Choosing Words in Computer-Generated Weather Forecasts

Ehud Reiter *,

Dept of Computing Science, University of Aberdeen, UK

Somayajulu Sripada,

Dept of Computing Science, University of Aberdeen, UK

Jim Hunter,

Dept of Computing Science, University of Aberdeen, UK

Jin Yu,

Dept of Computing Science, University of Aberdeen, UK

Ian Davy

Aerospace and Marine International, Aberdeen, UK

Abstract

One of the main challenges in automatically generating textual weather forecasts is choosing appropriate English words to communicate numeric weather data. A corpus-based analysis of how humans write forecasts showed that there were major differences in how individual writers performed this task, that is, in how they translated data into words. These differences included both different preferences between potential near-synonyms that could be used to express information, and also differences in the meanings that individual writers associated with specific words. Because we thought these differences could confuse readers, we built our SumTime-Mousam weather-forecast generator to use consistent data-to-word rules, which avoided words which were only used by a few people, and words which were interpreted differently by different people. An evaluation by forecast users suggested that they preferred SumTime-Mousam's texts to human-generated texts, in part because of better word choice; this may be the first time that an evaluation has shown that NLG texts are better than human-authored texts.

Key words: natural language processing, natural language generation, language and the word, information presentation, weather forecasts, lexical choice, idiolect

1 Introduction

A key aspect of connecting language to the world is choosing words that express non-linguistic data. This is part of the general problem of determining how linguistic constructs such as words relate to non-linguistic information such as sensor data. Most previous research on connecting language to the world has assumed that people use similar data-to-word mappings, but this is not necessarily the case. For example, it seems clear that there are substantial individual differences in how humans use colour terms such as *pink* [25]; in other words, the fact that one person considers an object to be *pink* does not necessarily mean that another person would also consider this object to be *pink*, even under identical lighting conditions. Such differences are one aspect of 'idiolect', that is how language is used by an individual. Idiolect differences raise an important question for computer Natural Language Generation (NLG) systems [27], that is software systems that produce texts in English (or other human languages) from non-linguistic input data; how can they ensure that the texts they produce are correctly interpreted by their readers, since words may be interpreted differently by different readers?

We have explored this issue in the context of SumTime-Mousam, which is an NLG system that generates weather forecast texts from numerical weather prediction data (SumTime-Mousam in fact is used by an Aberdeen company to help generate real forecasts, it is not just a research prototype [41]). A corpus analysis of human written forecasts showed that there was indeed considerable variation in how different forecasters choose words to express data; this reflected differences in preferences between near-synonyms, and also differences in meanings associated with words and phrases (especially time phrases). Because we thought human readers would value consistency, we built SumTime-Mousam to use a consistent set of data-to-word rules, which avoided words that only occurred in a few idiolects and words that had idiolect-dependent meanings (even if these words were common in the human corpus). An evaluation of SumTime-Mousam showed that human forecast readers preferred SumTime-Mousam texts to human-written texts; we are not aware of any previous evaluation showing that texts produced by an NLG system were superior to manually written texts. Human readers also rated human-written texts

* Corresponding author
Email addresses: ereiter@csd.abdn.ac.uk (Ehud Reiter), ssripada@csd.abdn.ac.uk (Somayajulu Sripada), jhunter@csd.abdn.ac.uk (Jim Hunter), jyu@csd.abdn.ac.uk (Jin Yu), ian@weather3000.com (Ian Davy).

edited to use SumTime-Mousam's words (but preserving the same content as the original human-written texts) as significantly easier to read than the human-written texts; this suggests that SumTime-Mousam does a better job of choosing words than human forecasters.

In the rest of this paper we present background information about SumTime-Mousam, describe our analysis of variability in human written forecasts, and summarise the results of our evaluation of SumTime-Mousam.

2 Background

2.1 Textual Summaries of Time-Series Data

The modern world is being flooded with data, and understanding and interpreting this data is a major challenge for many 21^{st} century professionals. For example, a typical gas-turbine produces 200MB of sensor data per day; one value per second from 250 sensors that measure temperatures, vibrations, fuel flows, power outputs, and so forth. A maintenance engineer may have one hour (per day) to attempt to understand this deluge of data. Similarly a doctor in an intensive care unit may be presented with megabytes of data (heart rate, blood pressure, etc) when making a treatment decision; assimilating all this data (especially under time pressure) is not easy. Such examples are everywhere in the modern world. Data is all pervasive and influences many decisions, and effective methods of data presentation are badly needed.

Currently time-series data is usually presented to people either numerically (tables) or graphically (visualisations) [36]. The goal of the Aberdeen SumTime project (which included SumTime-Mousam) was to develop better techniques for automatically generating textual summaries of time-series data; an overview of what we are attempting is shown in Figure 1. Whatever one thinks of the theoretical merits of textual versus graphical presentation of information, an enormous practical advantage of graphical presentations is that they can be produced automatically (and hence very cheaply), whereas textual summaries currently usually need to be written by a person (and hence are expensive). If good-quality textual summaries could be produced automatically, and hence cheaply and quickly, they would be much more attractive.

SumTime developed systems in three domains:

- *Gas Turbines*: Our SumTime-Turbine system [48] generates textual summaries of sensor data from a gas turbine.
- *Intensive Care*: Our SumTime-Neonate system [39] generates textual

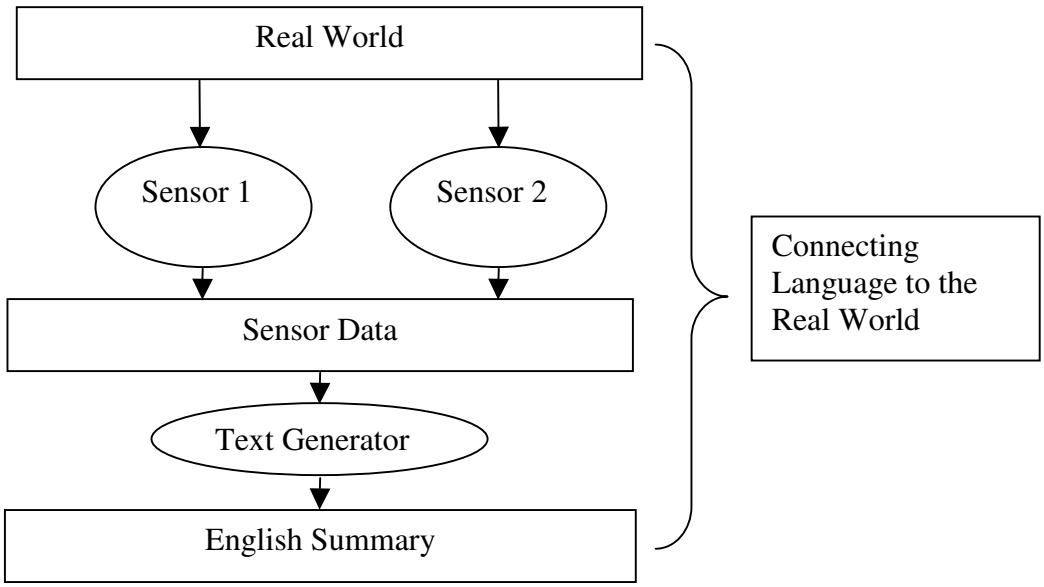


Fig. 1. Connecting Language to the World via Sensors

summaries of sensor data from a neonatal intensive care unit (ICU).
- *Meteorology*: Our SumTime-Mousam system [38] (the focus of this paper) generates weather forecast texts from numerical weather prediction data.

Perhaps the biggest difference between SumTime-Mousam and the other SumTime systems was that humans regularly read and write textual weather forecasts, whereas few people currently read or write textual summaries of gas-turbine and intensive-care data. Hence SumTime-Mousam could be informed by corpus analysis of existing human-written forecasts, and also could be evaluated by people who were used to reading textual weather forecasts.

Incidentally, some of our colleagues experimentally evaluated the effectiveness of human-written text summaries of ICU data against graphical presentations of the same data[20]. They did this by showing medical professionals either a text summary or a graphical presentation, and asking the medics to decide what action (if any) should be taken. When the medics were asked which presentation they preferred, they said they preferred the graphical presentation. However, when our colleagues evaluated the correctness of treatment decisions, they found that the medics made better treatment decisions when shown the textual summary. This supports the hypothesis that textual summaries can be an effective way of presenting data.

2.2 Weather forecasts

Modern weather forecasting is largely based on numerical weather predictions (NWP), which essentially are massive atmosphere simulations run on supercomputers. The output of NWP models is a set of predictions of meteorological parameters (wind speed, temperature, precipitation, etc) for various spatial locations and at various points in time.

Weather forecasting organisations take NWP data and modify it according to their local knowledge and expertise; for example they may know from previous experience that an NWP model tends to underestimate wind speeds at a certain location under some conditions. They also interpolate between the locations in the source NWP model, again using local knowledge and expertise. The result is a modified set of predicted numerical weather values, for locations of interest to their customers.

This data must then be presented to customers, who are assumed to primarily use it to assist in decision making. The first step in this process is selecting data that is important to the customer, given his likely decisions; for example a pilot who is landing an airplane at Aberdeen airport is very interested in visibility and wind speeds at various altitudes but may be less interested in temperature, while a farmer farming a field next to Aberdeen airport may be very interested in temperature and less interested in wind speed and visibility. The second step is to produce an actual summary of the data; this can be textual, tabular, or graphical, according to the customer's wishes.

In SumTime-Mousam, we primarily focused on generating forecasts for offshore oil rigs in the North Sea; these are used by staff on the rigs, on support boats, and on shore to make operational decisions. For example, a new pipeline can be laid only when the weather is relatively calm; if severe weather starts when a pipeline is only partially laid then it may be necessary to abandon it. Hence good forecasts are very important to deciding when to lay a pipeline, and for many other operational decisions as well.

Forecasts are normally issued twice a day for offshore rigs (although additional forecasts can always be requested). They are created by a team of forecasters; this essentially means that forecasts for a particular rig are written by different people on different days.

