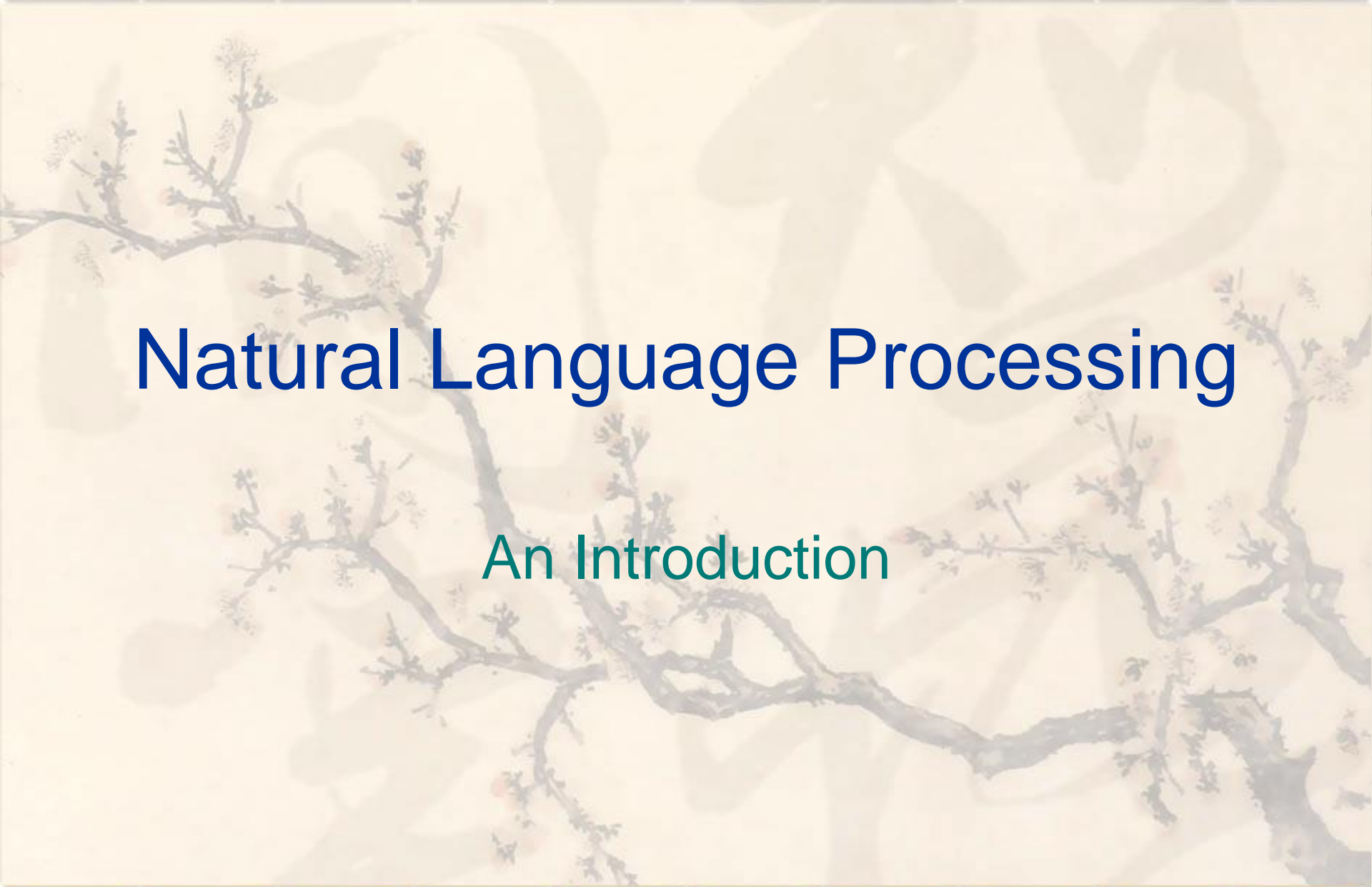



自然语言处理

杨沐昀

教育部-微软语言语音重点实验室

MOE-MS Joint Key Lab of NLP and Speech (HIT)



Natural Language Processing

An Introduction

Introduction

- ❖ A Preview of NLP
- ❖ Language: Linguistic Perspective
- ❖ NLP: Starting Line

A Preview of NLP

❖ ACL definition:

“The scientific study of **language** from a computational perspective.

