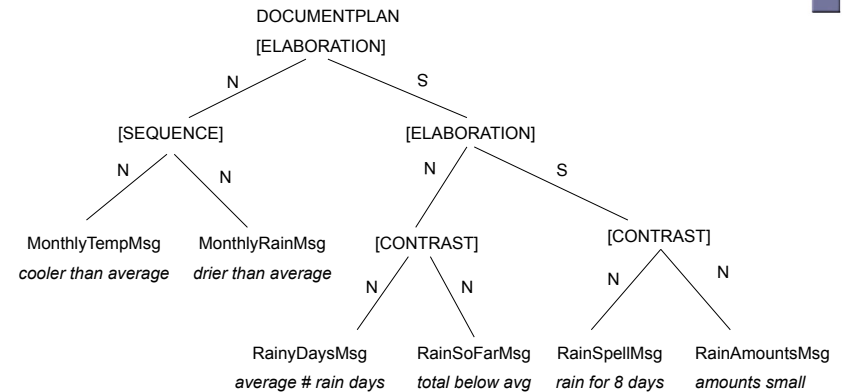# Sentence Planning 1: Lexical Choice

Lecture 9

February 18, 2012
Thanks to Ehud Reiter

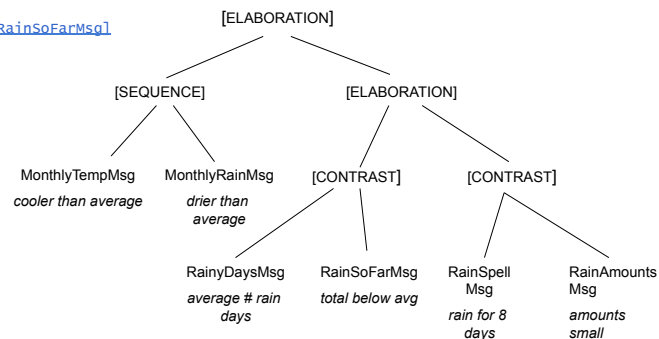*Reading: Chapter 5, Reiter and Dale*

A WeatherReporter Document Plan

```
                          DOCUMENTPLAN
                          [ELABORATION]
                      N                   S
              [SEQUENCE]                   [ELABORATION]
           N            N                 N            S
   MonthlyTempMsg   MonthlyRainMsg     [CONTRAST]      [CONTRAST]
   cooler than average  drier than average  N    N      N        N

                    RainyDaysMsg  RainSoFarMsg  RainSpellMsg  RainAmountsMsg
                    average # rain days  total below avg  rain for 8 days  amounts small
```

A WeatherReporter Document Plan

WeatherSummary →
    TemperatureInformation RainfallInformation
TemperatureInformation →
    MonthlyTempMsg [ExtremeTempInfo] [TempSpellsInfo]
RainfallInformation →
    MonthlyRainfallMsg [RainyDaysInfo]
    [RainSpellsInfo]
RainyDaysInfo →
    RainyDaysMsg [RainSoFarMsg]
...

```
                    [ELABORATION]
              [SEQUENCE]        [ELABORATION]
      MonthlyTempMsg  MonthlyRainMsg  [CONTRAST]   [CONTRAST]
      cooler than average  drier than average

                    RainyDaysMsg  RainSoFarMsg  RainSpell  RainAmounts
                                                 Msg        Msg
                    average # rain  total below avg  rain for 8  amounts
                    days                             days        small
```

A Very Simple Realizer

- Produce one output sentence per message in the document plan using *templates*
- A specialist fragment of code for each message type determines how that message type is realised

Example: For the MonthlyTemperatureMsg:

```
TempString = case (TEMP - AVERAGETEMP)
    [2.0 … 2.9]:    'very much warmer than average.'
    [1.0 … 1.9]:    'much warmer than average.'
    [0.1 … 0.9]:    'slightly warmer than average.'
    [-0.1 … -0.9]:  'slightly cooler than average.'
    [-1.0 … -1.9]:  'much cooler than average.'
    [-2.0 … -2.9]:  ' very much cooler than average.'
endcase
Sentence = 'The month was' + TempString
```

