Lexicalization

Issues in lexicalization

Search techniques

LEXICALIZATION

The task

- **Mostly understood as the choice of open class words (nouns, verbs, adjectives)**

- **Building lexicalized representations achieved by *structure mapping***

- **Related issue: closed class words (e.g., discourse markers)**

- **More distant issue: choice of conceptual predicates for specifications**

A fundamental issue – compositionality

- **Map all parts in the specification onto lexical item without gaps and overlaps**

- **Problems with non 1-to-1 mappings (avoid untractable residues)**

- **Mapping controled by type constraints in text structure to be built**

- **Concept "decide" may be mapped onto "decision" or onto "to decide"**

- **Attaching "important" only works for "decision", but not for "to decide"**
  -> *Generation gap*

PERFORMING LEXICAL CHOICE

Criteria

- **Formal versus colloquial ("policeman" versus "cop")**

- **Attitude towards an object/activity ("anarchist" versus "freedom fighter")**

- **Focus ("sell" versus "buy")**

- **Specificity ("dog" versus "poodle")**

- **Implicature (prototypicality, chosen specificity must be on purpose)**

- **Collocational constraints ("heavy smoker", "grand fumeur")**

Some phenomana

- **Head switching ("John likes to swim", "John schwimmt gerne")**

- **Distribution of information content {"John swam across the river")**

                              ("John a traversé la rivière en nageant")

METHODS

Organizing lexical knowledge

- **Phrasal lexica**

  Chunks of lexical items, covers larger input substructures

- **Discrimination networks**

  Constraints associated with internal nodes, combines generalizations

- **Lexica/taxonomic hierarchies**

  Conceptual coverage, constraints on type and relations to other items

Combination techniques

- **Incrementally building a text structure**

  Using some sort of unifications and back-tracking

- **Partitioning the problem and combining partial solutions**

  Constraint propagation

DISCOURSE RELATIONS

A example

Jones has lots of experience.

He has been on the board for 10 years.

And he 's refused bribes.

So he's honest.

He would really make a good president.

[Cohen 1987]

Diagnosis

- **Relations between facts/assertions not explicitly expressed**

- **Cue phrases (here: and, so) only contribute to a limited extent – ambiguous!**
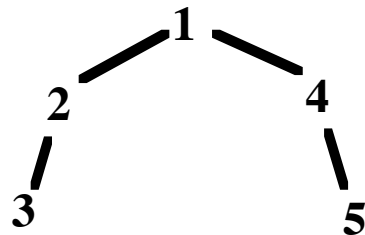
Challenges

- **Reconstructing the intended argumentative structure (in analysis)**

- **Presenting arguments in a natural and understandable form (in generation)**

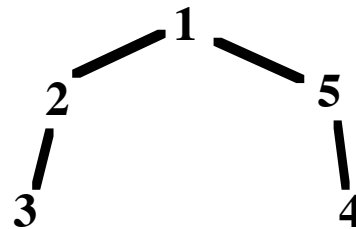# GENERATION – PRESENTING DISCOURSE RELATIONS

Some possible variations

PRE-ORDER	**HYBRID**	**POST-ORDER**
1. Jones would make a good president.	**1. Jones would make a good president.**	1. Jones has been on the board for 10 years.
2. He has lots of experience.	**2. He has lots of experience.**	2. He has lots of experience.
3. He has been on the board for 10 years..	**3. He has been on the board for 10 years.**	3. *And* he 's refused bribes.
4. And he's honest.	**4. *And* he 's refused bribes.**	4. *So* he's honest.
5. He's refused bribes.	**5. *So* he's honest.**	5. He would *really* make a good president.

```
PRE-ORDER                HYBRID                 POST-ORDER
       1                      1                        5
      / \                    / \                      / \
     2   4                  2   5                    2   4
     |   |                  |   |                    |   |
     3   5                  3   4                    1   3
```
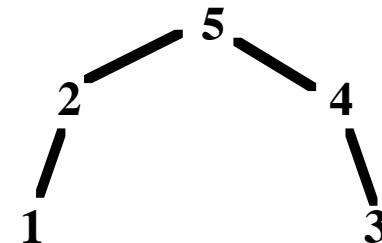
Methods

- **Ordering and cue-phrase selection, embedded in sentence planning**

  (e.g., [Grote, Stede 1998])

- **Decisions guided by heuristics expressing aspects of linguistic/rhetorical adequacy**

  (e.g., [Scott, de Souza 1992])

GENERATING ADVICE-GIVING PARAGRAPHS

Advisor II [Elhadad 1993]

Tasks addressed

Goal of the answer given as input to the sysetm

System extracts goal-relevant information from a knowledge base

Structuring of information into network, filtering and ordering it

Lexical selection involving choice of

- **scalar adjectives ("interesting" vs. "difficult")**
- **judgment determiners ("many" vs. "few")**
- **argumentative connectives ("but" vs. "so")**
- **quasi-synonyms differing in their argumentative connotation ("requires" vs. "includes")**

GENERATING ADVICE-GIVING PARAGRAPHS

Advisor II [Elhadad 1993]

Question: Should I take AI?