“Interested in providing **computational models** of various kinds of linguistic phenomena.”

一门以计算机为手段，通过建立语言现象的计算模型对自然语言进行研究和处理的学科。

Eliza: Weizenbaum (1966)

- ❖ User: You are like my father in some ways.
- ❖ ELIZA: What resemblance do you see
- ❖ User: You are not very aggressive but I think you don't want me to notice that.
- ❖ ELIZA: What makes you think I am not aggressive
- ❖ User: You don't argue with me
- ❖ ELIZA: Why do you think I don't argue with you
- ❖ User: You are afraid of me
- ❖ ELIZA: Does it please you to believe I'm afraid of you

Language

❖ A Linguistic Perspective

- ❧ What is language ?

- ❧ What do we know about language?

一个无解的疑问：语言起源

- ❖ 1、神授说：认为语言是上帝或神赐予人类
- ❖ ①印度：婆罗门教《吠陀》说语言是神赐给人的一种特殊的能力
- ❖ ②基督教《圣经》说耶和华创造了亚当，又由亚当给世间万物起了名字
- ❖ ③苗族：山神创造了人，并创造了语言

一个无解的疑问：语言起源

- ❖ 2、人创说：认为语言是人自己创造的，而不是上帝或神赐予的。
- ❖ ①摹声说：语言起源于人类对外界各种声音的摹仿。
- ❖ ②社会契约说：通过彼此约定，规定了事物的名称，因此产生了语言。
- ❖ ③手势说：在人类使用有声语言之前，经历了一个手势语言的阶段，人们用手势来表达思想，进行交际。
- ❖ ④感叹说：认为人类的有声语言是从抒发情感的各种叫喊演变来的。
- ❖ ⑤劳动叫喊说：认为人类的有声语言是从人们劳动时的叫喊声演变来的。注意到了语言的起源和劳动的关系；但无法解释劳动号子如何演变为语言。

一个无解的疑问：语言起源

- ❖ 恩格斯提出了劳动创造了语言，语言起源于劳动的观点。
- ❖ 普遍认为：人类有声语言的产生大约是在距今四五万年前的旧石器时代晚期，也就是晚期智人时期。
- ❖ 目前：学术会议不接受语言起源的论文！

Linguistic Perspective

❖ 语言和言语

∞ 言语：指说话这种行为和说出来的具体的话

❖ a. 具有个人特点，丰富多彩（嗓音、用词等）。

❖ b. 说话所用的词语和规则是全社会共有的、是语言的具体应用

∞ 语言：是从言语中概括出来的各言语要素的综合，是约定俗成的体系，有统一的语法规则和语音习惯，具有社会性。

❖ 两者关系：是个别和一般的关系。言语是对语言的具体运用，没有语言也就没有言语；另一方面，语言也不能脱离言语，语言存在与言语之中，而言语是语言的存在形式。

Language: Linguistic Perspective

语言的符号性

符号、能指、所指

用甲事物代表乙事物，而甲乙两事物之间没有必然的联系，甲事物就是代表乙事物的符号。甲事物就是符号的能指（形式），乙事物就是符号的所指（内容、意义）。

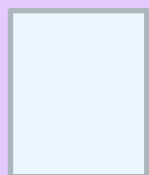
语言符号的主要特征

任意性

稳定性

渐变性

线性



关于语言符号

❖ 语言符号的性质：语言符号是音义结合的统一体。

- ∞ 语言是由语音和意义两个方面统一构成，语音是语言的物质外壳，是语言的存在形式；意义是语言的内容。
- ∞ 语音和意义在具体的语言中统一于一体的，密不可分，二者互为存在条件
- ∞ 语言符号的音义结合是社会约定俗成的。

