



教育部语言文字应用研究所

INSTITUTE OF APPLIED LINGUISTICS MINISTRY OF EDUCATION

WWW.IALC.CN

国家语委现代汉语语料库介绍

肖航

教育部语言文字应用研究所

2012

语料库建设

□ 国家语委语料库建设

- 1991年12月国家语言文字工作委员会提出立项；
- 1992年4月召开现代汉语语料库选材原则专家论证会；
- 1993年1月制订《现代汉语语料库选材原则》；
- 1993年9月召开现代汉语语料库选材专家审定会；
- 1998年底建成 7000万字的生语料库；
- 目前已完成1亿字生语料和5000万字标注语料；
- 语料库建设和加工工作还在继续进行。

□ 被列为国家语委“九五”、“十五”科研重大项目

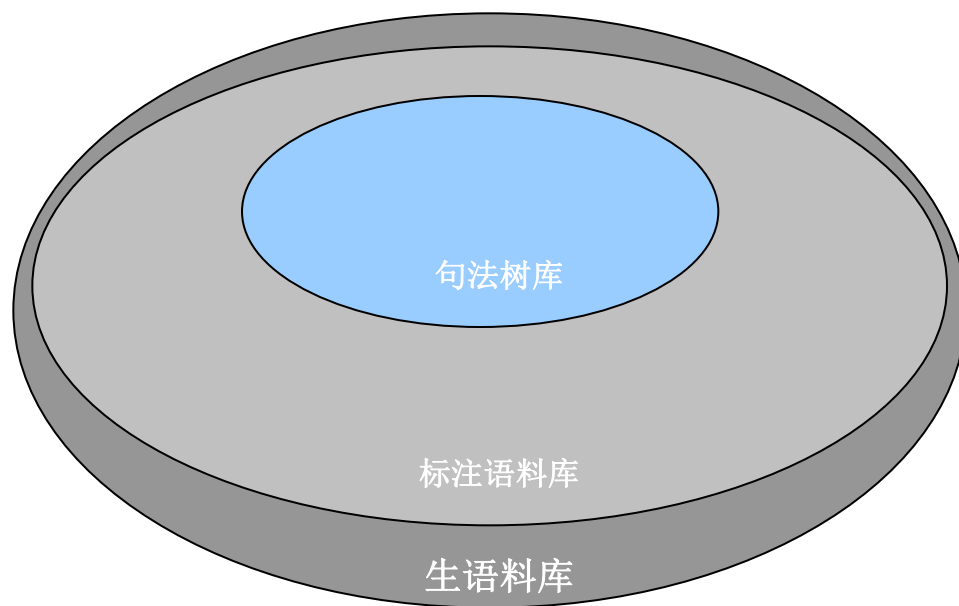
□ 得到国家科技部“863”、“973”计划多个项目的支持

- “智能中文信息处理平台”
- “图像、语音和自然语言理解”
- “中文信息处理应用基础研究”



语料库的主要内容

- 未经标注加工的生语料库
- 标注语料库
 - 词语切分
 - 词类标注
- 句法树库
 - 内部结构
 - 外部功能
- 分词词表
 - 88000词条
 - 词性标注
 - 频率信息
- 语料库加工标注规范
- 语料库软件工具



语料库的主要用途

□ 主要用途

- 语言文字的信息处理
- 语言文字规范和标准的制定
- 语言文字的学术研究
- 语文教育
- 语言文字的社会应用



语料来源

- 1993年以前的语料
 - 以人工录入印刷版本的语料为主
 - 约7000万字
- 1993~2002年的语料
 - 部分采用人工录入印刷版本语料
 - 约1500万字
 - 部分来源于网络电子文本
 - 约1500万字
- 2002以后的语料
 - 以网络电子文本为主
 - 约1000万字



语料分类

□ 三个主要类别

□ 人文与社会科学类

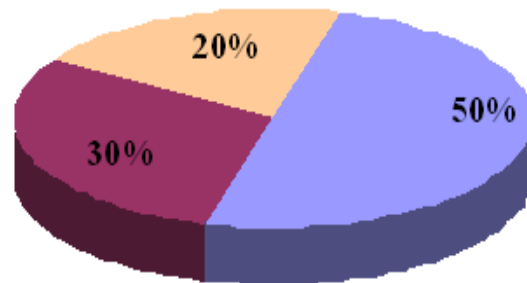
- 包括政法、历史、社会、经济、文学、艺术等类别语言材料

□ 自然科学类

- 自然科学的语言材料（含农业、工业、医学、电子、工程技术等），涉及科学技术发展的各个领域。

□ 综合类

- 应用文
- 难于归类的语料



■ 人文与社会科学类 ■ 自然科学类 ■ 综合类



教育部语言文字应用研究所

INSTITUTE OF APPLIED LINGUISTICS MINISTRY OF EDUCATION

WWW.IAL.CN

人文与社会科学类

- 人文与社会科学类划分为8个大类和30个小类：
 - 政法：哲学、政治、宗教、法律；
 - 历史：历史、考古、民族；
 - 社会：社会学、心理、语言文字、教育、文艺理论、新闻、民俗；
 - 经济：工业经济、农业经济、政治经济、财贸经济；
 - 艺术：音乐、美术、舞蹈、戏剧；
 - 文学：小说、散文、传记、报告文学、科幻、口语；
 - 军体：军事、体育；
 - 生活。
- 人文与社会科学类约占语料总量的50%



自然科学类

- 自然科学划分为6类：
 - 数理
 - 生化
 - 天文地理
 - 海洋气象
 - 农林
 - 医药卫生
- 自然科学类约占语料总量的30%，



综合类

- 综合类语料由应用文和难于归类的其他语料两部分组成。
- 应用文主要包括以下6类：
 - 行政公文：请示、报告、批复、命令、指示、布告、纪要、通知等；
 - 章程法规：章程、条例、细则、制度、公约、办法、法律条文等；
 - 司法文书：诉讼、辩护词、控告信、委托书等；
 - 商业文告：说明、广告、调查报告、经济合同等；
 - 礼仪辞令：欢迎词、贺电、讣告、唁电、慰问信、祝酒词等；
 - 实用文书：请假条、检讨、申请书、请愿书等。
- 综合类约占语料总量的20%



样例 语料分类

- | | | |
|--------------------|--------------------|--------------------|
| ■ 报纸 | ■ 人文与社会科学类·经济·财贸经济 | □ 人文与社会科学类·经济·工业经济 |
| □ 人文与社会科学类·经济·农业经济 | ■ 人文与社会科学类·经济·政治经济 | ■ 人文与社会科学类·军体 |
| ■ 人文与社会科学类·军体·军事 | ■ 人文与社会科学类·军体·体育 | □ 人文与社会科学类·历史·考古 |
| ■ 人文与社会科学类·历史·民族 | ■ 人文与社会科学类·社会 | ■ 人文与社会科学类·社会·教育 |
| ■ 人文与社会科学类·社会·民俗 | ■ 人文与社会科学类·社会·社会学 | ■ 人文与社会科学类·社会·文艺理论 |
| ■ 人文与社会科学类·社会·心理 | ■ 人文与社会科学类·社会·新闻 | □ 人文与社会科学类·社会·语言文字 |
| □ 人文与社会科学类·生活 | ■ 人文与社会科学类·文学·报告文学 | ■ 人文与社会科学类·文学·传记 |
| ■ 人文与社会科学类·文学·科幻 | ■ 人文与社会科学类·文学·口语 | ■ 人文与社会科学类·文学·散文 |
| ■ 人文与社会科学类·文学·小说 | ■ 人文与社会科学类·艺术 | ■ 人文与社会科学类·艺术·美术 |
| ■ 人文与社会科学类·艺术·舞蹈 | ■ 人文与社会科学类·艺术·戏剧 | ■ 人文与社会科学类·艺术·音乐 |
| ■ 人文与社会科学类·政法·法律 | ■ 人文与社会科学类·政法·哲学 | ■ 人文与社会科学类·政法·政治 |
| ■ 人文与社会科学类·政法·宗教 | ■ 自然科学类·海洋气象 | ■ 自然科学类·海洋气象·海洋 |
| ■ 自然科学类·海洋气象·气象 | ■ 自然科学类·农林 | ■ 自然科学类·生化 |
| ■ 自然科学类·生化·化学 | □ 自然科学类·生化·生物 | ■ 自然科学类·数理 |
| ■ 自然科学类·数理·电子 | ■ 自然科学类·数理·数学 | ■ 自然科学类·数理·物理 |
| ■ 自然科学类·天文地理 | ■ 自然科学类·天文地理·地理 | ■ 自然科学类·天文地理·地质 |
| ■ 自然科学类·天文地理·天文 | ■ 自然科学类·医药卫生 | ■ 综合类·礼仪辞令 |
| ■ 综合类·商业文告 | ■ 综合类·实用文书 | □ 综合类·司法文书 |
| ■ 综合类·行政公文 | ■ 综合类·章程法规 | |



