Natural Language Generation and Data-To-Text

Albert Gatt

Institute of Linguistics, University of Malta

Tilburg center for Cognition and Communication (TiCC)

Department of Computing Science, University of Aberdeen

Document Planning

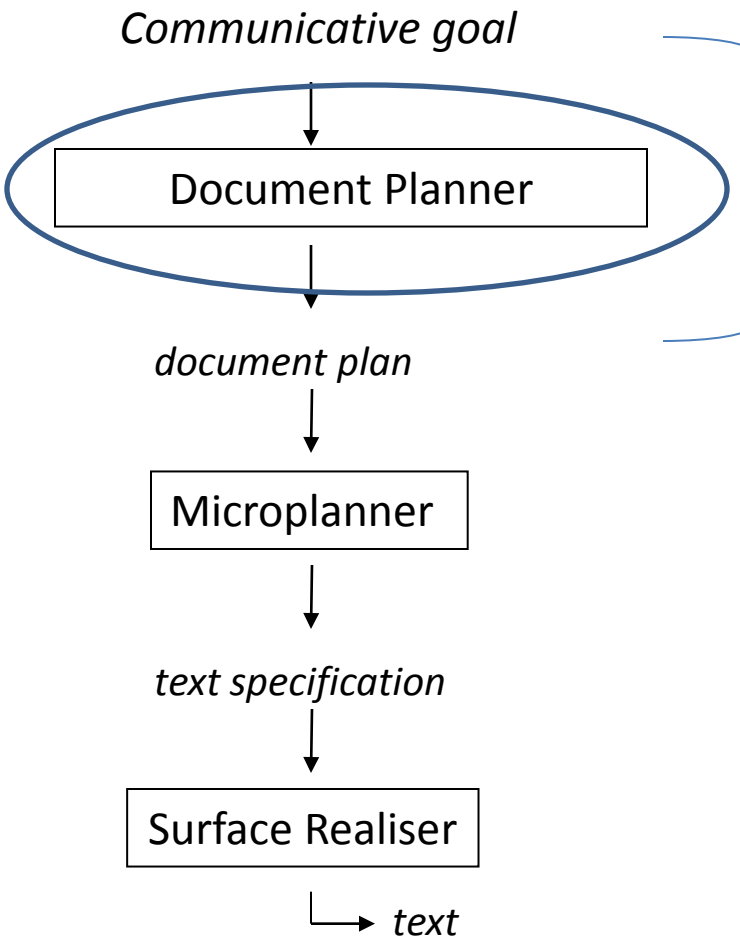
Content determination

- Given the input data, what should be included in a document?

Document structuring

- How should the selected information be structured in the document?

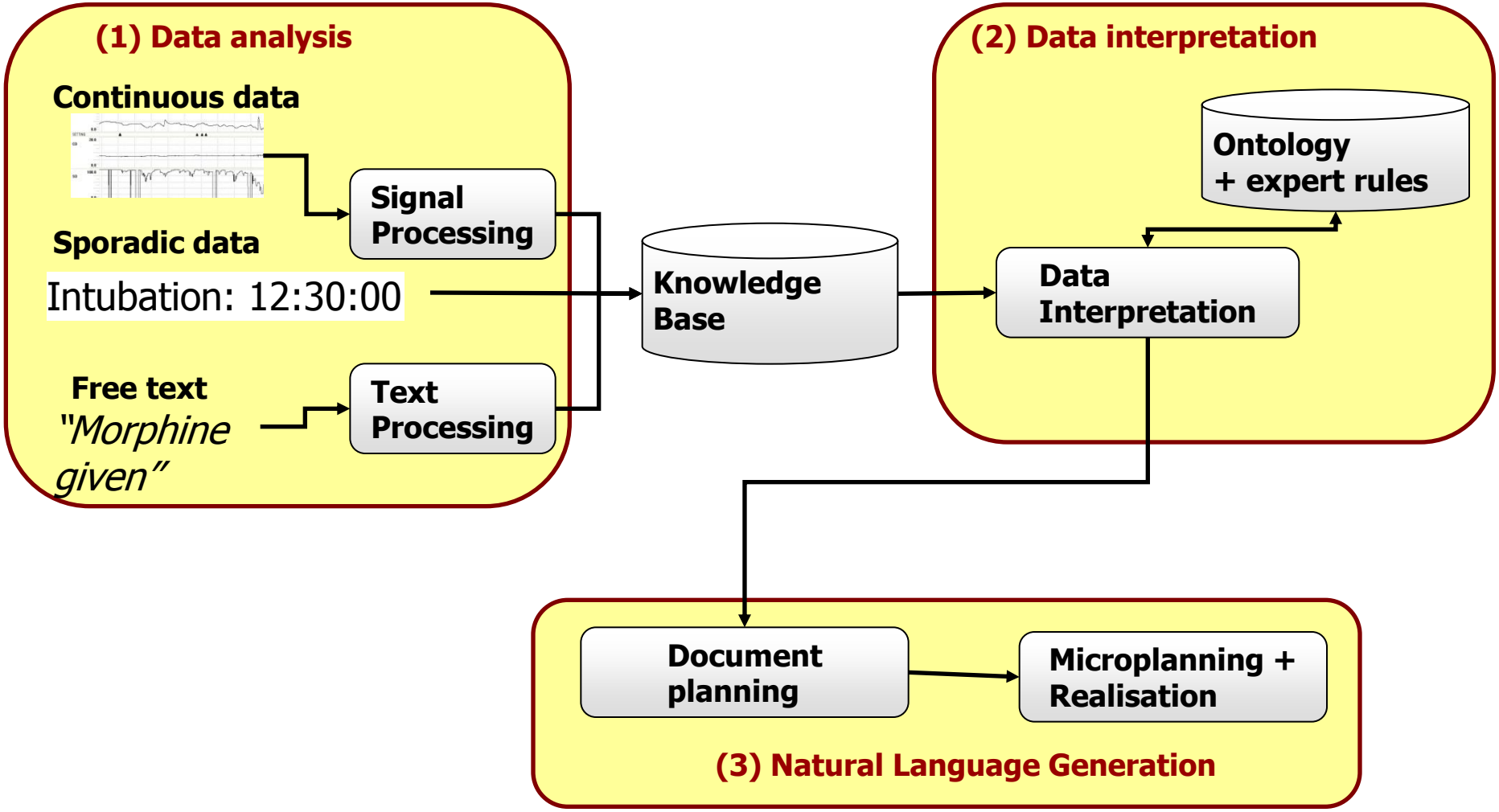
The “consensus” architecture



Tasks:

- Selecting the messages to be included
- Rhetorical structuring
- Ordering
- Segmentation

BabyTalk architecture



BabyTalk: Before document planning

- Data analysis has to extract the information units from the signals, the sporadic data and the free text data.

- Data interpretation has to:
  - Abstract elements into higher-order units
  - Link them using rules to establish:
    - Causality
    - Sequences
    - etc

Document planning

Two "traditional" approaches:

- Methods based on observations about common text structures, as reflected in a corpus.