- 语言的其他属性：
- 1. 民族性（洪堡特）
- 2. 生成性（乔姆斯基）
- 3. 模糊性（莱考夫）

语言系统

- 1.组合关系
- 若干较小的语言单位组合成较大的语言单位，其构成成分之间的关系就是组合关系，又称线性序列关系。

老师	分析	课文
小王	讲	故事
李明	写	文章
我	看	电视
他	学	英语



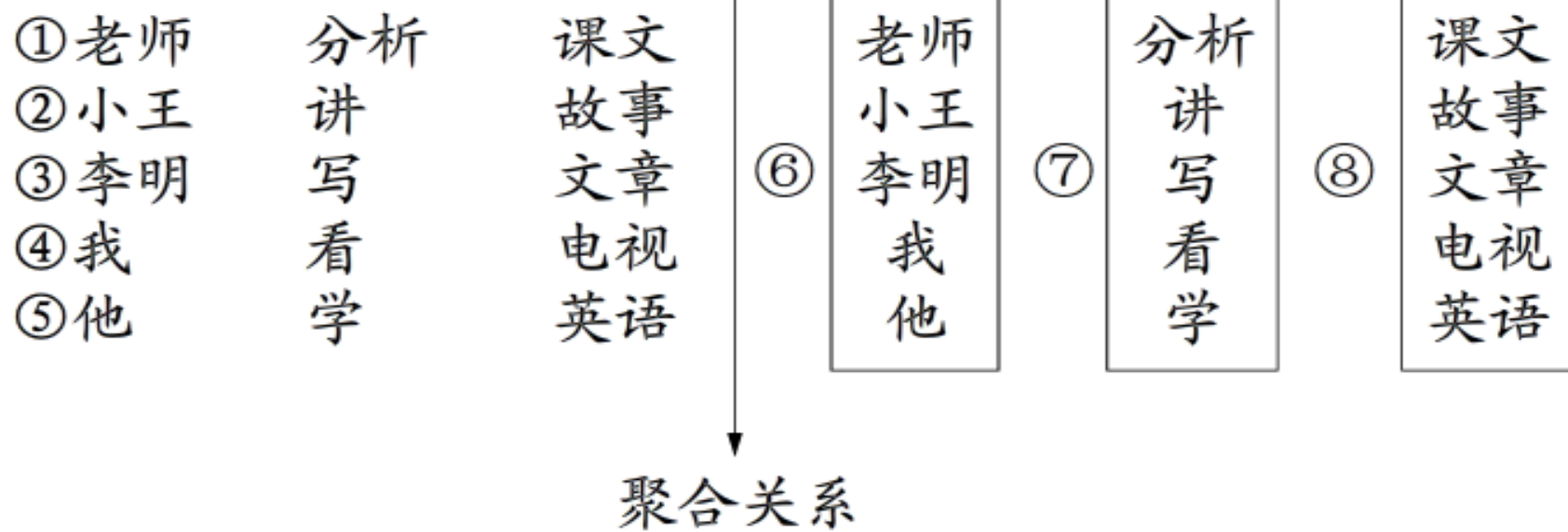
语言系统

2. 聚合关系

❖ 具有相同组合功能的语言单位之间的关系，就是聚合关系，又称联想关系。聚合关系专指那些具有替换关系的语言单位之间的关系。

