Towards Affective Natural Language Generation

van der Sluis, Ielka; Mellish, C.

Towards Affective Natural Language Generation: Empirical Investigations

Ielka van der Sluis and **Chris Mellish** [1]

Abstract. This paper reports on attempts to measure the differing effects on readers' emotions of positively and negatively "slanted" texts with the same basic message. The methods of "slanting" the texts are methods that could be used automatically by a Natural Language Generation (NLG) system. A pilot study and a main experiment are described which use emotion self-reporting methods from Psychology.

Although the main experiment was formulated with the benefit of knowledge obtained from the pilot experiment and a text validation study, nevertheless it was unable to show clear, statistically significant differences between the effects of the different texts. We discuss a number of possible reasons for this, including the possible lack of involvement of the participants, biases in the self-reporting and deficiencies of self-reporting as a way of measuring subtle emotional effects.

1 Introduction: Affective NLG

Much previous research in Natural Language Generation (NLG) has assumed that the purpose of generated texts is simply to communicate factual information to the user [9]. On the other hand, in the real world, texts vary enormously in their *communicative purpose*, for instance they may aim to persuade, amuse, motivate or console. In general, even when a text communicates information, it usually does so in order to affect the reader at a deeper level, and this has an impact on *how* the information should be communicated (the central task of NLG). As a consequence of this, De Rosis and Grasso have defined the notion of "affective NLG" as "NLG that relates to, arises from or deliberately influences emotions or other non-strictly rational aspects of the Hearer" [11]. In practice, however, work on affective NLG mostly emphasises the depiction of emotional states/personalities [8], rather than ways in which texts can induce different effects on readers. To build systems which, from a model of the reader, can intelligently select linguistic forms in order to achieve a particular deep effect, we need a scientific understanding of how the attributes of an individual reader (and the reading process for them) influence the effect that particular linguistic choices have. But in order to evaluate such understanding, we need to have ways of measuring the effects that texts have, beyond simply testing for the recall of facts. The work described in this paper is an initial attempt to find out whether it is possible to measure emotions evoked in the reader of a text. In particular, can we detect the difference between different wordings of a text (that an NLG system might produce) in terms of the emotions evoked in the reader? Although there has been some work on task-based evaluation in NLG cf. STOP [10] and SKILLSUM (Williams and Reiter, In Press), to our knowledge, measurement of emotions invoked in readers is not something that has been investigated before.

This paper is organised as follows: Section 2 introduces our approach to linguistic choice, the composition of affective text and a text validation study. Section 3 discusses potential psychological methods to measure the emotional effect of text and Section 4 a pilot study that was conducted to try these out on text readers. Finally Section 5 brings all together in a full Study in which the texts resulting from our text validation experiments and the most promising affect measurement methods are used to measure the affect of text invoked in readers. The paper closes with a discussion of findings and future work.

2 Linguistic Choice

We decided that a safe way to start would be to aim for large effects in primitive emotions, (e.g. positive versus negative emotions, such as sadness, joy, disappointment, surprise, anger), as opposed to aspects of contextual impact (e.g. trust, persuasion, advice, reassurance). Therefore, although there are many linguistic choices that an NLG system might explicitly control, we focus here on alternatives that relate to simple goals of giving a text a positive or negative "slant". Very often the message to be conveyed by an NLG system 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 so makes them feel more unhappy, confused etc. An NLG system 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 misrepresents the true situation. For instance, [2] 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). Although it may be inappropriate grossly to misrepresent the provided message, there may be more subtle 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, negative or neutral *polarity* in the context of the message.[2] This classification could, for instance, be derived from the information that a planning system could have about which propositions support which goals (e.g. to stay healthy one needs to eat

[1] Computing Science, University of Aberdeen, email: {i.v.d.sluis,c.mellish}@abdn.ac.uk

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

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 about when 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*”

(“a few rats died” has less impact than “many rats died”, whether or not rats dying is considered a good thing or not). 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, there is a question about whether in fact the two different wordings actually correspond to different *messages*, rather than different *wordings* that might be chosen by an NLG system. In this paper, we assume that the choice between these two possibilities would likely be implemented somewhere late in the “pipeline”, and so we think of it as being a choice of form, rather than content. This interpretation is supported by our text validation experiments described below.

