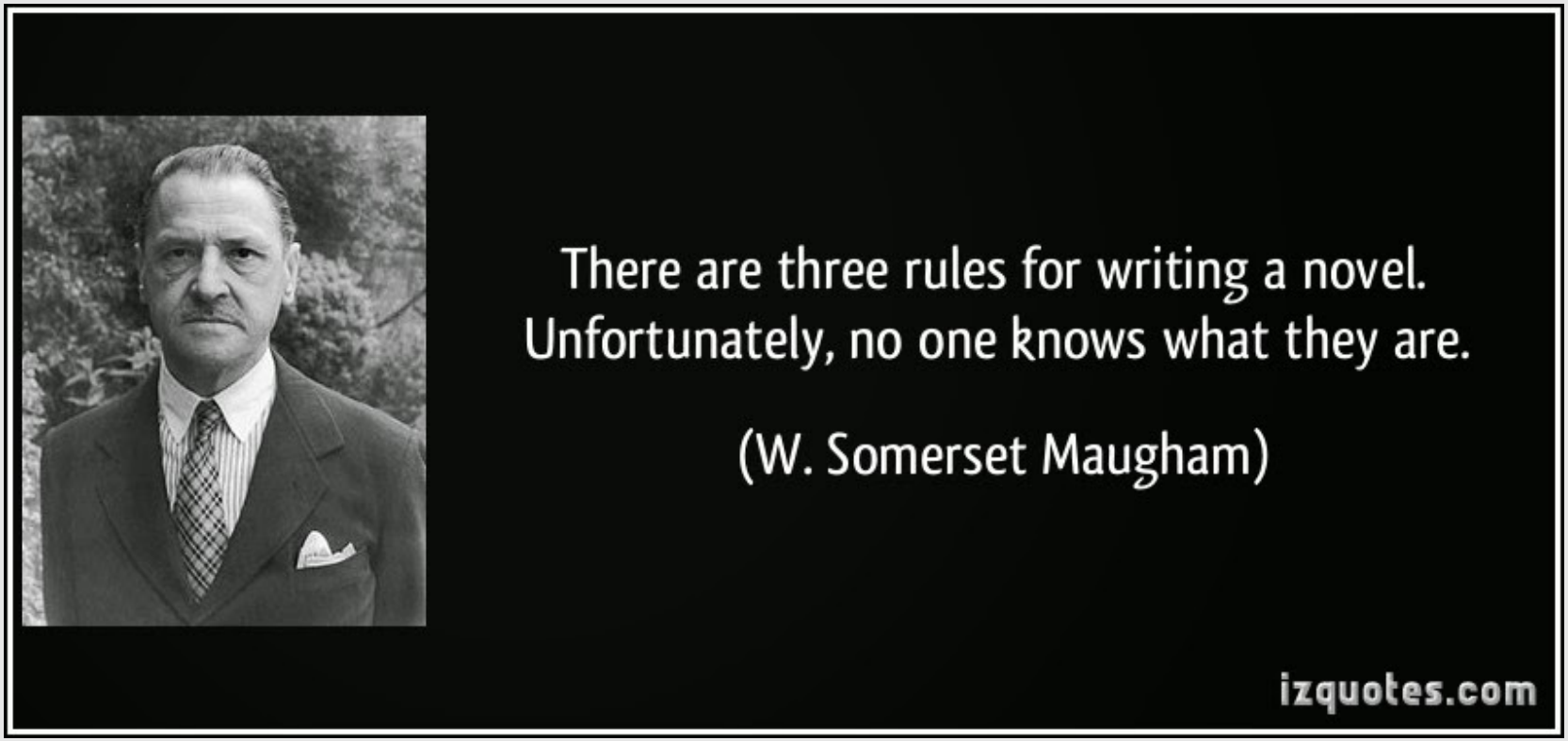
# COM6115: Text Processing

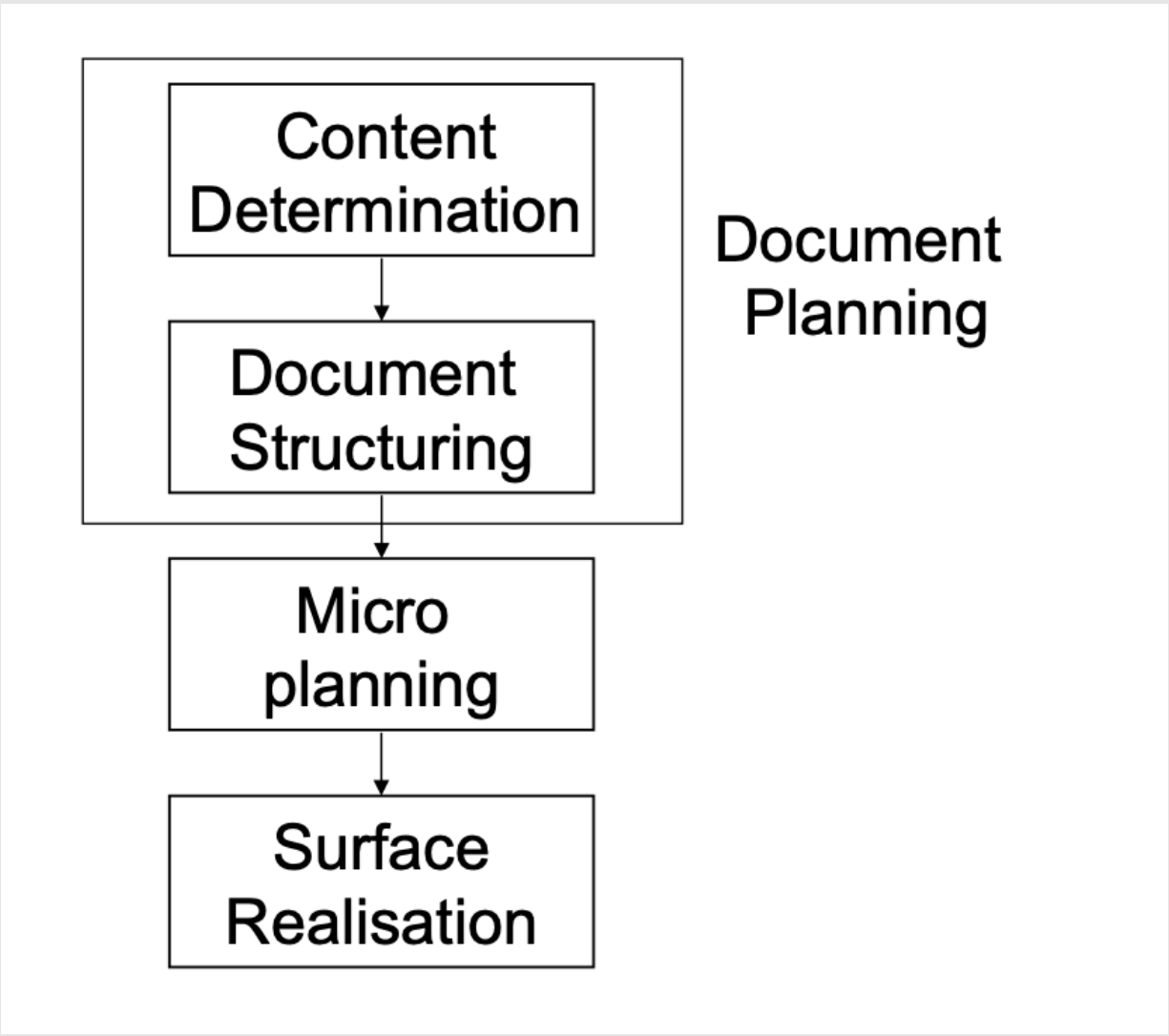
## *Natural Language Generation 2*

Chenghua Lin

Department of Computer Science
University of Sheffield


There are three rules for writing a novel.
Unfortunately, no one knows what they are.
(W. Somerset Maugham)
izquotes.com

# The Architectural View



# Document Planning

- Problem: Usually the output text can only communicate a small portion of the input data
  - Which bits should be communicated?
  - How should information be ordered and structured?

