Towards Emotional Variation in Speech-Based Natural Language Generation

Michael Fleischman and Eduard Hovy

USC Information Science Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
U.S.A.
{fleisch, hovy} @ISI.edu

Abstract

We present a framework for handling emotional variations in a speech-based natural language generation system for use in the MRE virtual training environment. The system is a first step toward addressing issues in emotion-based modeling of verbal communicative behavior. We cast the problem of emotion-based generation as a distance minimization task, in which the system chooses between multiple valid realizations for a given input based on the emotional distance of each realization from the speaker's attitude toward that input. We discuss evaluations of the system and future work that includes modeling personality and empathy within the same framework.

1. Introduction: the MRE

Emotion is an ever-present characteristic of human experience and behavior. As fundamental to the human condition as cognition, emotion has begun to pique the interest of those researchers in the Artificial Intelligence community concerned with simulating human behavior in embodied agents. Nowhere is this interest more prominent than in the domain of multi-modal, virtual training environments. In such environments, realistic modeling of emotion enhances the user's ability to suspend disbelief (Marsella & Gratch, 2001), and can be used as an additional parameter in creating more variable training scenarios.

The emotional NLG system that we present is designed within the Mission Rehearsal Exercise (MRE) virtual training environment (Swartout et al. 2001). The MRE is a large-scale collaborative research effort to develop a fully interactive training simulation modeled after the holodeck in Star Trek. The project brings together researchers working on graphics, 3-D audio, artificial intelligence, and Hollywood screenwriters to create a realistic virtual world in which human subjects can interact naturally with simulated agents. The agents are modeled using the Steve system of Rickel and Johnson (1999). They communicate through voice and gesture, reason about tasks and actions, and incorporate a complex model of their own emotions, as well as the emotional states of the other agents in their environment (Gratch & Marsella, 2001; Gratch, 2000). Users can query and interact with one (and eventually many) agent in real-time as they proceed through a scenario developed for the particular training mission at hand.

The scenario presently implemented is designed to train army lieutenants for eastern European peace keeping missions. The scenario centers around the trainee, a human lieutenant, who is attempting to move his platoon to a support position, when one of his drivers unexpectedly collides with a civilian car. A civilian passenger, a young boy, is critically injured and the boy's mother, as well as a crowd of local onlookers, is becoming increasingly agitated. The trainee must interact with his or her virtual platoon sergeant in order to stabilize the situation.

The MRE represents the integration of many fields in NLP. As the trainee interacts with the virtual agents in the environment, automatic speech recognition translates the user's speech into a text string that is passed to the natural

language understanding module. This module uses a finite state machine to convert the string into a case frame structure that is passed to a dialogue manager. At this point, the dialogue manager interacts with the task planner, the action selector, and the emotion model to initiate a particular response. The content of this response is then passed as an impoverished case frame to the NLG system. Generation converts the input into a tree structure that contains both syntactic and semantic information. The tree is then passed to a gesture module and is tagged with non-verbal information to control gaze and body movements. Finally, the tree is flattened, the gestures and visemes are synched using the BEAT system (Cassell, 2001), and the speech is synthesized.

2. Previous Work

While much attention has been paid to the effect of emotion on planning and non-verbal behavior (Marsella et al., 2001; Cahn, 1990), little work has been done on the effects of emotion on the verbal behavior of embodied agents. Most previous work focuses on intonation and non-verbal communication.

With respect to content and phrasing, the most relevant work is over 10 years old. In his thesis, Hovy (1988) implemented a 3-valued (positive, negative, neutral) system of emotional shades with a simple sign multiplication calculus to control affect laden text generation. The three values provided little flexibility to accommodate the more subtle nuances associated with different shades of affect.

Work by Bateman and Paris (1989) and Paris (1988) focus on variations of expert system output based on the reader's knowledge. Also here, the rules for combining ratings of sentence constituents was fairly simple and not easily extensible. Papers by Walker et al (1996) and Loyall and Bates (1997) explore aspects of style and emotion, but do not focus on the particulars of natural language generation.

In this paper we describe an integrated framework for modeling emotion in the speech-based natural language generation of embodied agents. It incorporates a distance calculus that adds flexibility and allows us to extend the emotional input from simple like/dislike to more complicated constructs.