Two alternatives with different goal generated from the same database:

Answer 1:	Answer 2:
AI has many programming assignments and it covers a lot of interesting topics, such as NLP, vision, and KR.	AI deals with many interesting topics, such as NLP, vision, and KR.
So it should be interesting.	But it has many assignments which consist of writing papers.
I would recommend it.	So it could be difficult.
	I would not recommend it.

MAPPINGS BETWEEN CONCEPTS AND WORDS

Floating Constraints **[MdKeown, Elhadad, Robin 1998]**

Sources for variations

- **Syntax - e.g., verb choice has consequences on realization of its arguments ("allow one to select" or "allow the selection")**

- **Semantics - word choice depends on concept and its taxonomisation ("rebound" vs. "point" as specializations of "performance" in basketball)**

- **The lexicon choice of words constrain use of other words in a sentence (collocations, such as "to grab rebounds" vs. "to score points")**

- **The domain - different meanings of the same verb in different domains ("rebounded from a 3 day loss" vs. "grabed 20 rebounds")**

- **Pragmatics - perspective, intention, background, previous discourse, … ("half full" vs. "half empty")**

MEANS TO EXPRESS ELEMENTS OF CONTENT

Inter-lexical constraints (operate on the lexicon)

    (restrictions on lexical co-occurrence)

Cross-ranking constraints (ranking among alternatives, expressing multiple units)

    (input concept network not isomorphic to linguistic structure)

- "Wall Street Indexes *opened strongly*"

  (time in *verb*, manner as *adverb*)

- "Stock indexes *surged at the start* of the trading day"

  (time in *PP*, manner as *verb*)

- "The Denver Nuggets beat the Boston Celtics with a narrow margin, 102-101."

  (game result in *verb*, manner as *PP*)

- The Denver Nuggets edged out the Boston Celtics 102-101."

  (game result and manner in *verb*)

VARIATIONS IN EXPRESSING PIECES OF CONTENT
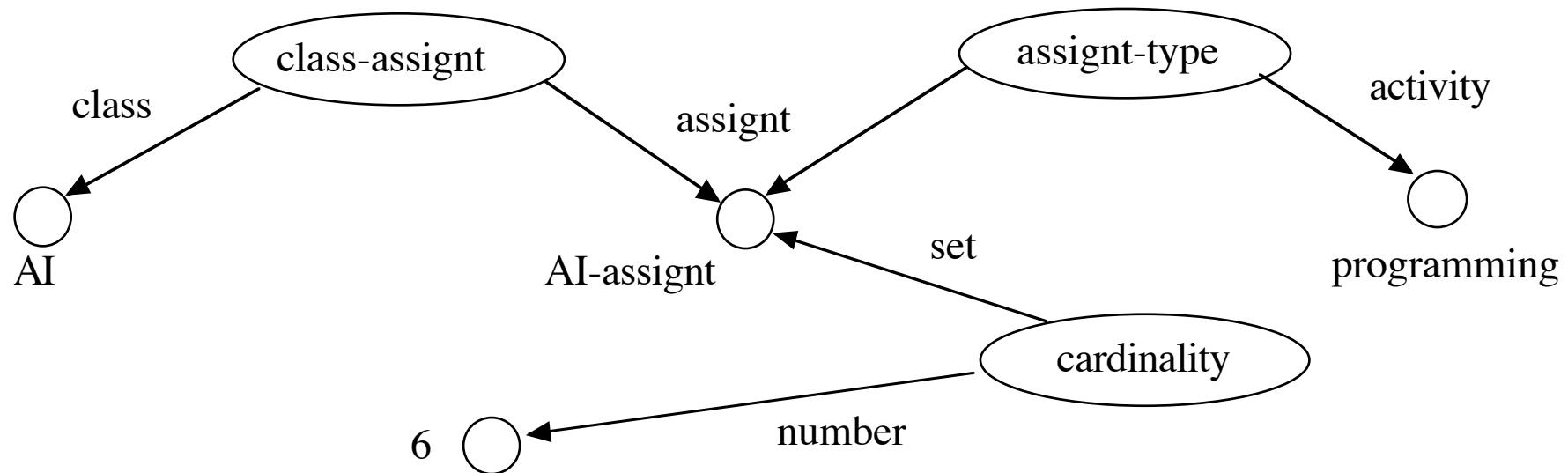
- **"The 6 AI assignments require programming"** (*main clause*)

