# 第 2 章　语料库与语言知识库

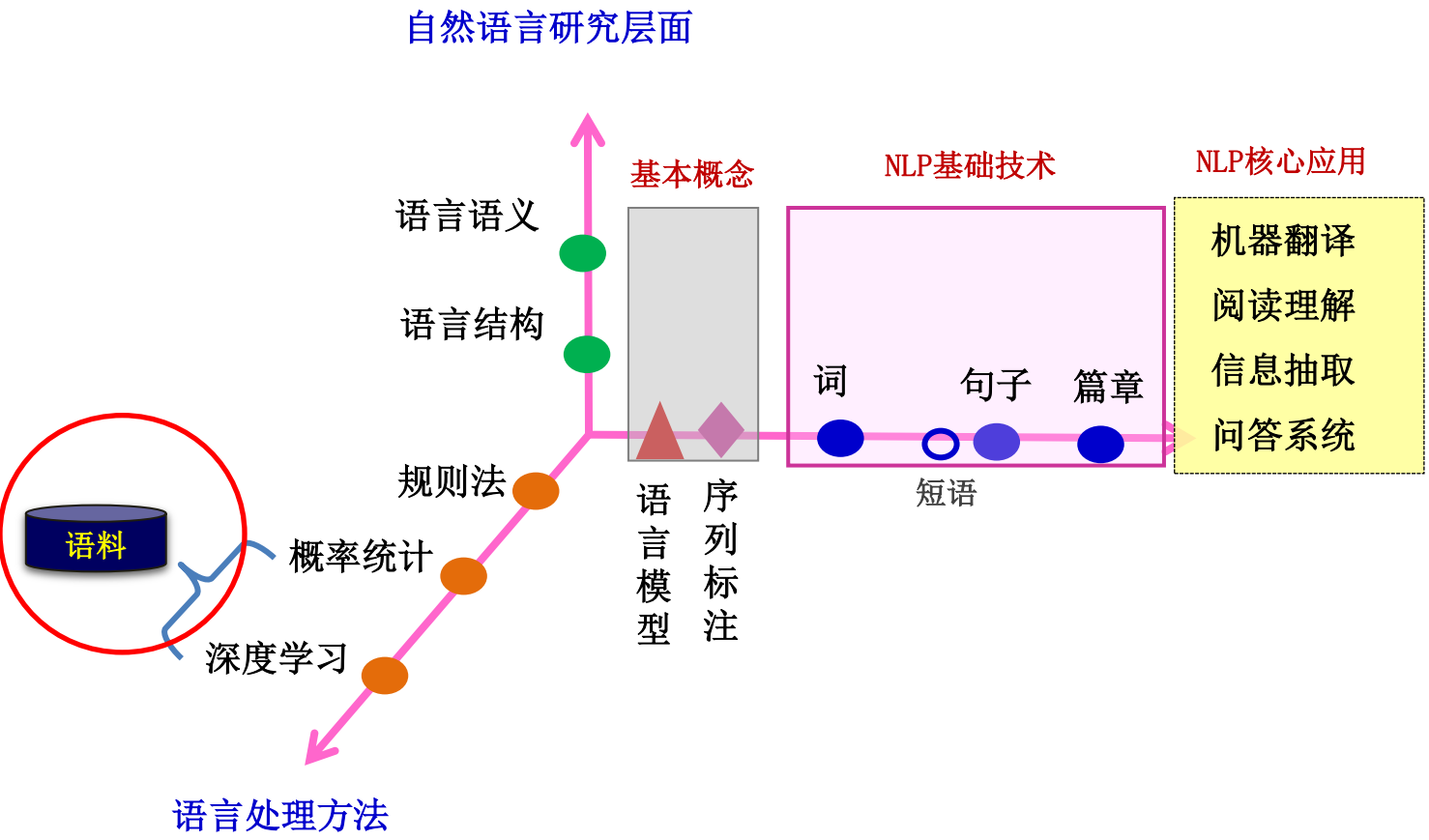
中科院信息工程研究所第二研究室

胡玥

huyue@iie.ac.cn

# 自然语言处理课程内容及安排

✧ **课程内容：**



# 内 容 提 要

2.1 语料库概述

2.2 语料库技术的发展

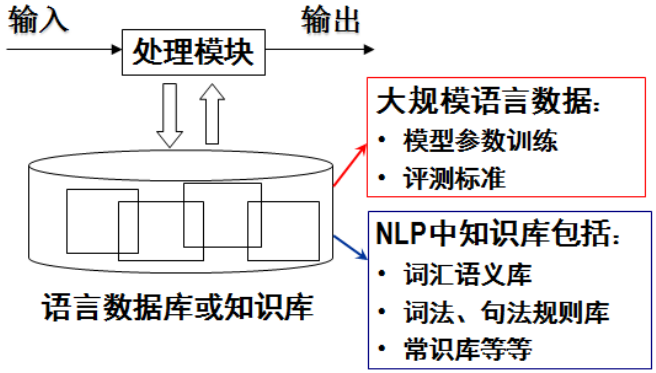
2.3 语料库资源

2.4 语料的收集与加工

2.5 语言知识库

# 2.1 语料库概述

**1. 语料库(corpus)**：存放语言材料的仓库现代的语料库是指存放在计算机里的原始语料文本 或 经过加工后带有语言学信息标注的语料文本。




以语言的真实材料为基础来呈现语言知识，反映语言单位的用法和意义，基本以知识的原始形态表现——语言的原貌

# 2.1 语料库概述

## 语料库有三点特征

- 语料库中存放的是在实际使用中真实出现过的语言材料；
- 语料库是以计算机为载体承载语言知识的基础资源，但并不等于语言知识；；
- 真实语料需要经过分析、处理和加工，才能成为有用的资源。

