语言数据和语料库

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)

自然语言: 数据视角

- 自然语言是一个系统
  - 语言是主要以呼吸器官发声为基础来传递讯息的符号系统，是人类重要的交际工具和存在方式之一
  - 用于表达事物、动作、思想、状态的一个系统
  - 人类共有的有意义的体系
- 自然语言是一个集合
  - 目前世界现存语言大约6909种，只有2000多种语言有书面文字，2500种语言濒危
- 自然语言的采样：语料库
  - 时间、空间、领域等

语料库

- 什么是语料库？
  - 语料库（corpus）一词在语言学上意指大量的文本，通常经过整理，具有既定格式与标记
- 相关概念：语料库语言学（Corpus Linguistics）
  - 概念
    - 根据篇章材料对语言的研究称为语料库语言学（K. Aijmer & B. Aitenberg, 1991）
    - 以语料为语言描写的起点或以语料为验证有关语言 的假说的方法称为语料库语言学（D. Crystal, 1991）
    - 基于现实生活中语言运用的实例进行的语言研究称 为语料库语言学（T. McEnery & A. Wilson, 1996）
  - 语料库语言学从上世纪60年代开始，发展至今已有50多年历史

语料库—发展历史

- 20世纪50年代中期之前：早期
  - 语料库在语言研究中被广泛使用：语言习 得、方言学、语言教学、句法和语义、音 系研究等
- 1957～20世纪80年代初期：沉寂时期
  - 1957年 Chomsky 的《句法理论》及其以后一系列著作的发表，根本改变了语料库语言学的发展状况。
  - Chomsky 及其转换生成语法学派批判早期的语料库研究方法：
    - 基于语料库的研究方法有误
    - 语料的不充分性

语料库—发展历史

- 20世纪80年代以后：复苏与发展时期
  - 特征之一：第二代语料库相继建成
  - 1983年英国Lancaster 大学建成 Lancaster-Oslo / Bergen Corpus (LOB语料库)：研究英国英语
  - 法国国家科学研究中心与美国芝加哥大学联合 建成法语语料库(Tremor de la Language Francaise, TLF语料库)
  - 芬兰赫尔辛基大学建成历史英语语料库(The Helsinki Corpus of Historical English)
  - 1988年伦敦大学建成国际英语语料库(The International Corpus of English, ICE)：语料来自所有英语国家

语料库—发展历史

- 特征之二：基于语料库的研究项目增多

起止年限	研究项目数目
1959-1965	10
1966-1970	20
1971-1975	30
1976-1980	80
1981-1985	160
1986-1991	320

语料库加工-文本处理

- 垃圾格式问题
  - 语料库内容来源复杂，存在杂质
  - 杂质包括：文档页眉和分隔符、排版代码、表和图
  - 如果数据来源于OCR，存在识别错误等问题
  - 需要过滤器过滤这些杂质
- 大小写
  - Brown语料库中都是大写字母
  - 简单方法：全部变为大写和小写
    - 问题：无法区分例如Richard Brown和brown paint中的Brown
  - 启发式方法：每个句子开头大写字母转为小写，把一串连续大写的词当做标题和副标题，可以认为其余的大写字母为名字，可以保留

语料库加工-文本处理

- 标记化（Tokenization）
  - 将文本切分成为词次（token）的组合，token可以是一个词，一个数字或一个标点符号
- 什么是一个词？
  - 前后有空格的连续字母组成的字符串，可以包含连字符和省略号，但是不能包含其他标点符号----Kucera and Francis（1967）
  - 反例：￥22.50

语料库加工-文本处理

- 句点
  - 词语前后不总存在空格
  - 标点符号常紧跟在词语后，如逗号，分号和句点
  - 对于句点来说去掉并不容易
    - 大多数句点的作用是表示句子结尾
    - 其他情况表示缩写，例如etc.
    - 这些表示缩写的句点应保留作为词语的一部分
    - 当etc.等缩写词语出现在句尾时，句子末尾只保留一个句点，这个句点可同时表示两种意思，在语法意义上，这种现象称为缩略

语料库加工-文本处理

- 单撇号
  - I’ll或isn’t，可以认为是两个词的缩写（切分），I will和is not，也可以认为是一个词（不切分）
  - S→NP VP
    - 不切分：I’m right，以上规则不适用
    - 切分：预料中会出现’s和n’t等词
- 连字符
  - 1.为排版美观，把词语分开，中间插入连字符以改进半片的对齐
    - 找到一行中最后的连字符，丢其它，把本行的词的一部分和下一行词的一部分连接起来
    - 缩略问题：处于行末的连字符不是此种情况下的连字符，连字符只出现了一次不是两次，上面方法不总是正确

语料库加工-文本处理

- 连字符
  - 2. E-mail, co-operate, so-called中的连字符为称为词汇连字符，通常被插入到小构词要素之前或之后，有时充当分裂原音序列的分隔符
  - 3. 用来帮助区分正确的词组
    - once-quiet, text-based, 26-year-old
- 相同形式表示不同的词语
  - 同形异义词
  - saw作为名词是锯子，但saw也是动词see的过去式

语料库加工-文本处理

- 词法
  - 词干化（stemming）
    - sit, sits, sat
  - 当性能评价指标是查询平均值时，词干化对经典IR系统的性能提高没有帮助
    - 某人输入了business，词干化将返回所有包含busy的文档，这不是一个好结果
    - 词法分析把一个词次切分成几个词次，然而，把密切相关的信息组合到一起也是值得的，尽管这将减少词汇量，实践中常把多字词组看成一个特殊的词次，这样可改善系统性能

语料库加工-文本处理

- 句子定义—启发式算法
  - 在 . ？！（和可能的 ； :）出现位置之后加一个假设的句子边界
  - 如果假设边界后面有引号，那么把假设边界移到引号后面
  - 除去以下情况中句点的边界资格：
    - 如果句点之前是一个不总出现在句子末尾的众所周知的缩写形式，而且通常后面会跟一个大写的名字，例如Prof.或者vs.
    - 如果句点前是一个众所周知的缩写形式，但是句点后面没有大写词。这样即可正确的处理像etc.或者Jr.这样大多数缩写用法，这些缩写一般出现在句子的中间或末尾
  - 如果下面条件成立，则除去?或者!的边界资格：
    - 这些符号后面跟着一个小写字母（或者一个已知名字）
  - 认为其他假设边界就是句子的边界

语料库加工-文本处理

- 句子边界的研究
  - Reley(1989)使用了统计学分类树来确定句子边界，分类树使用的特征包括句点前面和后面出现的词的大小写形式和长度，还有不同词出现在句子边界前后的先验概率
  - Palmer and Hearst(1994; 1997)通过简单的使用前后词的词性分布避免了获取上处概率数据的困难，使用神经元网络方法预测句子边界，构造了一个健壮的、很大程度上语言无关的高效率边界检测算法（正确率98%~99%）
  - Reynar and Ratnaparkhi(1997)以及(Mikheev 1998)开发了一个针对此类问题的最大熵系统，正确率达99.25%

语料库加工-文本处理

- 句子长度—新闻文本

长度	数量	百分比	累计百分比
1~5	1317	3.13	3.13
6~10	3215	7.64	10.77
11~15	5906	14.03	24.80
16~20	7206	17.12	41.92
21~25	7350	17.46	59.38
26~30	6281	14.92	74.30
31~35	4740	11.26	85.56
36~40	2826	6.71	92.26
41~45	1606	3.82	96.10
46~50	858	2.04	98.14
51~100	780	1.85	99.99
101+	6	0.01	100

语料库加工-格式标注

- 通用标记语言（Standard Generalized Markup Language, SGML ）
  - SGML是超文本格式的最高层次标准，是可以定义标记语言的元语言
  - HTML和XML同样派生于它：XML可以被认为是它的一个子集，XML的产生就是为了简化它，以便用于更加通用的目的。而HTML是它的一个应用

```
<QUOTE TYPE="example">
typically something like <ITALICS>this</ITALICS>
</QUOTE>
```

语料库加工-数据标注

- 语法标注
  - Brown标注集
  - 兰开斯特大学开发了一系列标注集，用来对Lancaster-Oslo-Bergen语料库进行标注
  - 英国国家语料库标注集：CLAWS1到CLAWS5
  - Penn树库标注集是Brown标注集的一个简化版本

标注集	基本大小	总标记数
Brown	87	179
Penn	45	
CLAWS1	132	
CLAWS2	166	
CLAWS c5	62	
London-Lund	197	

