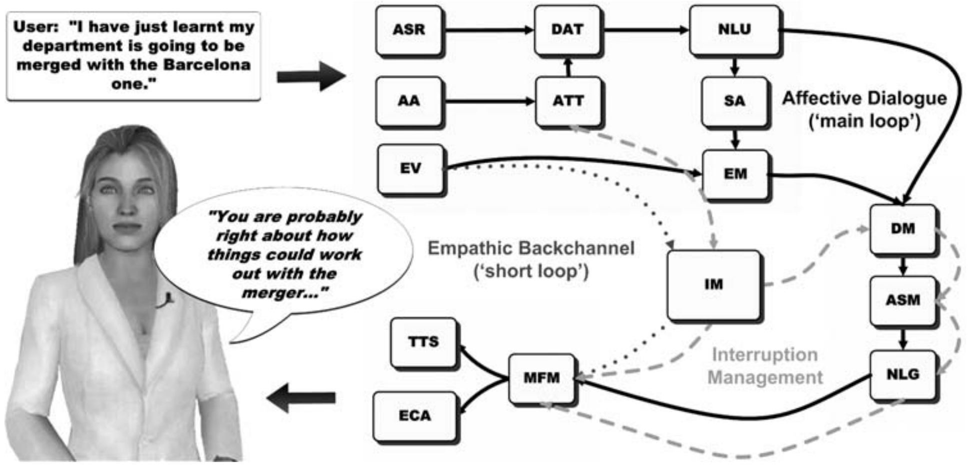

User: "I have just learnt my department is going to be merged with the Barcelona one."
ASR
DAT
NLU
AA
ATT
SA
Affective Dialogue ('main loop')
EV
EM
DM
"You are probably right about how things could work out with the merger..."
Empathic Backchannel ('short loop')
IM
ASM
TTS
MFM
Interruption Management
NLG
ECA

Figure 2. *System components with principal interaction loops (see text for details).*

emotional model performing multimodal fusion of affective categories. In this system, multimodality is primarily dedicated to affective aspects, both in terms of input (emotional contents of speech/voice and transcribed utterances) and output (ECA speech, facial expressions, and gestures).

Affective dialogue processing is led by the dialogue manager (DM), which supports traditional functions such as managing clarification dialogue and repair. It further makes use of the more specific affective strategy module (ASM) for generating complex affective utterances and a natural language generation (NLG) module for realizing replies into utterances for the multimodal affective output stage. The multimodal affective output is coordinated by the multimodal fission manager (MFM) which controls both the ECA and TTS modules. This is all overseen by an interruption management layer coordinated by the interruption manager (IM). The necessity to control turn-taking and interruptions has led to the incorporation of specific speech modules: the acoustic analysis (AA) and acoustic turn taking (ATT) modules, which input into a dialogue act tagger (DAT).

Natural language processing was also adapted to the objectives of affective dialogue and free conversation. The techniques used, including tagging, shallow parsing, called entity identification and contextual reference resolution, resemble information extraction and provide a robust coverage of the longer utterances, compared to previous dialogue systems, found in non-task-oriented conversations.

3 Interaction Strategies

The majority of language-enabled ECA have been developed in the context of task-based dialogue; this was dictated by both application constraints and linguistic coverage. However, the very idea of a companion agent assumes a level of conversation which is disconnected from any immediate task, and in particular is freed from strict constraints on the nature of dialogue.

Therefore, several traditional assumptions which have presided over the formalization of human–computer dialogue may need to be relaxed when exploring affective conversation. In everyday life, many inter-human conversations see one of the participants relating events through lengthy descriptions, without this corresponding to any specific request or encompassing speech act. Our objective was to support such free conversation, while still obtaining meaningful answers from the companion 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-topic conversation, in contrast both to small talk (Bickmore & Cassell, 1999) and ChatterBot approaches. It should be noted that even ChatterBots fail to depart from the conventions of human–computer dialogue, and most often feature dialogues in which user and agent utterances alternate rather strictly (De Angeli & Brahnam, 2008).

Our individual components seek to address some of the challenges of conversational dialogue: affective input, longer utterances, balancing clarification dialogue with long-form responses, and the generation of these long-form responses. Yet individual optimizations only tackle part of the problem and can often introduce further problems of their own. As such, we additionally sought a more holistic approach; several interaction strategies allowing the different components to work together effectively, with each strategy catering to different requirements of a companion.

In the following sections we look in detail at the interaction strategies available before going on to provide examples from our implemented system showing the various interaction strategies in operation.

3.1 Short Loop Interaction: An Empathic Backchannel

Previous work has amply demonstrated the importance of backchannels in human–agent conversation (Cassell & Thorisson, 1999; Morency, de Kok, & Gratch, 2008; Kopp, Stocksmeier, & Gibbon, 2007; Bevacqua, Mancini, & Pelachaud, 2008). In addition, the processing time required by the complete affective dialogue system, which includes reasoning upon the user's situation and the appropriateness of her emotional reaction, still exceeds the recommended response time for dialogue systems, being on average over 3 s. This makes it essential to provide a real-time (<700 ms) yet relevant backchannel to the user, able to acknowledge user interaction and provide an initial response appropriate to the affective context even without a full analysis of the utterance.

The short loop implements a fast alignment between the perceived emotional state of the user and the ECA's expression, as well as acknowledging user utterances (see Figure 2). This is achieved by matching the ECA's nonverbal response to the emotional speech parameters detected by the emotional speech recognizer EmoVoice and including an appropriate verbal acknowledgment (on a random basis to avoid acknowledging all user utterances). The short loop thus essentially aligns the ECA response on the user's attitude.

3.2 Main Loop Interaction: Affective Dialogue and Reasoning

The main interaction strategy consists in a complete end-to-end implementation of affective conversation (with a response time of under 3000 ms). It enacts the overall behavior of the companion as an affective dialogue system and involves its full response to the user utterance in terms of both verbal and nonverbal behavior (both gestures and facial expressions).

The main loop (see Figure 2) thus corresponds to an end-to-end implementation of affective conversation between the user and the agent. It is based on the identification of office life events, together with the affective context in which they are introduced. Following an appraisal step that determines the adequacy of the user's response to the situation he or she is facing (e.g., difficulties with colleagues, restructuring, redundancies), the companion will provide an affective response in the form of reassurance, advice, comfort (or, in some cases, warning) to positively affect the user's attitude. The content is, however, specific to the details of the situation reported and makes reference to the different causes and consequences of the reported events. Conversational dialogue further requires a degree of flexibility in juggling user utterances of varying lengths with shifting topics while accounting for affective aspects. The expectation is that the companion will be able to provide a response of appropriate length and tone in reply to the topic provided by the user. However, in order to do this effectively, the companion may be required to clarify information and elicit further information to support a meaningful response. The dialogue management thus needs to find a balance between employing clarification dialogue and generating appropriate responses to the information provided by the user.