语料库选材

□ 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。

□ 自然科学类

- 比较通用的中、小学各科教材。
- 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。

□ 教材

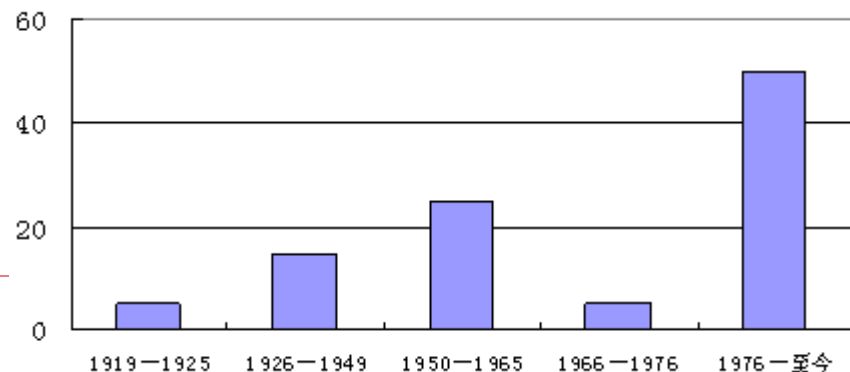
- 选取当时通用的教材为建库的语言材料。
- 中小学课本所选内容涉及的各个学科的基本知识
- 一般为典范的现代汉语作品，应具有相当的普及性、代表性。



语料库选材的历时性

□ 以人文与社会科学类为例

- 1919 — 1925年，约占5%。
 - 五四时期的白话文仍留有文言痕迹，选用少量的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。
- 1926 — 1949年，约占15%。
 - 白话文逐步脱离文言痕迹，现代汉语日趋成熟的时期。
- 1950 — 1965年，约占25%。
 - 新中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。
- 1966 — 1976年，约占5%。
 - 文革时期的作品，其中许多仅作为历史词语存于现代汉语之中。
- 1977 — 至今，占50%以上。
 - 新时期的语料代表了现代汉语的新发展。
- 每年按选材原则增补300万字左右的语料



语料的通用性原则

- 作为通用型语料库，应该比较真实地反映现代汉语在文字、词汇、语法、语义等方面的全貌。
 - 在语料的选择上，应当具有区别性特征。
 - 有别于专业性
 - 有别于地域性
 - 有别于纯口语性
- 尽可能地提高所选语料在采字、采词、采句和采义等方面的广度，要考虑到语料的时间层次、文化层次和社会使用面等层次。
 - 时间层次。
 - 文化层次。以具有高中文化程度的人能够阅读的语料为主。
 - 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；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。



语料抽样原则

■ 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。

■ 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。

■ 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。



语料抽样数量

■ 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。

■ 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。

■ 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。



语料抽样方式

- 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。



语料抽样的其他原则

- 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。
- 大学教材门类以国家规定的大学基础必修课为准。
- 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。



语料样本信息

□ 语料样本最多包含24个信息

a1 总号

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 出版日期

a12 期号

a13 版次(初版印数)

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 简繁体

a24 抽样文章



样例 语料相关信息

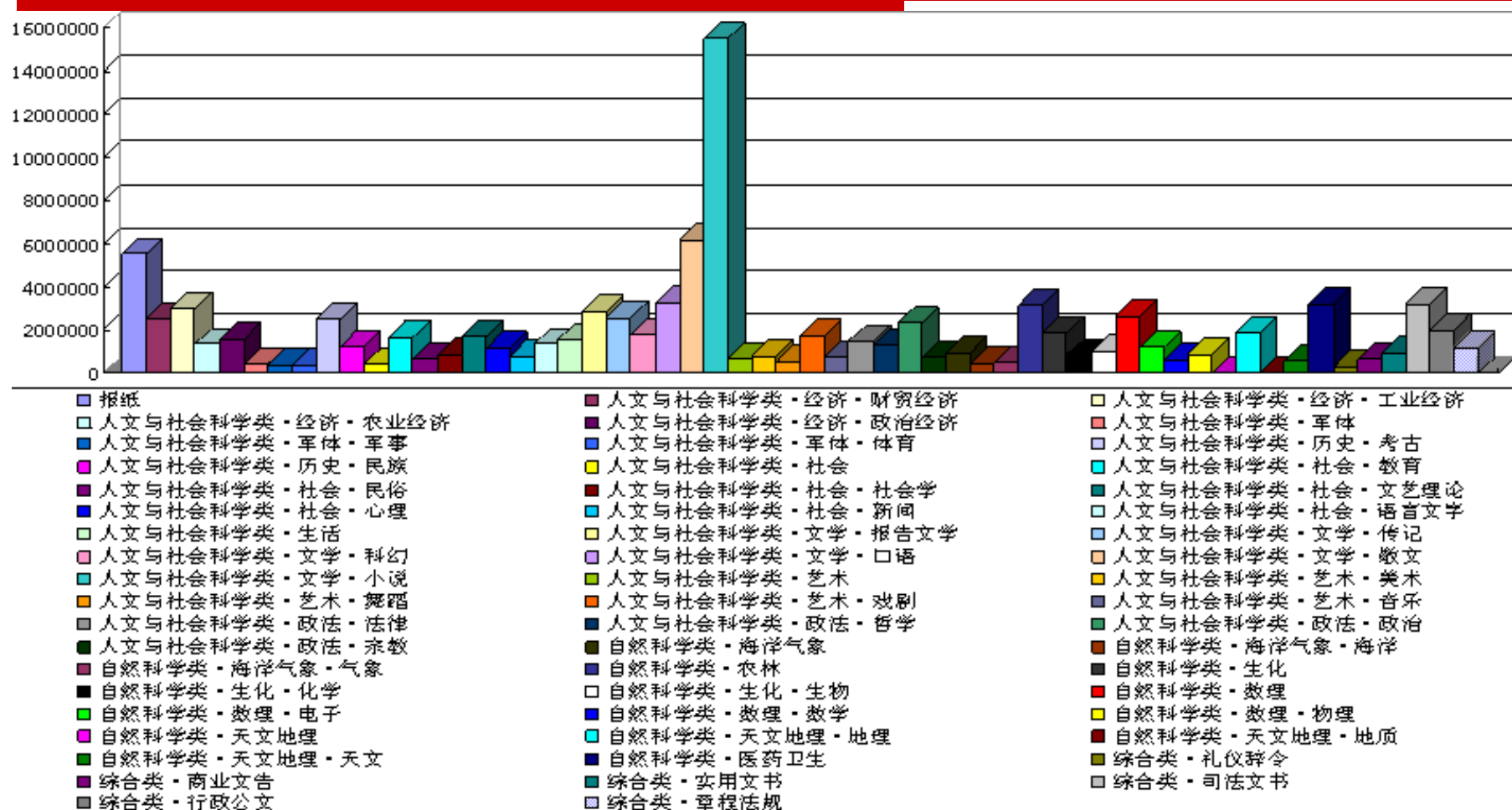
a2_分类号	a3_样本名称		a4_类别	a5_作者	a6_写作时间	a7_书刊名称	a8_编著者	a9_出版社	a10_所在省	a11_出版	
FH10005001	《我们的歌》节录		文学·小说	赵淑侠	1983-8-1	《我们的歌》	赵淑侠	友谊出版公司	北京	1983-9-1	
FH10005002	《我们的歌》节录		文学·小说	赵淑侠	1983-8-1	《我们的歌》	赵淑侠	友谊出版公司	北京	1983-9-1	
FH10005003	《我们的歌》节录		文学·小说	赵淑侠	1983-8-1	《我们的歌》	赵淑侠	友谊出版公司	北京	1983-9-1	
FH10005004	《我们的歌》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	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
FH10002205	1	14000	14000	186	0	简单随机抽样	88-91	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	68-70	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



