




Natural Language Processing

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)



序

❖ N元文法的巨大成功

- ❧ NLP任务：分词、词性标注、未登录词识别
- ❧ Speech任务：语音识别、语音合成
- ❧ 基于人工专家知识的规则系统不香了

❖ 更为深刻的问题

- ❧ MM&HMM 是随机过程分析的技术
- ❧ 语言本质上是概率分布吗？

Language

❖ A Linguistic Perspective

Q1 What is language ?(宏观)

Q2 What do we know about language?(微观)

∞ 语言的起源：目前无解

∞ 语言的研究：从古至今

∞ 语言的性质 (整体视角)

∞ 语言存在形式 (结构视角)

语言起源：目前无解

- 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

语言起源：目前无解

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。

语言起源：目前无解

- ❖ 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！

语言研究：从古至今

❖ 语言：古老的研究对象

- ❧ 中国：汉朝时产生了小学，包括文字、音韵和训诂

- ❧ 印度和希腊：4~3 B.C.，出现了语法学。

❖ 现代的语言学起源：18世纪初期

- ❧ 西方研究者开始梳理对比各种语言；

- ❧ 发现了印欧语言与梵语之间存在类似之处

- ❧ 产生了历史比较语言学：旨在寻找各种语言的原始语

语言研究：从古至今

❖ 现代语言学之父

❧ 费尔迪南·德·索绪尔

❖ Ferdinand de Saussure

❖ 1857-1913



❖ 把语言学塑造成为一门影响巨大的独立学科

❖ 他认为语言是基于符号及意义的一门科学

❧ 符号的研究：现在一般通称为符号学

❧ 从1907年始讲授“普通语言学”课程

语言研究：从古至今

❖ 语言学奠基的基础：系统性

❧ 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：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。

语言的性质

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）。

语言的性质

❖ 一般符号只要赋予意义即可：如交通信号

❖ 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。

语言的性质

- ❖ **共时性:** 就是研究语言在某个特定时期表现出的特点以及内在联系;
- ❖ **历时性:** 就是研究语言在整个历史长河中的变化,与其它时代语言特点的异同。
- ❖ **索绪尔观点:** 语言学研究应该是共时的。
 - ☞ 语言学的研究应该去除时间因素的干扰,语言学家要描述某个词的价值,并不需要求助于词源。换句话说,语言学是描述性的,应当描述当下语言系统中各个要素之间的关系;而不是站在历史的角度去研究系统中要素的变迁。
 - ☞ ~tip: 从他所处的历史实际理解他的观点

语言的性质

❖ 语言的社会性：地域、场合、身份都影响语言使用

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！”

语言的性质

❖ 语言的一些主要性质

∞ 任意性(arbitrary): 常说的“约定俗成”