The overall conversational loop is under the supervision of a DM which controls the various phases of dialogue and their timing, as well as the level of system initiative, in an integrated fashion. One of the main decisions it has to make is when to trigger lengthier utterances (which we have termed tirades, see, e.g., Figures 5, 6, and 7 discussed later in this paper), which correspond to an affective dialogue strategy aimed at influencing the user's attitude by means of a short narrative. The challenge for the DM is to shift between the various aspects of conversation: allowing long rants from the user, providing sympathetic feedback without shifting dialogue initiative toward itself, triggering clarification subdialogues, or regaining initiative through long utterances that provide advice and support in a more structured fashion. Some of these aspects may be covered by the identification of dialogue acts, but dialogue acts alone may not be able to deal with the contents of longer user utterances (>30 words). This is why one of the integrating principles adopted by our system is to also base dialogue control on event instantiation, thus relating it to information extraction (IE).

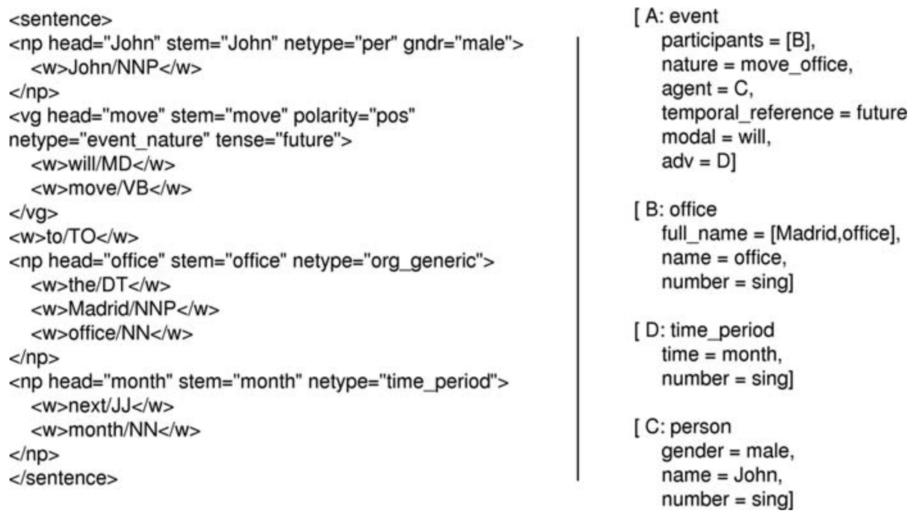
3.3 Information Extraction

Conversations may involve utterances of various lengths including utterances much longer (>50 words) than those typically found in task-oriented dialogues. Sentences may be ill-formed or highly elliptical. Furthermore, speech recognition under realistic conditions frequently results in a high word error rate, making the task of syntactic analysis even harder. The task of the NLU module is to recognize a specific set of events reported by the user. These events are formalized as objects consisting of feature-value pairs. The NLU (in collaboration with the DM) employs shallow processing methods that instantiate event templates. These methods resemble IE techniques (Grishman, 1997; Jönsson, et al., 2004).

The NLU takes the 1-best output from the speech recognizer, which has already been segmented into dialogue-act sized utterances. The utterances are then part-of-speech tagged and separated into noun phrase (NP) and verb group (VG) chunks which denote concepts in our domain. 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 customization 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 (NE) classes, some of which are the usual person, organization, time; and but others of which are specific to the scenario, as is traditional in IE; for example, salient events, expressions of emotion, and organizational structures, to name a few. NE classification, in the absence of domain specific training data, is carried out via hand-written pattern matching rules and gazetteers. The NPs and VGs are represented as unification grammar categories containing information about the internal structure of the constituents; for example, an utterance such as "John will move to the Madrid office next month" would yield results such as that on the left of Figure 3. F3

In the next stage of NLU processing, domain specific IE patterns are applied on NP and VG chunks which rely on their syntactic and semantic information to form constituents called objects. Examples could be, for example, "meeting with X about Y" where NE type of X is person, or "move to X" where NE type of X is org_generic. In the final stage, reference resolution for pronouns and definite NPs is performed. This module is based partly on the system described by Kennedy and Boguraev (1996), with the various weighting factors based on theirs. Each referring NP gives rise to a discourse referent, and these are grouped into coreference classes based on grammatical, semantic, and salience properties.

On its own, the NLU module is a large-coverage system which can tag, shallow parse, and resolve pronoun reference of any English sentence. Its coverage is most restricted by domain specific NE classes and IE patterns which must be introduced manually. The system covers more than 40 work-based topics of conversation; for example, discussions of meetings, problems with office equipment, relationships with colleagues, and even the weather. These are mostly represented as event objects. Complex objects such as these are created by a set of IE rules which attempt to cover a range of syntactic and semantic structures which denote identical content. In addition to event objects, the system covers objects of various NE types that relate to the events. For example, to refer to persons, the system may have to collect their



```
<sentence>
<np head="John" stem="John" netype="per" gndr="male">
  <w>John/NNP</w>
</np>
<vg head="move" stem="move" polarity="pos"
netype="event_nature" tense="future">
  <w>will/MD</w>
  <w>move/VB</w>
</vg>
<w>to/TO</w>
<np head="office" stem="office" netype="org_generic">
  <w>the/DT</w>
  <w>Madrid/NNP</w>
  <w>office/NN</w>
</np>
<np head="month" stem="month" netype="time_period">
  <w>next/JJ</w>
  <w>month/NN</w>
</np>
</sentence>
```

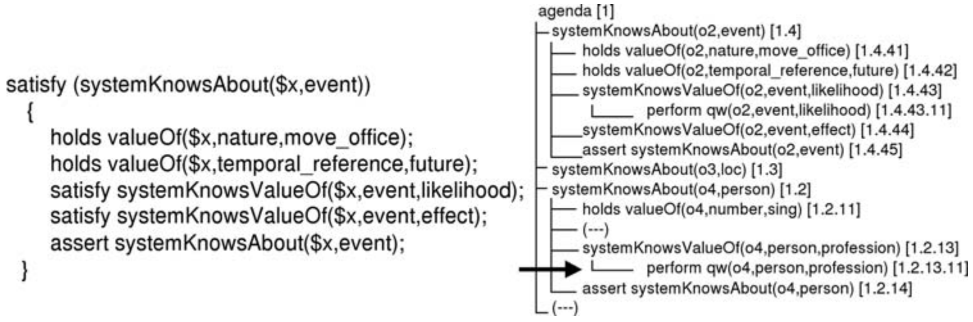
```
[ A: event
    participants = [B],
    nature = move_office,
    agent = C,
    temporal_reference = future
    modal = will,
    adv = D]

[ B: office
    full_name = [Madrid,office],
    name = office,
    number = sing]

[ D: time_period
    time = month,
    number = sing]

[ C: person
    gender = male,
    name = John,
    number = sing]
```

