

自我來黃州已過三寒
食年、欲惜春、意不
容惜今年又苦雨、月社
簫瑟、河海崇崇泥
污遊支雪、閣中偷負
多夜半、真有力何殊、少
年、病起頭、白
春江欲入戶、雨勢未
止、雨、小屋如漁舟、濛
水雲裏、空庭裏寒葉
破、竈燒酒、華、那
知是寒食、但見烏
銜、市、天門深
九重、清夢在、方寸、幾
哭、淪、窮、所、不、吹、不
起

右黃州寒食二首

计算语言学

Computational Linguistics

教师：孙茂松

Tel: 62781286


Email: sms@tsinghua.edu.cn

TA：林衍凯

Email: linyankai423@qq.com

郑重声明

- 此课件仅供选修清华大学计算机系研究生课《计算语言学》(70240052)的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中。未经孙茂松本人同意，任何人不得以任何方式扩散之（包括不得放到任何服务器上）。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。



第一章 引言

(Part 2)

1.4. The Different Levels of Language Analysis



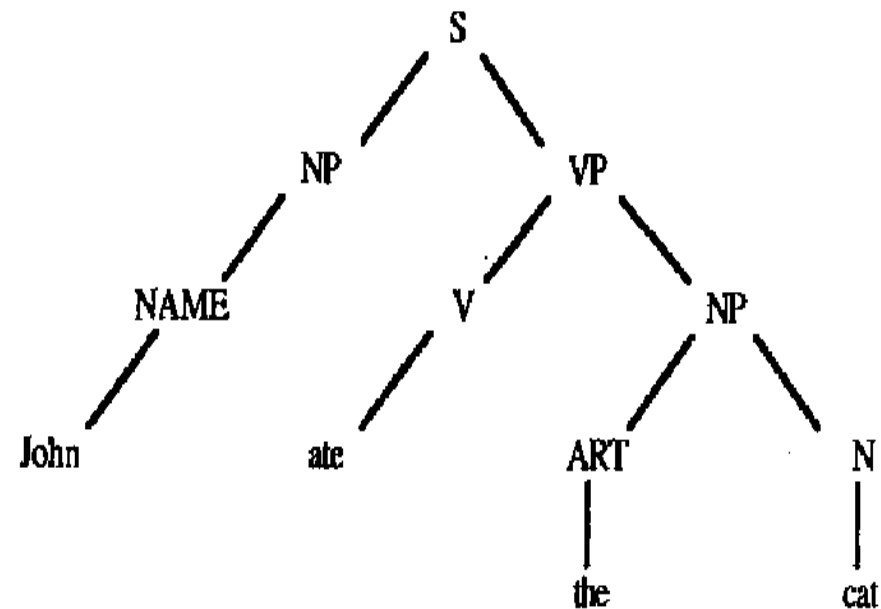
The different forms of knowledge relevant for natural language understanding:

- **Phonetic and phonological knowledge** - concerns how words are related to the sounds that realize them. Such knowledge is crucial for speech-based systems.
- **Morphological knowledge** - concerns how words are constructed from more basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language (for example, the meaning of the word "*friendly*" is derivable from the meaning of the noun "*friend*" and the suffix "*-ly*", which transforms a noun into an adjective).

1.4. The Different Levels of Language Analysis

- **Syntactic knowledge** - concerns how words can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of what other phrases.

Jhon ate the cat.



1.4. The Different Levels of Language Analysis



- **Semantic knowledge** - concerns what words mean and how these meanings -combine in sentences to form sentence meanings. This is the study of context-independent meaning - the meaning a sentence has regardless of the context in which it is used.

Eat (AGENT: John, PATIENT: the cat)

1.4. The Different Levels of Language Analysis

- **Discourse knowledge** - concerns how the immediately preceding sentences affect the interpretation of the next sentence. This information is especially important for interpreting pronouns and for interpreting the temporal aspects of the information conveyed.

Context of sentence

- (1) *Tom went to the bank.*
- (2) *He made an appointment with
Mary there.*
- (3) *Mary was a beautiful girl.*
- (4) *They walked along the river, ...*
- (5) *They fell in love.*

汉语承前省略现象

她弯着腰，
看看田里的水正合适，
不必再从河里车水进来。
又看看她手种的稻子，
 全很壮实，
摸摸稻穗，
 沉甸甸的。
再看看那稻草人，
 帽子依旧戴得很正，
 扇子依旧拿在手里，
 摇动着，
 发出啪啪的声音；
 并且依旧站得很好，
 直挺挺的，
 位置没有动，
 样子也跟以前一模一样。

例子引自宋柔老师指导的
张瑞朋博士论文《现代汉语
书面语中跨标点句句法
关系约束条件的研究》

承前省略构成一种独特的
树结构，这种结构不同于
句法结构，但对于汉语的
理解具有重要的作用

发现了问题，但缺乏足够
规模的语料库支持，无法
开展相关的研究工作

1.4. The Different Levels of Language Analysis



- **Pragmatic knowledge** - concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

Situation

Ticket seller on Beijing bus:

Imperial Palace 1 Yuan!

