Should Corpora Texts Be Gold Standards for NLG?

Ehud Reiter and Somayajulu Sripada
Dept of Computing Science, University of Aberdeen,
Aberdeen AB24 3UE, UK
{ereiter,ssripada}@csd.abdn.ac.uk

Abstract

There is increasing interest in using corpora in NLG, perhaps because of the success of corpus-based techniques in other areas of speech and language processing. Many uses of corpora in NLG implicitly assume that the human-authored texts in a corpora are a 'gold standard', in other words that the NLG system should produce texts similar to the corpora texts. However, our experience with several corpora raises questions about this assumption, because human authors make mistakes and because different people write differently.

1 Introduction

There is growing interest in using corpora of human-authored texts in Natural Language Generation (NLG), especially for knowledge acquisition. For example, several papers at ACL 2001 (Barzilay and KcKeown, 2001; Duboue and KcKeown, 2001; Hardt and Rambow, 2001) described how machine-learning techniques could be applied to a corpus to learn (respectively) rules for VP ellipsis, rules for ordering NP constituents, and paraphrase possibilities. There is also interest in using corpora for evaluation, for example Bangalore et al. (2000) describe techniques for evaluating an NLG system by comparing its output to human texts from a corpus.

All of the above papers assume that rules acquired from machine-learning on a corpora are good ones for NLG systems, and/or that an appropriate evaluation for an NLG system is to see how close its texts are to the human texts in a corpus. This in turn is based on the underlying assumption that NLG systems should attempt to generate texts that are similar to corpus texts, in other words that corpus texts are a 'gold standard' for NLG. However, our experiences with several corpora brings this assumption into question, because

- There are substantial variations between individual writers, which reduces the effectiveness of corpus-based learning.
- Human writers often make mistakes, especially if they are writing quickly; we do not want NLG systems to imitate these mistakes.

Note that variation and errors in a corpus may be desirable if the corpus consists of system *inputs*, and we expect such variations to occur in real-life inputs as well. Thus, for example, if we want a Natural Language Understanding (NLU) system to be able to deal robustly with varied texts which may contain grammatical and other mistakes, then it is useful if a corpus used to develop such a system contains examples of such texts (although generally we do not want variations and errors in parts of the corpus which represent annotations or system outputs). But in NLG, where textual corpora generally consist of potential system *outputs*, variations and mistakes in the corpus may not be desirable, because we usually do not want NLG systems to generate texts which are inconsistent or contain mistakes.

2 Background: Corpora in NLG

One of the biggest challenges in NLG is knowledge acquisition (Reiter et al., 2000). The NLG field is perhaps starting to have some partial understanding of appropriate architectures and algorithms, but we still have a very poor understanding of how to acquire the detailed knowledge and rules (content, discourse, lexical, etc.) needed to build real applications in real domains and genres. One solution is to use corpus techniques, that is to acquire a collection of human-authored texts in the domain/genre and use machine learning techniques to extract the knowledge and choice rules used by human writers;

this is essentially the strategy followed by the ACL-2001 papers cited above, and indeed many other recent NLG papers.

Evaluation is another big challenge in NLG for which corpus-based approaches seem attractive. Traditionally most NLG systems were evaluated somewhat informally, and the community is rightly insisting that evaluation should be more scientific and more rigorous. Testing the effectiveness of NLG systems on real users in a rigorous fashion can be extremely expensive and time consuming (Reiter et al., 2001). Hence there is interest in evaluating systems by comparing their output texts to texts produced by human writers from the same input data. This strategy is usually quicker and cheaper than user-based evaluations, and has been successfully used in related fields such as machine translation (Papineni et al., 2001).

Both of these uses of corpora in NLG are attractive as they provide solutions to difficult problems; but we must not lose sight of the fact that they only make sense if corpora texts are in fact similar to the texts that we would like NLG systems to produce.

Statistical and machine learning techniques are of course not restricted to corpus analysis. For example Rambow et al. (2001) asked humans to evaluate texts produced from randomly generated sentence plans and then used machine learning techniques to learn which sentence-planning choices led to good evaluation scores. In this paper we focus on issues with using corpora, and do not examine other uses of statistical and machine learning techniques in NLG.