3. NLG in MRE

Generation in the MRE is a hybrid process. The generator can take as input both highly elaborated case frames, for scenario specific utterances, and more impoverished frames, for use in interactive conversation. We discuss only the conversational aspect of the system.

The generator is, at this point, highly domain dependent, but has sufficient coverage to generate utterances for every task in the agents' task models. The generator is implemented in the SOAR programming language (Newell, 1990) and takes place in three stages: sentence planning, realization and ranking.

3.1 Sentence Planning

As seen in Figure 1(a), the inputs to this stage are received from the dialogue manager. These inputs contain minimal information about the state or event to be described, along with references to the actors and objects involved. A set of SOAR production rules converts this information into an enriched case frame structure, seen in Figure 1(b), which contains more detailed information about the events and objects in the input. The conversion process, which involves choosing the appropriate object case frames, relies heavily on the emotional decision engine.

3.2 Realization

Realization is a highly lexicalized procedure, and tree construction begins with the selection of main verbs (more on this below). Each verb in the lexicon carries with it slots for its constituents (e.g., agent, patient), which form branches in the tree. Once the verb is chosen, production rules recursively expand the nodes in the tree until no more nodes can be expanded. As each production rule fires, the relevant portion of the semantic frame is propagated down into the expanded nodes. Thus, every node in the tree contains a pointer to the specific aspect of the semantic frame from

```
^event collision
^time past
^speech-act assert
^agent driver
^patient mother
```

Figure 1a. Input from dialogue manager: input to sentence planning phase of generation

```
(<utterance>
        ^type assertion
        ^content <event>)
(<event>
        ^type event
        ^time past
        ^name collision
        ^agent <agent>
        ^patient <patient>)
(<agent>
        ^type agent
        ^name driver
        ^definite true
        ^singular true)
(<patient>
        ^type patient
        ^name mother
        ^definite true
        ^singular true)
```

Figure 1b. Expansion of input from dialogue manager; output of sentence planning

which it was created. For example, in Figure 1(c), the NP node of "the mother" contains in it a pointer to the frame <patient> from Figure 1(b). By keeping semantic content localized in the tree, we allow the gesture and speech synthesis modules convenient access to needed semantic information. This strategy is particularly convenient in a setting such as the MRE, where modules require increasing amounts of information as research continues.

For any given state and event, there are a number of theoretically valid realizations available in the lexicon. Instead of attempting to decide which is most appropriate at any stage, we adopt a strategy similar to that introduced by (Knight & Hatzivassiloglou, 1995), which puts off the decision until realization is complete. We realize all possible valid trees that correspond to a given semantic input, and store the fully constructed trees in a forest structure. After all such trees are constructed we move on to the final stage.

3.3 Ranking

In this stage we examine all the trees in the forest structure and decide which tree will be propagated further down the NLP pipeline. Each tree is given a rank score based upon the tree's information content and emotional quality. The score of each tree is calculated by recursively summing the scores of the nodes along the frontiers of the tree, and then percolating that sum up to the next layer. Summing and percolating proceeds until the root node is given a score that is equivalent to the sum of the scores for the individual nodes of that tree. The tree with the highest root node score is selected.

4. Emotional Variations

We cast the problem of emotional language generation as an optimization problem in which multiple acceptable realizations of a given semantic frame are produced. Given a set of valid realizations for a given frame, we output the sentence that most closely fits the emotional state of the speaker.

4.1 Speaker's Emotions

The emotion model employed by the MRE is based largely on various appraisal theories of emotion (Ortony, Clore, and Collins, 1988; Lazarus, 1991). Such models use the term appraisal to refer to the emotional evaluation of events. The MRE concretizes this notion of evaluation in terms of data structures, called construal frames[1], which represent relations between events and the dispositions of agents. In the MRE, dispositions are defined entirely in terms of an agent's plans and goals (Gratch, 2000). Thus, an agent's emotional state is predicated entirely on that agent's appraisal of an event in terms of how that event relates to its own set of goals, and plans toward those goals.

Conversely, the model allows us to describe events in the world in terms of their relationship to an agent's dispositions. Thus, each object and event in the agents' world model can be described as a vector of features relating that event to the agent's goals and plans. The features that describe these elements of the world model are relations such as whether the

[1] These are derived from the construal theory of Clark Elliot.

