# Building a Parallel Spatio-Temporal Data-Text Corpus for Summary Generation

**Ross Turner*, Somayajulu Sripada*, Ehud Reiter* and Ian P Davy****

* Dept of Computing Science, University of Aberdeen, {r.turner,yaji.sripada,e.reiter}@abdn.ac.uk

** Aerospace and Marine Intl. (UK) Banchory, Aberdeenshire, UK, idavy@weather3000.com

**Abstract**

We describe a corpus of naturally occurring road ice weather forecasts and the associated weather prediction data they are based upon. We also show how observations from an analysis of this corpus have been applied to build a prototype Natural Language Generation (NLG) system for producing road ice forecasts. While this corpus occurs in a narrow domain, it has much wider applicability due to the nature of its spatial descriptions, whose primary communicative goal is to describe the interaction between meteorological parameters and geographic features.

## 1. Introduction

Road ice weather forecasts are an essential aid for road engineers to base their salting and gritting application decisions during the winter months. They are important because consistent unnecessary treatment of a road can be as hazardous as leaving it untreated. Winter road maintenance operations also present a significant cost to local councils in the UK, where a single treatment of a road network can run into tens of thousands of pounds. Modern weather forecasting is driven by Numerical Weather Prediction (NWP) Models. Recent advances in technology have seen road ice models become increasingly localised as meteorological parameters can be measured at very fine grained spatial intervals along a road surface. One side affect of this is that the resultant output of the model becomes increasingly complex for a human expert to analyse and describe in a short textual summary.

Much previous work e.g. (Reiter et al., 2005; Boyd, 1998; Coch, 1998; Goldberg et al., 1994) has shown NLG Systems are particularly well suited for producing such textual reports; therefore, we have been developing RoadSafe, a NLG application for automatically generating road ice forecasts, in collaboration with a local weather forecasting company Aerospace and Marine International (AMI) UK. As part of the knowledge acquisition process for developing this application, we have built a parallel data-text corpus of naturally occurring road ice forecasts and associated input data, described in Section 2.. While extensively used in Machine Translation, there are few examples of using parallel corpora for NLG. (Snyder and Barzilay, 2007) describe an algorithm for automatically aligning textual descriptions to their corresponding database entries. While the Knowledge Acquisition (KA) studies carried out during the design of the SumTime Mousam weather forecast generator made extensive use of a parallel data-text corpus (Sripada et al., 2003).

The main aspect of RoadSafe that differs from other weather forecast generators, is its explicit handling of spatial data. We have been analysing the corpus to understand the process of summarising spatial data as well as understanding both the linguistic and non-linguistic requirements of the system. To this end we have annotated the corpus as described in Section 3.. In Section 4. we describe the corpus analysis process. Section 5. describes how the knowledge acquired from the corpus has guided the design of the RoadSafe summary generator. Finally we summarise our conclusions in Section 6..

## 2. Corpus Description

As stated in the previous section, the RoadSafe corpus is a parallel Data-Text corpus consisting of the output data from a NWP Road Ice Model[1] and the corresponding Road Ice Weather forecast. The corpus was collected between March 2006 and January 2008 for two local councils in England, UK, during their routine winter road maintenance operations, which last between October the 1st and April the 30th each year. The corpus consists of 431 data-text pairs with a of total of 29,857 words.

The model data and texts, described in more detail in Sections 2.1. and 2.2., comprise part of a road ice forecasting service provided by AMI UK. The aim of this service is to provide road engineers with continuous access to up to the minute weather information using various modes of presentation, such as graphs, graphics, tables and text. This information is delivered via a secure website to each council to help them plan how to grit and salt their local road network. To this end, we have developed a system to generate tabular summaries of the road ice model data in HTML format, shown in figure 1, which expert meteorologists at AMI UK can insert manually written textual wind and weather forecast statements into, via an online interface. The inserted text from the HTML files along with a file containing the raw Road Ice Model data, issued daily for both councils, form the basis of the corpus.