2.1 Test Texts

We started by composing two messages, a negative and a positive message, within a topic of general interest: food and health issues. 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 and to prevent cancer. The texts are set up in a similar way in that they both contain three paragraphs that address comparable aspects of the two topics. 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 extends 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. The slanting was done so that the positive and negative versions of the messages were still reporting on the same event. This resulted in four texts in total, two texts with a negative message one positively and one negatively phrased (NP and NN), and two texts with a positive message one positively and one negatively verbalised (PP and PN). For the negative message, the NP version is assumed to have less negative impact than the NN version. Likewise, the PN version of the positive message is assumed to have less positive impact than the PP version. To maximise the impact aimed for, various slanting techniques were used 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 can be coarsely classified as, on the one hand, created by the use of quantifiers, adjectives and adverbs to affect the conveyed magnitude of propositions, and, on the other hand, other techniques based on changing the polarity of the proposition (suggested by work on “framing” in Psychology [7];[15]) and changing the rhetorical structure to alter the prominence of propositions. Below, this variation is illustrated with examples taken from the two messages:

SLANTING EXAMPLES FOR THE NEGATIVE MESSAGE

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

Techniques involving adjectives and adverbs:

- “*A recall*” $<$ “*A large-scale recall*” of infected merchandise was triggered
- The substance is linked to “*a risk*” $<$ “*a significant risk*” of cancer

Techniques involving quantification:

- “*Some*” $<$“*Substantial amounts of*” contaminated food was withdrawn
- the substance was used in “*some*” $<$ “*many*” other products
- Since then “*more*” $<$ “*many more*” contaminated food products have been identified
- 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:

- “Scottish Water: “*A cancer-killer*” $<$ “*the Great Cancer-killer*”
- Neolite is a “*detoxifier*” $<$ “*powerful detoxifier*”

- Neolite is “*a possible*” $<$ “*an excellent*” cancer preventative
- Neolite has proven to be “*effective*” $<$ “*highly effective*” at destroying and preventing cancer cells

Techniques involving quantification:

- “*Cancer-killing Neolite*” $<$ “*Substantial amounts of cancer-killing Neolite*” was found in Scottish drinking water
- A campaign for the use of Scottish water in “*consumer products*” $<$ “*many more consumer products*”
- Waterwatch Scotland announced the start of an extensive campaign for the use of Scottish water in “*more*” $<$ “*many more*” consumer products

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 emotional effects of the textual variation between the four texts described above, a text validation experiment was conducted in which 24 colleagues of the Computing Science Department at the University of Aberdeen 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). 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. Below, the validations of the N and the P group are discussed separately.

N-Group Results 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 minus four fillers) participants unanimously agreed with our intuitions. In the other four 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”. The second sentence pair on which four participants disagreed was:

> you would be able to notice symptoms resulting from the substance “*just after*” $<$ “*already after*” ten years

Which we changed to:

> you would “*only*” $<$ “*already*” be able to notice symptoms resulting from the substance after ten years

because the original sentences seemed too complex to process. Overall, in 78.85 % of the cases the participants thought that both sentences in a pair could report on the same event.

P-Group Results indicated that in 82.46 % of the cases participants agreed with our intuitions about which one of the two sentences was most positive. In 4 of the 19 sentence pairs (23 minus 4 fillers) participants unanimously agreed with our intuitions. On average per sentence pair 2.11 of the 12 participants judged the sentences differently than what we expected. There were four cases in which a maximum of four participants did not agree with or were unsure about the difference we expected (i.e. in all other sentence pairs this number was less). In two of these cases we think that this disagreement was caused because the polarities of the sentences were more context-dependent than foreseen. We assumed that the larger amount/quantity the more positive the implications of the sentences:

- Scottish water is more beneficial for your health because it contains “*Neolite*” $<$ “*a large quantity of Neolite*”
- Scottish water is used in “*products*” $<$ “*a large number of products*” like,...

and yet some of the participants did not agree with this. Because of their context dependency and different judgements on similar cases on sentence pairs taken from the negative message texts, we decided to keep this variation. In the third case in which four people judged the sentences differently with respect to their positive impact, we thought the sentence too long. The sentence was split while the content was kept. In the fourth case our reasoning was that the larger the number of people that believed a particular fact the larger the impact:

> “*it is believed*” $<$ “*it is generally believed*” that taking Neolite is a cancer preventative

Because four participants in the text validation study disagreed with this assumption, the word ‘generally’ was removed from the positively slanted text. Overall, in 86.84 % of the cases the participants thought that both sentences in a pair could report on the same event.

3 Psychological Methods to Measure Emotions

The next step towards affective language generation is to find out what the best methods are 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. Indeed, standardised self-reporting questionnaires are widely used in psychological experiments. To measure these emotions we decided to try out three well-established methods that are used frequently in the field of psychology, the Russel Affect Grid [12], the Positive and Negative Affect Scale (PANAS) [18], and the Self Assessment Manikin (SAM) [5].

The PANAS test used in this pilot study is a scale consisting of a 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 consists of 81 cells which are arranged as a square of nine rows by nine columns with the rows defining the present level of arousal and the columns defining the present level of pleasure. By choosing the appropriate cell a participant simultaneously reports both aspects of his or her affective state.

The SAM test used in this study assessed the valence and arousal dimensions by means of two sets of graphical figures depicted in Figure 1. The participant ticks the 'dot' closest to the figure that represents his or her affective state best. The Russel Affect Grid and the SAM test were both used on a nine-point scale.

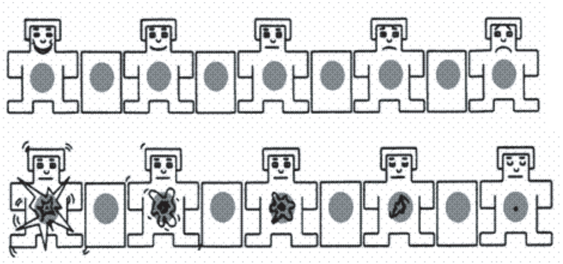

Figure 1. Self Assessment Manikin: the first row of pictures depicts valence the second row of pictures depicts arousal

4 Pilot Study

This section presents a pilot study that aimed to test a general experiment set up, and to help us find of the above methods the most promising ones to measure emotions evoked by text.

4.1 Method: Subjects, Stimuli and Setting

24 colleagues and students at the University of Aberdeen (other than the ones involved in the text validation experiments) participated as subjects in this pilot study in which they were asked to fill out 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 out 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 positive/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. The set up of the pilot study had nine phases as follows:

1. General information and instructions;
2. Consent form;
3. Questionnaire on participant's background and interests;
4. Russel Affect Grid to assess the participant's current emotional state;
5. Test text (NP or NN);
6. PANAS test to assess how the participants felt after reading the test text;
7. SAM test to assess how the participants felt after reading the test text;
8. Questionnaire to assess the participant's understanding and recall of the test text;
9. Debriefing which informed participants about the study's purpose and stated that the test text did not contain any truth.

4.2 Results

In general, the participants in the study indicated that they were interested in food. Before reading the text, they rated their interest in food 3.08 (std. 1.14) on a scale form 1 to 5. After reading the text, participants rated their interest in the topic of the text 2.96 (std. 1.30), the informativeness of the text 3.75 (std. 0.79) (all figures on a 5-point scale). The results of the emotion measurement methods used in the pilot study are presented in Table 1. Overall, the t-Test results failed to find significant differences between the two groups for any of the tests. The Russel test, which was taken before the participants read the test text, 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 (1.72 vs 1.51). 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.

