Comparing Rule-based and Data-driven Selection of Facial Displays

Mary Ellen Foster
Informatik VI: Robotics and Embedded Systems
Technische Universität München
Boltzmannstraße 3, 85748 Garching, Germany
foster@in.tum.de

Abstract

The non-verbal behaviour of an embodied conversational agent is normally based on recorded human behaviour. There are two main ways that the mapping from human behaviour to agent behaviour has been implemented. In some systems, human behaviour is analysed, and then rules for the agent are created based on the results of that analysis; in others, the recorded behaviour is used directly as a resource for decision-making, using data-driven techniques. In this paper, we implement both of these methods for selecting the conversational facial displays of an animated talking head and compare them in two user evaluations. In the first study, participants were asked for subjective preferences: they tended to prefer the output of the data-driven strategy, but this trend was not statistically significant. In the second study, the data-driven facial displays affected the ability of users to perceive user-model tailoring in synthesised speech, while the rule-based displays did not have any effect.

1 Introduction

There is no longer any question that the production of language and its accompanying non-verbal behaviour are tightly linked (e.g., Bavelas and Chovil, 2000). The communicative functions of body language listed by Bickmore and Cassell (2005) include conversation initiation and termination, turn-taking and interruption, content elaboration and emphasis, and feedback and error correction; non-verbal behaviours that can achieve these functions include gaze modification, facial expressions, hand gestures, and posture shifts, among others.

When choosing non-verbal behaviours to accompany the speech of an embodied conversational agent (ECA), it is necessary to translate general findings from observing human behaviour into concrete selection strategies. There are two main implementation techniques that have been used for making this decision. In some systems, recorded behaviours are analysed and rules are created by hand based on the analysis; in others, recorded human data is used directly in the decision process. The former technique is similar to the classic role of corpora in natural-language generation described by Reiter and Dale (2000), while the latter is more similar to the more recent data-driven techniques that have been adopted (Belz and Varges, 2005).

Researchers that have used rule-based techniques to create embodied-agent systems include: Poggi and Pelachaud (2000), who concentrated on generating appropriate affective facial displays based on descriptions of typical facial expressions of emotion; Cassell et al. (2001a), who selected gestures and facial expressions to accompany text using heuristics derived from studies of typical North American non-verbal-displays; and Marsi and van Rooden (2007), who generated typical certain and uncertain facial displays for a talking head in an information-retrieval system. Researchers that used data-driven techniques include: Stone et al. (2004), who captured the motions of an actor performing scripted output and then used that data to create performance

Proceedings of the Workshop on Embodied Language Processing, pages 1–8,
Prague, Czech Republic, June 28, 2007.

specifications on the fly; Cassell et al. (2001b), who selected posture shifts for an embodied agent based on recorded human behaviour; and Kipp (2004), who annotated the gesturing behaviour of skilled public speakers and derived "gesture profiles" to use in the generation process.

Using rules derived from the data can produce displays that are easily identifiable and is straightforward to implement. On the other hand, making direct use of the data can produce output that is more similar to actual human behaviour by incorporating naturalistic variation, although it generally requires a more complex selection algorithm. In this paper, we investigate the relative utility of the two implementation strategies for a particular decision: selecting the conversational facial displays of an animated talking head. We use two methods for comparison: gathering users' subjective preferences, and measuring the impact of both selection strategies on users' ability to perceive user tailoring in speech.

In Section 2, we first describe how we recorded and annotated a corpus of facial displays in the domain of the target generation system. Section 3 then presents the two strategies that were implemented to select facial displays based on this corpus: one using a simple rule derived from the most characteristic behaviours in the corpus, and one that made a weighted choice among all of the options found in the corpus for each context. The next sections describe two user studies comparing these strategies: in Section 4, we compare users' subjective preferences, while in Section 5 we measure the impact of each strategy on user's ability to select spoken descriptions correctly tailored to a given set of user preferences. Finally, in Section 6, we discuss the results of these two studies, draw some conclusions, and outline potential future work.