### 2.1. Model Data

The NWP data generated by the Road Ice Model is a large spatio-temporal data set (in order of Megabytes depending on the size of the area the model is being run for). The data contains predicted measurements of 9 Meteorological parameters (such as road surface temperature and wind

[1] The AMI UK GRIP model

Issued by Aerospace & Marine International:- Valid from noon on 26/12/2006

**24 Hour Forecast**

| Confidence/Comments | This initial analysis is carried out by the advanced RoadSafe system and is computer | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| All Routes | Min RST | Time <= 0c | Ice | Hoar Frost | Snow | Fog | MaxGusts | Rain | TS |
| Worst/Best | -1.1 /1.4 | 21:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/13 | NO /NO | NO |
| Wind (mph) | Light south to south-easterlies for the duration of the forecast period. Winds may become more moderate late morning on higher ground, but remaining southerly. (GT) | | | | | | | | |
| Weather | A mainly cloudy night, with foggy patches across much of the forecast area. Higher ground above the low cloud level could see temperatures drop below freezing during the late evening, with most western parts of the forecast area dropping below freezing by the morning. Urban areas are expected to remain marginal throughout the night. (GT) | | | | | | | | |
| Route | All routes summary worst/best | | | | | | | | |
| 1 | 0.4/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 2 | 0.7/2.0 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/10 | NO /NO | NO |
| 3 | 0.5/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 4 | 0.4/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/12 | NO /NO | NO |
| 5 | 0.7/1.9 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 6 | 0.7/2.1 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 7 | 0.9/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 8 | 0.8/2.1 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 9 | 1.4/2.1 | NA/NA | NO /NO | NO/NO | NO/NO | FOG/FOG | 13/9 | NO /NO | NO |
| 10 | 0.8/1.9 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 11 | 0.3/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 12 | -0.8 /1.5 | 22:40 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 13 | 0.1/1.6 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 14 | 0.7/1.7 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 15 | 0.4/1.7 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 16 | 0.3/1.9 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 17 | 0.4/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/9 | NO /NO | NO |
| 18 | 1.2/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/12 | NO /NO | NO |
| 19 | -0.6 /1.2 | 00:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FREEZING | 15/13 | NO /NO | NO |
| 20 | -0.9 /1.7 | 22:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 21 | -1.1 /1.4 | 21:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/13 | NO /NO | NO |
| 22 | -0.3 /1.6 | 03:20 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 23 | 0.3/1.4 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 24 | 0.0 /0.7 | 08:20 /NA | NO /NO | NO/NO | NO/NO | FREEZING/FREEZING | 15/13 | NO /NO | NO |
| 26 | -0.9 /1.2 | 21:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 27 | -0.5 /1.5 | 01:20 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 28 | -0.3 /1.6 | 00:40 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 29 | 0.3/1.3 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 30 | -0.2 /1.4 | 04:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 31 | 0.6/1.6 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 32 | 0.3/1.6 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/10 | NO /NO | NO |
| 33 | 0.2/1.2 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 34 | 0.4/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |

Route no. key:-
Green:- RST > 1.5 or RST > 1 and all routes dry.
Amber:- RST 0.5 - 1 or RST 0.5 - 1.5 if road wet or light sleet/snow present.
Red:- Any ice,hoar frost,heavy snow accumulations or RST <= 0.5 or RST < 1 with moderate snow.

Figure 1: Example Human Written Corpus Text and System Generated Table

speed) for several thousand geographical locations along a local council road network. Each geographical location is a point, indexed by a unique id, which ties it to a particular route[2], and a latitude longitude coordinate. For each point, each parameter is calculated at 20 minute intervals throughout a 24 hour forecast period. This means that there are 9 time series consisting of 72 time points associated with each point. An example of a small subset of the raw model data is given in Figure 2.

## 2.2. Forecast Texts

The forecast texts are written by 7 different expert meteorologists employed at AMI UK and consist of two paragraphs: one describing the wind conditions over the forecast area for the 24hr period between midday and midday the following day, and another describing the weather conditions. These paragraphs are generally short: typically around 1 sentence for wind descriptions and around 3 sentences for weather descriptions. The purpose of the texts are to complement the very fine grained details of the Road Ice Model presented in the tabular data with a more general weather overview. Essentially, the table is intended to present the user with worst/best conditions for specific points on each route, while the texts convey higher level information about the general area using spatial information not contained within the model. The table is also used by the experts as one of the resources to analyse the data; along with maps and graphs, and satellite imagery of the local area.

