Reference and Gestures in Dialogue Generation: Three Studies with Embodied Conversational Agents

Paul Piwek*

*ITRI
University of Brighton
Brighton, UK
Paul.Piwek@itri.brighton.ac.uk

Judith Masthoff†

†Dept. of Computing Science
University of Aberdeen
Aberdeen, UK
jmasthoff@csd.abdn.ac.uk

Malin Bergenstråhle‡

‡Dept. of Aeronautical and Vehicle Engineering
Royal Institute of Technology
Stockholm, Sweden
malbe@kth.se

Abstract

This paper reports on three studies into social presence cues which were carried out in the context of the NECA (Net-environment for Embodied Emotional Conversational Agents) project and the EPOCH network. The first study concerns the generation of referring expressions. We adopted an existing algorithm for generating referring expressions such that it could run according to an egocentric and a neutral strategy. In an evaluation study, we found that the two strategies were correlated with the perceived friendliness of the speaker. In the second and the third study, we evaluated the gestures that were generated by the NECA system. In this paper, we briefly summarize the most salient results of these two studies. They concern the effect of gestures on perceived quality of speech and information retention.

1 Introduction

In this paper, we describe a number of evaluation studies which were carried out in the context of the NECA project and EPOCH Network.[1] The studies evaluate a variety of strategies for generating social cues for embodied conversational agents. These strategies were implemented in the NECA system. The strategies which we discuss concern the generation of referring expressions, speaker gestures and hearer gestures. In Section 2, we first describe the NECA application and its requirements regarding social presence. The next two Sections, 3 and 4, describe strategies for generating referring expressions and gestures, respectively. Both sections consist of a description of the strategies followed by an overview of the evaluations that were carried out. In Section 3, we describe personality related strategies for generating referring expressions (definite descriptions and pronouns) which, to our knowledge, have not been proposed before. The generation of gestures has been studied by many before us. We did, however, obtain some interesting new results, in particular, regarding the relation between perception of speech and gestures and the effectiveness of hearer gestures. Finally, in Section 5 we provide our conclusions.

2 The NECA application and its requirements

The aim of the NECA project was to build a platform for web delivered performances of credible computer-generated characters. The project built on the pioneering work by André et al. (2000) on presentation teams of embodied conversational agents; but

Figure 1: Socialite Application

Figure 2: User Interface for Character's Personality/Mood Selection

see also the work by Cassell et al. (1994) on generating conversations for multiple animated agents. The members of such a presentation team engage in a dialogue with each other in order to inform and entertain the user. The user cannot directly interact with the characters, but does have the ability to set certain parameters before a dialogue/presentation takes place. These parameters partly determine the course of the dialogue. For instance, the user might be able to select the topic of conversation and certain personality traits of the interlocutors.

In the NECA project, two applications were implemented: *Socialite* and *eShowroom*. The Socialite application automatically generates multimodal dialogues between virtual students from a student area in Vienna known as 'der Spittelberg'. These dialogues are embedded in a webbased multi-user environment (Krenn et al., to appear). Rendering is performed using the Macromedia Flash Player, see Figure 1.

The eShowroom application automatically generates car-sales dialogues between a virtual seller and buyer. It allows a user to select a number of parameters –topic, personality and mood of interlocutors– which govern the automatically generated car-sales dialogues. Figure 2 shows one of the screens for making such selections. The presentations were originally generated using Microsoft Agents^{TM} (see Figure 3), but in the final version of the system the Charactor^{TM} player technology was used (see Figure 4).

The automatically generated presentations are intended to both inform and entertain the user. Information comes from the content of the dialogues, in which the interlocutors discuss the positive and negative attributes of one or more cars. The information provided goes beyond that pertaining to the specific cars under discussion: the interlocutors also connect
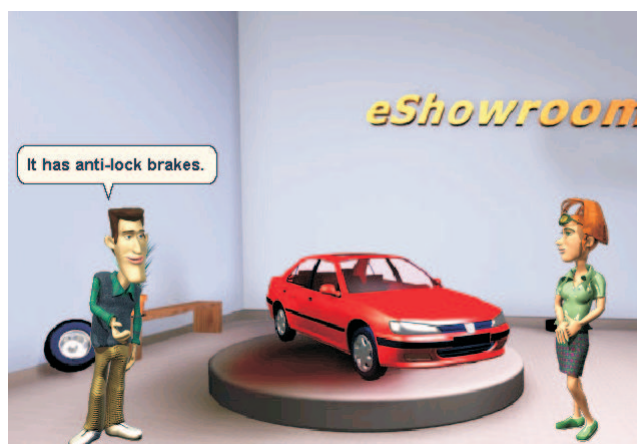

