Using Tactical NLG to Induce Affective States: Empirical Investigations

Ielka van der Sluis
Computing Science,
University of Aberdeen
i.v.d.sluis@abdn.ac.uk

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

Abstract

This paper reports on attempts at Aberdeen[1] to measure the effects on readers' emotions of positively and negatively "slanted" texts with the same basic message. The "slanting" methods could be implemented in an (NLG) system. We discuss a number of possible reasons why the studies were unable to show clear, statistically significant differences between the effects of the different texts.

1 Introduction: Affective NLG

"Affective NLG" 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). Although this term could cover a range of types of NLG, in practice, a lot of work on affective NLG emphasises the *depiction* of emotional states/personalities (Oberlander and Gill, 2004), rather than the *induction* of emotional effects on readers. However, there are many applications where the intention is, for instance, to motivate or discourage, as well as to inform.

How can NLG influence the emotions of its readers? It is apparent that strategical decisions ("what to say") can make a difference on how a reader responds emotionally to a text. If you tell someone good news, they will be happier than if you tell them bad news. On the other hand, much of NLG is concerned with tactical decisions ("how to say it"), and the affective relevance of these is less clear. Can tactical NLG choices be used to achieve goals in terms of the reader's emotions? In the area of affective computing, there has been some work on assessing the effects of interfaces on the emotions of their users, e.g. on their frustration levels (Prendinger et al., 2006) or their feelings of support/trust (Lee et al., 2007). In NLG there has been some work on task-based evaluation cf. STOP (Reiter et al., 2003) and SKILLSUM (Williams and Reiter, forthcoming). However, to our knowledge, there has not yet been any demonstration of tactical decisions making a difference on a reader's emotions.

The paper is organised as follows: Section 2 introduces the tactical choices we are studying, our test texts and a text validation study. Section 3 discusses a pilot study that was conducted to try out potential psychological measurement methods. Section 4 presents a full study to measure the affect of text invoked in readers. The paper closes with a discussion of the findings and their possible implications.

2 Tactical Choices

We decided that a safe way to start would be to choose primitive positive versus negative emotions (such as sadness, joy, disappointment, surprise, anger), as opposed to more complex emotions related to trust, persuasion, advice, reassurance. Therefore we focus here on alternatives that give a text a positive or negative "slant". These could be applied by an NLG system whose message has "positive" and "negative" aspects, where "positive" information conjures up scenarios that are pleasant and acceptable to the reader, makes them feel happy and cooperative etc. and "negative" information conjures up unpleasant or threatening situations and

[1] Ielka van der Sluis is now at the Department of Computer Science, Trinity College, Dublin

so makes them feel more unhappy, confused etc. For instance, (DeRosis et al., 1999) discuss generating instructions on how to take medication which have to both address positive aspects ('this will make you feel better if you do the following') and also negative ones ('this may produce side-effects, which I have to tell you about by law'). An NLG system in such a domain could make itself popular by only mentioning the positive information, but then it could leave itself open to later criticism (or litigation) if by doing so it clearly misrepresented the true situation. Although it may be inappropriate grossly to misrepresent the provided message, there are more subtle (tactical) ways to "colour" or "slant" the presentation of the message in order to emphasise either the positive or the negative aspects.

We assume that the message to be conveyed is a simple set of propositions, each classified in an application-dependent way as having positive or negative *polarity* according to whether the reader is likely to welcome it or be unhappy about it in the context of the current message.[2] In general, this classification could, for instance, be derived from the information that a planning system has about which propositions support which goals (e.g. to stay healthy one needs to eat healthy food). We also assume that a possible phrasing for a proposition has a *magnitude*, which indicates the degree of impact it has. This is independent of the polarity. We will not need to actually measure magnitudes, but when we make claims that one wording of a proposition has a smaller magnitude than another we indicate this with $<$. For instance, we would claim that usually:

"*a few rats died*" $<$ "*many rats died*"

Thus we claim that "a few rats died" has less impact than "many rats died", whether or not rats dying is considered a good thing (i.e. whether the polarity is positive or negative). In general, an NLG system can manipulate the magnitude of wordings of the propositions it expresses, to indicate its own (subjective) view of their importance. In order to slant a text positively, it can express positive polarity propositions in ways that have high magnitudes and negative polarity propositions in ways that have low magnitudes. The opposite applies for negative slanting. Thus, for instance, in an application where it is bad for rats to die, expressing a given proposition by "a few rats died" would be giving more of a positive slant, whereas saying "many rats died" would be slanting it more negatively.