∞ 稳定性: 短期、局部

∞ 渐变性: 长期、全局

∞ 线性: 书写、口述、理解, 都有先后过程

∞ 传承性和交际性: 人类文化得以传承和储存的有效载体

∞

语言存在形式

❖ 两种根本关系

- 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



语言存在形式

❖ 两种根本关系

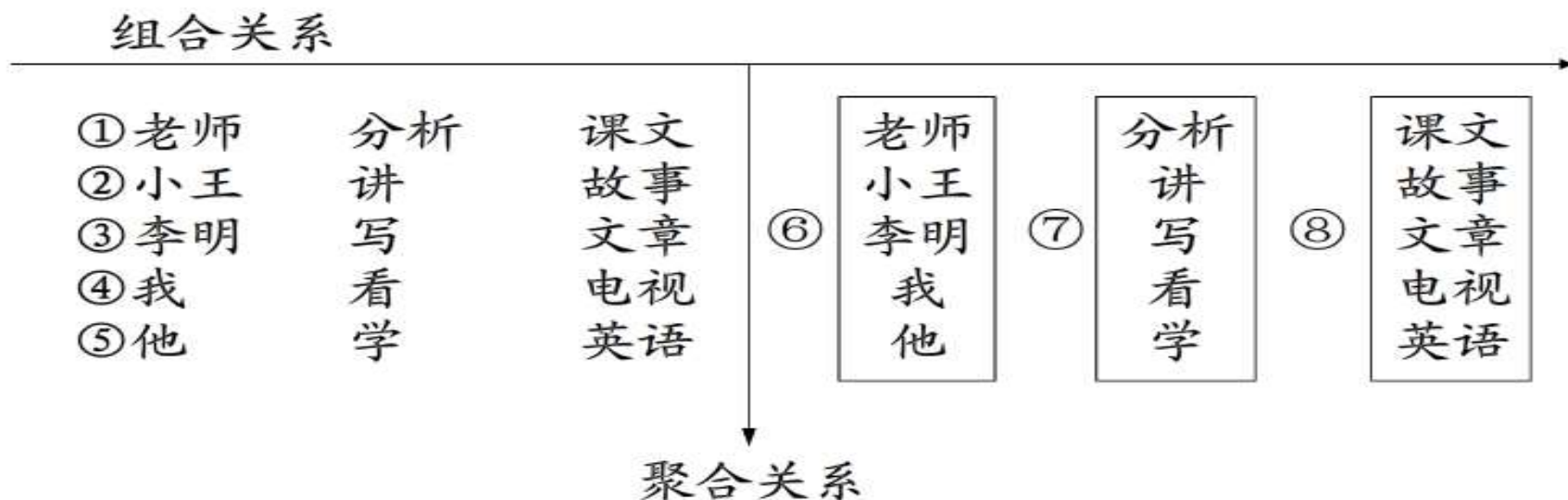
2. 聚合关系

- ❖ 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