Figure 3: eShowroom with MS Agents^{TM}

Figure 4: eShowroom with Charactor^{TM}

facts about the car with value judgements, as illustrated in the following dialogue fragment from an eShowroom dialogue:

Tina:	Does it have power windows?
Ritchie:	I'm afraid not.
Tina:	This car is not exactly very prestigious.
	What kind of luggage compartment does it have?

The eShowroom prototype is intended to demonstrate a new way of presenting information on the internet to potential car buyers. Most of the information could also have been presented by means of a plain text or even a table. The main reason for using a dialogue with embodied conversational agents is to make it more entertaining for the user to learn about a car. For this purpose, the dialogues need to be engaging. We tried to achieve this by giving the agents a distinct personality which is displayed through their use of language. In particular, the characters in the eShowroom demonstrator can be polite or impolite and good humoured or ill tempered (the user can decide which, see Figure 2). We also aimed at having the agents produce plausible gestures when speaking and listening. This should make the presentations more believable and also more lively and therefore more likely to capture the attention of the user.

3 Generating Personalized Referring Expressions

In the field of Natural Language Generation (Reiter and Dale, 2000; Belz et al., 2004), there is a common assumption that a natural generation system needs to make decisions on at least two levels when constructing natural language output:

1. Decisions on *what to say*, i.e., on the *content* of the current utterance and

2. decisions on *how to say it*, i.e., on the *form* of the current utterance.

Decisions on both levels have an impact on the social presence cues which a speaker emits. Decisions on the form level can, for instance, reflect whether a speaker is introvert or extrovert: one of the findings by Gill and Oberlander (2002) is that in emails, introvert people use 'hello' where extroverts use 'hi'. In Ball and Breese (2000) the use of Bayesian networks is proposed to implement decisions regarding form such as the aforementioned one. Fleischman and Hovy (2002) describe a generate-and-test algorithm

for emotion expression through lexical decisions regarding verb selection and object descriptions. Others have looked at decisions which are related to both content and form: e.g., Hovy's seminal work on pragmatic constraints in generation (Hovy, 1988), and more recent studies into politeness in generation (Walker et al., 1996; Porayska-Pomsta and Mellish, 2004).

Here we want to explore how social cues can be displayed through generation decisions for a specific class of expressions, i.e., referring expressions. These are phrases which are used to identify objects in a domain of conversation (the domain of conversation can encompass objects in the immediate environment of the interlocutors, but also include objects only accesible from memory) to the addressee (*cf.* Dale and Reiter (1995)). To our knowledge, there is no work so far in the natural language generation community on strategies for generating referring expressions to display social cues.[2] Our focus is on decisions regarding content selection in the generation of referring expressions.

3.1 Two strategies for generating referring expressions

As our starting point, we use an algorithm for generating referring expressions which is loosely based on Krahmer and Theune (2002). This algorithm implements the widely accepted idea that the content of a referring expression depends on whether the target object –the object which the speaker intends to refer to– has been referred to before in the discourse or is comparitively prominent for other reasons. In particular, the algorithm decides how much descriptive content a referring expresssion (e.g., 'the sports car' versus 'the car' versus 'it') should contain on the basis of the salience of both target object and the other objects in the domain of conversation with which it might be confused. For this purpose, all objects in the domain of conversation are assigned a number which

---

[2]Note though that psychologists have explored the assumption that, in particular, children show egocentric behaviour when performing referential communication tasks. Children are alleged to have difficulty conceptualizing a situation differently from their own perceptual view and therefore perform differently from adults on such tasks (Piaget and Inhelder, 1956; Glucksberg et al., 1966). Others have, however, contested this view, and explained the effects that were found in terms of other capabilities which children of a certain age lack (e.g., Maratsos (1973)).