Whenever one words a proposition in different ways, it can be claimed that a (perhaps subtle) change of meaning is involved. In an example like this, therefore, perhaps the content of the message changes between the two wordings and so this is in fact a strategic alternation. In this work, we take the view that it is legal to make changes that relate to the writer's attitude to the material of the text. The difference between "a few rats" and "many rats" is (in our view) that the number of rats is either less than or more than *the writer would have expected*. We can therefore choose between these alternatives by varying the writer, not the underlying message. Another reason for considering this choice as tactical is that in an NLG system, it would likely be implemented somewhere late in the "pipeline". Our claim that pairs such as this can appropriately describe the same event is also supported by our text validation experiments described below.

2.1 Test Texts

We started by composing by hand two messages containing mainly negative and positive polarity propositions respectively. The negative message tells the reader that a cancer-causing colouring substance is found in some foods available in the supermarkets. The positive message tells the reader that foods that contain Scottish water contain a mineral which helps to fight cancer. The first paragraph of both texts states that there is a substance found in consumer products that has an effect on people's health and it addresses the way in which this fact is handled by the relevant authorities. The second paragraph of the text elaborates on the products that contain the substance and the third paragraph explains in what way the substance can affect people's health.

To study the effects of different wordings, for each text a positive and a negative version was produced by slanting propositions in either a positive or a negative way. This resulted in four texts in total, two texts with a negative message one positively

[2] Note that this sense of "polarity" is not the same as the one used to describe "negative polarity items" in Linguistics

and one negatively phrased (NP and NN), and two texts with a positive message one positively and one negatively verbalised (PP and PN). To maximise the impact aimed for, various slanting techniques were used by hand as often as possible without loss of believability (this was assessed by the intuition of the researchers). The positive and negative texts were slanted in parallel as far as possible, that is in both texts similar sentences were adapted so that they emphasised the positive or the negative aspects of the message. The linguistic variation used in the texts was algorithmically reproducible and the techniques are illustrated below. A number of these were suggested by work on "framing" in Psychology (Moxey and Sanford, 2000; Teigen and Brun, 2003). Indeed, that work also suggests further variations that could be manipulated, for instance, the choice between using numerical and non-numerical values for expressing quantities.

SLANTING EXAMPLES FOR THE NEGATIVE MESSAGE

Here it is assumed that recalls of products, risks of danger etc. involve negative polarity propositions. Therefore negative slanting will amongst other things choose high magnitude realisations for these.

Techniques involving adjectives and adverbs:

- "*A recall*" $<$ "*A large-scale recall*" of infected merchandise was triggered

Techniques involving quantification:

- Sausages, tomato sauce and lentil soup are "*some*" $<$ "*only some*" of the affected items

Techniques involving a change in polarity

Proposition expressed with positive polarity:

- Tests on monkeys revealed that as many as "*40 percent*" of the animals infected with this substance "*did not develop any tumors*"

Proposition expressed with negative polarity:

- Tests on monkeys revealed that as many as "*60 percent*" of the animals infected with this substance "*developed tumors*".

Techniques manipulating rhetorical prominence

Positive slant:

- "So your health is at risk, but every possible thing is being done to tackle this problem"

Negative slant:

- "So although every possible thing is being done to tackle this problem, your health is at risk"

SLANTING EXAMPLES FOR THE POSITIVE MESSAGE

Here it is assumed that killing cancer, promoting Scottish water etc. involve positive polarity propositions. Therefore positive slanting will amongst other things choose high magnitude realisations for these.

