Finetuning NLG through experiments with human subjects: the case of vague descriptions

Kees van Deemter

ITRI, University of Brighton, Lewes Road, Brighton BN2 4GJ,
Kees.van.Deemter@itri.bton.ac.uk,
WWW home page: http://www.itri.brighton.ac.uk/∼Kees.van.Deemter

Abstract. This discussion paper describes a sequence of experiments with human subjects aimed at finding out how an NLG system should choose between the different forms of a gradable adjective. This case study highlights some general questions that one faces when trying to base NLG systems on empirical evidence: one question is what task to set a subject so as to obtain the most useful information about that subject, another question has to do with differences between subjects.

1 Introduction

NLG systems express information in linguistic form. As these systems become more and more sophisticated, a key problem is choosing between the different ways in which the same information can be expressed. Although intuition and general linguistic principles can go a long way, the problem is essentially an empirical one that can only be resolved by empirical study.

How can we *find out* whether one way of expressing a given input is better than another? One option is to look at corpora. But large corpora, such as the BNC, tend not to come with extensive semantic annotation, and even if they do, the domains with which the texts in the corpus are concerned tend not to have been formalised. (In NLG *jargon*: We don't know the input from which the corpus was generated.) In many cases, this is a crucial limitation, necessitating an alternative method. An alternative is to 'generate' corpora by logging human users' verbal behaviour during a controlled experiment (e.g., [2], [7]), resulting in corpora whose domains can be explicitly modelled, and where the truth value of all the relevant domain facts is known. The present paper focusses mostly on a slightly simpler method[1], by letting human subjects choose between the different ways in which the input information may be expressed. Rather than producing a corpus of extended discourse/dialogue, such experiments can tell us, at least in principle, which expression is favoured by subjects: the fact that explicit judgements are

[1] An partial exception is our third experiment, described in section 5, which could be seen as generating a corpus.

obtained from subjects might be even thought an advantage. We will focus on issues raised by previous work on referring expressions, focussing on the question how one might inform the generator's choice between the different forms of a gradable adjective in descriptions such as

a. – *'the fat chihuahua'* [base form]
b. – *'the fatter chihuahua'* [comparative form]
c. – *'the fattest chihuahua'* [superlative form]

After discussion of the experiments that we have done to inform this choice, we will discuss two problems: one problem has to do with experimental methodology (e.g., how does one find out what constitutes the optimal use of language), another problem has to do with differences between human subjects' preferences.

2 Generating vague descriptions; research hypotheses

It was reported in [12] how the module of an NLG system that generates referring expressions can use numerical information in the input to the generator to produce referring expressions like a-c above. Generating such *vague descriptions* is nontrivial partly because properties like 'fat', 'fatter', 'fattest' are not simply true or false of an animal: the fattest chihuahua is not necessarily the fattest dog, for example, or the fattest animal. Unlike the numerical properties that measure the animals' waistelines, properties like these are context dependent, and this has to be taken into account by the generation algorithm. If, for example, the knowledge base were to simply list *all* hippos and *no* chihuahuas as fat, then this would make it impossible to ever refer to a dog as 'the fat chihuahua', and this could even mean that the referent in question cannot be identified uniquely at all (i.e., if there happens to be no combination of other properties that distinguishes it from everything else in the domain). The problems surrounding gradable adjectives have been studied widely (e.g., [3], [10], [4]), but little of this is directly applicable to NLG, let alone to the choice between the different forms of the gradable adjective in descriptions such as a-c above.

In addition to presenting an algorithm for generating vague descriptions, [12] also described an informal experiment that supported the idea that for large classes of adjectives[2], the three forms a,b,c (above) of the vague description are *semantically equivalent* in the sense that, in those cases where they are used referentially, they are often interpreted as having the same reference.[3] The experiment involved numbers; the outcome suggested strongly that, normally, 'the

[2] *Evaluative* adjectives, such as 'brilliant' or 'stupid' are excluded [1].

