Using a Randomised Controlled Clinical Trial to Evaluate an NLG System

Ehud Reiter† **Roma Robertson**‡ **A Scott Lennox**‡ **Liesl Osman**§
Departments of Computing Science†, General Practice‡, and Medicine and Therapeutics§
University of Aberdeen, Aberdeen, Scotland, UK
{e.reiter, roma.robertson, s.lennox, l.osman}@abdn.ac.uk

Abstract

The STOP system, which generates personalised smoking-cessation letters, was evaluated by a randomised controlled clinical trial. We believe this is the largest and perhaps most rigorous task effectiveness evaluation ever performed on an NLG system. The detailed results of the clinical trial have been presented elsewhere, in the medical literature. In this paper we discuss the clinical trial itself: its structure and cost, what we did and did not learn from it (especially considering that the trial showed that STOP was not effective), and how it compares to other NLG evaluation techniques.

1 Introduction

There is increasing interest in techniques for evaluating Natural Language Generation (NLG) systems. However, we are not aware of any previously reported evaluations of NLG systems which have rigorously compared the task effectiveness of an NLG system to a non-NLG alternative. In this paper we discuss such an evaluation, a large scale (2553 subjects) randomised controlled clinical trial which evaluated the effectiveness of personalised smoking-cessation letters generated by the STOP system (Reiter et al., 1999). We believe that this is the largest, most expensive, and perhaps most rigorous evaluation ever done of an NLG system; it was also a disappointing evaluation, as it showed that STOP letters in general were no more effective than control letters.

The detailed results of the STOP evaluation have been presented elsewhere, in the medical literature (Lennox et al., 2001). The purpose of this paper is to discuss the clinical trial from an NLG evaluation perspective, in order to help future researchers decide when a clinical trial (or similar large-scale task effectiveness evaluation) would be an appropriate way to evaluate their systems.

2 Evaluation of NLG Systems

Evaluation is becoming increasingly important in NLG, as in other areas of NLP; see Mellish and Dale (1998) for a summary of NLG evaluation. As Mellish and Dale point out, we can evaluate the effectiveness of underlying theories, general properties of NLG systems and texts (such as computational speed, or text understandability), or the effectiveness of the generated texts in an actual task or application context. Theory evaluations are typically done by comparing predictions of a theory to what is observed in a human-authored corpus (for example, (Yeh and Mellish, 1997)). Evaluations of text properties are typically done by asking human judges to rate the quality of generated texts (for example, (Lester and Porter, 1997)); sometimes human-authored texts are included in the rated set (without judges knowing which texts are human-authored) to provide a baseline. Task evaluations (for example, (Young, 1999)) are typically done by showing human subjects different texts, and measuring differences in an outcome variable, such as success in performing a task.

However, despite the above work, we are not aware of any previous evaluation which has compared the effectiveness of NLG texts at meeting a communicative goal against the effectiveness of non-NLG control texts. Young's task evaluation, which may be the most rigorous previous task evaluation of an NLG system, compared

the effectiveness of texts generated by different NLG algorithms, while the IDAS task evaluation (Levine and Mellish, 1995) did not include a control text of any kind. Coch (1996) and Lester and Porter (1997) have compared NLG texts to human-written and (in Coch's case) mail-merge texts, but the comparisons were judgements by human domain experts, they did not measure the actual impact of the texts on users. Carenini and Moore (2000) probably came closest to a controlled evaluation of NLG vs non-NLG alternatives, because they compared the impact of NLG argumentative texts to a no-text control (where users had access to the underlying data but were not given any texts arguing for a particular choice).

Task evaluations that compare the effectiveness of texts from NLG systems to the effectiveness of non-NLG alternatives (mail-merge texts, human-written texts, or fixed texts) are expensive and difficult to organise, but we believe they are essential to the progress of NLG, both scientifically and technologically. In this paper we describe such an evaluation which we performed on the STOP system. The evaluation was indeed expensive and time-consuming, and ultimately was disappointing in that it suggested STOP texts were no more effective than control texts, but we believe that this kind of evaluation was essential to the project. We hope that our description of the STOP clinical trial and what we learned from it will encourage other researchers to consider performing effectiveness evaluations of NLG systems against non-NLG alternatives.