2 Corpus collection and annotation[1]

The recording scripts for the corpus were created by the output planner of the COMIC multimodal dialogue system (Foster et al., 2005) and consisted of a total of 444 sentences describing and comparing various tile-design options. The surface form of each sentence was created by the OpenCCG surface realiser (White, 2006), using a grammar that specified both the words and the intended prosody for the speech synthesiser. We attached all of the relevant contextual, syntactic, and prosodic information to each node in the OpenCCG derivation tree, including the user-model evaluation of the object being described (positive, negative, or neutral), the predicted pitch accent, the clause of the sentence (first, second, or only), and whether the information being presented was new to the discourse.

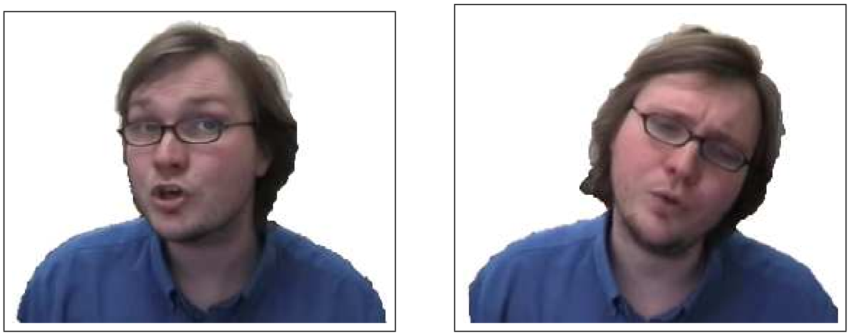
The sentences in the script were presented one at a time to a speaker who was instructed to read each out loud as expressively as possible into a camera directed at his face. The following facial displays were then annotated on the recordings: eyebrow motions (up or down), eye squinting, and rigid head motion on all three axes (nodding, leaning, and turning). Each of these displays was attached to the node or nodes in the OpenCCG derivation tree that exactly covered the span of words temporally associated with the display. Two coders separately processed the sentences in the corpus. Using a version of the β weighted agreement measure proposed by Artstein and Poesio (2005)—which allows for a range of agreement levels—the agreement on the sentences processed by both coders was 0.561.

When the distribution of facial displays in the corpus was analysed, it was found that the single biggest influence on the speaker's behaviour was the user-model evaluation of the features being described. When he described features of the design that had positive user-model evaluations, he was more likely to turn to the right and to raise his eyebrows (Figure 1(a)); on the other hand, on features with negative user-model evaluations, he was more likely to lean to the left, lower his eyebrows, and squint his eyes (Figure 1(b)). The overall most frequent display in all contexts was a downward nod on its own. Other factors that had a significant effect on the facial displays included the predicted pitch accent, the clause of the sentence (first or second), and the number of words spanned by a node.

3 Selection strategies

Based on the recorded behaviour of the speaker, we implemented two different methods for selecting facial displays to accompany synthesised speech. Both methods begin with the OpenCCG derivation

[1] Foster (2007) gives more details of the face-display corpus.



(a) Positive (b) Negative

Figure 1: Characteristic facial displays from the corpus

	Although	*it's*	*in*	*the*	*family*	*style,*	*the*	*tiles*	*are*	*by*	*Alessi.*
Original	nd=d	nd=d	nd=d		nd=d				nd=d,bw=u		
	.... ln=l										
Data-driven	nd=d				nd=d		. . tn=r . .				
Rule-based					ln=l,bw=d,sq						tn=r,bw=u

Figure 2: Face-display schedules for a sample sentence

tree for a sentence—that is, a tree in the same format as those that were used for the corpus annotation, including all of the contextual features. They then proceed top-down through the derivation tree, considering each node in turn and determining the display combination (if any) to accompany it.