[3] Notable exceptions arise when the description refers anaphorically or in combination with other gradable adjectives. For example, if an entity has been introduced as 'a large dog' then this legitimises later references to it as 'the large dog' even where other, larger dogs are in evidence. The second type of exception arises when, for example, 'the large dog' is juxtaposed with another vague expression such as 'an even larger dog'. For background and elaboration, see [13].

large number' is the same as 'the larger number', which is the same as 'the largest number'. What this experiment did *not* tell us is how to choose between a-c: for all we know, there might be situations in which the use of the superlative, for example, would be highly unnatural (even though it would tend to be interpreted correctly, as the experiment suggested). It is questions of the latter kind that the present paper addresses.

Intuition, supported by pilot experiments and informal study of corpora, suggested to us that there *are* situations where the base form is preferred over the superlative. We will henceforth refer to the difference in the relevant dimension between the intended referent and the nearest relevant object (the nearest 'distractor' removed by the adjective) as the *gap*. This gap itself can have different sizes. Roughly in line with Gricean principles, we hypothesised that when the gap is large, the base form is preferred; in these cases it would be unnecessarily elaborate to say 'the largest number'. Conversely, when the gap is small, we expected the superlative to be preferred. (In such cases, to use the base form sounds like an exaggeration.)

Hypotheses:
1. Small gap $\Rightarrow$ Superlative > Base
2. Large gap $\Rightarrow$ Base > Superlative

(*How* large a gap has to be to switch from superlative to base form would have to be assessed separately.) Our first experiment did not take the comparative form into account, but this was added in later experiments. All experiments asked whether subjects considered themselves native speakers of English, fluent, or none of the two; no major differences between the three groups were found.

3 First experiment: picking a numeral

25 academics at the University of Brighton were shown pairs of numerals, one of which appeared in brackets. We chose numerals because numbers come in all possible 'sizes', so when assessing the size or hight of a numeral, the only sensible comparison appears to be with other numerals that the experiment has presented. (Previously seen numerals can be ignored.) The gap between the two numbers was either large (in this case: gap size 7) or small (gap size 1). Eight different patterns involving the numbers $1, 2, 8, 9$ were offered, depending on whether the gap was large or small, in both possible orders, and involving numbers that are somewhat higher (2 and 9 when the gap is large; 8 and 9 when the gap is small) or lower (1 and 8 when the gap is large; 1 and 2 when the gap is small). They were asked

> *Which of the following statements offers the most natural description of this sequence?*
> – *'the large number appears in brackets'*, or
> – *'the largest number appears in brackets'?*

Thus, when confronted with the patterns [2 (9)], [(9) 2], [1 (8)], or [(8) 1], we expected to find that more subjects prefer the base form (because the gap is large); when confronted with the patterns [8 (9)], [(9) 8], [2 (1)], or [(1) 2], we expected to find that more subjects prefer the superlative.

Outcomes: The number of choices for each of the different types of adjective can be summarised in a contingency table:

	Base	Superlative
Large Gap	25	75
Small Gap	1	99

The table shows the gap size and adjective type to be non-independent of each other (Chi-square = 25.464 with *df* = 1 at p = 0.001.) More specifically, the experiment overwhelmingly supported Hypothesis 1. (Only one of the 25 subjects preferred the base form, and only on one occasion.) Equally clearly, however, the experiment failed to confirm Hypothesis 2: in fact, the superlative was chosen in the majority of cases even where the gap was large. *Post hoc*, the experiment suggests that base-form adjectives occur more often when the gap is large than when it is small. (Using the t-test, for example, this is highly significant, with p=.008361 and df=24.) This pattern was confirmed by later experiments.

Note that there was a striking difference between two groups of subjects: all the people in the IT faculty, except one, consistently chose the superlative, while people at the School of Languages showed more variety, including some who chose the base form *if and only if* the gap was large. We will return to the issue of differences between groups of subjects in section 7.2.

4 Second experiment: picking a number from a sequence

Confronted with this failure to confirm Hypothesis 2, it is not difficult to find excuses: maybe the gap between the numbers in the experiment had not been large enough, or maybe the question that was asked of the subjects (containing a possibly theory-laden expression like ‘natural’) was unclear; also, subjects were not offered the possibility of choosing the comparative form (‘the larger number’), which is sometimes thought to be preferable when there are only two things to compare. To remove these obstacles, we did the following experiment, which varies on the same numerical theme.

