

国家语委现代汉语通用平衡语料库

标注语料库数据及使用说明

肖航

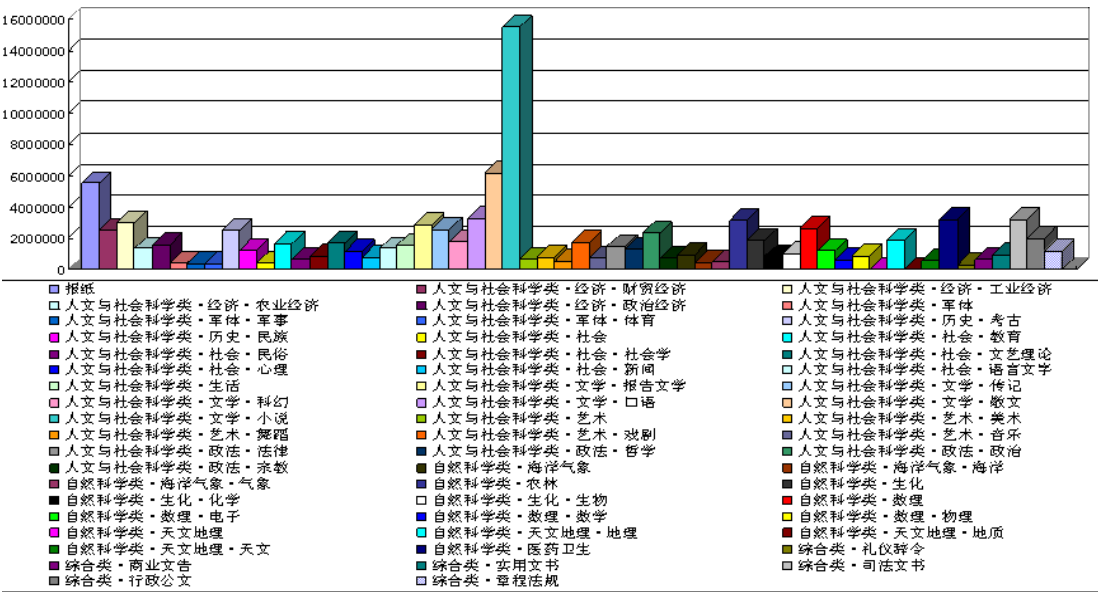
教育部语言文字应用研究所

1. 国家语委现代汉语通用平衡语料库

1.1 语料库全库

国家语委现代汉语通用平衡语料库全库约为 1 亿字符，其中 1997 年以前的语料约 7000 万字符，均为手工录入印刷版语料；1997 之后的语料约为 3000 万字符，手工录入和取自电子文本各半。

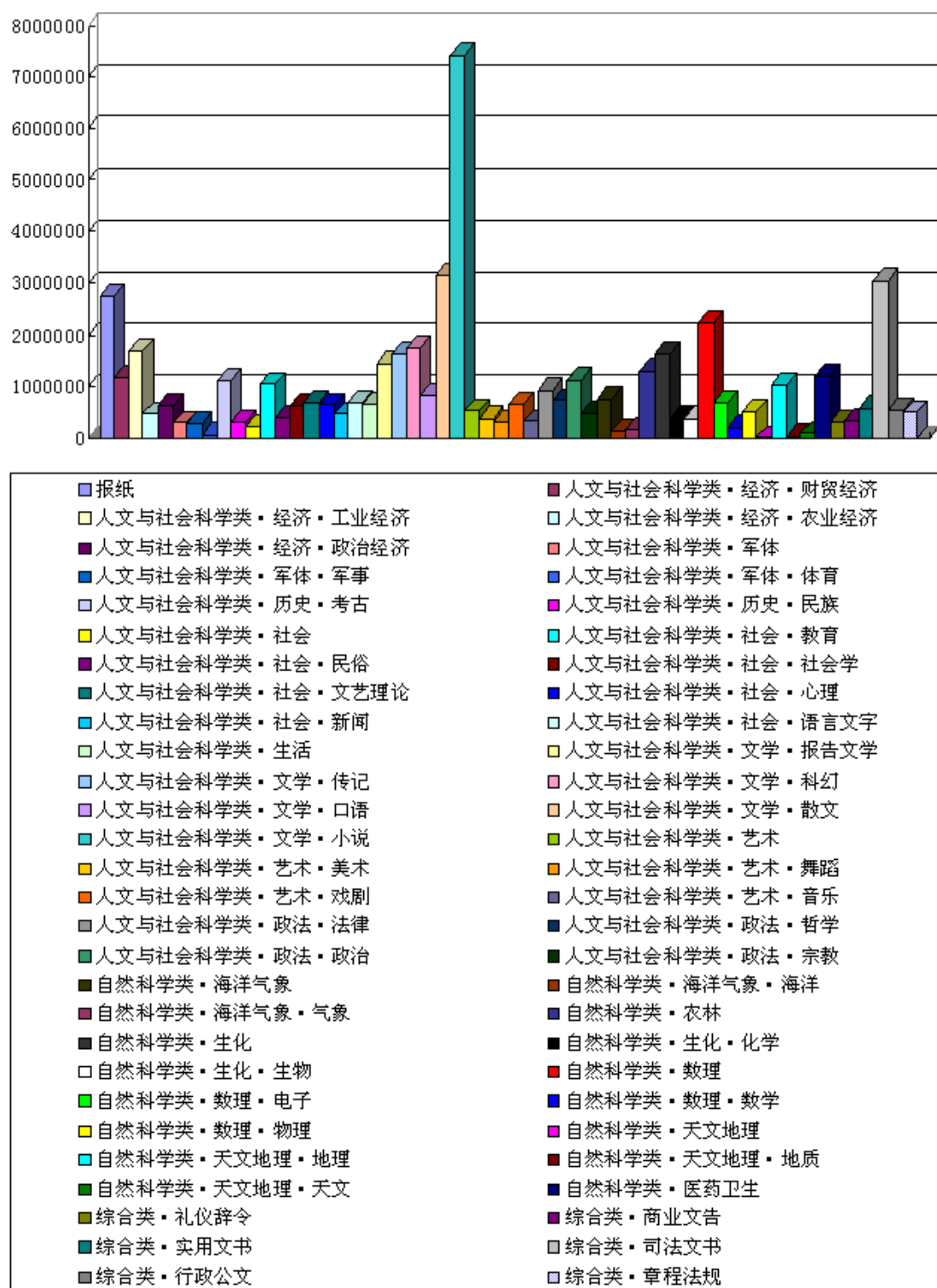
语料库的通用性和平衡性通过语料样本的广泛分布和比例控制实现。语料库类别分布如下所示：