- **World knowledge** - includes the general knowledge about the structure of the world that language users must have in order to, for example, maintain a conversation. It includes what each language user must know about the other user's beliefs and goals.

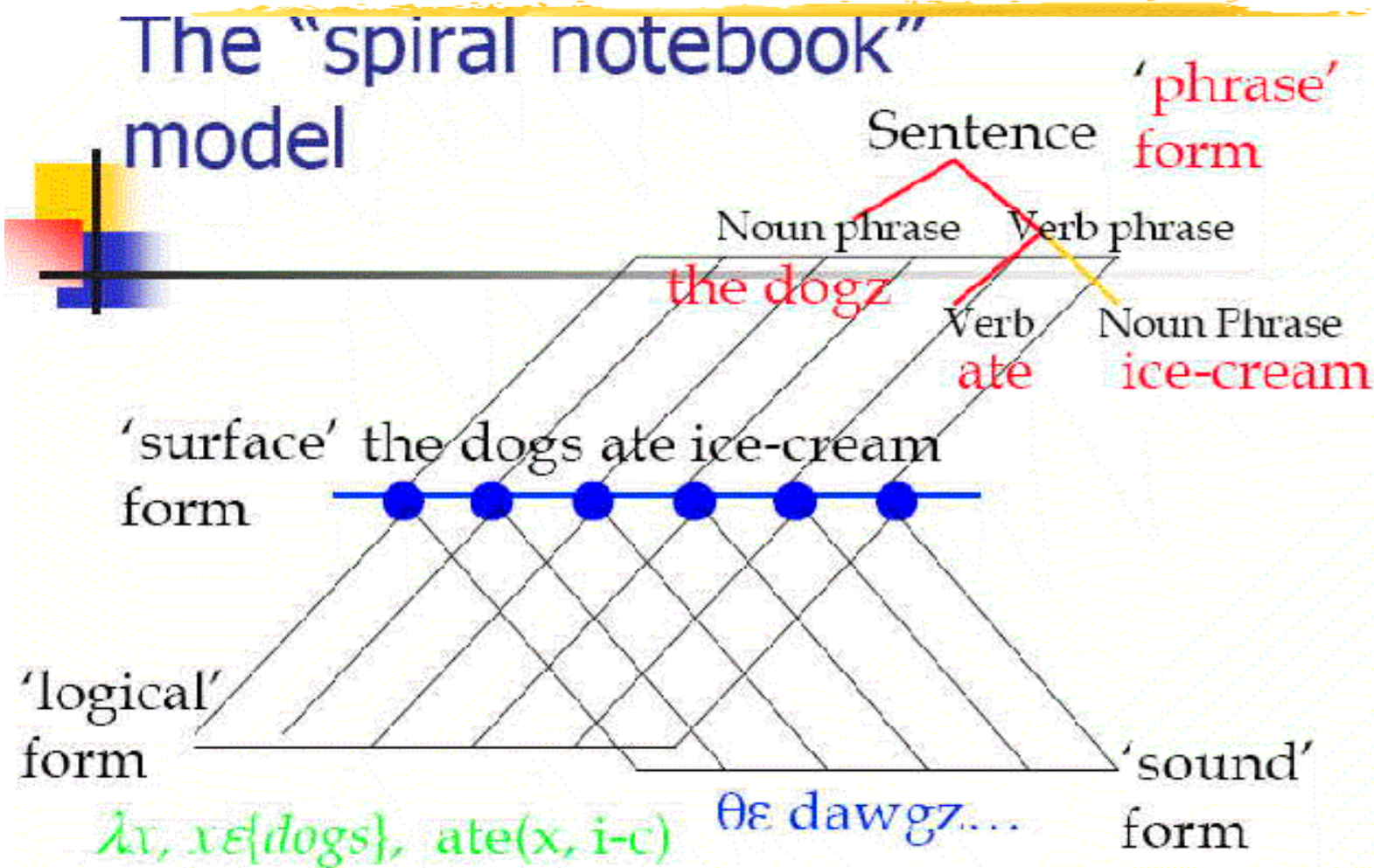
1.4. The Different Levels of Language Analysis