2.3 SumTime-Mousam

SumTime-Mousam [38] generates weather forecasts from weather prediction data. The input to SumTime-Mousam is numerical weather parameters from

Time	Wind Dir	Wind Speed 10m	Wind Speed 50m	Gust 10m	Gust 50m
06:00	W	10.0	12.0	12.0	16.0
09:00	W	11.0	14.0	14.0	17.0
12:00	WSW	10.0	12.0	12.0	16.0
15:00	SW	7.0	9.0	9.0	11.0
18:00	SSW	8.0	10.0	10.0	12.0
21:00	S	9.0	11.0	11.0	14.0
00:00	S	12.0	15.0	15.0	19.0

Table 1
Part of an input data set for SumTime-Mousam

an NWP model, as adjusted by forecasters to reflect local knowledge and expertise. Note that although these parameters are produced by a simulation, they are similar to real data produced by meteorological sensors. The output of SumTime-Mousam is a weather forecast. An extract from a typical SumTime-Mousam input data set is shown in Table 1, and an extract from the forecast generated by SumTime-Mousam from this input data set is shown in Figure 2.

We have concentrated on marine forecasts for offshore oil rigs, although a version of SumTime-Mousam is also used to help generate forecasts delivered to the public via telephone weather-information lines and SMS (mobile telephone text messages) weather-information services. SumTime-Mousam was developed in collaboration with Weathernews (UK) and Aerospace and Marine International (UK), and indeed Weathernews is currently operationally using SumTime-Mousam to help forecasters produce some types of forecasts; essentially SumTime-Mousam is used to generate draft forecasts, which are post-edited by human forecasters and then released to customers.

Like many NLG systems, SumTime-Mousam generates texts in three stages [27]:

- *Document Planning* decides on the content and structure of the generated text. SumTime-Mousam must in particular decide what information from the numeric weather data to communicate in a text; for example, the Wind (10M) text in Figure 2 mentions the fact that the wind direction changes to SW at 1500 (*mid afternoon*), but not the fact that the wind speed has changed to 7 knots at this time. SumTime-Mousam determines content using linear segmentation [17], which is adapted according to a pragmatic (Gricean) analysis of appropriate content for a weather forecast text [40]. The other key document planning task is deciding on the structure of gen-

Section 2. FORECAST 6 - 24 GMT, Wed 12-Jun 2002

Field	Text
WIND(KTS) 10M	W 8-13 backing SW by mid afternoon and S 10-15 by midnight.
WIND(KTS) 50M	W 10-15 backing SW by mid afternoon and S 13-18 by midnight.
WAVES(M) SIG HT	0.5-1.0 mainly SW swell.
WAVES(M) MAX HT	1.0-1.5 mainly SW swell falling 1.0 or less mainly SSW swell by afternoon, then rising 1.0-1.5 by midnight.
WAVE PERIOD (SEC)	Wind wave 2-4 mainly 6 second SW swell.
WINDWAVE PERIOD (SEC)	2-4.
SWELL PERIOD (SEC)	5-7.
WEATHER	Mainly cloudy with light rain showers becoming overcast around midnight.
VISIBILITY (NM)	Greater than 10.
AIR TEMP(C)	8-10 rising 9-11 around midnight.
CLOUD (OKTAS/FT)	4-6 ST/SC 400-600 lifting 6-8 ST/SC 700-900 around midnight.

Fig. 2. Extract from forecast generated by SumTime-Mousam, from the input data set partially shown in Table 1

erated texts; SumTime-Mousam does this using a schema [24] (pattern) which essentially imitates the structure of human-written forecasts.

- *Microplanning* decides on how abstract content and structure should be expressed linguistically. The most important aspect of this in SumTime-Mousam is *lexicalisation*, that is choosing which words should be used to express non-linguistic data; this is the focus of this paper. Also important in SumTime-Mousam is *aggregation*, that is deciding how to distribute information among sentences (for example, many short sentences or a few longer sentences). We do not discuss aggregation here; although it is important for producing acceptable texts, it is less fundamental to connecting language to the world. However, most of the observations we make about lexicalisation also apply to aggregation; there is considerable variation in

how human forecasters aggregate, and such variation may not be ideal for forecast readers. The third general microplanning task is *referring expressions generation*, that is deciding how to refer to entities introduced earlier in the generated text. This is quite straightforward in SumTime-Mousam texts.

- *Surface Realisation* generates an actual text according to the decisions made in document planning and microplanning, ensuring that the text conforms to the grammar of the target language. The key realisation challenge in SumTime-Mousam was to generate texts in 'weatherese' (that is, weather sublanguage [14] instead of conventional English); this was essentially done by building special grammar rules, based on an analysis of human-written forecasts.

2.4 Related work on Generating Textual Summaries of Data

A number of previous NLG systems have generated weather forecasts, including FoG [11] and MultiMeteo [6]. Other NLG systems which generated summaries of data include ANA [19], which generated summaries of stock market activity; LFS [18], which generated summaries of statistical data; SUMGEN [23], which generated summaries of events in a battle simulation; TEMSIS [3], which generated summaries of environmental data; and TREND [2], which generated summaries of historical weather data (not weather predictions). However, previous work on generating textual summaries of data has not (to the best of our knowledge) emphasised lexical choice and individual linguistic variability, which is the emphasis of this paper.

Quite a bit of research has been done on generating textual summaries of textual information [21]. Although the techniques used to summarise textual input are quite different from the techniques used to summarise data (systems that summarise text generally use information retrieval techniques to identify key phrases and sentences, and then stitch these together into a coherent summary using limited NLG), it is encouraging that evaluations of such systems have shown that users using summaries of texts can perform tasks much more quickly, and just as accurately, as people using raw unsummarised text [22]. As above, to the best of our knowledge research on 'text-to-text' summarisation has ignored individual variations in language use.

3 Word Choice in Human and SumTime-Mousam Forecasts

In this section we discuss our analysis of how human forecasters choose words for weather forecasts (focusing on wind statements), and also the lexical choice

rules that we implemented in SumTime-Mousam.

To give a concrete example of the problem of mapping data to words in SumTime-Mousam, consider the example Wind (10m) text from Figure 2, *W 8-13 backing SW by mid afternoon and S 10-15 by midnight*. Generating this text requires making several choices about which words to use, including

- Direction: should West be expressed as *W* or *W'LY*?
- Speed: should 8 knots be expressed as *8* or *08*?
- Verb: should *backing* or *becoming* be used to describe the change in wind direction?
- Time phrase: should *by evening*, *by late evening*, or *by midnight* be used to express the time 0000?

Human forecasters in fact make all of these choices. That is, we have seen some human forecasts where West is expressed as *W* and others where it is expressed as *W'LY*, some human forecasts where a speed of 8 knots is expressed as *8* and others where it is expressed as *08*, etc.

In SumTime we attempted to investigate these issues empirically, by analysing how people write forecasts. More specifically, we collected and analysed a corpus of human-written forecasts, and also a corpus of human post-edits to computer-generated forecasts. The most unexpected finding of this analysis was that the data-to-text mapping depended on contextual factors. Previous models of lexical selection in Natural Language Generation, and previous models of word meaning in language-and-world systems, have generally used fixed (non-context-dependent) models of meaning [43,33,44,27,34]. However, our work shows that the choice of which word(s) are used to describe a particular data set also depends on:

- *Preferences of individual writer:* Different people choose different words to describe the same chunk of data. Furthermore, people may change their word preferences over time. This is the most important factor, and the focus of this paper.
- *Linguistic context:* Choice of word is influenced by linguistic factors such as the position of a word in a sentence. Experiments with another SumTime system, SumTime-Turbine, show that word choice is also influenced by how similar information is lexicalised elsewhere in the text; we briefly review these experiments as well in this section.

Below we present detailed analyses of the choice of time phrase and verb in wind descriptors; we also briefly discuss contextual effects in SumTime-Turbine. We have also looked at the choice of other types of words in wind phrases, including numbers (e.g., *8* vs. *08*), directions (e.g., *S* vs. *S'LY*), connectives (e.g., *then* vs. *before*), and adverbs (e.g., *gradually*). Our conclusions about these words are similar; human writers choose them on the basis of

personal preferences (idiolect) and linguistic context, as well as the meaning they are trying to communicate.

3.1 Choosing Time Phrases

We needed a set of rules which told us which time phrase should be used in a weather forecast to communicate a numerical time from the NWP input data file. To create this algorithm, we first analysed how forecasters used time phrases (Section 3.1.1) and in particular if they differed in the meanings they associated with time phrases. We then tried to learn a classifier which told us which time phrase forecasters used in a given context (Section 3.1.2). Finally we created a set of time-phrase choice rules (Section 3.1.3), and then analysed post-edits made by human forecasters to our computer-generated forecasts (Section 3.1.4).

3.1.1 Analysis of Aligned Corpus

We wished to determine how forecasters used time phrases, and in particular what time phrases meant in terms of actual time (for example, does *by morning* mean 0600, 0900, or 1200?). To do this, we analysed a corpus of 1045 manually written forecasts and corresponding NWP data files. These were issued between June 2000 and May 2002, for offshore oil rigs in the North Sea, and were written by five different forecasters. We analysed time phrases in the corpus as follows (more details are given in [30]):

(1) We extracted from our corpus all wind at 10 meters statements which covered a time period of 24 hours or less (we excluded statements covering more than one day because we wanted to avoid the complication of day references such as *Saturday*).
(2) We parsed these statements using a simple parser tuned to the linguistic structure of wind statements.
(3) We aligned each phrase with an entry (that is, a time) in the corresponding NWP data file. Essentially we looked for a data file entry with the same wind direction as the parsed phrase, and a speed which was as close as possible to the midpoint of the speed range in the parsed phrase. Tests on time phrases with unambiguous meanings (such as *by midday*) suggested that the alignment process was 86% accurate.

