Affective Text: Generation Strategies and Emotion Measurement Issues

Ielka van der Sluis
Dept. of Computer Science
Trinity College Dublin
vdsluis@scss.tcd.ie

Chris Mellish
Dept. of Computing Science
University of Aberdeen
c.mellish@abdn.ac.uk

Gavin Doherty
Dept. of Computer Science
Trinity College Dublin
g.doherty@cs.tcd.ie

Abstract

In affective natural language generation (NLG) a major aim is to be able to influence the emotional effects evoked in the addressee through the intelligent use of language. While previous work has shown that varying the form of the language, while keeping the content the same, can have a measurable effect on the emotions of the addressee, we report here on work which investigated which linguistic techniques to give the text a more or less positive slant contribute to these emotional effects. We report on three studies in which texts that gave positive feedback on an IQ test performance were tested for emotional effects on the recipient. The first study followed a comparison method on the sentence level, and the second study compared the texts as a whole. In both of these, participants were asked to rate the emotional effects that they *thought* the texts would have. On the other hand, in the third study different types of feedback were evaluated in a context of use, where participants were asked to perform an IQ test, read their feedback and report on their *actual* emotional state. In the first two studies, participants confirmed that the texts contained essentially the same content. The positive slanting techniques generally resulted in texts that were judged to be either positive or equal to neutral texts, although the effects were less strong than in previous work, which employed a variety of techniques, and there were a number of exceptions which impact on the usefulness of these techniques. However the IQ-test experiment did not show any emotional effects arising from variation in the form of the feedback. We reflect on possible reasons for this outcome and what it might mean for further work on Affective NLG.

Introduction

There have been a number of applications that use speech or text with the intention to motivate or discourage, as well as to inform addressees. In the area of affective computing, there has also been some work on assessing the effects of interfaces on the emotions of their users, e.g. on their frustration levels (Prendinger, Becker, and Ishizuka 2006) or their feelings of support/trust (Lee et al. 2007). In Natural Language Generation (NLG, the task in natural language processing that involves the generation of natural language from a machine representation, such as a knowledge base or a logical form) there has been some work on task-based evaluation such as STOP (Reiter, Robertson, and Osman 2003) and SKILLSUM (Williams and Reiter 2008). In the area of 'Affective NLG', which has been defined as NLG that relates to, arises from or deliberately influences emotions or other non-strictly rational aspects of the Hearer (De Rosis and Grasso 2000), it is of great importance to find methods that allow us to measure the emotional effects as well as the strength of these effects of how particular content is formulated. Obviously people are affected by strategical decisions ('what to say'); telling good news makes people happier than telling bad news. But much of NLG is about tactical decisions ('how to say something') and the effects on the emotional state of the addressee of the particular way content is put into words is less obvious.

Previous work has suggested that tactical NLG choices can be used to achieve certain emotional effects in readers (van der Sluis and Mellish 2010). In their experiment, participants in Aberdeen worked through an IQ-test and received faked feedback on their IQ, either neutrally phrased or positively "slanted" using various techniques that could be implemented by an NLG system. Differences were measured between the resulting emotions of participants in the two conditions. The positive texts that were tested on readers combined a variety of techniques in order to make the reader happier (e.g. using rhetorical structures, adverbs, adjectives, choice of verbs, punctuation etc). We are interested in the relative power of these techniques – whether participants that receive a text that is more positively phrased using only one linguistic technique might be happier than participants that receive text slanted using another technique. We are thus aiming for a finer grained approach in which the emotional effects of particular linguistic techniques are tested in separation.

For this reason we composed three feedback texts: two positively slanted texts (each employing only one type of linguistic slanting) and one neutrally phrased baseline. To test the emotional effects of our feedback texts, we conducted three studies: (1) a text validation study in which we asked participants, out of context, to reflect on our materials at the sentence level, (2) a similar text validation study focussing on the whole feedback texts; and (3) an IQ test study to assess the feedback texts in an actual context of use. In the following we will describe the linguistic variations of interest and the experiments we conducted. The results of these experiments were out of line with previous work and we discuss why this might be and what this might tell us about the replicability of studies of this kind.