Levels:

- | | |
|---------------------------|----------|
| ● Phonetics and phonology | 语音与音位（学） |
| ● Morphology | 词法（学） |
| ● Syntax | 句法（学） |
| ● Semantics | 语义（学） |
| ● Discourse | 话语 |
| ● Pragmatics | 语用（学） |
| ● World knowledge | 世界知识 |

1.4. The Different Levels of Language Analysis



1.4. The Different Levels of Language Analysis



Syntax, Semantics, and Pragmatics

1	Language is one of the fundamental aspects of human behavior and is a crucial component of our lives	Well-formed syntactically, semantically and pragmatically
2	Green frogs have large noses.	Well-formed syntactically and semantically, but not pragmatically
3	Green ideas have large noses.	Well-formed syntactically, but not semantically and pragmatically
4	Large have green ideas nose.	ill-formed syntactically, semantically and pragmatically

1.4. The Different Levels of Language Analysis

各level之间相对独立，又相互作用

例如：一衣带水 (标准音步)

汉语“词/语”分界的韵律标准：右向音步 (2+1) 是构词音步，左向音步 (1+2) 是造语音步。

*皮工厂 (皮革厂) 读报纸 (*阅读报)

读报 读报纸 *阅读报 阅读报纸

句法：“强信息居后法则” (Behaghel 1909), “尾重原则” (Quirk 1972), “最后的最强” (赵元任 1968)

* 他们正在浇灌花

他们正在浇灌大白菜

1.4. The Different Levels of Language Analysis

XX/XXX

离离/原上草

XX/XX/XXX

且看欲尽花经眼

莫厌伤多酒入唇 (“伤” : 太)

X/XX/XX

XXX/XX

XX/X/XX

XX/XX/X

再如：中文分词

他从中学到了很多知识。

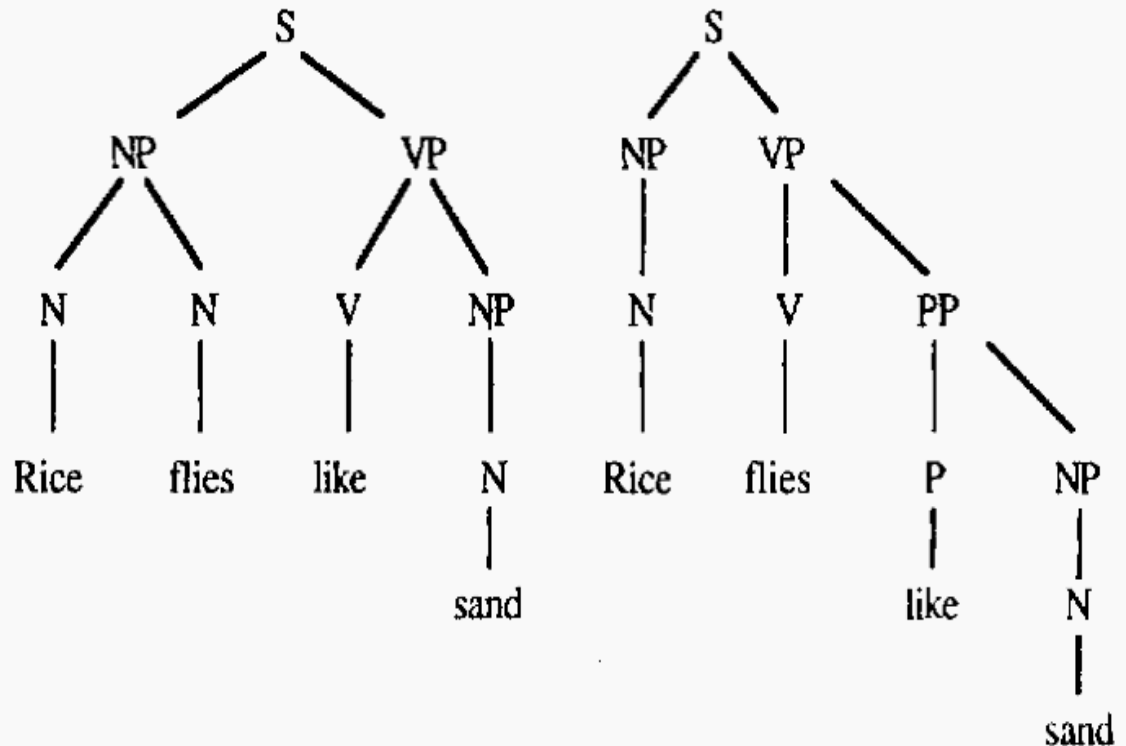
他从中学到大学学了很多知识。

1.5. Representations and Understanding

Representations must be precise and unambiguous!