Each text contain a number of spatial, temporal and spatio-temporal descriptions describing the various weather conditions. In particular, the spatial descriptions can vary between texts depending on the road network being forecast

[2]The route number each point belongs to is indicated by the 7th and 8th digits in the ID number. For example, each point in the example in Figure 2 belongs to Route 1

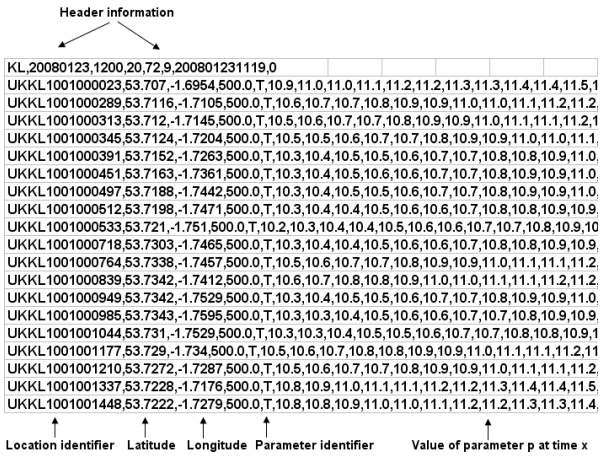


Figure 2: Raw Road Ice Model Data

for. The two road networks the forecast texts in the corpus describe belong to Hampshire and Kirklees councils in England, UK. These two areas of the UK have quite different geographies. The area covered by Kirklees council is in northern England and land locked, with some relatively high ground (above 500m) in the south west of the region. Hampshire is a much larger area, is a relatively flat county and is situated on the south coast of the UK.

## 3. Annotation Scheme

Our interest in using this corpus is to understand the requirements for building a natural language generation system to automatically produce the textual summaries from the input data. Therefore, our annotation scheme was designed to understand the content of the forecasts in terms of its underlying data; in particular, identification of message types, along with the events in the data they convey and their arguments.

Analysing a corpus (in tandem with other KA activities) is a fundamental part of deriving content selection rules when developing a NLG system for a given domain. (Callaway and Moore, 2007) describe a methodology for analysing a tutoring corpus to inform a content planning model of an Intelligent Tutoring System, while (Williams and Reiter, 2005) describe another methodology for selecting content for a system that generates personalised feedback reports. Our domain is quite different to the ones described in those papers. Therefore, the process to extract message types we used in our domain was more similar in spirit to the procedure outlined in (Geldof, 2003), which is applied to route descriptions. This procedure initially segments the corpus into message types, which are further refined on the basis of their occurring arguments.

A message in a road ice forecast typically involves an event, or series of events, that describe the state of a particular parameter or parameters. For example from Figure 1, 'Winds may become more moderate late morning on higher ground, but remaining southerly' or 'Urban areas are expected to remain marginal throughout the night'. An example of an annotated message is shown in Figure 3.

Events normally consist of a verb, describing the general trend or change in state of a parameter, along with an optional spatial phrase and/or temporal phrase describing the location and time of the event. E.g. 'Winds may become more moderate late morning on higher ground' in Figure 1. Events can also describe the general qualitative state of a parameter such as the two events that make up the message in Figure 4.

```
<MESSAGE ID="4612" MAINPARAMETER="wind">
 <EVENT ID="7706">
  <W>Winds</W>
  <W>may</W>
  <TREND ID="10380" TREND="increase">
    <VERB ID="13985">
      <W>become</W>
    </VERB>
    <W>more</W>
    <W>moderate</W>
  </TREND>
  <TEMPORALPHRASE ID="2167" TIMEPERIOD="morning">
   <W>late</W>
   <W>morning</W>
  </TEMPORALPHRASE>
  <SPATIALPHRASE AREA="part" ID="3864"
   PRIMITIVE="altitude" TYPE="geofeature">
   <W>on</W>
   <W>higher</W>
   <W>ground,</W>
  </SPATIALPHRASE>
 </EVENT>
 <W>but</W>
 <EVENT ID="7707">
  <TREND ID="13987" TREND="constant">
   <VERB ID="13986">
    <W>remaining</W>
   </VERB>
   <W>southerly.</W>
  </TREND>
 <W>(GT)</W>
 </EVENT>
</MESSAGE>
```

Figure 3: Annotation of the second sentence in the wind statement paragraph in Figure 1

```
<MESSAGE ID="1080">
 <EVENT ID="7708">
  <W>A</W>
  <W>mainly</W>
  <PARAMETER ID="2168" TYPE="cloudcover">
   <W>cloudy</W>
  </PARAMETER>
  <TEMPORALPHRASE ID="2169" TIMEPERIOD="aftermidnight">
   <W>night,</W>
  </TEMPORALPHRASE>
 </EVENT>
 <W>with</W>
 <EVENT ID="7709">
  <PARAMETER ID="2170" TYPE="fog">
   <W>foggy</W>
   <W>patches</W>
  <SPATIALPHRASE AREA="part" ID="4999" TYPE="other">
   <W>across</W>
   <W>much</W>
   <W>of</W>
   <W>the</W>
   <W>forecast</W>
   <W>area.</W>
  </SPATIALPHRASE>
  </EVENT>
 </PARAMETER>
</MESSAGE>
```