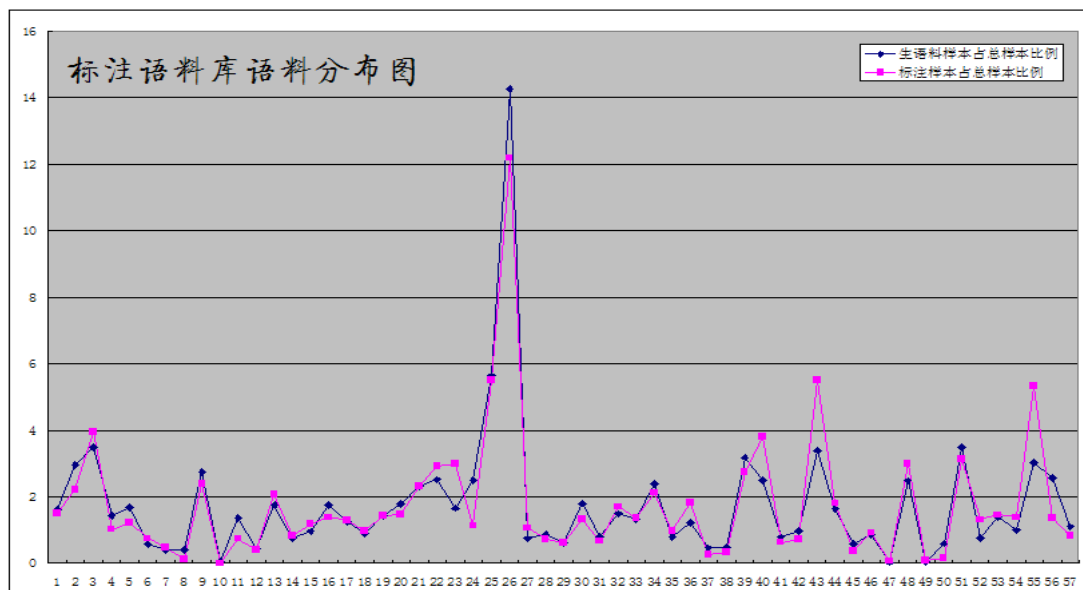
1.2 标注语料库

标注语料库为国家语委现代汉语通用平衡语料库全库的子集，约 5000 万字符。标注是指分词和词类标注，已经经过 3 次人工校对，准确率大于>98%。

语料库全库按照预先设计的选材原则进行平衡抽样，以期达到更好的代表性。标注语料库在样本分布方面近似于全库，不破坏语料选材的平衡原则。标注语料库类别分布如下所示：



标注语料库与全库的样本分布比较如下所示：



（蓝色曲线为语料库全库；红色曲线为标注语料库）

2. 国家语委现代汉语通用平衡语料库语料选材与样本分布

2.1 选材原则

依据材料内容，选材大体作如下分类：（下文字数为建库时数据）

2.1.1 教材

大中小学教材单作一类，约 2000 万字。

2.1.2 人文与社会科学的语言材料约占全库的 60%，共 3000 万字，包括：

- 政法（含哲学、政治、宗教、法律等）；
- 历史（含民族等）
- 社会（含社会学、心理、语言、教育、文艺理论、新闻学、民俗学等）；
- 经济；
- 艺术（含音乐、美术、舞蹈、戏剧等）；
- 文学（含口语）；
- 军体；
- 生活（含衣食住行等方面的普及读物）。

2.1.3 自然科学（含农业、医学、工程与技术）的语言材料，应涉及其发展的各个领域。拟从大、中、小学教材和科普读物中选取。其中，科普读物约占 6%，共 300 万字。教材字数另计。

2.1.4 报刊。以 1949 年以后正式出版的由国家、省、市及各个部委主办的报纸和综合性刊物为主，兼顾 1949 年以前的报纸和综合性刊物。这部分语料约占 26%，共 1300 万字。其中，报纸 900 万字，刊物 400 万字。

2.1.5 应用文。指各类政府公文、文告、书信、说明书、广告等。这部分语料约占 8%，共 400 万字。

2.2 选材年限及密度

2.2.1 教材类

选取现在通用的教材为建库的语言材料。中小学课本所选内容涉及各个学科的基本知识，一般为典范的现代汉语作品，具有相当的普及性、代表性。比较通用的具有通论性质的

高等师范院校和某些专科院校的基础必修教材，内容涉及各个学科的基础理论、基本术语，可为自然科学语汇的收集、科学术语的规范工作提供依据。

2.2.2 人文与社会科学类

以 1919 年为上限，选取五四以来的语言材料。对五四以来各个历史时期的语料采取不等密度选用的方式。

- 1919——1925 年。鉴于五四时期的白话文仍留有文言痕迹，拟选用少量的对后世影响较大的代表性作品。被选用的作品在行文上要尽量符合现代汉语的规范。这部分语料拟占人文与社会科学类的 5%。