- First stage of NLG
- Goals:
  - **Decide on content**: to determine what information to communicate
  - **Decide on rhetorical structure**: to determine how to structure the information to make a coherent text

# How to Choose Content

| Feature | P | R | F1 | Acc | P | R | F1 | Acc | P | R | F1 | Acc | P | R | F1 | Acc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Academic** | | | | **Conversation** | | | | **Fiction** | | | | **News** | | | |
| GloVe | 65.2 | 67.5 | 66.3 | 92.1 | 58.4 | 62.6 | 60.4 | 94.1 | 60.1 | 55.6 | 57.8 | 91.4 | 69.3 | 64.9 | 67.0 | 90.1 |
| ELMo | 65.1 | 74.1 | 69.3 | 92.4 | 67.6 | 65.1 | 66.4 | 95.2 | 62.3 | 68.4 | 65.2 | 92.3 | 72.6 | 73.4 | 73.0 | 91.6 |
| $BERT_{17}$ | 67.3 | 71.7 | 69.4 | 92.7 | 70.9 | 63.0 | 67.7 | 95.7 | 70.3 | 65.9 | 68.1 | 93.5 | 74.0 | 71.1 | 72.6 | 91.6 |
| GE | 66.9 | 74.6 | 70.5 | 92.8 | 63.3 | 69.3 | 66.1 | 94.9 | 65.8 | 65.5 | 65.7 | 92.8 | 73.1 | 74.5 | 73.8 | 91.8 |
| $GB_{17}$ | 64.7 | 77.2 | 70.4 | 92.5 | 68.1 | 67.5 | 67.8 | 95.7 | 70.3 | 67.6 | 68.9 | 93.7 | 74.3 | 71.5 | 72.9 | 91.7 |
| $EB_{17}$ | 71.8 | 72.3 | 72.0 | 93.5 | 69.9 | 66.3 | 68.1 | 95.8 | 72.9 | 64.8 | 68.6 | 93.6 | 76.1 | 70.5 | 73.2 | 91.9 |
| $GEB_{17}$ | 72.7 | 72.0 | **72.3** | 93.8 | 74.0 | 64.9 | **69.1** | 95.9 | 75.9 | 67.1 | **71.2** | 94.3 | 77.7 | 71.4 | **74.4** | 92.4 |
| | **Verb** | | | | **Adjective** | | | | **Noun** | | | | **Adverb** | | | |
| GloVe | 60.2 | 57.2 | 58.7 | 84.9 | 54.9 | 42.2 | 47.7 | 90.1 | 59.1 | 50.5 | 54.5 | 88.6 | 49.4 | 49.4 | 49.4 | 93.0 |
| ELMo | 62.7 | 70.3 | 66.3 | 86.6 | 46.7 | 54.9 | 50.5 | 88.5 | 61.5 | 58.6 | 60.0 | 89.5 | 57.6 | 51.9 | 54.6 | 94.0 |
| $BERT_{17}$ | 63.3 | 72.2 | 67.5 | 89.9 | 54.7 | 49.1 | 51.8 | 90.2 | 66.8 | 51.7 | 58.3 | 90.0 | 66.7 | 45.5 | 54.1 | 94.7 |
| GE | 62.4 | 68.9 | 65.5 | 86.4 | 56.9 | 58.7 | **57.8** | 90.8 | 62.4 | 59.9 | 61.1 | 90.1 | 53.7 | 56.5 | 55.1 | 93.6 |
| $GB_{17}$ | 64.7 | 69.1 | 66.8 | 87.1 | 58.4 | 53.8 | 56.0 | 90.9 | 65.0 | 57.7 | 61.1 | 90.1 | 61.3 | 49.4 | 54.7 | 94.3 |
| $EB_{17}$ | 66.9 | 69.0 | 67.9 | 87.8 | 53.7 | 53.2 | 53.4 | 90.1 | 73.4 | 49.5 | 59.1 | 90.8 | 63.3 | 49.4 | 55.5 | 94.5 |
| $GEB_{17}$ | 71.6 | 67.4 | **69.4** | 88.9 | 62.8 | 53.5 | **57.8** | 91.6 | 69.9 | 54.5 | **61.3** | 90.7 | 69.1 | 49.4 | **57.6** | 95.0 |

Table 3: Word embedding feature analysis on different genres and PoS of VUA-all-POS development set.

# Content Determination

- The most important aspect of NLG!
  - If we get content right, users may not be too fussed if language isn't perfect
  - If we get content wrong, users will be unhappy even if language is perfect
- Also the most domain-dependent aspect
  - Based on domain, user, tasks more than general knowledge about language

# How to Choose Content

- Theoretical approach: deep reasoning based on deep knowledge of user, task, context, etc
- Pragmatic approach: write schemas which try to imitate human-written texts in a corpus
- Statistical approach: use learning techniques to learn content rules from corpus

# Theoretical Approach

- Deduce what the user needs to know, and communicate this
- Based on in-depth knowledge
  - User (knowledge, task, etc)
  - Context, domain, world
- Use AI reasoning engine
  - e.g. applies logical rules to the knowledge base to deduce new information

- Not feasible in practice
  - Lack knowledge about user
  - Lack knowledge of context
  - Very hard to maintain knowledge base, e.g., new users, new regulations, etc.