3 Our Corpora

We have built up three corpora over the past few years, which contain input data as well as output texts. By far the largest of these is corpus of 1099 *weather forecasts for off-shore oil rigs*, written by five professional meteorologists (Sripada et al., 2001). The reports were primarily based on the output of a numerical weather simulation, and our corpus contains this information as well as the forecast texts. Each forecast is roughly 400 words long (depending on what is counted as a word), giving a total corpus size of about 400,000 words. Much of our analysis has focused on statements describing predicted wind speed and direction at 10 meters altitude during the next 72 hours. Each forecast contains 3 such statements, each of which is roughly 10 words long, hence there are about 30,000 words in our wind-statement subcorpus. These sizes are of course very small compared to many text-only corpora such as the British National Corpus (BNC), but we believe that our weather forecast corpus is one of the largest corpora in existence which contains both texts and (non-linguistic) specifications of what information the texts are intended to communicate.

The forecast corpus consists of naturally occurring texts, written by real forecasters for real users. We also have two much smaller corpora which were artificially constructed, in the sense that we gave domain experts the input data and asked them to write a text based on this data:

- A corpus of 33 *smoking-cessation letters* written by five medical professionals (Reiter et al., 2000). Each letter was based on a questionnaire about smoking habits, experiences, beliefs, etc. that was filled out by a smoker.
- A corpus of 50 descriptions of *sensor readings from a gas-turbine*, written by two software developers who were knowledgeable about the turbine (Yu et al., 2001). Each text was based on a graph of the sensor in question.

We will focus in this paper on our forecast corpus, because it is the largest and contains naturally occurring texts, but we will also refer to the other corpora to illustrate that the problems we encountered were not unique to the forecast corpus.

4 Detailed Example: Weather Time Phrases

Space does not permit a detailed description of all the problems we have had with corpora in all of our domains. Instead, we will in this section give a detailed description of the problems that arose during one analysis, of weather forecast time phrases. In Section 5 we will briefly summarise some of the other problems we have observed.

4.1 The Problem: Deciding How to Communicate Time

Our weather forecast corpus was gathered as part of the SumTime project, in order to enable us to write an NLG system which automatically produced weather forecast texts from the output of a numerical weather simulation. This is a similar application to FOG (Goldberg et al., 1994). Corpus analysis of weather forecasts was also performed by the FOG developers, incidentally, but this analysis just examined the actual weather forecast texts, it did not also look at the input data the texts were based on.

One thing we hoped to learn from the corpus was which time phrases we should use in the generated forecast texts. For example, if the input data showed that the wind speed increased at time 1500 (3PM), which time phrase should be used in the generated

FORECAST 00-24 GMT, FRIDAY, 10-Nov 2000

WIND(10M): NNW 06-10 BACKING W'LY 02-06 BY MID AFTERNOON THEN BACKING SE 06-10 BY LATE EVENING

Figure 1: Extract from 5-day forecast issued on 9-Nov-00

day	hour	wind dir	wind speed
10-11-00	0	NNW	8
10-11-00	3	NNW	8
10-11-00	6	NNW	7
10-11-00	9	NW	7
10-11-00	12	WNW	6
10-11-00	15	W	3
10-11-00	18	SSW	2
10-11-00	21	SE	4
11-11-00	0	SE	8

Figure 2: Wind (at 10m) extract from 9-Nov-00 data file

text to communicate this time? Note that it is rare for weather forecasts to explicitly mention numerical times such as 1500, and also that although there are standard terminologies for some meteorological phenomena such as cloud cover and precipitation, we are not aware of any standard terminologies for the use of time phrases in weather forecasts.

4.2 Corpus Analysis Procedure

In order to learn rules for choosing time phrases, we performed the following corpus analysis:

1. We parsed the wind description statements in the forecast texts, using a simple parser tuned to the simple linguistic structure of these texts.

2. We extracted from the parses all phrases which mentioned the wind changing speed and direction at a certain time, and which did not use qualifiers such as *mainly* or *occasionally*.