  → How are texts in this particular domain usually written?

- Methods based on reasoning about discourse coherence and the communicative goals of the text.

  → Given content X and content Y, what is the best way to combine them in the text?

Part 1

CONTENT DETERMINATION

Content Determination

We will assume that we're dealing with **MESSAGES** or **EVENTS**:

- the units of information in the text
- each message represents some portion of input data

What types of messages do we have?

- Based on a corpus of texts in a particular domain, we can identify what the core types of message are.
- To a significant extent, content determination is domain-dependent.

Example #1: Weather report app

- We want to build a system that generates a short weather report based on numerical weather data.

Method

1. Obtain a corpus of texts in this domain.
2. Analyse it to identify the information units (messages):
   a. Messages which are always present
   b. Messages which are optional

Acknowledgement: The weather report example is drawn from E. Reiter and R. Dale (1999). Natural Language Generation. Tutorial at EACL'99

Example #1: Weather report app

- Routine (obligatory) messages
  - MonthlyRainFallMsg,
    MonthlyTemperatureMsg,
    RainSoFarMsg,
    MonthlyRainyDaysMsg

- Always constructed for any summary to be generated

Example #1: Weather report app

A MonthlyRainfallMsg:

```
((message-id msg091)
  (message-type monthlyrainfall)
  (period ((month 04)
           (year 1996)))
  (absolute-or-relative relative-to-average)
  (relative-difference ((magnitude ((unit millimeters)
                                    (number 4)))
                        (direction +))))
```

Example #1: Weather report app

- Significant Event messages
  - RainEventMsg,
    RainSpellMsg,
    TemperatureEventMsg,
    TemperatureSpellMsg

- Only constructed if the data warrants their construction: eg if rain occurs on more than a specified number of days in a row

Example #1: Weather report app

A RainSpellMsg

```
((message-id msg096)
 (message-type rainspellmsg)
 (period ((begin ((day 04)
                  (month 02)
                  (year 1995)))
          (end ((day 11)
                (month 02)
                (year 1995)))
          (duration ((unit day)
                     (number 8)))))
 (amount ((unit millimetres)
          (number 120))))
```

Example #2: BabyTalk

BT-Nurse system

- Very large dataset, from a 12 hour shift, related to a specific patient.
- Much more complex and knowledge-intensive!

Method

- Use a corpus of texts written by a senior nurse to identify the elements of the ontology that are mentioned in the texts.
- Conduct interviews with experts to:
  - Build a knowledge base (ontology) to represent the information elements in the domain.
  - Rate the importance of the various information elements.

The structure of nurse summaries

Division by body system:

- Respiratory, Circulation/cardiovascular, ...
- BT-Nurse main focus on these two (there are others)
- Each section has 3 sub-sections:
  - Current status
  - Events during shift
  - Potential problems

Example human-written summary

Respiration

Current assessment

Respiratory effort reasonably good, his total resp rate being 40–50 breaths/minute while the ventilator rate is 20. […]

Events during the shift

[…] After blood gas at 23:00 ventilation pressure reduced to 14/4. CO2 was 4.1 and tidal volumes were 3.8–4 ml at that time. After a desaturation 3 hours later down to 65% pressures were put back to 16/4. He has had an oxygen requirement of 26% since this episode.

Potential problems

Small ETT could become blocked or dislodged – ongoing assessment of need for suction; ensure ETT is secure.

Corpus summary

- Shift summary written by a senior neonatal nurse;

Main properties

- Each subsection focuses on certain types of events or states.
- Also mentions other things which are clinically important.

Requirements for NLG

- Need a KB that contains the events, states, entities in the domain.
- Need a way to estimate importance.

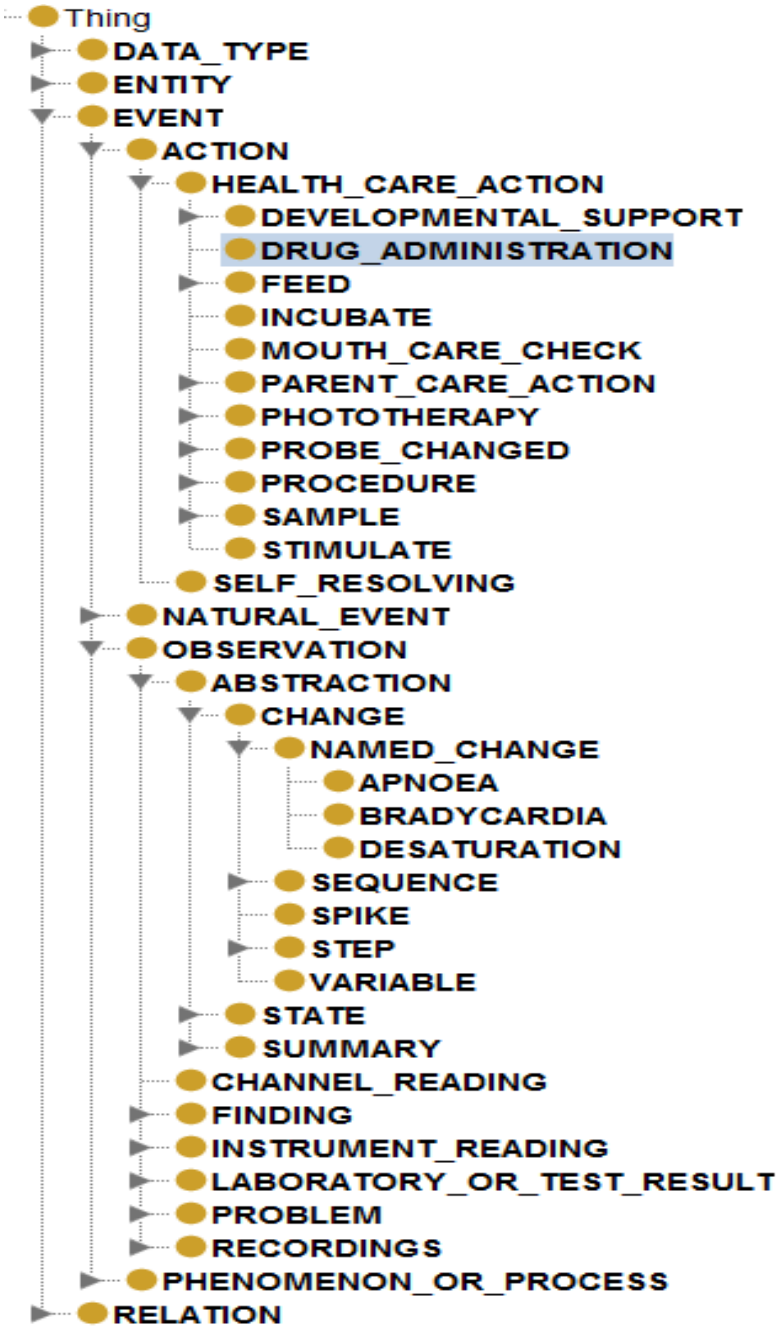
Example #2: BabyTalk

Knowledge is a crucial component in this system.

