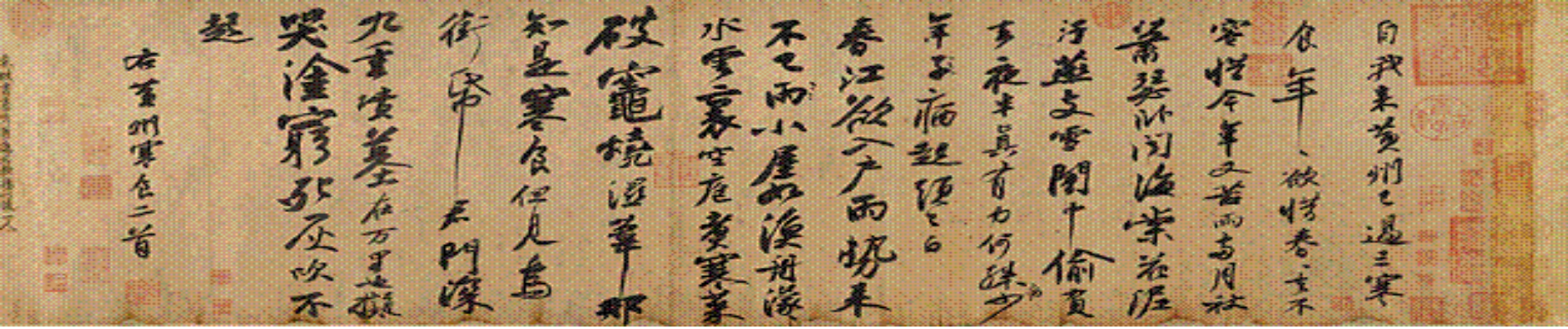


计算语言学
Computational Linguistics

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。

第二章 自然语言的特点及其计算复杂性

2.1. 自然语言的特点

- 结构性
- 无限性（递归性）

This is the cat.

This is the cat that caught the rat.

This is the cat that caught the rat that ate the cheese. ……

- This is the house
- This is the house that Jack built
- This is the grain that lay in the house that Jack built
- This is the rat that ate the grain that lay in the house that Jack built
- This is the cat that killed the rat that ate the grain that lay in the house that Jack built
- This is the dog that chased the cat that killed the rat that ate the grain that lay in the house that Jack built

2.1. 自然语言的特点

Recursive Structures

NP → NP PP The flight to Boston

VP → VP PP departed Miami at noon

Flights to Miami

Flights to Miami from Boston

Flights to Miami from Boston in April

Flights to Miami from Boston in April on Friday

Flights to Miami from Boston in April on Friday under $300.

Flights to Miami from Boston in April on Friday under $300 with lunch.

Conjunctions

S → S and S

NP → NP and NP

VP → VP and VP

2.1. 自然语言的特点

- 歧义性（ambiguity）

Lexical ambiguity

多音字(词)（polyphone）

朝辞白帝彩云间，
千里江陵一日还。
两岸猿声啼不住，
轻舟已过万重山。

Lucent Technologies
Bell Labs Innovations

TTS for Mandarin

2.1. 自然语言的特点

多义词 (polysemy)

同形异义字(词) （homograph）

“Minute”: (1) a unit for measuring time(noun); (2) to make a written record of what is said or decided. during a meeting(verb); (3) tiny(adj)

1a. One minute has sixty seconds.

1b. Part of the job of a secretary is to minute meetings.

1c. There is only minute difference between these pictures.

“编辑”

2.1. 自然语言的特点

Structural ambiguity

亚洲语言学会　　n+n+n　（句法结构歧义）

彩色铅笔盒子　　n+n+n　（句法结构歧义）

关于鲁迅的书　　prep+n+的+n（句法结构歧义）

他讲不清楚。　　v+不+adj　（句法结构歧义）

漂亮的姑娘和小伙子 adj+的+n+的+n（句法结构歧义）

小张的处理意见　　（语义结构歧义）

他在看病。　　（语义结构歧义）

他借我一本书。　　（语义结构歧义）

2.1. 自然语言的特点

中国队打败了。
中国队被打败了。
中国队打败了对手。

热爱人民的总理 v+n+的+n
咬死 猎人的鸡　　咬死 | 猎人的鸡
咬死鸡的 狗　　咬死鸡的 | 狗
咬死 猎人的狗　　咬死 猎人的 | 狗　咬死 | 猎人的狗

2.1. 自然语言的特点

- 统计性（Markov链）

  我爱吃红 _____

A brute force solution. Shannon proposed an interesting scheme to generate text according to a Markov model of order 1.

To construct [an order 1 model] for example, one opens a book at random and selects a letter at random on the page. This letter is recorded. The book is then opened to another page and one reads until this letter is encountered. The succeeding letter is then recorded. Turning to another page this second letter is searched for and the succeeding letter recorded, etc. It would be interesting if further approximations could be constructed, but the labor involved becomes enormous at the next stage.