# Statistical Approach

- Statistical/learning techniques (including deep learning)
  - ◇ Parse corpus, align with source data, use machine learning algorithms to learn content selection rules/schemas/cases
    - Modelling the coherence of discourse, Barzilay and Lapata, 2005
    - NBA boxscore-data, Wiseman et al, 2017
    - E2E Challenge, Dušek et al, 2019
- Worth considering if large corpora available

# Pragmatic Approach: Schema

- Analyse corpus texts (after aligning them to data), and manually infer content and structure rules.
- Typically based on imitating patterns seen in human-written texts (i.e., corpus)
  - ◇ Revised based on user feedback
- Specify structure as well as content

# Text Structure

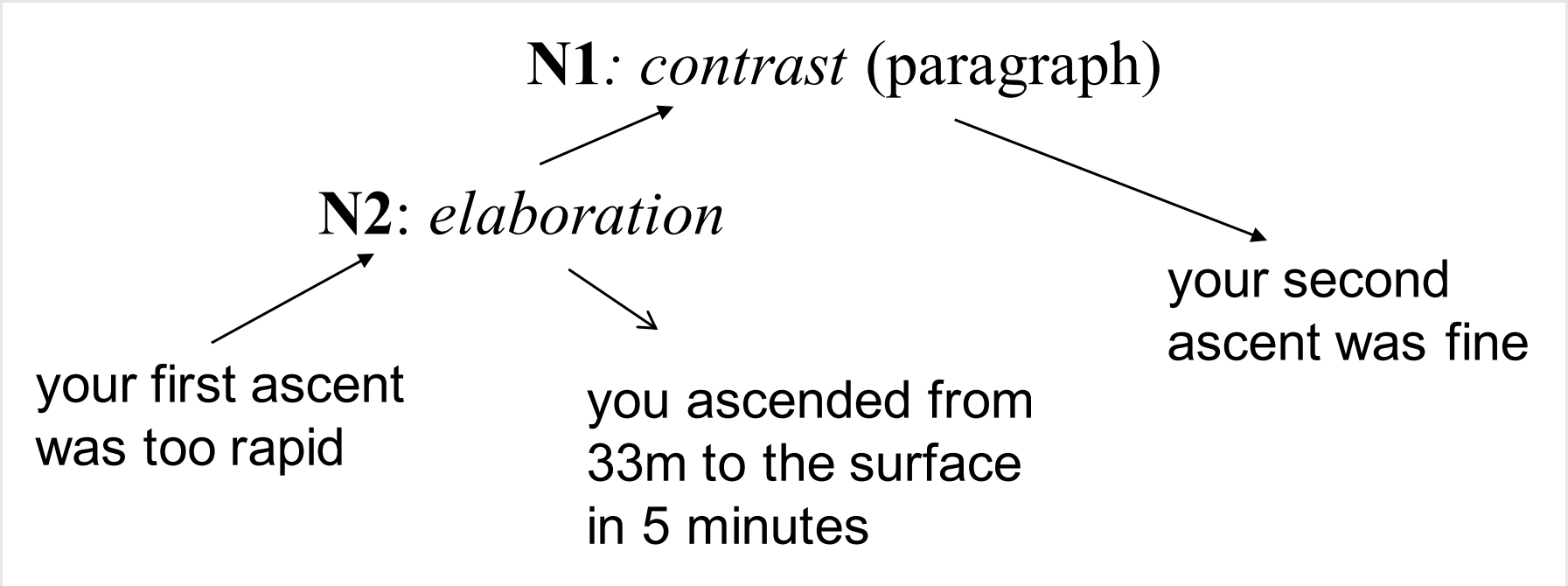
Rhetorical Relations: describe how the parts of a text are linked to each other. The common ones are:

- CONCESSION (although, despite)
- CONTRAST (but, however)
- ELABORATION (usually no cue)
- EXAMPLE (for example, for instance)
- REASON (because, since)
- SEQUENCE (and, also)