Figure 3. *NP and VG representation (left) and final semantic representation (right) used by the NLU.*



```
satisfy (systemKnowsAbout($x,event))
  {
     holds valueOf($x,nature,move_office);
     holds valueOf($x,temporal_reference,future);
     satisfy systemKnowsValueOf($x,event,likelihood);
     satisfy systemKnowsValueOf($x,event,effect);
     assert systemKnowsAbout($x,event);
  }
```

```
agenda [1]
├systemKnowsAbout(o2,event) [1.4]
│ ├── holds valueOf(o2,nature,move_office) [1.4.41]
│ ├── holds valueOf(o2,temporal_reference,future) [1.4.42]
│ ├── systemKnowsValueOf(o2,event,likelihood) [1.4.43]
│ │    └──── perform qw(o2,event,likelihood) [1.4.43.11]
│ ├── systemKnowsValueOf(o2,event,effect) [1.4.44]
│ └── assert systemKnowsAbout(o2,event) [1.4.45]
├ systemKnowsAbout(o3,loc) [1.3]
├ systemKnowsAbout(o4,person) [1.2]
│ ├── holds valueOf(o4,number,sing) [1.2.11]
│ ├── (---)
│ ├── systemKnowsValueOf(o4,person,profession) [1.2.13]
│ │    └──── perform qw(o4,person,profession) [1.2.13.11]
│ └── assert systemKnowsAbout(o4,person) [1.2.14]
└(---)
```

Figure 4. *Goal satisfaction rule (left) and Agenda (right) used by the DM.*

names, gender and profession, organization they work for, their colleagues, and the location where they live. In contrast to events, these objects mostly rely on recognition of NE classes.

The final output from the NLU in the format expected by the DM for the utterance "John will move to the Madrid office next month" is shown on the right of Figure 3.

3.4 Dialogue Management

The DM is based on work described previously (Boye & Gustafson, 2005; Boye, Gustafson, & Wirén, 2006; Boye, 2007), but has been substantially modified for the challenges of conversational dialogue. It receives user utterances from the NLU as semantic representations (right side of Figure 3). The DM first checks which information addresses the previous question or comment posed by the system in the dialogue and which information opens up new topics. The information constituting answers to system questions is integrated into the information state of the DM (called the object store), while new topics give rise to new conversational goals.

The DM keeps track of all the topics under discussion by maintaining a set of conversational goals; for example, (1) "Find out more about the possible office relocation to Madrid," or (2) "Make a comment about today's meeting." A number of goal-satisfaction rules (similar to the one on the left of Figure 4) specify how goals are F4
broken down into sequences of subgoals and system

utterances. For instance, finding out more about the office relocation (1) might amount to asking specific questions about whether the relocation will indeed take place, what the consequences would be for the user, and so on. The goal is considered satisfied when further information about the relocation has been collected.

The various possible topics of conversation are organized as in an ontology, so that it is known what attributes can be expected to be present for a particular object. For example, the value of the effect attribute of the event object must be another object of type event. Again this is reminiscent of IE, and the DM is in effect aiming to fill a template via clarification and supplementary questions (satisfy systemKnowsValueOf($x,event,effect)) to the point where it can be passed to the ASM.

The active goals are organized in a tree-structure, the so-called agenda, as shown on the right side of Figure 4. At any given point in time, the agenda might contain many topics, some old, some new (systemKnowsAbout (o2,event)), some completed (—), some still open for discussion, and some not yet addressed by the system (systemKnowsValueOf(o2,event,likelihood). For each turn of the clarification dialogue, the DM chooses which topic to pursue next by considering all the currently unsatisfied 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. In the example in Figure 4, the system considered it more important to find out about the person (o4 or 'John') than to find out about the event that the person is a participant of (o2 or 'move_office').

When sufficient information has been gathered from the user through the clarification dialogue, the DM will invoke the ASM so it can generate a suitable tirade. The DM makes the decision to invoke the ASM using heuristics that take into account, among other things, the emotional value of the user's utterances and the recency of the latest ASM invocation.

3.5 Affective Dialogue Strategies

Previous dialogue systems (Cavalluzzi, Carofiglio, & de Rosis, 2004; Bickmore & Sidner, 2006) have resorted to different models as a basis for influencing user behavior, such as the transtheoretical model (Prochaska, Di Clemente, & Norcross, 1992). However, in our current scenario, we are more interested in changes in attitudes rather than behavior (Tørning & Oinas-Kukkonen, 2009). In presenting a response to the user, it is first necessary to understand, or appraise, the situation that the user presents to the companion. This involves gaining an understanding of the events described and how these will affect the user. Further, the user's reaction to these events is also crucial in generating an appropriate tirade. The ASM centers its response on a main event, generally the focal event selected by the DM, and its consequences for the user.

An appraisal process determines the nature of the main event in terms of both its impact on the user and the appropriateness of the user's reaction. The impact depends on whether the event constitutes an improvement (promotion, payraise) or a deterioration (office-move, redundancy, increased-workload) to the user's situation. This is determined by using the NLU information to instantiate an event template which indicates both the event type (e.g., improvement) and anticipated outcome based on what the event is and the information available. Every possible NLU event has its own event template within the ASM and default knowledge is used to instantiate these templates where information is not available from the NLU.

Next, the user's mood, provided by the emotional model, is used to determine whether the user is showing an appropriate or inappropriate emotional reaction to the event, given the anticipated outcome. This is essentially whether the user is reacting positively to improvements and negatively to deteriorations.

These details are then used to determine the strategy employed by the companion. These strategies have been selected such that they cover the full range of possible situations a user can be in: a congratulatory strategy for when things are going well for the user, a sympathetic strategy for when they are not, encouraging or reassuring strategies for when the user's outlook is too negative, and warning or cautionary strategies for when the user's outlook is too positive. The appraisal process also analyses additional influences, be they positive or negative, for