The rule-based strategy specifies motions only on nodes corresponding to mentions of specific properties of a tile design: manufacturer and series names, colours, and decorations. The display combination is determined by the user-model evaluation of the property being described, based on the behaviours of the recorded speaker. For a positive evaluation, this strategy selects a right turn and brow raise; for a negative evaluation, it selects a left turn, brow lower, and eye squint; while for neutral evaluations, it chooses a downward nod.

In contrast, the data-driven strategy considers all nodes in the derivation tree. For each node, it selects from all of the display combinations that occurred on similar nodes in the corpus, weighted by the frequency. As a concrete example, in a hypothetical context where the speaker made no motion 80% of the time, nodded 15% of the time, and turned to the right in the other 5%, this strategy would select no motion with probability 0.8, a nod with probability 0.15, and a right turn with probability 0.05.

Figure 2 shows a sample sentence from the corpus, the original facial displays, and the displays selected by each of the strategies. In the figure, *nd=d* indicates a downward nod, *bw=u* and *bw=d* a brow raise and lower, respectively, *sq* an eye squint, *ln=l* a left lean, and *tn=r* a right turn.

4 Subjective preferences

As a first comparison of the two implementation strategies, we gathered users' subjective preferences between three different types of face-display schedules: the displays selected by each of the generation strategies described in the preceding section, as well as the original displays annotated in the corpus.

4.1 Participants

This experiment was run through the Language Experiments Portal,[2] a website dedicated to online psycholinguistic experiments. There were a total of 36 participants: 20 females and 16 males. 23 of the participants were between 20 and 29 years old, 9 were over 30, and 4 were under 20. 21 described themselves as expert computer users, 14 as intermediate users, and one as a beginner. 18 were native speakers of English, while the others had a range of other native languages.

[2] http://www.language-experiments.org/

Figure 3: RUTH talking head

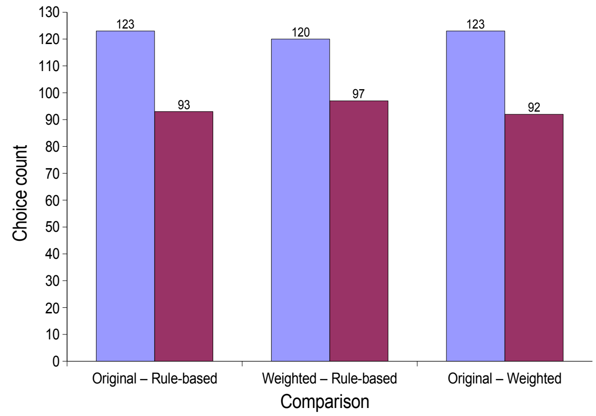


Figure 4: Subjective-preference results

4.2 Methodology

Each participant saw videos of two possible synthesised face-display schedules accompanying a series of 18 sentences. Both videos had the same synthesised speech, but each had a different different facial-display schedule. For each pair, the participant was asked to select which of the two versions they preferred. There were three different schedule types: the original displays annotated in the corpus, along with the output of both of the selection strategies. Participants made each pairwise comparison between these types six times, three times in each order. All participants saw the same set of sentences, in a random order: the pairwise choices were also allocated to sentences randomly.

4.3 Materials

To create the materials for this experiment, we randomly selected 18 sentences from the corpus and generated facial displays for each, using both of the strategies. The data-driven schedules were generated through 10-fold cross-validation as part of a previous study (Foster and Oberlander, 2007): that is, the display counts from 90% of the corpus were used to select the displays to use for the sentences in the held-out 10%. The rule-based schedules were generated by running the rule-based procedure from Section 3 on the same OpenCCG derivation trees. Videos were then created of all of the schedules for all of the sentences, using the Festival speech synthesiser (Clark et al., 2004) and the RUTH animated talking head (DeCarlo et al., 2004) (Figure 3).