## 语料库的作用

- 支持语言学研究和语言教学研究
- 支持NLP系统的开发

# 2.1 语料库概述

**例1：** 北京大学计算语言所富士通人民日报标注语料库样例：

- 历史/n 将/d 铭记/v 这个/r 坐标/n ：/w 北纬/b ４１．１/m 度/q 、/w 东经/b １１４．３/m 度/q ；/w
- 人们/n 将/d 铭 记/v 这/r 一/m 时刻/n ：/w １９９８年/t １月/t １０日/t １１时/t ５０分/t 。/w
- [中国/ns 政府/n]nt 顺利/ad 恢复/v 对/p 香港/ns 行使/v 主权/n ，/w 并/c 按照/p “/w 一国两制/j ”/w 、/w “/w 港人治港/l ”/w 、/w 高度/d 自治/v 的/u 方针/n 保持/v 香港/ns 的/u 繁荣/an 稳定/an 。/w

# 2.1 语料库概述

**例1：** 北京大学计算语言所语料库标记：

| 代码 | 名称 | 帮助记忆的诠释 |
|---|---|---|
| Ag | 形语素 | 形容词性语素。形容词代码为a，语素代码 g 前面置以A。 |
| a | 形容词 | 取英语形容词adjective的第1个字母。 |
| ad | 副形词 | 直接作状语的形容词。形容词代码a和副词代码d并 |
| an | 名形词 | 具有名词功能的形容词。形容词代码a和名词代码 |
| b | 区别词 | 取汉字“别”的声母。 |
| c | 连词 | 取英语连词conjunction的第1个字母。 |
| Dg | 副语素 | 副词性语素。副词代码为d，语素代码 g 前面置以 |
| d | 副词 | 取adverb的第2个字母，因其第1个字母已用于形容 |
| e | 叹词 | 取英语叹词exclamation的第1个字母。 |
| f | 方位词 | 取汉字“方” 的声母。 |
| g | 语素 | 绝大多数语素都能作为合成词的“词根”，取汉字“ |
| h | 前接成分 | 取英语head的第1个字母。 |
| i | 成语 | 取英语成语idiom的第1个字母。 |
| j | 简称略语 | 取汉字“简”的声母。 |
| k | 后接成分 | |
| l | 习用语 | 习用语尚未成为成语，有点“临时性”，取“临”的声 |
| m | 数词 | 取英语numeral的第3个字母，n，u已有他用。 |
| Ng | 名语素 | 名词性语素。名词代码为n，语素代码 g 前面置以 |
| n | 名词 | 取英语名词noun的第1个字母。 |
| nr | 人名 | 名词代码n和“人(ren)”的声母并在一起。 |
| ns | 地名 | 名词代码n和处所词代码s并在一起。 |
| nt | 机构团体 | “团”的声母为t，名词代码n和t并在一起。 |
| nz | 其他专名 | “专”的声母的第1个字母为z，名词代码n和z并在一起。 |
| o | 拟声词 | 取英语拟声词onomatopoeia的第1个字母。 |
| p | 介词 | 取英语介词prepositional的第1个字母。 |
| q | 量词 | 取英语quantity的第1个字母。 |
| r | 代词 | 取英语代词pronoun的第2个字母,因p已用于介词。 |
| s | 处所词 | 取英语space的第1个字母。 |
| Tg | 时语素 | 时间词性语素。时间词代码为t,在语素的代码g前面置以T。 |
| t | 时间词 | 取英语time的第1个字母。 |
| u | 助词 | 取英语助词auxiliary 的第2个字母,因a已用于形容词。 |
| Vg | 动语素 | 动词性语素。动词代码为v。在语素的代码g前面置以V。 |
| v | 动词 | 取英语动词verb的第一个字母。 |
| vd | 副动词 | 直接作状语的动词。动词和副词的代码并在一起。 |
| vn | 名动词 | 指具有名词功能的动词。动词和名词的代码并在一起。 |
| w | 标点符号 | |
| x | 非语素字 | 非语素字只是一个符号，字母x通常用于代表未知数、符号。 |
| y | 语气词 | 取汉字“语”的声母。 |
| z | 状态词 | 取汉字“状”的声母的前一个字母。 |

# 2.1 语料库概述

**例2：** London-Lund英语口语语料库样例：

^what a_bout a cigar\ette# . /
*((4 sylls))* /
*I ^w\on't have one th/anks#* - - - /
^aren't you .going to sit d/own# - /
^[/\m]# - /
^have my _coffee in p=eace# - - - /
^quite a nice .room to !s\it in ((actually))# /
*^\isn't* it# /
*^y/\es#* - - - /

# 2.1 语料库概述

**例2：** London-Lund英语口语语料库部分标记：

| 标记 | 含义 |
|---|---|
| # | 语调群的结束 (end of tone group) |
| ^ | 语音开始 (onset) |
| / | 上升型核心语调 (rising nuclear tone) |
| \ | 下降型核心语调 (falling nuclear tone) |
| ^ | 先升后降型核心语调 (rise-fall nuclear tone) |
| _ | 平型核心语调 (level nuclear tone) |
| [ ] | 不完整的词语和音节符号 (enclose partial words and phonetic symbols) |
| . | 标准重音 (normal stress) |
| ! | 高音高于前一个音节的重音 (booster: higher pitch than preceding prominent syllable) |
| = | 高音跟前一个音节相当的重音 (booster: continuance) |
| (( )) | 不清晰的音节 (unclear) |
| * * | 同步发音 (simultaneous speech) |
| - | 一个重音单位的停顿 (pause of one stress unit) |

# 2.1 语料库概述

## 2. 语料库的类型

- 按内容构成和目的划分（4种类型）
  - 异质的（heterogeneous）　　—[黄昌宁，2002]

    最简单的语料收集方法，没有事先规定和选材原则。
  - 同质的（homogeneous）

    与“异质”正好相反，比如美国的 TIPSTER 项目只收集军事方面的文本。
  - 系统的（systematic）

    充分考虑语料动态和静态问题、代表性和平衡问题以及语料库规模等问题。
  - 专用的（specialized）

    如：北美的人文科学语料库。

# 2.1 语料库概述

- 按语言种类划分
  - 单语的
  - 双语的或多语的

    平行语料库： 篇章对齐 / 句子对齐 / 结构对齐

例如，机器翻译中的双语对齐语料库

C: 早晨$_1$ 好$_2$ !$_3$

E: Good$_2$ morning$_1$ .$_3$

C: 早晨好！

E: Good morning.

