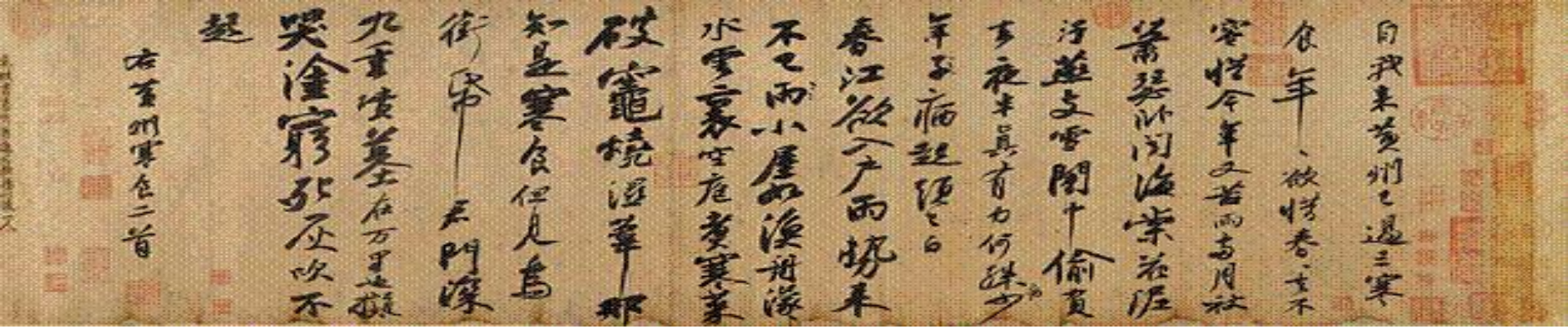


计算语言学
Computational Linguistics

教师：孙茂松

Tel:62781286

Email:sms@tsinghua.edu.cn

TA:李文浩

Email:wh-li20@mails.tsinghua.edu.cn

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（Part 2）

2.2. Complexity of Natural Languages

L0: (right) regular grammar

$S \rightarrow a\ S_1$ $S \rightarrow d$ $S_1 \rightarrow d$ $S_3 \rightarrow d$

$S \rightarrow b\ S_2$ $S_1 \rightarrow b\ S_2$ $S_2 \rightarrow c\ S_3$

$S \rightarrow c\ S_3$ $S_1 \rightarrow c\ S_3$ $S_2 \rightarrow d$

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:

$S \rightarrow a\ b \qquad S \rightarrow a\ S\ b$

ab, aabb, aaabbb,……

L2:不能用RG生成，可用CFG生成:

$S \rightarrow a\ S\ a \quad S \rightarrow b\ S\ b \quad S \rightarrow a\ a \qquad S \rightarrow b\ b$

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

L3不能用CFG生成，可用CSG生成:

$S \text{ -> } a\ S \qquad S \text{ -> } b\ S$

$aS \rightarrow a\ a$

a是集合$\{a,b\}$上的任意非空符号

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:

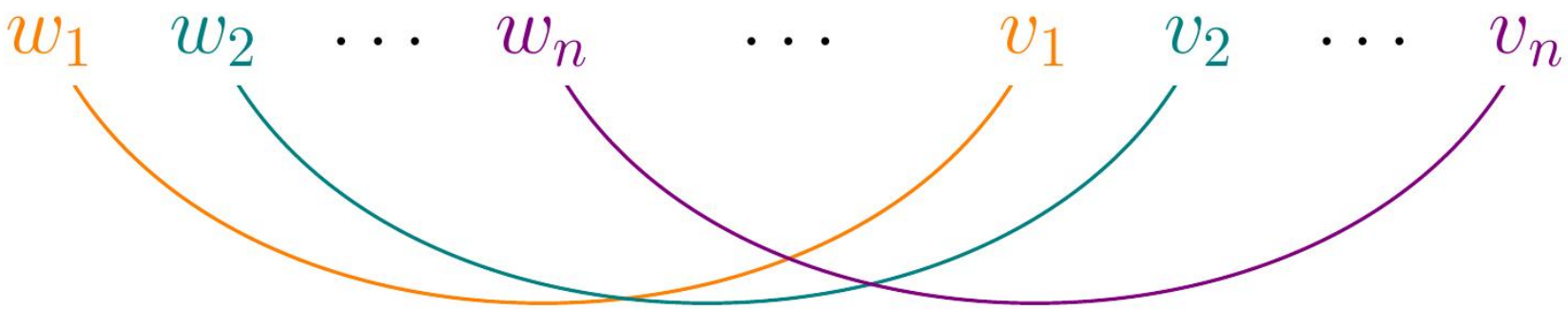
mer	d'chind	em Hans	es huus	lönd	hälfed	aastriiche
we	the children-ACC	Hans-DAT	the house-ACC	let	helped	paint
w	a	b	x	c	d	y
	N_1	N_2	N_3	V_1	V_2	V_3

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

L3

2.2. Complexity of Natural Languages

Cross-serial dependencies



*Shieber, Stuart (1985), "Evidence against the context-freeness of natural language", Linguistics and Philosophy, **8** (3): 333–343.*
http://www.eecs.harvard.edu/~shieber/Biblio/Papers/shieber85.pdf

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

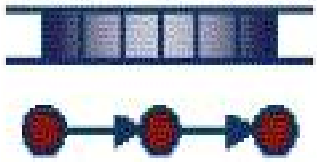
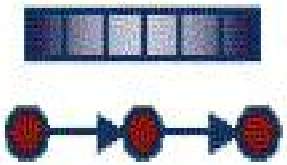
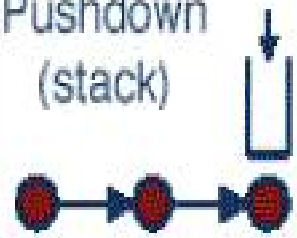
我书读书的甜头尝到了书读书的甜头

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 → A	Undecidable	Arbitrary
Context-Sensitive Languages	Linear-Bounded 	Context-Sensitive At → aA	NP-Complete 	Crossing 
Context-Free Languages	Pushdown (stack) 	Context-Free S → gSc	Polynomial 	Nested 
Regular Languages	Finite-State Machine 	Regular A → 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

推荐延展阅读图书（自愿阅读）

《语言本能：人类语言进化的奥秘》[The Language Instinct：How the Mind Creates Language] 当代伟大思想家、世界语言学家和认知心理学家之作 [美] 史蒂芬·平克（Steven Pinker）著

入选《美国科学家》（American Scientist）评出的20世纪100本合适的科学书籍，十分深刻、精彩!

与乔姆斯基的观点相呼应