2.1. 自然语言的特点

- 模糊性　　“下半旗”
- 文化因素

“lying on top of a bed in English and Chinese”:

“in bed” (English) and “在床上” (Chinese)

In Chinese, a bed means the bed frame but certainly not the duvet nor the blanket; whereas in the English sense of bed, pillows and duvets are often considered as part of the bed (that is why the English phrase “to make the bed” means to arrange the sheets and covers neatly on the bed instead of really constructing the bed from wood and/or metal). In this sense, if the preposition meaning “in” in Chinese is used instead, the meaning of the phrase will be distorted. For instance, if “John is in bed.” is translated to Chinese word by word, the resulting Chinese sentence will sound odd unless the bed is actually like a box where John hides himself in.

2.2. Complexity of Natural Languages

Grammar: $G = (V, T, S, P)$

V : Set of (meta) symbols, or variables

T : Set of terminal symbols

S : Start symbols, from V

P : Set of Production rules

2.2. Complexity of Natural Languages

$$G \qquad S \to aSb$$

$$S \to \lambda$$

$$G = (V, T, S, P)$$

$$V = \{S\} \qquad T = \{a, b, \lambda\}$$

$$P = \{S \to aSb, \quad S \to \lambda\}$$

2.2. Complexity of Natural Languages

Sentential Form:

A sentence that contains variables and terminals

Example:

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbb$$

Sentential Forms

sentence

2.2. Complexity of Natural Languages

We write: $S \overset{*}{\Rightarrow} aaabbb$ Derivations

Instead of:

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbb$$

In general we write: $w_1 \overset{*}{\Rightarrow} w_n$

If: $w_1 \Rightarrow w_2 \Rightarrow w_3 \Rightarrow \cdots \Rightarrow w_n$

2.2. Complexity of Natural Languages

Language of a Grammar

For a grammar G
with start variable S :

$$L(G) = \{w: \quad S \overset{*}{\Rightarrow} w\}$$

String of terminals

2.2. Complexity of Natural Languages

- **Chomsky hierarchy**

The Chomsky hierarchy is an ordering of types of grammar according to generality. The classification in fact only depends on the type of grammar rule (rewrite rule) used. The grammar types include:

unrestricted grammars(type 0): rules of the form $\alpha \rightarrow \beta$ with no restrictions on the sequence of symbols α and β .

context sensitive grammars(type 1): rules of the form $\alpha X \beta \rightarrow \alpha \Psi \beta$ where X is a non-terminal symbol, α and β are (possibly empty) sequences of symbols, and Ψ is nonempty sequence of symbols.

2.2. Complexity of Natural Languages

context free grammars(type 2): rules of the form $X \rightarrow \alpha$ where X is a single non-terminal symbol, α are nonempty sequence of symbols.

(right) regular grammars(type 3): rules of the form $X \rightarrow a$ and $X \rightarrow aN$ where X and N are nonterminal symbols, and a is a terminal symbol.
(left) regular grammars(type 3): rules of the form $X \rightarrow a$ and $X \rightarrow Na$ where X and N are nonterminal symbols, and a is a terminal symbol.

The Chomsky Hierarchy

- **unrestricted** or **type-0** grammars, generate the *recursively enumerable* languages, automata equals *Turing machines*

- **context-sensitive** grammars, generate the *context-sensitive* languages, automata equals *Linear Bounded Automata*

- **context-free** grammars, generate the *context-free* languages, automata equals *Pushdown Automata*

- **regular** grammars, generate the *regular* languages, automata equals *Finite-State Automata*

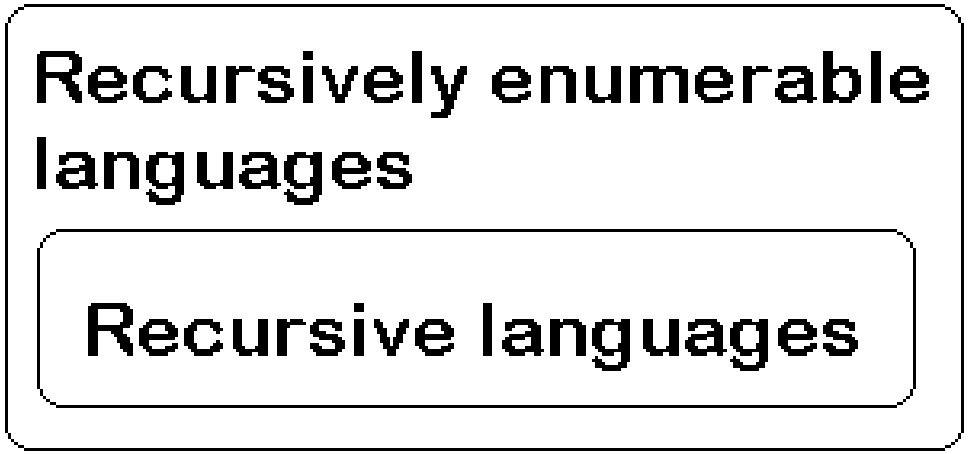
2.2. Complexity of Natural Languages

* A language is *recursively enumerable* if there exists a Turing machine that accepts every string of the language, and does not accept strings that are not in the language.