Figure 4: Annotation of the first sentence in the weather statement paragraph in Figure 1

Another important aspect of the forecasts and an important requirement of our system, are spatial descriptions. The

corpus contains many such descriptions, which generally do not consist of references to geographical landmarks such as towns or monuments. Instead they are vague, such as 'Most higher level and rural roads'. This is mainly due to the fact that when writing forecasts, forecasters try to avoid use of more specific spatial descriptions unless the pattern in the data is very clear cut, as weather system boundaries are inherently fuzzy. They also try to avoid ambiguity when they may not be aware of more provincial terminology used by road engineers and other users of the forecasts. This is challenging for an annotation scheme in this context as the spatial properties (i.e. altitude,urbanicity) used in such descriptions are not explicit in the input data, unlike temporal phrases whose value can be mapped to actual time values in the data (e.g. 0900 - 1140 for morning). Furthermore, vague concepts such as higher are open to interpretation and are dependent upon how the referent object is characterised (for example, the altitude resolution used).
From our other knowledge acquisition (KA) studies we found, as a first (and one-off) step when writing forecasts, forecasters build frames of reference to salient geographical features in the forecast region. This essentially involves familiarising themselves with parts of the region that may have an affect on the general weather pattern in the area, such as areas of high ground. This was accounted for in our annotation scheme, as spatial phrases were annotated for the spatial reference frame used. After some initial iterations spatial descriptions were classified into combinations of 4 main Frames of Reference:

- Altitude e.g. 'possible gale force gusts on higher ground'
- Direction
  - Absolute e.g. 'minimum temperatures around 5-6 degrees in the more Northern Routes'
  - Motion e.g. 'A band of rain moving across from the west'
- Population e.g. 'many urban routes will drop to be critical but remain above zero'
- Coastal Proximity - e.g. 'a few showers at first mainly along the coast'

In this context, a frame of reference is a particular perspective in which the domain can be observed; moreover, it is a set of related geographical features which allows the domain to be categorised into meaningful sub areas. How the domain is categorised is dependent upon the observer, but any classification should provide the ability to describe the location of domain entities in terms of its interaction with the underlying geography. For example, the population frame of reference is the set of town boundaries contained within the domain and its complement, classifying the domain into two sub areas: urban and rural. Altitude partitions the domain based on elevation at a chosen resolution, while coastal proximity characterises areas in terms of a specified distance from the coast, providing a binary classification: inland and coastal. Direction is based upon partitioning the domain into fixed compass bearings.

# 4. Corpus Anaylsis

Our initial methodology for analysing the corpus was to align individual words and phrases in the texts to specific data points in the input data files as described in (Reiter and Sripada, 2003). This method allows the meaning of meteorological terms to be 'grounded' in the data. However, this proved to be particularly difficult due to the sheer size of the input data and brevity of the texts. Essentially the data/text compression ratio accomplished by the texts is too high to reliably identify which part of the data set a particular phrase describes. Therefore, we had to rely on the data only as a means of comparison during the development of our system. We built a parser to parse the annotated corpus into its constituent parts. As mentioned in the previous section, our analysis described in Sections 4.1. and 4.2., concentrated on understanding the structure of message types and spatial descriptions.

## 4.1. Message Types

**Structure** Our annotation scheme described in the previous section, initially segmented the corpus into messages using full stops as a boundary. Rather than refining the messages further we concentrated on identifying the different types of events in the data that make up the sentence. Thus, in our system a message is a sentence, with event descriptions as clauses of that sentence. Event descriptions are characterised as predicate arguments structures similar to how Message types are defined by (Geldof, 2003). Table 1 shows an example of a corpus text split into events with their relevant arguments. The corpus contains a total of 1749 messages and 3598 events. The mean number of messages per forecast is 4.03 with a standard deviation of 1.165. The mean number of events per message is 2.08 with a standard deviation of 0.59.