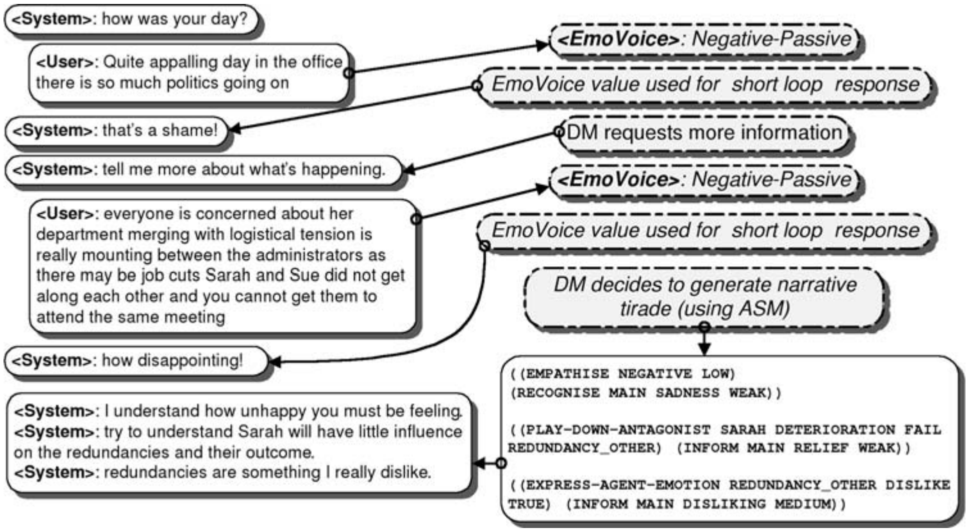

<System>: how was your day?
<User>: Quite appalling day in the office there is so much politics going on
<EmoVoice>: Negative-Passive
EmoVoice value used for short loop response
<System>: that's a shame!
DM requests more information
<System>: tell me more about what's happening.
<EmoVoice>: Negative-Passive
<User>: everyone is concerned about her department merging with logistical tension is really mounting between the administrators as there may be job cuts Sarah and Sue did not get along each other and you cannot get them to attend the same meeting
EmoVoice value used for short loop response
DM decides to generate narrative tirade (using ASM)
<System>: how disappointing!
((EMPATHISE NEGATIVE LOW) (RECOGNISE MAIN SADNESS WEAK))
((PLAY-DOWN-ANTAGONIST SARAH DETERIORATION FAIL REDUNDANCY_OTHER) (INFORM MAIN RELIEF WEAK))
((EXPRESS-AGENT-EMOTION REDUNDANCY_OTHER DISLIKE TRUE) (INFORM MAIN DISLIKING MEDIUM))
<System>: I understand how unhappy you must be feeling.
<System>: try to understand Sarah will have little influence on the redundancies and their outcome.
<System>: redundancies are something I really dislike.

Figure 5. *An example dialogue where the user discusses a negative situation and shows a correspondingly negative emotional state. Yet the companion detects this is just a potentially bad situation and employs a reassuring affective strategy.*

the events at hand. These will be used to enrich the companion's tirade, giving a more precise content to reassurance or warning statements.

In common with both narrative generation (Cavazza, Charles, & Mead, 2002) and text generation (Appelt, 1985), the ASM is based on planning technologies, more specifically a hierarchical task network (HTN) planner (Nau, Ghallab, & Traverso, 2004), which works through recursive decomposition of a high level task into subtasks until a plan of subtasks that can be directly executed is produced. The HTN planning process uses the information from the event templates along with results from the appraisal as heuristics to guide its decomposition. Combined with the fact that this heuristic selection process occurs at multiple levels of the HTN, it allows for greater complexity and variance than is achievable with a scripted approach.

The resulting plan of operators provides a set of communicative functions, each targeting different aspects of the user's utterance but unified under the overall affective strategy. For instance, various operators can emphasize or play down the event consequences or comment on additional factors that may affect the course of events. The planner uses a set of 40 operators, each with multiple parameters. Overall this supports the seamless generation of hundreds of significantly different influencing strategies from the base set of influence operators.

This plan is passed to the NLG module where each operator is realized as a sentence-forming part of the overall narrative utterance. The operators contain information supporting an FML-like language (Hernández et al., 2008) which allows full multimodal output composed of affective TTS, gestures, and facial expressions.

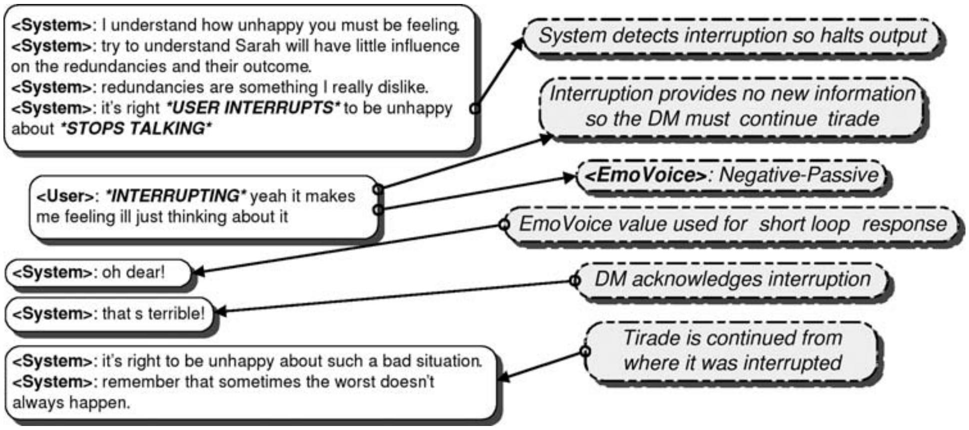
Figure 5 illustrates the operation of the ASM on an F5
excerpt from an actual dialogue. The companion first instantiates some basic information (a bad day event and discussion of office politics) from the first user utterance. However, this is not enough to meet the threshold for generating an affective tirade so the DM triggers a clarification step ("tell me more . . ."), which actually prompts a longer and more detailed reply from the user. From this reply, the system is able to instantiate further event templates, one about company restructuring, one about redundancies, and one about relationships between colleagues, with the DM determining that the redundancies event template is the most prominent event. The ASM then appraises this main event, determining (from the instantiated event template) that the redundancies have

not yet happened, and opting to perform a reassuring strategy. The ASM then generates a plan which shows different levels of empathy (one generic and one specific, mentioning the threat of redundancy), but also dissociates the two incidents by reminding the user that antagonistic colleagues will have no influence on redundancy decisions (this is achieved by looking for factors potentially influencing the key event, here company restructuring).

3.6 Handling Interruptions

Conversational flow in natural dialogues tends to be quite fluid, with partners frequently interrupting each other rather than observing the strict turn-by-turn structure of most current spoken language dialogue systems. Further, the generation of long, multi-sentence utterances by the ASM creates opportunities for the user to interrupt the companion while it is speaking. Indeed, the long ASM utterances may even provoke a user interruption given that they often include advice on dealing with difficult or stressful situations that the user has experienced. To resolve this, our companion includes interaction strategies for dealing with both barge-in interruptions and non-barge-in interruptions. When a user starts talking at the same time as the companion, interrupting the companion's reply, this is classed as a barge-in interruption. We now describe the handling process (see also Figure 2).

