

Towards Empirical Evaluation of Affective Tactical NLG

Ielka van der Sluis¹ and Chris Mellish²

¹ Dept. Computer Science, Trinity College Dublin, Ireland

ielka.vandersluis@cs.tcd.ie,

WWW home page: <https://www.cs.tcd.ie/Ielka.vanderSluis/>

² Dept. Computing Science, University of Aberdeen, Scotland, UK

c.mellish@abdn.ac.uk,

WWW home page: <http://www.abdn.ac.uk/~csc248/>

Abstract. One major aim of research in affective natural language generation is to be able to use language intelligently to induce effects on the emotions of the reader/ hearer. Although varying the *content* of generated language (“strategic” choices) might be expected to change the effect on emotions, it is not obvious that varying the *form* of the language (“tactical” choices) can do this. This chapter discusses two experiments carried out to show emotional effects of tactical variations. With the first experiment we were not able to show clear, statistically significant differences between the effects of the different texts in readers. We discuss a number of possible reasons and, building on our discoveries, we present a second experiment which does demonstrate such effects. This represents an important step towards the empirical evaluation of affective NLG systems.

1 Introduction

This chapter is about developing techniques for the empirical evaluation of affective natural language generation (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” [20]. It currently covers two main strands of work, the portrayal of non-rational aspects in an artificial speaker/writer (e.g. the work of [17] on projecting personality) and the use of NLG in ways sensitive to the non-rational aspects of the hearer/reader and calculated to achieve effects on these aspects (e.g. the work of [21] on generating instructions in an emotionally charged situation and that of [18] on producing appropriate tutorial feedback). Although there has been success in evaluating work of the first kind, it remains more problematic to evaluate whether work of the second type directly affects emotion or mood, or whether it succeeds or fails for other reasons. Existing work of this kind tends to be evaluated either by the reader’s performance on a task not directly related to emotions (e.g. [5]) or by an assessment of system outputs by experts (e.g. [8]). Neither of these tells us in a direct way whether emotions have been affected.

Since the work of [24], NLG tasks have been considered to divide mainly into those involving *strategy* (“deciding what to say”) and *tactics* (“deciding how to say it”). It seems clear that one can affect a reader’s emotion differently by making different strategic decisions about content (e.g. telling someone that they have passed an exam will make them happier than telling them that they have failed), but it is less clear that tactical alternations (e.g. involving ordering of material, choice of words or syntactic constructions) can have these kinds of effects. Unfortunately, the exact dividing line between strategy and tactics remains a matter of debate. For the purpose of this chapter, we take “strategic” to cover matters of basic propositional content (the basic information to be communicated) and “tactical” to include most linguistic issues, including matters of emphasis and focus, inasmuch as they can be influenced by linguistic formulation. It is important to know whether tactical choices can influence emotions because to a large extent NLG research concentrates on tactical issues (partly because strategic NLG remains a rather domain-specific activity).

Some light on the effects of tactical variations in text is shed by work in psychology, where there has been a great deal of work on the effects of the “framing” of a text [25, 19, 23]. 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. For instance, a piece of meat could be described as “75% lean” or “25% fat”, and arguably these are alternative truthful descriptions of the same situation. However, evaluation of this work has been primarily in terms of whether it affects people’s *choices* or *evaluations* of options available [15], or other aspects of task performance like motivation and beliefs [10, 2, 4]. As far as we know, nobody has detected *emotions* being affected in this way. There are therefore open questions about whether there can be non-rational effects of different tactical decisions on readers and whether it is possible to detect them. We believe that answering these is important for the further scientific development of affective NLG.

In the rest of this chapter, we first discuss our choice of a method for measuring emotions (section 2). This is followed by a section on the types of linguistic choices we consider. Then two attempts to measure emotional effects of tactical decisions in texts are described, in terms of the example texts we used, the text validation experiments we did to check our intuitions and the experimental set up and the results of the studies. Finally we reflect on the results in a concluding section.

2 Measuring Emotions

There are three broad ways of measuring the emotions of people – task performance methods, physiological methods and self-reporting. Task performance methods measure emotional effects via performance on a task that is known to be facilitated by particular emotions [6]. For instance, [5] measure persuasiveness of different textual realisations that may induce emotions. As indicated above, the trouble with such methods from our point of view is their indirectness – although

performance could indeed be influenced by emotions, it could also be affected by other factors. Physiological methods, on the other hand, whilst measuring something objective, unfortunately tend to have the problems of complex setup and calibration, which mean that it is hard to transport them between tasks or individuals. In addition, although emotional states are undoubtedly connected to physiological variables, it is not always clear what is being measured by these methods (cf. [14, 3]).