3 STOP and its Clinical Trial

The STOP system has been described elsewhere (Reiter et al., 1999). Very briefly, the system took as input a 4-page questionnaire about smoking history, habits, intentions, and so forth, and from this produced a small (4 pages of A5) personalised smoking cessation letter. All interactions with the smoker were paper-based; he or she filled out a paper questionnaire which was scanned into the computer system, and the resultant letter was printed out and posted back to the smoker. The first page of a typical questionnaire is shown in Figure 1, and part of the letter produced from this questionnaire is shown in Figure 2.[1] We wish to emphasise that producing personalised health information letters is not a new idea, many previous researchers have worked in this area; see Lennox *et al* (2001) for a comparison of STOP to previous work in this area.

The STOP clinical trial, which is the focus of this paper, was organised as follows. We contacted 7427 smokers, and asked them to participate in the trial. 2553 smokers agreed to participate, and filled out our smoking questionnaire. These smokers were randomly split among three groups:

- *Tailored.* These smokers received the letter generated by STOP from their questionnaire.
- *Non-tailored.* These smokers received a fixed (non-tailored) letter. The non-tailored letter was essentially the letter produced by STOP from a blank questionnaire, with some manual post-editing and tidying up. In other words, during the course of developing STOP we created a set of default rules for handling incomplete or inconsistent questionnaires; the non-tailored letter was produced by activating these default rules without any smoker data. Part of the non-tailored letter is shown in Figure 3.
- *No-letter.* These smokers just received a letter thanking them for participating in our study.

After six months we sent a followup questionnaire asking participants if they had quit, and also other questions (for example, if they were intending to try to quit even if they had not actually done so yet). Smokers could also make free-text comments about the letter they received. 2045 smokers responded to the followup questionnaire, of which 154 claimed to have quit. Because people do not always tell the truth about their smoking habits, we asked these 154 people to give saliva samples, which were tested in a lab for nicotine residues. 99 smokers agreed to give such samples, and 89 of these were confirmed as non-smokers.

[1] To protect patient confidentiality, we have changed the name of the smoker and her medical practice, and typed her handwritten responses.