Fourteen subjects (participants at the 2001 EACL conference in Toulouse) were shown sequences of two, three or four numbers, one or two of which appeared in brackets. There were 18 patterns, including, for example, [1 (59)] (large gap), [1 (59) (58)] (large gap), [55 (59) 54] (small gap), [(59) 55 (59) 55] (small gap), and [(59) 1 (59) 0] (large gap). Subjects were asked, in each of these cases, to say *Which description do you consider most likely to have been produced by a native speaker?*, where they could choose between

– *'the large numbers appear in brackets'* or
– *'the larger numbers appear in brackets'* or
– *'the largest numbers appear in brackets'*?

Based on the previous experiment, we refined our hypotheses. Given this subjects were offered a choice between three options, it seemed reasonable to compare the numbers of base forms with each of the two other forms separately:

Hypotheses:
1a. Small gap ⇒ Comparative > Base
1b. Small gap ⇒ Superlative > Base
2a. Large gap ⇒ Base > Comparative
2b. Large gap ⇒ Base > Superlative

Outcomes: The data can be summarised as follows:

	Base	Comparative	Superlative
Large Gap	43	37	42
Small Gap	2	65	59

Chi-square suggests a dependency between gap size and adjective type (at $p = 0.001$, $df = 2$, Chi-square = 47.851). Using a t-test, we see that, in line with the first experiment, Hypotheses 1a and 1b were confirmed but, crucially, Hypotheses 2a and 2b were not. The following table lists the number of choices for each of the three types of adjectives, for each of those 9 patterns of numbers where the gap was large. (The gap was small in patterns `b,c,d,e,h,i,k,m,q`.)[4]

[Large Gap]	Base	Comparative	Superlative
Pattern a	3	8	3
Pattern f	4	1	9
Pattern g	3	3	8
Pattern j	6	5	3
Pattern l	4	6	4
Pattern n	5	5	3
Pattern o	4	2	7
Pattern p	7	2	5
Pattern r	7	5	2

Focussing on Hypothesis 2a, we find that the two-tailed P value is not significant at 0.4774 ($t = 0.7276, df = 16$). Hypothesis 2b is even further off the mark with $P = 0.9120$ ($t = 0.1123$). As before, there were substantial differences between subjects, including some who used the base form *if and only if* the gap was small. Also as before, the base form was used much more often when the gap was large than when the gap was small.

[4] The numbers for patterns `n` and `o` do not add up to 14 because one subject had failed to enter a clearly legible choice.

5 Third experiment: describing triangles

Going by the two experiments reported so far, it appears that superlatives are preferred, even in situations where the gap is very large. Again, one can think of excuses. *Asking* subjects about their preferences may not be the best way of assessing them: perhaps speakers cannot always tell how they use language (just like most of us cannot say how they walk). Also, numbers have their peculiarities: contrary to what was assumed in section 2, one might argue that some numbers (the number 0, for example, given that we only used non-negative numbers) *are* intrinsically small. For these reasons, we shifted the subject matter from numbers to geometrical shapes, and we designed the experiment in such a way that subjects could produce their own referring expressions (instead of being offered a forced choice). 34 Subjects (students of an HCI module at the University of Brighton) were shown a piece of paper showing two isosceles, and approximately equilateral triangles,[5] one of which was circled. To encourage the use of size-related descriptions (like 'the large triangle', 'the biggest of the two figures', but unlike 'the figure on the top left'), the instructions asked subjects to imagine themselves on the phone to someone holding a copy of the same sheet of paper, but not necessarily with the same orientation (e.g., possibly upside down). To discourage lengthy descriptions, the space for text was limited:

```
Q: Which triangle on this sheet
   was circled?
A: The ............ triangle
```

Each of the 34 subjects saw six pairs of triangles, one of which was smaller than the other. Triangles came in three sizes, so the 'gap' between the two triangles could either be large (when the smallest of the three was juxtaposed to the largest of the three) or small (when the mid-size triangle was juxtaposed to either the smallest one or the largest one). The same hypotheses were used as in the previous experiment (see section 4).