1. As the user may speak at any time, the ATT module must decide whether this constitutes a genuine user interruption (as opposed to, say, backchannel). This decision is based on both the intensity and duration of the voice signal with the IM being informed when an interruption is detected.
2. The IM then requests that the ECA stop speaking and be given a look of surprise or irritation at being interrupted before broadcasting a notification of the interruption to all modules so they know the previous turn was not completed.
3. The DM determines how much of the ASM response was completed.
4. The ATT informs the IM when the interruption has ended. The IM then tracks the processing of the interrupting utterance through the system using a system state model implemented as a two-level finite state machine (Crook et al., 2010). Tracking the processing is necessary to ensure that the companion responds within a realistic time frame.
5. When triggered, the DM must decide how to respond to that interruption.
   - A. The DM would choose to continue the interrupted utterance if the user's utterance does not provide any new information. For example, if the interrupting utterance was "I couldn't agree with you more," then it would be reasonable for the DM to decide to continue the Companion's planned utterances from the point where the interruption took place. In Figure 6, the user interrupts the tirade in Figure 5, F6 causing the system to stop the tirade and process the interruption. After the short loop response, the DM determines that it is not necessary to revise information, and so will just continue, acknowledging the interruption, and resuming the tirade from the point of interruption (i.e., repeating the interrupted utterance).
   - B. The DM would choose to replan if the user's utterance provides new information. This would be the case, for example, if the user's interrupting utterance corrected what the system had just said. The replan is necessary because the current ASM plan was generated from a set of assumptions which have now been shown to be false or incomplete. In Figure 7, the user F7 also interrupts the tirade in Figure 5. This time, after the short loop response, the DM determines that it is necessary to replan. The user interruption is understood as correcting the main topic to that of an increased workload for the user rather than discussion of redundancies. The tirade is then regenerated using this new main topic (with the strategy of remaining reassuring). Note that it is not necessary to generate a full tirade for this new topic, as we have already relayed about half of the previous tirade, so we generate an equivalent to the remaining amount for the new tirade.



<System>: I understand how unhappy you must be feeling.
<System>: try to understand Sarah will have little influence on the redundancies and their outcome.
<System>: redundancies are something I really dislike.
<System>: it's right *USER INTERRUPTS* to be unhappy about *STOPS TALKING*

<User>: *INTERRUPTING* yeah it makes me feeling ill just thinking about it

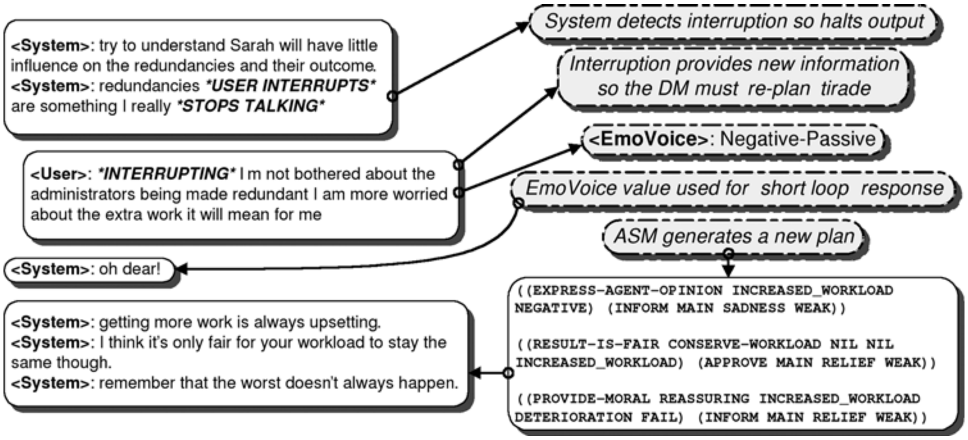
<System>: oh dear!

<System>: that s terrible!

<System>: it's right to be unhappy about such a bad situation.
<System>: remember that sometimes the worst doesn't always happen.

System detects interruption so halts output
Interruption provides no new information so the DM must continue tirade
<EmoVoice>: Negative-Passive
EmoVoice value used for short loop response
DM acknowledges interruption
Tirade is continued from where it was interrupted

Figure 6. *An example dialogue where the user interrupts without providing new information. The companion responds with "continue" interrupt handling.*



<System>: try to understand Sarah will have little influence on the redundancies and their outcome.
<System>: redundancies *USER INTERRUPTS* are something I really *STOPS TALKING*

<User>: *INTERRUPTING* I m not bothered about the administrators being made redundant I am more worried about the extra work it will mean for me

<System>: oh dear!

<System>: getting more work is always upsetting.
<System>: I think it's only fair for your workload to stay the same though.
<System>: remember that the worst doesn't always happen.

```
((EXPRESS-AGENT-OPINION INCREASED_WORKLOAD
NEGATIVE) (INFORM MAIN SADNESS WEAK))

((RESULT-IS-FAIR CONSERVE-WORKLOAD NIL NIL
INCREASED_WORKLOAD) (APPROVE MAIN RELIEF WEAK))

((PROVIDE-MORAL REASSURING INCREASED_WORKLOAD
DETERIORATION FAIL) (INFORM MAIN RELIEF WEAK))
```

System detects interruption so halts output
Interruption provides new information so the DM must re-plan tirade
<EmoVoice>: Negative-Passive
EmoVoice value used for short loop response
ASM generates a new plan

Figure 7. *An example dialogue where the user interrupts the companion with new information. The companion responds with "re-plan" interrupt handling.*

C. The DM chooses to abort if the user's utterance rejects the current dialogue strategy. An abort would be necessary if the interrupting utterance were something like "Don't talk to me about work, I'm not in the mood." An abort would discontinue the conversation until the user chose to continue by providing another utterance.

Handling non-barge-in interruptions is more straightforward as the user interrupts before the Companion has initiated its reply (i.e., the user continues speaking after the companion has registered the user turn as having finished, providing additional information after the companion has started processing the user turn, but before the companion has started delivering a response). The non-barge-in interruption can be summarized as follows:

1. The ATT detects an interrupt and informs the IM.
2. The IM informs the affective dialogue processing modules.
3. Affective dialogue processing modules disregard the current turn.
4. The DM continues, incorporating the previous turn into the next (i.e., merging the additional

Table 1. *Overview of User Testing Scenarios*

Scenario	Emotion	Events	Utterances	Emotional state
1a	Negative	Few	Short	Constant
1b	Positive	Few	Short	Constant
2	Negative	Many	Long	Constant
3	Negative to Positive	Many	Short	Mixed
4	1a + user defined	1a + user defined	User defined	User defined
5	User defined	User defined	User defined	User defined
6	Negative	Few	Short	Constant

information with the information previously being processed).

4 Evaluation

As the development of companion ECA requires the integration of a range of techniques to meet the challenges of conversational dialogue, so the evaluation of companion ECA requires new models of evaluation. We now present the results of an evaluation of our companion in which we concentrate on three main aspects of the system's functionality: main loop interaction, short loop interaction, and the use of the ECA. These aspects were considered in terms of the system's functional ability, the appropriateness of the companion's response, and the affective behavior of the companion.

In comparison to other evaluation methods (such as PARADISE; Walker, Litman, Kamm, & Abella, 1997) we do not attempt to reduce all features and parameters to a single, optimized figure but rather to target specific functionalities, capabilities, and behaviors within the system. This approach allows us to highlight the various strengths and weaknesses of the system while providing a means to characterize certain types of conversations that previously have proven difficult to reliably assess.

4.1 Testing Protocol and Scenario Design

The evaluations consisted of 12 extended sessions each taking approximately 2.5 to 3 hrs to complete. The 12 participants ranged from 22 to 54 years of age (with an average of 33), three were female and nine were male and all were native speakers of British English. After watching an introductory video, each participant trained the system (both EmoVoice and ASR) before undertaking a series of seven testing scenarios, concluding with a post-session questionnaire and interview.