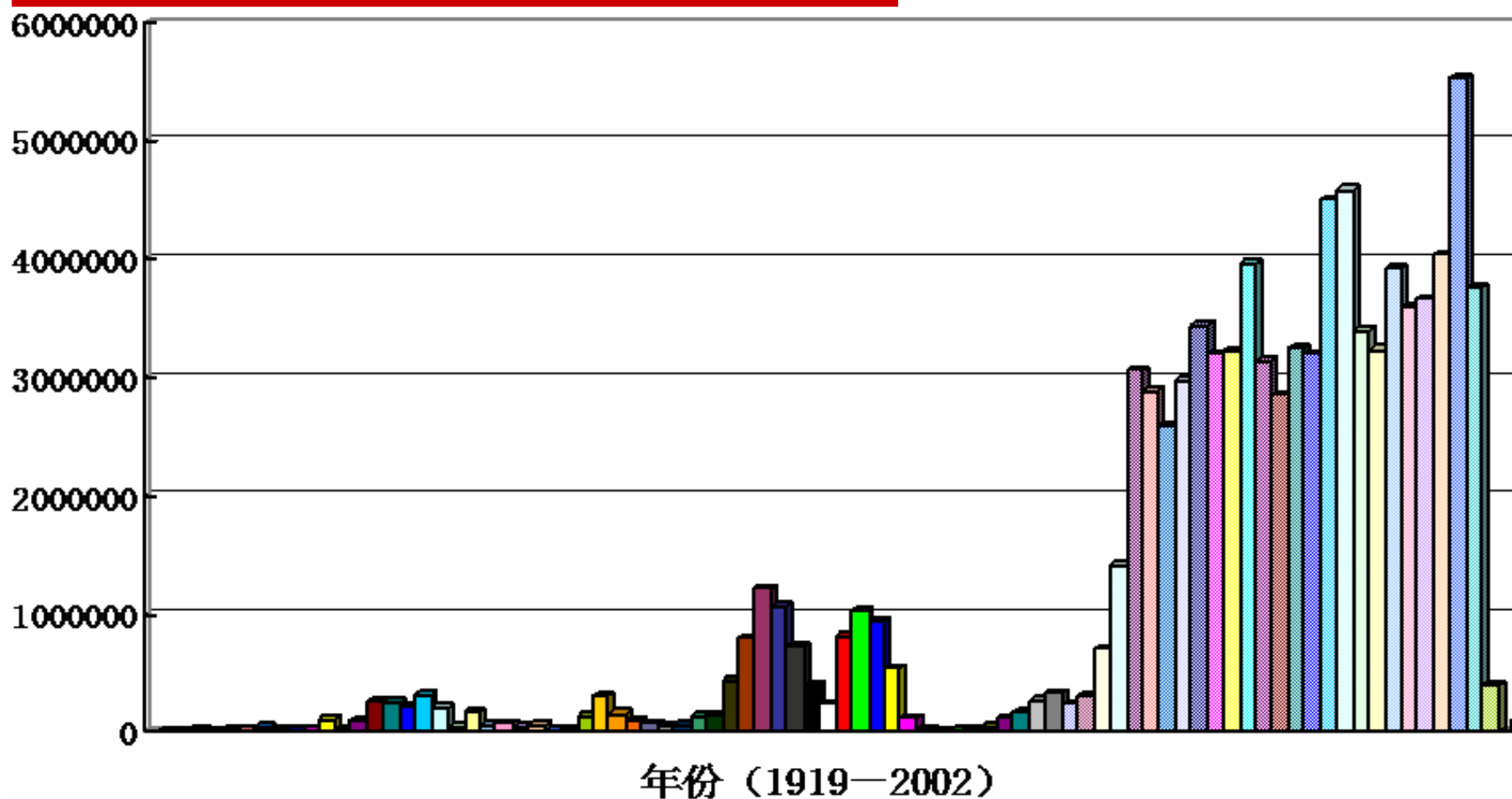
教育部语言文字应用研究所

INSTITUTE OF APPLIED LINGUISTICS, MINISTRY OF EDUCATION
WWW.AVERC.CN

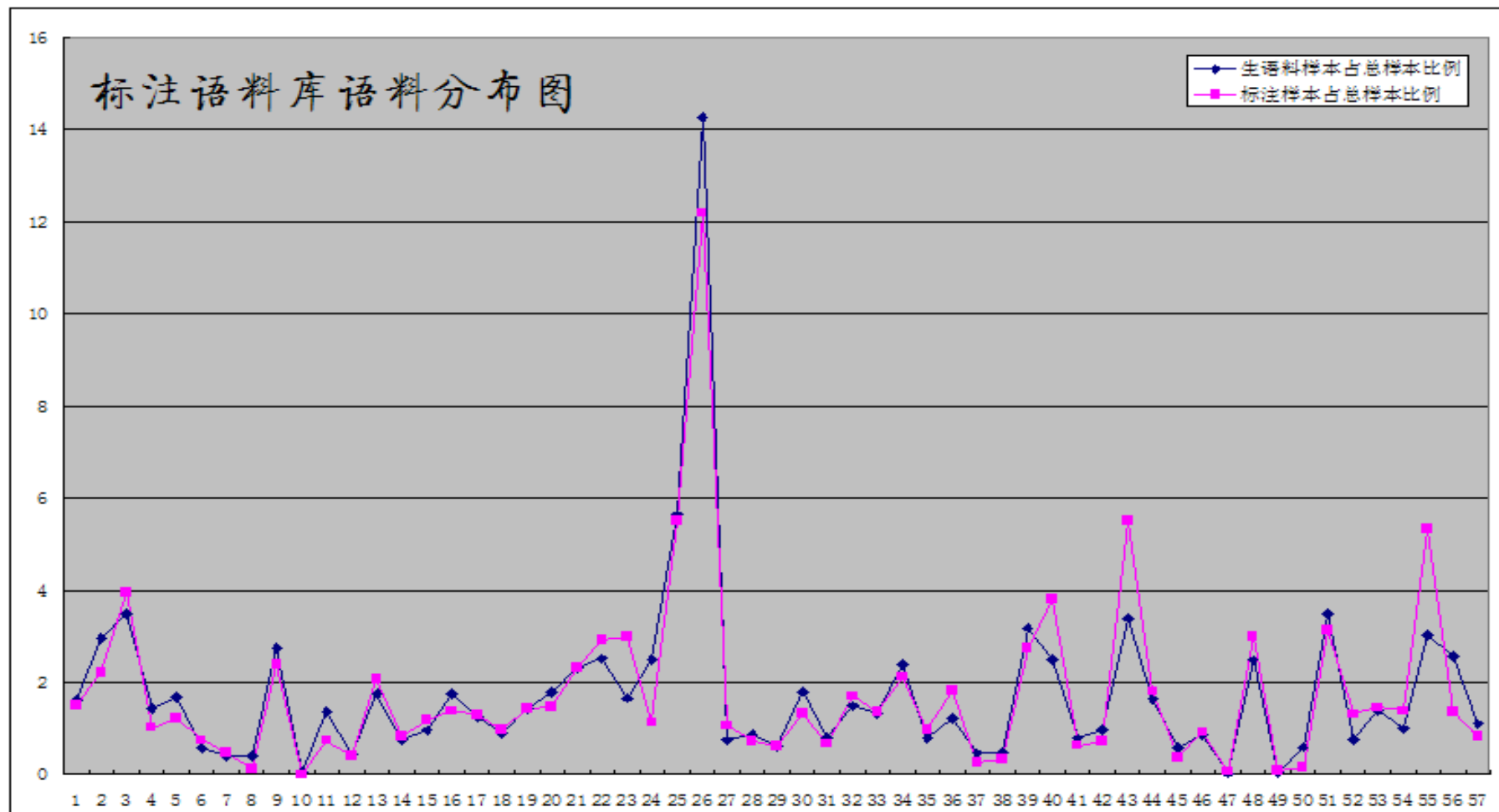
语料库样本分布一类别



语料库样本分布—时间



标注语料的类别分布



语料库的建设和加工

- 遵循国内外信息处理领域通用的语料库加工方式，标注分词和词性。
- 在面向信息处理的同时，重视为语言本体研究服务。
- 在加工过程中制定了《信息处理用词类标记集规范》等语料库建设规范。
- 采用机助人校的加工方式，通过开发专门软件工具来提高加工质量。
- 为兼容不同词语颗粒度，专门建立了层次化结构化的分词词表，可以提供不同颗粒度的语料。
- 语料库以每年增加约300万字的方式进行补充



语料库建设的主要成果

- 生语料库规模达到1亿字
- 可供使用的标注语料5000万字（分词和词性标注）
- 100万字（5万句）句法树库
- 现代汉语分词词表（88000词条）
- 语料切分标注加工规范
- 树库标记集规范
- 树库加工规范
- 语料切分和标注软件
- 树库标注软件
- 语料库校对加工软件
- 语料检索工具软件
- 语料统计工具软件
- 语料库管理软件



信息处理用现代汉语词类标记集规范

- 基本词类体系
- 基本词类体系的标记代码
- 《规范》的制定
 - 在国家社科基金“九五”重大项目《信息处理用现代汉语词汇研究》的子项目“信息处理用现代汉语词类标记集规范的基础上完成
 - 得到国家语委“九五”重大项目《现代汉语语料库建设》子课题“国家语委核心语料分词及词性标注加工”的支持。
- 2003年通过国家语委标准化委员会的审定
- 2006年成为国家标准，标准号GB/T 20532-2006
- 词类标记集规范的原则是有利于数据交换和资源共享



样例 分词和词性标注语料

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/n 的/u 世界/n ↵

杨栋/nh ↵

鸟/n , /w 是/vl 大自然/n 的/u 歌手/n , /w 鸟语/n 就是/vl 大自然/n 的/u 音乐/n 和/c 诗歌/n 了/u 。 /w ↵

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



样例 词类标记集

附表：词类及标记代码说明

序号	标记代码	词类名称	说明
1	n	名词	<u>n</u> oun
2	ng	普通名词	<u>n</u> oun-general
3	nt	时间名词	<u>n</u> oun- <u>t</u> ime
4	nd	方位名词	<u>n</u> oun- <u>d</u> irection
5	nl	处所名词	<u>n</u> oun- <u>l</u> ocation
6	nh	人名	<u>n</u> oun- <u>h</u> uman
7	nhf	姓	<u>n</u> oun- <u>h</u> uman- <u>f</u> amily name