语料库加工-数据标注

- 按几种不同的标注集对一个例句进行标注

句子	CLAWS c5	Brown	Penn 树库	ICE
she	PNP	PPS	PRP	PRON(pers,sing)
was	VBD	BEDZ	VBD	AUX(pass,past)
told	VVN	VBN	VBN	V(ditr,edp)
that	CJT	CS	IN	CONJUNC(subord)
the	AT0	AT	DT	ART(def)
journey	NN1	NN	NN	N(com,sing)
might	VM0	MD	MD	AUX(modal,past)
kill	VV1	VB	VB	V(montr,infin)
her	PNP	PPO	PRP	PRON(poss,sing)
.	PUN	.	.	PUNC(per)

语料库加工-搭配抽取

- 频率方法
  - 如果两个词在一起出现很多次，它们很有可能是搭配
  - 仅仅选择最频繁出现的二元组，结果并不理想，见右图
  - 除了New York以外，所有的二元组都是一对功能词

$C(W^1 W^2)$	W^1	W^2
80 871	of	the
58 841	in	the
26 430	to	the
21 842	on	the
21 839	for	the
18 568	and	the
16 121	that	the
15 630	at	the
15 494	to	be
13 899	in	a
13 689	of	a
13 361	by	the
13 183	with	the
12 622	from	the
11 428	New	York
10 007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

语料库加工-搭配抽取

- 频率方法
  - Justenson和Katz通过一个词性过滤器来过滤候选短语
  - A代表形容词，P代表前置词，N代表名词

标记模式	示例
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

语料库加工-搭配抽取

- 频率方法
  - 使用Justeson和Katz词性过滤器结果，如右图

$C(W^1\ W^2)$	W^1	W^2	标记模式
11 487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

语料库加工-搭配抽取

- 均值和方差方法
  - 基于频率的搜索方法可以很好的解决固定搭配的识别问题，但是很多搭配是两词搭配，并且彼此之间的关系非常灵活
  - 考虑动词knock和它的最频繁出现的搭配之一door

    a. she knocked on his door

    b. they knocked at the door

    c. 100 women knocked on Donaldson's door

    d. a man knocked on the metal front door

  - 在knocked和door之间出现的词是变化的，并且两个词之间的距离也不是固定的，所以不能识别出固定短语搭配来
  - Donaldson’s 算作三个词

语料库加工-搭配抽取

- 均值和方差方法
  - 上例中Knocked和door之间的平均偏移量

$$\frac{1}{4}(3+3+5+5)=4.0$$

  - 方差衡量的是单独的偏移量偏离均值的距离，可用下面的公式估计方差：

$$s^2=\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n-1}$$

  - 则上例中Knocked和door距离的方差为：

$$s=\sqrt{\frac{1}{3}((3-4.0)^2+(3-4.0)^2+(5-4.0)^2+(5-4.0)^2)}\approx 1.15$$

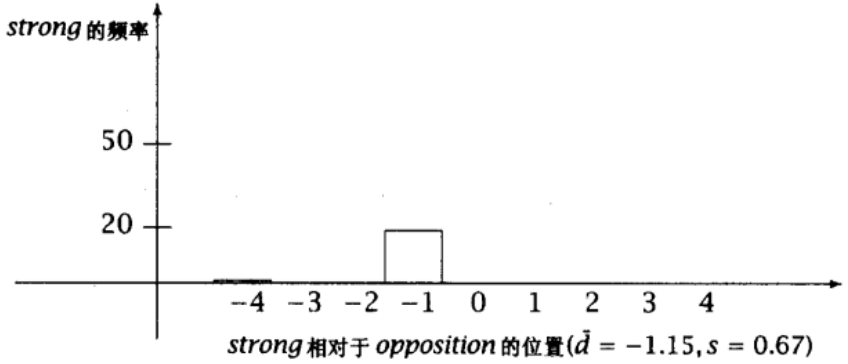
语料库加工-搭配抽取

- 均值和方差方法
  - 如果在所有的情况下样本的偏移量是相同的，那么方差为零
  - 如果偏移量是随机分布的（在这种情况下，两个词偶然同现，而不是由于特殊的关系），那么方差将会较大
  - 均值和方差特性化了语料库中两个词之间距离的分布
  - 可以使用这个信息来发现搭配，具体的方法是通过寻找带有低偏差值的词对，这意味着两个词通常会以大致相同的距离出现

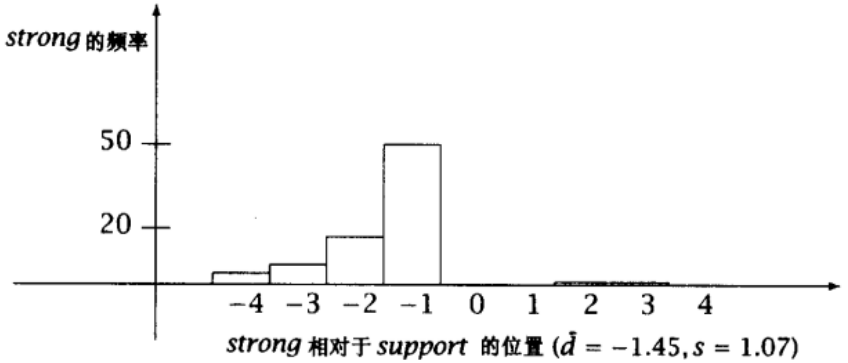
语料库加工-搭配抽取

- 均值和方差方法

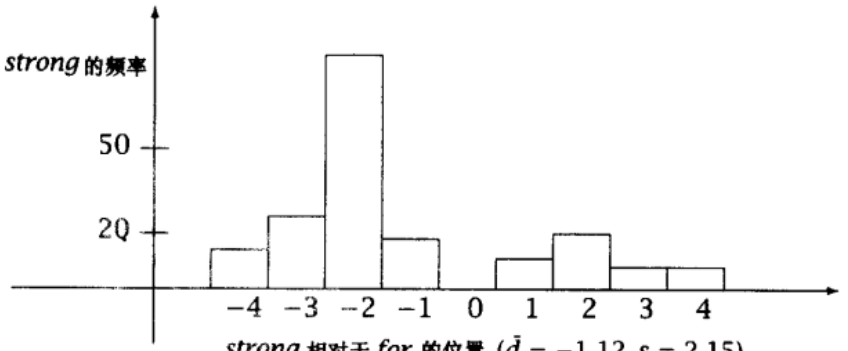
s	d	计数	第一个单词	第二个单词
0.43	0.97	11 657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said



strong 相对于 opposition 的位置($\bar{d} = -1.15, s = 0.67$)



strong 相对于 support 的位置 ($\bar{d} = -1.45, s = 1.07$)



strong 相对于 for 的位置 ($\bar{d} = -1.12, s = 2.15$)

语料库的性质

- 语料库可以使用机器进行处理
  - 利用机器自动管理语料库
  - 机器可以对语料库进行检索、统计等
  - 对语料进行自动或半自动的加工等
  - 过去也曾有过专书引得，或是为词典编纂、语言研究等而做的语料卡片库，但它们不能用机器处理，不是现代意义上的语料库
- 语料是自然话语材料
  - 是自然语言而非人造语言
  - 也不能经过语料收录者修改润色
  - 甚至要保留其中的不规范现象，如果修改都会损害语料的真实性

语料库的性质

- 语料库应具有一定规模
  - 语料库的规模是指收录语料的数量大小
  - 语料库以统计见长，没有数量规模基础无法发挥其特长
  - 语料库规模受制于三个因素
    - 语料获取的难度与成本
    - 计算机的存储能力、运算速度、网络运行及其成本
    - 不同类型语料库的应用需求，如用于常用字统计、常用词统计、一般语法现象研究的语料库，其规模就可以小一些；而用于罕用字统计、罕用词统计计、特殊语法现象研究的语料库，因数据稀疏因素的影响，其规模就需要大些

语料库的性质

- 语料库具有一定的结构
  - 存储结构
    - 一定规模的语料，在语料库中需要有一定的存放方式，语料在语料库中的存放方式形成了语料库的存储结构
  - 内容结构
    - 语料与语料之间会发生一定关联，如时间关联、使用领域关联、作者关联、主题内容关联等。语料之间的关联方式形成语料库的内容结构
  - 物理结构
    - 广义的语料库结构，还应包括处理语料的物理系统和软件系统，这是语料库的物理结构