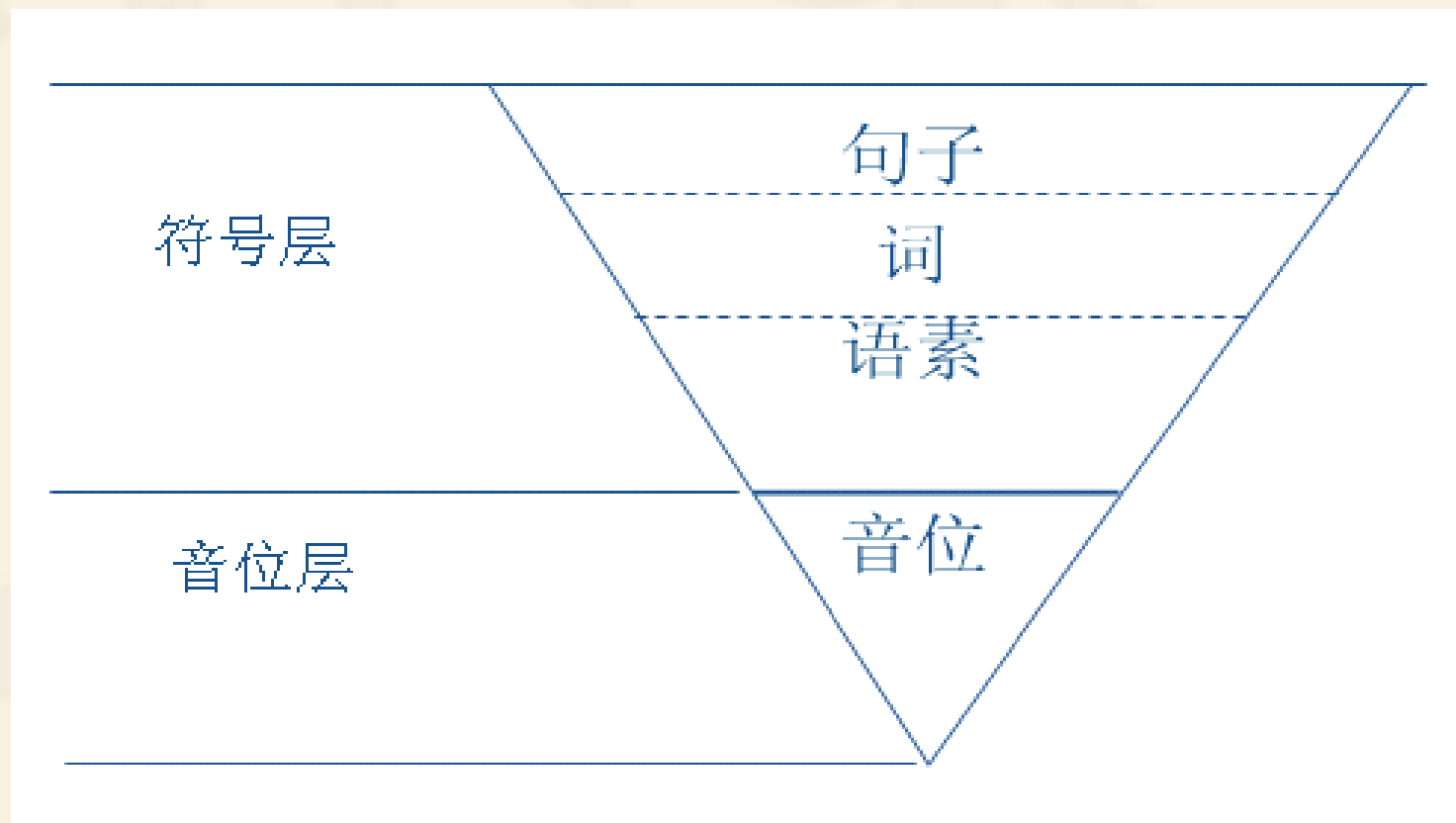
	老师	分析	课文
	小王	讲	故事
	李明	写	文章
	我	看	电视
	他	学	英语

组合关系



❖ 语言的组合关系和聚合关系是语言的两种根本关系，是语言系统的纲，把握了这个纲，就基本上把握了语言系统。

语言系统的层级体系



语言的社会性：交际工具

❖ 朱元璋做了皇帝，他从前相交的一班苦朋友照旧过着很穷的日子。有一天，他从前的一個苦朋友跑到南京求见，准见之后便说：

“我主万岁！当年微臣随驾扫荡芦州府，打破罐州城，汤元帅在逃，拿住豆将军，红孩儿当关，多亏菜将军。”