4.4 Results

The overall results of this study are shown in Figure 4. Not all participants responded to all items, so there were a total of 648 responses: 216 comparing the original corpus schedules to the rule-based schedules, 217 for the data-driven vs. rule-based comparison, and 215 for the original vs. data-driven comparison. To assess the significance of the preferences, we use a binomial test, which provides an exact measure of the statistical significance of deviations from a theoretically expected classification into two categories. This test indicates that there was a mildly significant preference for the original schedules over the output of each of the strategies ($p < 0.05$ in both cases). While there was also a tendency to prefer the output of the data-driven strategy over that of the rule-based strategy, this preference was not significant ($p \approx 0.14$). No demographic factor had a significant effect on these results.

4.5 Discussion

Although there was no significant preference between the output of the two strategies, the generated schedules were very different. The rule-based strategy used only the three display combinations described in Section 3 and selected an average of 1.78 displays per sentence on the 18 sentences used in this study, while the data-driven strategy selected 12 different display combinations across the sentences and chose an average of 5.06 displays per sentence. For comparison, the original sentences from the corpus used a total of 15 different combinations on the

(1) Here is a family design. Its tiles are from the Lollipop collection by Agrob Buchtal. Although the tiles have a blue colour scheme, it does also feature green.

(2) Here is a family design. As you can see, the tiles have a blue and green colour scheme. It has floral motifs and artwork on the decorative tiles.

Figure 5: Tile-design description tailored to two user models (conflicting concession highlighted)

same sentences and had an average of 4.83 displays per sentence. In other words, in terms of the range of displays, the schedules generated by the data-driven strategy are fairly similar to those in the corpus, while those from the rule-based strategy do not resemble the corpus very much at all.

In another study (Foster and Oberlander, 2007), the weighted data-driven strategy used here was compared to a majority strategy that always chose the highest-probability option in every context. In other words, in the hypothetical context mentioned earlier where the top option occurred 80% of the time, the majority strategy would always choose that option. This strategy scored highly on an automated cross-validation study; however, human judges very strongly preferred the output of the weighted strategy described in this paper ($p < 0.0001$). This contrasts with the weak preference for the weighted strategy over the rule-based strategy in the current experiment. The main difference between the output of the majority strategy on the one hand, and that of the two strategies described here on the other, is in the distribution of the face-display combinations: over 90% of the that the majority strategy selected a display, it used a downward nod on its own, while both of the other strategies tended to generate a more even distribution of displays across the sentences. This suggests that the distribution of facial displays is more important than strict corpus similarity for determining subjective preferences.

The participants in this study generally preferred the original corpus displays to the output of either of the generation strategies. This suggests that a more sophisticated data-driven implementation that reproduces the corpus data more faithfully could be successful. For example, the process of selecting facial displays could be integrated directly into the OpenCCG realiser's *n*-gram-guided search for a good realisation (White, 2006), rather than being run on the output of the realiser as was done here.

5 Perception of user tailoring in speech

The results of the preceding experiment indicate that participants mildly preferred the output of the data-driven strategy to that of the rule-based strategy; however, this preference was not statistically significant. In this second experiment, we compare the face-display schedules generated by both strategies in a different way: measuring the impact of each schedule type on users' ability to detect user-model tailoring in synthesised speech.

Foster and White (2005) performed an experiment in which participants were shown a series of pairs of COMIC outputs (e.g., Figure 5) and asked to choose which was correctly tailored to a given set of user preferences. The participants in that study were able to select the correctly-tailored output only on trials where one option contained a concession to a negative preference that the other did not. For example, the description in (1) contains the concession *Although the tiles have a blue colour scheme*, as if the user disliked the colour blue, while (2) has no such concession. Figure 6 shows the results from that study when outputs were presented as speech; the results for text were nearly identical. The first pair of bars represent the choices made on trials where there was a conflicting concession, while the second pair show the choices made on trials with no conflicting concession. Using a binomial test, the difference for the conflicting-concession trials is significant at $p < 0.0001$, while there is no significant difference for the other trials ($p \approx 0.4$).