- 1926——1949 年。白话文逐步脱离文言痕迹，现代汉语日趋成熟的时期。这部分语料拟占人文与社会科学类的 15%。

- 1950——1965 年。中华人民共和国的成立给社会文化生活带来巨大变化，新词新语大量涌现。这部分语料拟占人文与社会科学类的 25%。

- 1966——1976 年。文化大革命时期产生的作品，其中许多随着文革的结束而仅作为历史词语存于现代汉语之中。这部分语料拟占人文与社会科学类的 5%。

- 1977——。新时期的语料代表了现代汉语的最新发展。这部分语料拟占人文与社会科学类的 50%。

2.2.3 自然科学（含农业、医学、工程与技术）类

自然科学的发展具有较强的优胜劣汰的性质，故对这部分语料做共时性选取，选材范围包括：

- 目前比较通用的中、小学各科教材。
- 目前比较通用的具有通论性质的大学各科基础必修课程的教材。
- 涉及自然科学各个门类的科普读物。

2.3 现代汉语语料库选材字数的分布

2.3.1 人文与社会科学的语言材料占全部 5000 万字语料的 60%，为 3000 万字。这 3000 万字在各个学科的分布见表一。

2.3.2 文学的语言材料占人文与社会科学类的 50%，共 1500 万字。这 1500 万字在不同体裁、题材的语料的分布见表二。

2.3.3 长、中、短篇小说的选取比例大致为：长：中：短=1：2：3

2.4 语料的通用性原则和描述性原则

2.4.1 语料的通用性原则

2.4.1.1 作为通用型语料库，现代汉语语料库应真实地反映现代汉语在文字、词汇、语法、语义等方面的全貌。

2.4.1.2 现代汉语语料库在语料的选择上，应当具有区别性特征。

- 有别于专业性。该语料库的语料要有别于各类专业性的语料，但专业语词与通用语词并无严格的界限，一些专业的用语已经进入通用语言之中，该语料库应尽量涵盖这部分专业语料。

- 有别于地域性。部分方言语词已随社会交际的发展进入标准书面语，各类语料中方言语词也屡见不鲜，有些方言语词已与普通话语词无明显区别，但该语料库在选材上应做到有别于纯方言性的语料。

- 有别于纯口语性。口语语词随地域的不同而有所区别，它的使用范围是比较有限的，所以，该语料库的语料应当是书面语和表义连贯明确、能够用书面语转述的口语语料（如剧本、相声、谈话录、演讲录等），并以前者为主，后者为辅。

2.4.1.3 为确保 5000 万字语料的质量，尽可能地提高所选语料在采字、采词、采句和采义等方面的涵盖量，选材不仅要考虑到语料的时间层次、文化层次和社会使用面层次，还应采

取“抓住中心，其他补充”的方式。

- 时间层次。即指语料的历时性。选取 1919 年至今的各个时期的语料；以 1977 年至今的语料为主，其他各个时期的语料为辅进行补充。

- 文化层次。以具有高中文化程度的人能够阅读的语料为主，其他文化程度为辅。

- 社会使用面层次。以社会使用面较为广泛的语料为主，其他语料为辅进行补充；以人文与社会科学为主，自然科学为辅；以门类为主，以语体为辅，对门类进行补充。

2.4.2 语料的描述性原则

- 从现代汉语语料库建设的主要用途出发，语料应在必要的人工干预的前提下，做描述性选取，以便为语言文字的规范与科研提供客观的科学依据。

- 为了保证现代汉语的字、词、句、义在语料中具有合理的出现频率，语料的选择应在控制比例的前提下，尽量做到采样广泛。

2.5 抽样原则

2.5.1 语言材料的多样性

选用政论性文章、新闻报道、各类文学艺术作品、科普读物、通俗读物、学术专论及各种应用文语体等现代汉语作品。

2.5.2 语言材料的完整性

2000 字以下的文章原则上全篇采用。报纸可采取整篇文章、整版和整张相结合的方式。

2.5.3 语言材料的遍历性

选材要注意各学科，各学科分支，各行各业，以及社会生活各个领域的语言文字应用的代表性。

2.6 抽样要求

2.6.1 抽样的数量与方式

2.6.1.1 书籍

抽样数量一般占全书字数的 3——5%，字数最多不超过 10000 字。样本容量 2000 字，允许±500 字的伸缩。

2.6.1.2 报纸

采用整版（4 版或 8 版）选用的方式。不同的报纸选用不同的月份，以免内容重复。

报纸上的广告、启事等归在应用文类，不在报刊类语料的统计之列。

2.6.1.3 刊物

每本刊物上所选的总字数原则上不超过 5000 字。样本容量 2000 字，允许±500 字的伸缩。

对同一版面的不同文章，按从上至下、从左到右的顺序选取。

一个样本必为同一作者的同一篇文章，限字数不限样本数（报刊除外）。

每个样本之中必为连续的语料内容。

应用文（包括广告、说明书等）

2000 字以内的应用文宜整篇选用。对于篇幅较长的应用文，所选样本的容量为 2000 字，允许±500 字的伸缩。

2.6.2 抽样材料的取舍要求

- 每个样本头尾处小于句子的语言片段应删除。
- 书信中的落款、套语、日期等应删除。
- 图片的文字说明一律删除。
- 剧本中的人物名要删除。
- 作者、记者、实习生、通讯员、编辑、摄影、绘画等的名字一律删除。
- 报刊中的专栏及其承办、协办单位的名称一律删除。