Linguistic Variations

In this experiment, we were interested in two different ways to "slant" a text in a positive way. These are illustrated below by pairs consisting of a "neutral" text, N, and a corresponding "slanted" one, A or R.

A ("adjectives")- Adding vague adjectives and adverbs, possibly also with "to be". For example,

N: Your Baumgartner score is 7.38.
A: Your Baumgartner score of 7.38 is excellent!

N: This is better than the average score obtained by ...
A: This is distinctively better than the average score obtained by ...

N: You are one point below average on Visual Intelligence.
A: You are slightly below average (one point) on Visual Intelligence.

R ("rhetorical")- Adjusting order and introducing discourse markers to emphasise positive aspects. This included adding markers such as "in particular" and "on top of this", as well as reordering as in:

N: You are five points above average on Clarity of Thought and Spatial Intelligence. You are one point below average on Visual Intelligence.
R: Although you are one point below average on Visual Intelligence, you are five points above average on Clarity of Thought and Spatial Intelligence.

The idea is that an NLG system would employ methods of this kind in order to "slant" a message in a particular direction, rather that to present a message in a more neutral way. Our intention was to investigate whether effects on emotions could be achieved purely by manipulations of type A or R, and to get insight on which method was more powerful in its effects. Van der Sluis & Mellish (2010) used both of these methods, as well as other ones, and we wished to untangle the source of their results.

Three Studies

In what follows we describe three studies which aim to measure the emotional effects of the linguistic variations mentioned above. We conducted two text validation studies (Study 1 based on the sentence level and Study 2 considering full texts), in which the the materials were composed by hand, in order to answer two questions:

1. Are these variations really tactical, in the sense that they do not change the content significantly?
2. Is it indeed true that certain phrasings are expected to be received more positively than others? (e.g. do people agree with us about whether presented situations are better or worse?)

In addition, we conducted a third study in which we investigated the affect of the texts in a context of use.

Study 1: Text Validation on the Sentence Level

Study 1 was conducted with 13 colleagues (academic staff, researchers and PhD students at Trinity College Dublin), all native speakers of English. The participants were asked to comment on 17 sentence pairs. 12 pairs were made up from applications of strategies A and R that we intended to use in the main experiment. 5 additional filler pairs were included to distract the participants from the goal of the study. In the 12 sentence pairs that we were interested in either an A-sentence or an R-sentence was contrasted with a neutral N-sentence. The following analysis reports on our findings on these 12 sentence pairs only.

As in Van der Sluis & Mellish (2010), to test our intuitions about the tactical nature of the linguistic alternations, the participants were presented with descriptions of two fictitious teachers, Mary Jones and Gordon Smith, both completely honest but with very different ideas about teaching (Mary believing that any pupil can succeed, given encouragement, but Gordon believing that most pupils are lazy and have overinflated ideas about their abilities). In the experiment participants were told that they would be shown pairs of examples of (unattributed) feedback given by teachers such as these to a pupil who had just done an intelligence test. We then asked the following questions about each pair (where Q1 and Q3 are copied from Van der Sluis & Mellish (2010)):

Q1 "Is it possible that the two examples might actually be (honestly) giving different feedback to the same pupil on the same task?"

Q2 "If you were a pupil, receiving feedback for some specific task, which piece of feedback would make you feel happier?"

Q3 "If the two pieces of feedback were given to the same pupil (for the same task) and the pupil's parents found out, do you think they would have grounds to make a complaint that one of the teachers is lying?"

We expected that participants would answer "yes" to Q1 and "no" to Q3. Q2 could be answered with "sentence 1", "sentence 2" and "no difference". We expected A-sentences and R-sentences to make readers happier than N-sentences. A-R sentence pairs were not tested, as we had no hypothesis for such a comparison.