语言存在形式

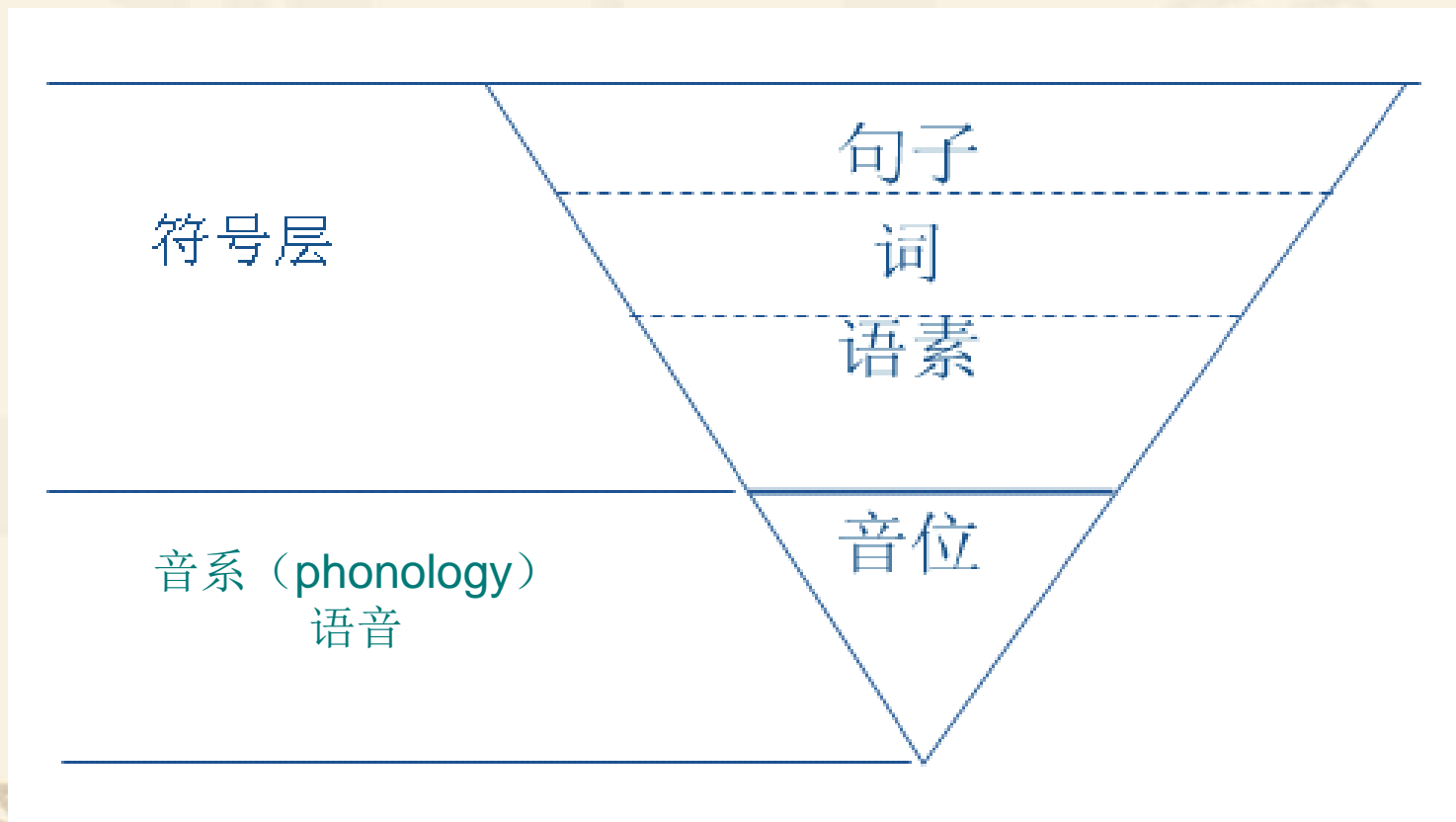
❖ 两种根本关系



❖ 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。

语言存在形式

❖ 层级体系：分层分粒度



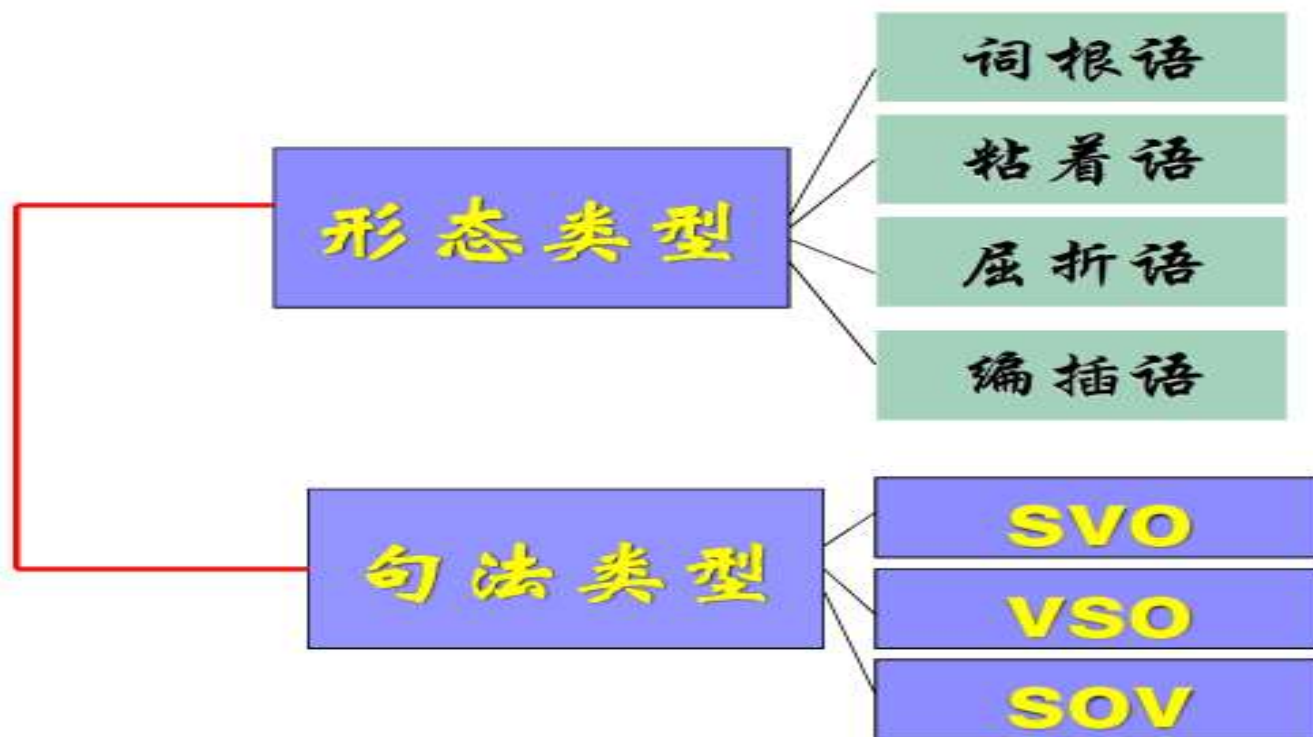
语言存在形式

- ❖ 语言的分类体系：语系语族语支语种
 - ❖ 语言间的近亲属关系
 - ❖ 印欧语系、日耳曼语族、英语
 - ❖ 汉藏语系、汉语族、汉语
 - ❖ 日语、韩语：（阿尔泰？）