C: 您能给我一杯咖啡吗？

E: Could you give me a cup of coffee?

… …

# 2.1 语料库概述

- 按是否加工处理过（标注）划分
  - 生语料库：未经加工的，没有任何切分、标注标记的原始语料库
  - 熟语料库：经过加工，带有切分、标注标记的语料库

—具有词性标注

—句法结构信息标注(树库)

—语义信息标注

# 2.1 语料库概述

## ❑ 共时语料库与历时语料库

- 共时语料库 是为了对语言进行共时(同一时段)研究而建立的语料库。研究大树的横断面所见的细胞和细胞关系，即研究一个共时平面中的元素与元素的关系。

- 历时语料库 是为了对语言进行历时研究而建立的语料库。研究大树的纵剖面所见的每个细胞和细胞关系的演变，即研究一个历时切面中元素与元素关系的演化。

# 内 容 提 要

2.1 语料库概述

2.2 语料库技术的发展

2.3 语料库资源

2.4 语料的收集与加工

2.5 语言知识库

# 2.2 语料库技术的发展

## 1. 语料库技术发展历史

前期（计算机发明以前），第一代语料库，第二代语料库，到第三代语料库

### 第一代（1970－80年代）

百万词级， 以语言研究为导向 。 如，Brown语料库 ， LLC语料库 等

### 第二代（1980－90年代）

千万词级， 词典编撰 --应用导向 。
如，COBUILD语料库 （2000万词级）Longman语料库

### 第三代（1990年代至今 ）

超大规模（上亿词级），标准编码体系，深度标注/多语种，NLP应用，
如， ACL/DCI语料库 ，UPenn树库 ，LDC 等

### 下一代（？ ）

互联网作为语料库

# 2.2 语料库技术的发展

## 2. 语料库发展趋势

时代：六，七十年代到八十年代及九十年代以来.

语料：从单语种到多语种.

数量：从百万级到千万级再到亿级和万亿级.

加工：从词法级到句法级再到语义和语用级.

文本：从抽样到全文

特征：　从固定容量、文本的时间段、范围或应用领域 到动态可变

# 内 容 提 要

2.1 语料库概述

2.2 语料库技术的发展

2.3 语料库资源

2.4 语料库的用途

2.5 语料的收集与加工

2.6 语言知识库

# 2.3 语料库库资源

## 1. 典型语料库

★布朗语料库 (Brown Corpus)

- 20世纪60s，Francis 和 Kucera 在布朗(Brown)大学建立，是世界上第一个根据系统性原则采集样本的标准语料库，100万词规模；
- 15种题材，共500个样本，每个样本不少于2000词；
- 1970s Greene 和 Rubin 设计了TAGGIT词性标注系统（词类标记81种，上下文约束规则3300条），自动标注正确率77％。

★ LLC口语语料库(London-Lund Corpus of Spoken English )

- 1960s 伦敦大学著名语言学家Quirk 组织，瑞典隆德(Lund)大学教授 Svartvik 主持录入计算机
- 87个文本，每个文本约5000词，最终规模 50万词
- 5大类：面对面交谈；电话交谈；讨论；采访；辩论，未经准备的当众评论、论证、演讲，经准备的当众演讲
- 标注：语调、节律、关键词(语段)，词类、出现次数、搭配关系等

# 2.3 语料库库资源

★朗文语料库（Longman Corpus）

- 朗文语料库委员会（Longman Corpus Committee）
- 选自1900～的20世纪英语：知识性(informative)文60%，想象性(imaginative)文本40%
- 2800 万词，10个分布广泛的领域：自然和纯科学、应用科学、社会科学、世界事务等

# 2.3 语料库库资源

★宾州(Pennsylvania)大学树库 (UPenn Tree Bank)
(http://www.ldc.upenn.edu/)

- 美国宾州大学计算机系 M. Marcus 教授主持
- 1993年完成约300万词次英语句子的语法结构标注
- 2000年完成第一版汉语树库，约10万词次，4185个句子
- Chinese Tree Bank (CTB) 中汉语词性(part-of-speech)被划分为33类，23类句法标记(Syntactic tags)