In general, for all 12 sentence pairs, participants agreed with our hypotheses for Q1 (85.26%). For 11 of the 12 sentence pairs only one or sometimes two participants found that the sentences gave different feedback. In one case, which contrasts an A-sentence with an N-sentence five participants found that the content was different:

N: Your scores on Imagination/Creativity and on Clarity of Thought are higher than average.
A: Your scores on Imagination/Creativity and on Clarity of Thought are great and considerably higher than average.

but we nevertheless left these sentences among those used for subsequent experiments. For all sentence pairs 97.44% found that there was no reason to make a complaint. We believe this supports our claim that the linguistic variations are purely tactical. Table 1 presents the responses to Q2 for each sentence pair, where the AN pairs compare A with N, and the RN pairs compare R with N. It can be seen that slanted sentences are generally seen as positive or neutral,

	Prefer slanted	No difference	Prefer Neutral
AN1	11	1	1
AN2	9	3	1
AN3	11	2	0
AN4	12	1	0
AN5	6	5	2
AN6	10	3	0
AN7	12	0	1
RN1	1	11	1
RN2	3	9	1
RN3	8	4	1
RN4	7	5	0
RN5	6	6	1
Adjectives	78.02%	16.48%	5.49%
Rhetorical	39.06%	54.69%	6.25%
Totals	61.94%	32.26%	5.81%

Table 1: Study 1, answers to Q2

but we note that the distribution is much less positive than the previous experiment using a variety of slanting techniques (van der Sluis and Mellish 2010), where over 97% of slanted texts were judged as more positive. We can also see that while the rhetorical slanting technique produces a somewhat more positive text, the effect is smaller than the adjective technique at the sentence level. This goes some way towards answering our questions regarding the contribution of different techniques. It can be seen from the results that some sentence pairs were problematic:

R1: Your Baumgartner score is better than the average score obtained by other people in your age group - whereas the average is 6.8, yours is 7.38.

N1: Your Baumgartner score is 7.38. This is better than the average score obtained by other people in your age group, which is 6.8.

N2: Your scores on Imagination/Creativity and on Clarity of Thought are higher than average.

R2: In particular, your scores on Imagination/Creativity and on Clarity of Thought are higher than average.

While the overall results show that individual slanting techniques are judged to work, in general, even at the sentence level, the fact that the above examples are not clearly judged to be positively slanted illustrates that neither the rhetorical nor the adjectival strategies can be relied upon all of the time. This has implications for the application of these techniques to small fragments of text that may not afford the option of applying a variety of slanting techniques simultaneously.

Study 2: Text Validation on Full Texts

The texts used for this study are presented in Figure 1. In the experiment the texts were given to three groups named "N", "A" and "R". The texts consist of 5 to 9 sentences with a direct correspondence between the sentences of the three texts. Note that the actual IQ scores are the same in all three texts. Study 2 was conducted with 20 undergraduate students from Trinity College Dublin, all native speakers of English. The participants were presented the same scenario as used in Study 1 describing the two teachers Mary Jones and Gordon Smith and were asked to read and compare two texts, either the A text to the N text (10 participants) or the R text to the N text (10 participants). The order in which the texts in each pair were presented was randomised. After reading the two texts, participants were asked to answer questions Q1, Q2 and Q3 above. Our hypotheses were the same as in Study 1: Q1 would be answered "yes", Q3 would be answered "no" and for Q2 we expected that participants would believe that the R and the A text would make them happier than text N. Results confirm our hypotheses for Q1 and Q3, 95% of the participants felt that the content of both texts were the same and 85% felt that there was no cause for complaints. For Q2, however, opinions were again divided (Table 2). The results are very similar to the overall results of the sentence level experiment. While this provides further evidence that both strategies contribute towards a more positive text even when used in isolation, the similarity is also surprising. With a greater volume of positively slanted text, we might have expected a shift towards a more positive interpretation of the slanted text. Three subjects even judged the neutral text to be more positive, which again indicates that the technique would need to be used with care.