Results: Rather pleasingly, all except one subject used brief, size-related descriptions throughout. There were a few occurrences of descriptions like 'the top triangle', 'the bottom triangle', 'the tiny triangle', yet the data allow a remarkably simple summary:

12 subjects chose **superlatives** throughout (i.e., 6 items)
6 subjects chose **comparatives** throughout
3 subjects chose **Base form** throughout
6 subjects oscillated between Superlative and Comparative
6 subjects oscillated between Base form and some other form

[5] Only two triangles were shown at a time, since the use of larger sequences in the second experiment had payed no obvious benefit. The lengths of the bases of the triangles were $5mm, 8mm$, and $16mm$ respectively.

As before, Hypotheses 2a and 2b were not confirmed (while Hypotheses 1a and 1b were confimed). Once again, these findings appear to suggest that superlatives are preferred regardless of circumstance.

6 Fourth experiment: recognising human behaviour

It seemed plausible that the fact of being under scrutiny (being tested!) might have encouraged 'pedantic' behaviour in our subjects, causing this unexpected run of results. We therefore made one more attempt at getting at subjects' normal behaviour, using a procedure reminiscent of the Turing test, which involves letting subjects judge whether a given pattern of behaviour is likely to have been produced by a real person as opposed to, for example, a computer or a linguist [11]. We showed results (or, in one case, fake results) from the previous experiment to 14 students of the School of Languages, at the University of Brighton. (Each subject in the third experiment gave answers to six answers to questions of the form *Which triangle (...)?*, so each subject produced a sequence of six answers, e.g., `smallest large smaller smaller larger larger`.) We told our new subjects that each output was produced by one of our subjects (**a**,**b**,**c**, and **d**) in an earlier experiment:

a: `smallest largest smallest smallest largest largest`
b: `smaller larger smaller smaller larger larger`
c: `small large smaller smaller larger larger`
d: `smaller larger small small large large`

(Subjects were also shown the geometrical patterns on which the four sets of judgments were based.) Subjects **a** and **b** show completely uniform output; **c** mixes base forms and comparatives in a manner consistent with our Hypotheses; **d** (constructed by hand, because no such output had been produced by a real subject) reverses the patterns produced by **c**, using a base form where **c** had produced a comparative and conversely. We asked:

> *Which of the four sequences of responses is* most likely *to have been produced by a real person, in your opinion? Circle* one *of the options: Subject* **a, b, c, d.**

In line with our earlier expectations, we hypothesised that **c** would be favoured over all others: that (i) **c** > **a**, that (ii) **c** > **b**, and that (iii) **c** > **d**

Results. The data are as follows:

- 9 subjects thought **b** was most likely
- 4 subjects thought **c** was most likely
- 1 subject thought **a** was most likely
- 0 subjects thought **d** was most likely

On the positive side, this suggests that a generator programmed to behave in accordance with our Hypotheses (leading to output as in **c**) would be judged to be quite natural. But the more striking conclusion is surely that subjects rated completely uniform behaviour even more highly. The number of subjects partaking in this fourth experiment may be too small to warrant firm conclusions but the experiment clearly does little to confirm Hypothesis (ii).

7 Discussion

Two themes emerge: the question how to set up experiments in such a way that their outcomes reflect the linguistic preferences of the subjects; and the question how to cope with differences between subjects.

7.1 Eliciting 'natural' use of language

Our experiments suggest that vague descriptions in the base form are almost always dispreferred in comparison with comparatives and/or superlatives. To us at least, this seems counterintuitive.[6] Otherwise, the experiments diverge. For example, superlatives came out on top of the third experiment, but were trumped by comparatives in the fourth. Note that such surprises are not easily explained along the lines of [5]. There, Oberlander invokes an asymmetry between speakers' inclinations and hearer's expectations, because speakers may be subject to factors that hearers cannot be aware of before the utterance is made (e.g., its propositional content). Our experiments, however, appear to make subjects aware of all the factors that writers/speakers would be aware of. (More specifically, the third and fourth experiments gave their subjects exactly the same information about the triangles and their properties.)

It is, of course, well known that questionnaire-type experiments are sensitive to subtle differences in formulations, and it is possible that our experiments have failed to capture spontaneous language use. One might therefore insist on naturalistic settings where subjects are given immersive real-world tasks to make them less self-conscious, and experiments of this kind (involving the narration of bedtime stories for children) are currently being prepared. The analysis of such experiments (involving unconstrained language use) tends to be difficult and time consuming, however. More importantly, one can always doubt the validity of the experiment by asking whether the task was immersive *enough*.