**Types and Arguments** Event description predicates can take any number of the arguments outlined in Table 1. The only mandatory argument is the parameter, where no Area or Timeperiod is provided the event description is taken to apply to the whole forecast area and forecast period respectively. Event descriptions may also depict a trend in a parameter, such as a fall in temperature or rise in wind speed, or a parameter remaining around a particular threshold for a substantial period of time. The FrameofReference argument describes the type of location phrase used and can assume the value of any the frames or reference described in Section 3. along with a combination of these or other.
Event descriptions can be split into two very general types: global and local, as denoted by the area attribute. A value of whole indicates the event is global and applies to the whole forecast area, whereas a value of part indicates a local event is being described. After manual inspection of the event texts and their arguments it was possible for event descriptions to be further broken down into 4 types:

1. **Overview(OV)** - Describe a general weather pattern in broad terms. Normally a forecaster has used some kind of domain knowledge to perform some simple interpretation of the data, e.g. 'rain will fall over most routes' or 'Very marginal night'.

| No. | County | Date | Statement | Trend | Timeperiod | Area | FrameofReference | Parameters | MainVerb |
|---|---|---|---|---|---|---|---|---|---|
| 1.1 | KIRKLEES | 26/12/2006 | WINDSTATEMENT | constant | forecastperiod | null | null | wind | |
| 2.1 | KIRKLEES | 26/12/2006 | WINDSTATEMENT | increase | morning | part | altitude | wind | become |
| 2.2 | KIRKLEES | 26/12/2006 | WINDSTATEMENT | constant | null | null | null | wind | remaining |
| 3.1 | KIRKLEES | 26/12/2006 | WEATHERSTATEMENT | null | aftermidnight | null | null | cloudcover | |
| 3.2 | KIRKLEES | 26/12/2006 | WEATHERSTATEMENT | null | null | part | other | fog | |
| 4.1 | KIRKLEES | 26/12/2006 | WEATHERSTATEMENT | decrease | evening | part | altitude | rst | drop |
| 4.2 | KIRKLEES | 26/12/2006 | WEATHERSTATEMENT | decrease | morning | part | absolute | rst | dropping |
| 5.1 | KIRKLEES | 16/11/2006 | WEATHERSTATEMENT | constant | aftermidnight | part | population | rst | remain |

Table 1: Parsed Events Table for the Corpus Text in Figure 1

| No. | Text | Event Predicate |
|---|---|---|
| 1.1 | Light south to south-easterlies for the duration of the forecast period. | TS(Constant(Light,SSE),ForecastPeriod,Wind) |
| 2.1 | Winds may become more moderate late morning on higher ground | ST(Increase(Moderate),Morning,Part,Altitude,Wind) |
| 2.2 | remaining southerly. | TS(constant(S),Wind) |
| 3.1 | A mainly cloudy night | OV(aftermidnight,CloudCover) |
| 3.2 | foggy patches across much of the forecast area | OV(part,other,Fog) |
| 4.1 | Higher ground above the low cloud level could see temperatures drop below freezing during the late evening | NST(decrease(Subzero),evening,part,altitude,RST) |
| 4.2 | most western parts of the forecast area dropping below freezing by the morning | ST(decrease(Subzero),morning,part,absolute,RST) |
| 5.1 | Urban areas are expected to remain marginal throughout the night | NST(constant(Marginal),aftermidnight,part,population,RST) |

Table 2: Corresponding Text Entries and Event Predicates for Table 1

2. **TimeSeries(TS)** - Describe the global state of a parameter during a particular time period. These descriptions are normally derived from a forecaster inspecting time series graphs of the input data, e.g.'Temperatures will drop away quickly into the night' and always.

3. **Stationary(ST)** - Describe events at specific time points in the data. These are normally the first appearance of a particular value of a parameter or condition, e.g. '35-40 gusts 55 mph in exposed places by 0600' or it's disappearance, e.g. 'showers mostly dying out by midnight'.

4. **Non-Stationary(NST)** - Describe local weather conditions developing over a period of time and is typically spatio-temporal, e.g. 'patchy rain spreading from the northwest around midnight'.

The results of this process for Table 1 is shown Table 2.

### 4.2. Spatial Descriptions

Our corpus contains a total of 857 spatial descriptions. Of that total, 662 refer to sub areas of the spatial domain, i.e. they do not refer to the whole area. We found a substantial number of descriptions that were entirely vague; such as 'in most areas','in many places' and 'on most roads' which we classified under other as they are simply expressing proportions which can be inferred from the generated table. We also found that many spatial descriptions often involved using combinations of frames of reference such as 'high ground in the south west', similar to a map overlay operation in a Geographic information system (GIS). Table 3 shows the proportions of spatial descriptions in the corpus based on frames of reference used. As the corpus is unevenly distributed in terms of the counties the texts describe (63% Kirklees and 37% Hampshire), the distribution is skewed towards altitude due to the fact that the dominant geographical feature affecting the weather in Kirklees is an elevated area to the soutwest of the region.

| Dir. | Pop. | Alt. | Coastal Prox. | Other |
|---|---|---|---|---|
| 19% | 5% | 34% | 7% | 35% |

