What is in a text and what does it do: Qualitative Evaluations of an NLG system – the BT-Nurse – using content analysis and discourse analysis.

Rahul Sambaraju
Queen Margaret Univ, UK
rsambaraju@qmu.ac.uk

Ehud Reiter
Univ of Aberdeen, UK
e.reiter@abdn.ac.uk

Robert Logie
Univ of Edinburgh, UK
rlogie@staffmail.ed.ac.uk

Andy McKinlay
Univ of Edinburgh, UK
hos.ppls@ed.ac.uk

Chris McVittie
Queen Margaret Univ, UK
cmcvittie@qmu.ac.uk

Albert Gatt
Univ of Malta, Malta
albert.gatt@um.edu.mt

Cindy Sykes
Edinburgh Royal Infirmary, U.K
Cindy.Sykes@luht.scot.nhs.uk

Abstract

Evaluations of NLG systems generally are quantiative, that is, based on corpus comparison statistics and/or results of experiments with people. Outcomes of such evaluations are important in demonstrating whether or not an NLG system is successful, but leave gaps in understanding why this is the case. Alternatively, qualitative evaluations carried out by experts provide knowledge on where a system needs to be improved. In this paper we describe two such evaluations carried out for the BT-Nurse system, using two different methodologies (content analysis and discourse analysis). The outcomes of such evaluations are discussed in comparison to what was learnt from a quantitiave evaluation of BT-Nurse. Implications for the role of similar evaluations in NLG are also discussed.

1 Introduction

Natural-Language Generation (NLG) systems are usually evaluated quantitatively, by measuring impact on task performance, human opinions on Likert-like scales, and/or similarity to a gold-standard corpus. While such evaluations are essential, we believe there is also a role for qualitative evaluations, especially when the goal of the evaluation is formative that is, assessing weaknesses and identifying how the NLG system could be improved.

In this paper we describe how we used two qualitative methodologies, content analysis and discourse analysis, to evaluate texts produced by the BT-Nurse system (Hunter et al., 2011). These methodologies require a human analyst to read and analyse the generated texts; and indeed for both types of analysis it is helpful to conduct a similar analysis of human-written corpus texts, so that generated texts can be compared to manually-authored texts. From a practical perspective this means that only a relatively limited number of texts can be analysed using these methodologies; but nevertheless we believe they can substantially help in formative evaluation of NLG systems.

2 Background

2.1 Evaluation in NLG

The great majority of published evaluations of NLG systems are quantitative: as described by Reiter and Belz (2009), they either measure the impact of a generated text on task performance, ask human subjects to rate generated texts on a Likert-like scale, or compare the similarity of generated texts to corpus texts using automatic metrics such as BLEU (Papineni et al., 2002). Reiter and Belz point out that many human-based quantitative NLG evaluations also solicit free-text comments from their subjects, and these are very helpful in diagnosing and fixing problems in generated texts. Soliciting such comments, however is usually a secondary goal of evaluations of NLG systems, the primary goal being quantitative.

One instance of the use of qualitative methodologies in evaluating NLG systems was that by McKinlay et al (2010) who used discourse analysis to analyse texts generated by the BT45 NLG system (Portet et al., 2009). The evaluation revealed certain problems with generated texts, such as a

poor narrative structure (Reiter et al., 2008). The discourse analysis work presented here uses a similar approach to McKinlay et al (2010).

2.2 BT-Nurse

The BT-Nurse system (Hunter et al., 2011) generates nursing shift handover reports for babies in a Neonatal Intensive Care Unit (NICU), from data stored in the baby's electronic medical record. The input data include numeric time-series data (e.g., heart rate), ad-hoc structured data (e.g., lab results), and descriptions of actions and observations of medical and nursing staff (such as administering drugs and performing surgical procedures). The handover report is produced at the end of a 12-hour nursing shift, and is given to the incoming nurse on the next shift as part of the handover process. Its purpose is to help the incoming nurse plan her care activities, and also ensure that she is aware of the baby's circumstances.

BT-Nurse is part of the BabyTalk family of systems (Gatt et al., 2009), and like other BabyTalk systems it combines signal analysis and pattern matching, data interpretation based on expert medical knowledge, and NLG techniques. It was developed in close consultation with NICU nurses, and used no input data other than what was stored in the electronic medical record.