One Message per Sentence

- The Result:

```
The month was cooler than average.
The month was drier than average.
There was the average number of rain days.
The total rain for the year so far is well below average.
There was rain on every day for 8 days from 11th to 18th.
Rainfall amounts were mostly small.
```

- What we'd really like:

```
The month was cooler and drier than average.  Even though
there was an average number of rain days this month, the
total rain for the year so far is well below average.
There was rain every day for 8 days from 11th to 18th, but
rainfall amounts were mostly small.
```
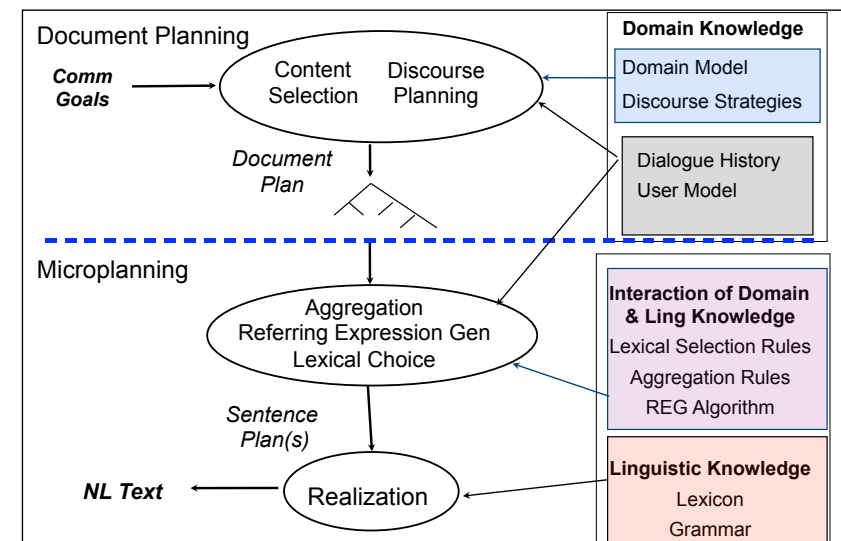
5

We Need

- **Lexicalisation**: choose words/phrases to express concepts and relations
- **Aggregation**: How/when to combine phrases into sentences
- **Referring expression generation**: choose words/phrases to refer to objects

Unfortunately, these tasks are not independent of one another!
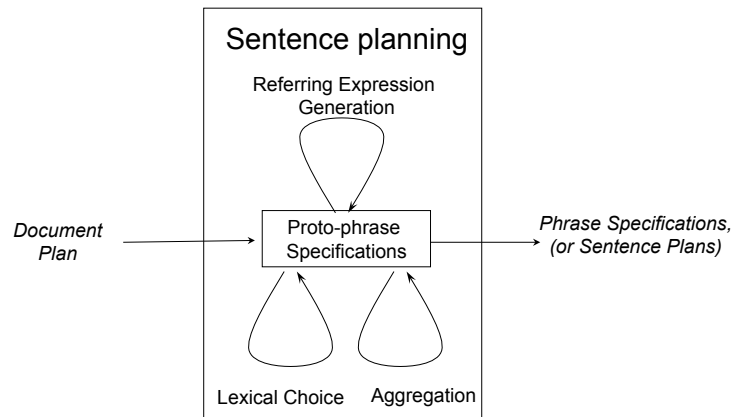
Back in the day

- Early (70s & 80s) NLG systems had two stages
  - Strategic component: *What to say.* Text content and structure
  - Tactical component: *How to say it.* Sentence level grammar and morphology

- Why microplanning?
  - Need for generating longer, richer, more fluent text
  - Reasoning systems not guaranteed to deliver info in chunks that correspond neatly to surface form sentences
- Where does referring expression generation belong?
- What about aggregation?