语料库的性质

- 语料库具有知识标记
  - 原始数据标记
    - 对语料的知识版权信息、载体发行信息、采样方式信息等元数据(Metadata)的标记，称为原始数据标记。
  - 语言知识标记
    - 语料库中各种语言单位的性质的标记、语言单位间语法关系和语义关系的标记、语言片段的语用标记等等，统称为语言知识标记。

语料库的种类

- 按语种划分
  - 单语种语料库和多语种语料库
- 按记载媒体划分
  - 单媒体语料库和多媒体语料库
- 按地域区别划分
  - 国家语料库和国际语料库

语料库的种类

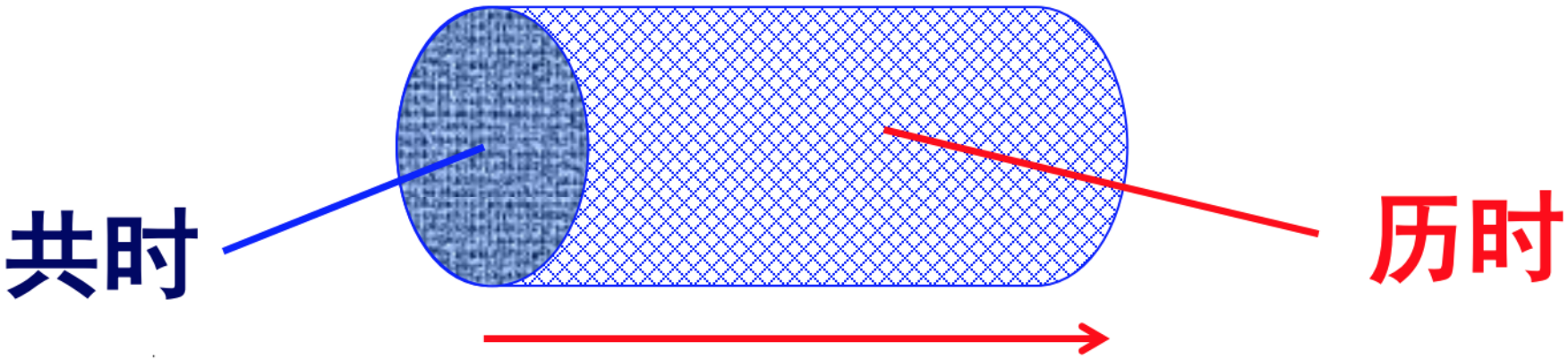
- 平衡语料库和平行语料库
  - 平衡语料库着重考虑的是语料的代表性和平衡性
    - 张普（2003）曾经提出语料采集的七项原则：语料的真实性、语料的可靠性、语料的科学性、语料的代表性、语料的权威性、语料的分布性和语料的流通性
    - 其中，语料的分布性还要考虑语料的科学领域分布、地域分布、时间分布和语体分布等
  - 平行语料库
    - 同种语言的语料上的平行，如国际英语语料库，共有20个平行的子语料库，分别来自英国、美国、加拿大等国家。其平行性表现为语料选取的时间、对象、比例、文本数、文本长度等几乎是一致的。建库的目的是对不同国家的英 语进行对比研究
    - 对平行语料库的另一种理解是指两种或多种语言之间的平行采样和加工。例如，机器翻译中的双语对齐语料库（句子对齐或段落对齐）

语料库的种类

- 通用语料库与专用语料库
  - 所谓的通用语料库实际上与平衡语料库是从不同角度看问题的结果，或者说是与专用领域对举的结果
  - 为了某种专门的目的，只采集某一特定领域、特定地区、特定时间、特定类型的语料构成的语料库就是专用语料库。例如，新闻语料库、科技语料库、中小学语料库、北京口语语料库等。
  - 一般把抽样时仔细从各个方面考虑了平衡问题的平衡语料库称为通用语料库

语料库的种类

- 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研究而建立的语料库，按照索绪尔的观点，共时研究是指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。无论所采集语料的时间段有多长，只要研究的是一个平面上的元素或元素的关系，就是共时研究，所建立的语料库就是共时语料库



语料库的种类

- 共时语料库与历时语料库
  - 历时语料库是为了对语言进行历时研究而建立的语料库。按照索绪尔的观点，历时研究是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
  - 根据历时语料库得到的统计结果就不像共时语料库的统计结果是一个频次点，而是依据时间轴的等距离抽样得到的若干频次变化形成的演变曲线，我们把这种曲线称为变化“走势图”
  - 香港城市大学建立的LIVAC（Linguistic Variations in Chinese Speech Communities）语料库是历时语料库的典型代表。该语料库自1995年开始构建，采集和处理了来自北京、上海、香港、台湾、澳门和新加坡六个泛华语地区有代表性的中文报章语料，累计收集了150万个词条，总字数达4亿汉字

语料库的种类

- 共时语料库与历时语料库
  - 张普认为，判断历时语料库有4条基本原则［张普，2003］：
    - 是否动态语料库：语料库必须是开放的、动态的
    - 语料库的文本是否具有量化的流通度属性：所有的语料都应来源于大众传媒，都具有采用不同计算方法与传媒特色相应的流通度属性。其量化的属性值也是动态的
    - 语料库的深加工是否基于动态的加工方法：随着语料的动态采集，语料也应进行动态加工
    - 是否取得动态的加工结果：语料的加工结果也应是动态和历时的

语料库的种类

- 生语料与标注语料库
  - 生语料是指没有经过任何加工处理的原始语料数据（corpora with raw data），如Chinese Gigaword和后面将要提到的BTEC口语语料库等
  - 标注语料库是指经过加工处理、标注了特定信息的语料库。根据加工程度不同，标注语料库又可以细分为分词语料库（主要指汉语）、分词与词性标注语料库、树库（tree bank）、命题库（proposition bank）、篇章树库（discourse tree bank）等。

典型语料库介绍

- 布朗语料库
  - 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
  - 选自1961年美国人撰写出版的普通语体的文本
  - 15种题材，共500个样本，每个样本不少于2000词
  - 1961年布朗大学出版了当代英语词频词典
  - 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%

典型语料库介绍

- 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
  - 2000小时的对话和广播等口语素材
  - 瑞典隆德（Lund）大学教授 Svartvik 主持录入计算机
  - 英语口语调查（The Survey of Spoken English, SSE）
  - SSE 于1981年完成，建成 London-Lund Corpus of Spoken English (LLC)
  - 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

典型语料库介绍

- 朗文语料库 (Longman Corpus)
  - 朗文语料库委员会（Longman Corpus Committee）
  - January 1981- November 1990
  - 设计原则：
    - 尊重本族语言者的直觉和语料库权威
    - 向研究人员提供语料（英国50%，美国40%，其它国家10%）
    - 书面语
  - 选自1900～的20世纪英语：知识性（informative）文本 60%，想象性（imaginative）文本40%
  - 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
  - 2800 万词