For example, the analysis of the example forecast shown in Figure 3 first broke this forecast up into four phrases, then parsed these phrases, and finally aligned each phrase as follows:

Example wind data:

Time	Wind Dir	Wind Speed
0000	SSW	12
0300	SSE	11
0600	ESE	18
0900	ESE	16
1200	E	15
1500	ENE	15
1800	ENE	18
2100	NNE	20
0000	NNW	26

Human forecast written from this data:

SSW 12-16 backing ESE 16-20 in the morning, backing NE early afternoon then NNW 24-28 late evening

Fig. 3. Example Wind Data and Human Forecast

- *SSW 12-16* (first phrase): this is only consistent with the 0000 data file entry, and hence is aligned with it.
- *backing ESE 16-20 in the morning* (second phrase): this is consistent with the 0600 and 0900 data file entries. We align with the 0600 entry, because the speed at 0600 (18 knots) is closest to the mid point of the speed range given in the text (16-20 knots). Hence we conclude that in this instance, *in the morning* means 0600.
- *backing NE early afternoon* (third phrase): We fail to align this, as no data file entry has a direction of *NE*.
- *then NNW 24-28 late evening* (fourth phrase): This is consistent with the 0000 entry only, hence we align it with this entry, and conclude that *late evening* in this case means 0000.

The result of this process was 2539 aligned (phrase, data file entry) pairs, which used 73 different time phrases. We analysed the association between time phrase and time in these pairs. Tables 2, 3, and 4 give details of the usage of the three most common non-contextual time phrases: *by evening*, *by midday*, and *by late evening* (contextual time phrases are phrases whose denotation we expect to vary with context, such as *soon* and *later*).

time	F1	F2	F3	F4	F5	total
0000	2	**9**	**80**	5	**14**	**110**
0300					1	1
0600					1	1
0900						0
1200		1				1
1500	2	1	1		2	6
1800	**30**	5	2	**27**	13	77
2100	13	6	8	2	11	40
total	37	22	91	34	42	236

Significance of differences: $p < 0.001$ (chi-square, ANOVA)
Table 2
Usage of *by evening*, by forecaster (mode in **bold**)

time	F1	F2	F3	F4	F5	total
0000			2	1		3
0300				1		1
0600				1		1
0900	3		1	7	2	13
1200	**23**		**71**	**86**	**11**	**191**
1500	7	**1**	9	5	2	24
1800			2	2	1	5
2100	1					1
total	34	1	85	103	16	239

Significance of differences: $p > 0.1$ (chi-square, ANOVA)
Table 3
Usage of *by midday*, by forecaster (mode in **bold**)

These tables also show the statistical significance of differences between forecasters, calculated with a chi-square test (which treats time as a categorical variable) and with a one-way ANOVA analysis (which compares mean time). This data suggests that

- *by evening* means different things to different people; for example, forecasters F1 and F4 primarily use this phrase to mean 1800, while F3 primarily uses this phrase to mean 0000.

time	F1	F2	F3	F4	F5	total
0000			**215**	**9**	**15**	**239**
0300					1	1
0600						0
0900						0
1200			1			1
1500						0
1800						0
2100			3	3	2	8
total	0	0	219	12	18	249

Significance of differences $p < 0.001$ (chi-square); $p = 0.06$ (ANOVA)
Table 4
Usage of *by late evening*, by forecaster (mode in **bold**)

- *by midday* was used in a very similar way by all forecasters (ignoring F2, who only used the term once).
- *by late evening* was used by all forecasters (who used this term) primarily to mean 0000. However, the usages of the different forecasters was still significantly different using the chi-square (categorical) test. This reflects a difference in the distribution of usage; in particular, F3 almost always (98% of cases) used this phrase to mean 0000, while F4 and F5 used this phrase to mean 0000 in about 80% of cases.

These patterns are replicated across the corpus: some phrases (such as *by midday* and *by morning*) are used in the same way by all forecasters; some phrases (such as *by evening* and *by late morning*) are used in different ways by different forecasters; and some phrases (such as *by late evening* and *by midnight*) have the same core meaning (e.g., 0000) but different distributions around the core.

We looked for seasonal variations in meaning. For example, since sunset in the North Sea can be as early as 3PM in the winter and as late as 11PM in the summer, *by evening* might be interpreted differently in summer and winter if it associated with the time of sunset. We found no evidence of such variation, and one of the forecasters explicitly told us that he did not vary the meaning of this phrase by season, because he interpreted it in terms of daily routine (e.g., when dinner is served), not in terms of sunset. However, our reader-based evaluation (Section 4.3.1) suggests that some readers may indeed take season into account when interpreting *by evening*.

3.1.2 Classifier analysis

Of course, what we really wanted to know was how to choose time phrases in SumTime-Mousam, not just the meanings forecasters associated with time phrases. We used the machine learning algorithm C4.5 [26] (as implemented in Weka's [46] J4.8 classifier) to learn classifiers which predicted which time phrase would be used in wind phrases extracted from the corpus; in other words, the class being predicted by the classifier was *by evening*, *by midday*, and so forth. The classifier was trained only on wind phrases which had been successfully aligned with the data file (Section 3.1.1), that is on the 2359 phrases which referred to a known time (subject of course to alignment error). More details about our classification analysis are given in Reiter and Sripada [31].

The classifier was always given the time being communicated as one of its features (for example, 0000). We experimented with giving it other features; indeed one of the main things we wished to learn was which features were most useful in predicting the choice of time phrase. These feature sets we experimented with were:

- *semantic* features: information from the data file, such as the actual wind speed and direction
- *author* feature: which forecaster wrote the text.
- *collocation* features: the preceding and subsequent words in the text.
- *repetition* feature: the previous word of this type (time phrase) in the text.
- *surface* features: the length and position in the sentence of the phrase.
- *temporal* features: when the forecast was issued, and how far in the future the prediction was from the forecast issue date.

For example, some of the features associated with the time phrase *IN THE MORNING* in the forecast shown in Figure 3 were

- *Class*: IN-THE-MORNING
- *Time*: 0600 (from alignment)
- *Wind-Speed*: 18
- *Author:* F5
- *Previous-Word:* Number (for this feature, all numbers were replaced by a generic Number token)
- *Previous-Time-Phrase:* None
- *Phrase-Position:* 2 (that is, this is the second phrase in the sentence)
- *How-Far-in-Future:* 1 day (that is, this forecast is for the day following the day the forecast was issued)

We used as a baseline a classifier which always chose the most common phrase for a time; for example, it always chose *by midday* for 1200. This classifier had a 67% error rate (all error rates are calculated using 10-fold cross-validation).

Adding the author feature reduced the error rate to 52%; this essentially incorporates the idiosyncratic preferences of authors. For example, when referring to 1500, F3 and F4 preferred to use *by mid afternoon*, F1 preferred *by afternoon*, and F5 preferred *early afternoon* (F2 did not have a clear preference).

Adding information about the position of a phrase in a sentence further reduced error rate to 48%; for example, F4 used *by midnight* to refer to 0000 in the middle of a sentence, but *later* to refer to this time at the end of a sentence. The other feature sets did not have a significant effect on classifier accuracy. We did not, for example, find any evidence that vaguer time phrases were used for predictions which were further in the future. Hence we concluded that author and the position of the phrase in the sentence were the most important features (at least of the ones we investigated) for predicting which time phrase would be used to refer to a time.

We did not explicitly check if alignment error had an impact on the above results. However, the fact that there were clear specific examples of the impact of author and position (as mentioned above) suggests that these features do effect the choice of time phrase, this is not just an artefact of alignment error.

3.1.3 Implementation

Our original hope had been that doing the above analysis would lead to a sophisticated and empirically-based set of choice rules for time phrases. In fact, the main outcome of our analysis was that time phrase choice (and meaning) seems to mostly depend on individual linguistic preferences, that is on the idiolect of the person writing the forecast. In principle we could have included a 'forecaster model' in SumTime-Mousam which recorded the preferences of individual forecasters. However, we decided not to do so, because we believed that forecast readers would dislike lexical variability, and prefer that lexical choice be consistently done regardless of author.

SumTime-Mousam as implemented dynamically decides whether a time should be expressed by a contextual (e.g., *later*) or non-contextual (e.g., *by midnight*) time-phrase, using information such as the position of the phrase in the sentence, as suggested by the analysis of Section 3.1.2. However, if SumTime-Mousam uses a non-contextual phrase, it always uses the same phrase; for example the only non-contextual time phrase that SumTime-Mousam uses for 0000 is *by midnight*, it never uses *by late evening*. We choose these non-contextual time phrases based on our corpus analysis, and also on a set of time phrases suggested to us by an expert meteorologist; essentially we tried to balance the goals of using common time phrases, avoiding ambiguity, and respecting the expert's advice. Table 5 shows the most common phrase in the corpus for each time, the expert's recommended time phrases, and the actual

time	most common phrase in corpus	phrase suggested by expert	phrase used in SumTime-Mousam
0000	by late evening	around midnight	by midnight
0300	tonight	in early hours	after midnight
0600	overnight	in early morning	by early morning
0900	by midday (**)	during morning	by (mid) morning (*)
1200	by midday	around midday	by midday
1500	by mid afternoon	in mid afternoon	by mid afternoon
1800	by evening	in early evening	by early evening
2100	by evening	during night	by (mid) evening(*)

(*) *by mid morning* is produced for 0900 if the text also includes the phrase *by early morning*, otherwise *by morning* is produced. Similarly *by mid evening* is produced for 2100 if the text also includes the phrase *by early evening*, otherwise *by evening* is produced.

(**) This is almost certainly due to alignment error; although only 5% of instances of *by midday* are aligned with 0900 (Table 3), this is a higher count that any other non-contextual time phrase, since most references to 0900 in the corpus use the contextual time phrase *soon*.

Table 5
Non-contextual time phrases used in SumTime-Mousam

time phrases used by SumTime-Mousam.

3.1.4 Post-edit Analysis

In addition to our corpus of manually written forecasts, we also collected a corpus of post-edited texts. This corpus consists of 2728 forecast texts produced by SumTime-Mousam, human-edited versions of these texts which were sent to real forecast users, and the raw NWP input data. We analysed this data [42] to see how the human forecasters edited SumTime-Mousam's time phrases. In particular, we hoped that post-edit analysis would shed light on whether forecasters were firmly committed to using particular time phrases, or whether they regarded choice of time phrases as unimportant; we assumed that they would only bother editing time phrases if they thought the choice of time phrase was significant.