NLG Tasks and Pipeline Architecture

Document Planning

Comm Goals → Content Selection   Discourse Planning

Document Plan

Microplanning

Aggregation
Referring Expression Gen
Lexical Choice

Sentence Plan(s)

NL Text ← Realization

Domain Knowledge
Domain Model
Discourse Strategies
Dialogue History
User Model

Interaction of Domain & Ling Knowledge
Lexical Selection Rules
Aggregation Rules
REG Algorithm

Linguistic Knowledge
Lexicon
Grammar

Microplanning Tasks Interact

Sentence planning

Referring Expression
Generation

Document
Plan → Proto-phrase
Specifications → *Phrase Specifications,
(or Sentence Plans)*

Lexical Choice   Aggregation

Examples

1 (a) Bob likes fishing and swimming.

 * (b) Bob likes fishing and to swim.

2 (a) Heavy rain fell from the 12th to the 14th.

 * (b) From the 12th to the 14th was the wettest three-day period so far this year.

Pipelined Microplanning

Document
Plan
↓
Lexical choice
↓ PPSs
Aggregation
↓ PPSs
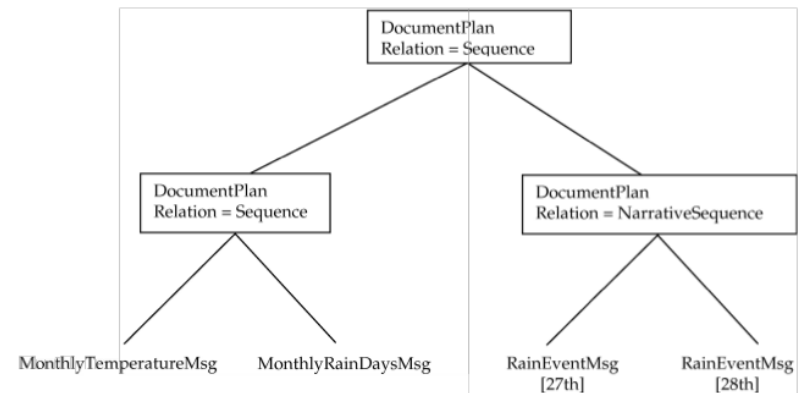Referring Expression Gen
↓
Phrase Specs
(Sentence Plans)
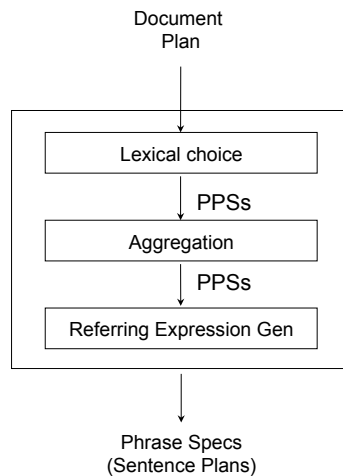
Lexical choice: selects words and syntactic structures to express messages. Result is proto-phrase-spec. PPSs may contain refs to domain entities

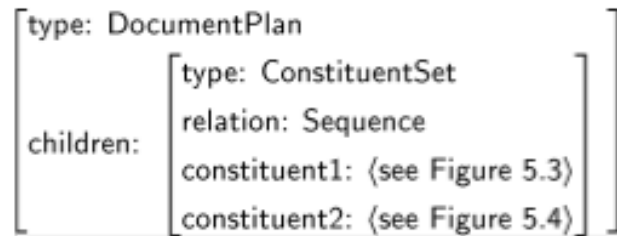**Aggregation:** combines multiple PPSs into single PPS