Techniques involving adjectives and adverbs:

- Neolite is a "*detoxifier*" $<$ "*powerful detoxifier*" preventing cancer cells

Techniques involving quantification:

- "*Cancer-killing Neolite*" $<$ "*Substantial amounts of cancer-killing Neolite*" was found in Scottish drinking water

Techniques involving a change in polarity

Proposition expressed with negative polarity:

- A study on people with mostly stage 4 cancer revealed that as many as "*40 percent*" of the patients that were given Neolite "*still had cancer*" at the end of the study.

Proposition expressed with positive polarity:

- A study on people with mostly stage 4 cancer revealed that as many as "*60 percent*" of the patients that were given Neolite "*were cancer free*" at the end of the study.

Techniques manipulating rhetorical prominence

Negative slant:

- "Neolite is certainly advantageous for your health, but it is not a guaranteed cure for, or defence against cancer"

Positive slant:

- "So Although Neolite is not a guaranteed cure for, or defence against cancer, it is certainly advantageous for your health"

2.2 Text validation

To check our intuitions on the effects of the textual variation between the four texts described above, a text validation experiment was conducted in which 24 colleagues participated. The participants were randomly assigned to one of two groups (i.e. P and N), group P was asked to validate 23 sentence pairs from the positive message (PN versus PP) and group N was asked to validate 17 sentence pairs from the negative message (NN versus NP). Each pair consisted of two sentences intended to differ in their magnitude but to be possible realisations of the same underlying content (as in the examples in the last section). Both the N and the P group sentence pairs included four filler pairs. The participants in group

P were asked which of the two sentences in each pair they thought most positive in the context of the message about the positive effects of Scottish water. The participants in group N were asked which of the two sentences in each pair they found most alarming in the context of the message about the contamination of food available for consumption. All participants were asked to indicate if they thought the sentences in each pair could be used to report on the same event (i.e. represented purely tactical variations).

Results in the N group indicated that in 89.75% of the cases participants agreed with our intuitions about which one of the two sentences was most alarming. On average, per sentence pair 1.08 of the 12 participants judged the sentences differently than what we expected. In 7 of the 13 sentence pairs (17 - 4 fillers) participants unanimously agreed with our intuitions. In the other sentence pairs 1 to, maximally, 4 participants did not share our point of view. In the two cases in which four participants did not agree with or were unsure about the difference we expected, we adapted our texts. One of these cases was the pair:

> "*just 359*" infected products have been withdrawn $<$ "*as many as 359*" infected products have been withdrawn "*already*"

We thought that the latter of the two would be more alarming (and correspond to negative slanting) because it is a bad thing if products have to be withdrawn (negative polarity). However, some participants felt that products being withdrawn was a good thing (positive polarity), because it meant that something was being done to tackle the problem, in which case the latter would be imposing a positive slant. As a consequence of the validation results, it was decided to 'neutralise' this sentence in both the NP and NN versions of the text to "359 infected products have been withdrawn". Overall, in 78.85% of the cases the participants thought that both sentences in a pair could report on the same event.

Results in the P group were similar. In 82.46% of the cases participants agreed with our intuitions about which one of the two sentences was most positive. In two cases, minor changes were made to make the texts clearer. Overall, in 86.84 % of the cases the participants thought that both sentences in a pair could report on the same event.

3 Pilot Study: Testing Psychological Methods to Measure Emotions

3.1 Psychological Methods

The next step was to determine plausible methods to measure the emotional effect of a text. There are two broad ways of measuring the emotions of human subjects – physiological methods and self-reporting. Because of the technical complications and the conflicting results to be found in the literature, we opted to ignore physiological measurement methods and to investigate self-reporting. To measure these emotions we decided do a pilot study to try out three well-established methods that are used frequently in the field of psychology, the Russel Affect Grid (Russell et al., 1989), the Positive and Negative Affect Scale (PANAS) (Watson et al., 1988), and the Self Assessment Manikin (SAM) (Lang, 1980). The PANAS test is a scale consisting of 20 words and phrases (10 for positive affect and 10 for negative affect) that describe feelings and emotions. Participants read the terms and indicate to what extent they experience(d) the emotions indicated by each of them using a five point scale ranging from (1) very slightly/not at all, (2) a little, (3) moderately, (4) quite a bit to (5) extremely. A total score for positive affect is calculated by simply adding the scores for the positive terms, and similarly for negative affect. The Russel Affect Grid and the SAM test both assess valence and arousal on a nine-point scale.