# 2.3 语料库库资源

## Penn Treebank词性标注集

| Tag | Description | Example | Tag | Description | Example |
|---|---|---|---|---|---|
| CC | Coordin. Conjunction | *and, but, or* | SYM | Symbol | +,%, & |
| CD | Cardinal number | *one, two, three* | TO | "to" | *to* |
| DT | Determiner | *a, the* | UH | Interjection | *ah, oops* |
| EX | Existential 'there' | *there* | VB | Verb, base form | *eat* |
| FW | Foreign word | *mea culpa* | VBD | Verb, past tense | *ate* |
| IN | Preposition/sub-conj | *of, in, by* | VBG | Verb, gerund | *eating* |
| JJ | Adjective | *yellow* | VBN | Verb, past participle | *eaten* |
| JJR | Adj., comparative | *bigger* | VBP | Verb, non-3sg pres | *eat* |
| JJS | Adj., superlative | *wildest* | VBZ | Verb, 3sg pres | *eats* |
| LS | List item marker | *1, 2, One* | WDT | Wh-determiner | *which, that* |
| MD | Modal | *can, should* | WP | Wh-pronoun | *what, who* |
| NN | Noun, sing. or mass | *llama* | WPS | Possessive wh- | *whose* |
| NNS | Noun, plural | *llamas* | WRB | Wh-adverb | *how, where* |
| NNP | Proper noun, singular | *IBM* | $ | Dollar sign | $ |
| NNPS | Proper noun, plural | *Carolinas* | # | Pound sign | # |
| PDT | Predeterminer | *all, both* | " | Left quote | (' or ") |
| POS | Possessive ending | *'s* | " | Right quote | (' or ") |
| PP | Personal pronoun | *I, you, he* | ( | Left parenthesis | ( [, (, {, <) |
| PPS | Possessive pronoun | *your, one's* | ) | Right parenthesis | ( ], ), }, >) |
| RB | Adverb | *quickly, never* | , | Comma | , |
| RBR | Adverb, comparative | *faster* | . | Sentence-final punc | (. ! ?) |
| RBS | Adverb, superlative | *fastest* | : | Mid-sentence punc | (: ; ... – -) |
| RP | Particle | *up, off* | | | |

- 33 类
- NN 名词、NR 专业名词、NT 时间名词、VA可做谓语的形容词、VC “是”、VE “有”作为主要动词、VV 其他动词、AD 副词、M 量词，等等。

# 2.3 语料库库资源

**例句：他还提出一系列具体措施的政策要点。**

分词标注：他/PN　还/AD　提出/VV　一/CD　系列/M　具体/JJ　措施/NN　和/CC　政策/NN　要点/NN　。/PU

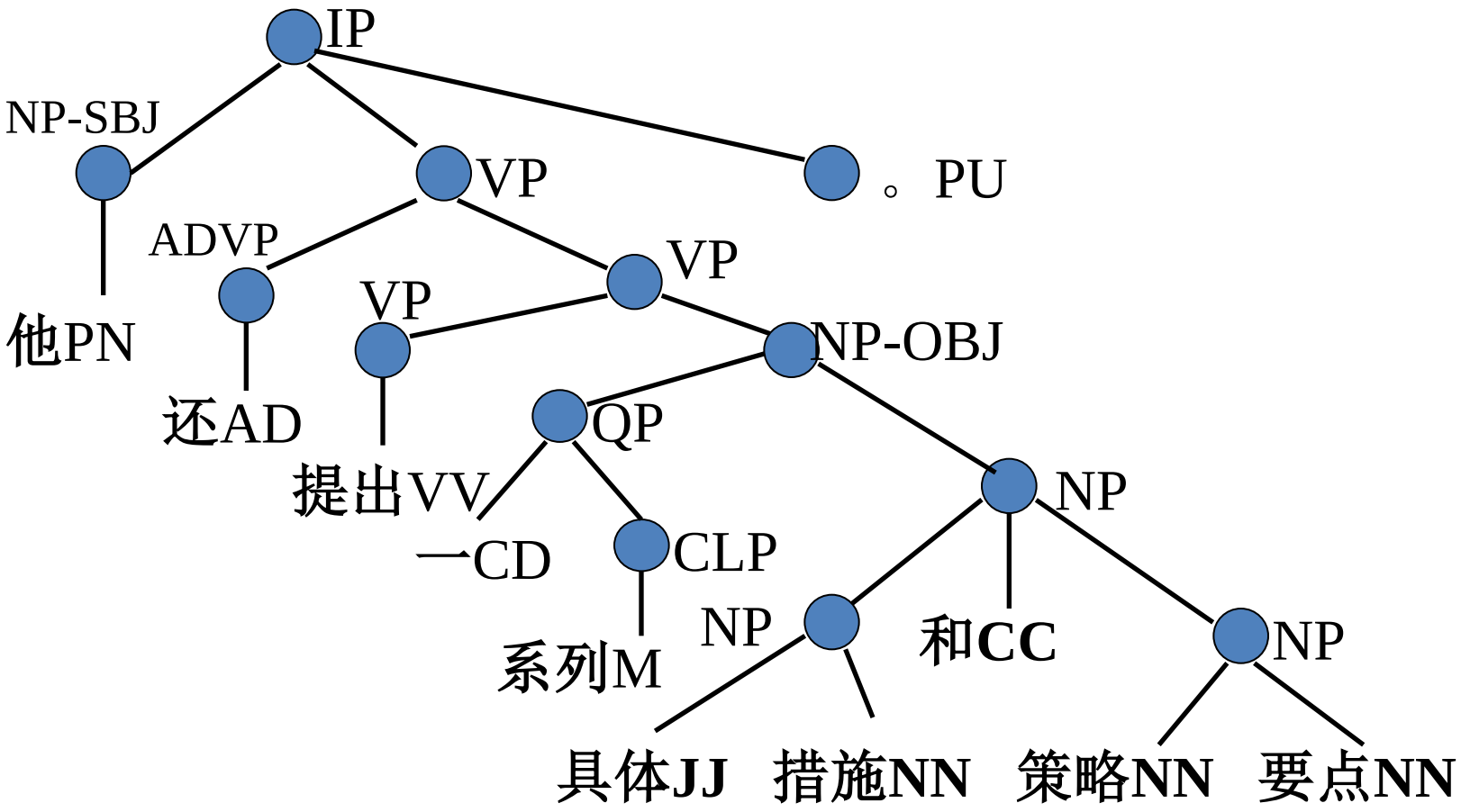
# 2.3 语料库库资源

句法标注：

```
( IP ( NP-SBJ (PN 他 ))
     ( VP (ADVP ( AD 还 ))
          ( VP ( VV 提出 ))
               ( NP-OBJ (QP ( CD 一)
                            ( CLP ( M 系列 )))
                        (NP ( NP ( ADJP ( JJ 具体)
                                        ( NP (NN 措施)))
                            ( CC 和)
                            ( NP ( NN 政策 )
                                 ( NN 要点 ))))))
     ( PU 。 ))
```

# 2.3 语料库库资源

句法树：



# 2.3 语料库库资源

★ 宾州树库的扩展

(1) PropBank （Proposition Bank）

起初是在宾州英语树库(Penn English Treebank) 的基础上增加语义信息后构建的“命题库”，其基本观点认为：树库仅提供句子的句法结构信息，对于计算机理解人类语言是不够的。因此，PropBank 的目标是对原树库中的句法节点标注上特定的论元标记 (argument label)，使其保持语义角色的相似性。

# 2.3 语料库库资源

例如1，John broke the window.