In this experiment, use the same experimental materials, but we use the talking head to present the system turns. This experiment allows us to answer two questions: whether the addition of a talking head affects users' ability to perceive tailoring in speech, and whether there is a difference between the impact of the two selection strategies.

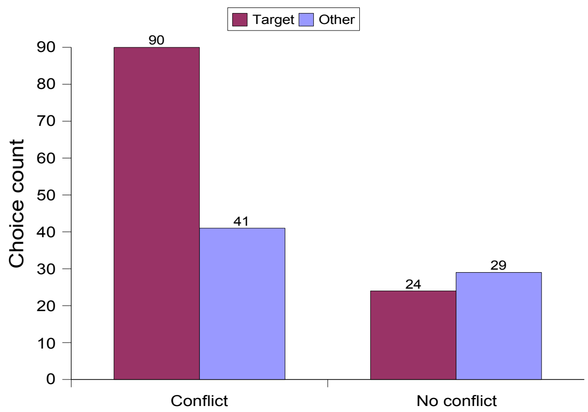


Figure 6: Results for speech-only presentation

5.1 Participants

Like the previous study, this one was also run over the web. There were 32 participants: 19 females and 13 males. 18 of the participants were between 20 and 29 years old, 10 were over 30, and 4 were under 20. 15 described themselves as expert computer users, 15 as intermediate users, and 2 as beginners. 30 of the participants were native English speakers.

5.2 Methodology

Participants in this experiment observed an eight-turn dialogue between the system and a user with specific likes and dislikes. The user preferences were displayed on screen at all times; the user input was presented as written text on the screen, while the system outputs were played as RUTH videos in response to the user clicking on a button. There were two versions of each system turn, one tailored to the preferences of the given user and one to the preferences of another user; the user task was to select the correctly tailored version. The order of presentation was counterbalanced so that the correctly tailored version was the first option in four of the trials and the second in the other four. Participants were assigned in rotation to one of four randomly-generated user models. As an additional factor, half of the participants saw videos with facial displays generated by the data-driven strategy, while the other half saw videos generated by the rule-based strategy.

5.3 Materials

The user models and dialogues were identical to those used by Foster and White (2005). For each sentence in each system turn, we annotated the nodes of the OpenCCG derivation tree with all of the necessary information for generation: the user-model evaluation, the pitch accents, the clause of the sentence, and the surface string. We then used those annotated trees to create face-display schedules using both of the selection strategies, using the full corpus as context for the data-driven strategy, and prepared RUTH videos of all of the generated schedules as in the previous study.