3.2 Method: Subjects, Stimuli and Setting

Our pilot study aimed to test a general experiment set up, and to help us find the most promising of the above methods to measure emotions evoked by text. 24 colleagues and students (other than the ones involved in the text validation experiments) participated as subjects in this pilot study in which they were asked to fill in a few forms about how they felt after reading a particular text. All, except three, were native or fluent speakers of English and none was familiar with the purposes of the study. The subjects were divided in two groups of 12 subjects each, and were asked to fill in some questionnaires and to read a text about a general topic with a particular consequence for the addressee. For this experiment, just the negative message texts illustrated in the previous section were used (i.e. “some of your food contains a substance that causes cancer”). One group of subjects, the NP-group, was given this negative message verbalised in a neutral way giving the impression that although there was a problem every possible thing was being done to tackle it. The other group, the NN-group, was given the same negative message presented in a negative way implying that although many things were being done to tackle the problem, there still was a problem. We expected that after the subjects had read the text, the emotions of the subjects in the NN-group would be more negative than the emotions of the subjects in the NP-group. We also expected the subjects in the NN-group to be more strongly affected than the subjects in the NP-group.

For ethical reasons, both in this experiment and the following one, the main experimental procedure was followed by a debriefing session in which the subjects were informed that they had been deceived by the texts presented and during which it was possible to provide support for subjects if their emotional reactions had been especially strong.

3.3 Results and Discussion

Overall, t-test results failed to find significant differences between the the NN-group and the NP-group for any of the emotion measurement methods used. The Russel test, which was taken before the participants read the test text[3], indicated that the participants in the NP group might be feeling slightly more positive and less aroused than the participants in the NN group. The results for the PANAS test, taken after the participants read the test text, show that the NP group might be feeling a little bit more positive that the NN group about the content of the text they just read. The Sam test, which the participants were also asked to fill out with respect to their feelings after reading the test text, indicates that the NP group might be feeling less positive and more aroused than the NN group.

How to interpret the outcomes of the pilot study? There are several factors that could have caused the lack of significant results. One reason could be that the differences between the NP and NN texts were not large enough. Yet another reason could be that the people that took part in the study were not really involved in the topic of the text or the consequences of the message. When looking at the three emotion measurement methods used, some participants did indicate that the SAM and Russel tests were difficult to interpret. Also some participants showed signs of boredom or disinterest while rating the PANAS terms, which were all printed on one A4 page; some just marked all the terms as ‘slightly/not at all’ by circling them all in one go instead of looking at the terms separately. Also, some participants indicated that they found it difficult to distinguish particular terms. For example the PANAS test includes both ‘scared’ and ‘afraid’. As a consequence, there were several things that could be improved and adjusted before going ahead with a full scale experiment in which all four texts were tested.

4 Full Study: Measuring Emotional Effects of Text

This section presents a full scale experiment conducted to assess the emotional effect invoked in readers of a text. The experimental set up attempts to take into account the results found in the pilot study presented in the previous section. However, there were obviously a number of things that could be improved after this study, and so many things were changed without any direct evidence that they would improve the experiment. Below the method, data processing and results are presented and discussed.

4.1 Method: subjects, stimuli and experimental setting

Based on the pilot results, the setup of this study was adapted in a number of ways. For instance, we decided to increase the likelihood of finding measurable emotional effects of text by targeting a group of subjects other than our sceptical colleagues. Because it has been shown that young women are highly interested in health issues and especially health risks (Finucane et al., 2000), we decided on young female students as our participants.

[3] Ideally we would have presented all tests both before and after the text was read, but we believed that this would overload the subjects and lead to distorted results.