REG: takes each PPS and replaces references to domain entities with a phrase spec corresponding to a noun phrase that will uniquely identify that entity to the reader
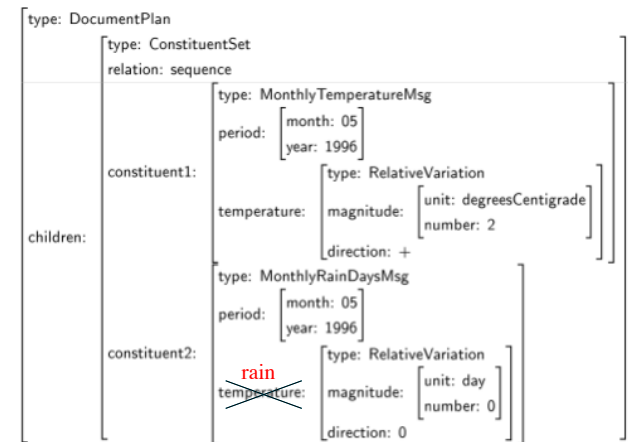
Example: Weather Reporter Document Plan

DocumentPlan
Relation = Sequence

DocumentPlan
Relation = Sequence

DocumentPlan
Relation = NarrativeSequence

MonthlyTemperatureMsg   MonthlyRainDaysMsg   RainEventMsg [27th]   RainEventMsg [28th]

The month was slightly warmer than average, with the average number of rain days. Heavy rain fell on the 27th and 28th.

Top-level Document Plan

```
[type: DocumentPlan
           [type: ConstituentSet
           relation: Sequence
children:  constituent1: ⟨see Figure 5.3⟩
           constituent2: ⟨see Figure 5.4⟩]
]
```

The month was slightly warmer than average, with the average number of rain days. Heavy rain fell on the 27th and 28th.

Doc Plan for Constituent 1


```
type: DocumentPlan
          [type: ConstituentSet
          relation: sequence
                              [type: MonthlyTemperatureMsg
                              period:    [month: 05
                                          year: 1996]
          constituent1:                            [type: RelativeVariation
                              temperature: magnitude:  [unit: degreesCentigrade
                                                        number: 2]
children:                                          direction: +]
                              [type: MonthlyRainDaysMsg
                              period:    [month: 05
                                          year: 1996]
          constituent2:                  rain
                              temperature: magnitude:  [unit: day
                                                        number: 0]
                                                   direction: 0]
```

The month was slightly warmer than average, with the average number of rain days.

Document Plans Specify Info Abstractly

<u>Do specify</u>
- Content to be included
- General semantic similarity
  - Constituent 1 includes general info messages
 - Constituent 2 includes all significant event messages

<u>Don't specify</u>
- Words to be used to communicate the notion of average temperature
- How to package content into sentences
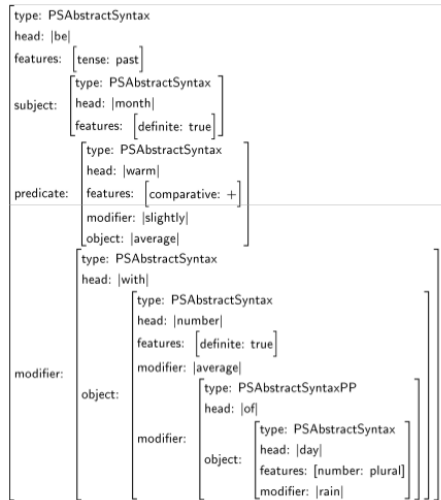- How to refer to fifth month of 1996

Top level Text Specification

Microplanner transforms document plans into a text specification made up phrase specifications

Continuing our example:

```
Type:  TextSpecification
Constitutent1:  <Sentence1>
Constitutent2:  <Sentence2>
```
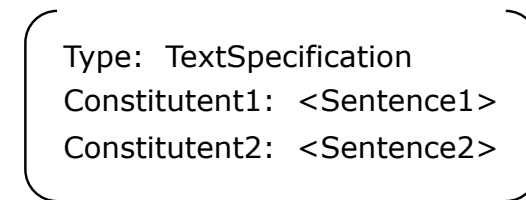
Consists of two phrase specifications (sentences)