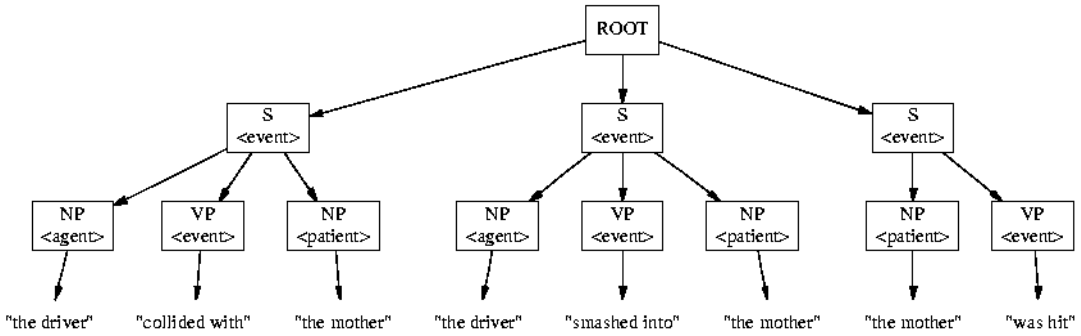

Figure 1c. A subset of the forest output of realization.

element represents a constraint on the agent's plan, to what extent the element leads to the achievement of a goal, and whether and how the element may affect the goals of others. This feature vector becomes arbitrarily long as the emotion model becomes more complex. For the purposes of this research, however, we choose to simplify the representation and describe each element of the world model by a single aggregate feature, namely, the attitude of the speaker toward the element. While this single feature representation is admittedly overly simplistic, it is useful pedagogically and, as will be discussed later, can be naturally extended to larger feature vectors.

We represent the emotional state of the speaker toward an element of the world model as an integer value (ranging from –5 to +5). Each value corresponds to the speaker's attitude toward a specific element of the input. For example, Figure 2(a) depicts an input describing an event (collision) with an agent (driver) and a patient (mother). Each element is further described by an emotional attitude representing how positively or negatively the speaker feels toward the element (agent: +4, patient: +1, event: -1). These values are calculated by the emotion model and passed as inputs to the generator, along with the semantic input, by the dialogue manager.

4.2 Emotional Distance

Previous emotion-based generation, such as PAULINE (Hovy, 1988) had trouble integrating multiple different emotion values into a single utterance. We therefore developed the following procedure. We calculate the fit of a sentence to the emotional state of the speaker as the distance between the speaker's emotional attitude toward an object and the default emotional shade of the lexical item or expression used to express that object. While the emotional attitudes of the speaker are given by the emotion module, the default emotional shades for the lexical items are stored in the lexicon.

Deciding what default value shade each lexical item is given is, at this point, a matter of linguistic intuition. However, empirical alternatives are discussed in later sections.

In order to avoid the memory explosion that comes with calculating distances for every possible valid sentence that represents a frame, we divide the task between two stages of generation: planning and ranking.

```
^time past
^speech-act assert
^event
  :reference collision
  :attitude  -1
^agent
  :reference driver
  :attitude  +4
^patient
  :reference mother
  :attitude  +1
```

Figure 2a. Input from dialogue manager showing speaker's attitudes toward objects and events

During planning, the impoverished input given by the dialogue manager is expanded into a semantic frame ready for realization. The task of expansion involves deciding which frame is to be chosen to represent each object in the input. For example, Figure 2(b), shows a number of possible frames that could be used to represent the agent "driver." The decision is based on the emotional shade of each semantic option. A distance is calculated between the

shade of each semantic frame representing the driver and the emotional attitude of the speaker toward the driver. The frame with the minimum distance is chosen for expansion. This is done for each of the objects associated with the event or state. Once all objects have been assigned a frame, planning is complete, and realization begins.

Lopez
(<agent>
^type agent
^name driver
^proper true
^singular true
^shade +5)
The driver
(<agent>
^type agent
^name driver
^definite true
^singular true
^shade 0)
A private
(<agent>
^type agent
^name private
^definite false
^singular true
^shade -2)

Figure 2b. Subset of possible case frame expansions for object "driver"; ^shade value represents emotional shade of using that frame to refer to "driver".