3. For each such phrase found, we searched the data file corresponding to the phrase's forecast text for the first instance where the wind was recorded as having the direction specified in the text, and a speed within the range specified in the text. We assumed that this time was the intended meaning of the time phrase in this particular case.

4. We performed the analyses described below in attempt to learn how time phrases were used in forecasts.

For example, the wind text in Figure 1, which is an extract from a 5-day forecast produced on 9 Nov 2000, includes two phrases which describe changes in the wind:

1. BACKING W'LY 02-06 BY MID AFTERNOON: The first entry in the data file (Figure 2) with a direction of W and a speed within the range of 2-6 is at 15 hours, hence in this case *by afternoon* is assumed to mean 1500 hours.

2. BACKING SE 06-10 BY LATE EVENING: The first entry in the data file with a direction of SE and a speed within the range of 6-10 is at 0 hours (on 11-11-00), hence in this case *by late evening* is assumed to mean 0000 hours. Note that at 2100 hours the wind has a direction of SE but its speed is not in the 6-10 range, hence this does not count as a match.

This process is not perfect, and in particular we encountered two problems that distorted the association between time and time phrase:

- If several records in the data file match the phrase, we use the earliest one, and this is not always correct. For example, if the data file in Figure 2 contained a (W, 3) record at both 0600 and 1500, then our analysis procedure will record the first of these, 0600, as being used to mean BY MID AFTERNOON.

- The forecasters sometimes adjust what they say based on their meteorological expertise and on information not available to the numerical weather simulation (such as satellite weather images). For example, even if the data file stated that the wind would become (W, 3) at 1500, the forecaster could decide that this change will in fact happen earlier, at 0900, and use a time phrase that communicates this, such as BACKING W'LY 02-06 BY MID MORNING. In this case our analysis procedure will record BY MID MORNING as meaning 1500, as this is the first entry in the data file which fits W'LY 02-06.

In order to determine the impact of errors, we looked at the usage of time terms for which we believed there was a clear and unambiguous interpretation (for example, *by midday* and *by midnight*). In all such cases 75-85% of the usages recorded by our analysis were the expected ones (for example, 0000 for *by*

hour	F1	F2	F3	F4	F5	total
0		**5**	**35**	1	3	44
3			1			1
6					1	1
9						0
12		1				1
15	5		2	3		10
18	**19**	3	1	**22**	4	**49**
21	7	**5**	22	3	**6**	43
total	31	14	61	29	14	149

Figure 3: How often *by evening* was used to refer to each time, for each forecaster (mode in bold font)

midnight). This suggests that the error rate in the analysis is about 20%. We experimented with more complex analysis procedures intended to reduce the error rate, but these had only marginally lower error rates and were less clear and intuitive, so we used the simple procedure described above.

4.3 Results

We analysed 1099 forecast texts. This analysis gave 1382 (time phrase, time) pairs involving 46 different time phrases. We regarded time phrases as different if they involved different head nouns, different prepositions (for example, *midday* and *by midday*) and/or different adjectives (for example, *by afternoon* and *by late afternoon*). However, we ignored determiners (for example, *by this evening* was regarded as the same phrase as *by evening*).

Of these 46 phrases, we removed 23 phrases which were used less than 10 times, and 4 phrases whose time denotation was context-dependent and hence was expected to vary (for example, *later*). For the remaining phrases, we looked at the most common time (mode) that each forecaster used the phrase for, and discovered that:

- 1 phrase (*by morning*) was only used by one forecaster.
- 9 phrases had the same modes for all forecasters, and hence seemed to have a consistent meaning across all forecasters.
- 9 phrases had different modes for different forecasters, and hence did not have a consistent meaning across forecasters.

The most common non-contextual time phrase was *by midday*, which all forecasters primarily used to mean 1200. The second most common non-contextual time phrase was *by evening*, whose meaning seemed less consistent across forecasters. The usage of *by evening* is shown in Figure 3.