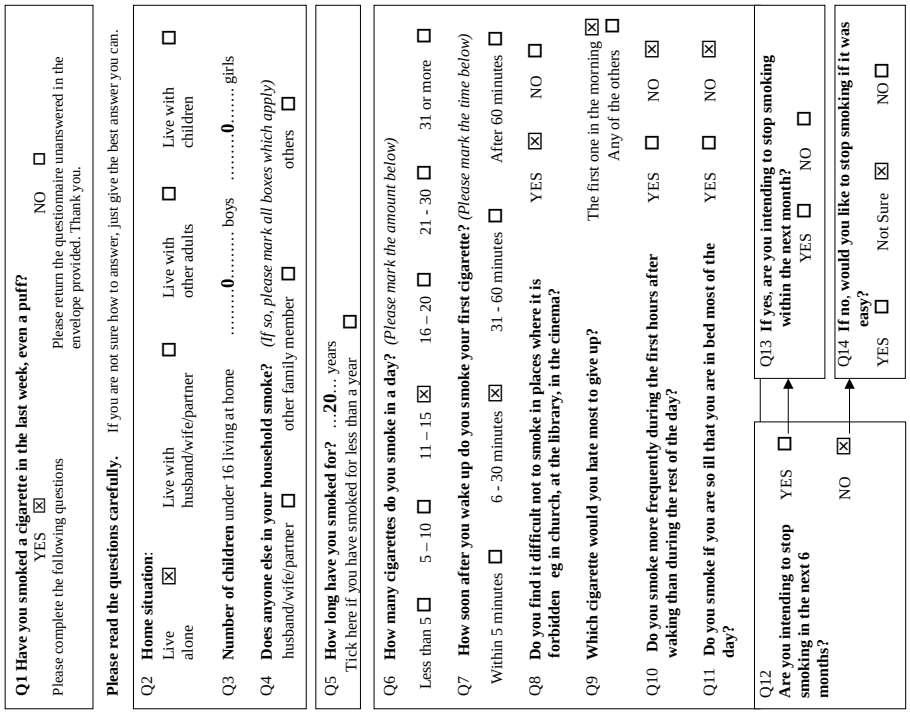


Figure 1: First page of a STOP questionnaire

3.1 Practical Aspects of the Clinical Trial

The STOP clinical trial took 20 months to run (of which the first 4 months overlapped software development), and cost about UK£75,000 (US$110,000). We believe the STOP clinical trial was the longest and costliest evaluation ever done of an NLG system. The length and cost of the clinical trial were primarily due to the large numbers of subjects. Whereas Levine and Mellish (1995), Young (1999), and Carenini and Moore (2000) included 10, 26, and 30 subjects (respectively) in their task effectiveness evaluations, we had 2553 subjects in our clinical trial. The cost of the trial was partially stationary and postage (we sent out over 10000 mailings to smokers, each of which included a reply-paid envelope), but mostly staff costs to set up the trial, perform the mailings, process and analyse the returns from smokers, and handle various glitches in the trial.

Another way of looking at the trial was that we spent about UK£30 (US$45) per subject (including staff time as well as materials). Perhaps the trial could have been done a bit more cheaply, but any experiment involving 2553 subjects is bound to be expensive and time-consuming.

The reason the trial needed to be so large was that we were measuring a binary outcome variable (laboratory-verified smoking cessation) with a very low positive rate (since smoking is a very difficult habit to quit). Young, in contrast, measured numerical variables (such as the number of mistakes made by a user when following textual instructions) with substantial standard deviations.

Another complication was that we wanted to use a representative sample of smokers in our trial, which meant that we could not (as Young and Levine and Mellish did) just recruit students and acquaintances. Instead, we contacted a representative set of GPs in our area, and asked them for a list of smokers from their patient record systems. This was the source of the 7427 initial smokers mentioned above.

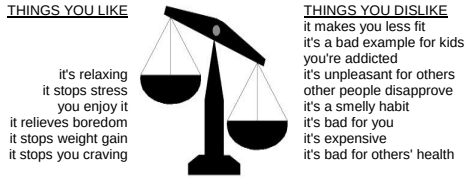
4 Results of the Clinical Trial

Detailed results of the STOP clinical trial, including statistical tables, have been published in the medical literature (Lennox et al., 2001). Here we just summarise the key findings which are of NLG

Smoking Information for Heather Stewart

You have good reasons to stop...

People stop smoking when they really want to stop. It is encouraging that you have many good reasons for stopping. The scales show the good and bad things about smoking for you. They are tipped in your favour.



You could do it...

Most people who really want to stop eventually succeed. In fact, 10 million people in Britain have stopped smoking - and stayed stopped - in the last 15 years. Many of them found it much easier than they expected.

Although you don't feel confident that you would be able to stop if you were to try, you have several things in your favour.

- You have stopped before for more than a month.
- You have good reasons for stopping smoking.
- You expect support from your family, your friends, and your workmates.

We know that all of these make it more likely that you will be able to stop. Most people who stop smoking for good have more than one attempt.

Overcoming your barriers to stopping...

You said in your questionnaire that you might find it difficult to stop because smoking helps you cope with *stress*. Many people think that cigarettes help them cope with stress. However, taking a cigarette only makes you feel better for a short while. Most ex-smokers feel calmer and more in control than they did when they were smoking. There are some ideas about coping with stress on the back page of this leaflet.

You also said that you might find it difficult to stop because you would *put on weight*. A few people do put on some weight. If you did stop smoking, your appetite would improve and you would taste your food much better. Because of this it would be wise to plan in advance so that you're not reaching for the biscuit tin all the time. Remember that putting on weight is an overeating problem, not a no-smoking one. You can tackle it later with diet and exercise.