Gloss	Agent Shade	Event Shade	Patient Shade
A bumped into P	0	-1	0
A collided with P	-1	-1	0
A ran into P	-2	-2	0
A hit P	-3	-2	+1
P was hit by A	-2	-2	+1
P was hit	Na	-2	+1
A crashed into P	-3	-3	+1
A smashed into P	-5	-4	+2
There was an accident	Na	-2	Na

Table 1. Valid lexical representations for event "collision" including the shades that the verbs apply to the objects related to the event.

During realization, semantic frames are expanded as described in Section 3.2. In this phase, all verbs in the lexicon that are valid representations of the input frame are used to create distinct trees. Each verb carries with it its emotional shade. This shade is expressed in two ways: by the overall emotional connotation of the verb itself, and by the emotional connotations that the verb imparts on its constituents. A sample of the lexicon for verbs that describe the event "collision" is shown in Table 1.

As seen in the entry, the verb "hit" casts a more negative emotional shade on the agent and event than the verb "smash." However, "smash" casts a more positive shade on the patient of the event than "hit." This effect is seen in the realizations: "The driver hit the mother's car" and "The driver smashed into the mother's car." While both verbs betray negativity toward the "driver," the latter is more severe than the former. Further, because of the intensity of the verb (and the negativity of the event), the patient is cast as more sympathetic in the latter sentence.

When the ranking phase begins, each tree formed of these verbs is ranked and compared, as described in Section 4.3. The tree finally selected is that in which the *total* emotional distance from the speaker's attitude is minimized across the event itself, as well as across all the constituents of that event. Thus, even if the speaker feels very negatively toward the event described in Figure 2(a), because the distances for each tree are summed across all of its constituents, the generator may still opt not to use the strong lexical item "smash" if the speaker has intensely positive feelings toward the agent.

4.3 Scoring and Ranking

Table 2 shows example calculations for three variations of the input given in Figure 2(a). The emotion scores for each variation are computed using the distance formulas below, where attitude(x) is the speaker's attitude toward x and shade(x) is the shade of the lexical item used to represent x. Though simple, this distance formula provides more intuitive results than various other obvious candidates.

$$\text{EmotScore}(x) = \text{Dist}(\text{verb}) + \sum_i \text{Dist}(\text{constituent})$$

$$\text{Dist}(x) = |\text{attitude}(x) - \text{shade}(x)|$$

This method of calculating emotional effect provides a great deal of variation with very little overhead. Once the lexicon is updated with items that carry emotional shadings, it is simply a matter of assigning the speaker attitudes, and

Event	Agent	Patient	Output
-2	-3	5	*A private crashed into the mother*
-2	-3	4	*A private hit the mother*
-2	-3	3	*A private hit the mother*
-2	-3	2	*The mother was hit by a private*
-2	-3	1	*The mother was hit by a private*
-2	-3	0	*A private ran into a woman*
-2	-3	-1	*A private ran into a civilian*
-2	-3	-2	*A private ran into a civilian*
-2	-3	-3	*A private ran into a civilian*
-2	-3	-4	*A private ran into a civilian*
-2	-3	-5	*A private ran into one of our "responsibilities"*

Figure 3a. Effect of varying the speaker's attitude toward the patient of an event; attitude toward the agent and the event itself are held constant.

applying a simple distance metric. The system will then automatically decide between possible realizations based not only on lexical choice, but also on sentence structure.

Verb	D(agent)	D(verb)	D(patient)	Score
was hit	**Na**	**|-2--1|=1**	**|1-1|=0**	**1**
collided	|-1-4|=5	|-1--1|=0	|0-1|=1	6
smashed	|-5-4|=9	|-4--1|=3	|2-1|=1	13

Table 2. Emotional scores for input shown in Figure 2(a). "was hit" obtains minimal distance score, and is selected.

As seen in Figure 3, the passive construction of the verb "hit" shades the elements of the event differently than the active construction of the same verb. Because the agent is not realized at all, the passive will be preferred when the attitude of the speaker is very positive toward the agent. This is because the event itself is such that it always shades the agent negatively. Thus, by not mentioning the agent at all, the speaker avoids having to say something negative about an object it regards positively. In extreme cases, the agent's attitude may even lead it to elide most of the sentence or to not speak at all.

