**1.3.1 基本概念**

语料库中存放的是在语言的实际使用中真实出现过的语言材料；语料库是以电子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](https://baike.baidu.com/item/%E5%8A%A0%E5%B7%A5/7353178" \t "https://baike.baidu.com/item/%E8%AF%AD%E6%96%99%E5%BA%93/_blank)（分析和处理），才能成为有用的资源。

**1.3.2 语料库类型**

根据不同的划分标准，语料库可以分为多种类型。

1. 通用语料库与专用语料库

通用语料库是指按照事先确定好的某种重要标准，把每个子类的文本按照一定比例收集到一起的语料库，在语料采集过程中需要仔细从各个方面考虑平衡问题（包括领域分别、地域分布，时间分析，语体分布等）。

专用语料库是指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、特定语体类型的语料构成的语料库，如新闻语料库、科技语料库、中小学语料库等。

2. 单语语料库与多语语料库

语料库根据它所包含的语言种类的数目分为单语语料库、双语语料库和多语语料库。双语和多语语料库按照语料的组织形式，还可以分为平行（对齐）语料库和比较语料库。

平行语料库指库中的两种或多种文本互相是对方的译文，因此可以用于翻译后者机器翻译研究。

比较语料库中两种或多种语言的文本不构成对译关系，只是邻域相同，主题相近，通常只能用于两种或多种语言的对比。

3. 共时语料库与历时语料库

共时语料库是指为了对语言进行共时研究而建立的语料库。中文五地共时语料库就是典型的共时语料库，由香港城市大学开发采集中国内地、中国香港、中国台湾、中国澳门地区及新加坡在1995~2005年10年内的报纸语料，每四天选一天的报纸，包括社论、头版、国际和地方版、特写、评论等内容，早期每天各地均采集二万字左右，其后同步增加至每天三四万字。

历时语料库是对语言文字的使用进行动态追踪，对语言的发展变化进行检测的语料库，也称为第三代语料库。它有两大特色：

1. 语料的动态性：语料是不断动态补充的
2. 具有量化属性“流通度”，可以通过观察和测量流通度的变化情况，从而追踪语言成分的产生、成长和消亡。

4. 生语料库与熟语料库

生语料库是指没有经过任何加工处理的原始语料数据。熟语料库是指经过加工处理、标注了特定信息的语料库。所谓语料库标注（加工）就是对电子语料（包括书面语和口语）进行不同层次的语言学分析，并添加相应的“显性”的解释性的语言学信息过程。或者说是把某种分类代码插入到计算机文件中，这种分类代码通常不是文件的组成部分，但是通过这些分类代码，可以了解文件的结构或者格式信息。

**1.3.3 典型语料库介绍**

1. Brown语料库

2. LOB语料库

3.Penn TreeBank

4.PropBank

5.NomBank

6.FrameNet

7.The Canadian Hansards语料库

8.LC-STAR语料

9. C-STAR口语语料

10.北京大学语料库

11.LDC中文树库

**1.3.4 语料处理的基本问题**

1.汉语方面

西方语言之间词与词之间用空格分开。汉语不实行按词连写，因此，在进行词法分析之前，首先要进行汉语自动分词

1. 英文方面

（1）空格围起了多个词

词+标点

词+单撇号

由连字符连接多个单词

（2）空格不是分界标志

电话号码，专有名词等