As part of the development process, an expert NICU nurse wrote a corpus of 32 example nursing summaries based on data in the medical record related to 10 babies collated over a period of 3 months. The babies concerned here were diagnosed to have a range of medical conditions affecting various body systems at differing levels of pathology. These texts differed from real-world existing handover reports in two ways: (1) they were much longer and more detailed (on-duty nurses do not have the time to write detailed shift-handover reports), and (2) they were purely based on the electronic patient record (and not, for example, on visual observation of the baby).

BT-Nurse was designed so that the output texts resemble corpus texts with the aim of complementing nurses engaged in their duties. In the remainder of this paper, *corpus text* refers to one of the specially written summaries above for the purposes of designing the system, and *actual handover text* refers to a real-world handover report written by an on-duty nurse for a baby she was looking after. At the time of analysis the BT-Nurse was focusing on producing texts that described only the baby's clinical history and respiratory system, so qualitative analyses were limited to these parts of the corpus texts, actual handover texts and BT-Nurse generated texts.

An extract from an actual handover text is shown in Figure 1, an extract from nurse-written corpus text is shown in Figure 2, and an extract from the corresponding BT-Nurse text is shown in Figure 3 (the complete texts are several pages long).

Nurse Shift Summary	
dbpatid	103362
working weight	840.0
nursedin	Incubator
Problems during shift	
resp distress	Oxygen requirement
stools	Changed
Respiratory	
resp support	SIMV
resp notes	rate 25 pressure 20/4
Fluids / Feeds	
prescribed daily feeds	175.0
type milk	Breast
vol mil per feed	6.0
feed given by	OGT
frequency of feeds	hly
adequate urine vol	Yes
stools	Changed
Medication	
drugs	
Social	
social visited	Mother|Father
length parent visit	1-3 hours
Developmental support	
developmental support	Incubator cover used|Positioning aids used|Containment given
Notes	
other notes	settled night, no change in ventilation. Obs stable in 438-44% O2. Suction x2 yeilding 2-3 M ET and Oral. Pud BO. Drugs given as prescribed.
signature designation	Staff nurse

Figure 1: Actual handover text

2.3 Quantitative Evaluation:

BT-Nurse was evaluated by deploying the system on-ward in the NICU, asking nurses to use it as part of the shift handover process, and soliciting ratings and free-text comments from nurses as to the understandability, accuracy, and helpfulness of BT-Nurse texts (Hunter et al., 2011). Overall, 90% of nurses thought BT-Nurse texts were understand-

able, 70% thought they were completely accurate, and 60% thought they were helpful. Free-text comments focused on specific content issues (requests for additional information, complaints about incorrect content, suggestions to remove content). There were fewer comments about language issues. These tended to be fairly specific when addressing microplanning issues (for example "*would prefer not to see the word 'since' with the date*"), but vaguer when addressing document-planning and narrative issues (for example, "*this summary does not convey the feeling that the baby has made progress*" and "*The above comments are accurate statements, however they do not present a 'picture' of current condition*").

This evaluation worked well from the perspective of getting some numbers on the system's perceived utility, which was its primary goal. However, from the perspective of diagnosing problems and suggesting enhancements, it worked much better for content and low-level phrasing issues than for document structure and narrative issues. This suggests that other methodologies, probably involving analysts with specialist expertise in narrative and structural issues, might be needed to diagnose and address these issues. In the remainder of this paper we describe how we used two such methodologies, content analysis and discourse analysis, to gain a better understanding of BT-Nurse's deficiencies from this perspective.

3 Evaluation using content analysis

3.1 Content Analysis

Content analysis is widely used as a data exploratory tool by qualitative researchers in psychology, linguistics and other social sciences. In content analysis, qualitative data, mainly texts, are coded according to some coding scheme which is usually predetermined, either from previous research or researcher expectations. Following this, frequencies can be calculated to enable a numerical comparison. A unit of analysis (sentence, paragraph or a page) is identified and classified according to specific codes. These codes could either be descriptive or analytic (Richards, 2009). The level of coding and what is done on these codes depends to a good extent on the research question (Saldaña, 2009). Here we were interested in what sorts of data representation was contained within BT-Nurse generated texts as compared to that in corpus texts. Therefore, the analysis had as its focus identifying content in these texts which was reflective of representations of data in textual form.