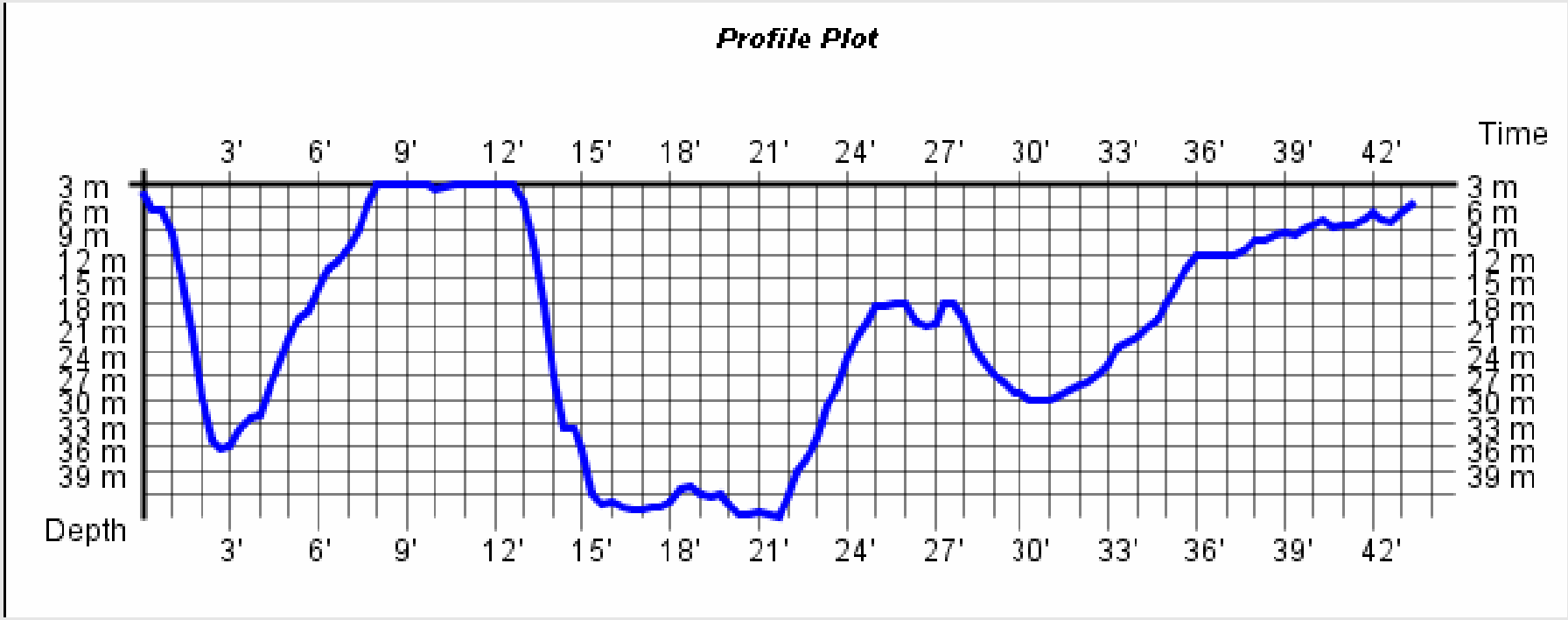
Research community does not agree; many different sets of rhetorical relations proposed.

# Example

- Your first ascent was too rapid; you ascended from 33m to the surface in 5 minutes. However, your second ascent was fine.