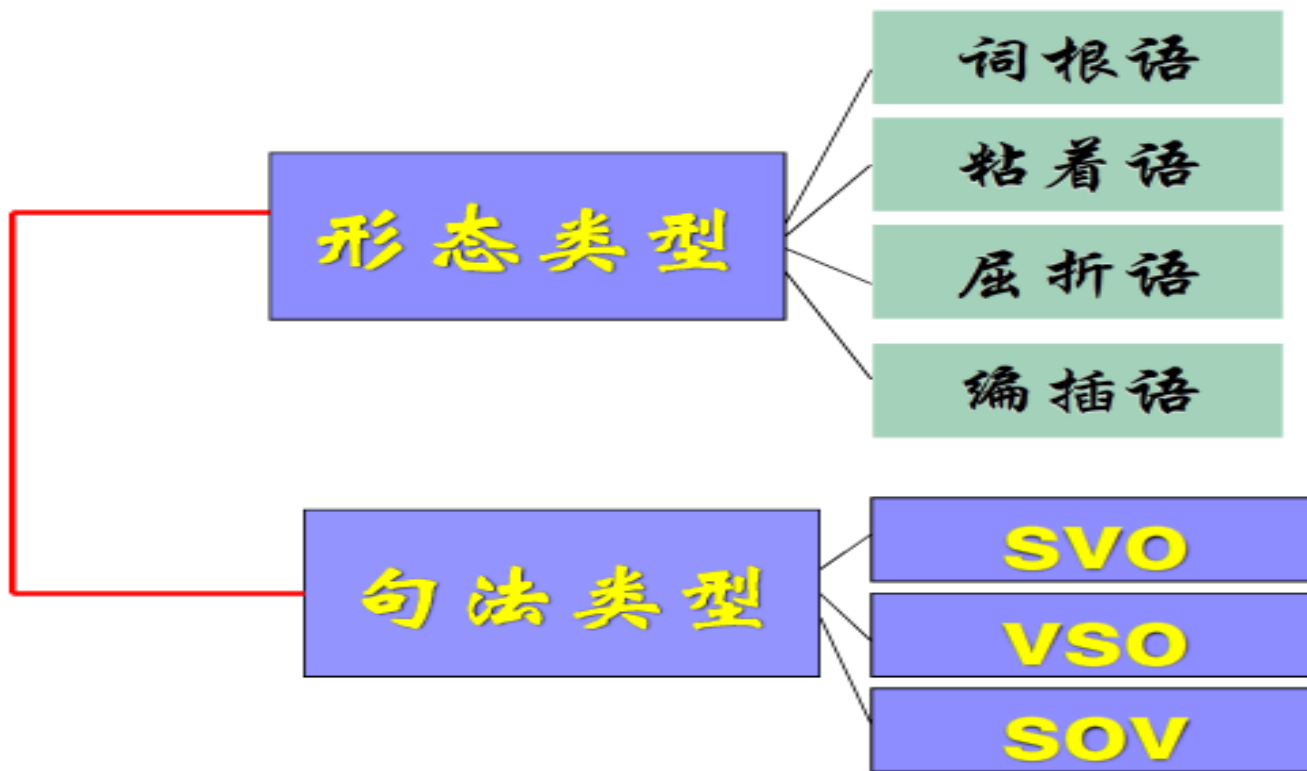
朱元璋一听，隐约觉得他的话中包含了一些从前的往事，见他说得好听，心里很高兴，所以立刻封他做了御林军的总管。这个消息让另外一个苦朋友听见了，他心想：“同是那时候一起玩的人，他去了有官做，我去当然也不会倒霉吧？”

“我主万岁！还记得吗？从前，你我都替人家看牛，有一天，我们在芦花荡里，把偷来的豆子放在瓦罐里煮着，还没等煮熟，大家就抢着吃，把罐子都打破了，撒下一地的豆子，汤都泼在泥地里。你只顾从地上满把地抓豆子吃，却不小心连红草叶子也送进嘴去。叶子梗在喉咙口，害得你哭笑不得。还是我出的主意，叫你用青菜叶子放在手上一拍吞下去，才把红草叶子带下肚子里去了。……”

朱元璋等不得听完就连声大叫：“推出去斩了！推出去斩了！”

语言的分类

语言的结构类型



形态类型

- 根据语言中形态变化是否丰富，以及形态变化的不同方式，一般将人类的语言划分位四种类型：
- 1、词根语（汉语、越南语、彝语、苗语、缅甸语等）
- 2、屈折语（印欧语系各语言、阿拉伯语等，德语、俄语）
- 3、粘着语（土耳其语、哈萨克、芬兰语、匈牙利、日语、朝鲜语、维吾尔语、蒙古语等）
- 4、编插语（美洲的各种印第安语、爱斯基摩人的一些语言以及古亚细亚语系的楚克奇语等。）