Because of the problems with the other two types of methods, we have turned to self-reporting methods as a way of measuring emotions. Although sometimes participants are not able objectively to report their own emotions (see our later discussion in section 6), such methods as the Russell Affect Grid [22], the Positive and Negative Affect Scale (PANAS) [27], and the Self Assessment Manikin (SAM) [13] are widely used in Psychology. 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 each term and indicate to what extent they experience(d) the emotion indicated by that term 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 Russell Affect Grid and the SAM test both assess valence and arousal separately on a nine-point scale.

3 Linguistic Choice

3.1 Polarity and Magnitude

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 so makes them feel more unhappy, confused etc. For instance, [21] 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 message³. 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, “eight of twenty rats died” by “a few rats died” would be giving more of a positive slant, whereas saying “many rats died” would be slanting it more negatively.

3.2 Tactical Methods

For our experiments, we have considered a number of different types of tactical methods that could be implemented straightforwardly in an NLG system⁴, as follows. Here, the word “positive polarity” is used to refer to propositions giving good news to the reader or attributes which give good news to the reader if they have high values (such as the reader’s intelligence). Similarly “negative polarity” refers to items that represent bad news, e.g. failing a test. In general, a further relevant concept is the “goal polarity” of a text in a context – which of positive and negative the text seeks to emphasise. A text which has a positive goal polarity will thus be “positively slanted” and one with a negative goal polarity will be “negatively slanted”.

A. Sentence emphasis - include explicit emphasis in sentences with the same polarity as the goal polarity (e.g. exclamation marks and phrases such as “on top of this”).

³ Note that this sense of “polarity” is not the same as the one used to describe “negative polarity items” in Linguistics

⁴ Though the choice about *when* to apply them might not be so straightforward.

- B. Choice of vague evaluative adjectives** - when evaluating attributes with the same polarity as the goal, choose vague evaluative adjectives that are more extreme over ones that are less extreme (e.g. “excellent”, rather than “ok” for positive polarity).
- C. Choice of vague adverbs** - provide explicit emphasis to propositions with the same polarity as the goal by including vague adverbs expressing great extent (e.g. “significantly”, rather than “to some extent” or no adverb).
- D. Choice of verbs** - for a proposition with the same polarity as the goal, choose a verb that emphasises the great extent of the proposition (e.g. “outperformed”, rather than “did better than”).
- E. Choice of realisation of rhetorical relations** - when realising a concession/contrast relation between a positive polarity proposition and one that is negative or neutral, word it so that the proposition with the goal polarity is in the nucleus (more emphasised) position (e.g. say “although you did badly on X, you did well on Y” instead of “although you did well on Y, you did badly on X” for a positive goal polarity).

Most of these methods have corresponding methods for when some polarity is opposite to the goal polarity. E.g. a sentence or adjective might be de-emphasised by the inclusion of a “hedge” in such cases.

The idea is that an NLG system would employ methods of this kind in order to “slant” a message in a particular direction, rather than to present a message in a more neutral way. This might be done, for instance, to induce positive emotions in a reader who needs encouragement or negative emotions in a reader who is over-confident.

We claim that these choices can be viewed as tactical, i.e. that they are “allowable” alternative realisations of the same underlying content. For instance, we believe a teacher could use such methods in giving feedback to a student needing encouragement without fear of prosecution for misrepresenting the same truth that would be expressed without the use of these methods.

Whenever one words a proposition in different ways, it can be claimed that a (perhaps subtle) change of meaning is involved. However, in these cases we claim that it is the *writer’s attitudes* that are being manipulated (and reflected in the text). We can therefore choose between these alternatives by varying the writer, not the underlying message. Our view is supported by a number of current accounts of the semantics of vague adjectives (though this is not an area without controversy). Many accounts of vagueness appeal to the idea that there is a norm which an adjective like “tall” implicitly refers to, and some of these argue both that the norm itself can be contextually determined and also that the amount by which the norm has to be exceeded has to be “significant” to a degree which is “relativized to some agent” [12]. For instance, with the phrase “John is tall”

“the property [...] attributed to John is not an intrinsic property, but rather a relational one. Moreover, it is not a property the possession of which depends only on the difference between John’s height and some norm, but also on whether that difference is a significant one. I take it that whether or not a difference is a significant difference does not depend only on its magnitude, but also on what our interests are” [9]

It is compatible with these accounts that different agents, with different interests and notions of what is noteworthy, can use vague adjectives in different ways⁵.

Probably the best way to check that we are using tactical alternations (according to our definition) is via some kind of text validation experiment in which human participants are asked to judge this. A good approximation seems to be to ask participants whether particular alternative sentences could be used to describe the same situation. Below we describe two studies in which we employ such text validation experiments, which provide strong support for our position.

4 Study I

4.1 Background for the Study

The goal of this first study was to experiment with emotion self-reporting methods as described in section 2 and to investigate whether we could reliably measure differences in emotions arising from tactical variations of texts.

