Endowing Spoken Language Dialogue Systems with Emotional Intelligence

Elisabeth André[1], Matthias Rehm[1], Wolfgang Minker[2], and Dirk Bühler[2]

[1] Multimedia Concepts and Applications, University of Augsburg, Germany,
{andre,rehm}@informatik.uni-augsburg.de,
[2] Information Technology, University of Ulm, Germany,
{wolfgang.minker,dirk.buehler}@e-technik.uni-ulm.de,

Abstract. While most dialogue systems restrict themselves to the adjustment of the propositional contents, our work concentrates on the generation of stylistic variations in order to improve the user's perception of the interaction. To accomplish this goal, our approach integrates a social theory of politeness with a cognitive theory of emotions. We propose a hierarchical selection process for politeness behaviors in order to enable the refinement of decisions in case additional context information becomes available.

1 Introduction

The last decade has seen a general trend in Human Computer Interaction (HCI) to emulate certain aspects of human-human communication. Computers are ever less viewed as tools and ever more as partners or assistants to whom tasks may be delegated. Empirical studies by Nass and colleagues [1] show that users tend to apply social norms to computers. They blame computers for mistakes, they try to be polite if the computer asks them for their opinion and feel flattered if the computer compliments them on a good performance.

In order to build dialogue systems that are able to communicate with the user in a more natural manner, the consideration of social aspects is inevitable. Martinovsky and Traum [2] demonstrated by means of user dialogues with a training system and a telephone-based information system that many breakdowns in man-machine communication could be avoided if the machine was able to recognize the emotional state of the user and responded to it more sensitively. Earlier experiments by Prendinger and colleagues [3] have shown that an empathetic computer agent can indeed contribute to a more positive perception of the interaction.

There has been an increasing interest in the development of spoken dialogue systems that dynamically tailor their conversational behaviors to the individual user and the current situation (see [4]). Most approaches focus, however, on the adaptation of the content to be conveyed and pay hardly any attention to stylistic variations. Furthermore, social and psychological aspects have been largely neglected so far.

Waibel and colleagues [5] present a first approach to adjust dialogue behaviors to the user's emotional state. For instance, they suggest that more explicit feedback should be given if the user is frustrated. Nevertheless, their approach relies on a few selection rules and is not based on a general framework for affective dialogue. Walker and colleagues [6] examine how social factors, such as status, influence the semantic content, the syntactic form and the acoustic realization of conversations. They consider the speaker's emotional state to parameterize the acoustic realization of speech acts. However, they don't attempt at manipulating the hearer's emotional state by the deliberate choice of dialogue strategies.

In the HCI community, there have been various attempts to create virtual agents that display emotions via mimics, gestures and speech. Most of this work was driven by the goal to increase the agents' believability (e.g., see [7,8,9]). There are only a few agents that try to elicit user emotions in a controlled manner. For instance, the COSMO agent [10] intentionally expresses emotions with the aim to encourage the student while the GRETA agent [11] deliberately decides whether or not to show an emotion. Prendiger and colleagues [12] apply so-called social filter rules to determine the intensity of the emotion to be displayed depending on factors such as the personality of the agents and the social distance between them.

The objective of our work is to endow dialogue systems with emotional intelligence. Emotional intelligence includes the ability to recognize the user's emotional state as well as the ability to act on it appropriately. In this paper, we investigate how the user's affective response to a system may be improved by the mitigation of face threats resulting from dialogue acts.

2 A Theory of Social Interaction

According to Brown and Levinson [13], politeness strategies are communicative devices for redressing the threats inherent in verbal and nonverbal utterances. Positive politeness aims at protecting the individual's desire to be evaluated positively, for example by expressing admiration for the addressee. Negative politeness accounts for the individual's desire to act free from impositions, for example, by emphasizing that the final decision is up to the addressee.

Walker and colleagues [6] have shown how the Brown and Levinson approach may be successfully applied to the implementation of simulated dialogues between conversational agents. Depending on the expected threat to the user's basic desires, different dialogue styles are realized. Walker and colleagues choose one of the following four strategies:

1. Do the speech act directly
2. Orient the realization of the act to the hearer's desire for approval (positive politeness)
3. Orient the realization of the act to the hearer's desire for autonomy (negative politeness)
4. Do the act off record by hinting facts and/or ensuring that the interpretation is ambiguous.

Each of these main strategies has a number of substrategies that may be realized by different linguistic means. According to Brown and Levinson, strategies come with constant thresholds. Consequently, the main and substrategies can be organized into a hierarchy according to their increasing thresholds.