In total 60 female students took part in the experiment and were paid a small fee for their efforts. The average age of the participants was about 20.57 (std. 2.41) years old. The participants were evenly and randomly distributed over the four texts (i.e. NN, NP, PN, PP) tested in this study, that is 15 participants per group. The texts were tailored to the subject group, by for example mentioning food products that are typically consumed by students as examples in the texts and by specifically mentioning young females as targets of the consequences of the message. On a more general level, the texts were adapted to a Scottish audience by, for instance, mentioning Scottish products and a Scottish newspaper as the source of the article. Although the results of the pilot study did not indicate that the texts were not believable, we thought that the presentation of the texts could be improved by making them look more like newspaper articles, with a date and a source indication.

To enhance the experimental setting, the emotion measurement methods were better tailored to the task. The SAM test as well as the Russel Grid were removed from the experiment set up, because they caused confusion for the participants in the pilot study. Another reason for removing these tests was to reduce the number of questions to be answered by the participants and to avoid bored answering. For the latter reason, also a previously used reduced version of the PANAS test (Mackinnon et al., 1999) was used, in which the number of emotion terms that participants had to rate for themselves was decreased from 20 to 10. This PANAS set, consisting of five positive (i.e. alert, determined, enthusiastic, excited, inspired) and five negative terms (i.e. afraid, scared, nervous, upset, distressed), was used both before and after participants read the test text. Before the participants read the test text, they were asked to indicate how they felt at that point in time using the PANAS terms. After the participants read the test text, they were asked to rate the affect terms with respect to their feelings about the text. Note that this is different from asking them about their current feelings, because we wanted to emphasise that we wanted to know about their emotions related to the content of the text they just read and not about their feelings in general. We expected that the reduced PANAS test would produce reliable results because of its previous successful use. Whereas in the pilot study each test was handled individually, the PANAS terms were now interleaved with other questions about recall and opinions to further avoid boredom.

4.2 Hypotheses

In this full study four texts were tested on four different groups of subjects. Two groups read the positive message (PP-group and PN-group) two groups read the negative message (NN-group and NP-group). Of the two groups that read the positive message, we expected the positive emotions of the participants that read the positive version of this message (PP-group) to be stronger than the positive emotions of the participants that read the neutral/negative version of this message (PN-group). Of the two groups that read the negative message, we expected the participants that read the negative version of this message (NN-group) to be more negative than the participants that read the positive version of the message (NP-group).

4.3 Results

Overall, participants in this study were highly interested in the experiment and in the text they were asked to read. Participants that read the positive message, about the benefits of Scottish water, appeared very enthusiastic and expressed disappointment when they read the debriefing from which they learned that the story contained no truth. Similarly, participants that read the negative message expressed anger and fear in their comments on the experiment and showed relief when the debriefing told them that the story on food poisoning was completely made up for the purposes of the experiment. Only a few participants that read a version of the negative message commented that they had got used to the fact that there was often something wrong with food and were therefore less scared. Table 1 shows some descriptives that underline these impressions. For instance, on a 5-point scale the participants rated the texts they read more than moderately interesting (average of *po-i* = 3.74). They also found the text informative (average of *inform* = 3.82) and noted that it contained new information (average of *new* = 4.05). These are surprisingly positive figures when we consider that the participants indicated only an average interest in food (average of

	PN	PP	NN	NP
pr-i	2.47(1.13)	3.07(1.03)	3.00(.85)	3.00(1.25)
inf	3.87(.83)	3.80(.94)	3.67(1.05)	3.93(.70)
pos	3.93(.96)	4.27(1.03)	1.67(.98)	1.67(.97)
neg	1.53(.64)	1.27(5.94)	4.07(1.22)	3.53(1.19)
new	4.13(1.18)	4.53(.64)	3.87(1.30)	3.67(1.59)
po-i	3.67(.82)	3.80(.78)	3.67(.72)	3.80(1.01)

Table 1: Means and Standard deviations (between brackets) for the PN, PP, NP and NN texts for various variables: *pr-i* interest in food before reading the text, the *inf*ormativeness of the message, the *pos*itive or *neg*ative polarity of the message, *new* information and the *po-i* post interest in the message. All measured on a 5-point Scale: 1 = not at all, . . ., 5 = extremely.