The idea that the choice of referring expression might not only depend on situational factors, but also on attributes of the speaker has been put forward by, for example, Piwek and Beun (2001) on the basis of an empirical study into referential behaviour in task-oriented dialogues.

represents the salience of the object.[3] At the outset of a conversation all objects will typically receive the salience value 0. We use the following two rules to update the salience values:

> INCR RULE (Increase): If an object is referred to in the most recent utterance, then increase the salience value of this object to 10.

> DECR RULE (Decrease): If an object is not referred to in the most recent utterance, then decrease the salience value of this object with 1, unless (i) the salience value is already 0 or (ii) the utterance is very short (less than 150 characters) and the utterance contains no referring expressions at all.[4]

The referring expressions generation algorithm takes as its input the current target object and a representation of the domain of conversation, which includes a salience value for each of the objects in the domain and also the properties which are true for each of the objects in the domain. Properties in the car sales domain were 'red', 'silver', 'car', 'for families', etc. The output of the algorithm consisted of one of the following three: 1. a set of properties, 2. the information that a pronoun should be used for reference, or 3. a failure message. The following provides an idea of how the algorithm computes the output, glossing over some details that are not relevant to the topic of this paper, in particular, regarding the preference ordering we used for selecting properties (Dale and Reiter's preferred attributes, see Dale and Reiter (1995)): The algorithm suggests to use a pronoun if the target object is the only object with salience 10. Otherwise, it tries to find a set of properties which distinguishes the target object from all other objects in the domain of conversation which are at least as salient as the target object. If no such set of properties can be found, the algorithm returns a failure message.

The main innovation which we introduce and which was implemented in the NECA multimodal generator (MNLG; see Piwek (2003b)) was to give each interlocutor their own record of salience values and personalized strategies for updating these values. More precisely, for each agent A there is function SV_A which maps objects in the domain of conversation to salience values (i.e., integers in $[0 - 10]$). The domain of conversation and the associated salience values can be seen as forming part of the common ground (*cf.* Clark (1996)) of the interlocutors. Ideally, both interlocutors share the same common ground and use the same strategies for updating it, such that for all objects in the domain of conversation the interlocutors have the same salience values. We would, however, like to investigate the supposition that egocentric versus non-egocentric speakers might differ with respect to their strategies.

An egocentric individual restricts its outlook or concern to its own activities. We propose that such an individual also behaves along these lines when updating the salience values of objects. In particular, such an individual will only increase the salience value of an object to 10, if she or he referred to the object. If it was the other agent who referred to the object, then this is not taken into account and rather than increasing the salience value, the utterance is treated as if it contained no references to the object at all, i.e., the salience value of the object is decreased by 1. Alternatively, a 'normal' individual is supposed to take the referring expressions of all interlocutors into account. Thus an agent A, depending on his or her personality, updates SV_A according to one of the following two strategies:

> EGOCENTRIC UPDATE STRATEGY: Try to apply the INCR RULE only after utterances by agent A. Try to apply the DECR RULE regardless of who the speaker of the utterance was.

> NEUTRAL UPDATE STRATEGY: Try to apply both the INCR RULE and the DECR RULE after each utterance, regardless of who the speaker was.

In a domain of conversation with two cars, one grey and the other a red sportscar, the following two dialogue fragments illustrate the differences between the two strategies. We have italicized the relevant referring expressions. This fragment was generated using the NEUTRAL UPDATE STRATEGY for both interlocutors:

Ritchie:	But what do you think about *this grey car*?
Tina:	Does *it* have anti-lock brakes?
Ritchie:	Certainly!
Tina:	Excellent! Does *it* have leather seats?
Ritchie:	I'm afraid not

---

[3] In Krahmer and Theune (2002), the term salience is used. This notion is, however, often associated with visual or auditory salience of a signal. A better term for the notion which Krahmer and Theune (2002) have in mind is that of 'accessibility' (see Ariel (1990)), since that notion is intended to also cover 'salience' of an object in memory.

[4] This second condition is intended to avoid that very short utterances such as 'ok', 'agreed', 'yes', 'no', etc. force the algorithm to produce a full description rather than a pronoun after such utterances. For instance, without the condition on utterance length we would generate 'A: Is the car safe? B: Yes. A: Does *the car* have airbags?' rather than 'A: Is the car safe? B: Yes. A: Does *it* have airbags?'. Note that this condition is not used in Krahmer and Theune (2002).