句法类型

1、SVO型语言

英语、法语、俄语、汉语、傣语、苗语、瑶语等就属于这种类型。如“他吃了饭。”

2、SOV型语言

日语、拉丁语、土耳其语、蒙语、藏语、彝语等

3、VOS型语言

阿拉伯语、威尔斯语、古诺尔都语等。

文字

❖ 一、文字是语言的书写符号系统

❧ 文字起源于图画

❖ 二、文字的类型

❧ 根据字符跟语言单位的语义还是语音相联系，文字分为表意文字、表音文字和意音文字。

(1) 表意文字：全部字符都是意符的文字。

(2) 表音文字：全部字符都是音符的文字。也叫拼音文字。如阿拉伯文字、希腊文字。

(3) 意音文字：一部分字符是意符，一部分字符是音符的文字。

文字

❖ 汉字的类型问题

❧ (1) 汉字是一种词语文字。

❧ (2) 汉字是一种意音文字。

❧ (3) 汉字是一种语素文字。

❧ (4) 汉字！=一种表意文字或象形文字。

❖ 表意文字是不能存在的。汉字也已经不象形了。

语法

- ❖ 盖一座大楼，如果砖瓦是词汇的话，语法就是把它们黏合在一起的规则。
- ❖ 说话要遵守该社会的语言规则



感知汉语语法特点

❖ 语序和虚词是汉语的重要语法手段

- ❧ 创作小说——小说创作
- ❧ 资本主义国家——国家资本主义
- ❧ 一会儿再谈——再谈一会儿
- ❧ 你今天能来吗——你能今天来吗
- ❧ 跑快——快跑
- ❧ 我看了书——我看着书——我看过书
- ❧ 我和老板——我的老板

感知汉语语法特点

❖ 句法同义现象。表达形式上的灵活性。

❧ 一杯水他喝了。

❧ 他喝了一杯水。

❧ 他把一杯水喝了。

❧ 一杯水被他喝了。

❧ 他一杯水喝了。

感知汉语语法特点

- ❖ 诗词中的超语法现象是汉语中一种独特的语言现象
不求有形，但凭心意，注重联想，以达意为目的。
- ❖ 楼船夜雪瓜洲渡，铁马秋风大散关。
——陆游《书愤》
- ❖ 枯藤老树昏鸭，小桥流水人家，古道西风瘦马。
——马致远《秋思》

语法和语法规则

❖ 一、什么是语法？

❖ 语法就是用词造句的规则，这种规则是客观存在于一种语言之中，是语言长期发展过程中形成的，说这种语言的全体成员必须共同遵守。

❖ 二、语法规则

❖ 语法规则是大家说话的时候必须遵守的习惯，不是语言学家规定的。语法的组合规则和聚合规则构成一种语言的语法规则。

语法单位

❖ 1、句子

- ❖ 句子是语言中最大的语法单位，又是交际中基本的表述单位。从形式上看，句子的最大特点是有一个完整的语调。
- ❖ 句子按其语气可以分为陈述、疑问、祈使、感叹等不同的类型，简称句型。一般来说，陈述句、祈使句和感叹句的语调在句末是下降的，而疑问句的语调则是上升的。

语法单位

❖ 2、词组

- ❖ 词组是词的组合，它是句子里面作用相当于词而本身又是由词组成的大于词的单位。
- ❖ 词组有自由词组和固定词组两种。固定词组中的成分一般不能更换、增删，次序不能颠倒，如成语和民间口头流传的词语。

语法单位

❖ 3、词

- ❖ 词是最重要的一级语法单位，它是造句的时候能够独立运用的最小单位。所谓独立运用，就是它在造句中能够到处作为一个单位出现；所谓最小，就是说不能分割和扩展，也就是说中间不能插入别的成分。从意义和作用看，词可以分为实词和虚词两大类。
- ❖ Part-of-speech: 名、动、形、副…
- ❖ 语法研究通常以词为界，词以上的规则叫句法，词以下的规则叫做词法。