- Syntax: representing sentence structure

Rice flies like sand



1.5. Representations and Understanding



Flying planes are dangerous.

Flying planes is dangerous.

- The logical form: The first order predicate calculus (FOPC)

John sold the book to Mary.

The book was sold to Mary by John.

Sell (AGENT: John, THEME: book, TO_POS: Mary).

1.5. Representations and Understanding

- The final meaning representation: task-oriented

Exa: an instruction to a robot:

“Come here, please”

FOPC: Come(*AGENT: robot, LOCATION: here*)

The final final meaning representation:

Move(*ACTOR: robot, DESTINATION: ?here*)

To do this:

(1) Speaker x ?

by speech recognition, face recognition, etc.

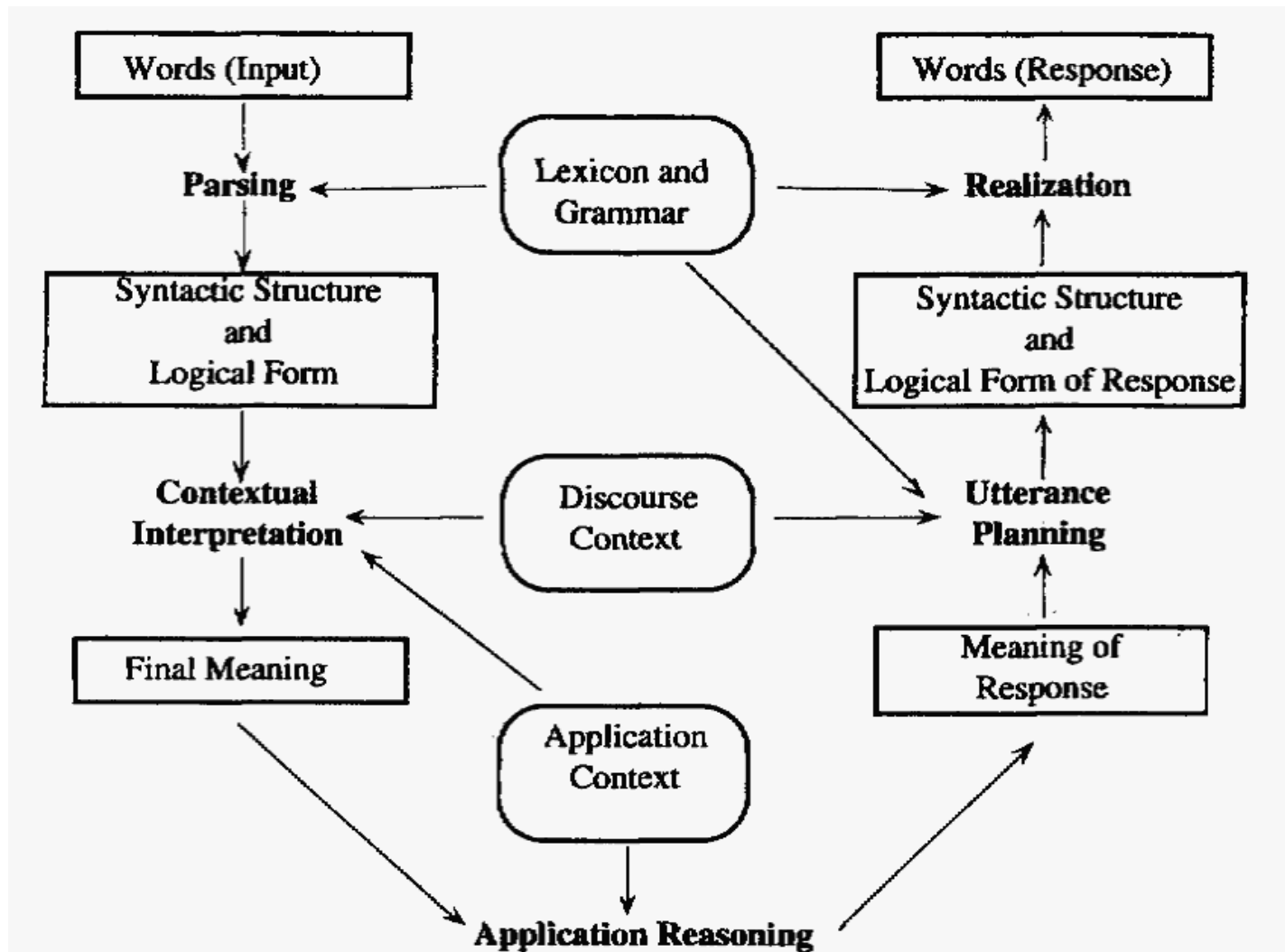
(2) location of x ?