3.2 Method

The corpus of 32 nurse written texts was first analyzed to come up with a coding scheme; this was then applied to corpus texts, BT-Nurse texts, and actual handover texts. The extent of corpus texts subjected to analysis was defined in two ways: (1) focus here was on identifying lexical items that communicated 'complex' information; for example temporal relations and causality (but not simple statements of parameter values such as heart rate) and (2) as mentioned above those parts of the texts relating to babies' clinical history and respiratory system only. Analysis led to the identification of various items that were abstracted into higher order 'codes' forming the coding scheme A sample of nurse written corpus text was made available to a doctorate student with brief notes on what was being looked for in those texts and to form some sort of codes relating to data representation. Codes identified were checked against the first authors' for agreement (Cohen's $\kappa = .74$), in line with common practices of doing such analyses with codes (Saldaña, 2009). The coding scheme presented here was used to calculate frequency of occurrence of each item in nurse written corpus texts, BT-Nurse summaries and actual handover texts.

3.3 Coding Scheme

Items on the coding scheme can be usefully differentiated into descriptive items that describe various particulars of information and inferential elements, which provide for inferences amongst data items.

1) Descriptive items in the coding scheme:

a) <u>*Temporal information*</u>: Data items, have time stamps, that is, they are presented as having occurred at *some* time:

i. <u>Specific clock times</u>: Temporal information is provided in terms of normative clock times – 11:00 or 13:30. E.g.: "The last blood gas was at *18:30* and no changes were made".
ii. <u>Vague temporal markers</u>: Temporal information provided in terms that do not readily specify the exact point in clock times, such as: 'morning', 'a few minutes ago' and others. E.g.: "but this *afternoon* he also looks pale".

iii. Shift time: shift start and end times are made use of as temporal markers. E.g.: "Insulin *just* commenced".

iv. References to other events: Clock time is provided for one event 'A' and another event 'B' is temporally located via references to 'A'. E.g.: "He *received morphine prior to intubation at 00:30*; no spontaneous respiratory effort *noted since being re-ventilated*".

b) *Time intervals*: Provision of temporal *information* for events that do not have a single temporal marker but two that refer to the start and end times, is made as unitary condensed entities. E.g.: "However, *over the day* his oxygen requirements generally have come down from 30% to 25%".

c) *Trends in parameter values*: Recordings of parameter values are made to capture changes in *the* parameter over a period of time providing the initial and final values along with the direction of such change. E.g.: "Baseline SpO2 *drifted down* from *95% to 88%* accompanied by *increasing* SpO2 variability associated with handling". [SpO2 – Oxygen saturation in blood]

d) *Evaluations of parameter values*: Parameter *values* are also evaluated either in terms of what is physiologically normal or in terms of what is locally taken to be normal for that particular shift and that particular baby. E.g.: "ABG at 23:10 showed CO2 *increased* from 7.7 to 9.27 in three hours". [ABG – Arterial Blood Gas; CO2 – Carbon Di-oxide]

e) *Events in temporal relation with other events*: *Information* about certain events and data items is presented as preceding or succeeding other events. E.g.: "Received one dose of surfactant *after* admission to NNU". [NNU – Neonatal Unit]

2) Inferential items in the coding scheme:

a) *Event characterizations*: Events are those data items that indicate a recording of a parameter value, a change in a parameter value, interventions and such.

i. Events 'marked up': Events are presented as important within the local context via providing clock times and describing other events in relation to this particular event. The use of one event 'A' as a temporal anchor for another 'B' presents it as consequential to 'A'. E.g.: *"Electively re-intubated at 00:30* to CMV rate 50, pressures 18/4 in 30% oxygen. *On ventilation*, oxygen requirement reduced to 30% and *ABG initially improved*". [CMV – Continuous Mandatory Ventilation]

ii. Event presented as forming a context for other events: Events are presented as occurring over a period of time and then other events are presented as having occurred in the contextual background of the former event. E.g.: "*While on BiPAP*, oxygen requirement increased to 50% by 23:00." [BiPAP – Bi-level Positive Airway Pressure]

b) *Evaluation*: The presentation of parameter values or medical interventions forms an evaluation of a prior event or parameter value. E.g.: "ABG taken 2 hours post-extubation was *reasonably good*: pH 7.33 and pCO2 7.08". [pCO2 – Partial pressure of Carbon Di-oxide]. Evaluative information together with the temporal marker anchored in 'extubation' serves to present changes in ABG as an evaluation of the outcomes of 'extubation'.

c) *Parameters grouped together*: Two or more dissimilar parameter descriptions are made together with a conjunctive indicating some sort of an association between the two parameters. E.g.: "*Desaturation* to 15% *with bradycardia* to 50-60s".