	Prefer slanted	No difference	Prefer Neutral
AN	6	2	2
RN	5	4	1
	55%	30%	15%

Table 2: Study 2, answers to Q2

Study 3: Measuring Affect in a Context of Use

Finally, we conducted a study using the set up described in (van der Sluis and Mellish 2010) in terms of procedure and type of participants, in which participants work through a fake IQ-test and receive feedback on their IQ. We are interested in whether participants that receive a text that is more positively phrased using only one linguistic technique display a larger change in their positive emotions than participants that receive neutrally phrased feedback. Such effects were found by (van der Sluis and Mellish 2010) when using multiple slanting techniques.

Participants 45 participants, all female students from Trinity College Dublin, took the IQ test. Participants were randomly assigned to a group and respectively received feedback which was either neutrally or positively phrased (but always based on the same scores). All participants except three were in age band 18-24. The exceptions were in age band 25-29 and did not receive the same feedback (i.e. were incidentally assigned to different groups in the study).

Materials The texts that we presented to our participants were portrayed as giving feedback on an IQ test that the participants had just taken. This feedback first explained the test and its type of scoring. This introduction was followed by one of the three texts in Figure 1. Before and after the participants took the IQ test, they were asked to fill out a questionnaire to assess their actual emotions and some questions for collecting demographical information. To test the participants' emotions we used a simplified version of the Positive and Negative Affect Scale (PANAS) (Watson, Clark, and Tellegen 1988) in order not to overburden the participants

N: Your Baumgartner score is 7.38. This is better than the average score obtained by other people in your age group, which is 6.8.
Your scores on Imagination/Creativity and on Clarity of Thought are higher than average. A factor analysis of your Baumgartner score gives more information.
You are five points above average on Clarity of Thought and Spatial Intelligence. You are one point below average on Visual Intelligence.
Your score is higher than average. There is a lot of variation in your age group.

A: Your Baumgartner score of 7.38 is excellent! This is distinctively better than the average score obtained by other people in your age group, which is 6.8.
Your scores on Imagination/Creativity and on Clarity of Thought are great and considerably higher than average. A factor analysis of your Baumgartner score gives more information. You outperformed most people in your age group with your scores for Imagination and Creativity (7.9) and Logical-Mathematical Intelligence (7.1).
You are an exceptional five points above average on Clarity of Thought and Spatial Intelligence. You are slightly below average on Visual Intelligence (one point). You outperformed most people in your age group with your exceptional scores for Imagination and Creativity (7.9) and Logical-Mathematical Intelligence (7.1).
Your score is significantly higher than average. There is a lot of variation in your age group.

R: Your Baumgartner score is better than the average score obtained by other people in your age group - whereas the average is 6.8, yours is 7.38.
In particular, your scores on Imagination/Creativity and on Clarity of Thought are higher than average. A factor analysis of your Baumgartner score gives more information.
Although you are one point below average on Visual Intelligence, you are five points above average on Clarity of Thought and Spatial Intelligence. On top of this you also outperformed most people in your age group with your scores for Imagination and Creativity (7.9) and Logical-Mathematical Intelligence (7.1).
There is a lot of variation in your age group, but your score is higher than average.

Figure 1: Linguistic variation used in the IQ test feedback

with questions and to avoid bored answering. In this test, which has been fully validated (Mackinnon et al. 1999) and which was found to be appropriate by Van der Sluis & Mellish (2010), participants have to rate only 10 instead of 20 terms: 5 for positive affect (i.e. alert, determined, enthusiastic, excited, inspired) and 5 for negative affect (i.e. afraid, scared, nervous, upset, distressed). Participants answered the PANAS questions using a slider on a 5 point scale, with two terms put at the extreme ends of the slider (i.e. 'very slightly/not at all' and 'extremely'). The materials, the test texts and questionnaires, as well as the experiment introduction and consent form were presented to the participants as a web experiment. For ethical reasons, participants received a debriefing about the aims of the study on paper from the experimenter in person.