The differences between forecasters in their usage of *by evening* is significant at $p < .001$ under both a chi-square test (which treats time as a categorical variable) and a One-Way ANOVA (which compares the mean time for each forecaster; for this test we recoded the hour 0 as 24). From a more qualitative perspective, we note that

- Forecasters F1 and F4 primarily used *by evening* to mean 1800.
- Forecaster F3 primarily used *by evening* to mean 0000 (midnight), although he also on many occasions used it to mean 2100.
- Forecasters F2 and F5 did not use *by evening* as often as the other forecasters. Bearing in mind the 20% expected error rate, all we can confidently say about them is that usually use this phrase to mean 1800, 2100, or 0000.

Of course, what an NLG system really needs to know is the mapping from time to time-phrase, not the mapping from time-phrase to time. We show in Figure 4 the most common non-contextual time phrase used to refer to each time. We also show a set of time phrases suggested by an expert, and the set of time phrases we currently used in SumTime; we have removed determiners from these phrases in order to be consistent with our corpus analysis.

One surprise was that *by midday* was the most common term for 0900. This was unexpected because the mode (most common) usage of this term, for every forecaster, was 1200; 75% (145 out of 192) of the usages of *by midday* were associated with 1200 and only 11% (22 out of 192) were associated with 0900. We investigated in more detail, and our best guess as to what is happening is that all forecasters do indeed regard *by midday* as meaning 1200, but because of the above-mentioned error factors the corpus analysis procedure incorrectly associates 22 (1 in 9) of the usages with 0900. Unfortunately, none of the phrases that genuinely refer to 0900 gets a usage count of more than 14, because different forecasters have different preferred terms (*during morning*, *by late morning*, etc.), and also because the forecasters refer to 1200 considerably more often than they refer to any other time (perhaps because it is in the middle of most forecast periods).

4.4 Discussion

The above analysis shows that different forecasters associate different meanings with time terms. In other words, we can learn from the corpus rules about what time individual forecasters associate with time phrases such as *by evening*, but we cannot

hour	most common phrase in corpus	phrase suggested by expert	phrase used in SumTime
0	by late evening	around midnight	by midnight
3	tonight(*)	in early hours	after midnight
6	overnight(*)	in early morning	by early morning
9	by midday	during morning	by morning
12	by midday	around midday	by midday
15	by mid afternoon	in mid afternoon	by mid afternoon
18	by evening	in early evening	by early evening
21	by evening	during night	by evening

(*) means the difference in usage between this term and the second-most-common term was 25% or less.

Figure 4: Suggested (non-contextual) time-phrases for each time

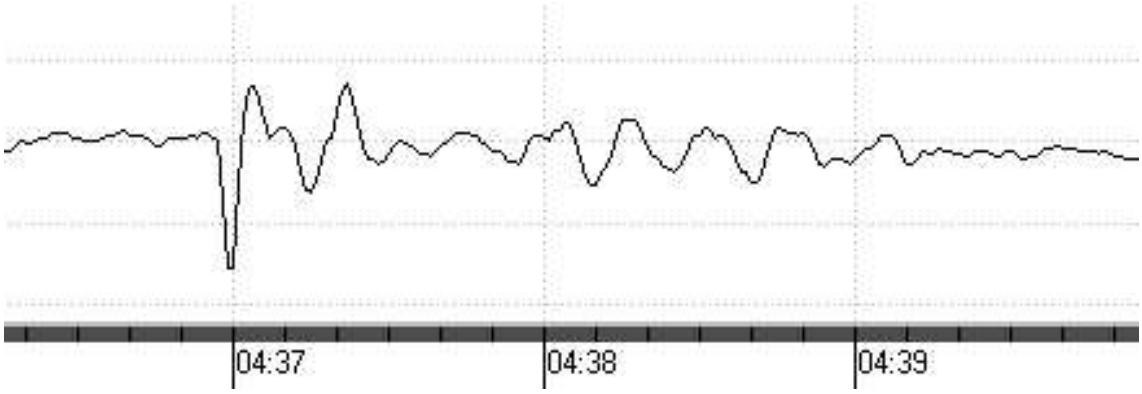

Figure 5: Signal fragment (gas-turbine exhaust temperature)

learn from the corpus forecaster-independent rules for what time phrases mean.