d) *Grouping similar events*: descriptions of two or more event descriptions are juxtaposed to each other. As above these descriptions are of their temporal status, outcomes and such. E.g.: "*Tried off CPAP once* but *put back on after 30 minutes* due to increased work of breathing; otherwise has not been off CPAP*".* [CPAP – Continuous Positive Airway Pressure]. Here, descriptions attend to two events: being on CPAP and being off CPAP. Including descriptions on these two events provides for inferences as to the reasons, outcomes and other such features of those events.

e) *Causation*: Events are presented to be causally related to each other either via an explicit discourse marker or presenting the parameter recordings or events in temporal relation to each other that makes relevant causal links between them. E.g.: "several episodes (about 3 per hour) of bradycardia with desaturation that *only resolved after* stimulation or increase in FiO_2".

3.4 Results and Discussion

Results shown in Table 1 include frequencies of coding items in corpus texts, BT-Nurse texts, and in actual handover texts. BT-Nurse texts score

	Coding item	Human Corpus	BT-Nurse	Actual Handover
1)	**Descriptive Items**			
a)	Temporal information			
i)	Vague	4	29	19
ii)	Shift times	3	27	19
iii)		8	2	4
b)	Time Periods	27	0	17
c)	Trends	19	99	31
d)	Evaluations	13	38	28
e)	Temporal relations	23	15	13
2)	**Inferential Items**			
a)	Event presentations			
i)	'Marked up'	8	2	10
ii)	Context forming	5	0	16
b)	Evaluations	8	0	16
c)	Grouping Parameters	14	10	12
d)	Grouping events	8	7	8
e)	Causation	8	0	17

Table 1: Frequencies of coding items.

more on descriptive items: they contain quantitatively more temporal information, higher reporting of trends in parameters, and more items of evaluation on parameter values. However, they do not contain representations of 'time intervals' (1 (b)). Representing time intervals can be thought of as using at least two time stamps on a temporal axis: the 'start' and 'end' (Adlassnig et al., 2006). BT-Nurse software apparently does not enable such representation, the outcome of which is reflected in item 2 (a) ii. The lack of representing an event 'B' as occurring over a period of time in BT-Nurse texts does not make for characterizing an event 'A' as occurring in the background context of the ongoing event 'B' (the event 'B' having 'start' time and an 'end' time). Although BT-Nurse texts do contain inferential items, overall these items are less frequent compared to nurse written corpus texts. Moreover, inferential items presented do not readily make it clear as to the nature of the relationship (see 4.3 below). These findings then reveal how representing temporal information has outcomes on other forms of data representation in BT-Nurse texts, and thus contribute to the design of the system.

Analysis of actual handover texts served to attend to issues of external validity of the evaluation. Results indicate that actual handover texts are more similar to nurse written corpus texts in containing more inference enabling items and more instances of explicit inferences. These results at one level are not very surprising as nurses engaged in doing their duties would arguably require information of this sort. In that sense, this evaluation has pointed to features of data-to-text systems that are indicative of the sorts of requirements users of these systems have. Thus, by providing more information on relevant parameters, a better trend detection ability and producing an easily usable textual document, BT-Nurse has significant potential to enhance nurse care planning in the NICU.

4 Evaluation using discourse analysis

4.1 Discourse analysis:

The other qualitative evaluation employed discourse analysis, which has as its focus pragmatic outcomes of texts. Discourse analysis specializes in the analysis of spoken or written discourse, as a topic of study in its own right (McKinlay et al., 2008). In contrast to content analysis, discourse analysis takes as its focus the action-orientation of discourse. The analyst focuses on identifying properties within the text, such as the design of individual discourse elements and how sets of such elements are sequentially organized in order to accomplish particular pragmatic outcomes in that, discourse is considered for the sorts of actions that ensue from specific forms of usage. Discourse analysis differs from other forms of linguistic analyses (such as those based on Rhetorical Structural Theory (Thompson et al., 1987) or Discourse Structural Relations (Hovy, 1993)) in focusing on the ways in which language gets used for specific outcomes, that is, the focus is on an analysis of discourse rather than on linguistic features of any fixed 'unit' of text. The analysis seeks to draw out those aspects of discourse production and reception which are treated by participants in a particular discursive interaction as 'everyday' or 'commonsense' but which are, at the same time, central to a full understanding of what is written. Outcomes of discourse analysis then are of a psychological nature than merely linguistic.