Perhaps most worryingly, our experiences cast doubt on previous findings: if questions of the form *Which description do you consider most likely to have been produced by a native speaker?* lead to outcomes that misrepresent writers/speakers' inclinations, then how do we know that the answers to other,

[6] As was explained in section 1, it would be no simple matter to test our hypotheses using a corpus like the BNC: even though base-form descriptions *are* much more frequent there than the two other types of descriptions, it would be no trivial matter to filter out attributive, anaphoric, and other uses which, collectively, are much more frequent than the referring expressions on which this paper has focussed.

similar questions that psycholinguists ask (e.g., 'How would you continue this sentence?' [9]; similarly for other types of questions relating to generation of referring expressions, e.g., [6]) are safe?

7.2 Honouring style differences

In recent work, Reiter and Sripada have argued that corpora can be a slippery basis on which to base the decisions of a generator [8]. Our own findings suggest a similar conclusion in connection with controlled experiments: one ignores style differences at one's peril. The underlying point is a simple mathematical one: Suppose a given input can be expressed in three ways: A, B, and C, where there are two different sub-styles, say **1** and **2**. Suppose expression A is most frequent in none of the two styles, while occupying a very solid second place in both:

Substyle **1**: $B > A >> C$
Substyle **2**: $C > A >> B$
Overall: $A > B > C$

If the construction of a generator was based on overall frequencies, it would express the input as A, the most frequent expression overall. But this would cause the generator to be a poor imitator of anyone's language use, since no existing substyle would be modelled. It would be better to choose either B or C since, at least, this would capture what is most frequent in one substyle.

8 Conclusion

When gathering data to inform natural language generation, controlled experiments are a natural complement to the study of corpora. This discussion paper has used a sequence of experiments to highlight two types of problems with the experimental approach, one of which (section 7.1) is specific to the experimental approach, while the other echoes similar problems in the study of language corpora (section 7.2).

9 Acknowledgments

Support for the TUNA project from the EPSRC under grant number GR/S13330 is gratefully acknowledged. Further thanks are due to Paul Piwek and to my TUNA colleagues: Albert Gatt, Emiel Krahmer, Richard Power, and Sebastian Varges, all of whom have contributed to this paper in some way or other.

10 References

1. Bierwisch, M.: The semantics of gradation. In M.Bierwisch and E.Lang (Eds.) *Dimensional Adjectives.* Berlin, Springer Verlag, (1989) pp.71-261.

2. Jordan, P.W.: Can Nominal Expressions Achieve Multiple Goals?: An Empirical Study. Procs. of ACL-2000, Hong Kong (2000).

3. Kamp, H.: Two theories about adjectives. In “Semantics for natural language”, ed. E. Keenan. Cambridge University Press (1975).

4. Klein E.: A semantics for positive and comparative adjectives. *Linguistics and Philosophy* **4** (1980).

5. Oberlander, J.: Do the right thing ... but expect the unexpected. Computational Linguistics, **24** (3), (1998) 501-507.

6. Pechmann, Th.: Incremental Speech Production and Referential Overspecification. *Linguistics* **27** (1989) 98-110.

7. Piwek, P. Cremers, A.: Dutch and English Demonstratives: A Comparison. Language Sciences **18** (3-4), (1996) pp. 835-851.

8. Reiter, E. Sripada, S.: Should Corpora Texts Be Gold Standards for NLG? In Proceedings of INLG-02 (2002) pages 97-104.

9. Stevenson, R.J., Crawley, R.A., and Kleinman, D.: Thematic roles, focus and the representation of events. *Language and Cognitive Processes* **9** (1994) 519-548.

10. Synthese. Special issue of the journal *Synthese* on semantic vagueness. *Synthese* **30** (1975).

11. Turing, A.: Computing Machinery and Intelligence. *Mind* **59**, No. 236 (1950) pp. 433-460.

12. van Deemter, K.: Generating Vague Descriptions. In Procs. of Int. Conf. on Natural Language Generation (INLG-2000) (2000) Mitzpe Ramon.

13. van Deemter, K.: Computational Generation of Vague Descriptions. Manuscript, ITRI, University of Brighton.