Procedure The participants went at their own pace through the various phases of the experiment as follows:

1. General introduction to the experiment, in which participants were told that the experiment was 'an assessment of a new kind of intelligence test which combines a number of well-established methods that are used as indicators of human brain power'. To make it more difficult for the participant to keep track of how well/poorly she performed over the course of the test, it said that the questions in the test had different weight factors in the overall score;
2. Consent form;
3. Questionnaire on participant's demographics interleaved with the emotion test to assess the participant's current emotional state (i.e. 'how do you feel right now?');
4. Message: Invitation to press a button to start the IQ test;
5. 30 IQ test questions randomly ordered (but with the same order for each participant) to be answered one at a time. The questions that were used for the test were carefully collected from the internet and included items from various tests and games;
6. Message: 'Please be patient while your answers are being processed and your test score is computed. After the result page, you will be asked another set of questions about the test, your performance and the way you feel about it. This information is very important for this study, so please answer the questions as honestly as possible.';
7. Feedback of type A, R or N (see Figure 1);
8. Questionnaire: emotion test to assess how the participants felt after reading their feedback (i.e. 'How do you feel right now knowing your scores on the test') interleaved with questions about the test, their expectations etc.;
9. Debriefing which informed participants about the study's purpose and stated that the IQ test was not real. Payment.

Hypotheses The hypotheses for this study were that:

- participants who received the "positively slanted" texts using the adjectives and adverbs (group A) would show a larger change in their positive emotions than the participants who received the neutrally phrased texts (group N).
- participants who received the "positively slanted" texts using the rhetorical structures (group R) would show a larger change in their positive emotions than the participants who received the neutrally phrased texts (group N).

Results Table 3 indicates that participants in all groups rated the positive emotions almost exactly the same before and after they undertook the IQ test. Although we were interested in the positive affect, we also checked the data for the negative PANAS terms and found that those present a similar picture. It is not useful to perform any further tests as no significant differences will be found.

	R	*N*	*A*
Pos emotions Before	3.45(.60)	3.26(.80)	3.04(.71)
Pos emotions After	3.50(.75)	3.26(.98)	3.03(.71)

Table 3: *Means(Standard deviations)* for the averages of the positive emotion terms in Study 3 using a scale ranging from 1 (i.e. 'very slightly/not at all') to 5 (i.e. 'extremely') *Before* the IQ-test and *After* the feedback of type *N*, *R* and *A* was processed.

Discussion

Results of Study 1 in which linguistic material was assessed on the sentence level showed that participants judged the materials as having the same content. In addition, participants thought that the positively slanted sentence would make them happier in 62% of the cases, and would be judged as neutral in 32% of cases. While this might be expected from slantings that are 'tactical', in that they do not change the content of the text, the results are very different to a very similar study carried out by (van der Sluis and Mellish 2010) in which the outcome was over 97% positive. One reason for our low agreement might have been that our participants were asked to look at tactical variation in isolation, and the positive cues might be too subtle to pick up from individual sentences. Therefore in Study 2 we investigated the same materials but as a larger passage of text. The results, as discussed above, were very similar to the sentence level experiment, and while both slanting techniques were found to contribute, the effect was no larger than at the sentence level. For both experiments, neutral texts were in some cases judged as more positive than the slanted texts. The fact that such counterexamples emerged from these relatively small samples (155 pairs in Study 1, 20 in Study 2) shows that individual slanting techniques cannot be fully relied upon, even on larger passages of text.

While study 1 and study 2 were useful to check whether our material targeted purely tactical variations (which is strongly supported by both studies), the intrinsic and reflective nature of text validation experiments means that they only give a partial insight on actual emotional effects (i.e. participants may think they will have a different response to what they will actually have). Therefore, in a similar fashion to (van der Sluis and Mellish 2010), Study 3 investigated the affective responses in readers of the feedback texts in a context of use. In this study, in contrast to the text validation studies, participants were not asked to compare different phrasings, but were confronted with only one type of positive feedback. Whereas (van der Sluis and Mellish 2010) found strong effects, we were unable to measure any differences between the groups regarding the strength of the participants' emotions before and after they read their feedback. In the following we will address a number of issues that may have played a role in our failure to replicate the previous results using more fine-grained linguistic distinctions. While a smaller effect might have been expected from the use of individual slanting strategies in isolation, the absence of even a trend forces us to consider a whole range of possibilities which will require further work to verify.