A prior use of such methodology in conducting an evaluation of another data-to-text system – BT-45 – showed that corpus texts written by domain experts had better narrative structures than system generated texts (McKinlay et al., 2010). These are considered to be desirable aspects in texts generated by NLG systems (Reiter et al., 2008), therefore we conducted an evaluation using this methodology.

This evaluation was in fact conducted on a preliminary version of BT-Nurse, and some changes were made to the final version of BT-Nurse based on this evaluation; for example the way 'causality' was expressed was changed in some cases to enhance clarity. The content analysis and quantitative evaluations, in contrast, were carried out on the final version of BT-Nurse.

4.2 Method

For reasons of illustration and space we provide here a comparative analysis of one nurse written corpus text and the corresponding BT-Nurse generated text for one 12 hour shift. This particular shift summary pair was randomly selected amongst the 32 pairs available. Analysis provided here aims to demonstrate the utility of discourse analysis in formative evaluations of NLG systems. The analysis was conducted by three of the authors on an extract taken from each of these texts that detailed occurrences within the shift related to baby's respiratory system. Analysis involved identification of lexical items (words, sentences and such) that were selected for inclusion and how they were sequentially combined within the summary. The identification of such was considered for the sorts of outcomes made available. Here, this led to the identification of three pragmatic discursive features present in nurse written corpus texts. These analytic findings were subsequently made use of in evaluating BT-Nurse output texts.

4.3 Analyses:

Figure 2 is an example nurse written corpus text that includes descriptions of baby's respiratory status. Figure 3 is the corresponding BT-Nurse generated text produced for the same baby for the same shift. It can readily be seen that they are similar in terms of producing a list of events that occurred during the said shift. The following comparative analysis aims to show the pragmatic outcomes of these two summaries. For the purposes of this paper, the analysis is presented along three main pragmatic features:

a) *Foregrounding the actor*:

The summary in Figure 2 begins with the admission of the baby and the status of his respiration. Through the use of 'he' at line 2, the author explicitly introduces the baby as a character. This first item also specifies a particular, desirable health status for the baby at that time: 'in air'. This provides a context for the rest of the description organized around the baby as a central character in a sequence of events. The final item selected for inclusion at lines 21-22 also makes explicit reference to the baby, thereby presenting a conclusion that is designed to highlight health of the central character at the end of the sequence.

Figure 3, however, begins at line 2 by describing an event, namely a decrease in oxygen saturation, occurring over an extended period of time which commences towards the beginning of shift. Thus, this account treats as the first reportable item a description of an event and not of the baby. It is not until line 7 of the summary that we see any mention of the baby himself. This relatively late introduction of the baby into the summary fails to foreground the baby himself as a central character in relation to the events that are being described. Additionally, the final item on the list makes no reference to the initial topic or a change in baby's respiratory status.

b) *Temporal organization of events*:

The description in Figure 2 begins at the start of the shift and concludes at the end, and the intervening events are temporally marked in a sequential order. The list begins at line 2 with a description located at the start of the period of observation. Subsequent items are designed in terms of their temporal connections to this starting point. The temporal marker 'Within an hour' at line 3 describes the next item on the list in relation to the commencement of observation. The next item at lines 8-14 is temporally indexed to be subsequent in the overall listing of events. Similarly, at lines 15-17, descriptions of the baby's respiratory status are temporally marked in relation to the time of occurrence and the age of the baby. Finally, at lines 21-22, concluding descriptions temporally mark events as occurring at the end of the shift by the use of 'now' (line 20).

Shift 23 written by Human Nurse

1 EVENTS THIS SHIFT
2 On admission he was in air.
3 Within an hour his respiratory rate was 63
4 with moderate recession, nasal flaring,
5 occasional grunting and SpO2 falling to
6 the low 80s. He was placed prone and put
7 into 24% incubator oxygen.
8 At 5 hours of age he was in 45% incubator
9 oxygen, and was electively intubated
10 (morphine and sux were given) and put
11 onto CMV ventilation: rate 50, pressures
12 19 / 5, iTime 0.3 in 30% oxygen, tidal
13 volumes were 5ml. ETT is size 3, and 8cm
14 at the lips.
15 At 16:20 (6 hours of age) surfactant was
16 given, 240 mg, first dose, and he was in air
17 within an hour after that.
18 Ventilation has been weaned with CBGs to
19 the present settings.
20 Recession is now just mild. Breathing has
21 settled and he is taking spontaneous
22 breaths.