语言存在形式

❖ 语言的分类体系：结构

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。

Language

❖ A Linguistic Perspective

Q1 What is language ?

Q2 What do we know about language? (微观)

∞ 语言学研究分支

∞ 文字

∞ 语法

∞

语言学：研究语言的学问

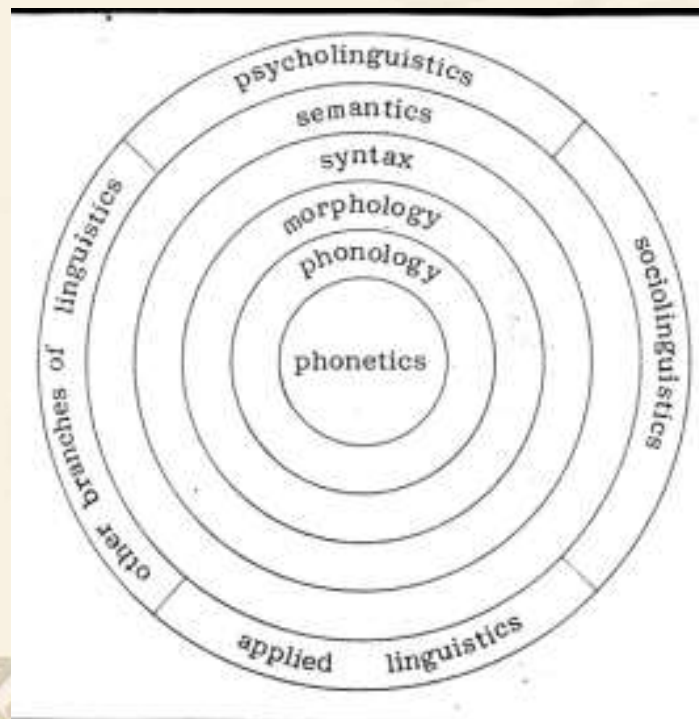
研究分支：Micro/ Intra -Linguistics:

- ❖ 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)

语言学：研究语言的学问

研究分支：Macro/ Inter -Linguistics:

- ❖ 社会语言学（sociolinguistics）
- ❖ 心理语言学（psycholinguistics）
- ❖ 认知语言学（cognitive linguistics）
- ❖ 计算语言学
- ❖ 神经语言学
- ❖ 数理语言学