8	nhg	名	<u>n</u> oun- <u>h</u> uman- <u>g</u> iven name
9	ns	地名	<u>n</u> oun- <u>s</u> pace
10	nn	族名	<u>n</u> oun- <u>n</u> ation
11	ni	机构名	<u>n</u> oun- <u>i</u> nstitution
12	nz	其他专名	<u>n</u> oun-“专”拼音的首字母
13	v	动词	<u>v</u> erb



样例 句法树库标记集（1）

□ （一）短语功能分类

功能名称↵	标记↵	示 例↵	说 明↵
名词短语↵	np↵	[FW 学校/n 里/nd]np ↵ [DZ 今年/nt 春天/nt]np ↵ [DZ[ZC 中国/ns 的/u]np 解放/v]np ↵ [DZ[ZC[DZ 小/a 时候/nt]np 的/u]np ↵ 我/r]np [LH 我/r 和/c 他/r]np↵	相当于名词或以名词为中心语。↵
数词短语↵	mp↵	[LH 七/m、 /w 八/m]mp↵	大部分已经并成数词，这里可以指联合式，如“七、八（个）”。↵
量词短语↵	qp↵	[SL 一/m 个/q]qp ↵ [SL 一/m 个个/q]qp ↵ [SL 这/r 个/q]qp ↵ [SL 这/r[SL 一/m 个/q]qp]qp↵	由代词或数词等+量词组成的短语。 量词重叠如“个个、条条”应该只是词，不算短语。↵
动词短语↵	vp↵	[CD 研究/v 研究/v]vp ↵ [ZZ 努力/a 学习/v]vp ↵ [ZZ[JB 从/p 北京/ns]pp[ZZ[JB 经/p 天津 ns]pp[SB 到/v 上海/ns]vp]vp ↵ [ZZ 十一/m[SB 至/v 十五/m]vp]vp↵	以动词为中心语。↵
形容词短语↵	ap↵	[BC 高兴/a 得/u 很/d]ap↵ [ZZ 不/d 好/a]ap↵	以形容词为中心语。形宾短语，如“高他一头(双宾结构)，(比它)长三尺，红了脸”标示宾结构形容词短语。↵



样例 句法树库标记集（2）

□ （二）短语结构分类

结构名称↵	标记↵	示 例↵	说 明↵
补充结构↵ (动补、形补)↵	BC↵	[BC 高兴/a 得/u 很/d]ap [BC 好/a 极/d 了/u]ap [BC 说/v 清楚/a]vp↵	由动词或形容词+（得+）补语（一般也是由动词或形容词充当）组成的动词或形容词短语。有“得”的时候是三分结构。↵
标号结构↵	BH↵	[BH 《/w[ZZ[JB 为/p 人民 /n]pp 服务/v]vp》 /w]np↵	凡是由一个普通短语+一个标点符号组成的短语。当标号为书名号时整个短语的功能类型为名词短语 np。↵
重叠结构↵	CD↵	[CD[SL 一/m 个/q]qp[SL 一 /m 个/q]qp]qp ↵ 研究研究↵	由完全相同的两部分（词或短语）组成的短语。↵
定中结构↵	DZ↵	[DZ 我们/r 学校/n]np [DZ 中国/ns 上海/ns]np [DZ[SL 一/m 丝/q]qp 温柔 /a]np↵	由定语+中心语组成的名词短语。↵
方位结构↵	FW↵	[FW 学校/n 里/nd]np [FW 出发/v 前/nd]np↵	由名词或动词等+方位名词组成的名词短语。↵