4.2 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 (i.e. a version with positive or negative goal polarity) 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 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 [19, 23].

⁵ Though there are certainly *some* limits on the situations where a word like "tall" can be truthfully used to describe a height

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”

4.3 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 be alternative 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 which were meant to keep participants alert. 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.

4.4 Experiment

Participants

Because a pilot experiment using 24 of our sceptical colleagues failed to produce clear patterns of behaviour, we decided to increase the likelihood of finding measurable emotional effects of text by targeting a group of participants likely to be especially affected by the subject material. It has been shown that young women are highly interested in health issues and especially health risks [7] and so we decided on young female students as our participants. 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.

Materials

For this experiment, we used versions of the NN, NP, NN and PP texts described above, with the modifications made after the text validation experiment. The texts were tailored in superficial ways to the participant 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. 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. Before and after the participants read a test text, they were asked to fill out a questionnaire with a PANAS test and some questions for collecting demographical information. All materials, the test texts and questionnaires, as well as the experiment introduction, consent form and debriefing were presented to the participants printed on A4 pages.

Procedure

Participants were asked to fill in questionnaires before and after reading a text about a general topic that would have particular consequences for them. For ethical reasons, both in this experiment and the following one, the main experimental procedure was followed by a debriefing session in which the participants were informed that they had been deceived by the texts presented and during which it was possible to provide support for participants if their emotional reactions had been especially strong.

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. As the SAM test and the Russell Grid had caused confusion for participants in our previous work, we elected to use a version of the PANAS test. However, in the

pilot study some participants had showed signs of boredom or disinterest while rating the PANAS terms 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’. For these two reasons, a reduced (but still validated) version of the PANAS test [16] 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. In the questionnaires, the PANAS terms were interleaved with other questions about recall and opinions to further avoid boredom.

Hypotheses

Four texts were tested on four different groups of participants. 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 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 have stronger negative emotions than the participants that read the positive version of the message (NP-group).

Results

The bar chart presented in Figure 1 illustrates the results of the PANAS questionnaire before and after reading the texts for each condition. The data resulting from the study did not confirm our hypotheses. The hoped-for results for the positive/negative slanting are also not forthcoming - t-tests show no significant differences, the PP-group did not report stronger positive emotions than the PN-group and the NN-group did not report stronger negative emotions than the NP-group. Cronbach Alpha scores show that participants consistently rate the positive and negative terms before (α Positive = .775; α Negative = .823) and after (α Positive = .791; α Negative = .911) reading the text.

Post-hoc we could analyse the data as follows: In general, subjects in all conditions reported similar positive and negative emotions before reading the texts; there were no outliers. After reading the text, in terms of the differences in message content (P* vs N*), there is a difference between the ratings of

the negative terms; as one would expect the NN-group and the NP-group report stronger negative emotions than the PP-group and the PN group. However, there is no such 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. 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 PP than for PN. When comparing the positive and the negative version of the negative message (NP vs NN), as expected, the NN-group had lower means for all 5 positive terms than the NP group. But none of these results reached statistical significance.