The following fragment was generated using the EGOCENTRIC UPDATE STRATEGY for Tina and the NEUTRAL UPDATE STRATEGY for Ritchie. Note that Tina only pronominalizes her references after *her own* first-mention reference.

Ritchie:	But what do you think about *this grey car*?
Tina:	Does *this grey car* have anti-lock brakes?
Ritchie:	Certainly!
Tina:	Excellent! Does *it* have leather seats?
Ritchie:	I'm afraid not

3.2 Evaluation

The strategies which we discussed in the previous section were implemented in the NECA MNLG (Piwek, 2003b). Our main aim with the evaluation of these strategies was to determine whether the strategies produced a noticable effect on the user's subjective impressions of the dialogues. We did not investigate whether the strategies actually corresponded to strategies employed by human dialogue participants. Also, our focus was on the effect which the strategies produce on an observer of a dialogue (as was the set-up of the NECA system); we did not concern ourselves with the effects they might produce on the dialogue participants themselves.

Method 40 undergraduate computing students participated in the evaluation (2 women and 38 men, mean age 23.9 years). Using the NECA MNLG we created two dialogues (fragments of these dialogues are given in the previous section) D_1 and D_2. For D_1, we used the NEUTRAL UPDATE STRATEGY for both Tina and Ritchie. For D_2, we used the NEUTRAL UPDATE STRATEGY for Ritchie and the EGOCENTRIC UPDATE STRATEGY for Tina. In all respects apart from the referring expressions, the dialogues were identical.[5] We divided the participants randomly between two groups: one group which was presented with dialogue D_1 and another group which was presented with dialogue D_2. Participants of both groups were asked to fill out a questionnaire with the following eight questions after they had seen the dialogue, where each answer was a value on a scale from 1 (e.g., 'not friendly at all') to 9 (e.g., 'very friendly'):

1. How friendly is Ritchie the car salesman?

2. How friendly is Tina the customer?
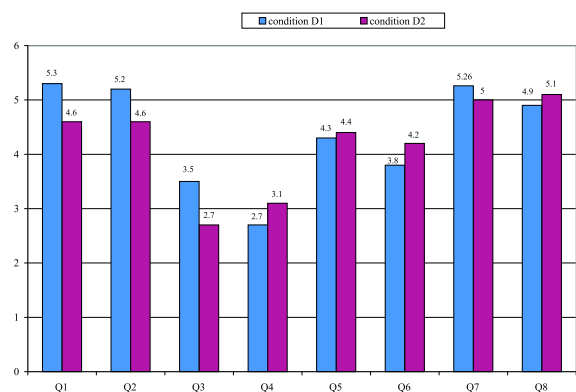
3. How smooth was the conversation?

Figure 5: Results evaluation of Reference Strategies

4. How entertaining did you think the dialogue was?

5. How agressive was Tina's attitude?

6. How aggressive was Ritchie's attitude?

7. How egocentric was Tina's attitude?

8. How egocentric was Ritchie's attitude?

Results The averages of the answers to the questions in the questionnaire can be found in Figure 5. We performed a t-test (one-tailed) to determine the statistical significance of the differences between the averages. We predicted that Tina would be perceived as less friendly, more aggressive and more egocentric when using the egocentric strategy in D_1. Only the difference between the averages for question 2 (concerning Tina's friendliness) turned out to be statistically significant with $df = 38$ and $t = 1.75$ at $p < 0.05$.[6] We also predicted that dialogues with Tina in egocentric reference mode would be perceived to be less smooth, but possibly more entertaining. The results were in the right direction (answers to Q3 and Q4), but not statistically significant.

Discussion The result for question 2 gives some weak credence to our hypothesized effect of the EGOCENTRIC UPDATE STRATEGY. Tina, when using the

---

[5]The materials we used are available from http://www.itri.bton.ac.uk/projects/neca/.

