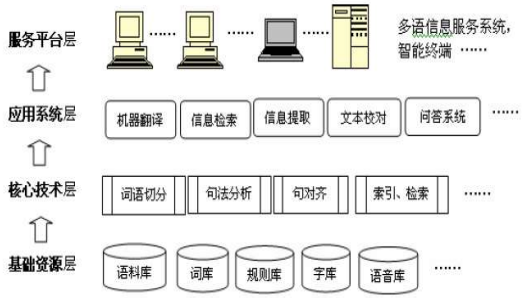
第二章 语料库



- 什么是语料库
- 语料库的类型
- 语料库的基本加工过程与规范
- 语料库的应用
- NLTK

2.1什么是语料库

- 语料库

  存放语言材料的数据库（文本集合），其复数形式为Corpora。库中的文本通常经过整理，具有既定的格式和标记，特指计算机存储的数字化语料库

- 语料库的基本认识
  - 存放的是在语言实际使用中真实出现过的语言材料
  - 以计算机为载体承载语言知识的基础资源
  - 真实语料需要经过加工（分析与处理）

2.1什么是语料库语言学

- 语料库语言学
  - 根据篇章材料对语言的研究称为语料库语言学。[Aijmer, 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。一[McEnery, 1996]
  - 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。[Crystal, 1991]

语料库语言学主要研究机器可读自然语言文本的采集、存储、检索、统计、语法标注、句法语义分析，以及具有上述功能的语料库在语言定量分析、词典编纂、作品风格分析、自然语言理解和机器翻译等领域中的应用。

2.1语料库语言学

- 语料库语言学研究的内容:
  - 语料库的建设与编纂
  - 语料库的加工和管理技术
  - 语料库的使用

语料库语言学的历史

- 第一代（70-80）
  - 百万词级
  - 以语言研究为导向
- 第二代（80-90）
  - 千万词级
  - 词典编纂-应用为导向
- 第三代（1990-）
  - –超大规模（上亿词级）
  - –标准编码体系
  - –深度标注/多语种
  - –NLP应用
- 第四代（？）
  - 互联网作为语料库

第一代语料库(1)

- Brown语料库
  - 始建于1960年代初
  - W.N.Francis和H.Kucera发起
  - 美国Brown大学建立
  - 世界上第一个根据系统性原则采集样本的标准语料库
  - 主要代表当代美国英语
  - 规模100万词次

第一代语料库(2)

- LOB语料库
  - 始建于1970年代初
  - 由英国Lancaster大学著名语言学家Geoffrey Leech倡议
  - 挪威Oslo大学StigJohansson主持完成
  - 安装在挪威Bergen大学挪威人文科学计算中心
  - 规模于Brown语料库相当
  - 主要代表当代英国英语

第一代语料库(3)

- London-Lund语料库
  - 1960年代初，由Randolph Quirk主持
  - 收集2000小时的谈话和广播等口语素材并整理成书面材料
  - 由瑞典Lund大学J. Svartvik主持全部录入计算机
  - 1975年建成

第二代语料库(1)

- COBUILD语料库
  - 建于1980年代
  - 以词典编撰为应用背景
  - 有英国Birminghan大学与Collins出版社合作完成
  - 规模达2000万词次
  - 基于该语料库出版的Collins Cobuild词典（1987）受到了广泛的好评

第二代语料库(2)

- Longman语料库
  - 建于1980年代
  - 包括三个语料库
    - LLELC语料库（Longman/Lancaster英语语料库）
    - LSC语料库（Longman口语语料库）
    - •LCLE（Longman英语学习语料库）
  - 目标是编撰英语学习词典，为外国人学习英语服务
  - 规模达5000万词次

第三代语料库(1)

- ACL/DCI语料库
  - 美国ACL倡议发起
  - 收集语料范围广泛
    - •华尔街日报
    - •Collins英语词典
    - •Brown语料库
    - •PennTreeBank
  - 既有已标注的语料，也有未标注语料
  - 制定了语料库文件的格式标注
    - 采用统一的SGML标注语言
    - 语料标注依照TEI（Text Encoding Initiative）标准

第三代语料库(2)

- PennTreeBank（宾州大学树库）
  - 美国Pennsylvania大学1980年代末开始发起
  - 由该校计算机系M.Marcus主持
  - 1993年，完成了对近300万英语词的句子语法结构标注
  - 2000年完成了中文树库（第一版）：10万词次，4185个句子