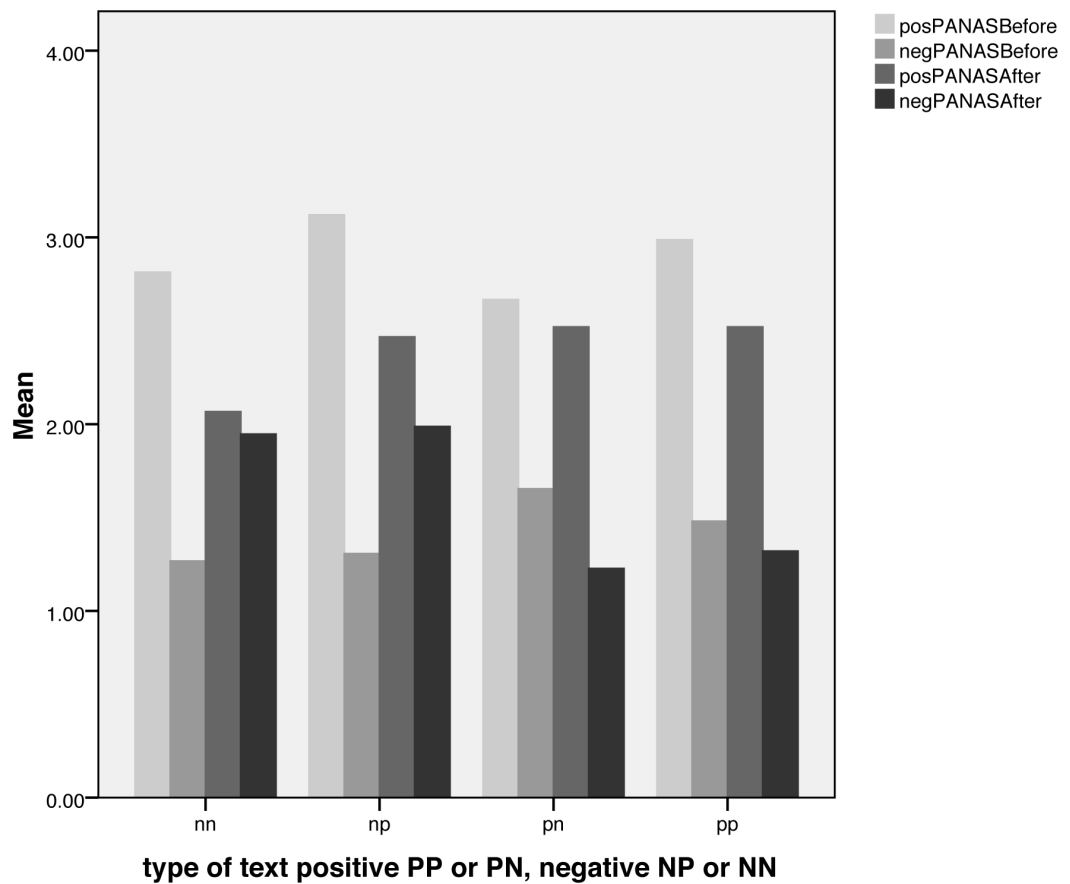


Fig. 1. Positive and negative PANAS means before and after the Participants read the test text.

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 $pr-i = 2.89$) 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.

	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(.59)	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 *informativeness* of the message, the *positive* or *negative* 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.

4.5 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.

Furthermore, participants seem to have a preference for reporting their positive emotions and focused less on their negative emotions. 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.

In this first attempt, we were not able to measure significant differences between the emotions evoked in readers dependent on the way their texts were phrased. There are several reasons that may have played a role in this. It may be that the emotion measuring methods we tried were 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 [1], 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 were just too subtle and/or that there was 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. It may not have helped that we asked for feelings about the text rather than simply current feelings. 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.

5 Study II

5.1 Background for the Study

In Study 1 we investigated different methods of measuring the effects of texts on emotions to demonstrate that tactical differences would lead to differences in effects. However, we were unable to show statistically significant results of

tactical variations and suggested various possible explanations. One involved the reliability of the reduced PANAS test, but this had been validated and used in multiple previous studies in Psychology, and so there is no reason to suggest that they would have fundamentally failed in this new context. Another was a problem with the limited granularity of the measurement method and partial use of the scale by participants. We addressed this in a simple way in Study II by having participants respond to the PANAS questions using a slider, rather than a five point scale. This means that only two terms were put at the extreme ends of the slider (i.e. ‘very slightly/not at all’ and ‘extremely’ were presented but not ‘a little’, ‘moderately’ or ‘quite a bit’).

Of the possible explanations for the Study I result, we believe that the possibility that the participants were not involved enough in the task to get strong emotions is the most compelling. It is very believable that the participants would fail to be really concerned by the texts in the experiments reported since the source was unclear, the message a general one not addressed to them individually and the topic (healthy and unhealthy food) one that occurs often enough in newspapers to fail to overcome natural boredom. Therefore the main innovation of Study II was in our method of seeking the emotional involvement of the participants. The texts that the participants read took the form of “feedback” on a (fake) IQ test that they undertook as part of the experiment. We selected university students as the participants, as they would likely be concerned about their intelligence, especially as compared to their peers. The texts appeared to be written individually for the participants and so sought to engage them directly.

5.2 Test Texts

For the experiment, we produced two feedback texts describing the same set of intelligence test results, one relatively neutral and one “positively slanted” using the above methods. For ethical reasons we did not use negative feedback. In the experiment, the feedback texts were given to participants in two groups, named “0” and “+” respectively. Each text consisted of 7 sentences, with a direct correspondance between the sentences of the two texts. Figure 2 presents the variations used in the feedback used in the experiment for group + (i.e. positively slanted) and group 0 (i.e. neutrally slanted). Note that the actual numbers are the same in both texts.

5.3 Text validation

A text validation study was conducted in which 15 colleagues participated. The participants were asked to comment on 12 sentence pairs, the 7 shown in Figure 2 and 5 additional filler pairs. The following analysis reports on our findings on the 7 sentence pairs shown in Figure 2 only.