语法单位

❖ 4、语素

- ❖ 语素是语言中音义结合的最小单位。就汉语来说，大抵一个汉字就是一个语素，但是也有两个字表示一个语素的，如：“咖啡”“玻璃”“葡萄”等。词由语素构成，有的词由一个语素构成，如火、山、人、咖啡等，有的由两个语素构成，如朋友、铁路等。
- ❖ 我们可以根据语素在词中的不同作用把它分成词根、词缀、词尾三类。
- ❖ 词根是词义的核心部分，词的意义主要是由它体现出来，它可以单独构成词，也可以彼此组合成词。如：桌、椅、水、电、英语read happy。汉语中绝大多数的词都是由词根构成的。

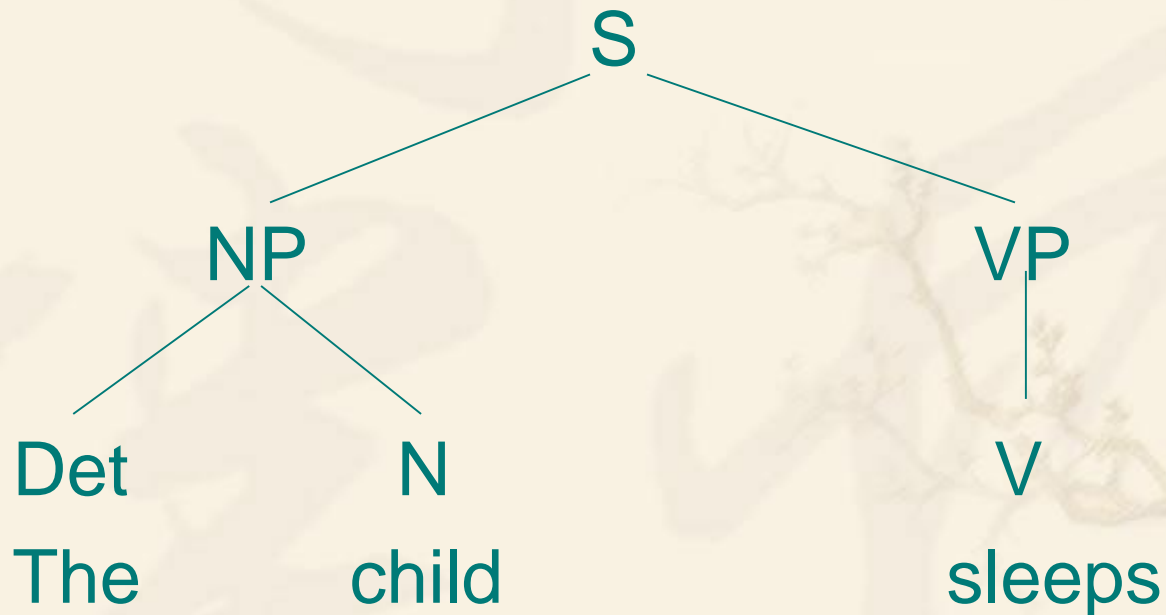


句法分析

- ❖ 目的：描述、解释、学习语言构造规律
- ❖ 成分分析法：主、谓、宾、定、状、补
- ❖ 层次分析法：对句法单位（包括短语和句子）的直接成分进行结构层次分析的方法，基本采用二分法
- ❖ 变换分析法：通过移位、添加、删除、替换等方法来考察两种句法结构之间的关系和变换规则的分析方法
 - ❧ 着眼于句法结构的外部分析，考察具有内在联系的不同句法结构之间的联系。

Representation

- ❖ tree structures used to represent sentences
- ❖ sentences broken down into constituents



语言学：研究语言的学问

phonetics (sounds in language)

- ❖ where sounds are made

phonology (sound patterns in language)

- ❖ what sound combinations are allowed

morphology (words and word structures)

- ❖ rules for combining morphemes

syntax (sentence and sentence structures)

- ❖ rules for coming words in a sentence

semantics (meaning)

Why Computer is applied to Language?