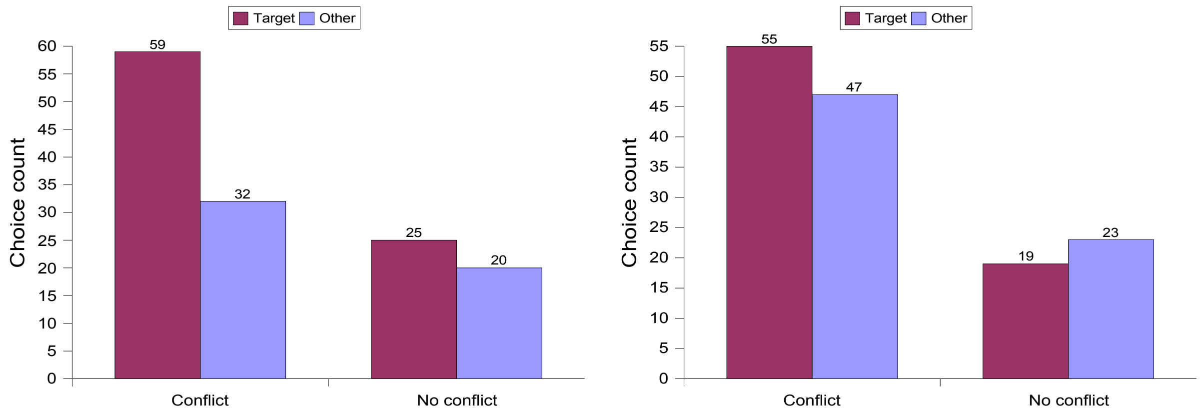
5.4 Results

The results of this study are shown in Figure 7: Figure 7(a) shows the results for the participants using the rule-based schedules, while Figure 7(b) shows the results with the data-driven schedules. Just as in the speech-only condition, the participants in this experiment responded essentially at chance on trials where there was no conflicting concession to negative preferences. For the trials with a conflicting concession, participants using rule-based videos selected the targeted version significantly more often ($p < 0.01$), while the results for participants using the data-driven videos show no significant trend ($p \approx 0.49$). None of the demographic factors affected these results.

To assess the significance of the difference between the two selection strategies, we compared the results on the conflicting-concession trials from each of the groups to the corresponding results from the speech-only experiment, using a χ^2 test. The results for the judges using the rule-based videos are very similar to those of the judges using only speech ($\chi^2 = 0.21$, $p = 0.65$). However, there is a significant difference between the responses of the speech-only judges and those of the judges using the weighted schedules ($\chi^2 = 4.72$, $p < 0.05$).

5.5 Discussion

The materials for this study were identical to those used by Foster and White (2005); in fact, the waveforms for the synthesised speech were identical. However, the participants in this study who saw the videos generated by the data-driven strategy



(a) Rule-based schedules

(b) Data-driven schedules

Figure 7: Results of the perception study

were significantly worse at identifying the correctly-tailored speech than were the participants in the previous study, while the performance of the participants who saw rule-based videos was essentially identical to that of the speech-only subjects.

The schedules selected by the data-driven strategy for this evaluation include a variety of facial displays; sometimes these displays are actually the opposite of what would be selected by the rule-based strategy. For example, the head moves to the right when describing a negative fact in 23 of the 520 data-driven schedules, and moves to the left when describing a neutral or positive fact in 20 cases. A description includes up to three sentences, and a trial involved comparing two descriptions, so a total of 75 of the trials (52%) for the data-driven participants involved at least one of these these potentially misleading head movements. Across all of the trials for the participants using data-driven videos, there were 38 conflicting-concession trials with no such head movement. The performance on these trials was essentially the identical to that on the full set of trials: the correctly targeted description was chosen 20 times, and the other version 18 times. So the worse performance with the data-driven schedules cannot be attributed solely to the selected facial displays conflicting with the linguistic content.

Another possibility is that the study participants who used the data-driven schedules were distracted by the expressive motions of the talking head and failed to pay attention to the content of the speech. This appears to have been the case in the COMIC whole-system evaluation (White et al., 2005), for example, where the performance of the male participants on a recall task was significantly worse when a more expressive talking head was used. On this study, there was no effect of gender (or any of the other demographic factors) on the pattern of responses; however, it could be that a similar effect occurred in this study for all of the participants.

6 Conclusions and future work

The experiments in this paper have compared the two main current implementation techniques for choosing non-verbal behaviour for an embodied conversational agent: using rules derived from the study of human behaviour, and using recorded human behaviour directly in the generation process. The results of the subjective-preference evaluation indicate that participants tended to prefer the output generated by the data-driven strategy, although this preference was not significant. In the second study, videos generated by the data-driven strategy significantly decreased participants' ability to detect correctly-tailored spoken output when compared to a speech-only presentation; on the other hand, videos generated by the rule-based strategy did not have a significant impact on this task.

These results indicate that, at least for this corpus and this generation task, the choice of generation strategy depends largely on which aspect of the system is more important: to create an agent that users like subjectively, or to ensure that users fully understand all aspects of the output presented in speech. If the former is more important, than an implementation that uses the data directly appears to be a slightly better option; if the latter is more important, then the rule-based strategy seems superior.

On the subjective-preference evaluation, users preferred the original corpus motions over either of the generated versions. As discussed in Section 4.5, this suggests that there is room for a more sophisticated data-driven selection strategy that reproduces the corpus data more closely. The output of such a generation strategy might also have a different effect on the perception task.