In order that we could test our intuitions about the tactical nature of the linguistic alternations (discussed in section 3.2 above), the participants were presented with a scenario where there were two different teachers, Mary Jones and Gordon Smith, both completely honest but with very different ideas about

- +1: Your Baumgartner score of 7.38 is excellent!
- 01: Your Baumgartner score of 7.38 is ok.
- +2: You did distinctively better than the average score obtained by other people in your age group.
- 02: You did somewhat better than the average score obtained by other people in your age group.
- +3: Especially your scores on Imagination/Creativity and on Clarity of Thought were great and considerably higher than average.
- 03: Your scores on Imagination/Creativity and on Clarity of Thought were good and a little higher than average.
- +4: A factor analyses of your Baumgartner score results in an overall excellent performance.
- 04: A factor analyses of your Baumgartner score results in an overall reasonable performance.
- +5: Although, compared to your peers, you have only slightly higher Spatial Intelligence (7.5 vs 7.0) and Visual Intelligence (7.2 vs 6.8) scores, your Clarity of Thought Score is very much better (7.2 vs 6.3).
- 05: Compared to your peers, you have a somewhat better Clarity of Thought Score (7.2 vs 6.3), but you have only slightly higher Spatial Intelligence (7.5 vs 7.0) and Visual Intelligence (7.2 vs 6.8) scores.
- +6: On top of this you also outperformed most people in your age group with your exceptional scores for Imagination and Creativity (7.9 vs 7.2) and Logical-Mathematical Intelligence (7.1 vs. 6.5).
- 06: You did better than most people in your age group with your scores for Imagination and Creativity (7.9 vs 7.2) and Logical-Mathematical Intelligence (7.1 vs. 6.5).
- +7: There is a lot of variation in your age group, but your score is significantly higher than average.
- 07: Your score is higher than average, but there is a lot of variation in your age group.

Fig. 2. Linguistic variation used in the IQ test feedback

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). Given a positively slanted sentence (e.g. +7) from Mary and a corresponding more neutrally slanted one (e.g. 07) from Gordon, addressed to one or more pupils, participants were asked to indicate:

1. "Is it possible that Mary and Gordon might actually be (honestly) giving different feedback to the *same* pupil on the same task?"
2. "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?"

The hypothesis was that (for the 7 pairs of sentences from Figure 2) in general participants would answer "yes" to question 1 and "no" to question 2. Indeed, for 6 pairs at least 14 out of the 15 participants answered as we had predicted. For the other pair (+4/04), 12 out of 15 agreed with both predictions. We see this

as very strong evidence for our position (the participants gave different answers for the filler pairs, and so were not just producing these answers blindly).

No alterations were made to the two feedback texts on the basis of the text validation results.

5.4 Experiment

Participants

30 participants, all female university students, took part in the experiment. All participants except two were in age band 18-24. The exceptions were in age band 25-29 (group +) and 30-34 (group 0).

Materials

As stated above, 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:

These various aspects of your intelligence contribute to an overall Baumgartner Score. The Baumgartner Score rates mall The Baumgartner test which you have just undertaken tests various kinds of intelligence, for instance, your visual intelligence, your logical-mathematical intelligence and your spatial intelligence. your intelligence on a 10-point scale with 10 as the highest possible score. Note that your Baumgartner Score can change over time dependent on experience and practice. Below your test score is presented in comparison with the average score in your age group.

The introduction to the test was followed by either the positively (+1..+7, Figure 2) or the relatively neutrally (01..07, Figure 2) phrased test results.

Before and after the participants took the IQ test, they were asked to fill out a questionnaire with a PANAS test and some questions for collecting demographic information. 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 could linearly traverse through the various phases of the experiment. An outline of the set up is given in Figure 3. In the general introduction to the experiment, 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 also said that the test consisted of open and multiple choice questions that had different weight factors in the calculation of the overall score and that would assess various aspects of their intelligence. Subsequently, the participant was asked to tick a consent form to participate in the study. Then a questionnaire followed in which the participant was asked about her age, gender and the quality of her English. She was also asked if she had any experience with IQ tests and how she expected to score on this one. These questions

were interleaved with an emotion assessment test (reduced PANAS) in which the participant was asked ‘how do you feel right now?’.

After filling out the questionnaire, the participant could start the “IQ test” whenever she was ready. The “IQ test” consisted of 30 questions which she had to answer one at a time. The participant could not skip a question and also had to indicate for each of the questions how confident she was about her answer. The questions that were used for the test were carefully collected from the internet and included items from various tests and games. Different types of questions were used: questions about logical truths, mathematical questions that required some calculations, questions about words and letter sequences, questions including pictures and questions about the participant’s personality. They were ordered randomly (but with the same order for each participant).

When the participant had finished the test, she was asked to wait patiently while the system calculated the test scores. When enough calculation time had passed the participant was presented with the test feedback (one of the two texts, regardless of their actual performance). After the participant had processed the feedback, she was asked to fill out one more questionnaire to assess her emotions (i.e. ‘How do you feel right now knowing your scores on the test’). This time the simplified PANAS test was interleaved with questions about the participant’s results, (e.g. were they as expected and how did she value them), the test (e.g. was it difficult, doable or easy?) and space for comments on the test and the experiment. Finally the participant was debriefed about the experiment and about the goal of the study.