Overall, forecasters edited SumTime-Mousam's time phrases in 17% of cases. The most commonly edited phrases were *by early evening* (for 1800), which was edited in 29% of cases; and *by early morning* (for 0600), which was edited in 34% of cases. In both of these cases, most edits consisted of removing the

modifier *early* (for example, converting *by early evening* to *by evening*). In both of these cases, the short version without *early* was in fact considerably more common in the corpus, but we decided to include *early* to help readers correctly interpret the phrase. For example, we thought using *by evening* to refer to 1800 might mislead people who usually used *by evening* to refer to 2100 or 0000 (Table 2), and *by early evening* was less likely to be misinterpreted.

If we ignore 'dropping *early*' edits, then the overall edit rate drops to 14%, with individual edit rates basically showing a normal distribution around this value, with a standard deviation of 3.5%. The most frequently edited time phrase is now *after midnight*, which SumTime-Mousam used to indicate 0300. This was edited in 19% of cases, most commonly into *later* (10% of cases) or *overnight* (6% of cases). The edit pattern is very idiosyncratic. Edit frequency by forecaster ranged from 0% (4 forecasters never changed *after midnight*) to 100% (2 forecasters always changed *after midnight* to something else). Actual edits were also idiosyncratic; for example, only one person replaced *after midnight* by *overnight* (but he did this a lot).

In short, the post-edit analysis suggests that at least some forecasters do care enough about time phrases to spend time post-editing the phrases chosen by SumTime-Mousam. Furthermore many of their edits seem to fit the pattern of changing SumTime-Mousam's time phrase to one that they personally prefer to use.

3.2 Verbs in Weather Forecasts

We also attempted to empirically develop a set of rules for choosing verbs in wind phrases (10 different verbs occurred in wind phrases in our corpus). We did not explicitly analyse the meanings associated with verbs (as we did with time phrases), because pilot analyses suggested that forecasters agreed on the meaning of verbs (unlike the situation with time phrases). We did build a classifier that predicted verb choice (Section 3.2.1), implement a set of lexical choice rules for verbs (Section 3.2.2), and use the post-edit data to evaluate our rules (Section 3.2.3), in a similar way to what we did with time phrases. We also explicitly asked forecasters to comment on how they made one verb choice (Section 3.2.4).

3.2.1 Classifier analysis

We learnt classifiers that predicted verb choice in wind phrases from various feature sets, using the same approach we used for learning classifiers that predicted time phrase choice (Section 3.1.2). We divided the choice of verb into three steps:

- *Verb type:* does the verb describe a change in wind speed (e.g., *rising*) or a change in wind direction (e.g., *veering*)?
- *Verb information:* What information does the verb communicate about speed or direction? For example, if the verb describes a change in wind direction, does it describe direction as shifting clockwise (e.g., *veering*), shifting counter clockwise (e.g., *backing*), or transitioning to/from the *variable* state (e.g., *becoming*).
- *Near-synonym choice:* If several verbs can be used to communicate the chosen type and information, which is chosen? For example, if the verb is needed to communicate that the wind speed is going down, will *decreasing*, *easing*, or *falling* be used?

Classifiers were built for each of these steps.

The results of our analysis were basically as follows (more details are given in [31]):

- Verb type is mostly determined by semantic information, that is what is happening to the wind. Basically (as one might expect), a speed verb is chosen if the change in speed is more significant than the change in direction, and a direction verb is chosen if the change in direction is more significant than the change in speed. A conjoined verb group (e.g., *backing and easing*) is used if there are major changes in both speed and direction.
- Verb information (assuming verb type is known) is almost completely determined by semantics, in particular the direction in which the wind's speed or direction is changing.
- Near-synonym choice (assuming verb type and information is known) is mostly determined by author. For example, when describing an increase in wind speed, F1 and F5 prefer to use *rising*; while F2, F3 and F4 prefer to use *increasing*; this simply reflects personal idiosyncrasies and writing styles. There are also a few cases of individuals associating idiosyncratic semantic connotations with verbs. For example, when describing an increase in wind speed, F4 normally uses *increasing*, but prefers *freshening* when the final wind speed (even after its increase) is 20 knots or less; no other forecaster does this.

3.2.2 Implementation

As with time phrases, our analysis of verbs suggested that these were largely determined by individual preferences. Table 4 shows the most common verb choice in the corpus for each (type, information) pair; the word recommended by our expert for each (type, information) pair; and the word used by SumTime-Mousam for each pair. In fact there was only one disagreement between the corpus and the expert, which was for the speed-down verb; here we opted for

verb type	verb information	most common corpus phrase	expert's suggestion	phrase used in SumTime
speed	going up	increasing	increasing	increasing
speed	going down	easing	decreasing	easing
direction	clockwise	veering	veering	veering
direction	counter clockwise	backing	backing	backing
direction	to/from variable	becoming	becoming	becoming

Fig. 4. Verbs used by SumTime-Mousam

the most common corpus word, *easing*. Verb type was determined by comparing the magnitude of the direction change to the magnitude of the speed change; if both were large then a conjoined verb group was generated. Verb information was determined by examining how the verb type parameter (speed or direction) was changing.

3.2.3 Post-edit analysis

We analysed verb edits in the post-edit corpus, using the procedure described in Section 3.1.3. The results are as follows:

- *Verb Type:* This was edited in only 3% of cases. Direction was changed to speed twice as often as speed was changed to direction, so our rules for deciding whether change in speed is more important than change in direction could probably benefit from some tweaking.
- *Verb Information:* Information was almost never changed for speed verbs. For direction verbs, information was changed 0.5% of cases, which is very low, but higher than we expected. Analysis showed errors were primarily due either to cases when the forecaster disagreed with SumTime-Mousam as to whether wind direction should be described as *variable* or not, and to errors in SumTime-Mousam's handling of large changes in direction (SumTime-Mousam determined the direction of change solely by looking at the beginning and end directions, which gave the wrong result when the direction changed by more than 180 degrees).
- *Near Synonym:* The only case where forecasters regularly post-edited a verb into a near-synonym was changing *easing* to *decreasing* (this was also the only disagreement between the corpus and our expert, see Figure 4). This happened in 15% of cases. Individual differences were very striking. We have post-edit data for 9 forecasters; 5 of these changed *easing* to *decreasing* less than 5% of the time, 2 made this change over 90% of the time, with the remaining two in between (30% and 75%).

With regard to changing *easing* to *decreasing*, we were especially surprised by forecaster F5, who made this change in 92% of cases. We have data from him in our corpus of manually authored forecasts, and in this corpus he used *decreasing* only 30% of the time and *easing* 69% of the time. However, F5's behaviour in this respect may have changed over time. The manual corpus was collected from July 2000 to May 2002, and while at the beginning of this period F5 definitely preferred *easing*, at the end of this period he seemed to prefer *decreasing* [30]. Since the post-edit corpus was collected in 2003, this behaviour change may explain the above discrepancy (we have observed changes in individual writing style in other projects as well [28]). In other words, not only do individuals have idiosyncratic preferences about near-synonym choice, but these preferences may change over time.

3.2.4 Forecaster's comments

We also asked the forecasters (anonymously) about the *easing* vs. *decreasing* choice. The comments received included

(1) "Personally I prefer *decreasing* to *easing*"
(2) "I tend to think of *easing* being associated with a slower decrease and or perhaps with lower wind speeds or heights"
(3) "*Easing* is used when trying to indicate a slight decrease when condition are bad ... it is not used when conditions are quiet"
(4) (from the Figure 4 expert) "On the whole it seems safer to say *decreasing*"

Note that (2) and (3) are in fact contradictory. The forecaster who said (3) associated *easing* with bad weather, which generally means high wind speeds; while the forecaster who said (2) associated *easing* with low wind speeds. This once again supports the idea that that the mapping from data to words is highly idiosyncratic.

Comment (4), that *decreasing* is the safest choice, presumably because it has the fewest connotations, is interesting. This is supported by another puzzling fact, which is that *increasing* was edited into a near-synonym (*rising* or *freshening*) in only 1% of cases. Yet in the manually written forecasts, *increasing* was less dominant in its cluster than *easing*; *increasing* was used in 58% of cases for wind-speed-increase, whereas *easing* was used in 71% of cases for wind-speed-decrease. One explanation is that *increasing* (unlike *easing*) is 'safe' in the sense of comment (4), and hence there is no need to change it. Safety is perhaps another factor that should be considered in near-synonym choice.

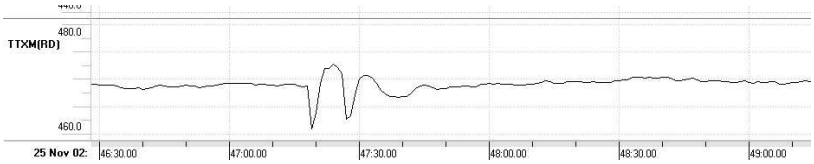


Fig. 5. Signal fragment (gas-turbine exhaust temperature): Is this an *oscillation*?