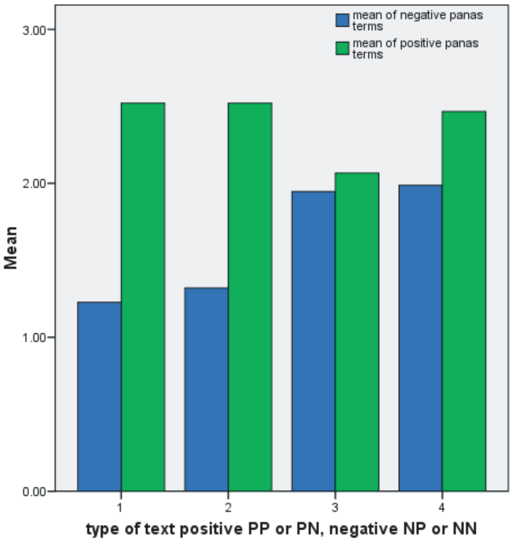

Figure 1: Positive and negative PANAS means after the Participants read the test text.

pr-i = 2.89) before they read the test text. The participants that read the negative messages (NN and NP) recognised that the message was negative (cf. *pos* and *neg* in Table 1). Moreover, the NN-group rated the text more negatively than the NP-group (4.07 vs 3.53). The participants that read the positive message found that they had read a positive message. The PP-group rated their text slightly more positive than the PN-group rated theirs.

The bar chart presented in Figure 1 illustrates the results of the PANAS questionnaire after reading the texts. In terms of the differences in message content (P* vs N*), there is a difference between the ratings of the negative terms, which is as expected. However, there is no significant difference for the positive terms, which were rated fairly similarly for all groups. Also, contrary to what was expected, the rating of the negative PANAS terms by both N* groups is lower than their rating of the positive terms. The hoped-for results for the positive/negative slanting are also not forthcoming - t-tests show no significant differences between the PN-group and the PP-group and no significant differences between the NN-group and the NP-group. All mean ratings stay far below 3, the ‘moderate’ average of the scale. When looking at these results in more detail, it appears that, of the positive PANAS terms, only ‘excited’ and ‘inspired’ had a higher mean for the positively worded message when comparing the positive and the negative version of the positive message (PP and PN). When comparing the positive and the negative version of the negative message (NP vs NN), as expected, the NN-group has lower means for all 5 positive terms than the NP group.

From this study various conclusions can be drawn. First of all, from the fact that only the lower half of the 5-point PANAS scale was used it can be concluded that the participants in this study seem to have difficulties with reporting on their emotions. This was the case both before and after the test text was read. Furthermore, participants seem to have a preference for reporting their positive emotions and focus less on their negative emotions. This can be inferred from the fact that the negative PANAS terms of the PP-group and the PN-group were lower than the means of the negative PANAS terms of the NN-group and the NP-group, but all groups had about the same means for the positive PANAS terms. The inference that self-reporting of emotions is troublesome is also indicated by the fact that the participants of this full study seemed highly interested and involved in the experiment and in what they read in the experiment texts. The participants generally believed the story they read and they expressed disappointment or relief when they were told the truth after the experiment. In addition, the descriptives in Table 1 show that participants generally correctly identified the text they read as either positive or negative. Note that in this respect the more fine-grained differences between the PP-group and the PN-group as well as the differences between the NN-group and the NP-group also confirm our expectations.

5 Conclusion and Discussion

This paper presented our efforts to measure differences in emotional effects invoked in readers. These efforts were based on our assumption that the wording used to present a particular proposition matters in how the message is received. Participants' judgements of the negative or positive nature of a text (in both the text validation and in the full study) are in accord with our predictions. In terms of *reflective analysis* of the text, therefore, participants behave as we expected. Although we strongly emphasised that we were interested in emotions with respect to the test text, our attempts to measure the *emotional effects* invoked in readers caused by tactical text differences did, however, not produce any significant results.