Although our particular experiment focussed on positive affect, we included the negative affect terms partly so that we could detect outliers in our participant set – people who were perhaps extremely nervous about the test or sensitive about their IQ. In fact, we did not find any such outliers.

The participants were randomly distributed over group + and group 0 and (for ethical reasons) did the test one by one in a one-person experiment room while the experimenter was waiting outside the room. As soon as the participant indicated that she had finished the task (i.e. stepped out of the experiment room), she was debriefed about the study by the experimenter and was paid with a voucher worth 5 pounds.

Hypotheses

Since the message of the feedback texts was relatively positive and there is no necessary correlation between positive and negative PANAS scores [26], we expected the main effects of the texts to be on the average evaluation of the positive PANAS terms. The hypothesis for this study was that participants who received the positively phrased feedback would show a larger change in their positive emotions than the participants who received the neutrally phrased feedback.

Results

The bar chart presented in Figure 4 illustrates the results of the PANAS questionnaire after reading the feedback texts. Table 2 indicates that on average after they had received their test results, participants in the +-group were more positively tuned than participants in the 0-group. Participants in the +-group

1. General introduction to the experiment;
2. Consent form;
3. Questionnaire on participant's background and familiarity with IQ-test interleaved with a PANAS test to assess the participant's current emotional state;
4. Message: 'Please press the next button at the bottom of this page whenever you are ready to start the intelligence test';
5. IQ test questions;
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 + or 0;
8. Questionnaire: PANAS test to assess how the participants felt after reading the test feedback interleaved with questions about the test, their expectations and space for comments;
9. Debriefing which informed participants about the study's purpose and stated that the IQ test was not real and that their test results did not contain any truth.

Fig. 3. Phases in the experiment set up

also rated the positive emotion terms higher than they had done before they undertook the IQ test. No such results were found for the 0-group. In contrast, compared to their responses before the IQ test, participants in the 0-group rated the positive terms slightly lower after they had processed their neutrally phrased feedback. With respect to the negative PANAS terms, participants in the +-group report slightly less negative emotions after they read their test scores, but none of the differences found in the negative PANAS scores were significant. Cronbach Alpha scores show that participants consistently rate the positive and negative terms before (α Positive = .797; α Negative = .883) and after (α Positive = .758; α Negative = .879) reading the feedback text.

A 2 (feedback type) * 2 (before/after) * 2 (positive/negative mean) repeated measures ANOVA was carried out on the average PANAS scores. This showed no main effect of feedback type (+ vs 0) and no main effect of before/after on average PANAS scores. However, there was a highly significant interaction between feedback type and before/after, which indicates that the change in PANAS mean before and after the text was strongly dependent on feedback type⁶ ($F(1, 28) = 10.246$, $p < .003$; power = .871). We interpret this to mean that the (after minus before) value is significantly greater for the +-group. A two-tailed, two sample t-test verifies this ($t = 3.2$, $p < 0.004$).

We did some post-hoc investigation in an attempt to understand the main result more fully. When looking at the positive PANAS scores in more detail (see Table 3), it turns out that only three of the five positive PANAS terms included in the simplified PANAS test render promising results. Interactions were found

⁶ An ANOVA test on the positive means only produces a similar result.