Table 3: Proportions of spatial descriptions by Frame of Reference used

Perhaps most interestingly, we found that certain frames of reference were used more frequently to describe certain parameters or certain events than others. For example, population is used almost exclusively (all but 14% of descriptions using the Population Frame of Reference) in descriptions of road surface temperature; while changes in precipitation type, i.e. rain turning to snow, are mainly described using altitude. These observations agree with our other KA studies we have carried out with the meteorologists at AMI where we have found that the spatial descriptions they make in the forecasts are expressing causality, i.e. the effect that a geographical feature has on a parameter, rather than being purely locative. For example a spatial description such as 'rain turning to snow in the north' may be geographically accurate, but a more useful spatial description would be 'rain turning to snow on higher ground' which also explains the cause of the meteorological event being described. Therefore, the observation of a link between Population and road surface temprature can be explained by the fact that urban roads tend to be warmer than rural roads due to their more frequent use, the population effect and their tendency to be at lower elevations; whereas changes in precipitation type are more commonly seen on higher ground where the air temperature is generally lower. The graph in Figure 5 shows the distribution of spatial de-

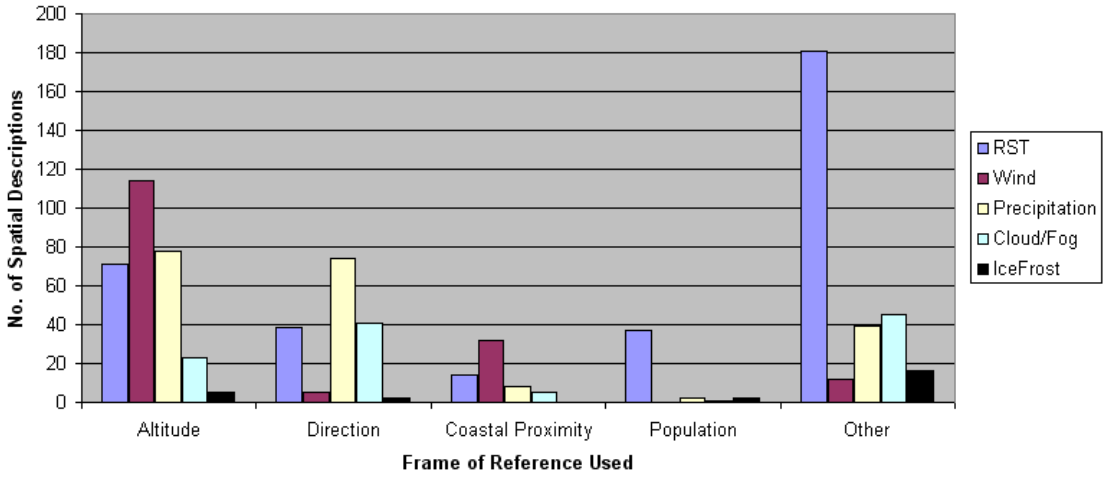


Figure 5: Spatial Description Frequencies by Frame of Reference Used and Parameter Described

scriptions in the corpus. The large number of spatial descriptions describing road surface temperature being classified under other can be accounted for by the fact that descriptions such as 'on most roads' and 'all roads' were included in this category. While the graph shows a clear preference for describing the location of events affecting wind in terms of altitude, the affect of the dominant geographical features affecting the weather pattern in a region becomes more clear when the distribution is viewed between the two counties. 93% of all spatial descriptions describing wind in Kirklees use the Altitude Frame of Reference as opposed to 30% in Hampshire which incorporates a coastline. Here the location of wind events are mainly described in terms of their proximity to the coast (54%).

From a linguistic perspective, the spatial descriptions in the corpus are relatively simple syntactically. They are normally a single prepositional phrase - 'on higher ground' or concatenation of prepositional phrases - 'in some places in the north'. As would be expected, vague quantifiers such as all,some,most,many and few are fairly commonly used (27% of all phrases). These are used where a forecaster is referring to the proportion of points contained within the bounding area of the event.

The number of variations in lexical choices for referring to values within a reference frame vary between each individual frame in the corpus. Direction values, as would be expected, are referred to using the standard points of the compass; coastal proximity, is also referred to using the standard distinctions between inland and coastal; within population, while urban and rural distinctions are mainly used, urban areas are also sometimes described using near synonyms such as suburb and residential as well as occasionally by proper names. The greatest variation is seen within references to altitudes, where the same elevated area can be referred using a modified noun phrase such as high ground, common nouns such as hills or moors, or its actual height value such as 300m.