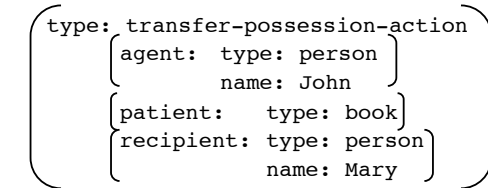## Phrase Specification for Sentence 1

- **Aggregation:** combines 2 messages from doc plan
- **Lexical Choice:** Expresses avg temp for month is 2 degrees higher than usual as: *"slightly warmer than average"*
- **REG:** uses *"the month"* to refer to May 1996
- Phrase Specs can be transformed to input to most realizers, e,g., HLDS

Sentence planning is hard

Problem: Many ways to express a message in words

Example:

$$\begin{bmatrix} \text{type: transfer-possession-action} \\ \text{agent:} \begin{bmatrix} \text{type: person} \\ \text{name: John} \end{bmatrix} \\ \text{patient:} \begin{bmatrix} \text{type: book} \end{bmatrix} \\ \text{recipient:} \begin{bmatrix} \text{type: person} \\ \text{name: Mary} \end{bmatrix} \end{bmatrix}$$

- *John gave the book to Mary.*
  - *John gave Mary the book.*
  - *Mary took the book from John.*
  - *John gave the book. Mary took it.*
  - *...*
- Which one should we use?

So many choices…

- Lexical choice: *sold* vs. *bought* vs. *purchased*
- Aggregation: one sentence vs. two
- Referring Expression Generation:
  - *the book, it, a novel, Emma, Jan Austen's fourth novel ?*

- How do we make these choices?
  - Rule-based
  - Statistically
- Which text is best?
  - Easiest to comprehend?
  - Most likely to be remembered?
  - Most persuasive?
  - Most appropriate style?

How to Decide

- Theoretical
  - Define what "best" means, make microplanning choices that optimise it
  - Hard to do in practice because we don't have good models of the effects of choices
    - bought vs. purchased
- Pragmatic
  - Imitate corpus
  - Problems:
    - corpus texts may not be ideal from a microplanning perspective
    - authors may produce what's easy for them to write, not what's easy for others to read (Shaw 1998)

Lexical choice

- Which word/phrase should be used to communicate a given domain concept or relation?
  - *bought* vs. *sold*
  - *took vs. bought*
  - *took vs. stole*
  - *bought vs. purchased*
  - *ascended* vs. *rose* vs. *surfaced*
  - *too fast* vs. *too rapidly*
  - *recommend* vs. *suggest*
  - *food vs. grub*

Near Synonyms

- Near synonyms grouped into clusters
  - Do-something-wrong: *Mistake, error, blunder, …*
  - Same core meaning
  - Peripheral distinctions
    - *Blunder* implies stupidity
    - Distinctions are usually semantic (meaning)
- No empirical testing or corpus work
  - Based on synonym dictionary

(Edmonds & Hirst, 2002)

Issues that affect lexical choice

- Frequency (affects readability)
  - *lie* vs. *prevarication*
- Formality:
  - *Error* vs. *howler*
- Connotation
  - *Error* vs. *mistake*
- Focus, expectations, alignment
  - *not many, few, a few, only a few* [students failed the exam]
- Technical terms
  - (statistics) *standard error*, not
  - *standard mistake*

Yet More Issues

- Linguistic context
  - *Answered questions correctly* vs.
  - *Answers are correct*
- Convention
  - Temperature *falls,* Wind speed *eases*
- Variation between users
  - *Grammar mistake* includes spelling?
  - *By evening* means 6PM or midnight?
  - *Blue baby* means skin is blue or O2 low
- Vary words to make texts interesting?
  - Standard writing advice
    - *I met John and then I ran into Joan*
    - *I met John and then I met Joan*

Example: Contextual and Pragmatic Factors

a. *It rained for seven days from the 20th to the 26th.*
b. *It rained for seven days from the 20th.*
c. *It rained from the 20th to the 26th.*
d. *It rained on the 20th, 21st, 22nd, 23rd, 24th, 25th, and 26th.*
e. *It rained all of Thanksgiving week.*

Default is *c*.