[6]The t-test tells us how likely it is that the means of the two populations are equal based on actual distance between the means and the within group variability of the two groups. The magnitude of t increases as the distance between the means increases and the within-group variability descreases. As t increases, the probability of the means being equal decreases. The non-significance of the results for Q1, even though the difference between the means of Q1 is larger than that of Q2, is explained by the fact that there was more within group variability for Q1 (i.e., for Q1 we had standard deviation $SD = 1.7$ for condition D_1 and $SD = 1.4$ for condition D_2, whereas for Q2 we had $SD = 1.2$ for condition D_1 and $SD = 1$ for condition D_2).

egocentric strategy is perceived to be less friendly than when she is using the neutral strategy. The answers to our other questions showed a tendency in the right direction, but were not statistically significant. There are some caveats when interpreting the results that we obtained. Firstly, most of the participants in the evaluation were male. It could be that the results do not generalize to a population of both male and female participants. Other studies with embodied conversational agents have found some effects related to whether the user is male or female (e.g., Buisine et al. (2004)). Additionally, we focussed on varying the behaviour of a single female agent. Thus our results are, so far, limited to perception of (mainly) male participants of two strategies in a female agent. In further experiments, we will need to verify whether the result extend to all combinations of male/female participants and male/female agents.

A further limitation of this study is the small amount of materials that was used. We had only two dialogues with instances of the independent variable (the dialogue strategy). We are planning to carry out further studies with a larger set of dialogues in order to verify that the reported effects are not due to random variation in the materials (*cf.* Dehn and van Mulken (2000)).

4   Gestures

The aim of this section is to discuss some interesting results we found when evaluating the gestures that are generated by the NECA eShowroom demonstrator. Our findings suggest that gestures, as cues of social presence, have to be added to an Embodied Conversational Agent with care, if one is to avoid unintended side-effects. The materials for the evaluations which we carried out are different from many existing studies in that we focussed on presentation teams of agents communicating with each other, rather than directly with the user.

4.1   Gesture Generation in NECA

The version of the eShowroom demonstrator (see Figure 3) that we discuss in this paper can insert three types of gestures:

1. Turn-taking signals: when a speaker has finished a turn, s/he looks at the other interlocutor and continues to do so whilst the other interlocutor speaks. When a speaker begins speaking, s/he looks slightly away from the other interlocutor.

2. Discourse function signals: these gestures are associated with the dialogue act type of an utterance. A distinction is made between, for instance, inform and request dialogue acts. The former cause the speaker to extend his/her hand to the hearer in a downward movement. The latter can cause the speaker to place their hands on their hips or raise a finger in the air. For a particular dialogue act, the generator selects at random a gesture from a set of suitable gestures. This approach is aimed at introducing some variation into the dialogue.

3. Feedback gestures: These are gesture by the hearer signalling attention to the speakers message or reflection on it, etc.

4.2   Evaluation

It is beyond the scope of this paper to fully describe the evaluation studies that we carried out regarding gestures. Rather, we highlight the, in our view, most salient results. For a full description of the studies we refer to the following two technical reports: Piwek (2003a) (for evaluation of speaker gestures) and Bergenstrahle (2003) (for evaluation of hearer gestures). The studies we carried out can be characterized as follows:[7]

- Speaker gestures study (28 participants): We compared a version of the system with speaker gestures (and no hearer/feedback gestures) with a version with no gestures at all.

- Feedback gestures study (12 participants): We compared a version of the system with speaker gestures (and no hearer/feedback gestures) with a version with both speaker and hearer gestures.

Method   In both studies, participants were divided into two groups according to conditions we wanted to compare (see above). After participants had been shown a dialogue, they were asked a number of subjective experience questions (e.g., how engaging/natural/etc. was the dialogue) whose answers were a point on a likert-type scale. Additionally, they were asked a number of multiple choice questions to test their retention of the information exchanged in the dialogue. Finally, there was also an open question asking participants for their views.

Results   For both studies, the results on subjective user experiences were not statistically significant. Note, however, that, in particular, for the study with feedback gestures, the number of participants was very small.

For the study on speaker gestures, we did however notice a surprising pattern in the answers to the open

---

[7]The materials we used are available from http://www.itri.bton.ac.uk/projects/neca/.

question: whereas in the *with speaker gestures* condition 41% of the subjects complained about the quality of speech synthesis only 9% the subjects did so in the *no gestures* condition.