We attempt to remedy a shortcoming of the approach by Walker and colleagues. They work with purely hypothetical values from 0 to 150 without accounting for the

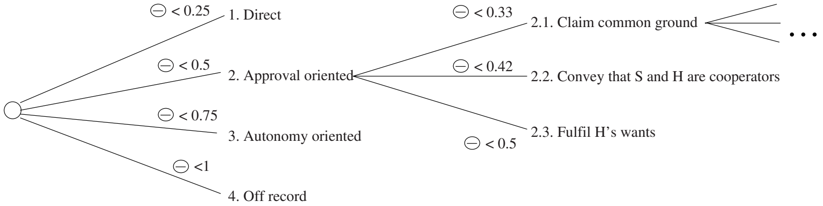

Fig. 1. A section of the hierarchy of strategies along with the thresholds.

origin of these values. We will start with a linear distribution of thresholds and check this assumption later against empirical evidence to come to a more realistic distribution. Moreover, our selection process is more fine-grained since we include the relevant substrategies in the selection process. Figure 1 gives an impression of the resulting hierarchy if a linear distribution of thresholds over the given strategies is assumed. For example, three approval oriented substrategies exist:

2.1. Claiming common ground,
2.2. Conveying that speaker (S) and hearer (H) are cooperators, and
2.3. Fulfilling H's want.

Each of these strategies can in turn be realized in different ways. Strategy 2.1, for example, may result in conveying that some desire of the speaker is also very interesting to the addressee or in claiming in-group membership with the addressee. To perform a speech act, the strategy with the highest threshold that does not exceed the threat is chosen. Consequently, the more serious a face-threat is the higher the threshold of the strategy should be. For instance, Strategy 2.1 would be chosen if the face threat was not too serious. Vice versa, using a higher-ranked strategy implies to the hearer that the threat is more serious because the interactors share knowledge about the organization of strategies.

The threat resulting from an utterance depends on the social distance between the speaker and the hearer D(S,H), the power that the hearer has over the speaker P(H,S) and, finally, on a ranking of imposition for the dialogue act under discussion R_α (α denotes the dialogue act). For instance, a command is ranked as a higher imposition than an offer.

In the setting of a computational system it is debatable whether D and P should be regarded as variables at all. Power definitely resides with the user that can simply switch the system off and attributing aspects like social distance to a technical system seems a bit far-fetched. But this perspective is only viable if users do not regard computational systems as interaction partners. Reeves and Nass [1] have shown in a convincing manner that this is actually the case and that users even contribute social behaviours to computers which they may or may not have in reality. Physical power (switching off) resides with the user but during the interaction between user and system the power of the roles played by system and user is the important aspect. In a tutoring scenario, e.g., the greater (social) power resides with the teacher who guides and challenges the pupil throughout the learning process. Consequently, in a tutoring system, the user has to accept the greater power of the system due to the social setting.

A similar effect can be observed for social distance. A user being confronted with a system for the first time is not yet familiar with its technical features and probably unsure how to interact with it. This is analogues to meeting a new collegue at work. If you talk to him for the first time, you might ask: "Excuse me, may you want to join me for lunch". The colloquial "Hey mate how about lunch" reflects a much lower social distance between the two conversational partners. In both cases, a lower bound for the two variables may be defined that is due to the roles of the interaction partners. Although teacher and student may come to know each other better over time, which will decrease the social distance between them, this will not result in peer-group relations.

3 Integration of an Affective Component

Walker and colleagues consider the speaker's emotional state to parameterize the acoustic realization of the speech act. However, the emotional state is not calculated due to situational or pre-defined personality factors and it has no effect on the choice of strategies at all. Instead, it is set as a fixed parameter for a given agent and only effects the acoustic output which becomes more variable in this way. Moreover, they don't attempt at manipulating the hearer's emotional state by the deliberate choice of dialogue strategies.

Our work starts from the assumption that the perceived threat resulting from a speech act heavily depends on the user's emotional state. For instance, if the user is already rather irritated due to communication problems, a proposal by the agent to input a long identification number is rather likely to be perceived as an impingement. Furthermore, knowledge about the causes for the user's emotions should guide the selection of politeness strategies. Consequently, the emotional state is a factor that emerges during the interaction and dynamically influences the ongoing dialogue.