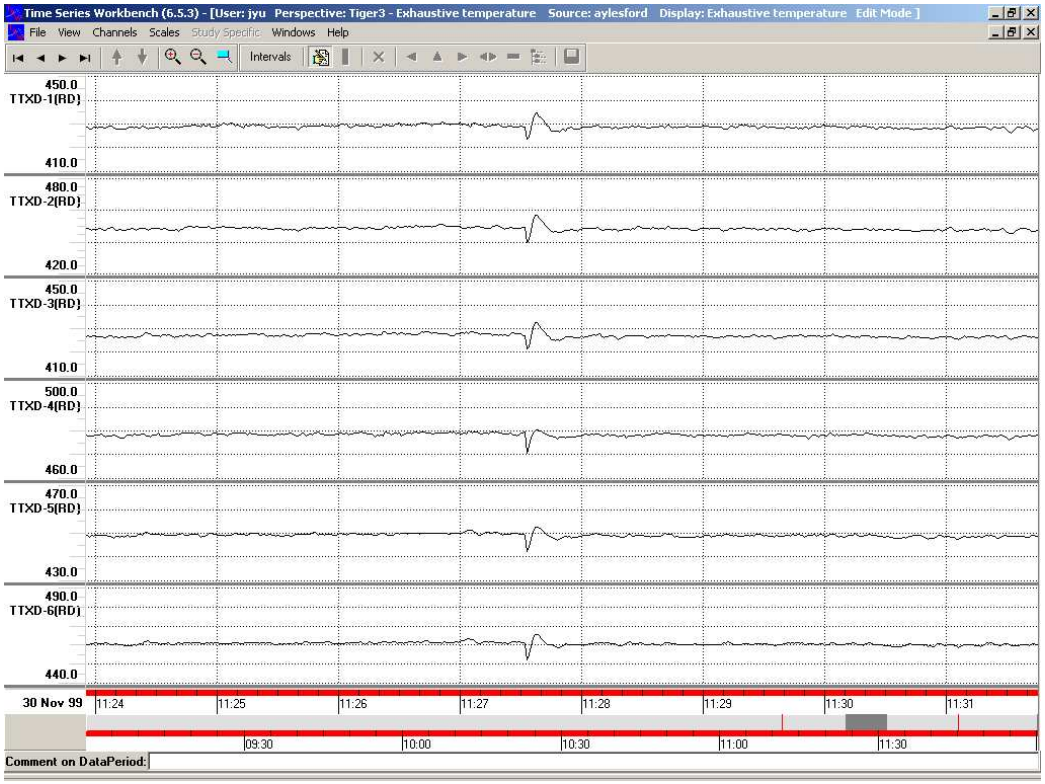
3.3 Pattern Name in SumTime-Turbine

Although this paper focuses on SumTime-Mousam, we would also like to present an interesting finding on lexical choice from another SumTime project, the SumTime-Turbine system that generates summaries of gas-turbine sensor data. One of the key lexicalisation tasks in SumTime-Turbine is choosing words to describe patterns in the time-series data, such as *spike*, *oscillation*, or *dip with oscillatory recovery*. As in SumTime-Mousam, we investigated this issue empirically, by asking two domain experts to name and describe shapes. The results of this exercise were consistent with SumTime-Mousam. There were differences between the experts in which words they preferred to use to describe a pattern (for example, *levelling out* versus *becoming steady*); and also differences in the meanings of words. For example, we specifically asked the experts if the signal shown in Figure 5 should be described as an *oscillation*; one said yes, and the other said no.

An interesting and unexpected contextual effect on lexical choice surfaced in the evaluation of SumTime-Turbine [47]. In this exercise, experts were asked to read, rate, and comment on texts produced by SumTime-Turbine. One common complaint about SumTime-Turbine's texts was lack of consistency in lexical choice between channels. For example, consider the spikes in the turbine data which is graphically shown in Figure 6. If describing the channels individually, SumTime-Turbine would call the shape in TTXD-4 a *downward spike*, and the shapes in the other 5 channels *erratic spikes*; this is because the TTXD-4 spike is almost purely downward, while the other spikes contain upward as well as downward components. SumTime-Turbine also generates combined descriptions of all these channels, and when doing this it uses the same lexicalisation strategy, hence producing the text

> At 11:26, there were erratic spikes in TTXD-1, TTXD-2, TTXD-3, TTXD-5, and TTXD-6, and a downward spike in TTXD-4.

The evaluators complained that this was misleading, because these shapes are similar enough that calling them by different names implies to the human reader a larger difference than really exists. In other words, because these shapes are similar, it is better to use the same name for all of them (probably *erratic spike* in this case), even though in isolation it would be inappropri-



Channels (from top to bottom) are TTXD-1, TTXD-2, TTXD-3. TTXD-4, TTXD-5, and TTXD-6.

Fig. 6. Shapes that Should be Given the Same Name

ate to describe the shape in TTXD-4 as an *erratic spike*. Such cases in fact were among the most common suggestions for improvements made by the SumTime-Turbine evaluators.

Hence the most appropriate word to describe a SumTime-Turbine pattern depends not only on the pattern being described, but also on how similar patterns are described elsewhere in the text.

3.4 Summary and Related Work on Lexical Choice

The strongest finding from our analysis is that the process of choosing words to describe data is idiosyncratic; different people do it in different ways. Differences include

- Different meanings associated with words; for example some people use *by evening* to mean 1800 while others use this phrase to mean 0000 (Sec-

tion 3.1.1).
- Different preferences about which phrase should be used to express a meaning; for example, some people prefer to express 1500 as *by mid afternoon* while others prefer to express this time as *by afternoon* (Section 3.1.2).
- Different fine-grained semantic connotations associated with words; for example some people think *easing* suggests low wind speed while others think this verb suggests high wind speed (Section 3.2.4).

In addition, people can change their preferences over time (Section 3.1.4); and the choice of descriptor is influenced by linguistic context and by which the words used to describe similar information elsewhere in the text. (Section 3.3). In short, the process of mapping data to words depends on the individual and on context, and not just on the data being mapped.

Perhaps the best-known previous research on choosing between words with similar meanings is Edmonds and Hirst's [8] model of near-synonym choice. However, while Edmonds and Hirst acknowledge the importance of dialect (they do not mention idiolect) and linguistic context, the focus of their research is on the impact of fine-grained semantic and pragmatic distinctions; for example the difference between *blunder* and *error* is that the former implies carelessness. Edmonds and Hirst also based their models on a published synonym dictionary [13], they did not analyse usage of words in a corpora (as discussed in this section) or conduct experiments with readers (as discussed in Section 4 of this paper).

A number of researchers have investigated the impact of collocation (one aspect of linguistic context) on lexical choice in NLG systems [7,16,35]. Gorniak and Roy [12] point out that the interpretation of spatial terms such as *middle* depends on visual context.

As far as we know, little has been written about the effect of idiolect on word choice. One partial exception is Roy [32], who learnt a model of word meanings for shape descriptors and then tested this model on new subjects by asking them to identify described shapes. Roy noted that one reason why subjects failed to identify shapes was because of variation in meanings associated with words such as colour terms. For example, an object might be described as having the colour *pink* by a subject in the training corpus, but an evaluation subject might have problems identifying the object when it was described as *pink*, because he did not consider it to have this colour. Parikh [25] has also noted that people use colour terms differently, and that this difference was not simply due to fuzzy terms in the sense of fuzzy logic.

The most in-depth analysis we know of on variation in language use is the Dictionary of American Regional English [4]. DARE is largely based on asking a representative set of Americans to respond to fill-in-the-blank questions

such as *When the wind begins to increase, you say it's* _____; in fact there were 228 different responses to this question! Unfortunately (at least from our perspective), DARE as published focuses on dialect and regional differences in word usage, not individual differences.

Several studies have been carried out in the psychological and medical communities on the differences in individual interpretations of words denoting risk and frequency. For example, Berry et al. [1] describe large differences in the numerical frequencies people associate with terms such as *common*, when these are used to describe side effects of drugs.

4 Evaluation

The post-edit analysis described above focused on the preferences of forecast writers, not forecast readers. In order to gain insight as to what was suitable for forecast readers, we conducted another experiment where we asked forecast readers to read different types of texts: human-written, computer-generated, and a hybrid which used human content but computer microplanning. This experiment suggested that consistent microplanning does indeed help forecast readers, who in fact preferred SumTime texts over human-written forecasts.

4.1 Method

We selected five short texts describing changes in the wind from forecasts which were issued in late 2000, for a particular oil rig (more precisely, we selected WIND(KTS) 10M texts, see sample forecast in Figure 2). For each of the five forecasters who were writing forecasts for that rig at this time, we selected the first forecast after 1 September 2000 which

- Started off with a prediction from 6AM to midnight, on the day the forecast was issued.
- Included in this prediction a wind statement (for 10 meters) which mentioned at least two changes in the wind.
- Was not based on numerical wind speed prediction data which stated that the wind speed was always above 20 knots, or always below 20 knots.
- Did not include typos or other obvious mistakes; and in general seemed consistent with the numerical prediction data.

For each of these wind statements, we created three variants

- *Human*: The original human-written text

Time	Wind Dir	Wind Speed 10m
06:00	S	18
09:00	S	19
12:00	S	22
15:00	SSE	23
18:00	S	24
21:00	S	22
00:00	SSW	20

Table 6
Wind data for Forecast 1 in Experiment

Human text:

S'LY 15-20 BECOMING 22-28 BY THIS EVENING. LATER VEERING S-SW 18-22

Computer text:

S 16-21 BACKING SSE 21-26 BY MID AFTERNOON, THEN VEERING S BY EARLY EVENING AND SSW 18-23 BY MIDNIGHT

Hybrid text

S 15-20 BACKING 22-28 BY EARLY EVENING, THEN VEERING SSW 18-22 BY MIDNIGHT

Fig. 7. The three texts for Forecast 1

- *Computer*: The text produced by SumTime-Mousam from the numerical prediction data
- *Hybrid*: The human-written text, manually edited to use the words (and other microplanning choices) that SumTime-Mousam would have used; in other words, the same content as the human texts, but expressed using SumTime-Mousam's language.

The purpose of the hybrid texts was to enable us to distinguish between the impact of SumTime-Mousam's content-determination and SumTime-Mousam's microplanning. Hybrid texts could have been generated automatically, by parsing the human texts into SumTime-Mousam's conceptual representation, and then regenerating these texts using SumTime-Mousam's microplanner and realiser. Because of time constraints we manually created hybrid texts for the experiment described here, but we subsequently modified SumTime-Mousam so that it can automatically generate hybrid texts in this fashion.

All texts were edited to remove information about gusts and showers; this was done in order to reduce the amount of numerical weather data that we needed

Please read the 5 short texts below, which describe the expected behaviour of the wind (at 10 meters altitude). Then, for various times in the day, tick whether you believe the wind will be less than 20kt, greater than (or equal to) 20kt, or are unsure. All forecasts cover an 18 hour period, from 6AM to midnight. You can assume that the forecasts are issued at 2AM on the same day (that is, 4 hours before the forecast period starts).

B1: S'LY 15-20 BECOMING 22-28 BY THIS EVENING. LATER VEERING S-SW 18-22.

Please cross (or tick) one box for each time

Time	<20kt	unsure	>=20kt
0600	☐	☐	☐
0900	☐	☐	☐
1200	☐	☐	☐
1500	☐	☐	☐
1800	☐	☐	☐
2100	☐	☐	☐
0000	☐	☐	☐

Fig. 8. Comprehension instructions and example question

to show subjects.