The testing scenarios were constructed to provide a suitable breadth to the evaluation and to cover the fullest range of functionality. The evaluation team started with a pilot phase of testing in order to determine the companion's anecdotal strengths and weaknesses. Based on these considerations, a set of 20 initial testing scenarios were developed by the evaluation team with the number of scenarios gradually being refined down to the final seven. These seven scenarios were chosen as encompassing the desired focus on particular behaviors and capabilities of the companion: handling of varying lengths of user utterance, handling a range of events within the domain of office work, handling the range of emotional information provided by the user, testing of interaction loops, and testing of interaction with the ECA. Most of the seven scenarios were scripted so that the content, emotional reaction, and length was fixed. This scripting consisted of a short descriptor with an additional emotional direction for each user turn. The intention was to guide the conversation while encouraging the participant to respond naturally and with his or her own phrasing.

The breakdown of the final testing scenarios is shown in Table 1. Emotion determines both the nature of the T1
events discussed and the advised emotional reaction. Through the nature of the HYWD scenario itself, the system is weighted toward negative events in its coverage

Table 2. *Summary of Dialogue Metrics Across Scenarios*

Scenario	Average words per user utterance	User turns	System turns	Average words per system utterance	Average concepts per user utterance	WER	CER
1a	8.12	13.60	16.60	6.97	1.31	0.37	0.31
1b	8.31	14.67	16.67	6.51	1.62	0.33	0.31
2	10.00	11.00	12.60	7.63	2.14	0.44	0.34
3	10.07	19.67	26.17	6.58	1.72	0.36	0.34
4	9.57	19.17	20.33	5.90	1.40	0.35	0.39
5	10.11	15.50	13.83	5.41	1.13	0.40	0.26
6	6.30	13.40	15.20	5.55	1.17	0.35	0.33
Average	8.92	15.29	17.34	6.36	1.50	0.37	0.33
Range	4–23	7–31	3–38	1–9.21	0.05–4.57	0.15–0.93	0–0.65

and the testing scenarios reflect this. Events determines the total number of events discussed in the dialogue (with this also affecting the overall dialogue length). Utterances determines the number of events included in a given user turn. Note that short utterances are merely short in comparison; due to the conversational nature of the companion, these utterances are still generally longer than utterances in a typical task-oriented dialogue system. Finally, the emotional state determines whether the nature of and emotional reaction to the events varies between user turns.

Scenarios 4 and 5 are not scripted so as to test the companion in free conversation, although Scenario 4 uses a correlate of Scenario 1a to explicitly prime the conversation before allowing the user to continue with the free-form dialogue. Scenario 6 is a repeat of Scenario 1a but with the additional user interface windows removed so the participant can only see the ECA during the interaction.

Following completion of the seven testing scenarios, the participant completes a questionnaire. This consists of 35 statements answered using a 5-point Likert scale of strongly agree, agree, undecided, disagree, and strongly disagree. The statements are structured around six themes exploring the notion of a companion (Benyon & Mival, 2008). The participant is then interviewed by an evaluator for 5–10 min on what they like and dislike about the companion, their thoughts on the concept, and any other matters they wish to discuss regarding their experience.

4.2 Dialogue Metrics

We collected various dialogue metrics during each session covering the seven testing scenarios. The principal measures were the word error rate (WER) and concept error rate (CER). (CER was calculated by ignoring the order of recognized concepts with substitution errors only used in cases where part of the recognized and actual concepts match.) Table 2 provides a summary T2 of the metrics collected for each Scenario.

The total average WER of 0.37 and CER of 0.33 represent very poor scores for speech recognition and present obvious difficulties for a speech-based dialogue system. This result is surprising given the use of a trained ASR system, but may be explained partly by the emotional variation of the participants' voices.

The response time of the system was also measured. This was on the basis of the time from the end of the user's utterance until the start of the audio output from the system. (The text response on the user interface would typically appear before the audio output.) The time for a short loop response was as low as 1.2 s and averaged 2.28 s. The main loop response averaged 6.47 s. Notably, in the participant interviews, the length of the delay in the response was considered far less of an issue

Table 3. *Accumulated Score for Selected Statements*

Statement	Score
The companion surprised me at times	13
The companion demonstrated emotion at times	11
I thought the companion acted independently	10
The companion was polite	10
I thought the conversation was appropriate	−10
I liked the behavior of the companion	−10
The companion anticipated my needs	−10
The companion got to know me during the conversation	−12
The conversation was coherent	−15
The conversation between myself and the companion felt natural	−16
The companion's responses were always appropriate	−19
The companion is rather like me	−19

than the timing of the response. Participants wanted feedback regarding the state of the companion during the response delay, specifically, whether the companion was indeed going to deliver a response or not (there are several utterances per dialogue that receive no reply). They reported that the length of the delay was less impactful than not knowing if and when a response was coming, and the largest frustration was when they started talking again but the companion then proceeded to talk over them.

4.3 User Metrics

At the end of each session, the participant completed a questionnaire and was interviewed. Results for T3 scores greater than 10 or less than –10 are shown in Table 3 (with scores being calculated as the sum of individual Likert values from +2 to –2). The participants' responses to the questionnaire indicated that they felt the conversation with the companion was unnatural. When interviewed, it appears it was not that the participants found the prototype itself to be unnatural, but rather elements of the conversation; specifically, the fact that they were having conversations about their work day with a computer. It was not that they thought this to necessarily be an inappropriate thing to do; but rather that it was novel. Several participants noted that combining the HWYD companion type discussion with the additional utility of scheduling data could prove both cathartic and very useful.

Although the participants felt the companion was nothing like themselves, they clearly felt it did have a personality and that it acted independently, demonstrated emotions, and that it was polite and friendly. This personality was felt to be the case despite the occasional lack of coherence in some of the companion's responses and incorrect assignment of emotion to a user utterance.

The participants also reported feeling that the companion understood them better when there was no textual feedback of either ASR result or emotion detection, as was the case in Scenario 6. They also highlighted that they felt the entire interaction felt far more natural as they focused on the ECA itself rather than the written response (which they had intuitively done in the previous scenarios). Confusion over turn-taking still occurred, but people would spontaneously stop speaking when the ECA started to respond.

A linked issue reported by every participant was the lack of communication to the user by the system as to its internal state, specifically, whether it was or was not thinking about what to respond (i.e., was still in a listening state and whether it was going to respond or not). This is a fairly typical usability issue within any computational system where user frustration is increased not by the specifics of user interface feedback or the time to receive that feedback, but by not knowing whether any feedback is actually going to come or not. A next step in companion development would be the incorporation of various nonverbal cues (i.e., gaze, head nodding) to indicate floor-grabbing behavior.

4.4 Appropriateness Analysis

In addition to these objective and subjective measures of analysis we carried out further analysis through appropriateness annotation (Webb, Benyon, Hansen, &

Mival, 2010). To capture this information, annotators scored every utterance (i.e., both system and user utterances) within a dialogue for its appropriateness in terms of the level of information it contains and the progression of the dialogue so far. The aim is to reward appropriate behavior (answering questions, using new knowledge correctly) and penalize mechanisms that are seen as inappropriate between humans (incorrect use of knowledge, asking unrelated or off-topic questions and over-verification).