中文语料库

- **中文语料库建设状况**

自1979年以来，中国就开始进行机器可读语料库的建设，早期在中国建立的主要机器可读语料库有：

汉语现代文学作品语料库（1979年），527万字，武汉大学
现代汉语语料库（1983年），2000万字，北京航空航天大学
中学语文教材语料库（1983年），106万8千字，北京师范大学
现代汉语词频统计语料库（1983年），182万字，北京语言学院

早期语料库有如下特点：

- 多数采用手工输入方式建立，耗时耗力，缺乏规范，规模较小，重用性差。
- 发现了汉语文本切分歧义的两种类型。即交集型歧义和多义组合型歧义。
- 建立了初步的分词规范。即GB13715《信息处理用现代汉语分词规范》。

中文语料库

1991年，国家语言文字工作委员会开始建立国家级的大型汉语语料库，计划期规模达到7000万字。

现在该课题已经结项，国建语委语言文字应用研究所成立了“汉语语料库深加工”的课题组，准备对国家级语料库的2000万字的核预料进行深加工，逐步把这个生语料库变为熟语料库。

1992年以来，大量的语料库在中国研究中文信息处理的单位建立起来，建设了大规模真实文本语料库的单位有：

《人民日报》光盘数据库；
北京大学计算语言学研究所；
北京语言文化大学
清华大学；
山西大学 等。

2.2语料库的类型

- 按照语料库的研究目的和用途
  - 通用的(systematic)
    - 充分考虑语料的动态和静态问题、代表性和平衡问题以及语料库的规模等问题。
  - 专用的(specialized)
    - 如:北美的人文科学语料库

2.2语料库的类型

- 按语言种类划分
  - 单语的
  - 双语的或多语的

    篇章对齐 / 句子对齐 / 结构对齐
- 是否标注?
  - 具有词性标注
  - 句法结构信息标注（树库）
  - 语义信息标注

两个术语:
- 生语料
- 熟语料

2.2语料库的类型

- 平衡语料库
  - 平衡语料库着重考虑语料的代表性与平衡性。
  - 语料采集的七项原则:语料的真实性、可靠性、科学性、代表性、权威性、分布性和流通性。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。[张普，2003]

2.2语料库的类型

- 平行语料库
  - 两种含义:一种是指在同一种语言的语料.上的平行，例如正在建立的“国际英语语料库”,共有20个平行的子语料库，分别来自以英语为母语或官方语言和主要语言的国家，如英国、美国、加拿大、澳大利亚、新西兰等。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。

2.2语料库的类型

➢ 另一种平行语料库是指在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

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?

… …

C: 早晨$_1$ 好$_2$!$_3$
E: Good$_2$ morning$_1$.$_3$

已有的双语资源

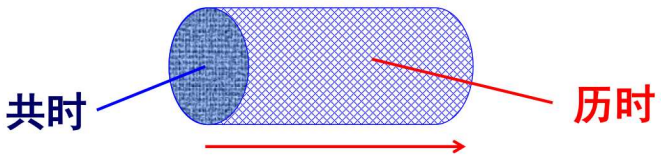
- 加拿大议会会议录 录 (Canadian Hansards) http://www.isi.edu/natural-language/download/hansard/
- 克姆尼茨英- 德翻译库 语料库 (Chemnitz E-G Translation Corpus) http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/transcorpus/index.htm
- 英语- 挪威语平行库 语料库 (ENPC) https://www.hf.uio.no/ilos/english/services/omc/enpc/
- 葡- 英双向平行库 语料库 (Compara) http://www.linguateca.pt/COMPARA/Welcome.html
- 香港立法委员会录 会议记录 (Hong Kong Hansards) http://catalog.ldc.upenn.edu/LDC2000T50
- 新闻 香港新闻 (Hong Kong News)
- 法律 香港法律 (Hong Kong Laws)

2.2语料库的类型

- 共时语料库与历时语料库.
  - 所谓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。

2.2语料库的类型

- 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



2.2语料库的类型

- 判断历时语料库的4 条原则 — [张普, 2003]
  - 是否 动态：语料库必须是开放的、动态的
  - 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， 具有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 。
  - 深加工是否基于动态的加工方法：随语料的动态变化采集， 并进行动态地加工 。
  - 是否取得动态的加工结果：语料的加工结果也应是动态的和历时的 。