样例 句法树

ChineseParser - Parser1_资产阶级夺取了革命的果实。 .rlt

文件(F) 编辑(E) 查看(V) 窗口(W) 帮助(H)

Parser1.inp

资产阶级/n 夺取/v 了/u 革命/v 的/u 果实/n 。 /w

Parser1_资产阶级夺取了革命的果实。 .rlt

分析结果1

- 句法结构
 - ROOT []
 - DJ [BH]
 - SP [ZW]
 - n [] 资产阶级
 - VP [SB]
 - VP [ZC]
 - v [] 夺取
 - u [] 了
 - NP [DZ]
 - NP [ZC]
 - v [] 革命
 - u [] 的
 - n [] 果实
 - w [] 。

文件(F) 操作(O) 查看(V) 帮助(H)

2 [BH] [ZW] 肖邦/nh [ZZ] [ZC] 愤慨/a 地/u]DP [ZZ] 。

节点编辑 上一句

文本 树(横) 树(竖) 问题 日志 保存句子 下一句

+ROOT

+DJ BH

+SP ZW

-nh

+VP ZZ

+DP ZC

+VP ZZ

+VP SB

-a -u -d

-v -v

肖邦 愤慨 地 断然 加以 拒绝 。

样例 结构化词表

| 序号↵ | 词↵ | 主要词类↵ | 结构↵ |
|-----|-------|-------|-----------------|
| 1↵ | 工具↵ | n↵ | 工具/n↵ |
| 2↵ | 工具包↵ | n↵ | [工具/n 包/n]/n↵ |
| 3↵ | 工具栏↵ | n↵ | [工具/n 栏/n]/n↵ |
| 4↵ | 工具书↵ | n↵ | [工具/n 书/n]/n↵ |
| 5↵ | 工具箱↵ | n↵ | [工具/n 箱/n]/n↵ |
| 6↵ | 工科↵ | n↵ | 工科/n↵ |
| 7↵ | 工矿↵ | jn↵ | 工矿/jn↵ |
| 8↵ | 工矿企业↵ | n↵ | [工矿/jn 企业/n]/n↵ |
| 9↵ | 工联↵ | jn↵ | 工联/jn↵ |
| 10↵ | 工龄↵ | n↵ | 工龄/n↵ |
| 11↵ | 工贸↵ | jn↵ | 工贸/jn↵ |
| 12↵ | 工贸结合↵ | n↵ | [工贸/jn 结合/v]/v↵ |
| 13↵ | 工农↵ | jn↵ | 工农/jn↵ |
| 14↵ | 工农兵↵ | jn↵ | 工农兵/jn↵ |
| 15↵ | 工农红军↵ | n↵ | [工农/jn 红军/n]/n↵ |
| 16↵ | 工农联盟↵ | n↵ | [工农/jn 联盟/n]/n↵ |
| 17↵ | 工农业↵ | jn↵ | 工农业/jn↵ |
| 18↵ | 工棚↵ | n↵ | 工棚/n↵ |
| 19↵ | 工期↵ | n↵ | 工期/n↵ |
| 20↵ | 工钱↵ | n↵ | 工钱/n↵ |

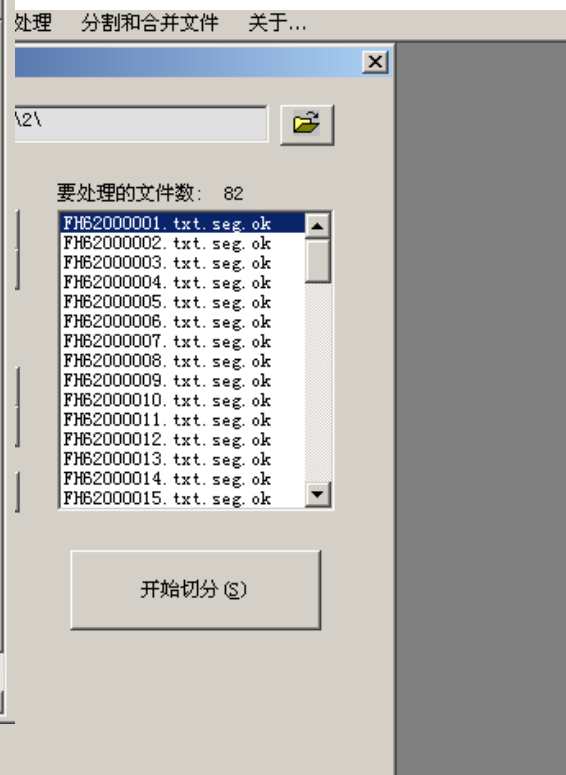
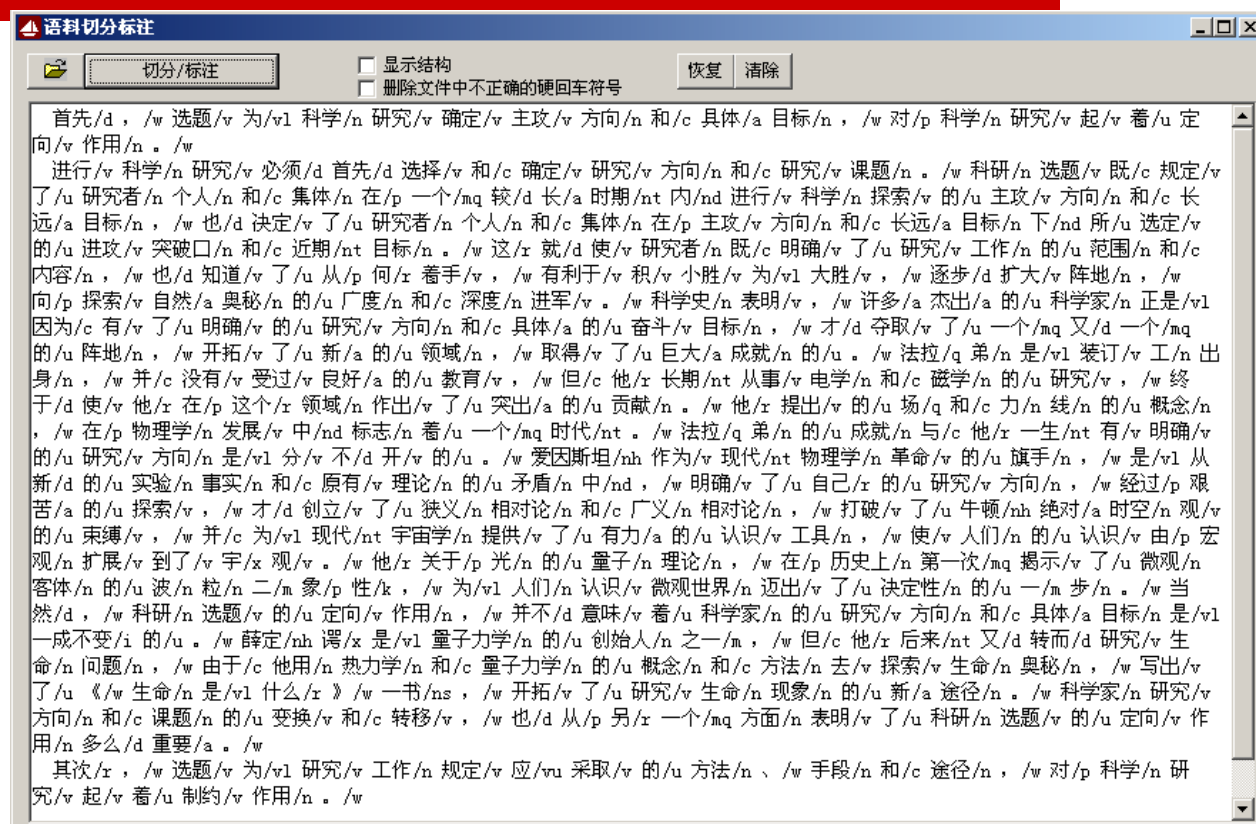