And finally...

We hope this letter will help you feel more confident about giving up cigarettes. If you have a go, you have a real chance of succeeding.

With best wishes,

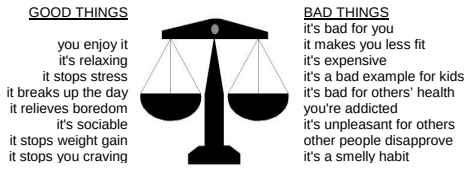
The Health Centre.

Figure 2: Inside pages of the STOP letter generated from the Figure 1 questionnaire

Information for Stopping Smoking

Do you want to stop smoking?

Everyone has things they like and dislike about their smoking. The decision to stop smoking depends on the things you don't like being more important than the things you do like. It can be useful to think of it as a balance. Have a look on the scales. What are the good and bad things for you?



Add any more that you can think of. Are you ready to stop smoking? If yes, maybe it's the right time to have a go. If no, think about the good and bad things about smoking. This might swing the balance for you.

You can do it.....

People who want to stop smoking usually succeed. 10 million people in Britain have stopped smoking - and stayed stopped - in the last 15 years. Many of them found it much easier than they expected!

Try it out.....

If you don't feel ready for an all-out attempt to stop smoking, there are some useful ways to prepare yourself. You could try some of the following ideas now. This will help you when you try to stop smoking.

- Delay your first cigarette of the day by half an hour.
- Stop smoking for 24 hours.
- Cut down the number you smoke by 5 cigarettes per day.

Planning will help.....

When you stop, it helps to plan ahead. Here are some things that have worked for others:

- Pick a day to stop, and let your family and friends know.
- Think of situations where you might feel tempted to smoke, and plan how you could avoid or deal with them.
- Get rid of all cigarettes and ashtrays the day before.
- When you do stop, take one day at a time; don't look too far ahead.

If it gets tough.....

Many people do hit rough patches; there are ways to deal with these. On the back page are some suggestions that other people have found useful.

If you do have a cigarette after a few days just put it behind you and keep on trying. Prepare yourself for another attempt, many people have more than one go before they stop for good!

With best wishes.

The Health Centre.

Figure 3: Inside pages of the non-tailored letter

(as well as medical) interest.

Of the 2553 smokers in the trial, 89 were validated as having stopped smoking. These broke down by group as follows:

- 3.5% (30 out of 857) of the tailored group stopped smoking
- 4.4% (37 out of 846) of the non-tailored group stopped smoking
- 2.6% (22 out of 850) of the no-letter group stopped smoking

The non-tailored group had the lowest number of heavy (more than 20 cigarettes per day) smokers, who are less likely to stop smoking (because they are probably addicted to nicotine) than light smokers; the tailored group had the highest number of heavy smokers. After adjusting for this fact, cessation rates were still higher in the non-tailored group than in the tailored group, but this difference was not statistically significant. We can see this if we look just at cessation rates in light smokers (few heavy smokers from any category managed to stop smoking):

- 4.3% (25 out of 563) of the light smokers in the tailored group stopped smoking
- 4.9% (31 out of 597) of the light smokers in the non-tailored group stopped smoking
- 2.7% (16 out of 582) of the light smokers in the no-letter group stopped smoking

The overall conclusion is therefore that recipients of the non-tailored letters were more likely to stop than people who got no letter[2] (p=.047 overall unadjusted; p=.069 overall after adjusting for differences between groups, such as heavy/light smoker split; p=.049 for light smokers). However, there was no evidence that the tailored letters were any better than the non-tailored ones in terms of increasing cessation rates.

[2]Note that while a 1% or 2% increase in cessation rates is small, it is medically useful if it can be achieved cheaply. See Law and Tang (1995) for a discussion of success rates and cost-effectiveness of various smoking-cessation techniques, and Lennox *et al* (2001) for an analysis that shows that sending letters is very cost-effective compared to most other smoking-cessation techniques.