"Does not accept" is *not* the same as "reject" -- the Turing machine could go into an infinite loop instead, and never get around to either accepting *or* rejecting the string.
The languages generated by unrestricted grammars are precisely the recursively enumerable languages.

* A language is *recursive* if there exists a Turing machine that accepts every string of the language and rejects every string (over the same alphabet) that is not in the language.

2.2. Complexity of Natural Languages



Decidable language (definition)
Definition: A language for which membership can be decided by an *algorithm* that halts on all inputs in a finite number of steps --- equivalently, can be recognized by a *Turing machine* that halts for all inputs.

2.2. Complexity of Natural Languages

- **Generative capacity of grammars**

Any grammar G that is a type n (> 0) grammar is also a type n-1 grammar.

Any language that is a type n (> 0) language is also a type n-1 language.

2.2. Complexity of Natural Languages

L0: (right) regular grammar

S -> a S1 S -> d S1 -> d S3 -> d

S -> b S2 S1 -> b S2 S2 -> c S3

S -> c S3 S1 -> c S3 S2 -> d

L1: $L_1 = \{a^n b^n\}, n \geq 1$

ab, aabb, aaabbb,……

L2: $L_2 = \{\alpha\alpha^*\}$

aa, bb, abba, aaaa,bbbb, aabbaa, abbbba,… 镜象语言

L3: $L_3 = \{\alpha\alpha\}$

aa, bb, abab,aaaa, bbbb, aabaab, abbabb, …

2.2. Complexity of Natural Languages

L1不能用RG生成，可用CFG生成:

S -> a b　　S -> a S b

ab, aabb, aaabbb,……

L2:不能用RG生成，可用CFG生成:

S -> a S a　S -> b S b

S -> a a　　S -> b b

aa, bb, abba, aaaa,bbbb, aabbaa, abbbba,… 镜象语言

L3不能用CFG生成，可用CSG生成:

S -> a S　　S -> b S

$\alpha S \rightarrow \alpha\alpha$

α是集合{a,b}上的任意非空符号串

aa, bb, abab,aaaa, bbbb, aabaab, abbabb, …

2.2. Complexity of Natural Languages

- 自然语言不能用RG完全生成

The rat disappeared.

a a

The rat the cat caught disappeared.

a b b a

The rat the cat the dog chased caught disappeared.

a b c c b a

L2

2.2. Complexity of Natural Languages

- Consider the following set of English sentences (strings)
  - S = If S_1 then S_2
  - S = Either S_3, or S_4
  - S = The man who said S_5 is arriving today
- Map *If, then* $\rightarrow a$ and *either, or* $\rightarrow b$. This results in strings like *abba* or *abaaba* or *abbaabba*

2.2. Complexity of Natural Languages

- 自然语言不能用CFG完全生成

- (Shieber, 1985) and (Huybregts, 1984) showed this using examples from Swiss-German:

mer	d'chind	em Hans	es huus	lönd	hälfed	aastriiche
we	the children-ACC	Hans-DAT	the house-ACC	let	helped	paint
w	a	b	x	c	d	y
	N_1	N_2	N_3	V_1	V_2	V_3

... we let the children help Hans paint the house

2.2. Complexity of Natural Languages

P. Postal（1964）发现，印第安的Mohawk语中:

“我读书”

我书读书

a a

“我喜欢读书”

我书读书喜欢书读书

b a b b a b

我尝到了读书的甜头

我书读书的甜头尝到了书读书的甜头

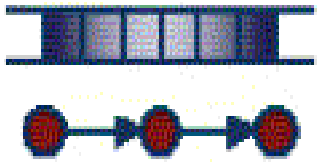
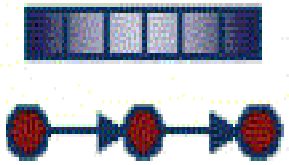
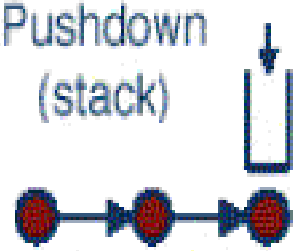
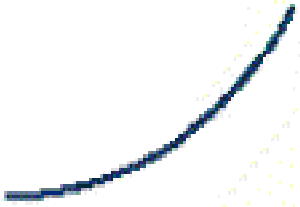
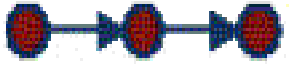
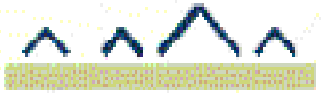
b a b c d b a b c d