The analysis also shows that a corpus analysis can produce inappropriate rules, because the time phrases in the 'most common phrase in corpus' column of Figure 4 are not a good set of lexicalisations for times. In particular, they suggest that the phrases *by midday* and *by evening* should both be used for two times, which would make them ambiguous to readers. We believe that in particular using *by midday* for 0900 would be counterintuitive to both users and forecasters. We also suspect that *tonight* and *overnight* are vague terms whose connotations might not be clear to readers.

The time phrases suggested by the expert (also in Figure 4), in contrast, were unambiguous and more precise. The actual time-phrases currently used in our SumTime weather report system are shown in the last column of Figure 4, and are in essence a blending of what the corpus analysis showed and what the expert suggested.

In other words, from our perspective the corpus analysis was certainly useful when combined with other forms of knowledge acquisition, such as advice from experts. However, it would have been a mistake to rely purely on corpus analysis when deciding on time phrase rules.

5 Other Observations

Substantial individual variations were observed in all of our corpora. For example, in the gas turbine domain, we asked two experts to write descriptions of the signal (gas-turbine exhaust temperature) shown in Figure 5. The experts produced the following descriptions

- Expert 1: *A damped oscillation followed by a few wiggles*
- Expert 2: *Sudden drop out, followed by oscillatory recovery and further oscillations*

Note that these texts differ in numerous ways, including

- Content: Expert 1 has divided this signal into two components (*a damped oscillation* and *a few wiggles*), while Expert 2 has divided it up into three components (*sudden drop-out*, *oscillatory recovery*, and *further oscillations*).

What are the *good* things for you about smoking?	very important	quite important	not important
it helps me to relax		X	
it helps to break up my working time		X	
it is something to do when I am bored			
it helps me to cope with stress			
I enjoy it	X		
it is something I do with my friends or family			
it stops me putting on weight			
it helps me concentrate			
other			

Figure 6: Extract from smoker questionnaire

- Lexical choice: Expert 1 refers to the signal from 04:38 to 04:39 as *wiggles*, while Expert 2 describes it as *oscillations*.
- Syntax: Expert 1 includes the determiner *a*, while Expert 2 does not.

Similar variations were observed in the other texts in this corpus; in no case did the two experts produce identical descriptions for a signal.

5.0.1 Variation and Learnability

Besides casting some doubt on what is actually being learned in a corpus analysis, human variation also makes the learning process more difficult. For example, with our weather forecast corpus we used the machine learning algorithm Ripper (Cohen, 1995) to attempt to learn a choice rule which stated when reductions in wind speed should be described with the verb *easing* and when they should be described with the verb *decreasing*. This analysis was carried out in a roughly similar way to the one described in Section 4, except that we used Ripper to learn a rule instead of simply looking for the most common usage. When we allowed Ripper to include author in its learned choice rule, we were moderately successful, and learned a rule with a 17% error rate (10-fold cross validation), as compared to a 21% error rate from a baseline rule which always choose the most common verb. But when we told Ripper to ignore forecast author, it had no success at all, and could not learn a rule with higher-than-baseline accuracy. In short, in this example successful learning was only possible when the learned choice rule could include author dependencies.

5.1 Mistakes

Human writers make mistakes, and not surprisingly there are many mistakes in our corpora, including spelling mistakes, grammar mistakes, lexical mistakes, and content mistakes. Many of these mistakes were unintentional 'careless' mistakes, but some mistakes perhaps reflected time pressure or lack of knowledge on the part of the writers. This is important because while one could argue that statistical techniques will automatically filter out random careless mistakes, they may not filter out consistently made mistakes that are caused by lack of knowledge or by time pressure.

Looking first at the good things about your smoking, you felt smoking helped you break up your working day, and helped you relax. Are there any other ways you could do this without smoking? We have included a sheet with some suggestions for other ways of relaxing which you might find useful if you were to stop smoking.