4.3 Discussion

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. It is also possible that the standard emotion measurement methods used in this study are not fine-grained enough to

	NP	NN	t(*p*)
Russel valence	4.75 (1.71)	4.33 (2.64)	.459(.651)
Russel arousal	4.25 (2.38)	5.08 (1.56)	-1.014(.322)
PANAS positive	1.72 (1.01)	1.51 (.51)	.655(.520)
PANAS negative	1.94 (.67)	1.91(.59)	.108(.915)
SAM valence	5.58 (1.68)	4.92 (1.83)	.930(.362)
SAM arousal	6.58 (2.23)	5.67 (2.93)	.861(.917)

Table 1. Comparing NP and NN texts: Means(Standard deviations) for each of the psychological emotion measurement methods used, as well as the t-test results and their (in)significance. SAM and Russel are measured on a 9-point scale with 1 = happy/aroused, . . ., 9 = sad/sleepy. PANAS is measured on a 5-point Scale: 1 = not at all, . . ., 5 = extremely.

detect the emotional effects invoked by text. 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 test was 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.

5 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 is adapted to the results found of the pilot study presented in the previous section. Below the method, data processing and results are presented and discussed.

5.1 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 [3], we decided on young female students of the University of Aberdeen as our participants. In total 60 female students took part the experiment and were paid a small fee for their efforts. The average age of the participants was about 20 years old (see Table 2). 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 inert answering. For the latter reason, also a previously used reduced version of the PANAS test [6] was used, with 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 feeling, 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 feeling in general. In this way outliers could be detected at the start of the experiment (i.e. highly positive or depressed participants) and changes in a participant's emotions could be measured. Differently from the strategy used in the pilot study in which each test was handled individually, the PANAS terms were now interleaved with other questions about recall and opinions to further avoid boredom.

The set up of the full-scale study had six phases, where phases 3a and 3b and phases 5a and 5b were interleaved as follows:

1. General information and instructions;
2. Consent form;
3. (a) Questionnaire on participant's background and interests;
   (b) Reduced PANAS test to assess the participant's current emotional state;
4. Test text (NP or NN);
5. (a) Reduced PANAS test to assess the participants emotions about the test text;
   (b) Questionnaire to assess the participant's understanding and recall of the test text;
6. Debriefing which informed participants about the study's purpose and stated that the test text did not contain any truth.

5.2 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).

5.3 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 2 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 *inf* = 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 *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 2). Moreover, the NN-group rated the text more negative 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.

	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)
age	20.00(2.39)	20.93(2.74)	20.80(2.27)	20.53(2.23)
pPs	1.65(.81)	1.48(.41)	1.27(.49)	1.31(.48)
nPs	2.67(.71)	3.00(.82)	2.83(1.03)	3.12(.68)

Table 2. Descriptive statistics for PN, PP, NP and NN texts Means(Standard deviations) for various variables: *pr-i* the participant's interest in food before reading the text, the *inf*ormativeness of the message, if the message contained a *pos*itive or a *neg*ative message, *new*information, *po-i* to indicate if the participant's interest in the message, the *age* of the participants and *pPs* and *nPs*, respectively, the positive and the negative PANAS terms that were rated before the participants read the test text. All measured on a 5-point Scale: 1 = not at all, . . ., 5 = extremely.

In Table 2 the means and standard deviations of the PANAS test show that participants felt more positive than negative over all conditions. The participants that were going to read a negative message that was negatively verbalised (NN-group) were the most negative (1.27) of all groups. The participants that were going to read a negative message that was worded in a positive/neutral way (NP-group) were the most positive (3.12). Overall, the participants in this study did not differ much in terms of their positive and negative emotions. Differences were minimal and no extreme outliers were detected. Note that all figures except for the most positive one (3.12) are between 1 and 3 and not using the upper part of the 5-point scale. These results are graphically illustrated with the bar chart presented in Figure 2.