However, the generator's need to convey information must be observed as well. We therefore compute a total rank score as a linear combination of the emotional distance and information content expressed by the tree:

$$\text{Total Score}(x) = \alpha\ \text{Info}(x) - (1-\alpha)\text{EmotScore}(x)$$

Here, the Info(x) is the number of slots from the input frame that are realized by x, and EmotScore(x) is as above. By changing the coefficient α, different weight will be given to the information content of the utterance versus its emotional shade. One can view an aspect of the personality of the speaker as a tendency toward a certain value for α: An agent who is more interested in the facts will always use a high α, while one who is more concerned with expressing emotion will use a low value.

5. Evaluation

In evaluating this system, we were particularly concerned with two points. First, how sensitive is the system to different inputs, and second, how well do the outputs actually mimic the emotional behavior of humans.

To determine the sensitivity of the system to different inputs we cycled through the parameters of the input space and observed the frequency of change in the output sentences. Because of the large number of possible inputs even for a simple frame such as in Figure 2(a) (i.e., the number of possible values raised to the power of the number of objects), we present results only for a subset of examples. Figure 3(a) shows the outputs of the generator when the attitude toward the patient is changed, holding all else constant; and Figure 3(b) shows the output when the attitude toward the agent is changed, holding all else constant. (Notice that the realization of the object being held constant does not change. This is because the frames that dictate the realizations are chosen at the sentence planning stage.

It is interesting to notice the difference in sensitivity between the two cases; changing the attitude toward the agent has more effect than changing the attitude toward the patient. This is because of the nature of the event "collision." As can be seen in Figure 3, the different realizations of the event vary mostly in their effect on the agent of the sentence. Thus, changing the attitude toward the patient has an

Event	Agent	Patient	Output
0	5	3	*A woman was hit*
0	4	3	*A woman was hit*
0	3	3	*Our driver bumped into a woman*
0	2	3	*Our driver bumped into a woman*
0	1	3	*One of our drivers bumped into a woman*
0	0	3	*The driver bumped into a woman*
0	-1	3	*The driver collided with a woman*
0	-2	3	*A driver ran into a woman*
0	-3	3	*One of our privates collided with a woman*
0	-4	3	*One of our privates ran into a woman*
0	-5	3	*A damn private collided with a woman*

Figure 3b. Effect of varying the speaker's attitude toward the agent of an event; attitude toward the patient and the event itself are held constant.

effect on the sentence only at the extremes of the range of attitudes. We conclude that using a distance measure as the basis for the emotion calculus is adequately sensitive.

Evaluating how a generated output correlates with human intuitions regarding the speaker's attitude is not an easy task. Judging the emotional state of someone based solely on their utterances is near impossible and presents many methodological challenges.

One way of determining such correlation is by having humans guess the attitudes of the system and comparing this to the system's emotional input. We asked subjects to rate the objects in the sentence on scales from 5 to –5 (where 5 means the speaker thinks most favorably about the object and –5 is most unfavorably). The correlation between what the subjects believed to be the attitudes of the speaker and the actual attitudes used for generation was statistically significant even with very few subjects ($r=0.659$, $n=10$), indicating that the expressiveness of the system is reliable.

An interesting prospect for future work is incorporating the procedure for evaluating the system into the system's actual design. We plan to examine the feasibility of using averages of human judgments as the shades for verbs in the lexicon. This is essentially the method that is employed now (using only the authors' intuitions), but by increasing the size of the sample, we suspect even more reliable outputs can be found.

6. Future Work

The system we present, while not complete, facilitates development on many fronts. The notion of a speaker's attitude toward an object or event, for example, while very simple in this implementation, can easily be expanded to fit the needs of the system. Because the decision method is a simple Euclidean distance metric, the single valued attitudes that we describe can easily be converted to the more complex vectors discussed above (Section 4.1). Once the lexicon is updated for the richer format, the distance metric need only be changed to operate on vectors.

Of further interest is the possibility of incorporating empathy into the generation process. In the current system, generation is based only on the emotions of the speaker. In the future, with more information from the emotion model, we will be able to generate sentences also taking into account the emotional attitudes of the hearers, by simply incorporating them into the distance calculations.