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。

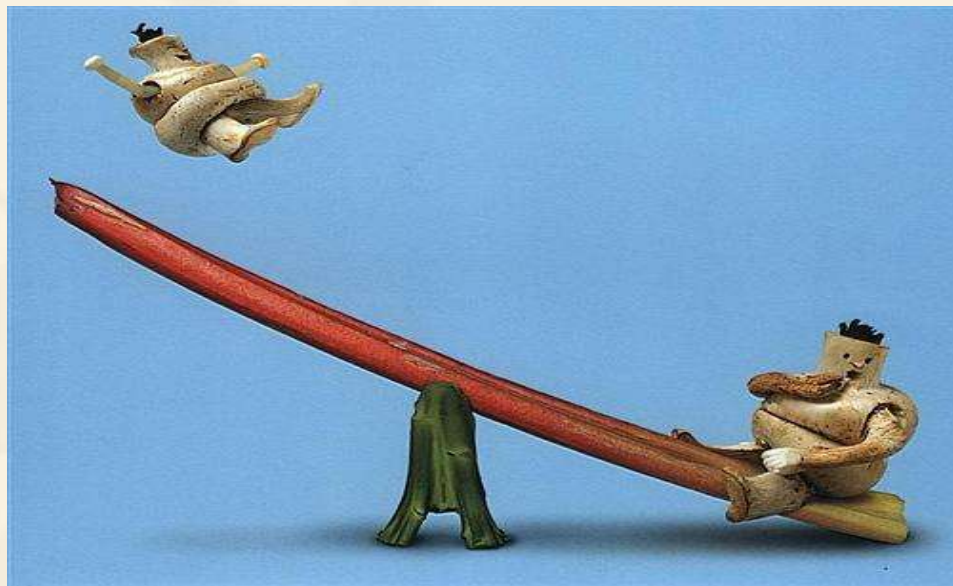
文字

❖ 汉字的类型问题

- ❧ (1) 汉字是一种意音文字。
- ❧ (2) 汉字是一种语素文字。
- ❧ (3) 汉字是一种词语文字。
- ❧ (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

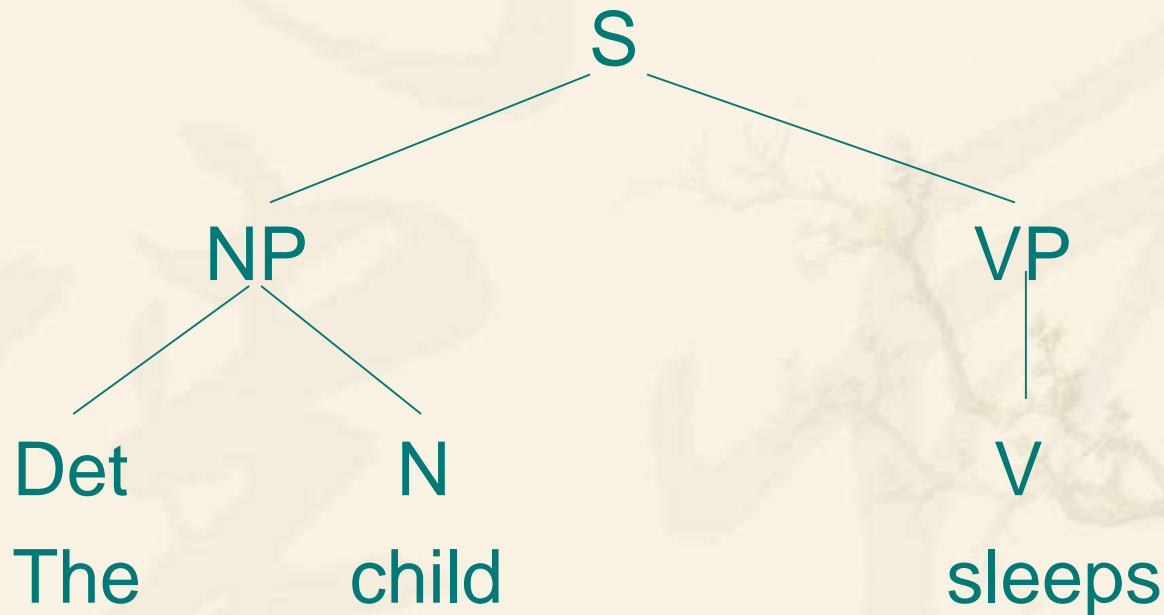
- ❖ 语素是语言中音义结合的最小单位。就汉语来说，大抵一个汉字就是一个语素，但是也有两个字表示一个语素的，如：“咖啡”“玻璃”“葡萄”等。词由语素构成，有的词由一个语素构成，如火、山、人、咖啡等，有的由两个语素构成，如朋友、铁路等。
- ❖ 我们可以根据语素在词中的不同作用把它分成词根、词缀、词尾三类。
- ❖ 词根是词义的核心部分，词的意义主要是由它体现出来，它可以单独构成词，也可以彼此组合成词。如：桌、椅、水、电、英语。汉语中绝大多数的词都是由词根构成的。

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



结语： NLP vs CL

❖ 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。

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

第一讲课后

❖ 阅读：

❧ Manning书：第3章；

❧ 宗成庆书：第一章；

❖ 思考题：

❖ 1. 为什么要讨论语言学对语言各种定义和概念？

❧ {提示：AI中的问题求解需要什么}

❖ 2. 语言的符号性和社会性决定了语言的客观性和主观性，这种说法对吗，谈谈你对这个问题的认识？