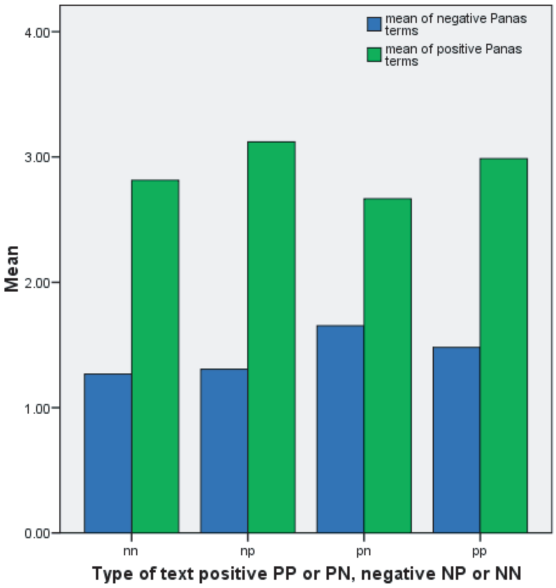

Figure 2. Positive and negative PANAS means before the Participants read the test text.

Table 3 presents the results of the PANAS questionnaire which the participants filled out after they read the test text usin a 5-point scale. The t-Test results show no significant differences between the PN-group and the PP-group and no significant differences between the NN-group and the NP-group. From the mean figures we can conclude that all groups rated the positive terms higher than the negative terms and that negative terms were rated higher by the participants that read the negative message than by the participants that read the positive message. Note that the average results with a maximum of 2.52 all stay far below 3, the 'moderate' average of the 5 point scale.

	negative PANAS terms	positive PANAS terms
PN	1.23 (.56)	2.52 (1.13)
PP	1.32 (.71)	2.52 (.80)
t(*p*)	.09 (.987)	.00 (1.00)
NN	1.95 (.81)	2.07 (.79)
NP	1.99 (.91)	2.47 (.88)
t(*p*)	.04 (.987)	.40 (.627)

Table 3. Descriptive statistics for PN, PP, NP and NN texts: Means(Standard deviations) for the positive and negative PANAS terms scored after the text was read, as well as the t-test results and their (in)significance. PANAS is measured on a 5-point Scale: 1 = not at all, . . ., 5 = extremely.

The bar chart presented in Figure 3 illustrates the results of the PANAS questionnaire and shows that the positive terms are rated similarly for the four texts. The NN-group rated the negative version of the negative message just .40 less in positive PANAS terms. Remarkably, and contrary to what was expected, the rating of the negative terms by both N* groups is still lower than the rating of the positive PANAS terms. Overall, the main difference between the groups is that the negative terms are rated lower by the PP-group and the PN-group than by the NN-group and the NP group.

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) (respectively, 2.60 vs. 2.33 and 2.67 vs. 2.40). Different from what was expected, the PN-group means of the positive terms ‘alert’, ‘determined’ and ‘enthusiastic’, were higher than the PP-group means (respectively, 2.33 vs. 2.47, 2.07 vs. 2.33 and 2.93 vs. 3.07). In addition, the means of the PP-group for the negative PANAS terms ‘scared and ‘nervous were higher than the means of the more neutrally verbalised version of this positive message (1.27 vs 1.13 and 1.53 vs 1.20). The means of the negative affect term ‘upset’ was the same for both groups (1.27). 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. In contrast, when looking at the negative terms, the mean of the NP group for ‘upset was higher than the NN-group mean for this term (2.07 vs 1.53).

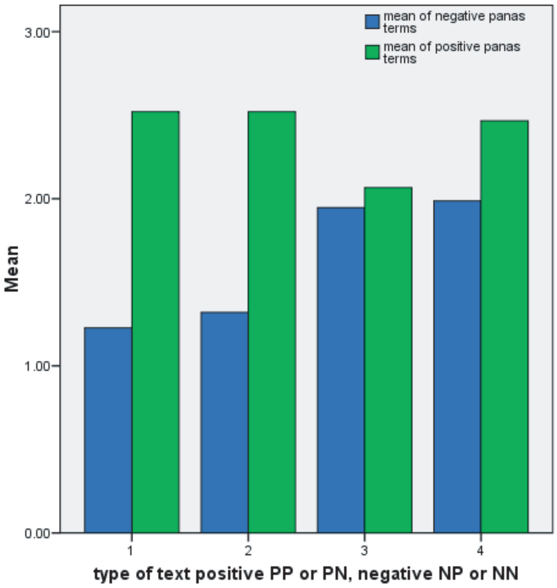

Figure 3. Positive and negative PANAS means after the Participants read the test text.