The study on speaker gestures yielded no significant results on the retention test, although there was tendency for the subjects in the *with speaker gestures* group to do better on the test than those in the *no gestures* group. A power test showed that we would need approximately 195 subjects to validate this effect (effect size was $r = 0.23$).

Interestingly, the study on feedback gestures did yield statistically significant results for the retention test (with $df = 10$, $t = 2.6$ and $p < 0.05$), despite the very small number of subjects. The result was, however, that the participants of the *no feedback gestures* condition did better than those of the *with feedback gestures* condition.

Discussion   Our results on the effect of speaker gestures showed that the addition of speaker gestures did not lead to a statistically significant improvement of the subjective user experiences and retention for our users. This result is disappointing, since it suggest that gestures need not have been included in the NECA system. What is more, inclusion of gestures for some reason seemed to make the participants more sensitive to the inadequacies of the speech synthesis (in the study, we used the L&H TruVoice^{TM} TTS with American English voices. This engine comes for free with Microsoft Agents^{TM}). We do, however, need to keep in mind that the latter effect could be due to rather artificial side-effects of the way Microsoft Agents^{TM} integrates speech and animations: animations often introduce pauses in the speech which are not there without the animations (and even though on the level of specification, no pauses are introduced explicitly either).

Additionally, the presence of speech balloons, see Figure 3, introduces a further complicating factor. It could be that the gestures detracted the attention from the speech balloon and that given the relatively low quality of the synthesized speech, this made it more difficult to understand/follow the speech for the participants.

It also has to be said, that our results regarding retention and subjective experience (see Piwek (2003a) for details) did go in the right direction. It might be that the effect is small, but still could be established in an evaluation design with significantly more subjects. We also have to point out that we restricted our study to short presentations of a single car. It might be that if users watch several dialogues, they do get

to appreciate the inclusion of gestures more.

The results on feedback gestures did cause us to remove these from the final NECA eShowroom demonstrator, given that these results indicated that inclusion of such gestures was counterproductive regarding retention. This results is not completely surprizing, e.g., (Craig et al., 2002, page 433) have also suggested that gestures can sometimes have a negative effect by distracting the user.

5   Conclusions

In this paper, we introduced three studies regarding social presence cues. The studies were exploratory in nature, and in this conclusions section we would like to briefly discuss how to proceed from here.

We hope that the study regarding the generation of referring expressions has introduced a new perspective on referring expressions generation. So far, work in this area has concentrated on finding a single optimal strategy for identifying the target object. The focus has been on ease of production of such expressions for the speaker and ease of interpretation for the hearer. Some work has been done on implicatures that might be generated by using particular expressions (e.g., Jordan (2000)), but to our knowledge no one has so far considered the implications that differing strategies might have for the perceived personality of the speaker. Our evaluation study provides us with some modest results on the effect of personalized referring expression generation strategies on dialogue observers. We intend to explore the effectiveness of personalized referring expression generation further in the future with more extensive experiments.

Our results from the studies on gestures tell us that in implemented systems gestures do not always have the intended effects. This can be due to the inevitable limitations of the technologies currently available. We also note that although many of our results were in the right direction, possibly due to small sample sizes, the statistical significance of these effects could not be established.

We would like to emphasize that the studies we carried out were primarily intended as *evaluations* of different incarnations of the NECA system. Such evaluations can help determine which version of the system is 'better' in certain respects and therefore the preferred choice for integration into an application. Such evaluations do *not* directly test general claims about the usefulness of embodied agents, gestures, etc.

Our agents are by no means the most optimal ones possible and therefore any conclusions about them do not generalize to future generations of embodied

agents. To test general claims about the usefulness of embodied agents, it might be better circumvent limitations of current technologies by working with human actors to compare, for example, information presentation through speech only and speech with gestures, etc. Results from such studies could function as a reference against which results for computer-generated embodied agents could be compared.

Further problems with obtaining results regarding computer-generated embodied agents concern the fact that the effects which we try to measure are potentially quite small. For the field to build up a body of results for such small effects, the development of standards and frameworks for evaluation are highly necessary (*cf.* Ruttkay and Pelachaud (2004)). This could, for instance, allow for the application of meta-analysis studies which is common in many fields from physics to behavioural studies. Such studies are also called for due to the fact that most evaluation work on embodied conversational agents is restricted to the system/prototype developed by specific research groups (*cf.* Dehn and van Mulken (2000)).