典型语料库介绍

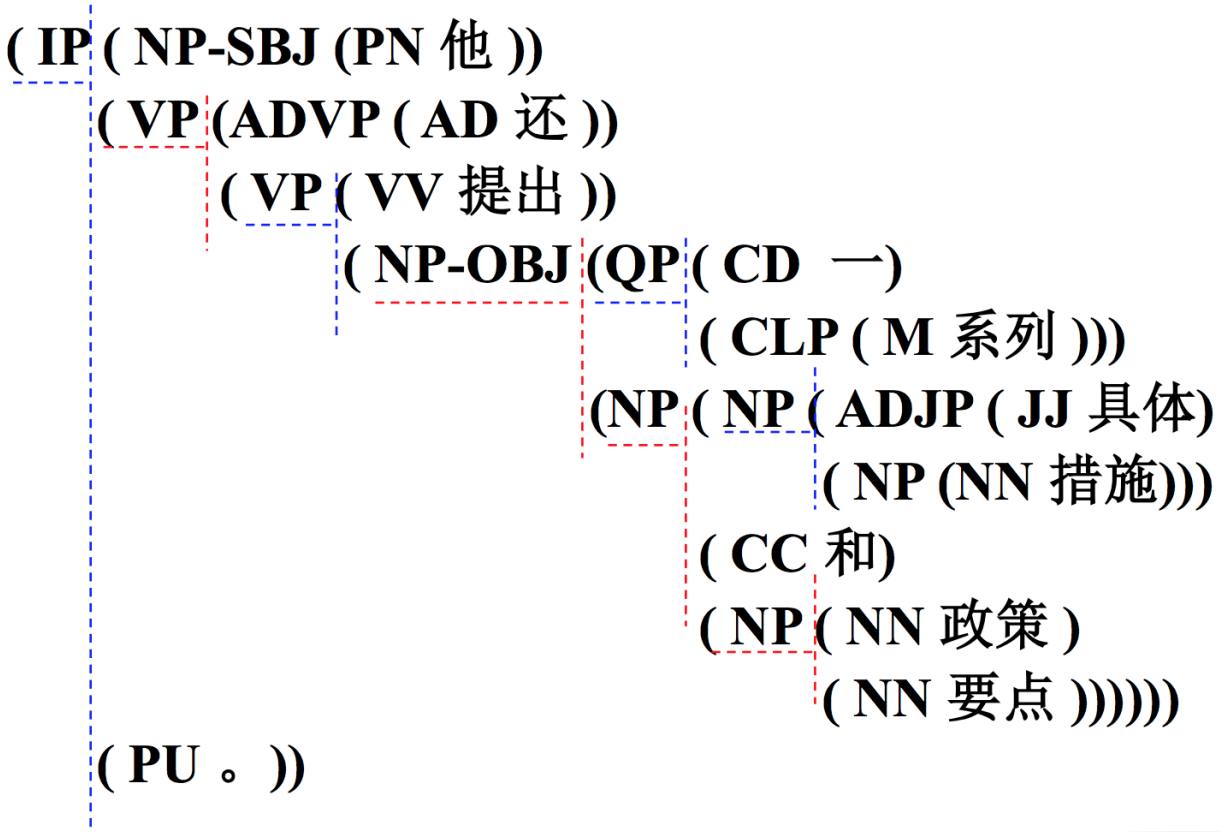
- 英国国家语料库（BNC）
  - 英国国家语料库（British National Corpus）是目前世界上最具代表性的当代英语语料库之一
  - BNC语料库书面语与口语并重，其规模超过一亿词，其中书面语语料占约90%，主要来自于报纸、期刊（包括学术期刊）、学术书籍、流行小说、信件、备忘录、论文和其他类型的文本
  - 口语语料约占百分之十，主要包括非正式对话（从不同年龄，地区和社会阶层中选出的志愿者记录）和在不同背景下收集的口语，从正式的商业或政府会议到电台节目和电话
  - BNC语料库第一版构建从1991年开始到1994结束，2001年的第二版（*BNC World*）和2007年的第三版（*BNC XML Edition*）均做了轻微的改动，并没有加入新的文本
  - BNC语料库是单语语料库，语言为当代英国英语，BNC也是共时语料库，主要收录二十世纪末期的语料
  - 官方网址： http://www.natcorp.ox.ac.uk/

典型语料库介绍

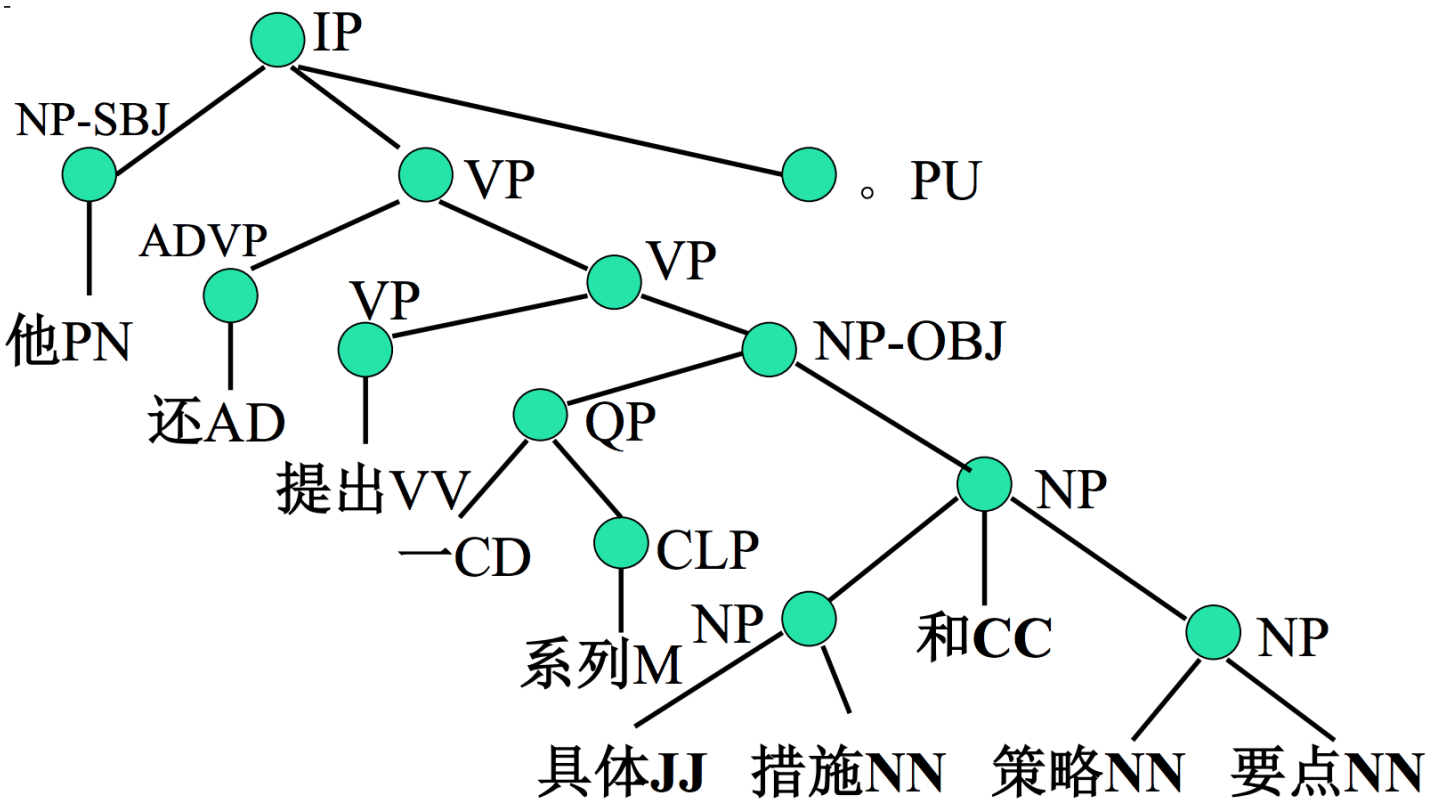
- 宾州(Pennsylvania)大学语料库 (UPenn Tree Bank)
  - 美国宾州大学计算机系 M.Marcus 教授主持
  - 1993年完成约300万词次英语句子的语法结构标注
  - 2000年完成第一版中文树库，约10万词次，4185个句子

例子：原始句子：他还提出一系列具体措施的政策要点。 词性标注：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

典型语料库介绍


(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 。))

典型语料库介绍


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

典型语料库介绍

- 北京大学语料库
  - 北大计算语言学研究所俞士汶教授主持，北大、富士 通、人民日报社共同开发
  - 《人民日报》1998年上半年全部文本（约1700万字）
  - 100万字切分及词性/注音标注
  - 完整的词语切分和词性标注信息

例子： 咱们/r 中国/ns 这么/r 大/a 的/u 一个/m 多/a 民族/n 的/u 国 家/n 如果/c 不/d 团结/a ，/w 就/d 不/d 可能/v 发展/v 经济/n ，/w 人民/n 生活/n 水平/n 也/d 就/d 不/d 可能/v 得到/v 改善/vn 和/c 提高/vn 。/w

典型语料库介绍

- 台湾中研院平衡语料库
  - 台湾中央研究院平衡语料库（Sinica Corpus）：世界上 第一个带有完整词类标记的汉语平衡语料库
  - 目标：500万词次汉语平衡语料库
  - 设计思想：
    - 遵循台湾计算语言学会的分词标准
    - 采样时以自然段落为准，不看文章长度
    - 语料采用多重分类法

典型语料库介绍

- Chinese LDC
  - 国家 973 项目资助（图象、语音、自然语言理解与知识 挖掘，编号：G1998030504）
  - 语音，文字（口语，书面语）
  - 单语：分词及词性标注语料，树库语料
  - 双语：汉英句子对齐
  - 规模：
    - 汉语通用词表 ：8－10万词
    - 汉语信息词典：2.5-3.0 万词
    - 分词词性标注语料：500万字
    - 汉语句法树库：100万字 … …