There is some very weak evidence that the tailored letter may have been better than the non-tailored letter among smokers for whom quitting was especially difficult. For example, among discouraged smokers (people who wanted to quit but were not intending to quit, usually because they didn't think they could quit), cessation rates were 60% higher among recipients of tailored letters than recipients of non-tailored letters, but the numbers were too small to reach statistical significance, since (as with heavy smokers) very few such people managed to stop smoking. Furthermore, among heavy smokers, recipients of the tailored letter were 50% more likely than recipients of the non-tailored letters to show increased intention to quit (for example, say in their initial questionnaire that they did not intend to quit, but say in the followup questionnaire that they did intend to quit) (p=.059). It would be nice to test the hypothesis that tailored letters were effective among discouraged smokers or heavy smokers by running another clinical trial, but such a trial would need to be even bigger and more expensive than the STOP trial, in order to have enough validated quitters from these categories to make it possible to draw statistically significant conclusions.

Recipients of the tailored letters were more likely than recipients of non-tailored letters to remember receiving the letter (67% vs 44%, significant at $p<.01$), to have kept the letter (30% vs 19%, significant at $p<.01$), and to make a free-text comment about the letter (20% vs 12%, significant at $p<.01$). However, there was no statistically significant difference in perceptions of the usefulness and relevance of the tailored and non-tailored letters.

Free-text comments on the tailored letters were varied, ranging from *I carried mine with me all the time and looked at it whenever I felt like giving in* to *I found it patronising . . . Smoking obviously impairs my physical health — not my intelligence!* The most common complaint about content was that not enough information was given about practical 'how-to-stop-smoking' techniques. STOP's tailoring rules only included such information in about one third of the letters; this was in accordance with the well-established Stages of Change model of smoking cessation (Prochaska and diClemente, 1992). Note that all

recipients of the non-tailored letter received such information. If practical advice was useful to more than one third of smokers, then the Stages-of-Change based tailoring rules which decided when to include such information may have decreased rather than increased letter effectiveness.

5 What Can be Learned from a Negative Result

One of the remarkable things about the NLG, NLP, and indeed AI literatures is that little mention is made of experiments with negative results. In more established fields such as medicine and physics, papers which report negative experimental findings are common and are valued; but in NLP they are rare. It seems unlikely that NLP experiments always produce positive results (unless the experiments are badly designed and biased towards demonstrating the experimenter's desired outcome); what is probably happening is that people are choosing not to report negative results.

One reason for this may be that it can be difficult to draw clear lessons from a negative result. In the case of STOP, for example, the clinical trial did not tell us why STOP failed. There are many possible reasons for the negative result, including:

1. Tailoring cannot have much effect. That is, if a smoker receives a letter from his/her doctor about smoking, then the content of the letter is only of secondary importance, the important thing is the fact of having received a communication from his/her doctor encouraging smoking cessation.

2. Tailoring could have an impact, but only if it was based on much more knowledge about the smoker's circumstances than is available via a 4-page multiple choice questionnaire.

3. Tailoring based on a multiple-choice questionnaire can work, we just didn't do it right in STOP, perhaps in part because we based our system on inappropriate theoretical models of smoking cessation.

4. The STOP letters did in fact have an effect on some groups (such as heavy or discouraged smokers), but the clinical trial was too small to provide statistically significant evidence of this.

In other words, did we fail because (1) what we were attempting could not work; (2) what we were attempting could only work if we had a lot more knowledge available to us; or (3) we built a poor system? Or (4) did the system actually work to some degree, but the evaluation didn't show this because it was too small? This is a key question for NLG researchers and developers (as opposed to doctors and health administrators who just want to know if they should use STOP as a black-box system), but the clinical trial does not distinguish between these possibilities.