- 事件是 “打碎 (breaking event)”
- John 为事件的 制造者 (instigator)
- window为 受事者 (patient)
- 窗户被打碎 (broken window)为事件的结果

# 2.3 语料库库资源

(3) 宾州语篇树库 (Penn Discourse Tree Bank, PDTB)

建造目标是开发一个标注语篇结构信息的大规模语料库，主要标注与语篇连通方式 (discourse connectives) 相关的一致关系 (coherence relation)。标注信息主要包括连通方式的论元结构、语义区分信息，以及连通方式和论元的修饰关系特征(attribution-related features) 等

# 2.3 语料库库资源

## ★北京大学开发的CLKB

- 现代汉语语法信息词典，含8万词的360万项语法属性描述；
- 汉语短语结构规则库，含600多条语法规则；
- 标注语料库1.5亿字，其中精加工的有5200万字，标注义项2800万字；
- 平行语料库，含对译的英汉句对100万；
- 多领域术语库，有35万汉英对照术语。

## ★台湾中研院平衡语料库（Sinica Corpus）

（http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/）

- 520万词次(789万汉字)汉语平衡语料库
- 语料选自1990年至1996年期间出版的哲学、艺术、科学、生活、社会和文学领域的文本
- 2003年增加了汉英平行语料库，含 2373 个汉英平行对照文本；北大现代汉语语料库，规模约为8500万汉字

# 2.3 语料库库资源

★布拉格依存树库（Prague Dependency Treebank，PDT）

（http://www.elsnet.org/nps/0040.html）

由捷克布拉格查尔斯大学(Charles University in Prague）组织开发，目前已经建成三个语料库：捷克语依存树库、捷克语－英语依存树库和阿拉伯语依存树库。有形态和句法分析层的标注工作和树库的深层语法层（tectogrammatical layer）的信息标注

★中国中文语言资源联盟(Chinese LDC) http://www.chineseldc.org

- 会员单位70多个
- 各类语言资源80余种
- 正式对外转让时间从2005年3月起
- 已共享资源超过133套，销售总额已经达到108万元人民币
- 授权评测单位使用超过40套

# 2.3 语料库库资源

★ One Billion Word Benchmark(Google)

http://www.statmt.org/lm-benchmark/

- 数据库含有大约 10 亿英语单词，词汇有 80 万
- 来源：沃尔玛(WMT)的新闻数据，最新的是2011年的数据

详细介绍：

https://github.com/tensorflow/models/tree/master/lm_1b

# 2.3 语料库库资源

★ Europarl(欧洲议会)

http://www.statmt.org/europarl/

- 包括21种欧洲语言的版本，现在每种语言达到6000万字
- 来源：欧洲议会的会议记录，时间跨度从1996年至2011年
- 目前这个语料库还在继续扩建中

# 2.3 语料库库资源

★ UMBC WebBase Corpus（马里兰大学）

http://ebiquity.umbc.edu/resource/html/id/351

- 语言为英文
- 处理后，规模30亿字
- 来源：斯坦福WebBase项目2007年抓取的1亿网页

# 2.3 语料库库资源

★ BCCWJ（日本国立国语研究所）

http://www.kotonoha.gr.jp/shonagon/search_form

- 语言为日语，规模为1亿500万字
- 来源为11个类型的数据，时间跨度为1976年至2008年

| | | | |
|---|---|---|---|
| 图书 | （1971-2005年， | 22058册， | 约6270万字） |
| 杂志 | （2001-2005年， | 1996册， | 约440万字） |
| 新闻 | （2001-2005， | 1473条， | 约1.4万字） |
| 白皮书 | （1976-2005年， | 1500册， | 约490万字） |
| 教材 | （2005- 2007年， | 412册， | 约90万字） |
| 报纸 | （2008年， | 354册， | 约380万字） |
| 雅虎问答 | （2005年， | 91445条， | 共10.3万字） |
| 雅虎博客 | （2008年， | 52680条， | 约10.20亿字） |
| 诗 | （1980-2005年， | 252册， | 约20万字） |
| 法律 | （1976-2005年， | 346条， | 约110万字） |
| 国会会议记录 | （1976-2005年， | 159册， | 约510万字） |

详细介绍 ： http://www.kotonoha.gr.jp/shonagon/

# 内 容 提 要

2.1 语料库概述

2.2 语料库技术的发展

2.3 语料库资源

2.4 语料的收集与加工

2.5 语言知识库

# 2.4 语料的收集与加工

## 1. 建库之前应考虑：

| 语料库三方面 | 属性 | 值 |
| --- | --- | --- |
| A. 语料本身 | 规模 | 百万词级 \| 千万词级 \| 亿万词级 \| … |
| | 领域 | 政治 \| 经济 \| 体育 \| 心理学 \| … |
| | 体裁 | 文学 \| 应用文 \| 新闻 \| … |
| | 时代 | 共时 \| 历时 |
| | 语体 | 书面语 \| 口语 |
| | 语种 | 单语 \| 双语 \| 多语<br>双语平行语料库 \| 双语比较语料库 |
| | 语言层次 | 语音（音节，韵律） \| 语法（词，句，…） |
| | | |
| B. 语料加工 | 数据形式 | Text文本 \| HTML文本 \| 数据库 \| … |
| | 编码体系 | TEI标准 \| 自定义编码体系 \| … |
| | 加工层次 | 词 性 \| 句 法 \| 语 义 \| 语 篇 \| …<br>双语句子对齐 \| 词对齐 \| … |
| | 加工方式 | 自动 \| 人机互助 \| 人工 |
| | | |
| C. 语料应用 | 应用领域 | 通用 \| 词典编纂 \| 机器翻译 \| … |
| | 辅助软件 | 检索工具 \| 人机界面 \| 数据接口 \| … |