Table 6 shows the wind speed and direction data for the first forecast in the experiment, and Figure 7 shows the three text variants (human, computer, hybrid) for the first forecast. Forecast texts are written in upper case only, as this is how the original texts were distributed to forecast readers.

Based on these texts, we then created our questionnaires. Each questionnaire had four parts, and subjects were asked to complete them in this order:

- *Background information:* We asked subjects how many marine forecasts they had read in 2004; for how many years they had been reading marine forecasts; for what purpose they usually read forecasts; and also if they usually read forecasts for a particular geographical area.
- *Comprehension:* We asked subjects to read a set of five forecast texts (either the human-authored texts, the computer-generated texts, or the hybrid texts). For each forecast text, we asked subjects to say if they thought the wind speed would be less than 20 knots or greater than (or equal to) 20 knots at various time points; an example is given in Figure 8. Note that 20

There are many ways of writing texts that describe numerical weather predictions. Below, we will show you 5 sets of numerical prediction data about wind speed and direction. For each of these sets, we will also show you two possible texts that could be written to describe the wind's behaviour during this period. The texts marked (a) are the same texts that you saw in Part B above. For each data set, please indicate which text you think is easiest to read, which text you think is most accurate, and more generally which text you think would be most appropriate for someone on an offshore oil rig. Comments on the texts are also welcome.

C1:

[Table 6 shown here]

Text (a)
S'LY 15-20 BECOMING 22-28 BY THIS EVENING. LATER VEERING S-SW 18-22

Text (b)
S 16-21 BACKING SSE 21-26 BY MID AFTERNOON, THEN VEERING S BY EARLY EVENING AND SSW 18-23 BY MIDNIGHT

Please cross (or tick) one box for each time

	(a)	(b)	both same
Which text is easiest to read?	☐	☐	☐
Which text is most accurate?	☐	☐	☐
Which text is most appropriate?	☐	☐	☐

Comments (if any)

Fig. 9. Preference instructions and example question

knots is an important threshold for many operational decisions on offshore oil rigs in the North Sea.

- *Preference:* We showed subjects two variants of each of the five forecast texts (the first of which was always the one they saw in the Comprehension section) along with the corresponding numerical forecast data, and asked which variant was easier to read, which was more accurate, and which was more appropriate; subjects could also make free-text comments about this. An example is shown in Figure 9.
- *General comments:* We asked subjects for general free-text comments about the experiment or forecasts in general.

We created three different questionnaires; these are described in Table 7. Subjects were of course not told whether the texts they were reading were Computer, Human, or Hybrid texts.

version	comprehension questions asked about:	preference questions compared:	numbers returned
version 1	Human texts	Human and computer texts	13
version 2	Computer texts	Human and computer texts	23
version 3	Hybrid texts	Human and hybrid texts	36

Table 7
Questionnaire versions

4.2 Participants

The questionnaires were distributed by staff at Aerospace and Marine International to people involved in marine and offshore oil rig operations; the people who initially received the questionnaires were also asked to ask interested colleagues to fill them out. People filled out the questionnaire at a time and place of their choosing, we did not ask them to come to the university and fill out questionnaires under controlled conditions (few of our subjects would have been willing to do this, as this would have required a much more substantial time commitment).

73 people returned questionnaires; two additional people sent us free text comments but did not fill out the questionnaire. One questionnaire was less than half completed, and was discarded; this left 72 valid questionnaires. The distribution between questionnaire versions is shown in Table 7. The imbalance is due to the fact that the people who were initially sent version 3 asked more of their colleagues to also fill out the questionnaire (equal numbers of each version were originally sent out).

The respondents in general were quite experienced. 82% had read at least 100 marine forecasts in 2004, and 78% had been reading marine forecasts for at least 10 years. 80% of the respondents primarily read forecasts to support offshore oil rig operations, mostly in the North Sea.

4.3 Results

We decided during the design phase that our primary hypotheses were that

- Readers think computer texts are more appropriate than human texts
- Readers comprehend computer texts better than human texts

As there are two primary hypotheses, we regard results for these as statistically significant if $p <= 0.025$ (0.05 divided by two).

Question	Computer/Hybrid	Human	same	p value
Computer vs. human texts				
More appropriate?	**43% (77)**	**27% (49)**	**30% (53)**	**0.021**
More accurate?	51% (90)	33% (59)	15% (29)	0.011
Easier to read?	41% (74)	36% (65)	23% (41)	>0.1
Hybrid vs. human texts				
More appropriate?	38% (68)	28% (50)	34% (62)	>0.1
More accurate?	45% (80)	36% (65)	19% (34)	>0.1
Easier to read?	**51% (91)**	**17% (30)**	**33% (59)**	**<0.0001**

Table 8
Preference results: statistically significant results are in **bold**. Bracketed numbers are the actual number of times this response was selected. Numbers in a row do not always add up to 180 because some subjects did not respond to all of the preference questions.

We also decided to test 6 secondary hypotheses: computer texts are easier to read, computer texts are more accurate, hybrid texts are easier to read, hybrid texts are more accurate, hybrid texts are more appropriate, hybrid texts are better comprehended (in all cases the comparison is against human texts). We set the statistical significance threshold for these tests to p <= 0.0001; this gives a familywise (FW) error rate of 0.0499 for the experiment as a whole.

4.3.1 Preference

The preference results are shown in Table 8. For each question, we have counted the number of times that subjects preferred the computer or hybrid texts, the number of times that subjects preferred the human texts, and the number of times subjects said they were the same (no difference). Significance is calculated using chi-square. We calculated a chi-square p value both on all three numbers (computer/hybrid, human, same) and on just the computer/hybrid and human numbers (ignoring sames), and report the larger of these two p values in the table.

Our results show that forecast readers think SumTime-Mousam texts are more appropriate than human-written texts, and that Hybrid texts are easier to read than human-written texts. There is also a suggestion that SumTime-Mousam texts are more accurate, although this does not reach statistical significance for a secondary hypothesis (p <= 0.0001). It is also striking that on every single measure, users numerically prefer computer or hybrid texts over human texts.

As one of the preference texts was also shown in the comprehension section, we checked if there was a general bias for or against such previously-encountered texts. In so far as such an effect exists (it is small and not statistically significant), the bias is against the text used in the comprehension questions. Since this effectively favours the human-written texts (fewer people were asked comprehension questions about human-written texts, see Table 7) we ignored this effect.

In the free-text comments, 9 people complained about the words used in the human texts. Specific words they complained about were:

- *by evening* and *by late morning*: in fact both of these phrases were identified by our corpus analysis (Section 3.1.1) as phrases that were used to mean different times by different forecasters.
- *later*: several people said that their understanding of meteorological terminology was that *later* should mean 'after 12 hours or more', and pointed out that *later* was not being used in this way in the human-written texts.
- *becoming*: one person said this was too vague, he preferred the more meaningful verb *backing* in the computer/hybrid texts.
- *rising* and *S'LY*: people said they preferred the computer/hybrid *increasing* and *S*, respectively.

One person also complained about punctuation in the human texts.

Another subject mentioned that the interpretation of *by evening* could depend on season. This is interesting because as mentioned in Section 3.1.1, the forecasters did not change the meaning of time phrases according to season. This could be another difference between forecast writers and at least some forecast readers. Somewhat to our surprise, one subject also said that the interpretation of time phrases could depend on location; for example, he would interpret *by early evening* as meaning 1800 if he was in the North Sea and 1500 if he was in the Caspian.

4 people complained about words used in the computer/hybrid texts, or praised words used in the human texts:

- *by early evening*: two people complained about this, essentially they thought it was too vague
- *S-SE*: two people liked the way the human forecaster used this to indicate that wind direction was somewhere between S and SE

One person also complained that there was too much elision in the computer/hybrid texts.

Text type	Average number of mistakes
Human	4.23
Computer	3.43
Hybrid	2.11

Table 9
Comprehension results

4.3.2 Comprehension

For each subject, we computed the total number of mistakes that he or she made, on all 35 comprehension questions (7 questions asked about each of five forecasts). Mistakes were situations when the actual wind speed from the numerical forecast data was less than 20 knots but the subject said the wind speed was greater than or equal to 20 knots, or vice-versa. If subjects said they were unsure, this was not counted as a mistake, since being unsure is quite reasonable given the fact that weather forecasts are predictions and hence not exact.

The average number of mistakes made for each type of forecast is shown in Figure 9. We computed the statistical significance of the differences, using SPSS General Linear Model, with the number of forecasts read in 2004 and the number of years reading forecasts as covariates (the group reading Hybrid forecasts was a bit more experienced than the other groups). The difference between the Human and Computer texts is not significant. The difference between the Human and Hybrid texts has a significance of $p = 0.050$; since this is a secondary hypothesis, we do not regard this as statistically significant.

The disappointing comprehension results for the Computer texts (we had hoped for statistically significant improvement in comprehension) is largely due to a single problem in content selection in the fifth forecast. In this forecast, the wind is at 16 kts at 0600, rises to 20 kts at 1200, and remains at 20 kts until 1800. The Human text (and hence the Hybrid text, which has the same content as the Human text) explicitly describes this pattern. The Computer text, however, simply states that the wind rises from 0600 to 1800 (*S 14-19 INCREASING 18-23 BY EARLY EVENING*), without giving further details; not surprisingly, 83% of subjects thought the wind was still less than 20 knots at 1200, and 36% thought it was still less than 20kt at 1500, which is incorrect. Without this problem, comprehension scores would have been similar on Computer and Hybrid texts.

The fact that only 13 people (18% of subjects) answered comprehension questions about the human texts (Table 7) also made it more difficult to obtain statistically significant results about comprehension. In this respect it is unfortunate that we did not receive more equally balanced numbers of responses

to the three versions of the questionnaires.

Overall, then, comprehension results are 'encouraging' and suggest that improved microplanning may enhance comprehension, but this cannot be regarded as statistically significant.