We represent emotions using a dimensional model (see [14]) which characterizes emotions by the two orthogonal dimensions valence and arousal (see Fig. 2). Valence indicates to which degree an emotion is positive or negative and arousal refers to the extent of the emotion. A given emotional state is then characterized by a point in this two dimensional space. Emotion dimensions can be seen as a simplified representation of the essential properties of emotions. For instance, anger can be described by high arousal and negative valence.

Apart from the ease of describing emotional states that cannot be distributed into clear-cut fixed categories, the two dimensions valence and arousal are well suited for emotion recognition (see [15]).

Given this two dimensional model of emotions, how does the user's emotional state influence the choice of an appropriate strategy? To calculate the weight of the face threat θ, Walker and colleagues follow Brown and Levinson's proposal which takes the variables social distance, power, and ranking of the speech act into account: $\Theta = D(S, H) + P(H, S) + R_{\alpha}$. Apart from D, P, and R, Brown and Levinson mention situational factors that may influence the given variables to fit specific contexts and situations, but leave aside the question of how this influence is reflected. We treat the emotional state of the user as such a situational factor. Integrating this information in the weight calculation renders them to multiplicative factors that influence the other

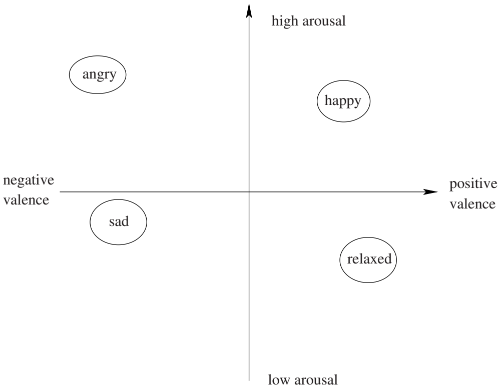

Fig. 2. The dimensional model with some example emotions.

variables. In our first approach, we assume the emotional state of the user E(H) to influence all three variables to the same amount according to the following heuristic:

- Positive valence: decrease the weight of threat θ
  - high arousal (e.g., joy): A possible threat is of no great significance to the user because he is in a very positive state. Thus, the decrease is high, $0 < E(H) < 0.5$.
  - low arousal (e.g., bliss): Like above, the user is in a positive state, the threat will affect him not with its full weight, but the decrease is low, $0.5 \leq E(H) < 1$.
- Negative valence: increase the weight of threat θ
  - high arousal (e.g., anger): The user is more sensitive to possible threats. The weight of the threat is increased by a large amount, $1.5 < E(H) \leq 2$.
  - low arousal (e.g., sorrow): The user is still sensitive to threats, but to a lesser degree then before. Accordingly, the increase is low, $1 < E(H) \leq 1.5$.

The question arises of whether the emotional state has a different influence on distance, power, and imposition. For example, anger might influence P(H,S) to a strong extent while sadness might have a significant impact on the variable R. Leaving these subtleties aside for the moment, our agent S estimates the threat Θ to the hearer H when performing the speech act α as:

$$\Theta = \begin{cases} 1 & : \quad \theta_E > 1 \\ \theta_E & : \quad otherwise \end{cases}$$

with θ_E defined as: $\theta_E = E(H) * \frac{1}{3}(D(S,H) + P(H,S) + R_\alpha)$

4 Illustration of the Model by Means of a Dialogue Example

To illustrate the approach, we start from an example taken from a Spoken Language Dialogue System (SLDS) developed by Bühler and colleagues [16] for appointment scheduling. Let us assume that the user has already fixed an appointment in the city of Heidelberg at a certain time and wishes to arrange another appointment the same day. Based on its reasoning process, the SLDS generates the following dialogue utterances:

System: *You already have an appointment in Heidelberg tomorrow.*
System: *Where do you want to go first, Stuttgart or Heidelberg?*

The user reacts by evaluating the first alternative as a hypothetical scenario:

User: *How much time would I have in Stuttgart if I went there first?*

After processing the travelling constraints, the system is able to provide the requested information:

System: *You would have at most 30 minutes.*

Now let us discuss how the enhanced dialogue system would interact with the user. Let us assume the social distance between system and user is very low and that the system has power over the user, i.e. D(S,H) = 0.1, P(H,S) = 0.3. The system decides to respond with an inform speech act to which a rather low imposition value is assigned: R = 0.3. We assume that the user has no preference for a specific schedule and is just exploring different options. Telling the user that there are only 30 minutes left would not necessarily result into a negative user emotion. Consequently, the face threat won't be influenced by the user's emotional state and the threat resulting from uttering an inform speech act is rather low: $\Theta = 0.23$. Therefore, the system would go for an effective solution and perform the act directly by responding:

(1) **System:** *You would have at most 30 minutes.*

Let us now assume that the user is in a rather bad emotional state due to repeated communication problems, i.e., negative valence and high arousal yield E(H) = 1.6. Consequently, $\Theta = 1.6 * 0.23 = 0.37$. Instead of a direct strategy, the application of an approval-oriented strategy is more advisable. According to Figure 1, the system decides to convey the impression that it and the user are cooperators:

(2) **System:** *I know this is quite irritating.*
System: *But, we would just have 30 minutes.*

Let us now start from a setting in which the social distance between system and user is high and in which the user has power over the system: D(S,H) = 0.8, P(H,S) = 0.8. Furthermore, we assume that the user insists on going to Stuttgart first.

User: *I want to go to Stuttgart first.*

In this case, the system would have to reject the user's request. The imposition associated with a refusal is ranked quite high: R = 0.8. In addition, the user will be in a rather bad emotional state since his goals cannot be accomplished: E(H) = 1.8. As a consequence, the threat resulting from the speech act is maximal: $\Theta = 1.0$. According to Fig. 1, the system would thus decide to use an off-record strategy. As already noted by [6], off-record strategies are most difficult to realize computationally. One possibility is to provide an evasive answer and to distract the user by mentioning the advantages of another option:

(3) **System:** *If you go to Heidelberg first, you will have enough time for the meeting in Stuttgart.*

Let us now assume that the user is in a rather relaxed and joyful state because he has just closed a profitable deal. Consequently, a change in the schedule won't have a great impact on his emotional state: E(H) = 0.7. In this case, the emotional state will decrease the threat resulting from the speech act: $\Theta = 0.8 * 0.7 = 0.56$. Therefore, the system would prefer an autonomy-oriented strategy.

(4) **System:** *Would it be ok if you had just 30 minutes in Stuttgart?*

5 Consideration of Additional Context Factors

Even though the theory by Brown and Levinson seems to provide a promising basis for the implementation of politeness strategies, it also creates a number of problems (see [17] for a more detailed discussion). First of all, the linear ordering of politeness strategies from direct over approval-oriented and autonomy-oriented to off-record may lead to inconsistencies. There is no doubt that the perceived politeness of any strategy may drastically vary depending on a number of context factors, such as the user's personality, his or her ability to perform a certain task etc. Indeed, a number of studies revealed that autonomy-oriented strategies are not always conceived as more polite than approval-oriented strategies. Furthermore, speakers tend to use several combinations of politeness strategies within one utterance. Even a dialogue act that is aimed at sustaining negative face of the hearer can be employed in an approval-oriented strategy. It is also questionable whether indirectness and vagueness are actually useful means to redress face threats. For instance, in the calendar scenario discussed above, vague system utterances might even increase the user's negative emotional state.

As a first step to improve the selection process, we consider not only the user's emotional state, i.e. the values for the dimensions valence and arousal, but rely on a specification of his or her personality profile and a classification of the events, actions and objects that caused it. This approach is in line with the OCC cognitive model of emotions [18] which views emotions as arising from some valenced reaction to events and objects in the light of agent goals, standards, and attitudes.

For example, a student that feels frustrated after repeatedly providing wrong answers to a tutoring system might interpret an autonomy-oriented strategy, such as "Why not try it again?", as pure irony. Since the event is not desirable to the user, but the user is responsible for the failure, the system would rather go for an approval-oriented strategy. On the other hand, if the failure is obviously caused by the system, the user might feel offended if the system applies an approval-oriented strategy and tries to convey the impression that they are collaborators, for instance by uttering: "Don't worry! Together, we will manage this!". Here, an excuse would be more appropriate.

The model that we have elaborated above can be regarded as the default case of emotional influence on strategy selection. It is employed if not enough knowledge about the interlocutor or the situation is available. For instance, a tutor might opt for an approval-oriented strategy because the student is lacking self-confidence. But such a strategy can still be realized in a number of different ways which mitigate face threats of different weights. Without any additional knowledge, we would now compute the expected face threat based on a refinement of the formula presented above.