Large ontology (OWL) of concepts, representing types of domain entities, events and relations.

Raw input data is mapped to ontology instances, during **signal analysis** and **data interpretation**.

Ontology instances are our **messages**.



Example #2: BabyTalk

SEQUENCE (BRADYCARDIA)

BRADYCARDIA (16:58:46) Imp: 31.64
BRADYCARDIA (17:01:15) Imp: 79.80
BRADYCARDIA (17:03:57) Imp: 80.21
BRADYCARDIA (17:04:30) Imp: 39.97
BRADYCARDIA (17:05:01) Imp: 34.60
BRADYCARDIA (17:06:03) Imp: 66.24

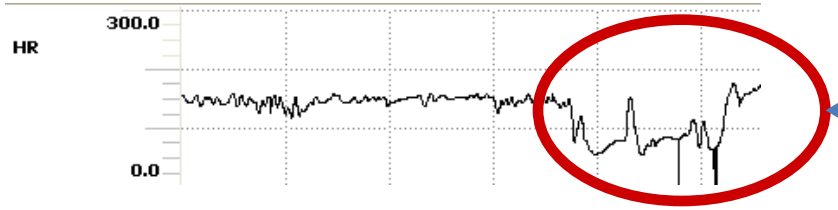
(1) Signal Analysis (pre-NLG)

- Identify interesting patterns in the data.
- Remove noise.

(2) Data interpretation (pre-NLG)

- Estimate the importance of events
- Perform linking & abstraction

Abstraction during data analysis and interpretation



Signal analysis identifies the patterns as bradycardias.

SEQUENCE (BRADYCARDIA)

BRADYCARDIA (16:58:46) Imp: 31.64
BRADYCARDIA (17:01:15) Imp: 79.80
BRADYCARDIA (17:03:57) Imp: 80.21
BRADYCARDIA (17:04:30) Imp: 39.97
BRADYCARDIA (17:05:01) Imp: 34.60
BRADYCARDIA (17:06:03) Imp: 66.24

Data interpretation concludes that the bradycardias are related: they are a **sequence**.

Also uses rules to assign **importance values** to these events.

- Thing
  - DATA_TYPE
  - ENTITY
  - EVENT
    - ACTION
      - HEALTH_CARE_ACTION
        - DEVELOPMENTAL_SUPPORT
        - DRUG_ADMINISTRATION
        - FEED
        - INCUBATE
        - MOUTH_CARE_CHECK
        - PARENT_CARE_ACTION
        - PHOTOTHERAPY
        - PROBE_CHANGED
        - PROCEDURE
        - SAMPLE
        - STIMULATE
      - SELF_RESOLVING
    - NATURAL_EVENT
    - OBSERVATION
      - ABSTRACTION
        - CHANGE
          - NAMED_CHANGE
            - APNOEA
            - BRADYCARDIA
            - [illegible]ON
          - SEQUENCE
          - SPIKE
          - STEP
          - VARIABLE
        - STATE
        - SUMMARY
      - CHANNEL_READING
      - FINDING
      - INSTRUMENT_READING
      - LABORATORY_OR_TEST_RESULT
      - PROBLEM
      - RECORDINGS
    - PHENOMENON_OR_PROCESS
  - RELATION

Linking during data analysis and interpretation

Data analysis

Data analysis has found an **intubation** event and a **decreasing trend** in heart rate.

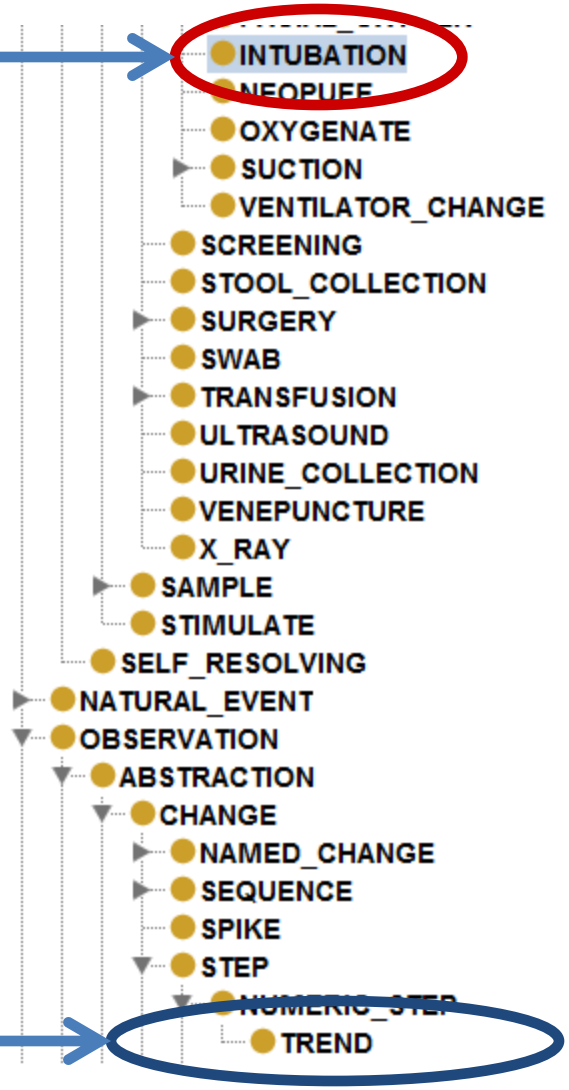
INTUBATION (12:30) Imp: 98

TREND (12:35) Imp: 100

Data interpretation

We know that intubation can affect heart rate. One of our data interpretation rules says:

IF E1 is an INTUBATION and
E2 is a TREND and
E2 channel is HEART RATE and
E2 is DECREASING and
E1 and E2 occur within *t* seconds of each other
THEN:
link E1 and E2 via CAUSE



Example #2: BabyTalk

- In summary, BT-Nurse relies on:
  - an ontology, which models the domain
  - procedures to identify trends and patterns in numerical data
  - rules to reason about the data and identify:
    - its **importance**
    - the **relationships between the information units** (e.g. SEQUENCE, CAUSE, etc).

Challenges

- There is too much data! Not all of it can be selected.
  - Can we use importance?
- The information has to be structured in the document.
  - Can we use the relations between events to help us?

Example #2: BabyTalk

- Once all the data has been processed, the content determination algorithm tries to:
  - Select the types of events which always need to be mentioned (based on corpus analysis).
  - Find other events to mention, based on importance.
- The Document Planner actually proceeds by planning each section of the document separately.
- Before we look into the details, we need to consider the problem of document structure.