Figure 2: Nurse written corpus text.

Shift 23 generated by BT-Nurse

1 Events During the Shift
2 Between 11:30 and 14:30, SaO2 decreased
3 from 93 % to 84 %.
4 A CBG was taken at 12:15. Parameters
5 were acceptable. pH was 7.37. CO2 was
6 7.02 kPa. Be was -1.9 mmol/L.
7 The baby was intubated at around 15:15
8 and was moved from Inc O2 to CMV. Vent
9 RR was 50 breaths per minute. Pressures
10 were 19/5 cms H2O. He was in air. Tidal
11 volume was 8.9.
12 At around 15:15, he was given morphine.
13 At about 15:15, he was given
14 suxamethonium.
15 At around 18:30, he was given a first dose
16 of 240 mg of surfactant.

Figure 3: BT-Nurse generated text.

Such temporal organization in Figure 3 however is limited. The initial description does make explicit reference to specific times and so marks the starting point for a temporally organised summary. As the listing of events continues, at a number of points specific events are also temporally marked in order to indicate their relationship to the chronological starting point of the description, ending at lines 15-16 with a description of drug administration presented as occurring towards the end of the period of observation. This sequence, however, is not organised entirely chronologically, in that the temporal reference at line 4 to '12.15' precedes the second such reference at line 2 which is to '14.30'. To the extent that the description provided is framed by reference to times near the start and end of the observational period it is presented in the form of a temporal sequence.

c) *Causal connectivity*:

Descriptions of events in Figure 2 highlight causal connectedness of preceding and subsequent events and actions. For instance, the description at lines 8-14 takes up as relevant the topic introduced at the conclusion of the preceding item, that of 'incubator oxygen'. This topic flow causally connects events described to that topic by detailing steps taken to support the breathing of the baby at that time. In addition, events found within this description are explicitly linked through the use of grammatical markers and the conjunctive 'and'. The parenthetical 'morphine and sux' at line 10 can be read as relevant to the immediately preceding description of intubation, making explicit for the reader the connection between these events. Following this, at lines 15-17 the description makes an explicit connection between two events, namely the medication given and the subsequent status of the baby. Further, the description of the baby as being 'in air' can be heard as a desirable state of affairs, in contrast to previous descriptions. This positive description provides a context for description of ventilation being 'weaned', which also suggests an improvement as a result of the actions taken. Finally, at lines 21-20, the summary concludes with a description that takes as its explicit topic 'breathing' and describes actions of the baby at this time. The reference here to 'taking spontaneous breaths' can be heard as desirable, and in so doing to be a continuation of the baby's breathing status set out previously. As such, the description draws together disparate elements – the baby as the actor in the events being described, his respiratory status, and the temporal context – in offering a hearably positive upshot to the sequence of events that occurred during the shift. Together, the continuation of topic

and linking of events presents the events being described as connected and as located within an ongoing narrative relating to the breathing of the baby over the course of the shift.

With respect to causal connectivity, in Figure 3 there is seen to be variation in how events are causally linked. First, some events are explicitly linked: at lines 7-8, the process of intubation is clearly marked as linked to the baby being moved from incubator oxygen to 'CMV' (CMV is a form of mechanical ventilation that follows from being intubated). Second, some are not marked in this way but can be read as being connected through the consecutive descriptions of particular actions and states: at lines 4-6, we see a description of a blood gas measurement being taken, an evaluation of parameters, and descriptions of particular measurements that allow them to be treated as consequentially relevant and the later descriptions to be treated as presenting the outcomes of the procedure. Third, the form of description works to suggest that there is no immediate connection between different events being described: at lines, 7-14, we are given a description of a process of intubation, of the baby being given morphine, and of the baby being given suxamethonium. Explicitly describing these events as occurring at a similar approximate time suggests that these are not related events occurring in a connected manner but rather are discrete events that simply happen to have occurred at the same time in the shift. This combination of descriptions that are explicitly linked, those that can be read as linked and those that are presented in an unconnected manner fails to provide a coherent ongoing causal narrative for the period of observation.