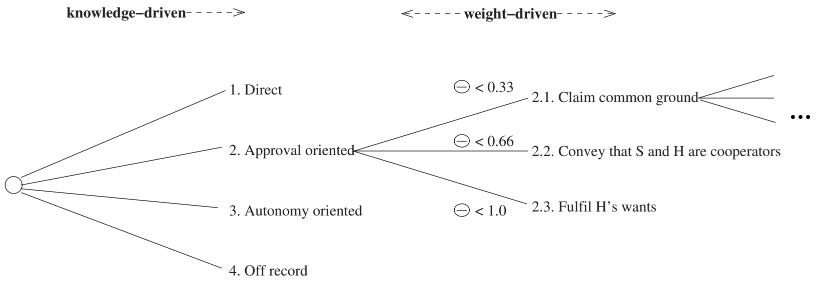

Fig. 3. A section of the modified hierarchy of strategies in the extended model along with the thresholds.

$$\Theta_i = \begin{cases} 1 & : \quad \theta_{E_i} > 1 \\ \theta_{E_i} & : \quad otherwise \end{cases}$$

with θ_{E_i} defined as: $\theta_{E_i} = E(H) * \frac{1}{3}(D(S,H) + P(H,S) + R_{\alpha_i})$ with i signifying either positive or negative politeness, i.e., $i \in \{pos, neg\}$.

Figure 3 gives a slightly modified version of the original choices. Depending on the available knowledge, strategies are either deliberately chosen by the speaker (knowledge-driven) or determined by the given thresholds (weight-driven). The thresholds are assigned dynamically in a linear fashion to the appropriate level in the hierarchy of strategies. Figure 3 gives an example where the choice between direct, approval-oriented, autonomy-oriented, and off-record is decided by the speaker in favor of an approval-oriented strategy leaving the application of the thresholds to the levels further down in the hierarchy. The dotted arrows indicate that the division between knowledge-driven and weight-driven decisions can shift in either direction.

Another attempt to consider a larger variety of context factors for the selection of politeness strategies has been proposed by Johnson et al. (see [19], this volume). They leave the choice between positive and negative politeness strategies to the deliberate decision of the speaker by calculating two FTA-weights that take different aspects of the speech acts and the choices of the speaker into account. Their extension of the Brown and Levinson model rests on observations of tutor-learner dialogues in a computer mediated learning environment. Consequently, the choice of an appropriate strategy in their model is driven by the knowledge of the tutor about the situation and about the learner. It is based on an analysis of the linguistic means employed by the tutor, leaving situational factors like the emotional display of the learner aside.

A straightforward combination of their and our approach would be to rely on their mechanism for choosing between approval- and autonomy-oriented strategies, but guide the selection of substrategies by the expected face threat on the user.

6 Conclusions and Future Work

In this paper, we have presented a new dialogue model that integrates a social theory of politeness with a cognitive theory of emotions. The objective of our work is to improve the user's subjective perception of the interaction by mitigating face threats that result from dialogue acts. The acceptance of a system that simulates certain aspects of social behaviors certainly depends to a large degree on the relationship between the user and the system. A user who views the system rather as a tool might be more interested in getting straight answers while a user who regards the system rather as a personal assistant may appreciate the system's attempts to exhibit certain kinds of social behavior.

To shed light on this question the next step of our work will be the conduction of an empirical study that compares the user's response to the affective and the non-affective version of the dialogue system by measuring and analyzing his or her physiological feedback. Therefore, while interacting with the two different system versions, the user is monitored by means of bio sensors, capturing skin conductivity, heart rate, respiration and muscle activity.

We aim at measuring the user's physiological feedback since self reports bear the disadvantage that users might not be able or willing to describe their true sensations. Furthermore, there is the danger that users might not remember the experience any more when asked after the interaction with the system. The advantage of bio signals as indicators of emotions, compared to external means of expression, is that they are hardly consciously controllable by the user and thus allow us to largely circumvent the artifact of social masking.

The goal of our studies will be to find correlations between patterns in the physiological state - as measured by means of bio sensors - and the threat resulting from the dialogue acts produced by the system. We hypothesize that an increased number of greater threats over a longer period of time will gradually result into user stress. Such a finding would also be in line with earlier experiments by [3] who showed by measuring the user's skin conductivity that users seem to be less stressed if the agent apologizes for delays. Even though biofeedback analysis may help to acquire knowledge about the user's emotional state while he or she is interacting with the system, our primary goal is to make use of this method to evaluate the affective impact of our politeness strategies and employ the resulting knowledge for guiding the selection and definition of politeness behaviors.