典型语料库介绍

- LC-STAR 项目 (NLPR-Nokia)
  - 14 国语言：英文、俄语、中文、西班牙语 …
  - 文本语料不少于100M words （中文约3000万字）
  - 领域：新闻 612万字，19%、财经418万字，14％、 文化娱乐374万字，12％、体育829万字，27％、 消费 499万字，16％、个人通讯 355万字，12％ 共计约：3087 万字
  - 抽取常用词汇：4.5 万词
  - 另外收集专用词汇：5000词
  - 人名：5 万个
  - 词典标注：拼音、词性等

国内外免费可用语料库

- 国外语料库

类型	时间	容量	语料	说明
SEU	1959年起	100万	书面语50% 口语50%	第一个大型计算机语料库
LLC	1975-1981	50万	口语	以计算机自动化处理方式获取SEU语料库的英语口语原始语料
BROWN	1960s	100万	书面语	用于研究当代美国英语
LOB	1970s	100万	书面语	用于研究当代英国英语
COBUILD	1980s	3.2亿	书面语75% 口语25%	在语料库支持下从事词典学研究
LONGMAN	1988-1990	2800万	书面和口语	编纂词典和供学术界使用
BNC	1991-1995	1亿	书面语90% 口语10%	其口语语料库可以精细分析语音研究
ICE	1988年起	2000万	书面语40% 口语60%	对讲英语的不同国家的英语进行对比研究

国内外免费可用语料库

- 国家语委
  - 现代汉语语料库http://www.cncorpus.org/
  - 古代汉语语料库http://www.cncorpus.org/login.aspx
- 北京大学计算语言学研究所
  - 《人民日报》标注语料库http://www.icl.pku.edu.cn/icl_res/
- 台湾中央研究院
  - 现代汉语平衡语料库http://www.sinica.edu.tw/SinicaCorpus
  - 古汉语语料库http://www.sinica.edu.tw/ftms-bin/ftmsw
  - 近代汉语标记语料库http://www.sinica.edu.tw/Early_Mandarin/
  - 树图数据库http://treebank.sinica.edu.tw/

国内外免费可用语料库

- 香港教育学院
  - 语言资讯科学中心及其语料库实验室 http://www.livac.org/index.php?lang=sc
- 中文语言资源联盟
  - 中文语言资源联盟http://www.chineseldc.org/
- 哈尔滨工业大学
  - 哈工大信息检索研究室对外共享语料库资源 http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm
- 中国传媒大学
  - 文本语料库检索系统http://ling.cuc.edu.cn/RawPub/

关于语料库的统计知识

- Zipf法则
  - 统计一种语言中所有的词在一个大型语料库中的出现次数，并出现次数大小排列顺序
  - 词出现的频率f和它的排列位置r之间的关系：

$$f \propto \frac{1}{r}$$

  - 或：

存在一个常数 k 使得 $f \cdot r = k$

关于语料库的统计知识

- Zipf法则
  - 例如，可以这样说，排在第50位的词出现次数大约是排在第150位的词的出现次数的3倍
  - Estoup在1916年首先发现了这种出现次数和排列位置之间的关系，但是Zipf正式定义并推广了这种关系，并以他的名字命名了这个关系
  - 不把这种关系看做一个规则，而是作为某些试验事实的一个比较粗糙的特性

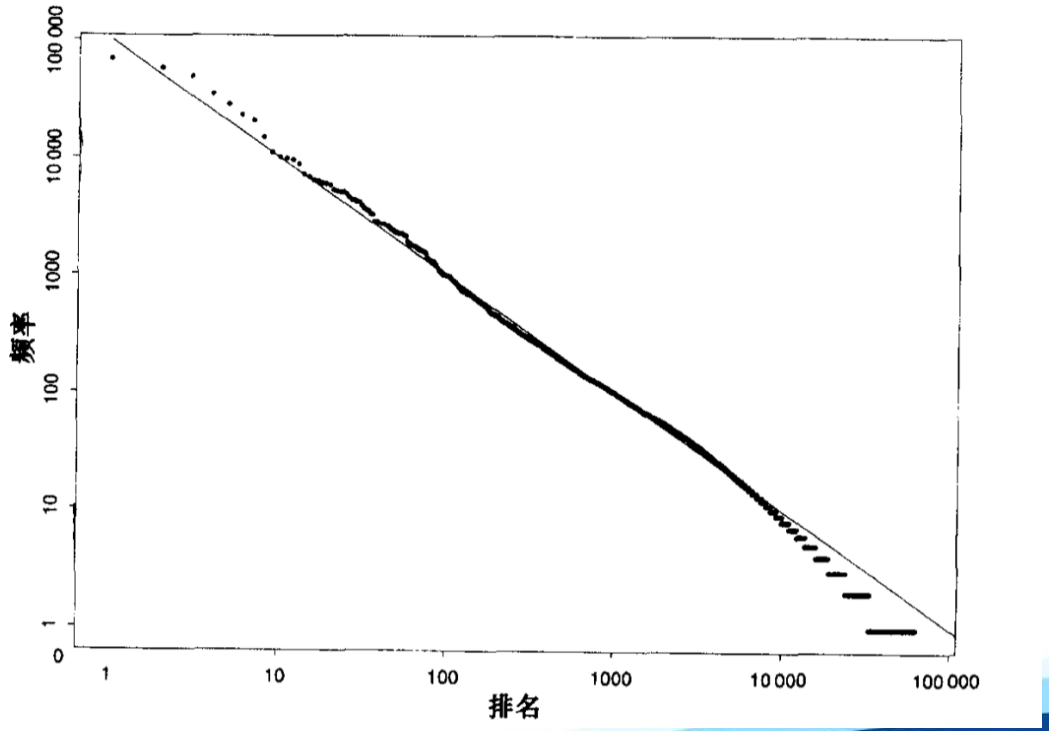
关于语料库的统计知识

- Zipf法则
  - Zipf法则在《汤姆索亚历险记》语料库的实验

单词	频率(f)	排名(r)	$f \cdot r$	单词	频率(f)	排名(r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10 200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10 440	Could	2	4000	8000
two	104	100	10 400	Applausive	1	8000	8000

关于语料库的统计知识

- Zipf法则
  - Zipf法则在Brown语料库的实验



关于语料库的统计知识

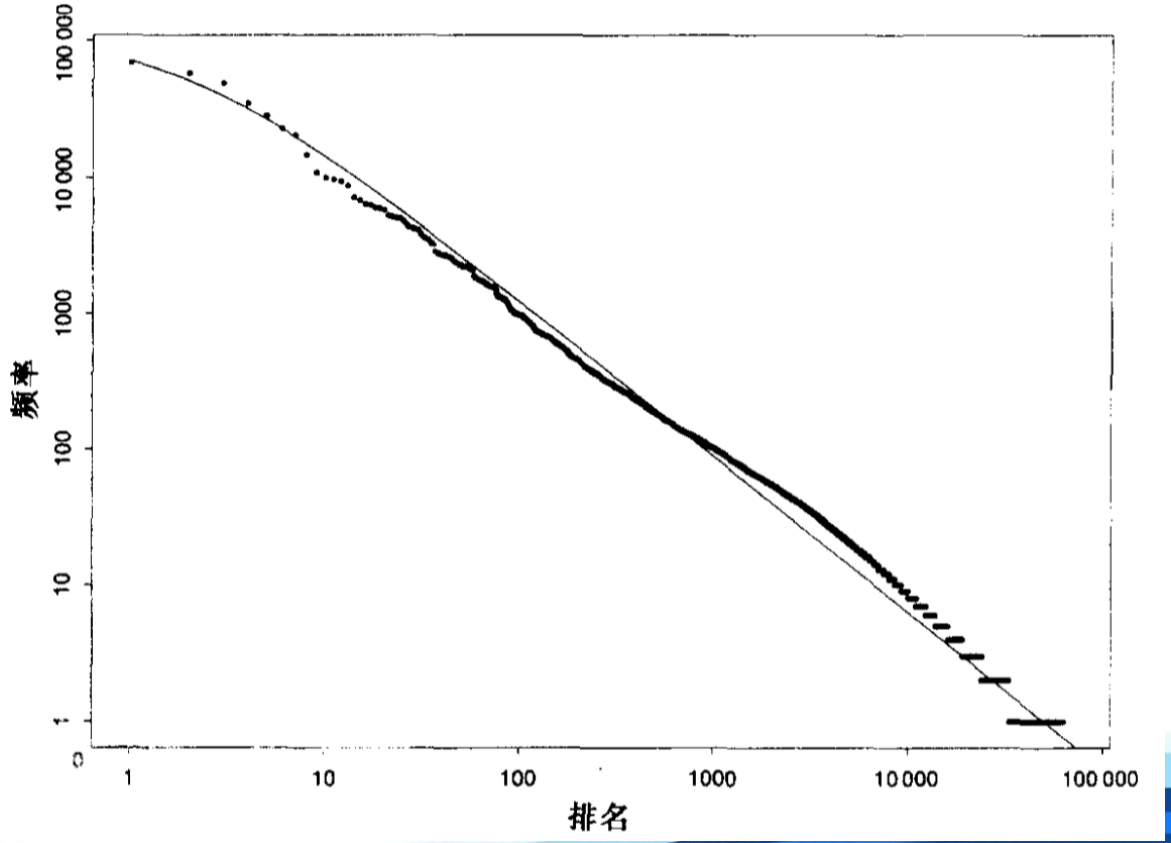
- Zipf法则
  - 为了得到更加接近词汇经验分布的结果，Mandelbrot得出了下面的更一般的排列和出现次数的关系：