- 旁注及旁注号一律删除。
- 报刊中标明稿件来源的字样，如“本报讯”、“通讯员”、“实习生”、“本报记者”、“新华社北京×月×日电”、“××杯散文特写征文”、“本报电视照片”、“插图”、“题图”、“本报编辑”、“责任编辑”等字样一律删除。
- 报纸上的电影广告、电视节目预告、体育比赛预告、戏剧节目预告等内容一律删除。
- 报刊上的报刊名称、日期、天气预报及版权说明部分一律删除。
- 报缝中的内容一律删除。
- 删除的内容一律用红笔加框。
- 印刷错误要改正。
- 字数采用通栏字数与行数的乘积去掉明显的空白的方式加以统计。
- 复印中字迹不清楚的当采用校对符号予以标明。

2.7 抽样的补充要求

2.7.1 在现代汉语语料库选材过程中，各承担任务的单位与个人应严格按照本原则所阐述的宗旨与规定进行，如遇确需改动的情况，须事先提出商量。

2.7.2 以上有关选材年限及密度的规定是着眼于科学的整体发展而制定的。各个学科的发展在不同的年代并不是齐头并进的，可根据具体情况适当调整依年限分布的比例、字数。调整的理由、调整后的比例和字数当详细说明，并作为附件收于清单之后。

2.7.3 大学教材门类以国家规定的大学基础必修课为准。

2.7.4 避免选取文言色彩较重的篇章作语料，例如鲁迅等作家的作品不宜用作语料。避免选取诗歌作语料；剔除篇章中诗歌形式的内容。

2.7.5 详细、准确无误地填写选材清单及选材卡片中的每一项。选材字数的统计精确到十位数。

2.8 分类别的样本分布示例

表一：人文与社会科学类

科目	比例	字数	1919 ~ 1925	1926 ~ 1949	1950 ~ 1965	1966 ~ 1976	1977 ~
			5%	15%	25%	5%	50%
哲学	8.3%	250	12.5	37.5	62.5	12.5	125
历史	8.3%	250	12.5	37.5	62.5	12.5	125
社会	8.3%	250	12.5	37.5	62.5	12.5	125
经济	8.3%	250	12.5	37.5	62.5	12.5	125
艺术	8.3%	250	12.5	37.5	62.5	12.5	125
文学	50%	1500	75	225	375	75	750
其他	8.3%	250	12.5	37.5	62.5	12.5	125

表二：文学类（含口语）

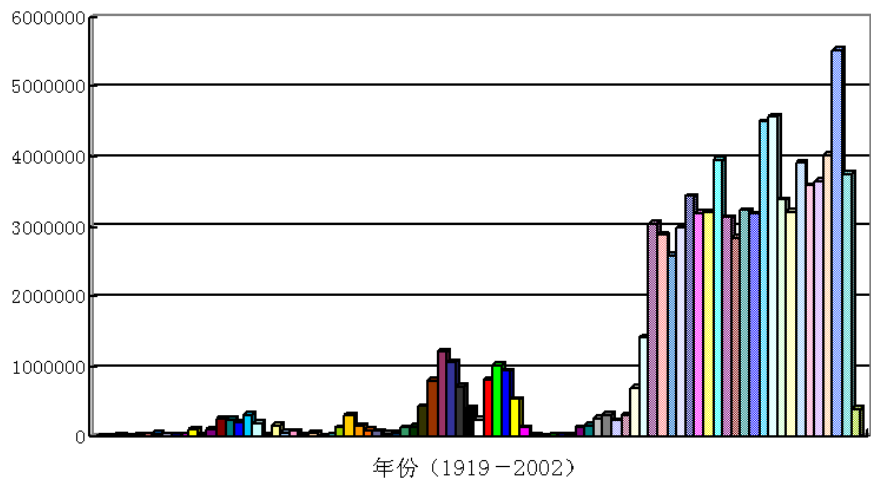
题材	比例	字数	1919 ~ 1925	1926 ~ 1949	1950 ~ 1965	1966 ~ 1976	1977 ~
			5%	15%	25%	5%	50%
小说	30%	450	22.5	67.5	112.5	22.5	225
散文（杂文）	20%	300	15	45	75	15	150
传记	10%	150	15	22.5	37.5	15	75
报告文学	10%	150	50			100	
科幻	10%	150	50			100	