2。3汉语语料库的基本加工规范

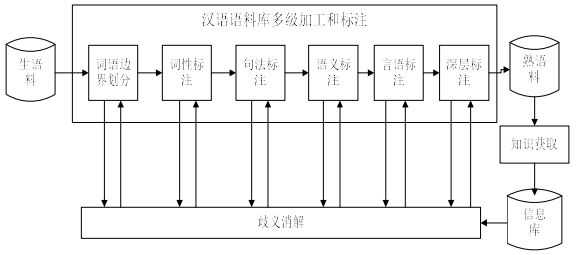
- 生语料与熟语料
- 汉语语料库的加工思路
- 汉语语料库的加工规范
- 汉语文本词性标注标记集

生语料与熟语料

文本通常有两种形式：生文本和标注文本。“标注”这个术语是指把某种分类代码插入到一个计算机文件中，这种分类代码通常并不是文件的组成部分，但是通过这些分类代码我们可以了解文件的结构或格式信息。

由于语料库的来源复杂语料中可能存在无法处理的各种各样的格式或者内容，如：文档页眉和分隔符、排版代码、表和图标等。因此在进一步对语料处理前需要一个过滤器来过滤掉这些无用内容。

2.3语料库加工过程



生语料与熟语料

制定《现代汉语语料库加工手册——词语切分与词性标注》的基本思路如下：

1.词语切分的规范应尽可能的同已有的中国国家标准GB13715《信息处理用现代汉语分词规范》保持一致。

2.词性规范使用小标记集。除了使用《语法信息词典》中的26个词类标记外，增加了专有名词分类标记、 语素g按其子类标注以及动词和形容词的某些功能标记。

3.与已有资源的配合。

汉语语料库加工规范

北大计算语言所制定的《现代汉语预料库加工——词语切分与词性标注规范》分为三个部分：**切分规范，标注规范，切分和标注相结合的规范。**

切分规范：主要规定将汉字串形式的句子切分为词序列的原则，即什么样的汉字组合可以作为一个切分单位。

进行切分通常要有一部**“分词词典”**。北大的《语法信息词典》收录的词条已超过7.3万，本规范规定词典的词条一般都是切分单位，这使得对“切分单位”的把握有了基本的参照。但处理大规模真实文本时总会遇到词典中没有的**“未定义词”**

汉语语料库加工规范

标注规范：标注规范包括一般词性的标注和专有名词的标注。

一般词性的标注即切分单位的标记。包括：

1.标记集以26个词类标记为基准，名动词、副动词 、名形词和专有名词的标记是在动词代码V 、形容词代码a 、名词代码n后增加一个小写字母，语素标记是在语素代码g前增加一个大写字母。

2.一个词语若在词法词典中已属于一个或若干个词类，标注时不要轻易增加词性。

3.当语法词典给某个词确定的词性确实不对或不完备时，当然也要修正或补充。

汉语语料库加工规范

4. 即使语法词典中的简称实际上是指团体、机构、组织名称或地名，标注时仍标以j，而不是改为nt或nz。
5. “唐朝” 、“宋代”等历史朝代虽然也是专名，因语法词典已作为时间收入，标注时仍标以t，不改为nz。

这里的专有名词标注，对“专有名词”含义进行了扩展。短语型的地名、团体机构名称及其他专有名词在词的切分基础上用ASCII的方括号括起来，并在右方括号后标注以相应的ns，nt，nz，方括号不嵌套。

汉语语料库加工规范

切分和标注相结合的规范：规范中还给出了一些基于词性描述的构词规则，规定了什么样的组合处理为一个切分单位，并给出了新组合词的词性。

由语素构造合成词的方式有“复合”、“附加”、“重叠”，但运用这三种方式将两个成分结合成一个较大的单位时，这个单位是否作为切分单位，不可一概而论。以“附加”方式的后接成分“者”为例：