No Actual Effect There are a number of possible reasons for the fact that we were unable to repeat the results obtained by Van der Sluis & Mellish (2010). One of them would be that there is no effect, the previous results were due to chance and affective NLG is not a productive research direction. This would mean that emotional effects in readers are caused solely by the content and any contribution by the phrasing is marginal. This seems unlikely. Some light on the effects of tactical variations in text is shed by work in psychology in the work on the effects of the "framing" of a text (Tversky and Kahneman 1981; Moxey and Sanford 2000; Teigen and Brun 2003). Other work in this area has been industrially funded, as there are considerable applications, for instance, in advertising. The alternative texts considered differ in ways that NLG researchers would call tactical, e.g. a description of a piece of meat as "75% lean" or "25% fat" are arguably alternative truthful descriptions of the same situation. Evaluation of this work has been primarily in terms of measuring the effects on people's *choices* or *evaluations* of options available (Levin, Schneider, and Gaeth 1998), or other aspects of task performance like motivation and beliefs (O'Hara and Sternberg 2001; Brown and Pinel 2003; Cadinu et al. 2005), but it seems not unreasonable to expect there to be effects on *emotions* as well.

Too Small Response For Measuring Method It could be that there was a small effect, but we were unable to measure it with the method we chose. In Study 1 and Study 2 a specific type of self-reporting was used, in that we asked people to first place themselves in a particular situation and then report on how they would feel. Hence, the participant's responses were not based on something that they actually had felt. In fact, it could have been a long time ago that they had found themselves in a similar situation. Moreover, the responses with the chosen method are dependent on the participants' capability of imagining themselves in a particular situation. However, as we did not find any effects in Study 3 possibly differences between our stimuli were just too subtle. Recall that Van der Sluis & Mellish (2010) used multiple linguistic techniques in the sentences they tested in their text validation, whereas we were interested in these techniques separately. Consequently, the emotional responses to those minimal differences may be marginal.

Culture Dependent Task Response Another reason could be that the task (taking an IQ test and getting feedback on it) elicits a culture-dependent response. It is possible that the participants in our study, who were based in a different country than the participants tested by Van der Sluis & Mellish (2010), had a different perception of the task they were asked to perform. The participants we tested may have perceived the task as uninteresting, artificial or, in the case of the third study, they did not become emotionally invested in the result of the IQ test. Alternatively, the content may have been more important for our participants than the linguistic style that was used for its presentation.

Cultural Dependent Measuring Method For the first two studies, it could be possible that the method for text validation we chose renders different results in different countries. Possibly, the participants tested by Van der Sluis & Mellish (2010), were better at imagining themselves in a particular situation and perhaps they were more comfortable reporting their imagined emotional responses. To investigate the possibility that our results for Studies 1 and 2 were obscured by cultural effects we repeated Study 1 at the University of Aberdeen where (van der Sluis and Mellish 2010) had carried out their work. 13 colleagues, all native speakers of English, took part. For question Q1 92.31% (compared to 85.26% at Trinity College Dublin) agreed with our intuitions that the content of the paired sentences was the same. Although with the group we tested at Trinity College there was one sentence pair that yielded a different result, in the new study all sentence pairs were considered to be tactical variations by at least 11 participants. This small difference does not seem to indicate a significant difference between the two populations. Q3 rendered similar results in both groups (97.44% vs. 98.71%) agreeing with us that there was no reason for complaints. Results for Q2 (i.e. which sentence would make them happier) were also very similar (61.94% vs. 66.66%). Thus overall it does not seem as if cultural effects were playing a role for Study 1. For Study 3, the PANAS questionnaire could have been a factor if participants differ between countries in their willingness to divulge personal information on emotional state.