口语	20%	300	15	45	75	15	150
----	-----	-----	----	----	----	----	-----

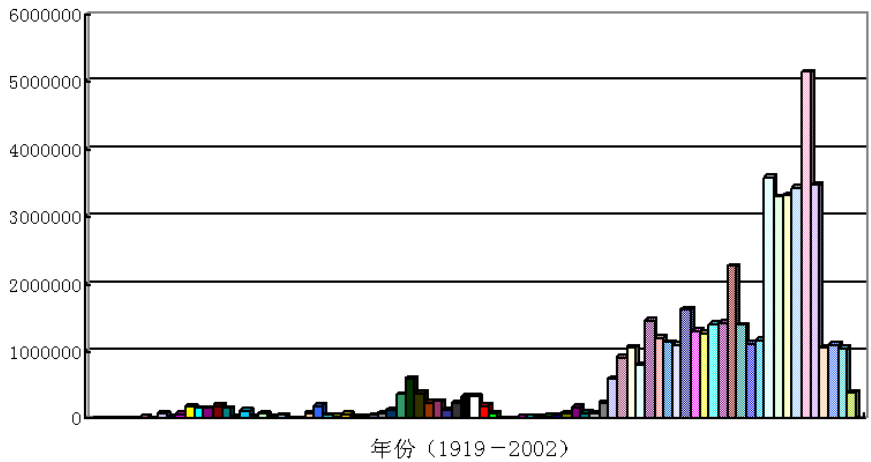
2.9 语料库样本的时间分布

语料时间跨度为 1919 年～2002 年，以近 20 年的语料为主。

语料库全库语料样本的时间分布：



标注语料库语料样本的时间分布



3. 语料库分词和词类标注

3.1 分词（切分单位）

切分单位定义为汉语信息处理使用的、具有确定语法功能的基本单位。它包括本标准的规则所限定的词、短语及其他单位。

切分单位包括词、短语和其他切分单位，如习用语、缩略语、前接成分、后接成分、语素字、非语素字、标点符号、非汉字符号等。

3.2 词类（词性）分词（切分单位）

词类指词的语法分类，主要是根据语法功能划分出来的类。

3.3 词类及分单位分类

词类划分为13个一级类，16个二级类；其他切分单位划分为7个一级类，13个二级类。
词类划分及标记代码如下：

3.3.1 名词(n)，表示人或事物的名称，在句子中主要充当主语和宾语。

3.3.1.1 普通名词(ng)，表示事物的名称。如：

人 马 书 教师 飞机 电冰箱 阿姨 桌子 木头
道德 理论 历史 思想 文化 因素 作风 哲学

3.3.1.2 时间名词(nt)，包括一般所说的时量词。如：

年 月 日 分 秒
现在 过去 昨天 去年 将来 宋朝 星期一

3.3.1.3 方位名词(nd)，表示位置的相对方向。如：

上 下 左 右 前 后 里 外 中 东 西 南 北
前边 左面 里头 中间 外部

3.3.1.4 处所名词(nl)，表示处所。如：

空中 高处 隔壁 门口 附近 边疆 一旁 野外

3.3.1.5 人名(nh)，表示人的名称的专有名词。

华罗庚 阿凡提 诸葛亮 司马相如 松赞干布 卡尔·马克思

3.3.1.6 地名(ns)，表示地理区域名称的专有名词。如：

亚洲 大西洋 地中海 阿尔卑斯山 加拿大
中国 北京 浙江 景德镇 呼和浩特 中关村

3.3.1.7 族名(nn)，表示民族或部落名称的专有名词。如：

回族 藏族 壮族 蒙古族 维吾尔族 哈萨克族

3.3.1.8 机构名(ni)，表示团体、组织、机构名称的专有名词。如：

联合国 教育部 北京大学 中国科学院

3.3.1.9 其他专有名词(nz)。如：

五粮液 宫爆鸡丁 桑塔纳

3.3.2 动词(v)，表示动作、行为，心理活动、生理状态及事物的存现、变化等，在句子中主要充当谓语。

3.3.2.1 及物动词(vt)，能够带宾语。如：

吃 打 擦 洗 喂 借 送 买 捧 提 填
喜欢 告诉 接受 羡慕 考虑 调查 同意 发动

3.3.2.2 不及物动词(vi)，不能够带宾语。如：

病 休息 咳嗽 瘫痪 游泳 睡觉

3.3.2.3 联系动词(vl)，表示关系的判断。如：

是

3.3.2.4 能愿动词(vu)，表示可能、意愿。如：

能够 能 应该 可以 可能 情愿 愿意 要

3.3.2.5 趋向动词(vd)，表示趋向。如：

(走)上 (趴)下 (进)来 (回)去
(跑)上来 (掉)下去 (提)起来 (扔)过去

3.3.3 形容词(a)，表示性质、状态，在句中主要充当谓语、定语、状语和补语。

3.3.3.1 性质形容词(aq)，表示性质。如：

好 高 美 大 勇敢 危险 漂亮 干净 伟大

3.3.3.2 状态形容词(as)，表示状态。如：

雪白 黧黑 通红 冰凉 绿油油 亮堂堂 白花花 冷冰冰

3.3.4 区别词(f)，表示事物的区别性特征，在句子中只能做定语修饰名词或跟助词“的”组成“的”字结构。如：

男 女 公 母 雌 雄 微型 国产 军用

3.3.5 数词(m)，表示数目和次序。如：

零 一 半 百 千 百万 一百零八
第一 第十八

3.3.6 量词(q)，表示人、事物或动作的单位。如：

个 条 片 匹 辆 尺 斤 两 吨 支 回 次 遍 千瓦时

3.3.7 代词(r)，起替代和复指作用。如：

我 你 他 这 那 谁 我们 你们 他们
这个 那个 大家 自己 什么 哪里 怎么 怎么样

3.3.8 副词(d)，修饰或限制动词和形容词，表示范围、程度等。在句子中做状语。如：

都 只 就 仅 很 再三 屡次 将 不 却
总共 正在 常常 重新 曾经 竟然 居然

3.3.9 介词(p)，引介名词性成分，不单独充当句子成分。如：

把 被 从 向 对 凭
按照 对于 为了 自从 关于

3.3.10 连词(c)，连接词、短语或句子，表示两者之间所具有的某种关系。如：

和 同 与 及 并 或
并且 而且 或者 因为 所以

3.3.11 助词(u)，附着在词、短语、句子后面表示某种附加意义。如：

的 地 得 了 着 过 等等 似的 一样

3.3.12 叹词(e)，表示感叹、呼唤或应答，可独立成句或在句中充当独立成分。如：

啊 嗯 唉 哎 哼 哦 哎哟 哎呀

3.3.13 拟声词(o)，模拟自然界事物的某种声音，不能单独成句。如：

砰 滴答 扑通 咕咚 丁丁当当

3.4 其他切分单位划分及标记代码

3.4.1 习用语(i)，一种相沿习用的定型短语。

3.4.1.1 名词性习用语(in)。如：

海市蜃楼 井底之蛙 蛛丝马迹

3.4.1.2 动词性习用语(iv)。如：

跑龙套 打官腔 吃老本 与时俱进 励精图治

3.4.1.3 形容词性习用语(ia)。如：

丰富多彩 艰苦朴素 光明正大

3.4.1.4 连词性习用语(ic)。如：

总而言之 由此可见 综上所述

3.4.2 缩略语(j)，专有名词或常用语的简缩形式。

3.4.2.1 名词性缩略语(jn)。如：

人大 五四 奥运

3.4.2.2 动词性缩略语(jv)。如：

调研 离退休

3.4.2.3 形容词性缩略语（ja）。如：

短平快 高精尖

3.4.3 前接成分（h），词根前面的附加构词成分。如：

阿 老 初 第

3.4.4 后接成分（k），词根后面的附加构词成分。如：

子 儿 头 化 们 式 性 者