We also looked specifically at how people interpreted time phrases. For example, the first phrase in the human text for the fifth forecast is *S 14-18 RISING 18-22 BY LATE MORNING*; we can judge how people interpret *by late morning* by analysing when they indicate they think the wind speed is >= 20 kts. Such analyses indicates that

- *by late morning* could be interpreted as either 0900 or 1200
- *by midday* was always interpreted as 1200
- *by mid afternoon* was always interpreted as 1500
- *by early evening* could be interpreted as either 1500 or 1800
- *by evening* could be interpreted as 1500, 1800, or 2100
- *by midnight* was always interpreted as 0000

In general these findings are quite similar to the results of our corpus analysis, which suggests that corpus analysis of how words are used in written texts can indeed provide a good first approximation of how words are interpreted by readers. Perhaps the main exception was *by early evening*, which our corpus analysis suggested usually meant 1800, but which readers thought could mean 1500 as well. In fact only 2 of the 5 forecasters who contributed to our original corpus used this phrase, so in retrospect we did not have lot of evidence that its meaning was consistent across idiolects.

It is also interesting that the time phrases that our subjects complained about in their free-text comments (*by early evening*, *by evening*, *by late morning*) are exactly the time phrases which different subjects interpreted differently. This further supports our hypothesis that generated texts should avoid words whose meaning varies substantially in different idiolects.

4.3.3 General Comments

Subjects were also given the opportunity to make general comments. The most common such comment was about textual versus tabular or graphical forecasts. Several people said they preferred tables or graphics, and others said textual and tabular forecasts should be integrated; for example, texts should just give a high-level summary, and detailed information should be presented in tables or graphs.

Several people also commented that they wanted more information about uncertainty and confidence. For example, forecast texts should explicitly indicate

when a particular prediction was very likely or not very certain. One could argue that communicating uncertainty could be a valid motivation for using vague time phrases such as *by evening*. However, as mentioned in Section 3.1.2, we found no evidence that forecasters varied time phrases according to how far in the future a forecast is for (which is a major influence on how uncertain it is).

Another person commented that some speed ranges were easier to understand than others; e.g., he found *18-22* easier to understand than *17-23*. We wondered if this could partially reflect frequency; in our main corpus of 1045 human-written forecasts, *18-22* occurs 1048 times while *17-23* occurs 0 times. *17-23* was not actually used in any of the forecast texts shown in our experiment (although it could of course have been used in another forecast which this subject had recently read).

There were also a number of comments about the content of the SumTime-Mousam texts, which are perhaps less relevant to this paper. In very general terms several people made suggestions for improving the content of the SumTime-Mousam's texts. Another person essentially commented that he was concerned about inconsistencies in the way human forecasters chose content, and wanted consistent content-determination in the forecasts that he read.

4.4 Discussion

Overall, the evaluation suggests that forecast readers prefer wind texts generated by SumTime-Mousam over human-written wind texts; this is shown by the fact that readers find SumTime-Mousam texts more appropriate by a statistically significant margin, and indeed by the fact that on every measure (including number of complaints in free-text comments), SumTime-Mousam texts score better than human-written texts. We believe this is the first time that an evaluation has shown that NLG texts are superior to human-written texts.

The comparison of human and hybrid texts suggest that much of the advantage of SumTime-Mousam texts comes from better linguistic choices, especially better lexical choices. In other words, as was suggested by our corpus analysis, human writers do not always do a good job of picking the best words to communicate the information in the forecast. We hypothesise that this could be due to the fact that many human writers write texts that they themselves would like to read; and hence have no hesitation about using *by evening* (for example) as long as they themselves interpret the term unambiguously. In other words, many writers write for their own idiolect, and do not consider

variations in their readers' idiolect. But idiolect does vary amongst readers, and hence a writer can (unintentionally!) mislead his readers if he simply assumes that they share his idiolect. SumTime-Mousam 'wins' because we have at least a crude idea of common idiolect variations, and hence can avoid terms that are known to be ambiguous, and also terms that are only used by a few people.

Psycholinguists have shown that participants in dialogues align the language that they use with each other [9,10]; that is, dialogue participants will start using similar words, syntactic structures, and so forth. Presumably this mechanism ameliorates the impact of idiolect differences in situations where two people talk to each other face to face. In the context of agreeing on how words communicate data, it presumably also helps if the dialogue participants both have access to the data being discussed (for example, they are discussing a visual scene that both can see). But alignment may not help in situations like weather forecasting, when one group of people (forecasters) produce texts for another group (forecast users), with no real-time feedback or interaction. It would be interesting to experimentally analyse words used by speakers in a face-to-face dialogue about a shared visual scene, and see if idiolect differences between speaker and hearer still create problems in such a context.

Standardised terminologies are of course another attempt at solving this problem; they essentially specify word-meaning mappings in a domain. However, standard terminologies only work if both writers and readers know and conform to them, and the comments about the use of *later* suggest that this is not always the case in weather forecasts. We explicitly asked one forecaster about *later*, and he said that he interpreted it to mean towards the end of a forecast period (which is indeed how the word is used in our corpus of human-authored forecasts). However, he added that *later* was sometimes used to mean after a period of 12 hours in shipping forecasts (which are different from forecasts for offshore oil rigs), although he believed this usage might be fading. In other words, *later* was used to mean one thing in one type of forecast (for offshore oil rigs) and something else in another type of forecast (shipping forecast); perhaps it is not surprising that people who read both types of forecasts may get confused about what *later* is supposed to mean. Incidentally, human forecast writers do not usually write both of these kinds of forecasts, so they may be less aware of this problem.

Problems with standardised terminologies are not unique to weather forecasts. For example, Berry et al. [1] show that a proposed standard terminology for frequencies of drug side-effects does not match how people actually use the proposed terms; and Cushing [5] discusses several cases where deviations from standard terminology in communications between pilots and air traffic controllers contributed to airplane crashes.

From a pragmatic applications perspective, the task of producing a weather forecast has essentially three steps: (1) adjust the NWP data based on local knowledge and expertise, (2) decide what changes in the weather to mention to the readers, and (3) write a text which describes these changes. As above, we believe SumTime-Mousam does a better job than human forecasters at step (3), but it does not even attempt the step (1), and we suspect its performance is comparable to humans (not better than humans) at step (2). This suggests that perhaps the optimal way to produce forecasts is as follows:

- Step 1 (adjust NWP data): done solely by human forecasters.
- Step 2 (content determination): done initially by SumTime-Mousam, but humans allowed to post-edit if they think they can improve this.
- Step 3 (microplanning and realisation): done by SumTime-Mousam, humans discouraged from post-editing.

The above breakdown was also more or less suggested by Goldberg et al [11].

Presumably, the fact that human forecasters seem to be better at meteorology than at writing in part reflects the fact that they are trained as meteorologists, not as writers. We wonder if similar patterns might be seen when other technically-trained professionals (such as doctors and engineers) are asked to write texts. It is interesting in this regard that the SureGen-2 NLG system [15], which is operationally used to generate some types of surgical reports, only does microplanning and realisation, and asks human doctors to specify content (via a GUI).

Our results also suggest that human texts in a corpus should not automatically be considered to be 'gold standard' texts which an NLG system should attempt to replicate [29], unless there is evidence (for example from an evaluation study with readers) that the corpus texts are in fact readable and effective. SumTime-Mousam texts would have been less readable and effective if we had blindly imitated what we saw in our corpus.

From the perspective of connecting language to the world, our results emphasise that there are substantial differences in how people select words to communicate non-linguistic information. Hence it can be dangerous to build language-to-world models purely on the basis of texts produced by a single speaker, which seems to be the case in a number of projects. Indeed, as noted in Section 3.4, Roy [32] acknowledges that this was a major cause of errors in the evaluation of his system.

5 Future Work

We are currently investigating many extensions to our SumTime research, and describe a few of these here.

5.1 Tailoring Texts to Individual Linguistic Models

SumTime-Mousam's approach to idiolect is to use a set of linguistic (especially lexical) choices which are appropriate for as many idiolects as possible. A more radical approach to idiolect differences, which we would like to explore in future research, is to attempt to tailor texts according to the idiolect of specific readers; in other words, every reader gets a text which is personalised to his or her personal idiolect. We would like to try this, and see if idiolect-tailoring substantially text readability or effectiveness. We believe that such tailoring could be especially useful for people with limited literacy skills, or for people who mainly read and write texts in unusual forms of English (such as mobile phone SMS text messages).

We believe such idiolect models could be represented as constraints and preferences on microplanning; this is the approach used by Williams [45], who attempted to build generic microplanning choice models for people with limited literacy. The biggest challenge may be acquiring the idiolect models. Some possible ways of doing this include psychometric experiments, analysis of texts written by an individual, and directly asking someone about his or her idiolect preferences; determining which acquisition technique(s) work best will be a key aspect of this research.

Individual linguistic models could be especially useful when communicating medical information such as risks of operations and frequency of side effects of medication. It is important to do this well so that patients can make informed choices about their treatment, and many people (especially those with poor maths skills) have difficulties correctly interpreting numeric probabilities and frequencies. Generic (non-tailored) texts are not effective because people interpret words such as *common* quite differently [1]. Hence there seems to be a real opportunity here to use individual linguistic models to improve the accuracy with which this important type of information is communicated to people.

5.2 Text and Graphical Presentations

We would like to explore in more depth the difference in effectiveness between textual and graphical (or tabular) presentations of data, and also how textual and graphical/tabular presentations can be integrated to produce an effective combined information presentation. Comments along this line were among the most common comments made in the SumTime-Mousam evaluation; and the experiment with intensive care data mentioned above [20] suggest that there may be a difference between what people prefer and what is best for supporting decision-making. Certainly with regard to weather forecasts, we believe the future lies with such integrated presentations. Perhaps (as suggested by some of our subjects), text should be used to provide a high-level overview, and graphs/tables should be used to provide details; or perhaps (as has been suggested to us in another domain), texts should be used to provide qualitative information (background domain knowledge, non-numeric properties, history, etc) while graphics should be used to provide quantitative data. We believe the first step in exploring this area is to conduct more experiments with professional data users (such as doctors and ship masters), we hope to do this in the future.