5.4 Discussion

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. In the remainder of this section we will focus on the PANAS test results that were obtained 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 2 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.

6 Conclusion and Future Directions

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. This assumption was tested and confirmed with the text validation experiments discussed in Section 2.2. The results of these experiments showed that participants generally agree on the relative magnitude or impact of different phrasings of propositions (depending on, for instance, quantifiers and adjectives used), while still allowing these phrasings to report on the same event. Also participants’ judgements of the negative or positive nature of a text 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 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 (cf. [1]; [14]), 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. 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) of the kind we have investigated here.

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. One could also try to measure emotions indirectly, for instance, by measuring whether people are more inclined to perform a particular action after reading a particular text (c.f. [4]). Another option is to use an objective observer during the experiment (e.g. videotaping the participants) to judge if the subject is affected or not.

Two other aspects that will be addressed in our follow up study are framing and multimodality. Inspired by [16] and [13], we aim to look at the impact of the context in which the feedback is presented. For instance, it might make difference to the emotions of the participants whether they are confronted with how well their peers are doing on the same task or whether they are shown the course of their own performance over time. The follow up study also aims to address emotional effects of multimodal presentations, as graphs and illustrations are believed to ease the interpretation process of a text. Yet another possibility might be to try to strengthen the impact of the feedback by asking the participant to read the text aloud instead of in silence (cf. [17]).

ACKNOWLEDGEMENTS

This work was supported by the EPSRC platfrom grant ‘Affecting people with natural language’ (EP/E011764/1). 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 and the anonymous reviewers of our paper.

REFERENCES

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

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

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

[4] E. Krahmer, J. van Dorst, and N. Ummelen, ‘Mood, persuasion and information presentation: The influence of mood on the effectiveness of persuasive digital documents’, *Information Design Journal and Document Design*, **12**(3), 40–52, (2004).

[5] P. Lang, *Technology in Mental Health Care Delivery Systems*, chapter Behavioral Treatment and Bio-behavioral Assessment: Computer Applications, 119 137, Norwood, NJ: Ablex, 1980.

[6] A. Mackinnon, A. Jorm, H. Christensen, A. Korten, P. Jacomb, and B. Rodgers, ‘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, (1999).

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

[8] J. Oberlander and A. Gill, ‘Individual differences and implicit language: Personality, parts-of-speech and pervasiveness’, in *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, (2004).

[9] E. Reiter and R. Dale, *Building Natural Language Generation Systems*, Cambridge, 2000.

[10] E. Reiter, R. Robertson, and L. Osman, ‘Lessons from a failure: Generating tailored smoking cessation letters’, *Artificial Intelligence*, **144**, 41–58, (2003).

[11] F. De Rosis and F Grasso, ‘Affective natural language generation’, in *Affective Interactions*, ed., A. Paiva, Springer LNAI 1814, (2000).

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

[13] S. Sher and C. McKenzie, ‘Information leakage from logically equivalent frames’, *Cognition*, **101**, 467–494, (2006).

[14] S. S. Stevens, ‘On the psychophysical law.’, *Psychological Review*, **64**, 153–181, (1957).

[15] K. Teigen and W. Brun, ‘Verbal probabilities: A question of frame’, *Journal of Behavioral Decision Making*, **16**, 53–72, (2003).

[16] A. Tversky and D. Kahneman, ‘rational choice and the framing of decisions’, *Journal of Business*, **59**(4, Part 2), 251–278, (1986).

[17] E. Velten, ‘A laboratory task for induction of mood states’, *Behavior Research & Therapy*, **6**, 473–482, (1968).

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

[19] S. Williams and E. Reiter, ‘Generating basic skills reports for lowskilled readers’, *To appear in Journal of Natural Language Engineering*.