Pragmatic Goal: Use *d* to emphasize unusual rainy period this month

User knowledge: Use *e* if user knows when Thanksgiving week is

Repetition vs. conciseness: Use *a* to reduce possibility of misunderstanding. But, use *b* or *c* where space is limited

Weather Reporter Example - Cont'd

f. *It snowed for five days from the 10th, and it rained from the 20th to the 26th.*
g. *It snowed for five days from the 10th, and it rained for seven days from the 20th.*

Consistency with previous text:

- Use realization that has been used in a previous message for another instance of the same concept, as in *g*
- Allows user to compare the two spells
- May also make aggregation easier

Expressing Discourse Relations

h. *The month was much wetter than average, but rainfall for the year is still very depleted.*
i. *The month was much wetter than average, however rainfall for the year is still very depleted.*
j. *The month was much wetter than average, although rainfall for the year is still very depleted.*
k. *Although the month was much wetter than average, rainfall for the year is still very depleted.*
l. ** However the month was much wetter than average, rainfall for the year is still very depleted.*

Note:

Rules for deciding which cue word to use to express a discourse relation depend on subtle pragmatic and domain factors

But and **although** cannot be used interchangeably *(Knott, 1996)*

Approaches

- Rule-based
  - Use knowledge of word meaning to choose one that best fits context
- Corpus-based
  - Identify words human authors use in corpus of texts of the kind to be generated
  - Use frequencies from large, domain independent corpus (e.g., BNC, ANC, Switchboard, WSJ, …)
- Statistical
  - Collocations
  - N-gram language models

Rule Based Approach: Example

RainSpellMsg Template

$$\begin{bmatrix} \text{type: PPSAbstractSyntax} \\ \text{head: } |\text{fall}| \\ \text{features: } [\text{tense: past}] \\ \text{subject: } \begin{bmatrix} \text{type: PPSAbstractSyntax} \\ \text{head: } |\text{rain}| \\ \text{features: } [\text{definite: false}] \\ \text{modifier: lexicalise(rainType)} \end{bmatrix} \\ \text{modifier: lexicalise(spell)} \end{bmatrix}$$

To lexicalise(spell):

if spell.begin.day = spell.end.day **then**
    return a proto-phrase spec for *on the ith*
else if spell.begin.day is day before spell.end.day **then**
    return a proto-phrase spec for *on the ith and jth*
else
    return a proto-phrase spec for *from the ith to the jth*

Proto-phrase spec for *on the ith*

$$\begin{bmatrix} \text{type: PPSAbstractSyntax} \\ \text{head: } |\text{on}| \\ \text{object: } \begin{bmatrix} \text{type: PPSAbstractSyntax} \\ \text{head: spell.begin.day} \\ \text{features: } \begin{bmatrix} \text{definite: true} \\ \text{inflection: ordinal} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Proto-phrase spec for *on the ith and jth*

$$\begin{bmatrix} \text{type: PPSAbstractSyntax} \\ \text{head: } |\text{on}| \\ \text{object: } \begin{bmatrix} \text{type: PPSAbstractSyntax} \\ \text{head: } |\text{and}| \\ \text{conj1: } \begin{bmatrix} \text{type: PPSAbstractSyntax} \\ \text{head: spell.begin.day} \\ \text{features: } \begin{bmatrix} \text{definite: true} \\ \text{inflection: ordinal} \end{bmatrix} \end{bmatrix} \\ \text{conj2: } \begin{bmatrix} \text{type: PPSAbstractSyntax} \\ \text{head: spell.end.day} \\ \text{features: } \begin{bmatrix} \text{definite: ellided} \\ \text{inflection: ordinal} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Other syntactic categories?

- What if we need to express the concept of `spell` as a different syntactic category, e.g.,

  The 12th, 13th and 14th was the wettest three-day period seen so far this year.