# 2.4 语料的收集与加工

## 语料的选取标准

- 精品原则
- 有影响力原则
- 随机挑选原则
- 高流通度原则
- 典型性原则
- 易于获得原则
- 具有统计样本意义原则
- 符合语言规范原则
- 语料库中各类文本的比例均衡原则
- 专业语料库的建设应有专业领域的专家参与

# 2.4 语料的收集与加工

## 2. 语料的收集：

### 获取语料的途径

- 纸质媒介 → 人工录入 → 光学扫描、OCR软件
- 电子语料：光盘语料 + 互联网语料
- 双语平行语料库：
- 大型国际组织（联合国、欧盟）
- 双语社会（加拿大、新加坡、香港）

### 语料文件的数据格式

- 文件格式：.doc，txt，pdf，ps，rtf
- 采用纯文本文件格式存放语料，便于计算机处理
- 采用关系数据库组织语料，直接利用数据库的检索、统计等功能
- 要考虑字符编码方式

# 2.4 语料的收集与加工

## 3. 语料的编码：

语料库的编码体系

- SGML（标准置标语言）

http://www.w3.org/MarkUp/SGML/

- XML（可扩展的置标语言）

http://www.w3.org/TR/REC-xml

- TEI（文档编码计划）

http://www.tei-c.org/

- CES（语料库编码标准）

http://www.tei-c.org/Applications/index-co02.html

# 2.4 语料的收集与加工

## 4. 语料的加工：

语料库加工/标注：隐形信息→显性信息

- 词性标记（Part-of-speech tagging）
- 句法标记（Grammatical parsing）
- 词义标记（Word sense tagging）
- 篇章指代标记（Anaphoric annotation）
- 韵律标记（Prosodic annotation）

制定标注集

双语(平行)语料库的对齐：

- 段落对齐
- 句子对齐
- 词对齐
- 短语对齐

# 内 容 提 要

2.1 语料库概述

2.2 语料库技术的发展

2.3 语料库资源

2.4 语料的收集与加工

2.5 语言知识库

# 2.5 语言知识库

**语言知识库：**从大量的实例语料中提炼、抽象、概括出来的系统的语言知识，如电子词典、句法规则库、词法分析规则库等。

# 2.5 语言知识库

## 1. WordNet (http://wordnet.princeton.edu/)

- 普林斯顿大学(Princeton University) 认知科学实验室 George A. Miller教授领导开发。
- 开发目的：解决词典中同义信息的组织问题
- 目前规模：95600 英语词条，其中，51500个简单词，44100 个搭配词。70100个词义(同义词集合)。
- 五大类词汇：名词、动词、形容词、副词、虚词。(实际上 WordNet 中仅包含前4类)
- 特色：根据词义（而不是词形）组织词汇信息，从某种意义上讲，它是一部语义词典。
- WordNet 按语义关系组织：语义关系看作是同义词集合之间的一些指针，语义关系是双向的。如果词义{$x_1$, $x_2$, …} 和{$y_1$, $y_2$, …}之间有一种语义关系R，则在{$y_1$, $y_2$, …}和{$x_1$, $x_2$, …} 之间也有语义关系R。属于这两个同义词集合的单词之间的关系也是R。

# 2.5 语言知识库

- 4 种语义关系：
  - 同义关系 (synonymy)
  - 反义关系 (antonymy)
  - 上下位关系(hypernymy)或称从属/上属关系：如：{枫树}是{树}的下位，{树}是{植物}的下位。
  - 部分关系(meronymy)或称部分/整体关系。

# 2.5 语言知识库

➢名词的25个独立起始概念：

{动作，行为，行动}、{自然物}、{动物，动物系}、{自然现象}、{人工物}、{人，人类}、{属性，特征}、{植物，植物系}、{身体，躯体}、{所有物}、{认知，知识}、{作用，方法}、{信息，通信}、{量，数量}、{事件}、{关系}、{直觉，情感}、{形状}、{食物}、{状态，情形}、{团体，组织}、{物质}、{场所，位置}、{时间}、{目的}

- 21000个动词词形、约8400个词义，14个文件：

  照顾动词，功能动词，变化动词，认知动词，通信动词，竞争动词，消费动词，接触动词，创作动词，感情动词，运动动词，感觉动词，占用动词，社会交往动词，天气变化动词。

- 19500个形容词词形，近10000个词义

  描述性形容词，参照修饰形容词，颜色形容词，关系形容词。

# 2.5 语言知识库

## WordNet 的应用

词汇消歧，语义推理，理解等。

例如：食堂 没 地方，我 在 饭馆 吃 了 蛋 炒饭。

“地方”的三种意思：

＃指地理位置　　如：在祖国各个地方

＃指空间　　如：没地方

＃指部分　　如：他说的有些地方不对

三个含义在两棵不同的名词集成语义树上：

# 2.5 语言知识库

## 2. 知网(HowNet) (http://www.keenage.com)

➢1988年由董振东教授提出，4个基本观点：

(1)NLP系统最终需要更强大的知识库的支持。

(2)知识是一个系统，是一个包含着各种概念与概念之间的关系，以及概念的属性与属性之间的关系的系统。

(3)关于知识库建设，他提出应首先建立一种可以被称为知识系统的常识性知识库。它以通用的概念为描述对象，建立并描述这些概念之间的关系。

(4)首先应由知识工程师来设计知识库的框架，并建立常识性知识库的原型。在此基础上再向专业性知识库延伸和发展。专业性知识库或称百科性知识库主要靠专业人员来完成。这里很类似于通用的词典由语言工作者编纂，百科全书则是由各专业的专家编写。