Conclusion and Future Directions

In this paper we have reported on three studies that were conducted to find out if particular slanting techniques could be used to influence emotions in readers. From Study 1 and 2 we can conclude that the two slanting techniques we investigated in separation are judged as purely tactical. Although results included more neutral responses than found in previous work, the studies also showed that the slanting techniques generally produce more positive text even when used individually. However, there were exceptions at both the sentence level and above which may impact on the usefulness of these techniques in some contexts. No emotional effects were found when employing the texts in a context of use (Study 3), while using a similar set up as used in previous work in which such effects were found. We discussed a number of possible reasons for this unexpected outcome. For future work in affective NLG, where we want to measure readers' emotional responses to subtle tactical methods used in text, it could be that more sensitive measuring instruments are needed. Self-reporting (although the most valid) may not be the best method due to thresholding. Physiological methods may be an alternative, but unfortunately tend to have the problems of a complex set up and calibration. In addition, it is not always clear what is being measured by these methods (cf. (Lazarus, Kanner, and Folkman 1980; Cacioppo et al. 2000)). One way forward would be to combine multiple measuring instruments in Affective NLG.

Acknowledgments

This work was supported by the EPSRC platform grant 'Affecting people with natural language' (EP/E011764/1) and the SFI grant (07/CE/I1142) for the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin.

References

Brown, R., and Pinel, E. 2003. Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology* 39:626–633.

Cacioppo, J.; Bernston, G.; Larson, J.; Poehlmann, K.; and Ito, T. 2000. The psychophysiology of emotion. In Lewis, M., and Haviland-Jones, J., eds., *Handbook of Emotions*. New York: Guilford Press. 173–191.

Cadinu, M.; Maass, A.; Rosabianca, A.; and Kiesner, J. 2005. Why do women underperform under stereotype threat? *Psychological Science* 16(7):572–578.

O'Hara, L., and Sternberg, R. 2001. It doesn't hurt to ask: Effects of instructions to be creative, practical, or analytical on essay-writing performance and their interaction with students' thinking styles. *Creativity Research Journal* 13(2):197–210.

Lazarus, R.; Kanner, A.; and Folkman, S. 1980. Emotions: A cognitive-phenomenological analysis. In Plutchik, R., and Kellerman, H., eds., *Emotion, theory, research, and experience*. New York: Academic Press. 189–217.

Lee, J.-E.; Nass, C.; Brave, S.; Morishima, Y.; Nakajima, H.; and Yamada, R. 2007. The case for caring co-learners: The effects of a computer-mediated co-learner agent on trust and learning. *Journal of Communication* 57(2):183–194.

Levin, I.; Schneider, S.; and Gaeth, G. 1998. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behaviour and human decision processes* 76(2):149–188.

Mackinnon, A.; Jorm, A.; Christensen, H.; Korten, A.; Jacomb, P.; and Rodgers, B. 1999. A short form of the positive and negative affect schedule: evaluation of factorial validity and invariance across demographic variables in a community sample. *Personality and Individual Differences* 27(3):405–416.

Moxey, L., and Sanford, A. 2000. Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology* 14(3):237–255.

Prendinger, H.; Becker, C.; and Ishizuka, M. 2006. A study in users' physiological response to an empathic interface agent. *International Journal of Humanoid Robotics* 3(3):371–391.

Reiter, E.; Robertson, R.; and Osman, L. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence* 144:41–58.

De Rosis, F., and Grasso, F. 2000. Affective natural language generation. In Paiva, A., ed., *Affective Interactions*. Springer LNAI 1814. 204–218.

Teigen, K., and Brun, W. 2003. Verbal probabilities: A question of frame. *Journal of Behavioral Decision Making* 16:53–72.

Tversky, A., and Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science* 211:453–458.

van der Sluis, I., and Mellish, C. 2010. Towards empirical evaluation of affective tactical NLG. In Krahmer, E., and Theune, M., eds., *Empirical Methods in Natural Language Generation*, volume LNCS 5980. Springer, Berlin/Heidelberg.

Watson, D.; Clark, L.; and Tellegen, A. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology* 54:1063–1070.

Williams, S., and Reiter, E. 2008. Generating basic skills reports for lowskilled readers. *Journal of Natural Language Engineering* 14(4):495–525.