## 5. Implications for Summary Generation

For building a prototype road ice forecast generator, it was clear from observing how the input data mapped to the output texts that a number of extensions to the architecture for data-to-text systems proposed in (Reiter, 2007), were required to facilitate the handling of spatial data. We describe these extensions in Section 5.1. and highlight how observations from the corpus analysis influenced Content Section in our system in Section 5.2.. Finally, Figure 8 provides an example of a weather statement generated by the RoadSafe system.

### 5.1. General System Architecture

As mentioned in the previous section, our system extends the general architecture outlined in (Reiter, 2007). The extended architecture used in our RoadSafe prototype, the details of which are to be published elsewhere, contains 3 extensions; a Geographic Characterisation stage, a Spatial Reasoning module and a Spatial Database, which we describe next.

**Geographic Characterisation** Characterisation is defined as finding compact descriptions of data (Miller and Han, 2001). As the only spatial information contained within the input data to our system described in 2.1., is a latitude-longitude coordinate pair, it is necessary to characterise the data in terms of the frames of reference we have identified in the output text. We treat this as a one-off preprocessing step for each new forecast area in the initial data analysis stage of our architecture. This process is a form of data enrichment (Miller and Han, 2001) performed by combining the data with other external spatial data sources. Essentially each point in the input data set is assigned additional spatial properties as shown in Figure 6.

**Spatial Database** The spatial database acts as a store for the external spatial data information used in the Geographic Characterisation step. Frames of reference are stored as thematic layers in the database. For example, the altitude frame of reference is stored as the set of all elevation contour lines at a given resolution for the area, while the population frame of reference is the set of all town boundary

| ID | Latitude | Longitude |
|---|---|---|
| UKKL1001000023 | 53.707 | -1.6954 |
| UKKL1001000289 | 53.7116 | -1.7105 |
| UKKL1001000313 | 53.712 | -1.7145 |
| UKKL1001000345 | 53.7124 | -1.7204 |
| UKKL1001000391 | 53.7152 | -1.7263 |
| UKKL1001000451 | 53.7163 | -1.7361 |
| UKKL1001000497 | 53.7188 | -1.7442 |
| UKKL1001000512 | 53.7198 | -1.7471 |
| UKKL1001000533 | 53.721 | -1.751 |

| ID | Latitude | Longitude | Height | Direction | UrbanRural |
|---|---|---|---|---|---|
| UKKL1001000023 | 53.707 | -1.6954 | 0 | NORTHNORTHEAST | Dewsbury |
| UKKL1001000289 | 53.7116 | -1.7105 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000313 | 53.712 | -1.7145 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000345 | 53.7124 | -1.7204 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000391 | 53.7152 | -1.7263 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000451 | 53.7163 | -1.7361 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000497 | 53.7188 | -1.7442 | 100 | NORTHNORTHEAST | RURAL |
| UKKL1001000512 | 53.7198 | -1.7471 | 100 | NORTHNORTHEAST | RURAL |
| UKKL1001000533 | 53.721 | -1.751 | 100 | NORTHNORTHEAST | RURAL |

Figure 6: Geographic Characterisation of input data

polygons and it's complement within the area. The spatial database also allows the system to perform topological queries on the stored data.

**Spatial Reasoning module** The Spatial Reasoning module acts as layer between the main system and the spatial database. It performs two main functions: the first is to perform the geographic characterisation of the input data, the second is to provide functionality for the rest of the system to perform higher level spatial queries. These queries can range from combining frames of reference to adding location information to system events. Together with the Spatial database, the Spatial Reasoner acts in a similar way to a limited GIS system.

### 5.2. Content Selection

Our spatio-temporal analysis method, described in more detail in (Turner et al., 2007), explicitly takes into account the geographic characterisation of the forecast region. The data is clustered according to the frames of reference we have identified in our corpus providing results that can be easily mapped to spatial descriptions. The clustering method is also density based, applying proportions to each cluster that provide a simple mechanism for the system to include vague quantifiers in the generated spatial descriptions, such as those described in Section 4.2..

As the spatial descriptions generated by our system should express the effect of geographic features on weather conditions, our system implements a preference ordering over the way it selects the frame of reference it uses in the description. This is dependent on the type of event and parameter being described and is implemented based upon our corpus analysis described in Section 4.2.. An example of the preference ordering for describing road surface temperature is shown in Figure 7. This approach has similarities to the one described in (Kelleher and Kruijff, 2006), as it takes the context into account when considering which properties to use in the resulting description. Lexical choice is done simply by choosing the option for the reference frame value with the highest frequency in the corpus. The only exception to this is altitudes, which are described using the actual height value as requested by experts.