“者”接在语素或词的后面构成合成词，自然为切分单位，标记/n，如：死者/n，笔者/n，当局者/n，旁观者/n，屡教不改者/n

“者”接在较长短语或句子后，却应分开，将“者”单独标记为/k，如：经过/d 苦苦/d 追求/v 而/c 获得/v 幸福/a 者/k 。

汉语文本词性标注标记集

北大计算语言所制定的标注集

表 5.1 中文词性标注集

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，语素代码为 g 前面置以 A
A	形容词	取英语形容词 adjective 的第一个字母
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 并在一起
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 n 并在一起
b	区别词	取汉字“别”的声母
c	连词	取英语连词 conjunction 的第一个字母
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
d	副词	取 adverb 的第 2 个字母，因其第 1 个字母已用于形容词
e	叹词	取英文叹词 exclamation 的第 1 个字母
f	方位词	取汉字“方”
g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母
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	习惯用语	习惯用语尚未成为成语，有点“临时性”，取“临”的声母
m	数词	取英语 numeral 的第 3 个字母
Ng	名语素	名词性语素。名词代码 n，语素代码 g 前面置以 N

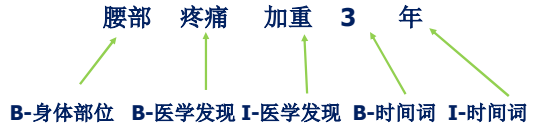
汉语文本词性标注标记集

表 5.1 中文词性标注集（续）

代码	名 称	帮助记忆的诠释
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 t，名词代码 n 和 t 并到一起
Nz	其他专名	“专”的声母的第一个字母 w 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 2 个字母
p	介词	取英语介词 prepositional 的第 1 个字母
q	量词	取英语 quanity 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	取英语组词 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 前置以 V
v	动词	取英语动词 verb 的第 1 个字母
Vd	副动词	直接做状语的动词。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	非语素词	非语素字只有一个符号，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

其他标注规范

- 如：实体抽取

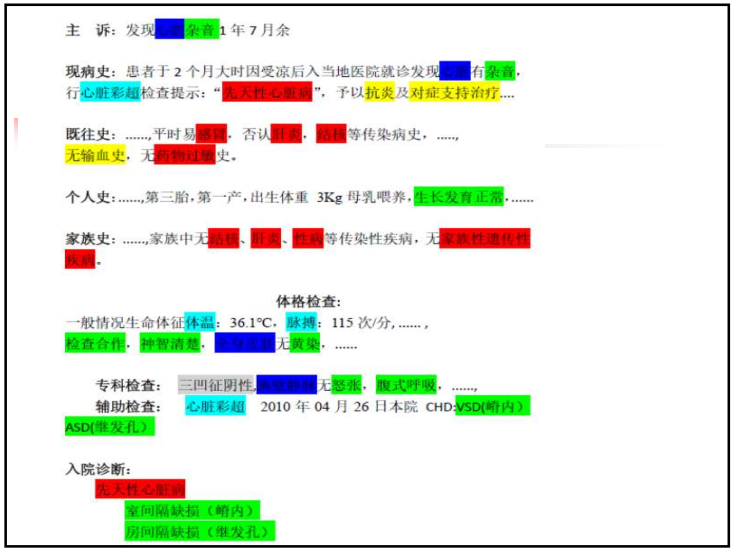


自定义规范

- **身体部位**（Body part）
- 标注时由蓝色字体标出，身体部位作为一类典型信息可以在实际应用中用于分析发病的部位，对临床研究和系统检索大有用处。为凸显部位的特征，这一实体在标注时不包含具体的方位，举例如下：

1) 口唇无发绀=（“口唇”）
2) 心尖搏动位于第4肋间左锁骨中线内0.5cm=（“心尖”）
3) 双肺叩诊音清=（）
4) 脊柱四肢正常=（“脊柱”，“四肢”）

- 这里有两点约定，第一点是在疾病、药物、检查/检验、治疗中的部位名称不标注为部位，而是将整体标注为疾病、药物、检查/检验或者是治疗，如“双肺叩诊”是检查方式、“肺结核”是疾病，这是因为很多都是固定搭配，如“肺结核”、“肝功能”等，或者是跟部位搭配一起时意义才能完整；第二点是当多个部位连在一起时若两个部位是包含关系或者只是固定名称时则连同标注，如“口唇”，“肝脾肋”,否则分开标注，如“脊柱四肢”。



主　诉：发现[心脏]杂音1年7月余

现病史：患者于2个月大时因受凉后入当地医院就诊发现[心脏]有杂音，行心脏彩超检查提示：“先天性心脏病”，予以抗炎及对症支持治疗....