语料库应用软件

- 语料切分和标注软件
- 树库标注软件
- 语料库校对加工软件
- 语料检索工具软件
- 语料库查询与统计工具软件
- 基于互联网的语料库例句检索



样例 语料切分和标注软件



样例 语料库校对和质量检查

现代汉语语料库切分标注校对工具 - [xiaohang] - [D:\wangxy\1\BR00105301.txt.Seg.rtf]

文件操作 环境设置 查找替换 自由编辑 撤销 帮助 字符总数: 6969 当前位置: 4459 文本字数: 2334

打开新文件(Open File) 保存当前文件(Save File) 打开最近一次编辑文件(Open recent) 启用自动保存的文档(GetAutosave) 关闭文件(Close) 退出(Exit)

环境/nt 6/m 月/nt 开始/v , /w 全/a/a 市/ns/n 先后/d 涌现/v 出/vd 多/a 名/n/q 污
他们/r 暗中/d 检查/v 企业/n 非污/v 情况/n , /w 随时/d 向/p 环保部门/n
局/n 现/v/d 已/d 兑现/v 奖金 /v m 多/a 万元/nh/mq , /w 处罚/v 企业/n 数百
/v 设施/n 的/u 正常/a 运转/v 从/p 原来/d 的/u 35/m %/w 提高/v 到/v 现
一度/d 污浊/a 不堪/v 的/u 河水/n 重新/d 变/v 清/a 了/u 。 /w 江苏省/ns 在/p
总结/v 富/a 阳/n/ns 经验/n 的/u 基础上/n1 , /w 于/p 今年/nt 2/m 月/nt 在/p 全国/n 率先/d 实行/v 了/u
全省/n 范围/n 内/nd 的/u 污染/v 有奖/v 举报/v 。 /w 省/n 环保/j 厅/n 和/c 13/m 个/q 省辖市/n 、 /w 所
有/a 县/n (/w 市/n) /w 的/u 环保局/n , /w 分别/d 向/p 社会/n 公布/v 了/u 污染/v 有奖/v 举报/v 电话/n
。 /w 截至/v 7/m 月底/nt , /w 江苏省/ns 各级/n 环保部门/n 收到/v 的/u 污染/v 举报/v 多达/v 15451/m 次/q
 , /w 共/d 查处/v 违法/v 企业/n 11959/m 家/n/q (/w 次/q) /w 。 /w 眼下/nt , /w 这/r 一/m 措施/n 已/d
在/p 许多/a 省份/n 推广/v 。 /w 国家/n 环保/j 总局 统一/v 使用/v 的/u 举报/v
电话/n : /w 12369/m , /w 目前/nt 已/d 在/p 38/m 万绿色/n 社区/n : /w
把 /w 你/r 拍/v 一/n 家/n 分类/v 扔/v 垃圾/n
 ; /w 白/v 二/m , /w 垃圾/n
 ; /w 分类/v 歌/n 》 /w
 , /w 因为/c 它/r
 n 行动/n 的/u 真实/a 记录/v/n /w 建功/v 南/nd 甲/nd/ns 小区/n 是/v
 建/v 小区/n , /w 有
 n 文化/n 中心/n
 /w 宣传/v 教育/v
 的/u 生活/n ” /w 的/u 口号/n , /w 从/p 垃圾/n 分
 , /w 发动/v 市民/n 参与/v 环保/j 。 /w 许静林/nh
 办法/n : /w 洗/v 菜/n 淘/v 米/v/n 的/u 水/n 用来
 桶/n , /w 洗澡/v 的/u 水/n 用来/v 拖地/v 。 /w 他
 吨/q 降至/v 4/m 吨/q 左右/n 。 /w 北京市/ns 精神文
 心/n 联合/v 编制/v 了/u 《 /w 绿色/n 社区/n 手册/n 》 /w , /w 将/d 绿色/n 小区/n 建设/v/n 纳入/v 规范
 化/v 、 /w 标准化/v 的/u 轨道/n 。 /w 绿色/n 社区/n 在/p 其他/r 城市/n 也/d 开展/v 得/u 如火如荼/i 。 /w

查找和替换

查找内容: 查找下一个(F) 替换为: 替换(R) 向下 向上 全部

自由编辑

月/nt 开始/v , /w 全/a/a 市/ns/n 先后/d 涌现/v

确定(C)

月/nt 开始/v , /w 全/a/a 市/ns/n 先后/d 涌现/v

格式检查发现校对错误, 请校正

| ID | 标注单位 | 校对标记 |
|----|-------|-------------|
| 1 | 府 | 府<D#/j#D> |
| 2 | 为/pp | 为 |
| 3 | 全/a/a | 全<E#/a/a#E> |



样例 语料库例句检索

现代汉语语料库例句检索 CylkSearch

检索方式 检索设置 保存结果 显示/隐藏查询条件窗口 显示/隐藏例句信息窗口 帮助

设定主查询条件

请输入字符:
半导体

☒ 按词检索
☐ 按词/词类检索
☐ 按词类检索
☐ 按字符串检索

开始检索 (Start)

附加查询条件

同现字/词/词类:
+
☒ 前后均可
☐ 在前 ☐ 在后
间隔 50 个字符内

同现词列表:
清除所有 移除一项

检索结果 [第 01 页] 下一页 [Next Page]