Part 2

DOCUMENT STRUCTURING VIA SCHEMAS

Document Structuring via Schemas

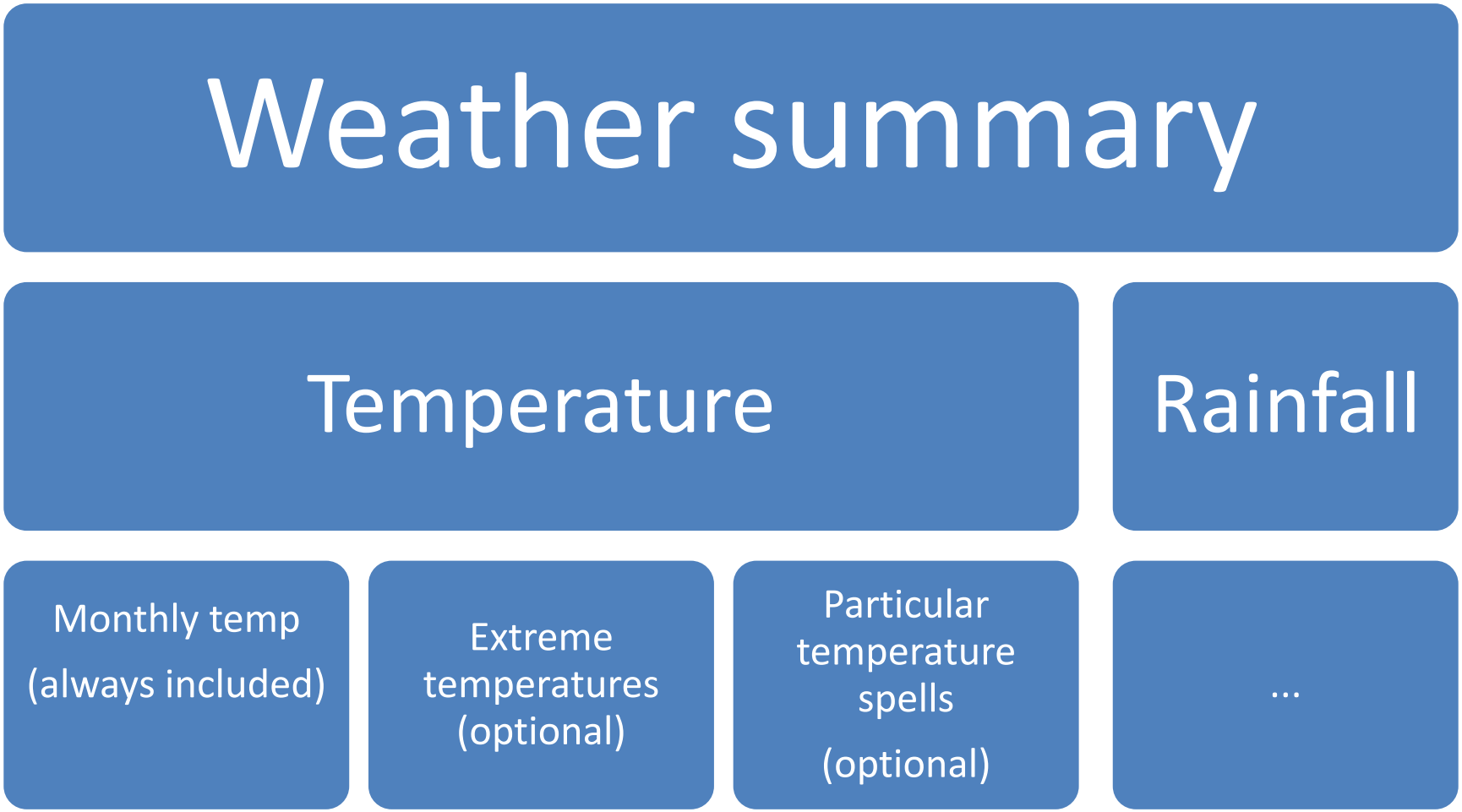
Basic idea (after McKeown 1985):

- Texts often follow conventionalised patterns.

- These patterns can be captured by means of ‘text grammars’ that:
  - Specify where certain content goes
  - Ensure coherent structure

- Can specify many degrees of variability and optionality.

Example #1: weather report



Document Structuring via Schemas

Implementing schemas:

- Simple schemas can be expressed as grammars.

- More flexible schemas can also be implemented as macros or class libraries on top of a conventional programming language, where each schema is a procedure.

Deriving Schemas From a Corpus

Using a corpus:

- Take a small number of similar corpus texts.
- Identify the messages, and try to determine how each message can be computed from the input data.
- Propose rules or structures which explain why message *x* is in text A but not in text B — this may be easier if messages are organised into a taxonomy.
- Discuss this analysis with domain experts, and iterate.

Example #1: weather document schema

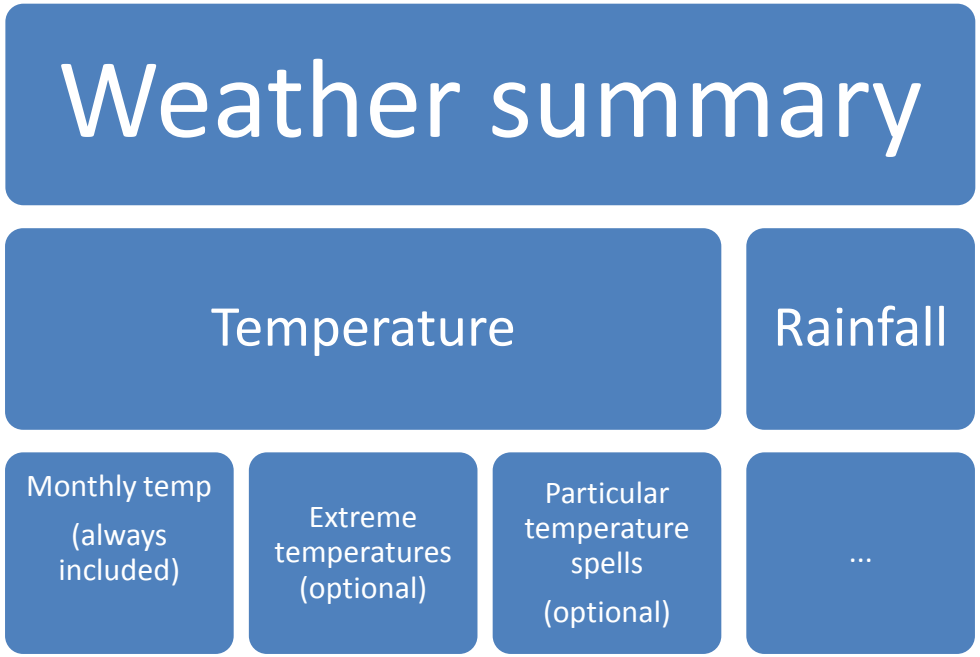
A Simple Schema in grammar notation

```
WeatherSummary →
    TemperatureInformation
    RainfallInformation

TemperatureInformation →
    MonthlyTempMsg
    [ExtremeTempInfo]
    [TempSpellsInfo]

...
```

The schema as a tree



Schemas: Pros and Cons

Advantages of schemas

- Computationally efficient
- Allow arbitrary computation when necessary
- Naturally support genre conventions
- Relatively easy to acquire from a corpus

Disadvantages

- Limited flexibility: require predetermination of possible structures
- Limited portability: likely to be domain-specific

Document Structuring via Explicit Reasoning

- Texts are coherent by virtue of relationships that hold between their parts — relationships like narrative sequence, elaboration, justification ...