For example, in the MRE, when the agent is asked about the status of the boy lying, bleeding in the street, it knows that the boy is critically injured and that this information will upset the boy's mother. Taking this into account, empathic generation is blocked from saying: "the boy is dying," and chooses the more appropriate: "the boy needs a doctor, quick," instead.

Such empathy is easily implemented in our framework by replacing the vector of the speaker's attitudes with a linear combination of the speaker's attitudes and the hearers' attitudes. This treats empathic generation not as a decision for sentence planning, but rather as an alternation in realization. Such treatment is particularly intuitive if one takes the stance that, given the circumstances of the utterance, the content being conveyed to the human in the

above situation does not differ between utterances.

Under this formulation, the personality of the speaker can be partially described in terms of the weights with which one performs this combination (much like the α used in ranking, see section 4.3). For example, a speaker who is *sensitive* is just someone who tends to give higher weight to the attitudes of the hearers than to their own. On the other hand, an *indifferent* speaker would be one who ignores the attitudes of the hearers when generating an utterance.

We believe that this framework is a simple and convenient method for treating emotion in language. As virtual environments become more common, and the population of virtual characters in those environments explodes, the need for such emotional generation becomes more apparent. While our system is not complete, it is a simple and intuitive method for dealing with a necessary and under-explored area of natural language generation.

Acknowledgments

The authors would like to thank Manisha Joshi for her invaluable help with the system. We would also like to thank John Gratch and Stacy Marcella, whose own work on emotion has made this work possible. The project depicted is sponsored by the U.S. Army. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Bateman, J. and Paris, C.L. (1989). Phrasing a text in terms the user can understand. *Proc. of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, Michigan.

Cahn, J.E. (1990). The Generation of Affect in Synthesized Speech. Journal of the American Voice I/O Society, July.

Cassell, J., Vilhjlmsson, H., & Bickmore, T. (2001). BEAT: The Behavior Expression Animation Toolkit, *Proc. of SIGGRAPH,* ACM Press.

Gratch, J., (2000). Emile: marshalling passions in training and education. *Proc. of the Fourth International Conference on Intelligent Agents*, Barcelona, SPAIN.

Gratch, J. and Marsella, S. (2001). Tears and Fears: Modeling emotions and emotional behaviors in synthetic agents. *Proc. of the 5th International Conference on Autonomous Agents,* Montreal, Canada.

Hovy, E. H. (1988). Generating Natural Language under Pragmatic Constraints. Lawrence Erlbaum, Hillsdale, New Jersey.

Knight, K. & Hatzivassiloglou, V. (1995). Two-level, many-paths generation. *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, Mass, June 1995, pp. 252-260.

Lazarus, R.S., (1991). *Emotion and Adaptation*. Oxford Press.

Loyall, A.B., and J. Bates. 1997. Personality-rich believable agents that use language. *Proc. of the First International Conference on Autonomous Agents* (106--113).

Marsella, S. and Gratch, J.(2001) Modeling the Interplay of Emotions and Plans in Multi-Agent Simulations. *Proc. of the 23rd Annual Conference of the Cognitive Science Society,* Edinburgh, Scotland.

Marsella, S., Gratch, J., and Rickel, J. (2001). The Effect of Affect: Modeling the Impact of Emotional State on the Behavior of Interactive Virtual Humans. *Proc. of the Agents2001 Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents,* Montreal, Canada.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press.

Ortony, A., Clore, G. L., and Collins, A. (1988). The Cognitive Structure of Emotions. Cambridge: Cambridge University Press.

Paris, C.L. (1988). Tailoring Object Descriptions to a User's Level of Expertise. *Computational Linguistics* 14(3): 64-78.

Rickel, J., & Johnson, W. (1999). Virtual Humans for Team Training in Virtual Reality. *Proc. of the Ninth International Conference on AI in Education*, pp. 578-585. IOS Press.

Swartout, W. et al. (2001). Towards the Holodeck: Integrating Graphics, Sound, Character and Story. In *Proc. of the Fifth International Conference on Autonomous Agents*, Montreal, Canada.

Walker, M., J. Cahn, and S. Whittaker. (1996). Linguistic style improvisation for lifelike computer characters. *AAAI Technical Report* WS-96-03