$$f = P(r+\rho)^{-B} \quad \text{或} \quad \log f = \log P - B\log(r+\rho)$$

  - 其中P, B和ρ是文本的参数，他们总体衡量了文本中词汇使用的广度
  - 如果设置参数B=1, ρ=0，Mandelbrot公式就简化为Zipf法则

关于语料库的统计知识

- Mandelbrot公式在Brown语料库上的实验



语料库研究-研究方向

- 语料库的建设与编纂
  - 出发点：如何使得在其基础上展开的语言调查是合理可靠的。
  - 因此Kennedy(1998)指出了语料库设计师所面临的基本问题：这个语料库所采集的语言数据是否真正代表了某种期望的语言或语体
  - 在语料库的建设和编撰过程中因考虑的问题包括：
    - 静态与动态
    - 代表性和平衡性
    - 规模

语料库研究-研究方向

- 语料库的加工和管理技术
  - 主要是指用于语料分析、标注、维护和检索的工具
  - 语料库不仅仅是文本的集合，它应具有良好的存取性能，以便研究人员能从中检索自己需要的信息
  - 语料库检索工具----AntConc
    - 下载：http://www.laurenceanthony.net/software/antconc/
    - 视频教程：https://www.youtube.com/playlist?list=PLiRIDpYmiC0Ta0-Hdvc1D7hG6dmiS_TZj
    - 支持的检索方式：
      - 逐词索引（concordance）
      - 词簇（cluster）
      - 搭配（collacates）
      - 词表（word list）
      - 关键词表（keyword list）

语料库研究-研究方向

- 语言研究中语料库的使用
  - 言语研究
    - 语言学理论
    - 语言史研究
    - 句法、词法及自动语法分析
  - 词汇研究
  - 语义学
  - 语用学和话语分析
  - 社会语言学
  - 心理语言学
  - 外语教学
- 在计算语言学中的应用

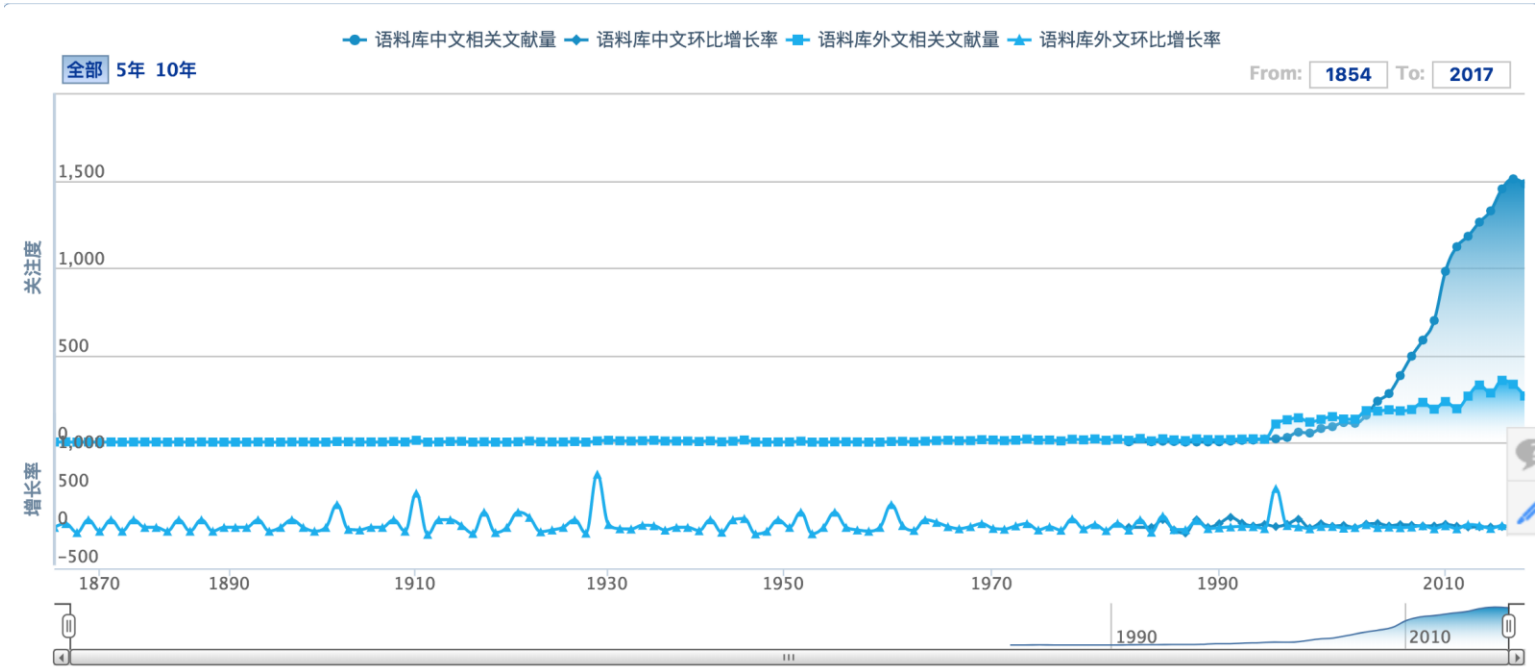
语料库研究

- 近几年语料库研究数量统计(CNKI)