3.4.5 语素字（g），汉字字符集中一般不单独使用的汉字。

3.4.5.1 名词性语素字（gn）。如：

民 农 材

3.4.5.2 动词性语素字（gv）。如：

抒 究 涤

3.4.5.3 形容词性语素字（ga）。如：

殊 遥 伟

3.4.6 非语素字（x），汉字字符集中单独使用时不具有意义的汉字，如：

垃 琵 蜘 踌 鸯 蜻

3.4.7 其他（w）

3.4.7.1 标点符号（wp），如：

， 。 、 ； ？ ！ ： “ ” ……

3.4.7.2 非汉字字符串（ws），如：

office windows

3.4.7.3 其他未知的符号（wu）。

3.5 词类及其他切分单位标记代码表

表 词类及其他切分单位标记代码表（按标记代码的字母顺序排列）

序号	标记代码		类别名称	代码说明
	一级类	二级类		
	a		形容词	<u>a</u> djective
		aq	性质形容词	<u>a</u> djective- <u>q</u> uality
		as	状态形容词	<u>a</u> djective- <u>s</u> tate
	c		连词	<u>c</u> onjunction
	d		副词	<u>a</u> dverb
	e		叹词	<u>e</u> xclamation
	f		区别词	<u>d</u> ifference
	g		语素字	“根”的汉语拼音首字母
		ga	形容词性语素字	“根”的汉语拼音首字母-adjective
		gn	名词性语素字	“根”的汉语拼音首字母-noun
		gv	动词性语素字	“根”的汉语拼音首字母-verb
	h		前接成分	<u>h</u> ead
	i		习用语	<u>i</u> diom
		ia	形容词性习用语	<u>i</u> diom- <u>a</u> djective
		ic	连词性习用语	<u>i</u> diom- <u>c</u> onjunction
		in	名词性习用语	<u>i</u> diom- <u>n</u> oun

		iv	动词性习用语	<u>i</u> diom- <u>v</u> erb
	j		缩略语	“简”的汉语拼音首字母
		ja	形容词性缩略语	“简”的汉语拼音首字母- <u>a</u> djective
		jn	名词性缩略语	“简”的汉语拼音首字母- <u>n</u> oun
		jv	动词性缩略语	“简”的汉语拼音首字母- <u>v</u> erb
	k		后接成分	依据通常做法
	m		数词	<u>n</u> um <u>e</u> ral
	n		名词	<u>n</u> oun
		nd	方位名词	<u>n</u> oun- <u>d</u> irection
		ng	普通名词	<u>n</u> oun- <u>g</u> eneral
		nh	人名	<u>n</u> oun- <u>h</u> uman
		ni	机构名	<u>n</u> oun- <u>i</u> nstitution
		nl	处所名词	<u>n</u> oun- <u>l</u> ocation
		nn	族名	<u>n</u> oun- <u>n</u> ation
		ns	地名	<u>n</u> oun- <u>s</u> pace
		nt	时间名词	<u>n</u> oun- <u>t</u> ime
		nz	其他专有名词	<u>n</u> oun-“专”的汉语拼音首字母
	o		拟声词	<u>o</u> nomatopoeia
	p		介词	<u>p</u> reposition
	q		量词	<u>q</u> uantity
	r		代词	<u>p</u> ronoun
	u		助词	<u>a</u> uxiliary
	v		动词	<u>v</u> erb
		vd	趋向动词	<u>v</u> erb- <u>d</u> irection
		vi	不及物动词	<u>v</u> erb- <u>i</u> ntransitive
		vl	联系动词	<u>v</u> erb- <u>l</u> inking
		vt	及物动词	<u>v</u> erb- <u>t</u> ransitive
		vu	能愿动词	<u>v</u> erb- <u>a</u> uxiliary
	w		其他	依据通常做法
		wp	标点符号	依据通常做法
		ws	非汉字字符串	“w”- <u>s</u> tring
		wu	其他未知符号	“w”- <u>u</u> nknown
	x		非语素字	依据通常做法

说明：1）以上词类标记体系为推荐性，标注语料库并未标注表中出现的全部二级分类代码，例如vt不及物动词；2）少量切分单位被标注为组合代码，例如数量词mq。

词类标记集规范详见附录：GBT20532-2006 信息处理用现代汉语词类标记规范.doc

4. 语料库数据格式描述

4.1 语料样本的基本信息

每个语料样本包含 24 个信息，如下表所示：

字段名称	内容	类型	大小
a1_总号	总号	数字	双精度型

a2_分类号	分类号	文本	12
a3_样本名称	样本名称	文本	100
a4_类别	类别	文本	22
a5_作者	作者	文本	50
a6_写作时间	写作时间	文本	8
a7_书刊名称	书刊名称	文本	40
a8_编著者	编著者	文本	50
a9_出版社	出版社	文本	30
a10_所在省	所在省	文本	10
a11_出版日期	出版日期	文本	8
a12_期号	期号	数字	双精度型
a13_版次(初版印数)	版次(初版印数)	数字	双精度型
a14_本版印数	本版印数	数字	双精度型
a15_总印数	总印数	数字	双精度型
a16_总页数	总页数	数字	双精度型
a17_开本	开本	文本	20
a18_选择方式	选择方式	文本	20
a19_起止页数	起止页数	文本	12
a20_样本字数	样本字数	数字	双精度型
a21_样本总字数	样本总字数	数字	双精度型
a22_文章总字数	文章总字数	数字	双精度型
a23_简繁体	简繁体(录入时全部为简体)	文本	10
a24_抽样文章	抽样文章	备注	
a24_抽样文章_标注	标注语料	备注	