➢ 知网的哲学

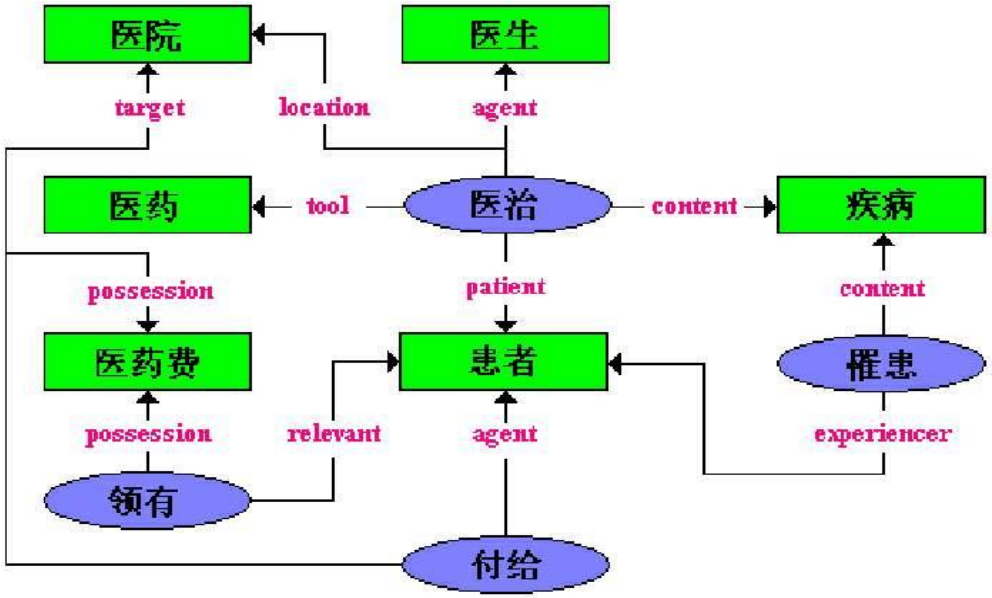
世界上一切事物（物质的和精神的）都在特定的时间和空间内不停地运动和变化。它们通常是从一种状态变化到另一种状态， 并通常由其属性值的改变来体现。

# 2.5 语言知识库

➢ 知网的特色

知网作为一个知识系统，名副其实是一个网而不是树。它所着力要反映的是概念的共性和个性，例如：对于“医生”和“患者”，“人”是它们的共性。

同时知网还着力要反映概念之间和概念的属性之间的各种关系。



# 2.5 语言知识库

- 知网描述了下列各种关系：
  - (a) 上下位关系（由概念的主要特征体现）
  - (b) 同义关系
  - (c) 反义关系
  - (d) 对义关系
  - (e) 部件-整体关系
  - (f) 属性-宿主关系
  - (g) 材料-成品关系
  - (h) 施事/经验者/关系主体-事件关系（如“医生”，“雇主”等）
  - (i) 受事/内容/领属物等-事件关系（如“患者”，“雇员”等）
  - (j) 工具-事件关系（如“手表”，“计算机”等）
  - (k) 场所-事件关系（如“银行”，“医院”等）
  - (l) 时间-事件关系（如"假日"，"孕期"等）

# 2.5 语言知识库

(m) 值-属性关系（如“蓝”，“慢”等）

(n) 实体-值关系（如“矮子”，“傻瓜”等）

(o) 事件-角色关系（如“购物”，“盗墓”等）

(p) 相关关系（如“谷物”，“煤田”等）

# 2.5 语言知识库

词语例子：

NO.=000001
W_C=打
G_C=V
E_C=~酱油，~张票，~饭，去~瓶酒，醋~来了
W_E=buy
G_E=V
E_E=
DEF=buy|买

NO.=015492
W_C=打
G_C=V
E_C=~毛衣，~毛裤，~双毛袜子，~草鞋，~一条围巾，~麻绳，~条辫子
W_E=knit
G_E=V
E_E=
DEF=weave|辫编

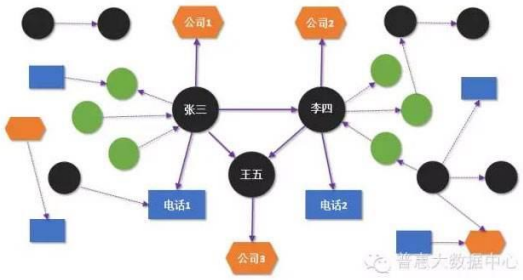
# 2.5 语言知识库

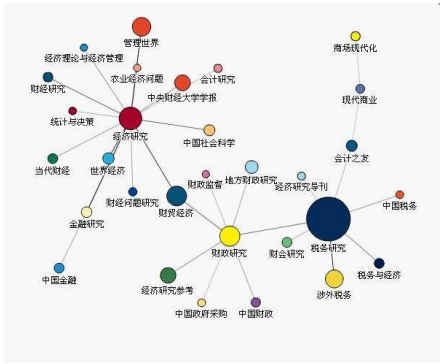
## 3. 知识图谱

知识图谱，也称为科学知识图谱，它通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。为学科研究提供切实的、有价值的参考。