对齐方式: 关键词居中

/n 做/v 了/u 一套/mq ' /w 的确/n ' /w 衬衫/n 衬裤/n , /w 送/v 一/m 台/q 收音机/n 。 /w
人们/n 只要/vu 手/n 握/v 一/m 只/q 形/n 如/v 半导体/n 的/u H J /ws 型/k 通用/v 袖珍/a 货币/n 鉴别/v 器/n , /w 按/v 亮/a 上面/nd 的/u 紫/a
我国/n 有/v 了/u 最大/a 的/u 半导体/n 研究/v 主楼/n
由/p 市城建四公司/ni 承建/v 的/u 我国/n 目前/nt 最大/a 的/u 半导体/n 研究/v 主楼/n 近日/nt 在/p 本/r 市/n 北部/nl 建成/v 。 /w
它/r 格/d 为/p 加快/v 发展/v 我国/n 的/u 半导体/n 科学/n 技术/n 和/c 大规模集成电路/i 的/u 研究/v 提供/v 现代化/v 的/u 实验/n 基地/n 。 /w
8 /m 年/nt 批准/v , /w 在/p 本市/n 北部/nl 新建/v 一/m 座/q 现代化/v 的/u 半导体/n 研究/v 基地/n , /w 建筑/n 总面积/n 为/vl 7 5 0 0 0 /m 多/m 平方/m 米/q , /w 由/p 三十一/m 个/q 子项
行政/n 经费/n 一万零八百六十四/m 元/q 购买/v 了/u 毯子/n 、 /w 雨衣/n 、 /w 半导体/n 收音机/n 、 /w 皮鞋/n 、 /w 雨鞋/n 、 /w 煤/n 火炉/n 、 /w 被褥/n 、 /w 蚊帐/n 、 /w 工作服/n 、 /w 手3
元件/n 中/nd 大功率/n 晶体管/n 过去/nt 都/d 用/p 金属/n 封/v 装/v , /w 上海有色金属压延厂/ni 研
半导体/n 激光/n 治疗/v 羊/n 的/u 关节炎/n 、 /w 羔羊/n 的/u 脐带/n 炎/n 及/c 羔羊/n 的/u 进步/v , /w 也/d 获
半导体/n 、 /w 集成电路/n 、 /w 大规模/n 和/c 超/v 大规模集成电路/i 几/m 代/nt 的/u 进步/v , /w 计算机/n 的
半导体/n 集成电路/n 是/vl 电子/n 信息/n 产品/n 的/u " /w 心脏/n " /w , /w 美国/ns 半导体/n 工业/n 协会/n 研
部/n 商务/n 经理/n 任雪梅/nh 在/p 接受/v 记者/n 采访/v 时/nt 也/d 强调/v , /w 建设/v 大规模集成
里/nd 播放/v 着/u " /w 血染/v 的/u 风采/n " /w , /w 又/d 一/m 次/q 勾起/v 了/u 她/r 的/u 伤心/v
半导体/n 工业/n 自/p 9 0 /m 年代/nt 初/n 恢复/v 对/p 日本/ns 的/u 世界/n 霸主/n 地位/n 以来/nt , /w 美国/
半导体/n 工业/n 附加值/n (/w 销售/v 收入/n 扣除/v 材料/n 、 /w 水电/n 等/u 费用/n) /w 由/p 1 1 2 亿/m 美
半导体/n 产品/n 价格/n 持续/v 下降/v 所/u 造成/v 的/u 消费者/n 购买力/n 提高/v , /w 就/d 相当于/u 美国/ns
半导体/n 技术/n 不仅/c 以/p 自身/n 的/u 知识/n 创新/v 直接/a 改变/v 了/u 经济/n 增长/v 质量/n , /w 而且/c
半导体/n 技术/n 所/u 具有/v 的/u 高性能/n 与/c 低/a 价格/n 的/u 连锁/n 推进/v 机制/n , /w 早期/nt 这些/r
" /w 小/a 与/c 大/a " /w 、 /w " /w 高/a 与/c 低/a " /w 辩证/a 关系/n 或/c 摩尔定律/n 作用/v 下/nd ,
半导体/n 技术/n 创新/v , /w 则/c 不仅/c 以/p 自身/n 的/u 集成/v 技术/n 叩开/v 了/u 知识经济/n 大门/n , /w
半导体/n 技术/n 、 /w 超/v 平面/n 显像/v 技术/n 、 /w 等离子/n 显像/v 技术/n 以及/c 节能/v 照明/v 技术/n 等
半导体/n 制造/v 处理/v 工艺/n , /w 其/r 独特/a 功能/n 可/vu 帮助/v 厂商/n 生产/v 手机/n 、 /w 个人/n 数字/r
半导体/n 材料/n 、 /w 能源/n 材料/n 、 /w 环境/n 材料/n 、 /w 纳米/n 材料/n 、 /w 超导材料/n 、 /w 生物/n 及/c
半导体/n 器件/n 、 /w 中子/n 辐照/n 育种/v 、 /w 中子/n 治疗/v 癌症/n 、 /w 中子/n 调/v 井/n 等/u 等/a 。 /w
半导体/n 材料/n 及/c 电子/n 材料/n , /w 新/a 金属/n 材料/n 陶瓷/n 材料/n , /w 复合材料/n , /w 高分子/n 材料/
半导体/n 叩开/v 知识经济/n 大门/n (/w 院士/n 园地/n) /w (/w 附/v 图片/n 1 /m 张/q) /w 半导体/n 叩开/
半导体/n 工程/n 技术/n 开发/v 和/c 工业/n 发展/v 工作/n , /w 8 0 /m 年代/nt 后/nd , /w 探索/v 科研/n 生产
半导体/n 技术/n 的/u 发展/v , /w 使/v 所有/a 其他/r 技术/n 都/d 相形见绌/i , /w 它/r 作为/v 数字化/v 革命
半导体/n 表面/nd 物理/n 时/nt , /w 在/p 一块/n 硅/n 片/n 上/nd , /w 发现/v 了/u 既/c 有/v 电子管/n 跨/v 1
半导体/n 研究/v 和/c 发现/v 晶体管/n 效应/n " /w 方面/n 的/u 贡献/n , /w 把/p 当年/nt 的/u 物理学/n 奖/n 3
半导体/n 科学/n 技术/n 却/d 获得/v 了/u 辉煌/a 成就/n 。 /w
半导体/n 实验室/n 以及/c 随后/d 的/u 一系列/a 裂变/v 与/c 繁衍/v , /w 触发/v 了/u 半导体/n 工业/n 的/u 包
半导体/n 技术/n 发展/v 中/nd , /w 有/v 两/m 个/q 反映/v 创新/v 的/u 规律/n , /w 一个/mq 是/vl 指导/v 技术
半导体/n 器件/n 尺寸/n 从/p 早期/nt 的/u 十几/m 微米/q , /w 缩小/v 到/v 现在/nt 的/u 深/a 亚/a 微米/q ; /
半导体/n 发展/v 的/u 这些/r 规律/n , /w 在/p 体现/v 科学/n 理论/n 和/c 技术/n 商品化/v 的/u 有机/a 结合 ;
半导体/n 激光/n 用于/v 蔬菜/n 栽培/v 揭/p 《 /w 日/j 经/v 产业/n 新闻/n 》 /w 报道/v , /w 日本东海大学/ni 3
半导体/n 材料/n 及/c 半导体/n 物理/n 专家/n , /w 她/r 的/u 科研/n 成果/n 为/p 我国/n 微电子/n 和/c 光电子
半导体/n , /w 才能/c 抛弃/v 电子管/n , /w 生产/v 集成电路/n 。 /w
半导体/n 材料/n 。 /w
半导体/n 材料/n 。 /w
半导体/n 技术/n 和/c 生物材料/n " /w 联结/v " /w 的/u 结晶/n , /w 由于/c 它/r 可能/vu 形成/v 巨大/a 产业/n
半导体/n 技术/n 和/c 生物材料/n 专家/n , /w 她/r 的/u 科研/n 成果/n 为/p 我国/n 微电子/n 和/c 光电子
半导体/n 激光/n 照射/v 培育/v 水稻/n 取得/v 成功/a 。 /w

#[例句出处]:
[样本名称]: 半导体叩开知识经济大门; [作者]: 许居衍; [写作时间]: N/A; [出版时间]: 1998-9-1; [书刊名称]: 人民日报; [编著者]: N/A; [出版社]: 人民日报社
#[上下文信息]:
[1]: 半导体发展的这些规律,在体现科学理论和技术商品化的有机结合中,反映了半导体技术独特的“小与大”、“高与低”新技术经济辩证法,即在尽可能小的材料空间内,集成尽可能强大的功能,以实现高性能与低价格的正反馈良性推进关系。
[2]: 正是这种辩证关系,促使人类经济特征由“回报递减”(Diminishing returns)过渡到“回报递增”上来。
[3]: 例如,美国半导体工业自90年代初恢复对日本的世界霸主地位以来,美国经济也正好同期出现了繁荣增长的趋势。
[4]: 从1987年到1996年,美国半导体工业附加值(销售收入扣除材料、水电等费用)由11.2亿美元增加到41.6亿美元(1996年销售额为709亿美元),增加值达30.4亿美元,比制药业、汽车业同期增加值高出50%以上;10年间年均增长率达15.7%,比整个经济快3倍多。
[5]: 近20年来,因半导体产品价格持续下降所造成的消费者购买力提高,就相当于美国国内生产总值的5%,从而使知识经济最发达的美国保持了高增长、低通胀的良好发展态势。

样例 语料库例句检索 2

语料库查询及例句抽取工具 - V2.0

语句抽取 (Extract) 批处理 (Batch) 说明 (Readme) 关于 (About)

国家语委现代汉语平衡语料库语句抽取

输入查询字符串: 语言

附加查询条件: 同现词: 文字

查询倾向: ☒ 查准 ☐ 查全

停止 (Stop)