4.4 Discussion

Taken together, these pragmatic features function to present descriptions in Figure 2 in a recognisably narrative form. Figure 3, however differs from Figure 2 in the following ways. The selection of reportable events, particularly the first and last items in the summary, differs markedly from those in Figure 2. The first reported item provides little, if any, context for the descriptions to follow and makes no reference to the baby as the focus of the summary. Further, the causal organization of the events being described is variable, making some connections explicit, other connections inferable, and failing to make relevant causality in instances where it might be appropriate. In these respects, the text produced by the NLG system does not have the narrative form seen in the nurse-written corpus text in Figure 2. However, temporal organization of events and inclusion of some causal elements provide a more coherent organization of descriptions and thus make available at least some causal connections between events. To this extent, the NLG system appears to have produced text that more closely resembles that produced in nurse written corpus text. These findings show that discourse analysis represents a useful tool for evaluation of NLG systems. The analyses identified a range of pragmatic features which are desirable features in a text which seeks to describe in an efficient and useable manner the sequence of events and occurrences which can arise in nursing shifts in an NICU.

These findings have implications for the design of NLG systems. First, in terms of content selection, corpus texts show that the nurse does not merely select items as being topically relevant, but *treats* these items as topically relevant in terms of how descriptions of actions and events are designed and of how these descriptions are sequentially organized. In this respect, topical relevance must be viewed not as an objective feature of the situations being described, but rather as a pragmatic outcome of texts themselves. Second, it is apparently important to carefully select those items that are reported at the very start and the very end of the text. The first and last entries function to introduce the topic of the summary and offer an upshot of the matters at the end, that is these items take up *functional* 'slots'. Third, the human nurse expert attends to the topic flow: the sequential organization of a text to provide for readily recognizable shifts from one topic to another; this is absent from the text produced by the BT-Nurse system.

These issues come together in the issue of narrative structure. Narrative can be viewed as a form of talk or text in which descriptions of events are sequentially ordered so as to tell a story about those events. The human nurse's text contains pragmatic features such as identifying the baby as an actor in events, and indicating causal relationships among the actions and events being described, which features make it likely for it being treated as having a narrative form (Daiute et al., 2003).

5 General Discussion

5.1 Findings on BT-Nurse:

In terms of types of content, BT-Nurse texts have more instances of trend detection and recordings of parameter values, and fewer instances of inference enabling data representations than the corpus texts. This is perhaps a natural consequence of the differences in capabilities between a computer (good at crunching numbers) and a person (good at making domain inferences). It probably makes sense to accept this distinction and try to determine how a computer-generated text can most usefully support a nurse: an improved analysis of numeric data.

The evaluation using discourse analysis showed that BT-Nurse texts are deficient from a narrative perspective. They show a minimal foregrounding of the baby as a central character, inconsistent temporal organization of events and variable causal connectivity. Narrative form is a desirable feature of texts from an understandability and utility perspective (Reiter et al., 2008) more so because narratives are a pervasive feature of human interaction (Jefferson, 1978; Sacks, 1992).

5.2 Implications for NLG systems:

A content analysis of corpus texts reveals various ways in which domain experts represent various domain relevant types of information. For instance, here we see various ways in which both temporal markers and events are presented in corpus texts which can inform ways in which inferential items can potentially be included in NLG system generated texts. Knowledge of this sort then is certainly useful in designing NLG systems to produce texts which present information in appropriate ways for the domain.

Discourse analysis differs from content analysis in providing an understanding of ways in which users engaged in their daily duties present summaries or similar texts as part of their duties and helps in producing texts that take up such concerns. Here, aspects of presenting the baby as a central character was one feature of producing corpus texts. This is readily seen to be relevant for activities performed by nurses in that their duties are about caring and/or providing nursing care for one particular party, namely 'the baby'. To see that human users take up aspects such as these to be relevant features is knowledge useful in the design of NLG systems that are to be deployed in specific domains. Another finding of relevance is the role of items that occupy the start and final positions in a text. The inclusion of specific items at certain points in a text by human users allows them to do specific functions: doing an introduction, offering an upshot and others. Of note is that such features *serve* to make the text more of a narrative.

The interesting thing about the above findings is that they did *not* arise from the quantitative evaluation of BT-Nurse. To us, this suggests that such findings are more likely to arise from a qualitative evaluation conducted by analysts with expertise in discourse analysis or content analysis; they are not likely to be spontaneously suggested by subjects who have domain expertise but no expertise in analysis of texts.

5.3 Limitations:

Although, the extent of texts covered in these analyses is limited, outcomes of such evaluations are useful and a complete analysis is likely to throw up further useful knowledge. For instance, across the corpus texts foregrounding the baby as a central character and how descriptions offered are made in ways to make overall evaluations of the baby's status, such as being 'okay' or 'deteriorating' are seen to be consistent features.