References

E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. The automated design of believable dialogues for animated presentation teams. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, pages 220–255. MIT Press, Cambridge, Massachusetts, 2000.

M. Ariel. *Accessing Noun-Phrase Antecedents*. Routledge, London, 1990.

G. Ball and J. Breese. Emotion and personality in a conversational agent. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*. MIT Press, Cambridge, Massachusetts, 2000.

A. Belz, R. Evans, and P. Piwek, editors. *Natural Language Generation: Third International Conference (INLG 2004)*, volume 3123 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, 2004.

M. Bergenstrahle. Feedback gesture generation for embodied conversational agents. Technical Report ITRI-03-22, ITRI, University of Brighton, UK, 2003.

S. Buisine, S. Abrilian, and J.C. Martin. Evaluation of multimodal behaviour of embodied agents. In Zs. Ruttkay and C. Pelachaud, editors, *From Brows to Trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers, 2004.

J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer Graphics and Interactive Techniques*, pages 413–420, 1994.

H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

S. Craig, B. Gholson, and D. Driscoll. Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology*, 94(2):428–434, 2002.

R. Dale and E. Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263, 1995.

D. Dehn and S. van Mulken. The impact of animated interface agents: a review of empirical research. *Int. J. Human-Computer Studies*, 52:1–22, 2000.

M. Fleischman and E. Hovy. Towards emotional variation in speech-based natural language generation. In *Proceedings of the Second International Natural Language Generation Conference (INLG02)*, pages 57–64, NY, USA, Jully 2002.

A. Gill and J. Oberlander. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368, Fairfax VA, 2002.

S. Glucksberg, R. Krauss, and R. Weisberg. Referential communication in nursery school children: Method and some preliminary findings. *Journal of Experimental Child Psychology*, 3: 333–342, 1966.

E. Hovy. *Generating Natural Language Under Pragmatic Constraints*. Lawrence Erlbaum, Hillsdale, New Jersey, 1988.

P.W. Jordan. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from and Empirical Study*. PhD thesis, University of Pittsburgh, Pennsylvania, 2000.

E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing*, pages 223–264. CSLI, Stanford, California, 2002.

B. Krenn, M. Grice, P. Piwek, M. Schroeder, M. Klesen, S. Baumann, H. Pirker, K. van Deemter, and E. Gstrein. Generation of multi-modal dialogue for net environments. In *Proceedings of KONVENS-02*, pages 91–98, Saarbruecken, Germany, 2002.

B. Krenn, B. Neumayr, E. Gstrein, and M. Grice. Life-like agents for the internet: A cross-cultural case study. In S. Payr S and R. Trappl, editors, *Agent Culture: Designing Human-Agent Interaction in a Multicultural World*. Lawrence Erlbaum Associates, NJ, to appear.

M. Maratsos. Nonegocentric communication abilities in preschool children. *Child Development*, 44:697–700, 1973.

J. Piaget and B. Inhelder. *The child's conception of space*. Routledge and Kegan Paul, London, 1956.

P. Piwek. The effect of gestures on the perception of a dialogue between two embodied conversational agents: a pilot study. Technical Report ITRI-03-09, ITRI, University of Brighton, UK, 2003a.

P. Piwek. A flexible pragmatics-driven language generator for animated agents. In *Proceedings of EACL (Research Notes)*, Budapest, Hungary, 2003b.

P. Piwek and R.J. Beun. Multimodal referential acts in a dialogue game: From empirical investigations to algorithms. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue (IPNMD-2001)*, Verona, Italy, December 2001.

K. Porayska-Pomsta and C. Mellish. Modelling politeness in natural language generation. In Anja Belz, Roger Evans, and Paul Piwek, editors, *Natural Language Generation: Third International Conference (INLG 2004)*, volume 3123, pages 141–150, Berlin, 2004. Springer.

E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, 2000.

Zs. Ruttkay and C. Pelachaud, editors. *From Brows to Trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers, Dordrecht, 2004.

M.A. Walker, J.E. Cahn, and S.J. Whittaker. Linguistic style improvisation for lifelike computer characters. In *Proceedings of the AAAI Workshop on AI, ALife and Entertainment*, Portland, 1996.