L3

2.2. Complexity of Natural Languages

大姐、二姐 、三姐分别是二十 、十八和十六岁。

a b c a b c

Language	*Automaton*	*Grammar*	*Recognition*	*Dependency*
Recursively Enumerable Languages	Turing Machine 	Unrestricted $Baa \rightarrow A$	Undecidable	Arbitrary
Context-Sensitive Languages	Linear-Bounded 	Context-Sensitive $At \rightarrow aA$	NP-Complete 	Crossing 
Context-Free Languages	Pushdown (stack) 	Context-Free $S \rightarrow gSc$	Polynomial 	Nested 
Regular Languages	Finite-State Machine 	Regular $A \rightarrow cA$	Linear 	Strictly Local 

2.2. Complexity of Natural Languages

Natural language 属于CSG, 接近于CFG

CFGs are very **important** because:

* powerful enough to describe most of the structure in natural languages;
* restricted enough so that efficient parsers can be built to analyze sentences.

2.2. Complexity of Natural Languages

- **Chomsky 范式**

任何上下文无关语言都能由那样的文法产生，其中所有规则的形式或者是U → XY 或者是U → T, 这里X,Y,U属于VN, T属于VT.

- 上下文无关语言的可判定性

- 文法的二义性问题是不可判定的(上下文无关文法)
  寻找充分条件

- DFA vs. NFA

2.2. Complexity of Natural Languages

Noam Chomsky
Institute Professor; Professor of Linguistics
Linguistic Theory, Syntax, Semantics,
Philosophy of Language, MIT
http://web.mit.edu/linguistics/www/chomsky.home.html

(December 7, 1928)

The Chomsky hierarchy is a containment hierarchy of classes of formal grammars that generate formal languages. This hierarchy was described by Noam Chomsky in 1956.

2.2. Complexity of Natural Languages

Chomsky has written and lectured widely on linguistics, philosophy, intellectual history, contemporary issues, international affairs and U.S. foreign policy. His works include: Aspects of the Theory of Syntax; Sound Pattern of English (with Morris Halle); Language and Mind; American Power and the New Mandarins; At War with Asia; For Reasons of State; Peace in the Middle East?; Reflections on Language; The Political Economy of Human Rights, Vol. I and II (with E.S. Herman); Rules and Representations; Lectures on Government and Binding; Towards a New Cold War; Radical Priorities; Fateful Triangle; Knowledge of Language; Turning the Tide; Pirates and Emperors; On Power and Ideology; Language and Problems of Knowledge; The Culture of Terrorism; Manufacturing Consent (with E.S. Herman); Necessary Illusions; Deterring Democracy; Year 501; Rethinking Camelot: JFK, the Vietnam War and US Political Culture; Letters from Lexington; World Orders, Old and New; The Minimalist Program; Powers and Prospects; The Common Good; Profit Over People; The New Military Humanism; New Horizons in the Study of Language and Mind; Rogue States; A New Generation Draws the Line; 9-11; and Understanding Power.

2.2. Complexity of Natural Languages

乔姆斯基“言语获得装置”(language acquisition device)：认为儿童的大脑里有一种天生的“言语获得装置”。这是人类头脑中固有的内在的语法规则。儿童运用这种普遍语法，就很容易掌握这种语言。

1871年，达尔文首先提出语言是一种本能的理论。“牙牙学语”…

2005年,英国的《展望》（Prospect）和美国的《外交政策》（Foreign Policy）两本杂志联合进行了一次跨大西洋两岸的读者投票，以期选出全球最著名的公众知识分子。共两万余名读者填写了选票，最后生成了一份百人大榜。 乔姆斯基位列头名。

目前人文领域被引次数最高的十位作家之一。超过黑格尔，紧跟马克思、列宁、莎士比亚、《圣经》、亚里士多德、柏拉图和弗洛伊德之后，唯一在世

2.2. Complexity of Natural Languages

2001年“9·11”事件发生，当月，他的《9·11》一书便告上市。此后在美国主流媒体上，很难再见到他，主流知识分子——无论左右，往往也与他这样的好斗者保持距离。他的声音更多是通过校园演讲、“油印”小报、海外报刊、互联网，以及小出版社的出版物达于外界。他在MIT（乔氏从1955年起任教于此）的讲座，从来都是人满为患，其场面之热烈，堪与校园内的摇滚音乐会媲美。讲话时声调不高，更像是在做学术报告，而非发表反战演说，但极其雄辩，锋芒毕露又不失机巧