说明：1）格式为 Access 格式数据库时，每一信息对应一个字段；2）格式为文本文件时，a24 对应的语料内容放于以 a2_分类号命名的文本文件中，其他信息通过 a2_分类号和语料清单文件对应；3）a2_分类号为语料样本的编号，每一个样本有且仅有一个对应分类号；4）不是每一个语料样本都具有上述 24 项信息，小部分样本存在个别信息不全的情况，原因是样本不具有该项信息或采样时未采集该项信息；5）a1_总号已经不再使用，不作任何用途。

语料基本信息数据如下表所示：

a2_分类号	a3_样本名称	a4_类别	a5_作者	a6_写作时间	a7_书刊名称	a8_编著者	a9_出版社	a10_所在省	a11_出版
FH10005001	《我们的歌》节录	文学·小说	赵淑侠	1983-8-1	《我们的歌》	赵淑侠	友谊出版公司	北京	1983-9-1
FH10005002	《我们的歌》节录	文学·小说	赵淑侠	1983-8-1	《我们的歌》	赵淑侠	友谊出版公司	北京	1983-9-1
FH10005003	《我们的歌》节录	文学·小说	赵淑侠	1983-8-1	《我们的歌》	赵淑侠	友谊出版公司	北京	1983-9-1
FH10005004	《我们的歌》节录	文学·小说	赵淑侠	1983-8-1	《我们的歌》	赵淑侠	友谊出版公司	北京	1983-9-1
FH10005005	《我们的歌》节录	文学·小说	赵淑侠	1983-8-1	《我们的歌》	赵淑侠	友谊出版公司	北京	1983-9-1
FH10000101	《大地》节录	文学·小说	秦兆阳	1983-10-1	《大地》		人民文学出版社	北京	1984-6-1
FH10000102	《大地》节录	文学·小说	秦兆阳	1983-10-1	《大地》		人民文学出版社	北京	1984-6-1
FH10000103	《大地》节录	文学·小说	秦兆阳	1983-10-1	《大地》		人民文学出版社	北京	1984-6-1
FH10000104	《大地》节录	文学·小说	秦兆阳	1983-10-1	《大地》		人民文学出版社	北京	1984-6-1
FH10000105	《大地》节录	文学·小说	秦兆阳	1983-10-1	《大地》		人民文学出版社	北京	1984-6-1
FH10002201	《柳暗花明》节录	文学·小说	欧阳山	1981-5-1	《柳暗花明》		人民文学出版社	北京	1981-9-1
FH10002202	《柳暗花明》节录	文学·小说	欧阳山	1981-5-1	《柳暗花明》		人民文学出版社	北京	1981-9-1
FH10002203	《柳暗花明》节录	文学·小说	欧阳山	1981-5-1	《柳暗花明》		人民文学出版社	北京	1981-9-1
FH10002204	《柳暗花明》节录	文学·小说	欧阳山	1981-5-1	《柳暗花明》		人民文学出版社	北京	1981-9-1
FH10002205	《柳暗花明》节录	文学·小说	欧阳山	1981-5-1	《柳暗花明》		人民文学出版社	北京	1981-9-1

a13_版次	a14_本版印数	a15_总印数	a16_总页数	a17_开本	a18_选择方式	a19_起止页数	a20_样本字数	a21_样本总字数	a22_文章总字数	a23_简繁
1	14000	14000	186 0		简单随机抽样	68-91	1777	6490	121000	繁
1	14000	14000	186 0		简单随机抽样	128-131	1672	6490	121000	繁
1	12000	12000	74 0		简单随机抽样	2-4	1478	6790	45000	繁
1	12000	12000	74 0		简单随机抽样	34-36	1616	6790	45000	繁
1	12000	12000	74 0		简单随机抽样	37-39	1530	6790	45000	繁
1	12000	12000	74 0		简单随机抽样	57-59	1707	6790	45000	繁
1	4800	4800	79 0		简单随机抽样	4-6	1858	9700	56000	繁
1	4800	4800	79 0		简单随机抽样	7-9	1883	9700	56000	繁
1	4800	4800	79 0		简单随机抽样	40-42	1919	9700	56000	繁
1	4800	4800	79 0		简单随机抽样	43-45	1999	9700	56000	繁
1	4800	4800	79 0		简单随机抽样	66-70	1927	9700	56000	繁
1	2000	2000	106 0		简单随机抽样	8-10	1984	9700	74000	繁
1	2000	2000	106 0		简单随机抽样	39-41	2051	9700	74000	繁
1	2000	2000	106 0		简单随机抽样	64-66	1842	9700	74000	繁
1	2000	2000	106 0		简单随机抽样	67-69	1875	9700	74000	繁
1	2000	2000	106 0		简单随机抽样	87-89	2017	9700	74000	繁