Annotators worked with the ASR output, so appropriateness is with respect to the information the system receives rather than what the participant actually said, marking each utterance with a code depending on whether it was a system or user utterance. User utterances could be a direct response to the system (RTS), eliciting a response from the system (RES), providing no response with this being appropriate (NRA), and providing no response with this being deemed inappropriate (NRN). System utterances could be one of nine categories: filled pauses (FP), requests for repair (RR), appropriate responses (AP), appropriate questions (AQ), appropriate new initiatives (INI), appropriate continuations (COM), and finally, inappropriate utterances containing inappropriate emotion (NAPE), content (NAPC), or some other defect (NAPF).

It is important to note that in this stage of the development and application of this evaluation methodology, we do not believe that the total score (or indeed individual annotation scores) are of the utmost usefulness.
T4 Instead, we believe that comparative scores (as in Table
F8 4), and label distributions across dialogues (as in Figure 8) are the most useful measures.

We start with a quick breakdown of the distribution of annotation labels across the entire evaluation (results marked average in Figure 8) which shows that the majority of utterances in the evaluation sessions (almost 30% overall) are responses by the user to system utterances (RTS). Unsurprisingly, the second largest category is appropriate questions asked by the system (AQ). Looking at the utterances labeled as inappropriate, we see that 3.22% of inappropriate labels are caused by incorrect emotional output (i.e., responding to a negative event with a positive utterance), and that 8.31% are caused by incorrect semantic content (i.e., a user states that he or she is working on the companion's project, and the next system question is, "What's the name of the project?"). If we take just the inappropriate utterances as a whole, we see then that around 30% of all errors are caused by inappropriate emotion handling, and the remaining 70% are from inappropriate content. Most utterances marked as appropriate responses by the system (AP) are in fact emotional statements made by the companion in response to user input. As an estimation of emotional system performance (from a subjective point of view), we can sum all appropriate responses (7.05%) with the inappropriate emotional responses (3.22%), and see that 10.27% of all utterances from the system contain some emotional output. In this context, we see that around 69% of all emotional output is deemed appropriate, with 31% being inappropriate given the context of the dialogue.

Table 4. *Appropriateness Scores per Scenario*

Scenario	Number of utterances	Average score	Score per utterance
1a	23.27	17.86	0.77
1b	27.75	16.63	0.6
2	20.09	13.59	0.68
3	35.5	25.13	0.71
4	33.45	20.18	0.6
5	23.58	10.63	0.45
6	28.5	19.05	0.67
Average	27.49	17.56	0.64

Examining the average score for each Scenario normalized for the length of the dialogue (score per utterance in Table 4) we find that our baseline condition, Scenario 1a, outperforms the average. Scenario 1b, by comparison, underperforms the average, despite the only difference being the polarity of events in the scenario. Most noticeably, scenarios involving any deviation from the script (Scenario 4 with slight deviation, and Scenario 5 with no script) score lower than average.

In terms of label distribution (Figure 8), our baseline condition, Scenario 1a, correlates strongly with the average. In Scenario 1b, we find that there are a greater num-

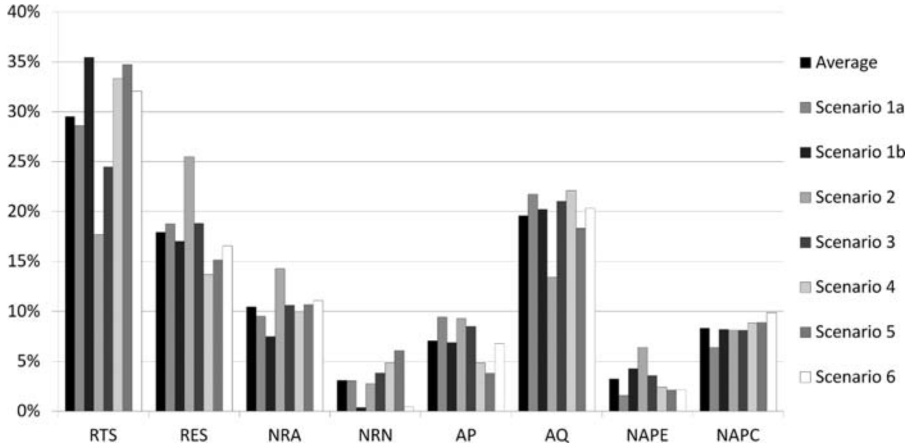


Figure 8. *Label distributions (as a percentage) across each scenario.*

ber of responses to the system than in Scenario 1a, as users give more information in response to systems questions. Also, where Scenario 1a had very few inappropriate emotional responses, the number in Scenario 1b is above average. This indicates the system struggled to recognize positive emotional events compared to the negative events used in Scenario 1a, and consequently had a hard time responding appropriately to clear, positive user events.

In Scenario 2 we find that the number of responses to the system is way below the average, as this scenario requires users to use longer utterances. As a consequence of receiving more information in longer utterances, the system has fewer questions to ask and the user gives longer, more involved responses to single questions. A trade-off to this is that the emotional response is harder to identify here, resulting in a greater than average number of inappropriate emotional responses, as perhaps it is harder to detect the overall emotional value in long utterances than in the shorter, clearer utterances.

Scenario 4 represents the first scenario where free-form user input is permissible (following a short script similar to Scenario 1a). To that end, we find a similar distribution to that in Scenario 1a and, although the system asks a greater number of appropriate questions and the user gives responses to those, we note a slight increase in inappropriate content (not recognizing the information exchanged from user to system) is also found. In Scenario 5, where users have completely free access to the system, although they are implicitly guided by prior interactions, we find an increase in utterances from the user that appear to warrant some response from the system, yet return nothing (i.e., the system is silent in response to some question or emotional comment from the user). We also find a corresponding drop in appropriate responses from the system, and fewer appropriate questions, all of which cause a drop in overall score for this scenario. (Encouragingly, given that users were free to interact as they saw fit, we do not see any significant increase in inappropriate responses.) This seems to indicate as the users deviate from the scripts (and by inference, the underlying template structure of the domain), the system has less to ask or respond with that is within the topic of the conversation. Consequently, it appears the system chooses to say nothing. We saw in previous evaluations using this appropriateness measure (Webb et al., 2010) that the use of simple conversational mechanisms found in ChatterBots may help to address these issues.

In the final scenario, Scenario 6, we see little deviation from the pattern found in Scenario 1a. This is merely confirmation that, in terms of appropriateness scores, this scenario performs equally well to our baseline. This scenario was designed to test various UI parameters, and consequently shows that the users and system performed more or less equally, whether the user had access to visual feedback from the system or not. In conjunction with the user feedback from subjective surveys, this would indicate that the best course of action is to remove the additional visual feedback for future trials and focus on the ECA.