- **"AI has 6 assignments which involve programming"** (*relative clause*)

- **"AI has 6 assignments of programming nature"** (*PP)*

- **"AI has 6 programming assignments"** (*predicative adjective)*

- **"AI has 6 implementation assignments"** (*noun-noun modifier)*

THE LEXICALIZATION PROCEDURE (EXCERPT)

Handling constituent structure

> **1. Unify top-level input with the lexicon**
>
> **2. Identify constituents in result**
>
> **3. Recursively unify each constituent with the lexicon**
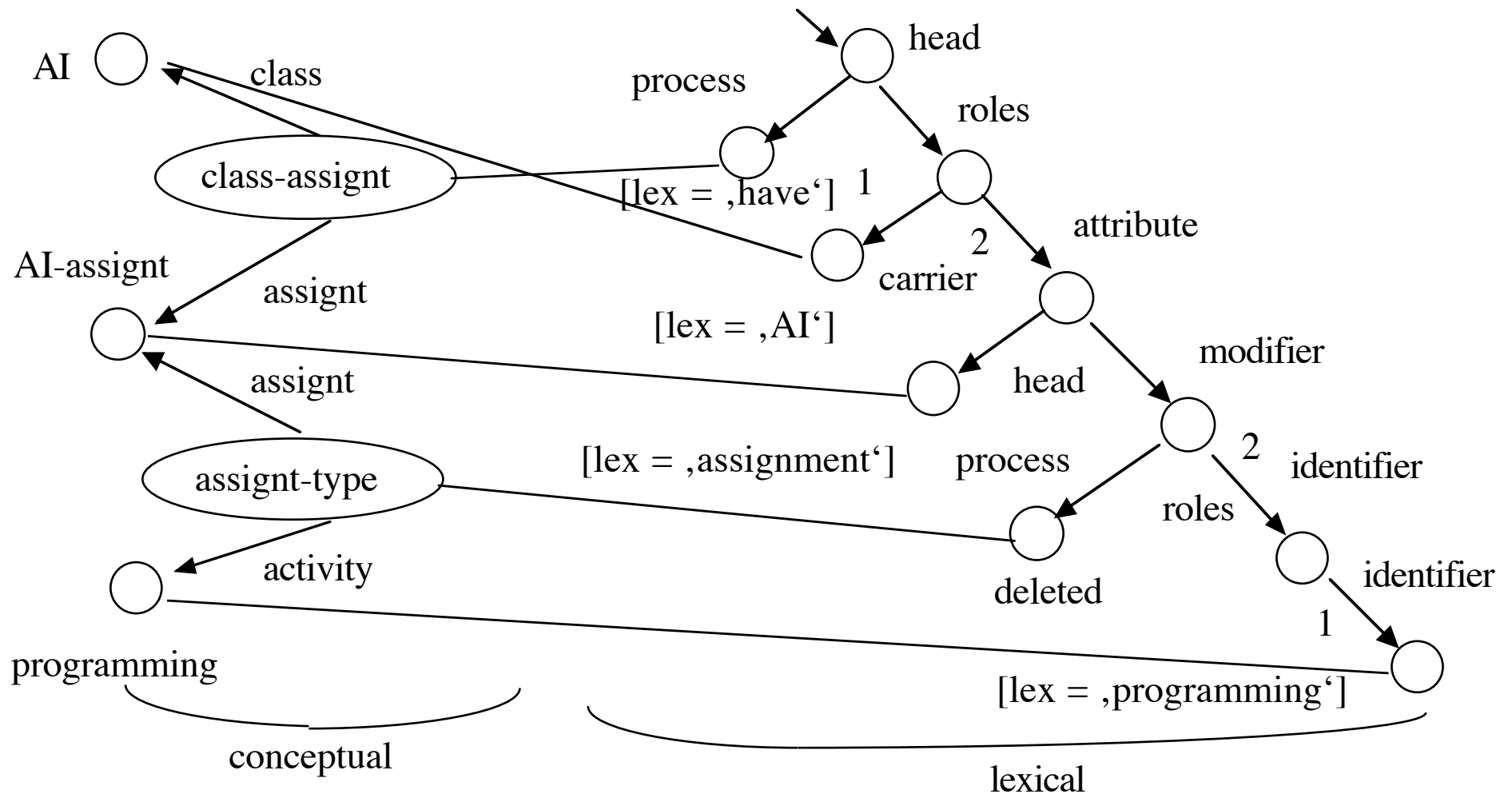
Stages of the lexical chooser