There are several reasons that may have played a role in this. It may be that the emotion measuring methods we tried are not fine-grained enough to measure the emotions that were invoked by the texts. As mentioned above, participants only used part of the PANAS scale and seemed to be reluctant to record their emotions (especially negative ones). Other ways of recording levels of emotional response that are more fine-grained than a 5-point scale, such as magnitude estimation (Bard et al., 1996), might be called for here. Carrying out experiments with even more participants might reveal patterns that are obscured by noise in the current study, but this would be expensive.

Alternatively, it could be that the differences between the versions of the messages are just too subtle and/or that there is not enough text for these subtle differences to produce measurable effects. Indeed, we are not aware of PANAS being used to assess purely textual effects before. Perhaps it is necessary to immerse participants more fully in slanted text in order to really affect them differently. Or perhaps more extreme versions of slanting could be found. Perhaps indeed the main way in which NLG can achieve effects on emotions is through appropriate content determination (strategy), rather than through lexical or presentation differences (tactics).

Another reason could still be a lack of involvement of the participants of the study. Although the participants of the full study indicated their enthusiasm for the study as well as their interest in the topic and the message, they may have felt that the news did not affect them too much, because they considered themselves as responsible people when it comes to health and food issues. We are designing a follow up experiment in which, to increase the reader's involvement, a feedback task is used, where participants play a game or answer some questions after which they receive feedback on their performance. The study will aim to measure the emotional effects of slanting this feedback text in a positive or a negative way. As in such a feedback situation the test text is directly related to the participants' own performance, we expect an increased involvement and stronger emotions.

As argued above, the results of our study seem to indicate that self-reporting of emotions is difficult. This could be because participants do not like to show their emotions, because the emotions invoked by what they read were just not very strong or because they do not have good conscious access to their emotions. Although self-reporting is widely used in Psychology, it could be that participants are not (entirely) reporting their true emotions, and that maybe this matters more when effects are likely to be subtle. In all of these situations, the solution could be to use additional measuring methods (e.g. physiological methods), and to check if the results of such methods can strengthen the results of the questionnaires. Another option is to use an objective observer during the experiment (e.g. videotaping the participants and observing the duration of smiles or frowns) to judge whether the subject is affected.

Yet another possibility would be only to measure emotional effects via performance on a task that is known to be facilitated by particular emotions. For instance, one could use the methods of (Carenini and Moore, 2000) to measure persuasiveness of different textual realisations that may induce emotions.

Acknowledgments

This work was supported by the EPSRC grant 'Affecting people with natural language' (EP/E011764/1) and also in part by Science Foundation Ireland under a CSET grant (NGL/CSET). We would like to thank the people who contributed to this study, most notably Louise Phillips, Emiel Krahmer, Linda Moxey, Graeme Ritchie, Judith Masthoff, Albert Gatt and Kees van Deemter.

References

E. G. Bard, D. Robertson, and A. Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

G. Carenini and J. D. Moore. 2000. An empirical study of the influence of argument conciseness on argument effectiveness. In *Proceedings of the 38th annual meeting of the Association for Computational Linguistics*.

F. DeRosis, F. Grasso, and D. Berry. 1999. Refining instructional text generation after evaluation. *Artificial Intelligence in Medicine*, 17(1):1–36.

M. Finucane, P. Slovic, C. Mertz, J. Flynn, and T. Satterfield. 2000. Gender, race, and perceived risk: the 'white male' effect. *Health, Risk & Society*, 2(2):159 – 172.

P. Lang, 1980. *Technology in Mental Health Care Delivery Systems*, chapter Behavioral Treatment and Biobehavioral Assessment: Computer Applications, page 119 137. Norwood, NJ: Ablex.

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

A. Mackinnon, A. Jorm, H. Christensen, A. Korten, P. Jacomb, and B. Rodgers. 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.

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

J. Oberlander and A. Gill. 2004. Individual differences and implicit language: Personality, parts-of-speech and pervasiveness. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

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

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

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

J. Russell, A. Weiss, and G. Mendelsohn. 1989. Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, 57:493–502.

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

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

S. Williams and E. Reiter. forthcoming. Generating basic skills reports for lowskilled readers. *Journal of Natural Language Engineering*.