既往史：.....,平时易感冒，否认肝炎、结核等传染病史，.....,无输血史，无药物过敏史。

个人史:.....,第三胎,第一产,出生体重 3Kg 母乳喂养,生长发育正常.....

家族史：.....,家族中无结核、肝炎、性病等传染性疾病，无家族性遗传性疾病。

体格检查:

一般情况生命体征体温：36.1℃，脉搏：115 次/分,.....,检查合作，神智清楚，[全身皮肤]无黄染，.....

专科检查：三凹征阴性,[呼吸节律]无急促，腹式呼吸，.....,

辅助检查：心脏彩超 2010 年 04 月 26 日本院 CHD:VSD(嵴内)ASD(继发孔)

入院诊断:

先天性心脏病

室间隔缺损（嵴内）

房间隔缺损（继发孔）

表 3.1 医学发现修饰举例

例句	实体	修饰		
		时间性	肯否性	正异性
到当地医院检查，发现心脏杂音	杂音	过去	肯定	异常
双侧直腿抬高试验（±）	（±）	现在	可能	正常
双膝反射未引出	引出	现在	否定	正常
否认血吸虫疫水接触，否认毒物接触史	血吸虫疫水接触，毒物接触	过去	否定	异常
B超:多发性肝囊肿	多发性肝囊肿	过去	肯定	异常

表 3.2 药物修饰举例

例句	实体	修饰	
		时间性	肯否性
自2017年在我院使用中成药治疗	中成药	过去	肯定
目前服用厄贝沙坦降压治疗	厄贝沙坦	现在	肯定
长期未使用支气管扩张药物	支气管扩张药物	过去	否定

其他标注规范

- 当前主要的标注工具：
  - Ehost
  - **中文文本标注工具Chinese-Annotator**

https://blog.csdn.net/m0epnwstyk4/article/details/79083806

2.4语料库的应用

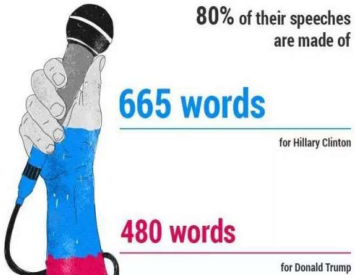
喜欢用短而常见的词，反复使用：真的（really），很好（nice），非常好（great），问题（problem）偏向于国际问题

倾向于较为罕见的词。希拉里•克林顿提到“美国”的次数大大多于特朗普：27：5。更加着眼于国内事物

2.4语料库的应用

- 所有的人都在说 “特朗普”

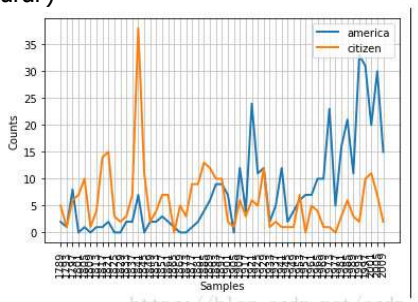
2.4语料库的应用



特朗普的演讲大概13%的词汇是独特的（全文7460个单词中有965个不同的主干词）。平均每个词重复7.7次。而希拉里有17%的独特词，每个词平均重复约6次。区别很明显：特朗普的演讲之80%只需要480个词，而希拉里需要665个词！多出来38%。

2.4语料库的应用

- import nltk
- nltk.download('inaugural')