--- 百度百科

知识图谱本质上是一种语义网络。其结点代表实体(entity)或者概念(concept)，边代表实体/概念之间的各种语义关系；用来描述真实世界中存在的各种实体和概念，及实体、概念之间的关联关系。



# 2.5 语言知识库

★ Freebase™

http://freebase.com

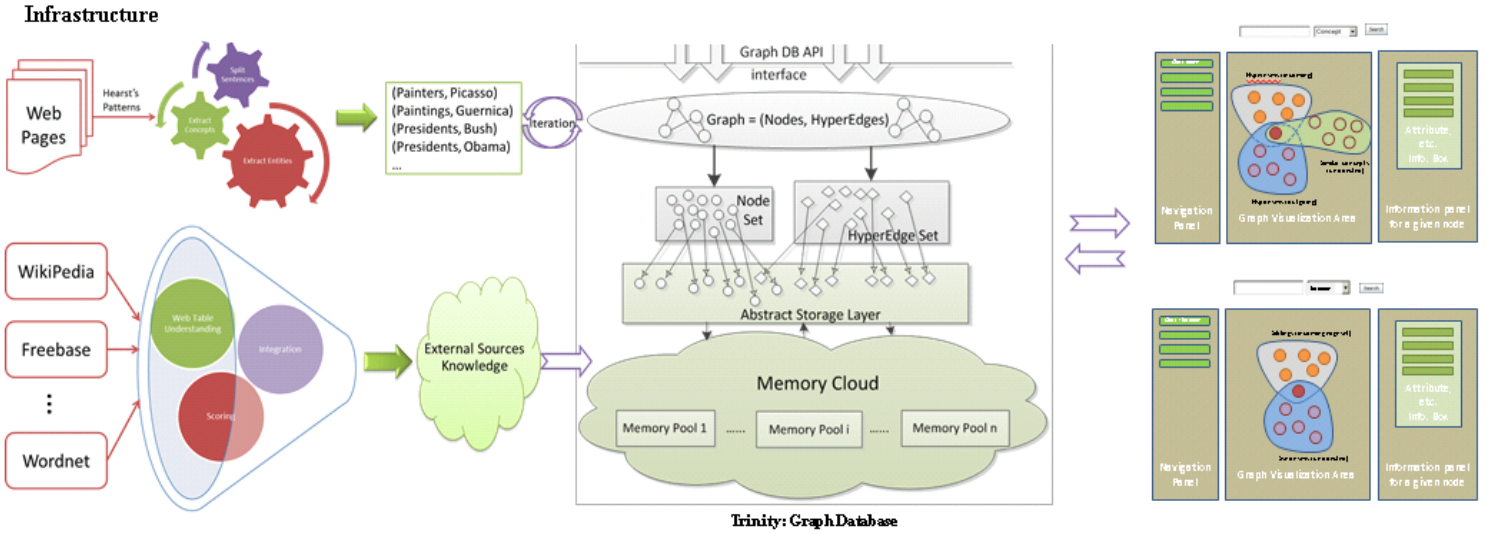
- 包含3900万个实体和18亿条实体关系
- 允许任何人创建、修改、查询的知识库，即众包模式。
- 存储的是结构化良好、机器也可读的数据格式。
- 2010年7月被Google收购。2015年，Google宣布将逐步关停Freebase, Freebase原有的数据迁移至WikiData。

# 2.5 语言知识库

★ ProBase

http://research.microsoft.com/enus/projects/probase/

- 微软构建的知识图谱。
- 目标是使机器“意识到”人类的精神世界，使机器能更好地了解人类的沟通。
- 包含5,376,526个唯一概念，12,501,527个唯一实例和85,101,174个 IsA关系。



# 2.5 知识图谱

## ★ *Microsoft Concept Graph*

https://concept.research.microsoft.com/Home/Download

- 微软亚洲研究院2016年10月27日正式发布，用于帮助机器更好地理解人类交流并且进行语义计算。
- 知识图谱包含了超过540万条概念。
- 包含的知识来自于数以亿计的网页和数年积累的搜索日志，可以为机器提供文本理解的常识性知识。

# 2.5 语言知识库

★ 


http://dbpedia.org

- 由德国莱比锡大学等机构发起的项目，从维基百科中抽取实体关系，包括1千万个实体和14亿条实体关系。
- 数据集以多达125种不同语言表示。
- DBpedia项目使用资源描述框架（RDF）来表示提取的信息，包括30亿个RDF三元组：从维基百科的英文版提取的5.8亿和从其他语言版本提取的24.6亿。

# 2.5 语言知识库

知识图谱列表：

| 类别 | 名称 | 网址 |
| --- | --- | --- |
| 基于维基百科 | DBPedia | http://dbpedia.org |
| | YAGO | http://yago-knowledge.org |
| | Freebase | http://freebase.com |
| | WikiTaxonomy | http://www.hits.org/english/research/nlp/download/wikitaxono my.php |
| | BabelNet | http://babelnet.org |
| 开放知识抽取 | KnowItAll | http://openie.cs.washington.edu |
| | NELL | http://rtw.ml.cmu.edu |
| | Probase | http://research.microsoft.com/enus/projects/probase/ |

# 思考与练习

- 查阅并了解

  1. 宾州大学语料库（UPenn Tree Bank）

  2. 北京大学语料库

- 自学（复习）

  第3章 ： 概率论&信息论基础知识

# 参考文献：

宗成庆，统计自然语言处理（第2版）课件

http://wenku.baidu.com/link?url=LCyTYzgORW7MUxDqtUmgos1NwkwGydBEVMLS6EIDNhIp7MKxbXYlcrpOGfSFooev4AZhYSO_ypkRWsR62u1WcpMUvsMQCmPGPg6iZGMaiTS

秦志君，常见语料库使用入门

**在 此 表 示 感 谢 ！**

信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

# 谢谢各位！