- Need lexicalisation templates that express `spell` as a noun phrase, in addition to those for expressing `spell` as a prepositional phrase
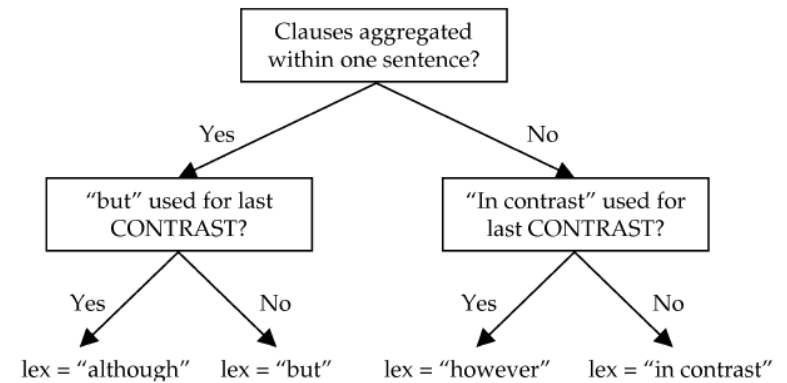
Another Rule-Based Approach: Decision Trees

ACT = INGEST

Obj = solid?
- y → "eat" (7264) "consume" (515)
- n → Obj = liquid?
  - y → "drink" (7674) "quaff" (9) "imbibe" (11)
  - n → Obj = gas?
    - y → Obj = air?
      - y → "breathe" (1061) "inhale" (73)
      - n → "inhale"
    - n → "ingest" (31)
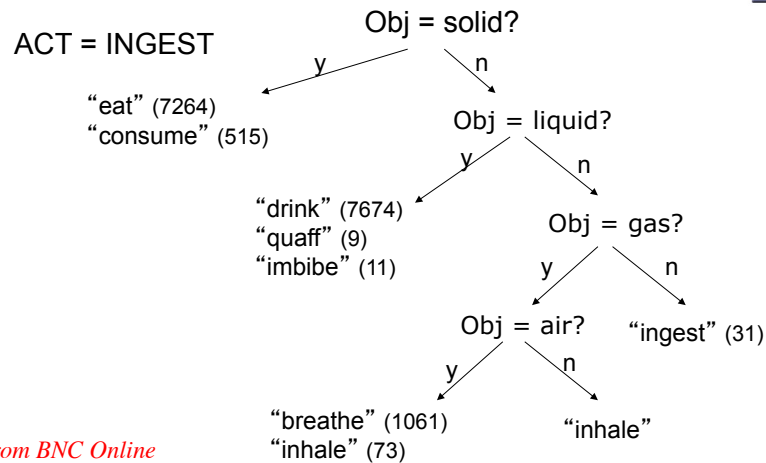
From BNC Online
http://bncweb.info/

Decision Tree for Realising CONTRAST Relation

Clauses aggregated within one sentence?
- Yes → "but" used for last CONTRAST?
  - Yes → lex = "although"
  - No → lex = "but"
- No → "In contrast" used for last CONTRAST?
  - Yes → lex = "however"
  - No → lex = "in contrast"

Rule Based Approach: Problems

- Hard to get to work in practice
  - Hard to get good data about context, user, level of formality, etc
  - Unexpected results
- Example: Application that tests language ability of low-skilled readers and generates report *(Williams and Reiter, 2008)*
  - Should it generate
    - *You got 15 questions right*   or
    - *You got 15 questions correct*
  - Authors thought *right* would be better (13x more common)
  - But, 90% of users preferred *correct*

Corpus-Based

- Works OK, not ideal *(Reiter et al. 2005)*
  - Problem: many domain experts use words that readers may not interpret correctly
    - Use words not familiar to reader (e.g., *grammar*)
    - Use words idiosyncratically (e.g., *later*)
- Need to check whether readers understand the generated text

Reiter et al. approach

- Initially imitate words in corpus
- Look for problem cases
  - Words that users may not understand
  - Words that different authors use differently
- Check these with users, change if appropriate
- Vary connectives, nothing else

Statistical Approach: Collocations

- Collocation: a sequence of words or terms that co-occur more often than would be expected by chance
- Restrictions on how words can be used together, e.g.,
  - which prepositions are used with particular verbs
  - which adjectives and nouns are used together
- Example:
  - *strong tea* vs. *powerful tea*
  - *strong computers* vs. *powerful computers*