# Scubatext Example: Input



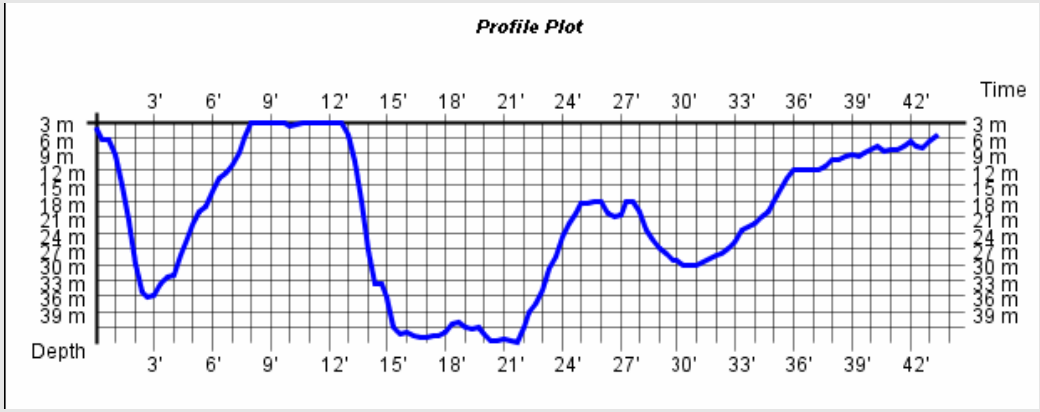
# Input Segments

| diveNo | segNo | iTime | iValue | fTime | fValue |
|---|---|---|---|---|---|
| 1460 | 1 | 0 | 1.3 | 60 | 6.3 |
| 1460 | 2 | 60 | 6.3 | 140 | 32.2 |
| 1460 | 3 | 140 | 32.2 | 480 | 0 |
| 1460 | 4 | 480 | 0 | 760 | 0 |
| 1460 | 5 | 760 | 0 | 920 | 38.9 |
| 1460 | 6 | 920 | 38.9 | 1300 | 41.6 |
| 1460 | 7 | 1300 | 41.6 | 1500 | 15.5 |
| 1460 | 8 | 1500 | 15.5 | 1860 | 27.2 |
| 1460 | 9 | 1860 | 27.2 | 2160 | 9.2 |
| 1460 | 10 | 2160 | 9.2 | 2600 | 2.7 |

# Corresponding Corpus Text

- Your first ascent was a bit rapid; you ascended from 33m to the surface in 5 minutes, it would have been better if you had taken more time to make this ascent. You also did not stop at 5m, we recommend that anyone diving beneath 12m should stop for 3 minutes at 5m. Your second ascent was fine.

# Align corpus text with data




Input: 1460 3 140 32.2 480 0
Output: (representation of)
Your first ascent was a bit rapid; you ascended from 33m to the surface in 5 minutes, it would have been better if you had taken more time to make this ascent.

Input: 1460 10 2160 9.2 2600 2.7
Output: (representation of)
Your second ascent was fine.

# Possible Content Rules

- Describe segments that end (near) 0
  - And that don't start at 0
  - Also segment at end of dive
- Give additional info about such segments whose slope is too high
  - Explain risk
  - Say what should have happened

# Possible Ordering Rules

- Break up dive into sections, where each section starts and ends at surface
- For each section, start with most important safety issue (or say dive was fine if no safety issue)
- Then add less important safety issues
- Then say something about what was done well
- ...

# More Examples

- We have just looked at one example here!
- Need to repeat process for at least 20-30 examples, which cover spread of possible cases (including special cases)
- Merge rules and deal with conflicts
  - Often causes by different corpus authors writing differently;
  - may give priority to one particular author, and imitate his style

# Pseudocode example

```
Schema ScubaSchema
  for each ascent A in data set
     if ascent is too fast
              add unsafeAscentSchema(A)
     else
              add safeAscentSchema(A)
  set rhetorical relation
```

# Creating Schemas

- Usually just written as code in Java or other standard programming languages
- Creating schemas is an **art**, no solid methodology (yet)
- Problems
  - Corpus texts likely to be inconsistent
    - Especially if several authors wrote texts
  - Some cases not covered in the corpus
    - Unusual cases, boundary cases

# Advanced: User-Adaptation

- Texts should depend on
  - User's personality
  - User's domain knowledge (how much do we need to explain)
  - User's vocabulary (can we use technical terms in the text)
  - User's task (what does he need to know)
- Hard to get this information...

# Personality and Perspectives

- Text can communicate perspectives, e.g.,
  - Smoking is killing you
  - If you keep on smoking, your health may keep on getting worse
  - If you stop smoking, your health is likely to improve
  - If you stop smoking, you'll feel better
- How to choose between these?
- Depends on personality of reader
  - Some people react better to positive messages, others to negative messages
  - Some react better to short direct messages, others want these weakened ("may", "is likely to")
  - Hard to predict...

# Conclusion

- Content determination is the first and most important aspect of NLG
  - ◇ What information should we communicate?
- Mostly based on imitating what is observed in human-written texts
  - ◇ Using schemas, written in Java
- Also decide on structure
  - ◇ Tree structure, rhetorical relations