年份	数量
2018（1.1-9.21）	467
2017	1156
2016	772
2015	676
2014	633

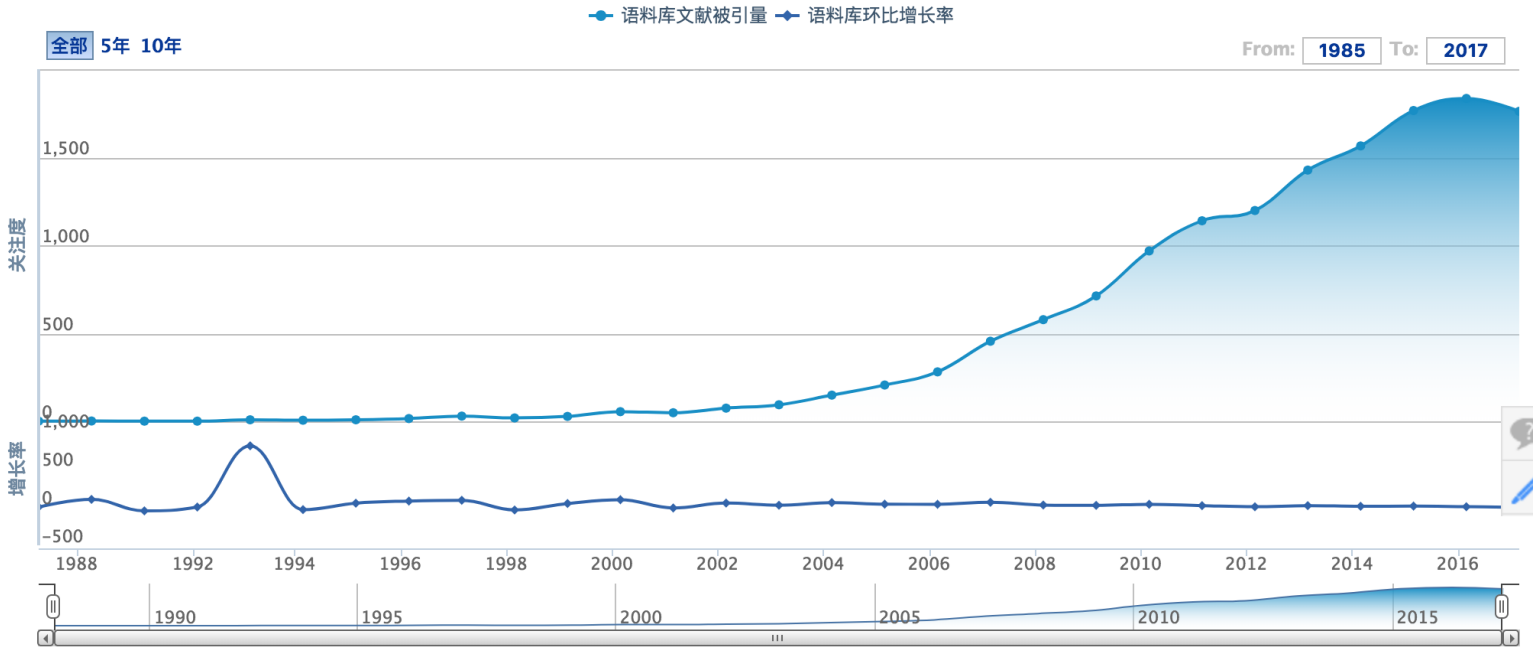
语料库研究

- 语料库学术关注度

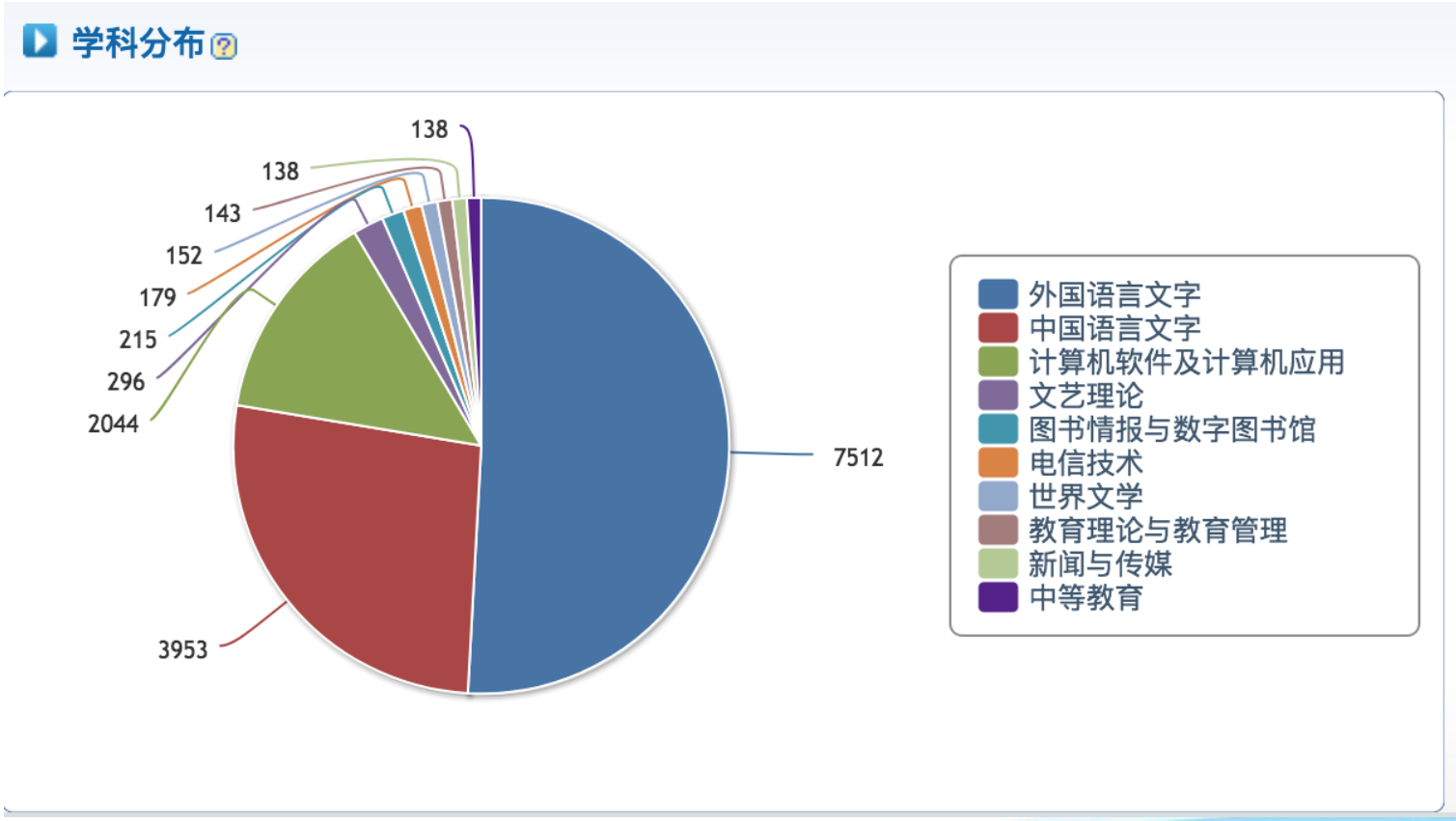


语料库研究

- 语料库学术传播度

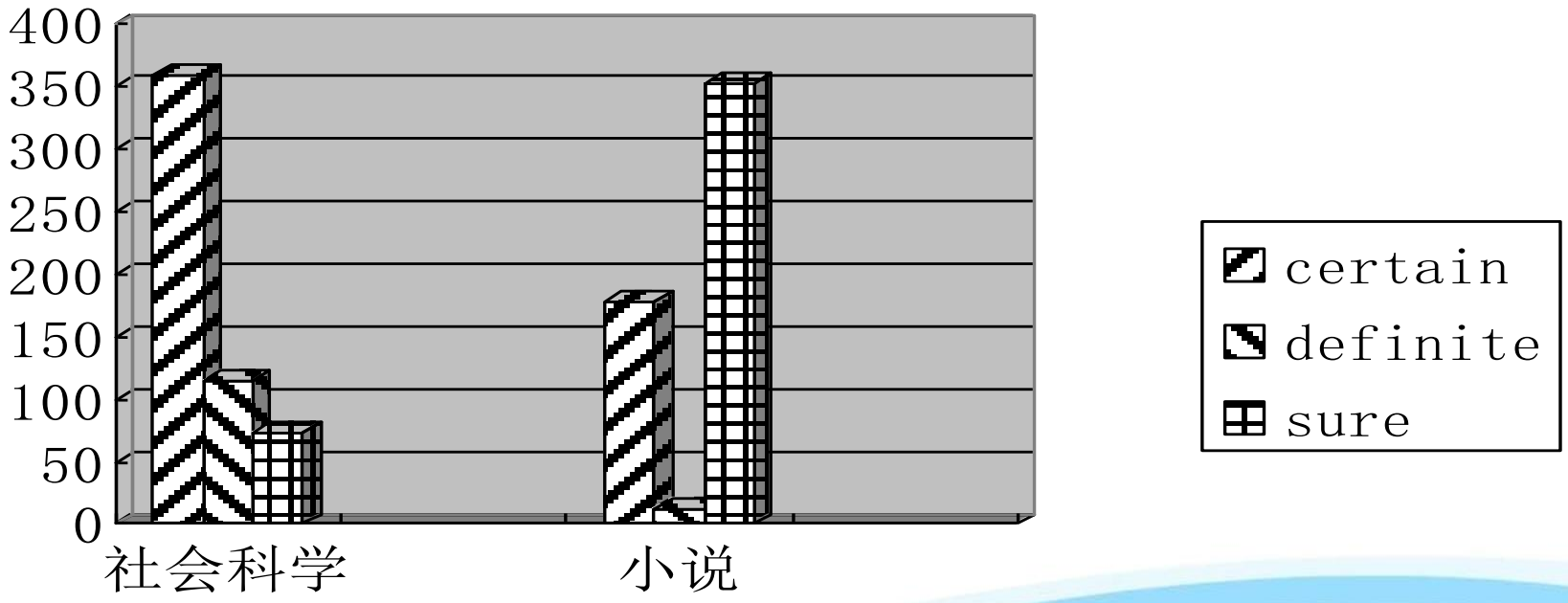


语料库研究


学科分布
138
138
143
152
179
215
296
2044
7512
3953
外国语言文字
中国语言文字
计算机软件及计算机应用
文艺理论
图书情报与数字图书馆
电信技术
世界文学
教育理论与教育管理
新闻与传媒
中等教育