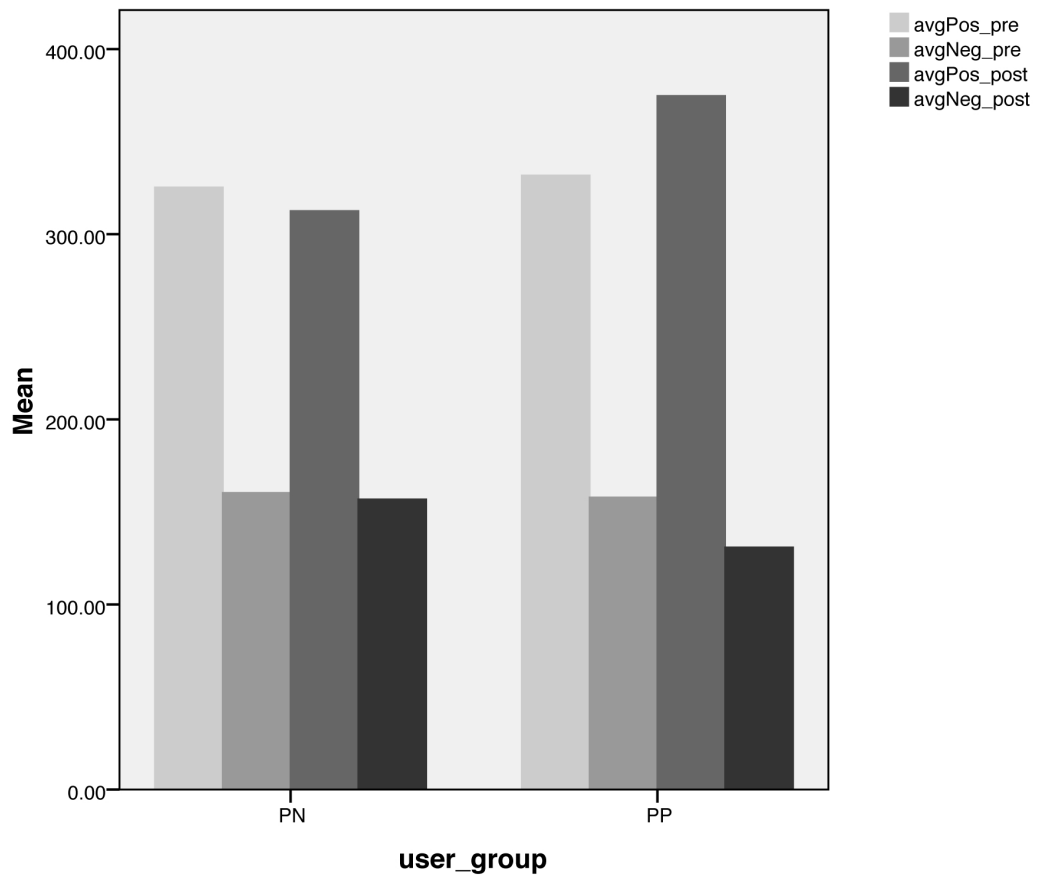


Fig. 4. Positive and negative PANAS means before and after the Participants read the feedback about their test results.

for the terms ‘alert’ ($F(1, 28) = 10.291, p < .003$) and ‘enthusiastic’ ($F(1, 28) = 5.651, p < .025$). No interactions were found for the terms ‘determined’ and ‘inspired’. For ‘inspired’ however, we found a main effect of feedback type : ($F(1, 28) = 8.755, p < .006$), which indicates that participants in the +group could have been more inspired because of their test scores than participants in the 0-group. Not all of these results would be significant if Bonferroni corrections were made.

5.5 Discussion

Compared with Study I, we expected participants to indicate stronger emotional effects, because the text participants were asked to read was about their own

	<i>0-group</i>	<i>+group</i>
Negative PANAS terms Before	1.60(.76)	1.58(.68)
Negative PANAS terms After	1.57(.68)	1.31(.45)
Positive PANAS terms Before	3.25(.78)	3.32(.55)
Positive PANAS terms After	3.13(.58)	3.75(.55)

Table 2. Means and Standard deviations (between brackets) for the negative and positive PANAS terms as indicated before and after the IQ test undertaken by participants that received neutral and participants that received positive feedback on their performance.

	<i>0-group</i>	<i>+group</i>
Alert Before	3.96(.80)	3.17(.99)
Alert After	3.45(.76)	3.65(.75)
Determined Before	3.49(1.02)	3.60(.50)
Determined After	3.50(1.13)	3.74(.61)
Enthusiastic Before	3.52(1.05)	3.49(.72)
Enthusiastic After	2.97(.81)	3.84(.66)
Excited Before	2.74(.97)	3.28(.61)
Excited After	2.64(.75)	3.69(.83)
Inspired Before	2.56(1.21)	3.06(.77)
Inspired After	3.06(1.05)	3.81(.78)

Table 3. Means and Standard deviations (between brackets) for the positive PANAS terms as indicated after the IQ test undertaken by participants that received positive and participants that received neutral feedback on their performance.

capabilities instead of about something in the world around them which they could think would not affect them. Indeed, this seems to have been the case. In Study I, all responses used the lower half of the scale, whereas with the slider our participants indicated values up to both extremes of the range available. Unfortunately, the fact that one set of values is discrete and the other continuous means that it is hard to carry out a simple statistical comparison.

6 Conclusions and Future Directions

This chapter 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.

In Study I, 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.

In Study II, we aimed to increase the reader's involvement. We used the technique of using a feedback task, where participants play a game or answer some questions after which they receive feedback on their performance. The study aimed to measure the emotional effects of slanting this feedback text in a positive or a neutral way. As in such a feedback situation the test text is directly related to the participants' own performance, we expected an increased involvement and stronger emotions.

The fact that we have been able to show a significant difference in the emotions induced by the two texts in Study II is very encouraging. It suggests that there *is* a possible methodology for directly evaluating affective NLG and that the tactical concerns with which much of NLG research is occupied are relevant to affective NLG. A similar methodology could perhaps now be used to determine the effectiveness of specific NLG methods and mechanisms in terms of inducing emotions. This is an important first step, but there is still a lot more to do. Although we have now shown that NLG tactical decisions can affect emotions, it remains to be seen what kind of changes in strategy, learning, motivation, etc., can be induced by positive or negative affect and thus how these framing decisions can best be made by an NLG system.