Our approach leaves a lot of room for extensions. First of all, face threats are inherently multi-modal. Dressing up a threat in a joke usually only works if the speaker shows in his whole appearance (facial expression, body posture) that he is telling a joke. Otherwise the threat might be even more severe than it is. To identify multi-modal conversational behaviours to control the strength of perceived face threats, we are currently preparing a corpus study which will also help us to ground the selection of politeness behaviors in empirical data.

Acknowledgments. We would like to thank Lewis Johnson for fruitful discussions on the implementation of politeness behaviors.

References

1. Reeves, B., Nass, C.: The media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press (1996)

2. Martinovski, B., Traum, D.: Breakdown in human-machine interaction: the error is the clue. In: Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems. (´2003) 11–16
3. Prendinger, H., Mayer, S., Mori, J., Ishizuka, M.: Persona effect revisited. using bio-signals to measure and reflect the impact of character-based interfaces. In: Fourth International Working Conference on Intelligent Virtual Agents (IVA-03), Springer (2003) 283–291
4. Litman, D.J., Pan, S.: Designing and Evaluating an Adaptive Spoken Dialogue System. User Modeling and User-Adapted Interaction **12** (2002) 111–137
5. Polzin, T.S., Waibel, A.: Emotion-sensitive human-computer interfaces. In: ITRW on Speech and Emotion. ISCA (2000) 201–206
6. Walker, M.A., Cahn, J.E., Whittaker, S.J.: Improvising linguistic style: Social and affective bases of agent personality. In Johnson, W.L., Hayes-Roth, B., eds.: Proceedings of the First International Conference on Autonomous Agents (Agents'97), Marina del Rey, CA, USA, ACM Press (1997) 96–105
7. Marsella, S.C., Johnson, W.L., Labore, C.: Interactive pedagogical drama. In Sierra, C., Gini, M., Rosenschein, J.S., eds.: Proceedings of the Fourth International Conference on Autonomous Agents, Barcelona, Catalonia, Spain, ACM Press (2000) 301–308
8. Gratch, J., Marsella, S.: Tears and fears: modeling emotions and emotional behaviors in synthetic agents. In Müller, J.P., Andre, E., Sen, S., Frasson, C., eds.: Proceedings of the Fifth International Conference on Autonomous Agents, Montreal, Canada, ACM Press (2001) 278–285
9. Paiva, A., Chaves, R., Piedade, M., Bullock, A., Andersson, G., Höök, K.: Sentoy: a tangible interface to control the emotions of a synthetic character. In: Proceedings of the second international joint conference on Autonomous agents and multiagent systems, ACM Press (2003) 1088–1089
10. Lester, J.C., Towns, S.G., Callaway, C.B., Voerman, J.L., FitzGerald, P.J.: Deictic and emotive communication in animated pedagogical agents. In Cassell, J., Prevost, S., Sullivan, J., Churchill, E., eds.: Embodied Conversational Agents. The MIT Press (2000) 123–154
11. Pelachaud, C., Carofiglio, V., De Carolis, B., de Rosis, F., Poggi, I.: Embodied contextual agent in information delivering application. In: Proceedings of the first international joint conference on Autonomous agents and multiagent systems, ACM Press (2002) 758–765
12. Prendinger, H., Ishizuka, M.: Social role awareness in animated agents. In: Proceedings of the fifth international conference on Autonomous agents, ACM Press (2001) 270–277
13. Brown, P., Levinson, S.C.: Politeness - Some universals in language use. Cambridge University Press (1987)
14. Lang, P.J.: The emotion probe: Studies of motivation and attention. American Psychologist **50** (2002) 372–385
15. Bosma, W., André E.: Exploiting emotions to disambiguate dialogue acts. In: Proceedings of the 9th International Conference on Intelligent User Interface, ACM Press (2004) 85–92
16. Bühler, D., Minker, W.: An architecture for logic-based human-machine dialogue. In: ISCA ITRW Workshop on Spoken Multimodal Human-Computer Dialogue in Mobile Environments, Kluwer (2002)
17. Knapp, M.L., Daly, J.A.: Handbook of Interpersonal Communication. Sage Publications (2003)
18. Ortony, A., Clore, G.L., Collins, A.: The cognitive structure of emotions. Cambridge University Press (1988)
19. Johnson, L., Rizzo, P., Bosma, W., Kole, S., Ghijesen, M., van Welbergen, H.: Generating socially appropriate tutorial dialog. In: Tutorial Workshop on Affective Dialogue Systems, Springer (2004) this volume.