实例—运用语料库进行外语研究

- 不同语域词频差异的调查
  - certain, sure, definite在社会科学和小说中的频率分布图



实例—运用语料库进行外语研究

- 在社会科学类文本中，用的最多是certain（1百万词中359次），其次是definite（114次），sure最不常见（74次）；而在小说类型的文本中，sure比certain要常见得多（353次对179次），而definite就极为少见（仅11词次）。这说明在表达比较严谨的文本中，更倾向于使用certain和definite，而在表达相对自由的小说中，较为口语化的sure用得更多。不同语域同义词的取舍有一定的指导意义，形成一定的优先原则。如在社会科学类的文本中，可优先考虑certain，其次为definite；而在小说中，则sure将是首选，其次才为certain。

实例—运用语料库进行外语研究

- 根据搭配调查语义差异
  - take a job 和take on a job

1)to pay off, she cannot now	**take**	a job **paying** less than pounds 12,000 a year.
2) iver. He is now leaving to	**take**	a job in Brussels as a **European commissioner**.
3) a kitchen assistant before	**taking**	a job as a **pizza delivery driver** 18 months a
4)x years. Three years ago I	**took**	a **part-time** job and have received my tax allow.
5)eir boy to be a lawyer. He	**took**	a **job** with the Ministry of the Interior but sp
6)se neuroses. At 16, Moore	**took**	a **summer job** working on the chassis line at GM
7)er moving to New York, she	**took**	a **modeling job** and, while doing an ad for Oli
8)block any move for him to	**take**	another job in football." Little would see a r

实例—运用语料库进行外语研究

观察take a job索引例句的搭配，尤其是右搭配，发现与它共现的词有：

- (1)工作类别：as a European commissioner，as a pizza delivery，with the Ministry of the Interior，modeling，in football
- (2)工作时间：part-time，summer
- (3)工作报酬：paying

由此可以看出，take a job多指“干什么样的具体工作”，与之相关的有“工作付多少报酬，工作是全职或兼职”等，核心意思是“就业”。

实例—运用语料库进行外语研究

同样观察take on a job的索引例句可以看到与其共现的词语有：

- (1)工作内容（并非职业）：scrapping excess capacity, compiling the electoral register, defending, grain preparation
- (2)工作压力：stressful job-loads，demanding，stress-loaded
- (3)无报酬：unpaid
- 可见，take on a job多表示“把责任赋予某项工作，不管有无报酬”，其它未在此列出的搭配词，还有诸如role(s), responsibility/ies, task(s), work, commitment(s), burden(s), challenge(s)等，它们都显示出take on a job的核心在于“责任”。

实例—运用语料库进行外语研究

- 根据搭配调查语义韵差异：cause和lead to
  - cause多与表示疾病、伤害、不佳情绪、问题、困难等含义的词语一起出现，几乎全含有否定和消极的意味，这说明cause导致的基本都是坏的结果，语义韵特征上倾向于否定和消极。
  - rash, greater injury, complaint, irritation, severe embarrassment, shortages, initial problems, fluid retention, styling problems, difficulty, fatal problems, all sorts of havoc, anxiety, slowdown in deficiency disease, more violence, later harm, poor weather, a host of problems, terrible damage, heart attack, tension, cancer damage, trouble, sorrow, confusion, lack.

实例—运用语料库进行外语研究

- 根据搭配调查语义韵差异：cause和lead to
  - 而对lead to 搭配的观察，则没有发现明显的倾向性：
  - 其客体既可是肯定的，如“notable improvement, great successes, permanent opportunities, new developments, professional qualification, improved human health, happiness, formation, specification”等；
  - 也可是否定的，如“loss of life, more problems, unfair advantage and conflict, immediate withdrawal, drug taking and crime, anxiety attacks, serious problems, water loss and damage, scarring, holes, prosecution increased risk”等。
  - 就整体分布而言， 两者几乎平分秋色。从这一点来看，lead to不存在语义韵的显著差别，既可引起好的结果，也可导致坏的结果。

汉语语料库建设中的问题

- 语料库建设的规范问题
  - 如果没有公认的、统一的语料库加工规范，语料库的建设和利用势必会受到严重制约
  - 语料资源长期无法共享，大量语料库处于小规模、低水平、重复建设状态
- 现有规范
  - 信息处理用GB13000.1字符集汉字部件规范　国家语委（1997.12.5）
  - GB12200.1—90汉语信息处理词汇01部分：基本术语　国家技术监督局（1993）
  - GB/T12200.2—94汉语信息处理词汇02部分：汉语和汉字　国家技术监督局（1994）
  - GB13715信息处理用现代汉语分词规范

汉语语料库建设中的问题

- 产权保护和国家语料库建设问题
  - 产权保护两个方面
    - 文本的知识产权
      - 1990年9月7日《中华人民共和国著作权法》
    - 语料库的知识产权
      - 至今在著作权法、语言文字法、计算机软件保护等相 关法规和实施条例中有关语料库知识产权的条款都是空白
  - 国家语料库
    - 国家语料库的建设、开发、保护应该是一种国家行为，在信息社会和数字化生存时代，我们要把语言资源的收集、 保护、开发提高到一种对待国家资源的高度来认识

参考阅读

- Biber, Douglas, Susan Conrad & Randi Reppen. 1998. Corpus Linguistics. Cambridge: Cambridge University Press. （外研社引进）
- Granger, S. et al. (eds.). 2003. Corpus-based Approaches to Contrastive Linguistics and Translation Studies《基于语料库的语言对比和翻译研究》. Amsterdam: Rodopi. （外研社引进）
- Gries, Stefan Thomas. 2004. Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement. Beijing: Peking University Press. （北大出版社引进）
- Hunston, Susan. 2002. Corpora in Applied Linguistics. Cambridge: Cambridge. University Press. （世界图书出版社引进）
- Kennedy, Graeme. 1998. An Introduction to Corpus Linguistics. London: Longman. （外研社引进）

参考阅读

- Nattinger, James R. & Jeanette S. DeCarrico. 1992. Lexical Phrases and Language Teaching. Oxford: Oxford University Press. （外教社引进）
- Sinclair, John. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press. （外教社引进）
- Thomas, Jenny & Mick Short. 1996. Using Corpora for Language Education. London: Pearson Education. （外研社引进）
- Zanettin, F., et al. (eds.). 2003. Corpora in Translator Education《语料库与译者培养》. Manchester: St. Jerome Publishing. （外研社引进）

参考阅读

- 蔡金亭，2003，《语言因素对英语过渡中使用——一般过去时的影响》。北京：外语教学与研究出版社。
- 何安平（主编），2004，《语料库在外语教育中的应用：理论与实践》。广州：广东高等教育出版社出版。
- 何安平，2004，《语料库语言学与英语教学》。北京：外语教学与研究出版社。
- 华南师范大学外国语学院编，2005，《语料库语言学的研究与应用》。长春：东北师范大学出版社。
- 黄昌宁，李涓子著，2002，《语料库语言学》。北京：商务印书馆。
- 濮建忠，2003，《学习者动词行为：类联接、搭配及词块》。开封：河南大学出版社。
- 王建新，2005，《计算机语料库的建设与应用》。北京：清华大学出版社。

参考阅读

- 王克非等，2004，《双语对应语料库研制与应用》。北京：外语教学与研究出版社。
- 王立非、梁茂成等，2007，《计算机辅助第二语言研究方法与实用》。北京：外语教学与研究出版社。
- 卫乃兴，2002，《词语搭配的界定与研究体系》。上海：上海交通大学出版社。
- 卫乃兴，李文中，濮建忠等，2005，《语料库应用研究》。上海：上海外语教育出版社。
- 文秋芳、王立非、梁茂成，2005，《中国学生英语口笔语语料库》。北京：外语教学与研究出版社。
- 杨达复，2000，《英语错误型式分析》。西安：陕西人民出版社。
- 杨惠中、桂诗春，2003，《中国学习者英语语料库》。上海：上海外语教育出版社。
- 杨惠中、卫乃兴，2005，《中国学习者英语口语语料库建设与研究》。上海：上海外语教育出版社。
- 杨惠中等（主编），2005，《基于CLEC语料库的中国学习者英语分析》。上海：上海外语教育出版社。
- 杨惠中主编，2002，《语料库语言学导论》。上海：上海外语教育出版社。

THANKS