5.3 More Knowledgeable Content Selection

Although this is not directly relevant to this paper, we believe that content-determination in SumTime-Mousam could be substantially improved by including more knowledge in SumTime-Mousam about how people use forecasts. For instance, as mentioned in Section 4.3.2, the disappointing performance of SumTime-Mousam texts in the comprehension evaluation was due to a single mistake, when SumTime-Mousam did not describe the change in the wind in sufficient detail. We believe that the level of detail chosen by SumTime-Mousam in this case would have been appropriate if the wind had been rising from 12 kt to 16 kt, because this difference has relatively little impact on operational decisions in offshore rigs; but it was not appropriate when the wind rose from 16kt to 20kt, because 20kt is a key threshold for many decisions. In general the human forecasters are more aware of how forecasts are used than SumTime-Mousam, and we believe that incorporating such knowledge into SumTime-Mousam could significantly improve its content selection.

Also, SumTime-Mousam currently determines the content of each forecast field (wind, wave, weather, cloud, etc.) independently, without considering the overall meteorological situation. We believe that content could perhaps be better if the system first created a qualitative overview of the weather as

a whole, and used this to help decide on content [37]. For example, if the overview was "initially good, but cold front moving in around 1200", the system might then decide to explicitly mention the weather at 1200 (in other words, *by midday*) in all forecast fields, even if the meteorological parameters described by some fields did not seem to change much at this time.

6 Conclusion

The NLP, AI, and indeed cognitive science communities have paid little attention to individual differences in language usage, and indeed in many (probably most) cases simply ignored this issue completely. But people do vary considerably in both the words they prefer to use, and in what they think words 'mean' in terms of non-linguistic data.

Because we had a crude understanding of how language use varied in our domain, we were able to build an NLG system, SumTime-Mousam, which produced texts which (at least to some degree) avoided words which only occurred in one idiolect, and words whose meanings varied in different idiolects. Clearly we could have done a better job in this regard; for example SumTime-Mousam probably should not have used the time phrase *by early evening*. But despite these flaws, human subjects still considered SumTime-Mousam's texts to be more appropriate than human-written texts.

We believe that having a good knowledge of idiolect should similarly enable other NLG systems to be built which produce better-than-human texts, especially in applications (such as weather forecasting) where texts are usually written by scientific or technical experts, not professional writers. In the long term, it may be possible to tailor texts specifically to the idiolect of specific readers; but in the shorter term, much can be achieved simply by avoiding words that are problematical from an idiolect perspective, as we did in SumTime-Mousam. We look forward to the day when NLG systems are routinely used to produce texts that communicate technical information, because texts produced by NLG systems are known to be better, as well as cheaper, than texts produced by human writers.

Acknowledgements

Our thanks to the many individuals who have discussed this work with us, of which there are too many to list here. Special thanks to the meteorologists at Weathernews, and Aerospace and Marine International for their help, and to all the subjects who participated in our experiment; without whom this work

would have been impossible! Our thanks also to the anonymous reviewers for their helpful suggestions, and for encouraging us to perform the evaluation described in Section 4. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC), under grant GR/M76881.

References

[1] D. Berry, P. Knapp, T. Raynor, Is 15 per cent very common? informing people about the risks of medication side effects, International Journal of Pharmacy Practice 10 (2002) 145–151.

[2] S. Boyd, TREND: a system for generating intelligent descriptions of time-series data, in: Proceedings of the IEEE International Conference on Intelligent Processing Systems (ICIPS-1998), 1998.

[3] S. Busemann, H. Horacek, Generating air-quality reports from environmental data, in: S. Busemann, T. Becker, W. Finkler (Eds.), DFKI Workshop on Natural Language Generation, DFKI Document D-97-06, Saarbruecken, 1997.

[4] F. Cassidy, J. Hall (Eds.), Dictionary of American Regional English, Belknap, 1996.

[5] S. Cushing, Fatal Words, University of Chicago Press, 1994.

[6] J. Coch, Interactive generation and knowledge administration in MultiMeteo, in: Proceedings of the Ninth International Workshop on Natural-Language Generation (INLG-1996), 1998, pp. 300–303.

[7] L. Danlos, The Linguistic Basis of Text Generation, Cambridge University Press, 1987.

[8] P. Edmonds, G. Hirst, Near-synonymy and lexical choice, Computational Linguistics (2002) 105–144.

[9] S. Garrod, A. Anderson, Saying what you mean in dialogue: A study in conceptual and semantic co-ordination, Cognition 27 (1987) 181–218.

[10] S. Garrod, M. Pickering, Why is conversation so easy?, TRENDS in Cognitive Science 8 (2004) 8–11.

[11] E. Goldberg, N. Driedger, R. Kittredge, Using natural-language processing to produce weather forecasts, IEEE Expert 9 (2) (1994) 45–53.

[12] P. Gorniak, D. Roy, Grounded semantic composition for visual scenes, Journal of Artificial Intelligence Research 21 (2004) 429–470.

[13] P. Gove (Ed.), Webster's New Dictionary of Synonyms, Merriam-Webster, 1984.

[14] R. Grishman, R. Kittredge (Eds.), Analysing Language in Restricted Domains: Sublanguage Description and Processing, Lawrence Erlbaum, 1986.

[15] D. Hüske-Kraus, Suregen 2: A shell system for the generation of clinical documents, in: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003) (Research Notes and Demos), 2003, pp. 215–218.

[16] L. Iordanskaja, R. Kittredge, A. Polguère, Lexical selection and paraphrase in a meaning-text generation model, in: Proceedings of the 4th International Workshop on Natural Language Generation (INLG-1988), 1988, pp. 1019–1023.

[17] E. Keogh, S. Chu, D. Hart, M. Pazzani, An online algorithm for segmenting time series, in: Proceedings of IEEE International Conference on Data Mining, 2001, pp. 289–296.

[18] L. Iordanskaja, M. Kim, R. Kittredge, B. Lavoie, A. Polguère, Generation of extended bilingual statistical reports, in: Proceedings of the 14th International Conference on Computational Linguistics (COLING-1992), Vol. 3, 1992, pp. 1019–1023.

[19] K. Kukich, Design and implementation of a knowledge-based report generator, in: Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL-1983), 1983, pp. 145–150.

[20] A. Laws, Y. Freer, J. Hunter, R. Logie, N. McIntosh, J. Quinn, Generating textual summaries of graphical time series data to support medical decision making in the neonatal intensive care unit, Journal of Clinical Monitoring and Computing. In press.

[21] I. Mani, Automatic Summarization, John Benjamins, 2001.

[22] I. Mani, G. Klein, D. House, L. Hirschman, SUMMAC: A text summarization evaluation, Natural Language Engineering 8 (2002) 43–68.

[23] M. Maybury, Generating summaries from event data, Information Processing and Management 31 (5) (1995) 735–751.

[24] K. McKeown, Text Generation, Cambridge University Press, 1985.

[25] R. Parikh, Vagueness and utility: The semantics of common nouns, Linguistics and Philosophy 17 (1994) 521–535.

[26] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1992.

[27] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, 2000.

[28] E. Reiter, R. Robertson, L. Osman, Knowledge acquisition for natural language generation, in: Proceedings of the First International Conference on Natural Language Generation, 2000, pp. 217–215.

[29] E. Reiter, S. Sripada, Should corpora texts be gold standards for NLG?, in: Proceedings of the Second International Conference on Natural Language Generation, 2002, pp. 97–104.

[30] E. Reiter, S. Sripada, Learning the meaning and usage of time phrases from a parallel text-data corpus, in: Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data, 2003, pp. 78–85.

[31] E. Reiter, S. Sripada, Contextual influences on near-synonym choice, in: Proceedings of the Third International Conference on Natural Language Generation, 2004.

[32] D. Roy, Learning visually grounded words and syntax for a scene description task, Computer Speech and Language 16 (2002) 353–385.

[33] J. Siskind, A computational study of cross-situational techniques for learning word-to-meaning mappings, Cognition 61 (1996) 39–91.

[34] J. Siskind, Grounding the lexical semantics of verbs in visual perspection using force dynamics and event logic, Journal of Artificial Intelligence Research 15 (2001) 31–90.

[35] F. Smadja, K. McKeown, Automatically extracting and representing collocations for language generation, in: Proceedings of the 28th conference on Association for Computational Linguistics, 1990, pp. 252–259.

[36] R. Spence, Information Visualization, ACM Press (Pearson), 2001.

[37] S. Sripada, E. Reiter, J. Hunter, J. Yu, A two-stage model for content determination, in: Proceedings of ENLG-2001, 2001, pp. 3–10.

[38] S. Sripada, E. Reiter, I. Davy, SumTime-Mousam: Configurable marine weather forecast generator, Expert Update 6 (3) (2003) 4–10.

[39] S. Sripada, E. Reiter, J. Hunter, J. Yu, Summarising neonatal time-series data, in: Proceedings of EACL-2003, 2003, pp. 167–170.

[40] S. Sripada, E. Reiter, J. Hunter, J. Yu, Generating English summaries of time series data using the Gricean maxims, in: Proceedings of KDD-2003, 2003, pp. 187–196.

[41] S. Sripada, E. Reiter, I. Davy, K. Nilssen, Lessons from deploying NLG technology for marine weather forecast text generation., in: Proceedings of PAIS-2004, 2004, pp. 760–764.

[42] S. Sripada, E. Reiter, L. Hawizy, Evaluation of an NLG system using post-edit data: Lessons learnt, in: Proceedings of ENLG-2005, 2005, forthcoming.

[43] M. Stede, Lexicalization in natural language generation: a survey, Artificial Intelligence Review 8 (1995) 309–336.

[44] L. Wanner, Lexical choice in text generation and machine translation, Machine Translation 11 (1996) 3–35.

[45] S. Williams, Natural language generation of discourse relations for different reading levels, Ph.D. thesis, University of Aberdeen, Department of Computing Science (2004).

[46] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2000.

[47] J. Yu, Sumtime-turbine: A knowledge-based system to generate english textual summaries of gas turbine time series data, Ph.D. thesis, University of Aberdeen, Department of Computing Science (2004).

[48] J. Yu, E. Reiter, J. Hunter, C. Mellish, Choosing the content of textual summaries of large time-series data sets, Natural Language Engineering 11 (2005). In press.