$y \Leftarrow \text{Where-is}(\text{SPEAKER: } x)$

by a number of sensors

(3) MOVE(*ACTOR: robot, DESTINATION: y*)

1.6. The Organization of Natural Language Understanding



1.6. The Organization of Natural Language Understanding

- Analysis (input: texts):

Morphological analyzer ==> parser ==>

Semantic interpreter ==>

discourse and situational analyzer

supported by resources:

lexicon, grammar, semantic constraints,
world knowledge, ...

A variety of internal representations generated:

tokens, parsing tree, FOPC,

The final meaning representations, ...

- Generation(in reverse order of Analysis in general):

Generation instead of Analysis

machine translation, question-answering systems, ...

1.7. History of NLU



- * Weaver Warren, a memorandum, entitled simply “Translation,” which he wrote in July 1949;
co-author (together with Claude Shannon) of the landmark work on communication, The Mathematical Theory of Communication (1949, Urbana: University of Illinois Press).
- * 1950s
 - "MT is rather easy" (MIT, Georgetown)
 - No real NLP research
 - Structuralism reigns in linguistics (Harris)

1.7. History of NLU



The spirit is willing but the flesh is weak.

The vodka is good but the meat is rotten.

“精神上乐意接受的，但肉体上却很虚弱”(马太福音26:41)

Rotten: 腐烂的, 恶臭的, 堕落的, (岩石等)风化的, 虚弱的, 无用的

hydraulic ram (液压油缸)

water *sheep*

Hydraulic 水力的, 水压的; *ram* 公羊, 撞锤, 猛击, 撞

1.7. History of NLU

Conclusion: machine translation is impossible

Bar-Hillel, a mathematician, claimed that machine translation could not be done by mapping one set of strings to another (aka: dictionary lookup).

The pen is in the box. vs. The box is in the pen.

The correct translations require knowledge of:

1. typical sizes of boxes, writing instruments, and enclosures
2. the meaning of "in"
3. spatial reasoning

As a result of disappointing early results, machine translation research and funding essentially stopped!

* 1960s

- "MT is too hard." (ALPAC report)
- CL/NLP research starts (e.g. ACL)
- Transformational paradigm in linguistics (Chomsky)

1.7. History of NLU


* 1970s

- "MT Winter" in US
- Parsing comes of age: CFGs, ATNs, Case-frames,...
- Applications: NL for DB query
- Golden age of American Linguistics: Chomsky, Lakoff, Fillmore, Bresnan/Kaplan, Gazdar, McCawley
- Speech understanding starts

* 1980s

- "MT Spring", first in Japan, then US
- Core NLP research: parsing (TAGs, Joshi; GLR, Tomita,...), generation (TAGs, Whalster; Systemic grammars), and Anaphora/Ellipsis (Grosz, Carbonell,...)
- Applications: Query/Command, MT
- Linguistics: UG, LFG, GPSG, ...
- Speech: Continuous, NLP + Speech

1.7. History of NLU



* 1990s

- Parsing: Spontaneous dialog (e.g. GLR*)
- Proto-Applications boom:
 - + Speech-NL interfaces
 - + MT (KBMT, PC-based, EBMT...)
 - + IR and fact extraction
 - + Speech dictation
- Linguistics: not much new
- Statistical NLP (MT, tagging, ...)

1.7. History of NLU

* 2000s

- MT: Speech-speech MT comes of age
- Corpus-based paradigm booms:
 - + EBMT, statistical MT
 - + Corpus-based statistical parsers (treebank trained)
 - + Corpus-based fact extractors
- Speech recognition: research → commodity
- Linguistics: corpus-based revolution (e.g. Bresnan)
- IR boom: Google, translingual IR, summarization,...
- Return to basic unsolved problems contemplated:
 - + Dialog: beyond "planners"
 - + Definite Reference resolution
 - + Metaphor, Metonymy,...

* 2010s and beyond

阅读作业(不用交)



- (1) 阅读Allen 1995 : Natural Language Understanding
- “Chapter 1: Introduction to Natural Language Understanding”
- “Chapter 2: Linguistic Background: An Outline of English Syntax”