语料库的应用

-词频统计与Zipf定律

词频统计

1998全年《人民日报》语料中频度最高的28个词（带词性）
北京大学计算语言学研究所俞士汶教授提供

word	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	year
的/u	54454	57239	59625	63097	62699	61038	65511	59153	63778	62656	63729	63823	736802
在/p	11447	12358	12078	13353	12746	12821	13446	12796	13870	13363	13520	12844	154642
和/c	10693	10930	12743	11829	12298	12240	13474	12180	12600	13054	12167	12629	147837
了/u	10235	11291	10682	12143	11654	11689	12628	11449	12387	12535	12799	12385	141784
是/v	10192	10894	12070	12369	12301	11816	11847	10810	11328	11870	11916	11656	138769
一/m	7085	7392	7355	8108	7811	7569	8055	7288	7812	7256	7383	6939	90053
不/d	4836	5197	4928	5709	5810	5483	5426	5000	5444	5163	5427	4967	62691
有/v	4549	4723	5276	5654	5139	5154	5404	4998	5263	4714	5359	4677	60910
对/p	3683	4227	4586	4803	4725	4830	4985	4348	4489	4482	4703	4532	54393
中国/ns	3356	3914	3753	3917	4314	4121	4022	3198	3918	4946	4882	5288	49629
中/f	3236	3400	3674	3808	3711	3543	4049	3658	3914	3837	3769	3816	44415
这/r	3205	3608	3617	4030	3792	3673	3747	3570	3815	3619	4357	3289	44119
要/v	2943	3039	4630	3338	3489	3450	3656	3503	3435	3255	3137	3273	41148
上/f	3017	3142	3204	3561	3370	3250	3658	3190	3533	3574	3669	3568	40736
他/r	2826	3793	3034	3584	3555	3036	3303	3112	3751	3846	3382	2420	39641
也/d	2776	2927	3098	3292	3306	3175	3106	3024	3184	3063	3251	3093	37295
人/n	2702	2711	2999	3413	3057	2910	3188	3060	3269	2981	3044	2795	36129
个/q	2633	2507	2865	3213	2912	3007	3237	3046	3015	3213	3164	2957	35769
等/u	2717	2642	2746	2966	2963	2876	3378	2883	2966	3063	3097	2936	35222
说/v	2583	2845	3466	2911	3096	2975	2851	2549	2833	2874	2930	2441	34354
经济/n	2597	2431	3094	2718	2415	2603	2601	2602	2685	3106	3497	3480	33829
就/d	2282	2633	2522	2996	2681	2643	2763	2753	2805	2739	2806	2320	31943
工作/vn	2120	2488	3474	2267	2540	2685	2680	2786	2822	2533	2588	2849	31802
年/q	2416	2365	2890	2699	2353	2407	2661	2122	2577	2799	2900	3444	31633

齐普夫（Zipf）定律

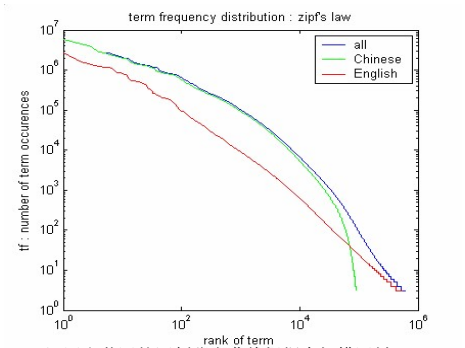
- 齐普夫定律是美国学者G.K.齐普夫于本世纪40年代提出的词频分布定律。它可以表述为：

  如果把一篇较长文章中每个词出现的频次统计起来，按照高频词在前、低频词在后的递减顺序排列，并用自然数给这些词编上等级序号，即频次最高的词等级为1，频次次之的等级为2，……，频次最小的词等级为D。

  若用f表示频次，r表示等级序号，则有：

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

齐普夫（Zipf）定律－实验



汉语和英语的词频分布曲线根据大规模语料库进行自动分词产生

词语分布均匀度

1998年《人民日报》语料库部分词语分布
北京大学计算语言学研究所俞士汶教授提供

Word	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Year
编织袋/n	2	2	3			6	12	109	23		5	2	164
舰长/n	3	9					75	36			1	2	126
九江/ns	1	4	11	6	5	12	60	230	83	52	11	4	479
九江市/ns	2	3	8	1	2	4	16	59	34	5	4	0	138
抗洪/v		2	3	4	5	10	145	1125	620	317	44	45	2320
抗洪/vn	1	5	1	1	7	10	211	2004	1326	852	140	66	4624
抢险/v	20	8	15	10	8	11	98	807	411	220	12	32	1652
抢险/vn	10	5	4	17	9	11	105	893	497	281	22	19	1873
水情/n				1		3	16	61	28	14	2	3	128
滔滔/z	3	4	1	3	2	1	16	65	46	27	6	5	179
新加坡/ns	42	80	38	37	41	40	60	57	41	71	89	31	627

题目

- 到http://www.icl.pku.edu.cn下载北京大学《人民日报》切分标注语料库（1个月），研究汉语“动词＋名词”可能构成哪些歧义结构？