Resulting Approach:

- Segment knowledge of what makes a text coherent into separate rules.
- Use these rules to dynamically compose texts from constituent elements by reasoning about the role of these elements in the overall text.

Document Structuring via Explicit Reasoning

- Typically adopt AI planning techniques:
  - Goal = desired communicative effect
  - Plan constituents = messages or structures that combine messages (subplans)
- Can involve explicit reasoning about the user's beliefs
- Often based on ideas from **Rhetorical Structure Theory**

Part 3

OVERVIEW OF RHETORICAL STRUCTURE THEORY

Rhetorical Structure Theory

What makes a text coherent?

- What kinds of coherence relations exist?
  - *Cause, elaboration, justification...*
- How can these relations be recognised?
  - E.g. Are they associated with particular kinds of phrases?
- Which text segments do these relations apply to?
  - *Sentences, clauses, ...*
- How does the theory view the structure of the text?
  - *As a sequence of units, as a tree, as a graph...*

Rhetorical Structure Theory

- RST (Mann and Thompson 1988) is a theory of **text structure**
  - **not** concerned with the topic of a text **but**
  - **how** bits of the underlying content of a text are structured so as to hang together in a **coherent** way.

- The main claim of RST:
  - Parts of a text are related to each other in predetermined ways.
  - There is a finite set of such relations.
  - Relations hold between two spans of text
    - **Nucleus**
    - **Satellite**

A small example

John was sacked this morning.
He was caught stealing from the cash register.

- The first sentence seems to be the “main” one here.
  → we would call this the **nucleus.**

- The second sentence is “subordinate”; it explains what **CAUSE**d the event described by the first sentence.
  → we could call this the **satellite**

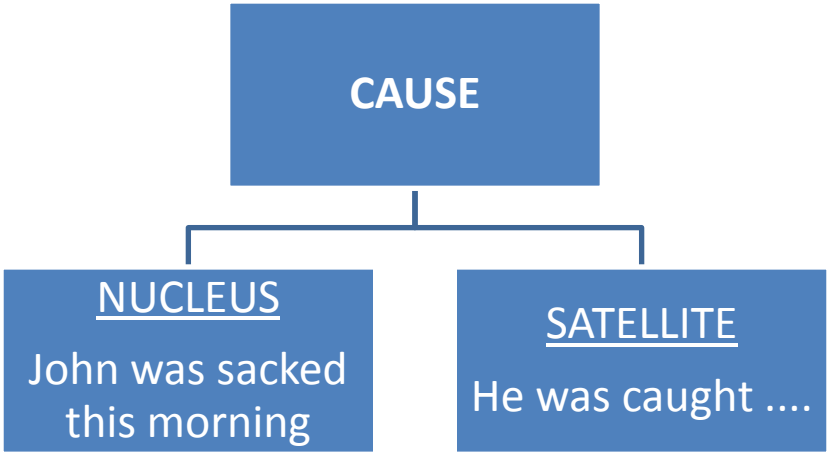
A small example

This text can be viewed as a tree.

John was sacked this morning.

He was caught stealing from the cash register.

- Note that the relation in this example is **implicit**

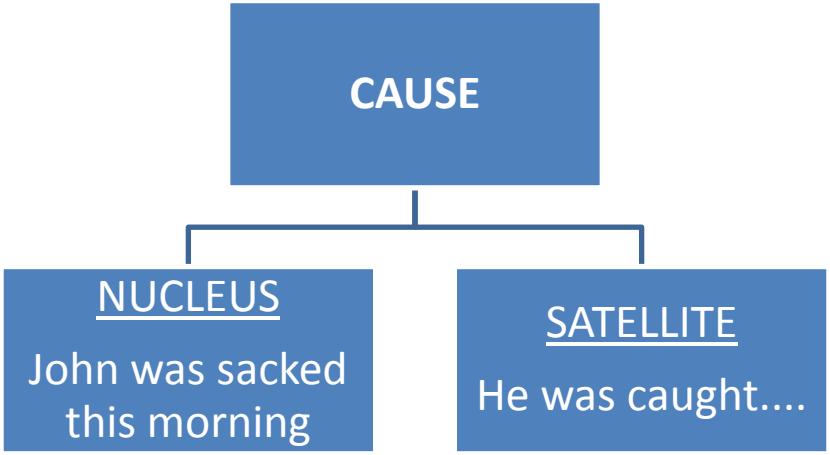


A small example

The relation is independent of the realisation.

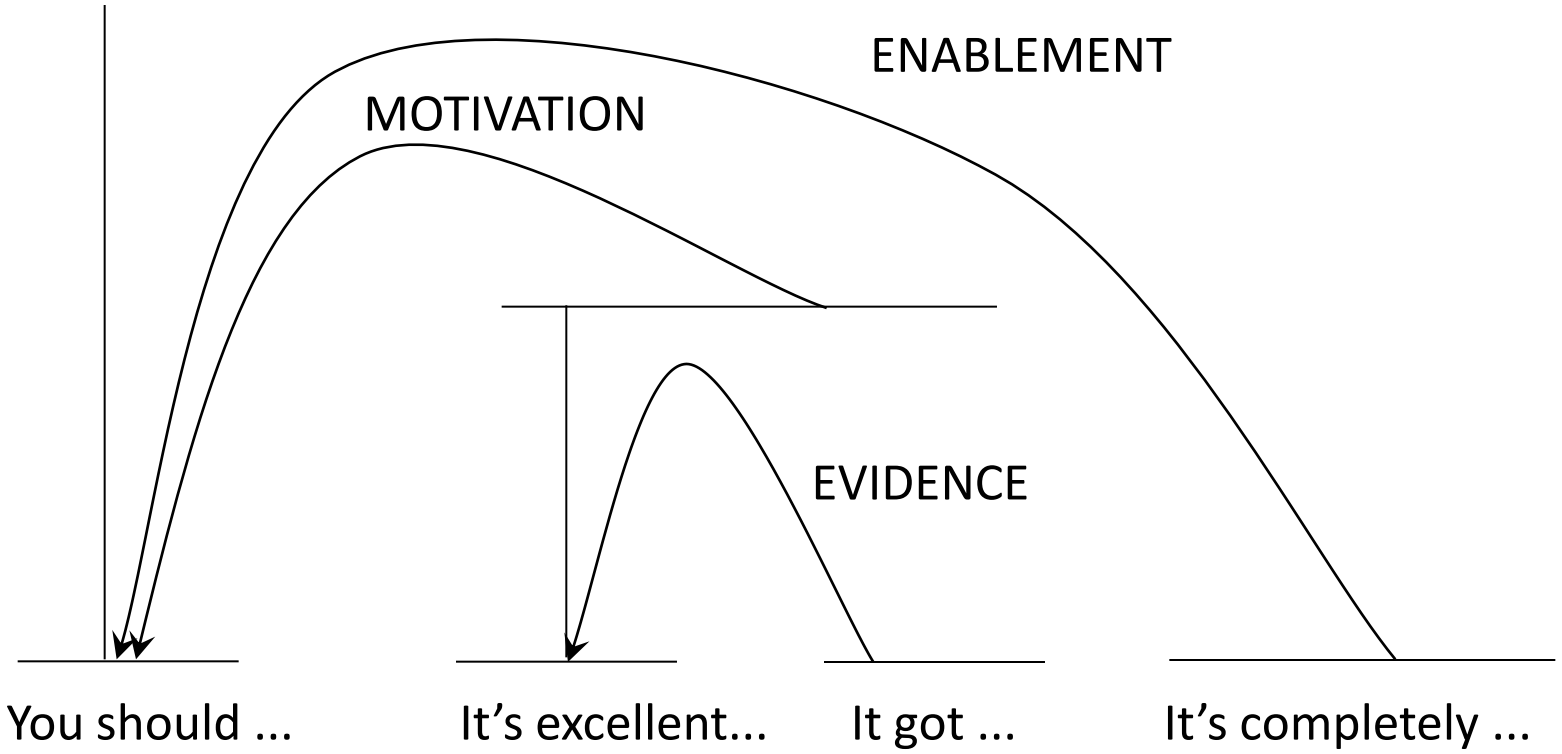
John was caught stealing from the cash register, so he was sacked this morning.