Determining collocations

- Measure of association evaluates whether the co-occurrence is purely by chance or statistically significant
- Association scores used to rank results
- Commonly used measures of association
  - mutual information
  - t-test
  - log-likelihood

Bake vs. Roast

roast	bake
Beef, pork, chicken, *ham, *gammon	Bread, cake, pies
	Ham, fish
Potato, chestnuts turnip, crab apples, ?apple	Apple, potato
	Pork chops with apple
coffee beans	Clay, concrete, pottery
metal oar	

N-grams for generation

- N-gram statistics can be used to make many decisions
  - Tense
  - Gender, number agreement
  - Syntactic subcategorization decisions
  - Word choice

  (Langkilde & Knight, 1998)

 - Lots more about this when we talk about statistical approaches to NLG

Summary: Lexical Choice

- Involves mapping domain concepts and relations to words or phrases
- Many issues affect lexical choice
- Use corpora with caution!

References

- P. Edmonds and G. Hirst (2002). "Near-synonymy and lexical choice." *Computational Linguistics, 28(2):* 105--144.
- A Knott (1996). "A Data-Driven Methodology for Motivating a Set of Coherence Relations." PhD thesis, Department of Artificial Intelligence, University of Edinburgh.
- I Langkilde and K. Knight (1998). "The Practical Value Of N-Grams in Generation." In Proc. of 9th International Workshop on Natural Language Generation.
- E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy (2005). "Choosing Words in Computer-Generated Weather Forecasts." *Artificial Intelligence 167*: 137-169.
- J. Shaw (1998). "Clause aggregation using linguistic knowledge." In Proc of 9th International Workshop on Natural Language Generation, pp. 138-147.
- S. Williams and E. Reiter (2008). "Generating basic skills reports for low-skilled readers." *Natural Language Engineering* 14:495-535.

Document Plan as Feature Structure
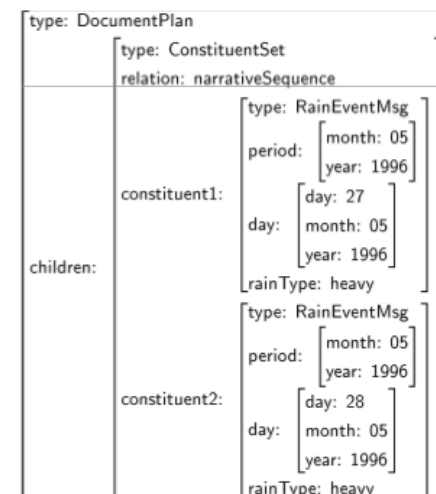
Second constituent of document plan on previous slide

Slide 1

$$\begin{bmatrix} \text{type: RainEventMessage} \\ \text{period:} \begin{bmatrix} \text{month: 05} \\ \text{year: 1996} \end{bmatrix} \\ \text{day:} \begin{bmatrix} \text{day: 27} \\ \text{month: 05} \\ \text{year: 1996} \end{bmatrix} \\ \text{rainType: heavy} \end{bmatrix}$$

Figure 5.13: A simple message

What affects lexical choice?

- Edmonds and Hirst (2002)
  - *Error, mistake, blunder, slip, lapse, faux pas, howler*
  - *Lie, fib, misrepresentation, untruth, prevarication, falsehood*
  - *Correct, correctly, right*
- Class's opinions on choosing between these?

Case Creation

- Additional structure is required to realize the meaning of the semantic representation

```
(*A-KICK
  (AGENT *O-JOHN)
  (PATIENT *O-BALL))
```

"John propelled the ball with his foot"

Case Absorption

- Word chosen to realize a semantic head also implies the meaning conveyed by a semantic role

```
(*A-FILE-LEGAL-ACTION
  (AGENT *O-BOB)
  (PATIENT *O-SUIT)
  (RECIPIENT *O-ACME))
```

"Bob sued Acme"