Arguments can be made for all three of the above possibilities. For example, we could argue for (1) on the basis that brief discussions about smoking with a doctor have about a 2% success rate (Law and Tang, 1995), and this may be an upper limit for the effectiveness of a brief letter from a doctor. If so, then letters cannot do much better that the 1.8% increase in cessation rates produced by the STOP non-tailored letter. Or we could argue for (2) by noting that when we asked smokers to comment on STOP letters in a small pilot study, many of their comments were very specific to their particular circumstances For example, a single mother mentioned that a previous attempt to stop failed because of stress caused by dealing with a child's tantrum, and an older woman discussed the various stop-smoking techniques she had tried in the past and how they failed. Perhaps tailoring according to such specific circumstances would add value to letters; but such tailoring would require much more information than can be obtained from a 4-page multiple-choice questionnaire. We could also argue for (3) because there clearly are many ways in which the tailored letters could have been improved (such as having practical 'how-to-stop' tips in more letters, as mentioned at the end of Section 4); and for (4) on the basis of the weak evidence for this mentioned in Section 4.

We do not know which of the above reason(s) were responsible for STOP's failure, so we cannot give clear lessons for future researchers or developers. This is perhaps true of many negative experimental results, and may be a reason why people do not publish them in the NLP community. Again there is perhaps a different attitude in the medical community, where papers describ-

ing experiments are taken as 'data points' and more theoretically minded researchers may look at a number of experimental papers and see what patterns and insights emerge from the collection as a whole. Under this perspective it is less important to state what lessons or insights can be drawn from a particular negative result, what matters is the overall pattern of positive and negative results in a group of related experiments. And like most such procedures, the process of inferring general rules from a collection of specific experimental results will work much better if it has access to both positive and negative examples; in other words, if researchers publish their failures as well as their successes.

We believe that negative results are also important in NLG, NLP, and AI, even if it is not possible to draw straightforward lessons from them; and we hope that more such results are reported in the future.

6 Other Evaluation Techniques in STOP

The clinical trial was by far the biggest evaluation exercise in STOP, but we also performed some smaller evaluations in order to test our algorithms and knowledge acquisition methodology (Reiter, 2000; Reiter et al., 2000). These included:

1. Asking smokers or domain experts to read two letters, and state which one they thought was superior;
2. Statistical analyses of characteristics of smokers; and
3. Comparing the effectiveness of different algorithms at filling up but not exceeding 4 A5 pages.

These evaluations were much smaller, simpler, and cheaper than the clinical trial, and often gave easier to interpret results. For example, the letter-comparison experiments suggested (although they did not prove) that older people preferred a more formal writing style than younger people; the statistical analysis suggested (although again did not prove) that the tailoring rules should have been more influenced by level of addiction; and the algorithmic analysis showed that a revision architecture outperformed a conventional pipeline architecture.

So, these experiments produced clearer results at a fraction of the cost of the clinical trial. But the cheapness of (1) and (2) were partially due to the fact that they were too small to produce statistically solid findings, and the cheapness of (2) and (3) were partially due to the fact that they exploited data sets and resources that were built as part of the clinical trial. Overall, we believe that these small-scale experiments were worth doing, but as a supplement to, not a replacement for, the clinical trial.

7 When is a Clinical Trial Appropriate?

When is it appropriate to evaluate an NLG system with a large-scale task or effectiveness evaluation which compares the NLG system to a non-NLG alternative? Certainly this should be done when a customer is seriously considering using the system, indeed customers may refuse to use a system without such testing.

Controlled task/effectiveness evaluations are also scientifically important, because they provide a technique for testing applied hypotheses (such as 'STOP produces effective smoking-cessation letters'). As such, they should be considered whenever a researcher is interested in testing such hypotheses. Of course, much research in NLG is primarily theoretical, and thus perhaps best tested by corpus studies or psycholinguistic experiments; and much work in applied NLG is concerned with pilot studies and other hypothesis formation exercises. But at the end of the day, researchers interested in applied NLG need to test as well as formulate hypotheses. While many speech recognition and natural-language understanding applications can be tested by comparing their output to a human-produced 'gold standard' (for example, speech recogniser output can be compared to a human transcription of a speech signal), this to date has been harder to do in NLG, especially in applications such as STOP where there are no human experts (Reiter et al., 2000) (there are many experts on personalised oral communication with smokers, but none on personalised written communication, because no one currently writes personalised letters to smokers). In such applications, the only way to test hypotheses about the effects of systems on human users may be to run a controlled task/effectiveness evaluation.