- In this case, the relation is more explicit. We have used *so*...



A slightly bigger example

You should visit the new exhibition. It's excellent. It got very good reviews. It's completely free.



An RST relation definition (example)

MOTIVATION

- Nucleus represents an action which the hearer is meant to do at some point in future.
  - *You should go to the exhibition*
- Satellite represents something which is meant to make the hearer want to carry out the nucleus action.
  - *It's excellent. It got a good review.*
  - Note: Satellite need not be a single clause. In our example, the satellite has 2 clauses. They themselves are related to each other by the EVIDENCE relation.
- Effect: to increase the hearer's desire to perform the nucleus action.

RST relations more generally

- An RST relation is defined in terms of the
  - **Nucleus** + constraints on the nucleus
    - Nucleus is the core content of the discourse unit.
    - (e.g. Nucleus of motivation is some action to be performed by H)
  - **Satellite** + constraints on satellite
    - Satellite is additional information, related to the nucleus in a specific manner.
  - A desired effect on the reader/listener, arising as a result of the relation.

Some further RST Examples

With its distant orbit – 50 percent farther from the sun than Earth – and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -70 degrees Fahrenheit at the equator, and can dip to -123 degrees C near the poles.

Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure. Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide.

- **CAUSE**: the nucleus is the result; the satellite is the cause
  - ...**any liquid water would evaporate because of the low atmospheric pressure**
- **ELABORATION**: the satellite gives more information about the nucleus
  - **With its distant orbit [...] and slim atmospheric blanket**, Mars experiences frigid weather conditions.
- **CONCESSION**: satellite expresses possible "exceptions" or apparent counter-examples to the rule expressed by the nucleus
  - **Although the atmosphere holds a small amount of water** [...] most Martian weather involves blowing dust...

Source: I. Mani (2001). *Automatic Summarization.* Amsterdam: John Benjamins

Other relations

- Circumstance
  - Satellite describes the temporal, spatial or situational framework in which the reader should interpret the situation presented in the nucleus.
    - ***As your portable drive spins, your computer is copying data to it.***

- Purpose
  - Satellite describes an effect that the action described by the nucleus is intended to achieve.
    - **You have to move to hit the ball**

- Solutionhood
  - Nucleus is a solution to a problem posed in the satellite.
    - **What should you do if your portable drive stops working? Call customer care...**

Implications for NLG

- Given:
  - Events, messages etc in the input that we wish to express
  - Sufficient knowledge to work out the relationship between these messages
- Then we can:
  - Design rules that identify the relations between events
  - Use the relations to plan the structure of the text.
  - (We can use traditional AI planning techniques for this, but this isn't the only way.)

Part 4

PLANNING DOCUMENTS WITH RST

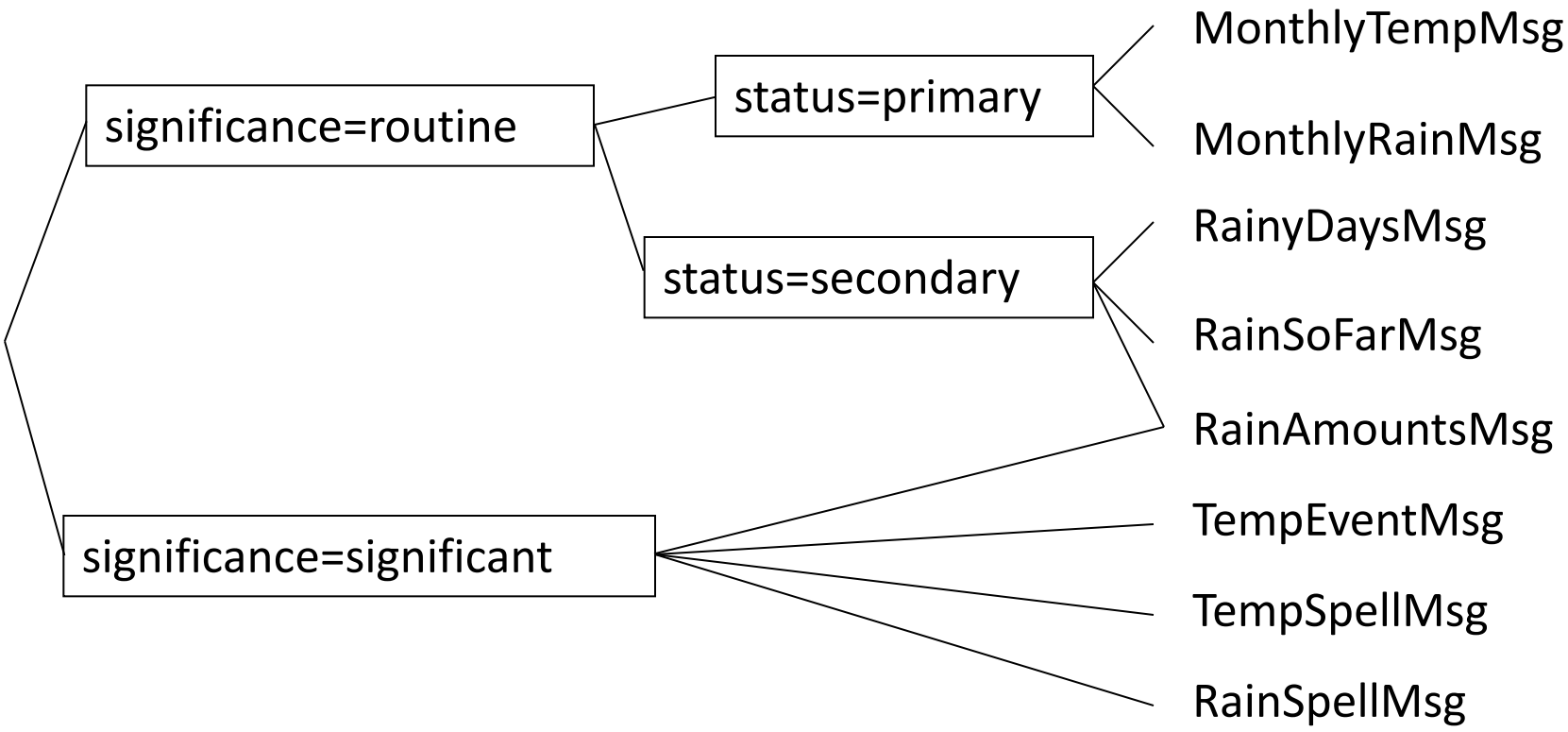
Example #1: Weather report

Three basic rhetorical relationships:

- SEQUENCE
- ELABORATION
- CONTRAST

Applicability of rhetorically-based planning operators determined by attributes of the messages.

Message Attributes



Example #1: weather report

SEQUENCE planning rule

- Two messages can be connected by a SEQUENCE relationship if both have the attribute

  message-status = primary

Example #1: weather report

ELABORATION planning rule

- Two messages can be connected by an ELABORATION relationship if:
  - they are both have the same *message-topic*
  - the nucleus has *message-status = primary*

Example #1: weather report

CONTRAST

- Two messages can be connected by a CONTRAST relationship if:
  - they both have the same message-topic
  - they both have the feature *absolute-or-relative = relative-to-average*
  - they have different values for *relative-difference:direction*

Example #1: algorithm

- Select a start message
- Use rhetorical relation operators to add messages to this structure until all messages are consumed or no more operators apply
- Start message is any message with

  message-significance = routine

Example #1: algorithm

The algorithm:

DocumentPlan = StartMessage

MessageSet = MessageSet - StartMessage

repeat

– find a rhetorical operator that will allow attachment of a message to the DocumentPlan

– attach message and remove from MessageSet

until MessageSet = 0 or no operators apply

Example #1: algorithm

The Message Set:

MonthlyTempMsg ("cooler than average")

MonthlyRainfallMsg ("drier than average")

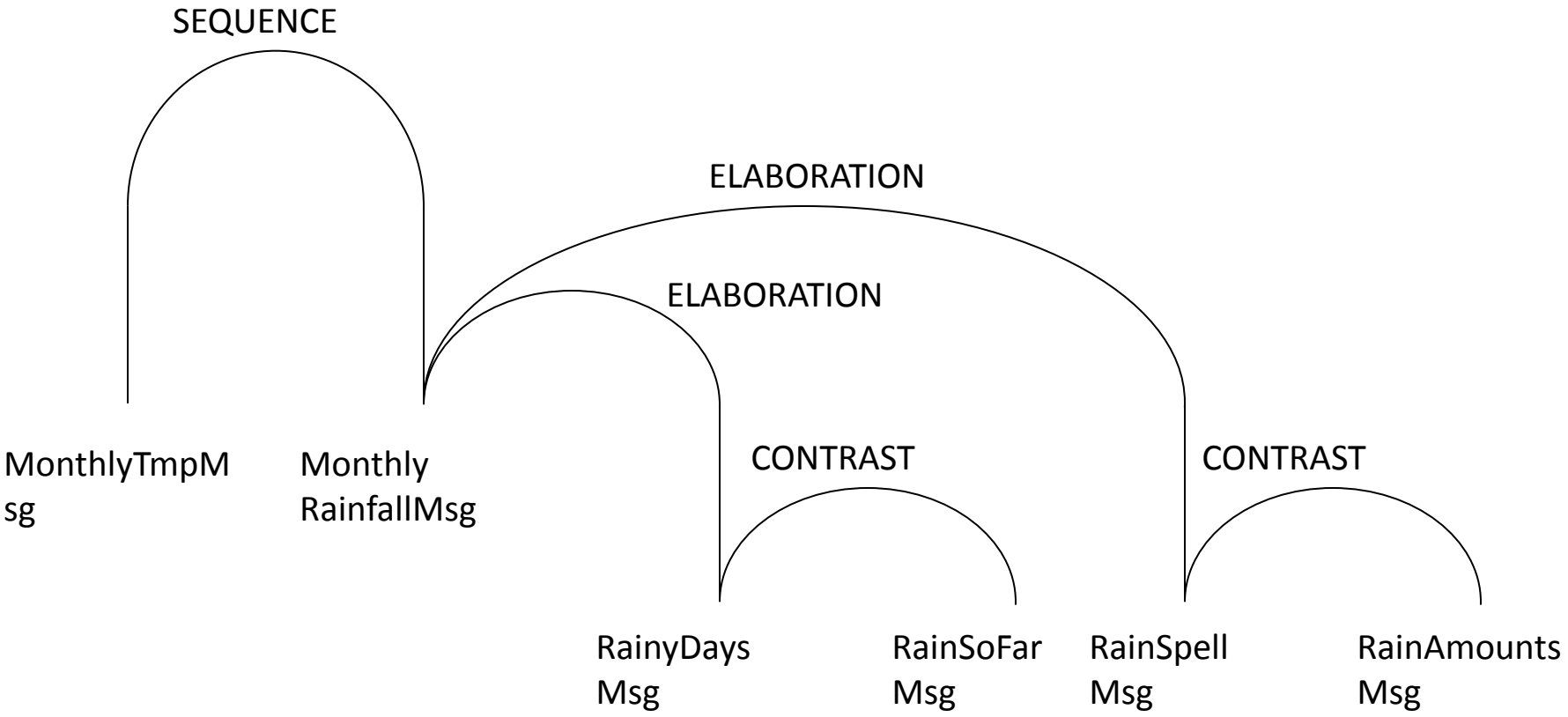
RainyDaysMsg ("average number of rain days")

RainSoFarMsg ("well below average")

RainSpellMsg ("8 days from 11th to 18th")

RainAmountsMsg ("amounts mostly small")

Example #1: Document plan



Part 5

COMBINING REASONING AND SCHEMAS IN BT-NURSE

Hybrid document planning

Schemas

- Some parts of a summary are fixed. They always contain roughly the same types of messages.
- Here, we rely on a schema-based approach.

Reasoning

- Other parts are more dynamic. Their content depends on what happened which is of clinical importance.
- Here, we rely on heuristics that:
  - Select events based on their importance.
  - Look at the relations between them (identified during data interpretation)
  - Try to structure the text coherently prioritising important events.

Example human-written summary

Respiration

Current assessment

Respiratory effort reasonably good, his total resp rate being 40–50 breaths/minute while the ventilator rate is 20. […]

Content here is fixed. Every summary needs to report on things like respiratory rate, ventilator rate, etc.
A schema is ideal here.