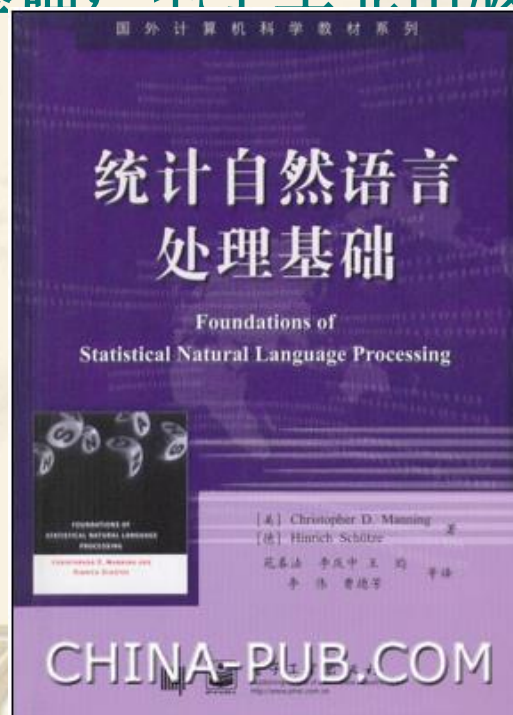
One reason for studying language - and for me personally the most compelling reason - is that it is tempting to regard language, in the traditional phrase, as a “mirror of mind”.



Chomsky, 1975

参考书目

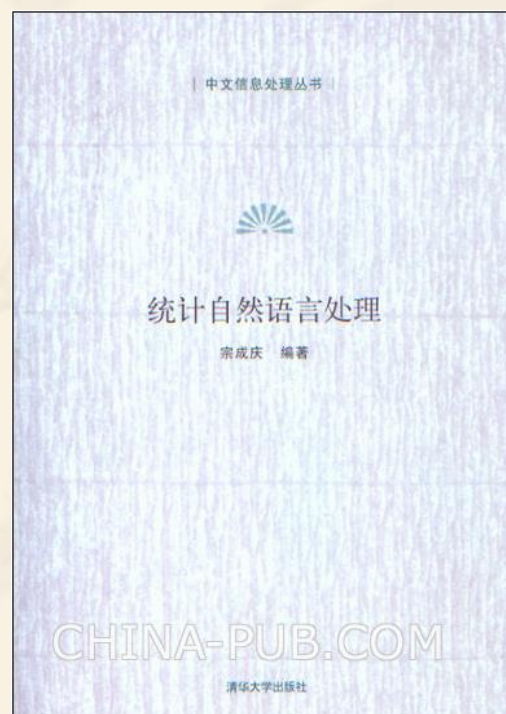
- ❖ Christopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999
- ❖ 苑春法等译，统计自然语言处理基础，电子工业出版社，2005.1



- ❖ Daniel Jurafsky, James H.Martin, **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**, Prentice Hall Press, 2000
- ❖ 冯志伟, 孙乐译, **自然语言处理综论**, 电子工业出版社, 2005.6



❖ 宗成庆，统计自然语言处理，清华大学出版社，2008.5



阅读论文——大部分在网上

- ❖ Proceedings of major conferences:
 - ❧ ACL (Assoc. of Computational Linguistics)
 - ❧ COLING (Intl. Committee of Computational Linguistics)
 - ❧ EACL (European Chapter of ACL)
 - ❧ ANLP (Applied NLP)
- ❖ But I suggest you read more papers, esp. classical papers.

课程 考评

考核环节	主选比例	考核/评价细则	辅选比例
芝麻开门	10分	1) 根据课堂回答问题的次数, 考量参与课堂讨论的程度, 每人每次记1分; 2) 每人获得分数上限为10分;	平时成绩10%
项目实践	40%	1. 两项项目开发各占15%; 2) 提交三份报告, 分别获得3%, 3% 和 4% 成绩; 参照主流学术论文的内容要素完整度评价 3) 每人自带10分, 最后根据贡献自动调配, 不能相同;	0
期末考试	60%	卷面考试	90%
课程最终成绩 = (1) + (2) + (3)			

课程群 (qq) : 658240901

第一讲课后作业（选作）

- ❖ 调研汉字编码方案
- ❖ 调研汉字输入法
- ❖ 追加思考题：为什么要讨论语言学对语言各种定义和概念？
- ❖ {结合AI中的问题求解}