4.2 语料库数据格式

4.2.1 语料数据库为 Access 数据库文件。

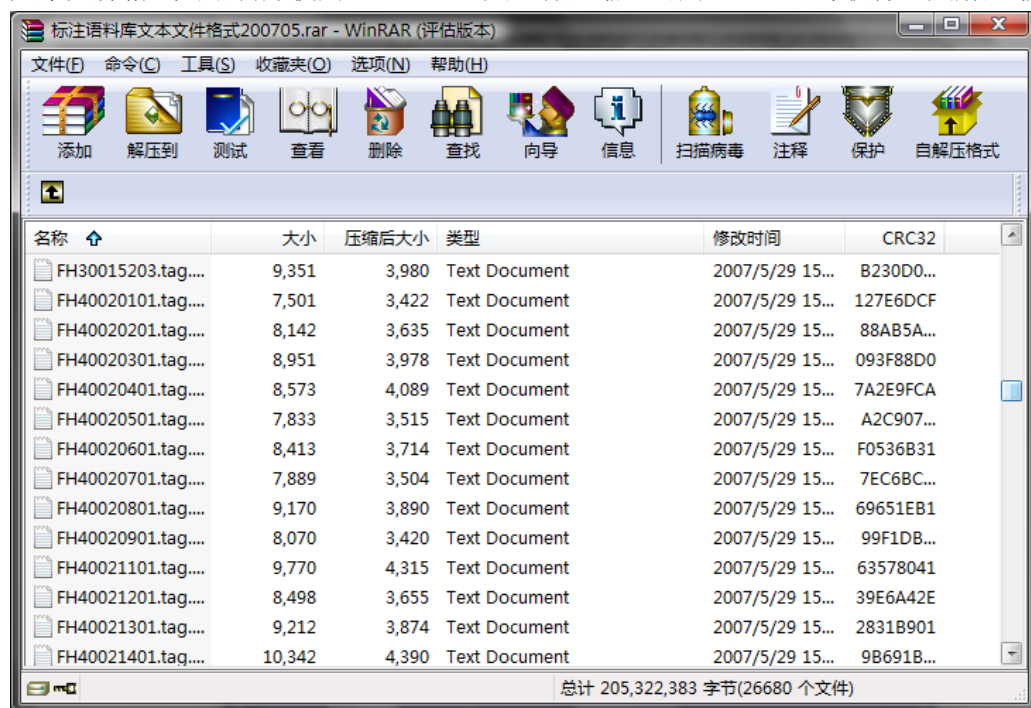
Access 数据库格式(后缀.MDB),可用 Microsoft Office 2000/XP/2003/2007 中的 Microsoft Access 软件打开,也可以使用其他的数据库工具打开。

Access 数据库格式语料库文件名称为:国家语委标注语料库(5000 万字).mdb。

4.2.2 文本文件格式。

格式为文本文件(后缀.TXT)的,语料文本和样本信息分开存放。文件名(例如 FH30015203)与语料数据库中字段“a2_分类号”一一对应(可通过分类号到数据库中获取该文本的详细信息)。每个语料有且仅有一个分类号,分类号不重复

文本文件格式的语料库使用 WinRAR 工具进行压缩,可用 WinRAR 等软件工具解压缩。



4.2.3 标注语料样例

样本编号: BF29701101

样本名称: 鸟的世界

类别: 文学·散文

作者: 杨栋

出版时间: 1997-12-11

书刊名称: 人民日报

鸟/的/世界/

杨栋/

鸟/，/w 是/vl 大自然/的/歌手/，/w 鸟语/就是/vl 大自然/的/音乐/和/c 诗
歌/了/。/w

山村/里/nd 的/鸟/除了/p 麻雀/，/w 就/d 数/v 燕子/多/a 了/。/w 村/人/
对/p 燕子/很/d 爱护/v，/w 说/v 它/r 吃/v 庄稼/的/害虫/，/w 常/a 吓唬/v 孩子/
们/k 不要/vu 去/v 玩/v 燕子/，/w 会/vu 坏/v 自己/r 的/眼睛/。/w 有时/r 光/v 屁股
/的/小/a 燕/掉/v 下来/vd，/w 也/d 要/vu 送回/v 燕/窝/里/nd 去/v。/w

说明: 1) 词(切分单位)之间以空格分隔; 2) 词(切分单位)与词类标记之间以“/”号
分隔。

其他样例:

样本: 10010za (哲学)

毛/nhf 泽东/nhs 同志/n 指出/v，/w 在/p 事物/的/内在/f 矛盾/n 中/nd，/w
存在/v 着/u 主要/a 的/矛盾/n 和/c 矛盾/n 的/主要/a 方面/n；/w 这/r 进一步/v
说明/v 任何/r 事物/的/内在/f 矛盾/n 是/vl 复杂/a 的/u，/w 严密/a 的/u，/w
它们/r 的/u 运动/v 发展/v 必然/d 是/vl 有/v 规律/n 的/u，/w 不/d 是/vl 可以/vu
任意/d 变动/v 的/u。/w 上述/n 唯物辩证法/n 的/u 基本/a 概念/n 已/d 从/p 人类/
的/u 社会/n 实践/n 和/c 科学/n 实践/n 取得/v 丰富/a 的/u 证明/n。/w 一切/r 事
物/n 都/d 有/v 内在/f 矛盾/n 这/r 概念/n，/w 是/vl 以/p 物质/n 可/vu 分割/v 性
/k 和/c 不可/vu 穷尽/v 性/k 的/u 认识/n 为/vl 基础/n 的/u。/w

样本: 10010303jc (历史)

到/v 公元前/nt 2/m 世纪/nt，/w 罗马/ns 已经/d 占有/v 地中海/ns 周围/nd 从
/p 西班牙/ns 到/v 小亚细亚/ns 的/u 许多/a 地方/n、/w 北非/ns 的/u 一/m 部分/n
和/c 地中海/ns 上/nd 的/u 许多/a 岛屿/n，/w 一/d 跃/v 而/c 为/vl 地中海/ns 的/u
一/m 大/a 强/a 国/n。/w 奴隶/n 制度/n 随着/p 对/p 外/nd 扩张/v 高度/a 发展/v
起来/vd。/w 罗马/ns 每/r 征服/v 一/m 个/q 地方/n，/w 常常/d 把/p 那里/r 的/u 人
/n 全部/n 卖/v 为/vl 奴隶/n。/w 海盗/n 又/d 常/d 拐/v 掠/v 人口/n 卖/v 作/v 奴
隶/n。/w 因此/c 奴隶/n 数量/n 很/d 多/a，/w 售价/n 也/d 低/a。/w

Tel: 010 65592936

Email:exiaohang@sina.com

教育部语言文字应用研究所

北京市东城区朝阳门南小街 51 号（100010）

2009/10/22