In Study II, a number of different techniques (e.g. emphasis, vague adjectives and adverbs) were used to phrase the various propositions in the feedback. In future work we aim to identify the relative importance of the individual techniques. We also aim to investigate more fully the algorithms that might be used by an NLG system wishing to employ these slanting techniques. In particular, techniques such as "monitoring" [11] might be useful to ensure that they are used enough, but not excessively, for a given purpose.

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 participant is affected.

Acknowledgments

This work was supported by the EPSRC platform grant ‘Affecting people with natural language’ (EP/E011764/1) and also in part by Science Foundation Ireland under a CSET grant (CNGL/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, Kees van Deemter and Nikiforos Karamanis.

References

1. Bard, E.G., Robertson, D., Sorace, A.: Magnitude estimation of linguistic acceptability. *Language* 72(1), 32–68 (1996)
2. Brown, R., Pinel, E.: Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology* 39, 626–633 (2003)
3. Cacioppo, J., Bernston, G., Larson, J., Poehlmann, K., Ito, T.: The psychophysiology of emotion. In: Lewis, M., Haviland-Jones, J. (eds.) *Handbook of Emotions*, pp. 173–191. New York: Guilford Press (2000)
4. Cadinu, M., Maass, A., Rosabianca, A., Kiesner, J.: Why do women underperform under stereotype threat? *Psychological Science* 16(7), 572–578 (2005)
5. Carenini, G., Moore, J.D.: 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*. pp. 150–157
6. E. Wilson, C.M., Campbell, L.: The information processing-approach to emotion research. In: Coan, J., Allen, J. (eds.) *Handbook of emotion elicitation and assessment*. New York: Oxford University Press (2007)
7. Finucane, M., Slovic, P., Mertz, C., Flynn, J., Satterfield, T.: Gender, race, and perceived risk: the ‘white male’ effect. *Health, Risk & Society* 2(2), 159–172 (2000)
8. Fleishman, M., Hovy, E.: Towards emotional variation in speech-based natural language generation. In: *Proceedings of the Second International Natural Language Generation Conference (INLG’02)*
9. Graff, D.: Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics* 20, 45–81 (2000)
10. O’Hara, L., Sternberg, R.: 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 (2001)
11. Hovy, E.: Pragmatics and natural language generation. *Artificial Intelligence* 43(2), 153–198 (1990)

12. Kennedy, C.: Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30, 1–45 (2007)
13. Lang, P.: Behavioral treatment and bio-behavioral assessment: Computer applications. In: Sidowske, J., Johnson, J., Williams, T. (eds.) *Technology in Mental Health Care Delivery Systems*, pp. 119–137. Norwood, NJ: Ablex (1980)
14. Lazarus, R., Kanner, A., Folkman, S.: Emotions: A cognitive-phenomenological analysis. In: Plutchik, R., Kellerman, H. (eds.) *Emotion, theory, research, and experience*, pp. 189–217. New York: Academic Press (1980)
15. Levin, I., Schneider, S., Gaeth, G.: All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behaviour and human decision processes* 76(2), 149–188 (1998)
16. Mackinnon, A., Jorm, A., Christensen, H., Korten, A., Jacomb, P., Rodgers, B.: 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)
17. Mairesse, F., Walker, M.: Trainable generation of big-five personality styles through data-driven parameter estimation. In: *Proceedings of the 46th Annual Meeting of the ACL: HLT*. pp. 165–173 (2008)
18. Moore, J., Porayska-Pomsta, K., Vargas, S., Zinn, C.: Generating tutorial feedback with affect. In: *Proceedings of the 7th International Florida Artificial Intelligence Research Symposium Conference (FLAIRS)*. pp. 123–130
19. Moxey, L., Sanford, A.: Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology* 14(3), 237–255 (2000)
20. De Rosi, F., Grasso, F.: Affective natural language generation. In: Paiva, A. (ed.) *Affective Interactions*, pp. 204–218. Springer LNAI 1814 (2000)
21. De Rosi, F., Grasso, F., Berry, D.: Refining instructional text generation after evaluation. *Artificial Intelligence in Medicine* 17(1), 1–36 (1999)
22. Russell, J., Weiss, A., Mendelsohn, G.: Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology* 57, 493–502 (1989)
23. Teigen, K., Brun, W.: Verbal probabilities: A question of frame. *Journal of Behavioral Decision Making* 16, 53–72 (2003)
24. Thompson, H.: Strategy and tactics: A model for language production. In: *Proceedings of the Chicago Linguistics Society*. Chicago (1977)
25. Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. *Science* 211, 453–458 (1981)
26. Watson, D., Clark, L.: *Manual for the Positive and Negative Affect Schedule - Expanded Form*. The University of Iowa (1999)
27. Watson, D., Clark, L., Tellegen, A.: Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology* 54, 1063–1070 (1988)