> **1. Phrase planning - processes semantic input to determine syntactic category**
>
> **2. A skeletal functional description of that category enriches input**
>
> **3. Selecting a head word (unifying the top structure with the lexicon)**
>
> **4. Propagating lexical constraints to semantic subconstituents, plus recursion**

Processing techniques

> • **Default control regime is breadth-first**
>
> • **Dependency-directed back-tracking**

MAPPING CONCEPTUAL SPECIFICATIONS - EXAMPLE

AI

class-assign

AI-assignt

class

process

head

roles

[lex = ‚have']   1

2

attribute

carrier

[lex = ‚AI']

assignt

assignt

assign-type

[lex = ‚assignment']   process

head

modifier

2   identifier

roles

activity

deleted

identifier

programming

[lex = ‚programming']

1

conceptual

lexical

REVISION-BASED PLANNING (Robin 1997)

Motivation – corpus observations from sports reports

Concise linguistic forms

Complex sentences (50 words, parse tree depth of 10)

Optional and background facts opportunistically slipped as modifiers

High paraphrasing power

Measurements

Increasing the number of content planning and linguistic realization options

Modeling the mutual constraints relating these options

Applying a dedicated two-pass procedure that

       generates a sentence in a revision-based fashion

THE INFORMATION TO BE CONVEYED

All facts expressed in simple sentences

Charles Barkley scored 42 points. Those 42 points equal his best scoring performance of the season. Danny Ainge is a teammate of Barkley. They play for the Phoenix Suns. Ainge is a reserve player. Yet he scored 21 points. The high scoring performance of Barkley and Ainge helped the Suns defeat the Dallas Mavericks. The Mavericks played on their homecourt in Texas. They had already lost their 12 previous games there. No other team in the league has lost so many games in a row at home. The final score was 123-97. The game was played Sunday.

Assessment

- **Sounds odd and cumbersome, although expressed in a coherent discourse**

- **Relatively simple to generate, with limited lexical material**

- **Much lower quality than corpus texts**

GENERATION TECHNIQUES

Two-pass planning process

1. Simple draft sentences with obligatory information

2. Opportunistically adding new facts by applying revision rules

Revision rules

Complement an already included fact

Justify the relevance of a fact by providing its historical background

Some of these rules are non-monotonic ! (reword also the original fact)

Process control

Popping additional facts from a priority list to integrate them in the sentence

Stopping the process when empirically observed complexity limits are reached

INCREMENTAL REVISION (EXAMPLE) (I)

Initial draft

"Dallas, TX – Charles Barkley *scored* 42 points Sunday as the Phoenix Suns defeated the Dallas mavericks 133-97."

Adjunction of Created into Instrument

"Dallas, TX – Charles Barkley *tied a season high with* 42 points Sunday as the Phoenix Suns defeated the Dallas mavericks 133-97."

Coordination Conjoin of Clause

"Dallas, TX – Charles Barkley tied a season high with 42 points and *Danny Ainge added* 21 Sunday as the Phoenix Suns defeated the Dallas mavericks 133-97."

INCREMENTAL REVISION (EXAMPLE) (II)

Absorb of Clause in Clause as Result with Agent Control

"Dallas, TX – Charles Barkley tied a season high with 42 points and *Danny Ainge came off the bench to add* 21 Sunday as the Phoenix Suns defeated the Dallas Mavericks 133-97."

Nominalization with Ordinal Adjoin

"Dallas, TX – Charles Barkley tied a season high with 42 points and Danny Ainge came off the bench to add 21 Sunday as the Phoenix Suns handed the Dallas Mavericks *their 13th straight home defeat* 133-97."

Adjoin of Classifier to NP

"Dallas, TX – Charles Barkley tied a season high with 42 points and Danny Ainge came off the bench to add 21 Sunday as the Phoenix Suns handed the Dallas Mavericks their *league worst* 13th straight home defeat 133-97."

STREAK'S ARCHITECTURE

Game Score Box   Historical Statistics Database

Fact Generator   *Sentence Planner*   *Lexicalizer*

Fixed Facts

First Flat Conceptual Net | First Semantic Tree | First Lexicalized Skeleton Syntactic Tree

First Draft

Stack of Floating Facts   *Sentence Reviser*

Final Draft

Final Flat Conceptual Net | Final Semantic Tree | Final Lexicalized Skeleton Syntactic Tree

SURGE

Natural Language Summary