1. Altitude
2. Population
3. Coastal Proximity
4. Direction

Figure 7: Preference Ordering for Road Surface Temperature Events

*Road surface temperatures will fall below zero on most routes by evening. Wintry precipitation will affect some routes during the afternoon and evening clearing for a time by evening, falling as snow flurries in some places above 400M at first. Snow clearing by 21:40. Road surface temperatures will fall slowly during the mid afternoon and evening, reaching zero in some places above 400M by 18:00. Ice and hoar frost will affect most routes during the evening and tonight. Hoar frost turning heavy by evening except in areas below 100M. Rain will affect all routes during tonight and tomorrow morning turning heavy tomorrow morning except in far southern and north western areas.*

Figure 8: Example Weather Statement Generated by RoadSafe

## 6. Future Work and Conclusions

We have described a parallel data-text corpus of spatio-temporal NWP data and its corresponding output texts. The corpus was built as part of the KA studies carried out during the development of RoadSafe, a prototype NLG system that automatically produce road ice forecasts. The system is currently installed at AMI's premises and generating draft forecast texts which are being post edited by forecasters before being released to clients. The purpose of this evaluation is to further improve the quality of the texts in preparation for a full scale user evaluation in which we plan to compare the quality of the generated texts with human written ones.

As the input data in the corpus is large and complex, aligning words and phrases in the texts to actual data points was not possible. An analysis of the spatial descriptions in the corpus found that it contained no definite descriptions and almost no reference to named landmarks. Instead of named landmarks, spatial descriptions describe the location of events in the data in terms of frames of reference, which are sets of related geographical features that affect the general weather pattern. We found that this is because the purpose of the spatial descriptions is to communicate the effect of the geography on the data rather than be purely locative. Our observations led us to identify 3 additional modules necessary to incorporate into the architecture of our NLG system: geographic characterisation, spatial reasoning and a spatial database.

## Acknowledgments

Many thanks to our collaborators at Aerospace and Marine International UK, especially Keith Thomson and the other Meteorologists, for their helpful feedback and comments.

The RoadSafe project is supported jointly by Aerospace and Marine International UK, and the UK Engineering and Physical Sciences Research Council (EPSRC), under a CASE PhD studentship.

## 7. References

S. Boyd. 1998. Trend: a system for generating intelligent descriptions of time-series data. In *IEEE International Conference on Intelligent Processing Systems (ICIPS1998)*.

Charles B. Callaway and Johanna D. Moore. 2007. Determining tutorial remediation strategies from a corpus of human-human tutoring dialogues. In *Proceedings of the 11th European Workshop on Natural Language Generation*, Schloss Dagstuhl, Germany, June.

J. Coch. 1998. Multimeteo: multilingual production of weather forecasts. *ELRA Newsletter*, 3(2).

S. Geldof. 2003. Corpus analysis for nlg. In Reiter E., Horacek H., and van Deemter K., editors, *9th European Workshop on NLG*, pages 31–38.

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

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 1041–1048.

Harvey J. Miller and Jiawei Han. 2001. Geographic data mining and knowledge discovery: An overview. In *Geographic Data Mining and Knowledge Discovery*, chapter 1, pages 1–32. Taylor & Francis.

Ehud Reiter and Somayajulu Sripada. 2003. Learning the meaning and usage of time phrases from a parallel text-data corpus. In *HLT-NAACL 2003 Workshop: Learning Word Meaning from Non-Linguistic Data*, pages 78–85, Edmonton, Alberta, Canada.

E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy. 2005. Choosing words in computer-generated weather forecasts. In *Artificial Intelligence*, volume 67, pages 137–169.

E. Reiter. 2007. An architecture for data-to-text systems. In *ENLG07*, pages 97–104.

Benjamin Snyder and Regina Barzilay. 2007. Database-text alignment via structured multilabel classification. In Manuela M. Veloso, editor, *IJCAI*, pages 1713–1718.

S. Sripada, E. Reiter, J. Hunter, and J. Yu. 2003. Exploiting a parallel text-data corpus. In *Proceedings of Corpus Linguistics*, pages 734–743, Lancaster, UK.

R. Turner, S. Sripada, E. Reiter, and I. Davy. 2007. Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data. In *Applications and Innovations in Intelligent Systems XV*, pages 75–88.

S. Williams and E. Reiter. 2005. Deriving content selection rules from a corpus of non-naturally occurring documents for a novel nlg application. In *Corpus Linguistics 2005 Workshop on Using Corpora for Natural Language Generation*.