Figure 7: Extract from 1997 smoking-cessation letter

The things you like about smoking are that you enjoy it, it helps you to relax, and it helps you break up your working day. Certainly many people enjoy smoking, and there is no easy answer to missing the enjoyment of smoking. Perhaps you could use some of the money you save to do something else you enjoy. But you may simply have to accept giving up the pleasure of smoking as the price to be paid for the benefits of stopping.

Figure 8: Extract from 1999 smoking-cessation letter

For example, in the smoking domain we performed an experiment where we asked a doctor to repeat a letter-writing exercise in February 1999 which he had first done in November 1997; that is, on both occasions we gave him the same smoker questionnaires and asked him to write letters for these smokers. The 1999 letters were clearly different than the 1997

letters, and when we showed one pair of letters to a group of seven smokers, they preferred the 1999 letter five to one over the 1997 letter (with one smoker expressing no preference). We suspect that the doctor had become a better letter writer in 1999 because of his involvement in our project; this of course means that the 1997 letters he wrote for our corpus were not as good as they could have been.

We asked the doctor to comment on the differences between his 1997 and 1999 letters, and he explicitly stated that one decision he had made in 1997 was probably a mistake. He had been writing a letter for a smoker who had said in his questionnaire that he really enjoyed smoking (see questionnaire extract in Figure 6), and he decided to ignore this (the relevant part of the 1997 letter is shown in Figure 7). When the doctor wrote a new letter for this smoker in 1999 and subsequently reviewed his old 1997 letter, he stated that ignoring the fact that the smoker enjoyed smoking was a mistake, and the letter should instead explicitly acknowledge this fact; this is shown in Figure 8, which is an extract from his 1999 letter.

Another example comes again from our weather forecast corpus. We noticed when we were analysing time phrases that forecasters often omitted a time phrase when some parameter changed in a more or less steady fashion throughout a forecast period. For example, if a S wind rose steadily in speed from 10 to 20 over the course of a forecast period covering a calendar day, many forecasters would write S 8-12 RISING TO 18-22, instead of S 8-12 RISING TO 18-22 BY MIDNIGHT. A corpus analysis showed that the time phrase (such as BY MIDNIGHT in this example) was omitted in 50-60% of such cases (the uncertainty is due to the error factors mentioned in Section 4). Accordingly, we programmed our system to omit the time phrase in such circumstances. However, when we showed our system to forecast managers during an initial informal evaluation, they immediately noticed this behaviour, and told us that it was incorrect, and that forecasts were more useful to end users if they included explicit time phrases and did not rely on the readers remembering when forecast periods ended. In other words, in this example the forecasters were doing the wrong thing in most cases, which of course meant that the rule produced by corpus analysis was incorrect.

We don't know why the forecasters are doing this, but discussions with the forecast managers about this and other mistakes (such as forecast authors describing wind speed and direction as changing at the same time, even when in fact they change at different times) suggest that one possible cause is time pressure. Forecasters write under considerable time pressure, which perhaps encourages them to produce shorter texts, even if these texts are not optimal for users. In fact most of the repeated mistakes we have observed do involve forecasters writing shorter texts than would be optimal for the user.

Thus, people don't always do the right thing, even if they haven't made a careless mistake. Many, perhaps most, experts have gaps in their understanding, and also experts writing under time pressure will be tempted to cut corners; such behaviour should not be imitated by NLG systems.

6 Corpus Analysis for Different Types of Knowledge

A few colleagues have pointed out to us that a perhaps unusual aspect of our corpora analyses is that we are largely interested in content and lexicalisation decisions. Most previous usages of corpus analysis, in contrast, have focused on realisation and expression issues. This raises the question of whether corpus analysis is better suited to grammatical and linguistic questions than to semantic and content questions – perhaps because humans are more consistent and make fewer mistakes with grammatical decisions than with content decisions? This is an interesting conjecture, which perhaps deserves to be further explored and developed.