In other words, there's probably no point in conducting a large-scale task/effectiveness evaluation of an NLG system if you're interested in formulating hypotheses instead of testing them, or if you're interested in theoretical instead of applied hypotheses. But if you want to test an applied hypothesis about the effect of an NLG system on human users, the most rigorous way of doing this is to conduct an experiment where you show some users your NLG texts and other users control texts, and measure the degree to which the desired effect is achieved in both groups.

Large-scale evaluation exercises also have the benefit of forcing researchers and developers to make systems robust, and to face up to the messiness of real data, such as awkward boundary cases and noisy data. Indeed we suspect that STOP is one of the most robust non-commercial NLG systems ever built, because the clinical trial forced us to think about issues such as what we should do with inconsistent or improperly scanned questionnaires, or what we should say to unusual smokers.

In conclusion, large-scale task/effectiveness evaluations are expensive, time-consuming, and a considerable hassle. But they are also an essential part of the scientific and technological process, especially in testing applied hypotheses about the effectiveness of systems on real users. We hope that more such evaluations are performed in the future, and that their results are reported whether they are positive or negative.

Acknowledgements

Many thanks to the rest of the STOP team, and especially to Ian McCann and Annette Hermse for their work in the clinical trial. Thanks also to Yaji Sripada, Sandra Williams, and the anonymous reviewers for their comments on drafts of this paper. This research was supported by the Scottish Office Department of Health under grant K/OPR/2/2/D318, and the Engineering and Physical Sciences Research Council under grant GR/L48812.

References

Guiseppe Carenini and Johanna Moore. 2000. An empirical study of the influence of argument conciseness on argument effectiveness. In *Proceedings of ACL-2000*.

José Coch. 1996. Evaluating and comparing three text production techniques. In *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING-1996)*.

Malcolm Law and Jin Tang. 1995. An analysis of the effectiveness of interventions intended to help people stop smoking. *Archives of Internal Medicine*, 155:1933–1941.

A Scott Lennox, Liesl Osman, Ehud Reiter, Roma Robertson, James Friend, Ian McCann, Diane Skatun, and Peter Donnan. 2001. The cost-effectiveness of computer-tailored and non-tailored smoking cessation letters in general practice: A randomised controlled study. *British Medical Journal*. In press.

James Lester and Bruce Porter. 1997. Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics*, 23(1):65–101.

John Levine and Chris Mellish. 1995. The IDAS user trials: Quantitative evaluation of an applied natural language generation system. In *Proceedings of the Fifth European Workshop on Natural Language Generation*, pages 75–93, Leiden, The Netherlands.

Chris Mellish and Robert Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech and Language*, 12:349–373.

James Prochaska and Carlo diClemente. 1992. *Stages of Change in the Modification of Problem Behaviors*. Sage.

Ehud Reiter. 2000. Pipelines and size constraints. *Computational Linguistics*, 26(2):251–259.

Ehud Reiter, Roma Robertson, and Liesl Osman. 1999. Types of knowledge required to personalise smoking cessation letters. In Werner Horn et al., editors, *Artificial Intelligence and Medicine: Proceedings of AIMDM-1999*, pages 389–399. Springer-Verlag.

Ehud Reiter, Roma Robertson, and Liesl Osman. 2000. Knowledge acquisition for natural language generation. In *Proceedings of the First International Conference on Natural Language Generation*, pages 217–215.

Ching-Long Yeh and Chris Mellish. 1997. An empirical study on the generation of anaphora in chinese. *Computational Linguistics*, 23(1):169–190.

Michael Young. 1999. Using Grice's maxim of quantity to select the content of plan descriptions. *Artificial Intelligence*, 115:215–256.