Both of these studies used the RUTH talking head (Figure 3), which has no body and, while human in appearance, is not particularly realistic. We used this head to investigate the the generation of a limited set of facial displays, based on contextual information including the user-model evaluation, the predicted prosody, the clause of the sentence, and the surface string. More information about the relative utility of different techniques for selecting non-verbal behaviour for embodied agents can be gathered by experimenting with a wider range of agents and of non-verbal behaviours. Other possible agent types include photorealistic animated agents, agents with fully articulated virtual bodies, and physically embodied robot agents. The possibilities for non-verbal behaviours include deictic, iconic, and beat gestures, body posture, gaze behaviour, and facial expressions of various types of affect, while any source of syntactic or pragmatic context could be used to help make the selection. Experimenting with other combinations of agent properties and behaviours can improve our knowledge of the relative utility of different mechanisms for selecting non-verbal behaviour.

References

R. Artstein and M. Poesio. 2005. Kappa^3 = alpha (or beta). Technical Report CSM-437, University of Essex Department of Computer Science.

J. B. Bavelas and N. Chovil. 2000. Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology*, 19(2):163–194. doi:10.1177/0261927X00019002001.

A. Belz and S. Varges, editors. 2005. *Corpus Linguistics 2005 Workshop on Using Corpora for Natural Language Generation*. http://www.itri.brighton.ac.uk/ucnlg/ucnlg05/.

T. Bickmore and J. Cassell. 2005. Social dialogue with embodied conversational agents. In J. van Kuppevelt, L. Dybkjær, and N. Bernsen, editors, *Advances in Natural, Multimodal Dialogue Systems*. Kluwer, New York. doi:10.1007/1-4020-3933-6_2.

J. Cassell, T. Bickmore, H. Vilhjálmsson, and H. Yan. 2001a. More than just a pretty face: Conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14(1–2):55–64. doi:10.1016/S0950-7051(00)00102-7.

J. Cassell, Y. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich. 2001b. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*. ACL Anthology P01-1016.

R. A. J. Clark, K. Richmond, and S. King. 2004. Festival 2 – build your own general purpose unit selection speech synthesiser. In *Proceedings of the 5th ISCA Workshop on Speech Synthesis*.

D. DeCarlo, M. Stone, C. Revilla, and J. Venditti. 2004. Specifying and animating facial signals for discourse in embodied conversational agents. *Computer Animation and Virtual Worlds*, 15(1):27–38. doi:10.1002/cav.5.

M. E. Foster. 2007. Associating facial displays with syntactic constituents for generation. In *Proceedings of the ACL 2007 Workshop on Linguistic Annotation (The LAW)*.

M. E. Foster and J. Oberlander. 2007. Corpus-based generation of conversational facial displays. In submission.

M. E. Foster and M. White. 2005. Assessing the impact of adaptive generation in the COMIC multimodal dialogue system. In *Proceedings of the IJCAI 2005 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

M. E. Foster, M. White, A. Setzer, and R. Catizone. 2005. Multimodal generation in the COMIC dialogue system. In *Proceedings of the ACL 2005 Demo Session*. ACL Anthology W06-1403.

M. Kipp. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Dissertation.com.

E. Marsi and F. van Rooden. 2007. Expressing uncertainty with a talking head. In *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007)*.

I. Poggi and C. Pelachaud. 2000. Performative facial expressions in animated faces. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 154–188. MIT Press.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press. doi:10.2277/052102451X.

M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Lees, A. Stere, and C. Bregler. 2004. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics (TOG)*, 23(3):506–513. doi:10.1145/1015706.1015753.

M. White. 2006. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75. doi:10.1007/s11168-006-9010-2.

M. White, M. E. Foster, J. Oberlander, and A. Brown. 2005. Using facial feedback to enhance turn-taking in a multimodal dialogue system. In *Proceedings of HCI International 2005*.