Additionally matters that appear to be of a quantitative nature were revealed as relevant aspects of these texts only posterior to qualitative analyses. For example, the content analysis showed a difference in the frequency of trend descriptions of parameter values between corpus texts and BT-Nurse texts. This could probably be tested using quantitative techniques; this would require annotating the texts, and the annotation scheme could be based on the scheme used in content analysis. In theory a task evaluation study could even be performed to evaluate the impact of having more trend descriptions, although this would be an expensive undertaking.

6 Conclusion

The qualitative evaluations presented above make use of two different but complementary methodologies. Content analysis provides us with knowledge on the sorts of items present in a text. Discourse analysis on the other hand moves a step further and makes clear aspects of ways in which these items are presented in the service of certain

actions (making the baby a central character, for instance). In particular, content analysis is appropriate in showing what goes into a text and discourse analysis reveals what the texts are designed to do.

Qualitative analyses described above identified many differences between generated texts and corpus texts. Some of the differences identified may be desirable, such as the fact that BT-Nurse texts contain more trend descriptions than corpus texts. Other differences are probably not desirable, such as narrative deficiencies in the generated texts. However, the key point is that qualitative analyses have identified these differences, so that developers are aware of them and can decide what action to take.

Acknowledgements

This research was funded by the UK Engineering and Physical Sciences Research Council under grants EP/D049520/1, EP/D05057X/1, and EP/E011764/1. The authors would like to thank other members of the BABYTALK project including François Portet, Jim Hunter, Somayajulu Sripada, and Neil McIntosh for their help.

References

K Adlassnig, C Combi, A Das, E Keravnou, and G Pozzi. 2006. Temporal representation and reasoning in medicine: Research directions and challenges. *Artificial Intelligence in Medicine, 38*(*2):* 101 - 113.

C Daiute and C Lightfoot. 2003. *Narrative analysis: Studying the development of individuals in society*. London: Sage.

A Gatt, F Portet, E Reiter, J Hunter, S Mahamood, W Moncur, and S Sripada. 2009. From data to text in Neonatal Intensive Care Unit: Using NLG technology for decision support and information management. *AI Communications, 22:* 153 - 186.

E Hovy 1993. Automated Discourse Generation Using Discourse Structure Relations. *Artificial Intelligence, 63*(*1-2):* 341 - 386.

J Hunter, Y Freer, A Gatt, E Reiter, S Sripada, C Sykes, and D Westwater. 2011. BT-Nurse: Computer generation of natural language shift summaries from complex heterogenous medical data., *Journal of American Medical Informatics Association* **18**:621-624.

G Jefferson 1978. Sequential aspects of storytelling in conversation. In, J. Schenkein (Ed), *Studies in the organization of conversational interaction*. London: Academic Press.

A McKinlay, C McVittie, E Reiter, Y Freer, C Sykes, and R Logie. 2010. Design issues for socially intelligent user-interfaces: A Discourse analysis of a data-to-text system for summarizing clinical data. *Methods of Information in Medicine, 49*(*4):* 379 - 387.

A McKinlay and C McVittie. 2008. *Social Psychology & Discourse.* Sussex: Wiley-Blackwell.

K Papineni, S Roukos, T Ward, and W Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *Proceedings of ACL-2002,* pages 311-318.

F Portet, E Reiter, A Gatt, S Sripada, Y Freer, and C Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence, 173*(*7,8):* 789 - 816.

E Reiter and A Belz. 2009. An investigation into the validity of some metrics for automatically evaluating Natural Language Generating systems. *Computational Linguistics, 35:* 529 - 558.

E Reiter, A Gatt, F Portet, and M van der Meulen. 2008. The importance of narrative and other lessons from an evaluation of an NLG system that summarizes clinical data. *Proceedings of INLG-08,* pages 147-155

L Richards. 2009. *Handling Qualitative Data: A Practical Guide.* London: SAGE.

H Sacks. 1992. *Lectures on Conversation.* Oxford: Blackwell's.

J Saldaña. 2009. *The Coding Manual for Qualitative Researchers.* London: SAGE.

S Thompson and W Mann. 1987. Rhetorical Structure Theory: A framework for the Analysis of Texts. *IPRA Papers in Pragmatics, 1:* 79 - 105.