5 Conclusion

We have presented a fully-implemented prototype of an ECA supporting affective dialogue under a truly conversational paradigm, which allows longer utterances both from the user and the agent, mixed-initiative as well as user interruptions. We conclude that our approach to the integration of conversational and affective aspects rests with the definition of interaction loops, all under the control of a top-level DM, orchestrating elementary dialogue steps (e.g., clarification), narrative utterances for advice giving, and user interruptions. It has reached maturity as a proof-of-concept system and is now the object of public demonstrations (Cavazza, Santos de la Camara, Turunen, & The Companions Consortium, 2010).

With respect to results, we have presented an evaluation using a new evaluation methodology designed specifically to highlight the strengths and weaknesses of conversational dialogue with a companion. The evaluation shows that, despite poor speech recognition performance and the novelty of the application, participants were able to use the system and felt the companion was polite, friendly, and exhibited a sense of personality. Further, through appropriateness annotation, we were able to compare various aspects of the interaction with the companion embedded in the seven testing scenarios used for the evaluation. This helped to identify the areas where the companion performed best and those areas requiring improvement. It also helped to suggest areas of further development of the companion such as new conversational mechanisms and a greater focus on the ECA versus additional visual feedback.

Acknowledgments

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 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 from the Companions Consortium include: Wei Wei Cheng, Morena Danieli, Carlos Sanchez Fernandez, Debora Field, Mari Carmen Rodriguez Gancedo, Jose Relano Gil, Ramon Granell, Jaakko Hakulinen, Sue Harding, Topi Hurtig, Roger Moore, Lei Ye, and Enrico Zovato.

References

André, E., Dybkjær, L., Minker, W., & Heisterkamp, P. (Eds.). (2004). Affective Dialogue Systems. *Lecture Notes in Computer Science, Vol. 3068, Proceedings of a Tutorial and Research Workshop.* Berlin: Springer-Verlag.

Appelt, D. E. (1985). *Planning English sentences.* Cambridge, UK: Cambridge University Press.

Benyon, D., & Mival, O. (2008). Landscaping personification technologies; From interactions to relationships. *Proceedings of the Conference on Human Factors in Computing Systems, CHI2008.*

Bevacqua, E., Mancini, M., & Pelachaud, C. (2008). A listening agent exhibiting variable behaviour. In *Lecture Notes in Computer Science: Vol. 5208, Intelligent Virtual Agents 2008* (pp. 262–269). Berlin: Springer-Verlag.

Bickmore, T., & Cassell, J. (1999). Small talk and conversational storytelling in embodied interface agents. *Proceedings of the AAAI Fall Symposium on Narrative Intelligence,* 87–92.

Bickmore, T., & Sidner, C. L. (2006). Towards plan-based health behavior change counseling systems. *Proceedings of AAAI Spring Symposium on Argumentation for Consumers of Healthcare.*

Boye, J. (2007). Dialogue management for automatic troubleshooting and other problem-solving applications. *Proceedings of the 8th SIGDial Workshop on Discourse and Dialogue.*

Boye, J., & Gustafson, J. (2005). How to do dialogue in a fairy-tale world. *Proceedings of the 6th SIGDial Workshop on Discourse and Dialogue.*

Boye, J., Gustafson, J., & Wirén, M. (2006). Robust spoken language understanding in a computer game. *Journal of Speech Communication, 48*, 335–353.

Cassell, J., & Thorisson, K. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence: An International Journal, 13*(4–5), 519–538.

Cavalluzzi, A., Carofiglio, V., & de Rosis, F. (2004). Affective advice giving dialogs. *Lecture Notes in Computer Science: Vol. 3068, Affective Dialogue Systems 2004* (pp. 77–88). Berlin: Springer-Verlag.

Cavazza, M., Charles, F., & Mead, S. J. (2002). Character-based interactive storytelling. *IEEE Intelligent Systems, 17*(4), 17–24.

Cavazza, M., Santos de la Camara, R., Turunen, M., & The Companions Consortium. (2010). How was your day? A Companion ECA. *Proceedings of AAMAS 2010.*

Crook, N., Smith, C., Cavazza, M., Pulman, S., Moore, R., & Boye, J. (2010). Handling user interruptions in an embodied conversational agent. *Proceedings of the AAMAS International Workshop on Interacting with ECAs as Virtual Characters,* 27–33.

De Angeli, A., & Brahnam, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with Computers, 20*(3), 302–310.

Englemore, R., & Morgan, T. (1988). *Blackboard systems.* New York: Addison-Wesley.

Grishman, R. (1997). Information extraction: Techniques and challenges. *Lecture Notes in Artificial Intelligence: Vol. 1299, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology,* (pp. 10–27). Berlin: Springer-Verlag.

Hernández, A., López, B., Pardo, D., Santos, R., Hernández, L., Relaño Gil, J., *et al.* (2008). Modular definition of multimodal ECA communication acts to improve dialogue robustness and depth of intention. *Proceedings of AAMAS 2008 Workshop on Functional Markup Language.*

Jönsson, A., Andén, F., Degerstedt, L., Flycht-Eriksson, A., Merkel, M., & Norberg, S. (2004). Experiences from combining dialogue system development with information extraction techniques. In: *New Directions in Question Answering.*

Kennedy, C., & Boguraev, B. (1996). Anaphora for everyone: Pronominal anaphora resolution without a parser. *Proceedings of COLING 1996,* 113–118.

Kopp, S., Stocksmeier, T., & Gibbon, D. (2007). Incremental multimodal feedback for conversational agents. *Lecture Notes in Computer Science: Vol. 4722, Intelligent Virtual Agents 2007* (pp. 139–146). Berlin: Springer-Verlag.

Moilanen, K., & Pulman, S. (2007). Sentiment composition. *Proceedings of the Recent Advances in Natural Language Processing International Conference (RANLP 2007),* 378–382.

Morency, L.-P., de Kok, I., & Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. *Lecture Notes in Computer Science: Vol. 5208, Intelligent Virtual Agents 2008* (pp. 176–190). Berlin: Springer-Verlag.

Nau, D., Ghallab, M., & Traverso, P. (2004). *Automated planning: Theory & practice.* San Mateo, CA: Morgan Kaufmann.

Prochaska, J., Di Clemente, C., & Norcross, H. (1992). In search of how people change: Applications to addictive behavior. *American Psychologist, 47,* 1102–1114.

Tørning, K., & Oinas-Kukkonen, H. (2009). Persuasive system design: State of the art and future directions. *Proceedings of the 4th International Conference on Persuasive Technology, Persuasive 2009.*

Vogt, T., André, E., & Bee, N. (2008). EmoVoice—A framework for online recognition of emotions from voice. *Proceedings of the Workshop on Perception and Interactive Technologies for Speech-Based Systems.*

Walker, M., Litman, D., Kamm, C., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics,* 271–280.

Webb, N., Benyon, D., Hansen, P., & Mival, O. (2010). Evaluating human-machine conversation for appropriateness. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010).*