Events during the shift

[…] After blood gas at 23:00 ventilation pressure reduced to 14/4. CO2 was 4.1 and tidal volumes were 3.8–4 ml at that time. After a desaturation 3 hours later down to 65% pressures were put back to 16/4. He has had an oxygen requirement of 26% since this episode.

Content here is dynamic as it depends on what actually happened.
Need to rely on data interpretation to estimate event importance, and causal relationships.

Dynamic document planning

Content selection algorithm

Input: the KB, with ontology instances derived from data, and their relationships

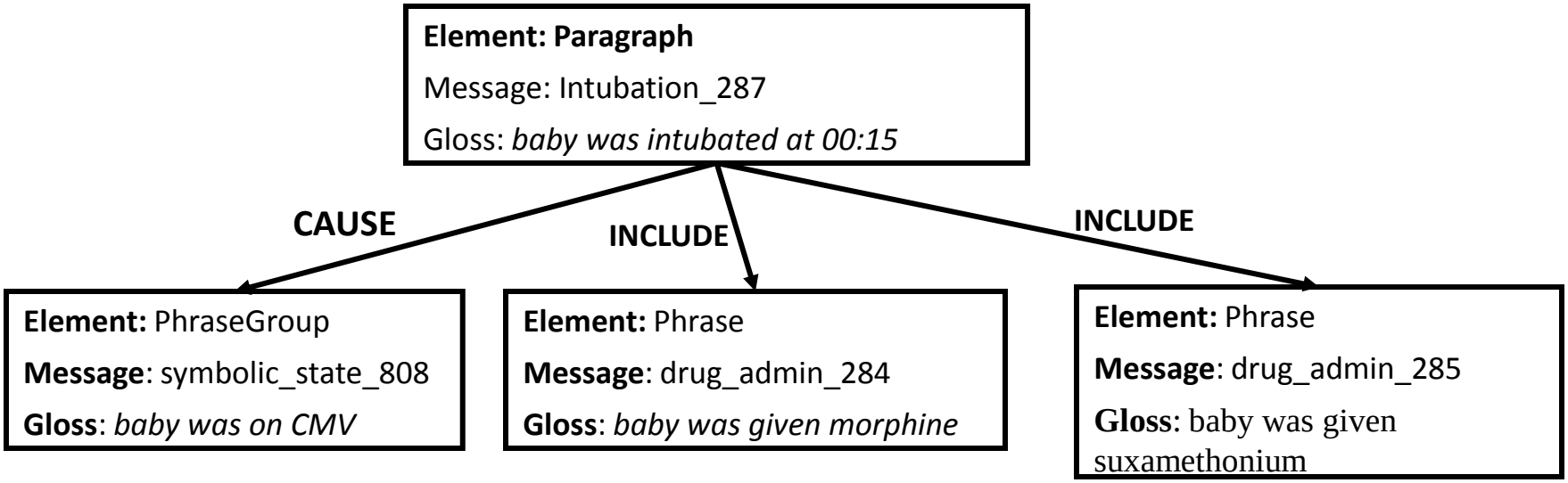
1. Identify a small number of **key events** (events with high importance).
2. Generate a paragraph for each key event.
3. For each key event, select a number of **related events** and populate the relevant paragraph.

Information ordering

- In BT-Nurse, information in the document is ordered in two ways:
  - Body system (respiratory, cardiovascular etc)
  - Time:
    - Key events within each body system section are ordered by time.
    - This however means that key events are consecutive, but events within the paragraph need not be consecutive.

      → Big challenge to handle time and express it properly!

Document plan (part)



The baby was intubated at 00:15 and was on CMV. He was given morphine and suxamethonium.

Things to note

- Messages are instances in the ontology, created during signal analysis.
- Relations between instances are determined during data interpretation, and mapped to rhetorical relations during document planning.

Part 6

CHALLENGES AND OPEN QUESTIONS

Problem 1: Continuity

- Choosing events based only on importance can have bad consequences.

Example from the BT-45 prototype system:

TcPO2 suddenly decreased to 8.1*. SaO2 suddenly increased to 92.* ***TcPO2 suddenly decreased to 9.3.***

- System selected two TcPO2 events that were important, but didn't include one event that involved a rise in TcPO2 before the second decrease.

Problem #1: Continuity

Solution in BT-Nurse

- Events which aren't important "on their own" can become important in context!
- After importance-based content selection, check for narrative discontinuities, and insert any missing events.

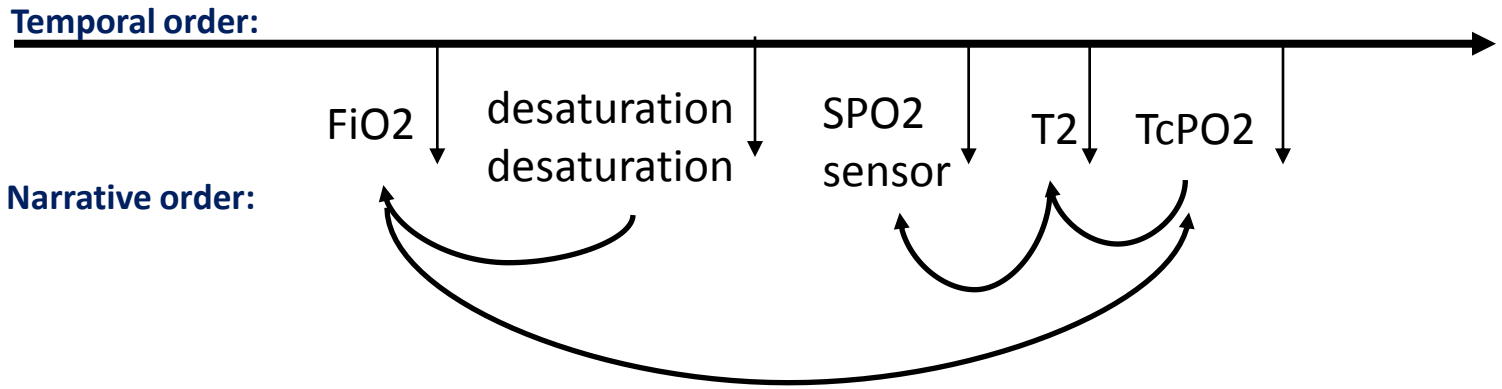
TcPO2 suddenly decreased to 8.1. SaO2 suddenly increased to 92. ***After increasing to 19, TcPO2 suddenly decreased to 9.3***.

Problem #2: Time

- Key events are ordered chronologically, but events within a paragraph are not.

Example, from the BT-45 prototype system:

By 14:40 there had been 2 successive desaturations down to 68. Previously FIO2 had been raised to 32%. Tcpo2 decreased to 5.0. T2 had suddenly increased to 33.9. Previously the SPO2 sensor had been re-sited.



Problem #2: Time

- In language, we often do not report things in the order in which they happened.
- This is also the case in the BT systems.
- But the challenge is to find a way to express temporal relations in a way that is easy to understand...