7 Implications: Corpora and NLG

What are the implications of the above analysis? Firstly, we encourage people using corpora in NLG to construct their corpora in a way which minimises the above problems. For example, if we could redo the corpus-building exercises which produced the artificial corpora mentioned in Section 3, we might use fewer authors and also give the authors more time for each text. This would reduce the size of the corpus but would increase its quality, by reducing variation and problems due to hurried writing. We believe that for many NLG applications a smaller corpus of high-quality texts is more useful than a larger corpus of problematical texts. After all, we usually want NLG systems to produce high-quality texts, and it seems plausible that we are more likely to acquire rules for generating good quality texts if we analyse a corpora of such texts.

Secondly, we believe that the results of corpus-based knowledge acquisition should be treated as hypotheses which need to be integrated and compared with the results of other types of KA (such as working with experts), and then evaluated with experts or users. We should not assume a corpus-derived rule is always true, and we should not rely solely on corpus-

based evaluation to determine if a corpus-derived rule is correct. As hopefully has been made clear in this paper, we have seen many cases where corpus-derived rules (such as expressing 0900 as *by midday*, or omitting end-of-period time phrases) seem to be incorrect, but we would never have detected these problems if we had not performed expert-based KA and evaluation as well as corpus-based KA and evaluation.

Finally we believe that more research is needed to determine how well corpus-based evaluation is correlated to the results of user-based evaluations. We applaud the initial efforts of Bangalore et al. (2000) along these lines, but believe that larger and more comprehensive experiments that involve a complete NLG system are needed, perhaps similar to those carried out with the BLEU evaluation technique for machine-translation systems (Papineni et al., 2001).

In summary, we believe that corpus-based knowledge acquisition and evaluation can be very useful in developing NLG systems, but they are not a panacea. In particular they suffer from the problem that we usually do not want NLG systems to imitate some aspects of human-written texts, such as inconsistency and corner-cutting due to time pressure. Corpus-based techniques can be very valuable when combined with other KA and evaluation techniques, but we believe it is inappropriate to rely purely on corpus-based KA and evaluation.

Acknowledgements

Numerous people have given us comments and feedback on this work (not all of them agreeing with our conclusions), too many to acknowledge here. But we would like to give special thanks to our colleagues in STOP and SUMTIME who helped us create our corpora, and to the experts who wrote the corpora texts. Special thanks as well to Sandra Williams for reading and commenting on several versions of this paper. This research was supported by the UK Engineering and Physical Sciences Research Council (EPSRC), under grants GR/L48812 and GR/M76881.

References

Srinavas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the First International Conference on Natural Language Generation*, pages 1–8.

Regina Barzilay and Kathleen KcKeown. 2001. Extracting paraphrases from a parallel corporus. In *Proceedings of the 39th Meeting of the Association for Computation Linguistics (ACL-01)*, pages 50–57.

William Cohen. 1995. Fast effective rule induction. In *Proc. 12th International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann.

Pablo Duboue and Kathleen KcKeown. 2001. Empirically estimating order constraints for content planning in generation. In *Proceedings of the 39th Meeting of the Association for Computation Linguistics (ACL-01)*, pages 172–179.

Eli Goldberg, Norbert Driedger, and Richard Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

Daniel Hardt and Owen Rambow. 2001. Generation of VP-ellipsis: A corpus-based approach. In *Proceedings of the 39th Meeting of the Association for Computation Linguistics (ACL-01)*, pages 282–289.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. Technical Report RCS22176, IBM Thomas Watson Research Center, Yorktown Heights, NY 10598, USA.

Owen Rambow, Monica Rogati, and Marilyn Walker. 2001. Evaluating a trainable sentence planner for a spoken dialogue system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, pages 426–433.

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.

Ehud Reiter, Roma Robertson, Scott Lennox, and Liesl Osman. 2001. Using a randomised controlled clinical trial to evaluate an NLG system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, pages 434–441.

Somayajulu Sripada, Ehud Reiter, Jim Hunter, Jin Yu, and Ian Davy. 2001. Modelling the task of summarising time series data using KA techniques. In *Applications and Innovations in Intelligent Systems IX*, pages 183–196. Springer-Verlag.

Jin Yu, Jim Hunter, Ehud Reiter, and Somayajulu Sripada. 2001. An approach to generating summaries of time-series data in the gas-turbine domain. In *Proceedings of ICII-2001*, pages 44–51. IEEE Press.