15 个满足条件

正在查询... 02%

国家语委现代汉语平衡语料库语句抽取

输入查询字符串: 语言

附加查询条件: 同现词: 文字

查询倾向: ☒ 查准 ☐ 查全

执行 (Execute)

27 个满足条件

导出查询结果到文件 (Export)

| 序号 | 例句 |
|----|---|
| 1 | 语言文字工作是我国现代化建设事业中不可缺少的组成部分, 需要规范化和标准化。 |
| 2 | 但长期以来, 由于缺少相应的法律、法规, 语言文字应用不规范、不标准的现象十分严重, 亟需要把 |
| 3 | 国家语委建议由本委员会牵头起草《语言文字法》, 本委同意这一建议, 决定承担牵头起草该法的任 |
| 4 | 从某种意义上讲, 伊斯兰教的传播, 既是语言文字的传播, 也是文化的传播。 |
| 5 | 我们的事业需要大批的宣传家、教育家, 即使如此, 在联系实际方面, 在语言文字的风格方面还是可 |
| 6 | 维吾尔族有自己的语言文字。 |
| 7 | 彝族有自己的语言文字。 |
| 8 | 白族有自己的语言, 白语属汉藏 |
| 9 | 柯尔克孜族有自己的语言和文字。 |
| 10 | 《玛纳斯》不仅是一部以描写战 |
| 11 | 第三, 城市少数民族的风俗习惯 |
| 12 | 政府出台了各种政策措施和相应 |
| 13 | 各种媒体如语言、文字、画图、 |
| 14 | 在此过程中, 文字不断繁衍、滋 |
| 15 | 他们垄断祭祀、卜筮, 秉承国君 |
| 16 | 这种迥异于其他语言文字的特征 |

| 序号 | 来源文章名 | 作者 | 写作日期 | 书刊名称 |
|----|-------------|------|-----------|------------|
| 1 | 全国人大教育科学文化 | 全国人大 | | 《全国人大常委会 |
| 2 | 全国人大教育科学文化 | 全国人大 | | 《全国人大常委会 |
| 3 | 全国人大教育科学文化 | 全国人大 | | 《全国人大常委会 |
| 4 | 伊斯兰教在中亚的传播 | 常玲 | | 东欧中亚研究 历 |
| 5 | 《哲人哲思——漫步遐 | 程先达 | | 哲人哲思——漫步 |
| 6 | 维吾尔族 | 郭明轩 | 1996-6-29 | 《人民日报》海外 |
| 7 | 彝族 | 郭明轩 | 1996-9-14 | 人民日报 (海外版) |
| 8 | 白族 | 郭明轩 | 1997-5-3 | 人民日报 (海外版) |
| 9 | 柯尔克孜族 | 郭明轩 | 1998-8-8 | 光明日报 |
| 10 | 柯尔克孜族 | 郭明轩 | 1998-8-8 | 光明日报 |
| 11 | 少数民族流动人口与城 | 周竟红 | 2001-4-1 | 民族研究 |
| 12 | 少数民族流动人口与城 | 周竟红 | 2001-4-1 | 民族研究 |
| 13 | 人与历史 | 王维明 | 1999-1-1 | 文史杂谈 |
| 14 | 原始时代: 中国文化的 | | | |
| 15 | 原始时代: 中国文化的 | | | |
| 16 | 原始时代: 中国文化的 | | | |
| 17 | 透视网上聊天 | 彭 兰 | 1999-7-28 | 人民日报 |
| 18 | 讲究语言艺术 | 张永琇 | 2001-2-12 | 光明日报 |
| 19 | 简论语言文字立法的意 | 江蓝生 | 2001-1-16 | 光明日报 |
| 20 | 简论语言文字立法的意 | 江蓝生 | 2001-1-16 | 光明日报 |
| 21 | 简论语言文字立法的意 | 江蓝生 | 2001-1-16 | 光明日报 |
| 22 | 简论语言文字立法的意 | 江蓝生 | 2001-1-16 | 光明日报 |
| 23 | 简论语言文字立法的意 | 江蓝生 | 2001-1-16 | 光明日报 |
| 24 | 简论语言文字立法的意 | 江蓝生 | 2001-1-16 | 光明日报 |



样例 语料库查询统计工具

语料库统计-检索语料

第一步 选择语料库

选择语料库: 国家语委语料库-标注语料库

☐ 全语料库

☒ 选择语料库

语料库选项

可选项:

- ☒ 时间
- ☐ 作者
- ☐ 出版社
- ☒ 类别

时间值

写作时间:

从: 1919-1-1

至: 2005-3-22

出版时间:

从: 1919-1-1

至: 2005-3-22

格式: 年月日 (如: 2005-1-1)

语料库查询结果

语料库信息:

| 序号 | 分类号 | 样本名称 | 类别 | 作者 |
|----|------------|----------------------|----|------|
| 1 | FE10005202 | 《工业企业管...》第2, 3, 1 | 经济 | 李铁城 |
| 2 | FE20005801 | 《农业经济学讲义》第2, 4, 5, 6 | 经济 | 中国人民 |
| 3 | FE20005802 | 《农业经济学讲义》第2, 4, 5, 6 | 经济 | 中国人民 |
| 4 | FE20005803 | 《农业经济学讲义》第2, 4, 5, 6 | 经济 | 中国人民 |

语料库统计-检索语料

第一步 选择语料库

选择语料库: 国家语委语料库-标注语料库

☐ 全语料库

☒ 选择语料库

第二步 语料库统计

语料库: 统计项目

开始统计 (Execute)

导出 (Export)

语料库统计-检索语料

语料库信息:

| 序号 | 分类号 | 样本名称 | 类别 |
|----|------------|-------------------|----|
| 1 | PH10027701 | 《危栏》节录 | 文学 |
| 2 | PH10027901 | 《绝症》节录 | 文学 |
| 3 | PH10105304 | 《瓦砾堆》节录 | 文学 |
| 4 | PH10105305 | 《瓦砾堆》节录 | 文学 |
| 5 | PH10105306 | 《瓦砾堆》节录 | 文学 |
| 6 | PH10105401 | 《山雨》第6, 16, 28章节录 | 文学 |
| 7 | PH10027601 | 《主义》节录 | 文学 |
| 8 | PH10001904 | 《鹿回头》节录 | 文学 |
| 9 | PH10001905 | 《鹿回头》节录 | 文学 |
| 10 | PH10002201 | 《柳暗花明》节录 | 文学 |
| 11 | PH10002202 | 《柳暗花明》节录 | 文学 |
| 12 | PH10086601 | 《药娘和药女》节录 | 文学 |
| 13 | PH10090304 | 《皇城根》节录 | 文学 |
| 14 | PH10090305 | 《皇城根》节录 | 文学 |
| 15 | PH10090306 | 《皇城根》节录 | 文学 |
| 16 | PH10057501 | 《暗礁》节录 | 文学 |
| 17 | PH10057601 | 《父亲》节录 | 文学 |

语料库统计-检索语料

统计结果:

| 序号 | 字词 | 出现次数 | 频率 |
|----|----|------|--------|
| 1 | 的 | 550 | 4.9937 |
| 2 | 是 | 260 | 2.3846 |
| 3 | 了 | 265 | 2.3579 |
| 4 | 他 | 247 | 2.1977 |
| 5 | 不 | 218 | 1.9397 |
| 6 | 一 | 182 | 1.6194 |
| 7 | 有 | 120 | 1.0677 |
| 8 | 她 | 113 | 1.0054 |
| 9 | 们 | 107 | .952 |
| 10 | 她 | 106 | .9431 |
| 11 | 人 | 99 | .8809 |
| 12 | 我 | 94 | .8364 |
| 13 | 说 | 88 | .783 |
| 14 | 在 | 78 | .694 |
| 15 | 这 | 76 | .6762 |
| 16 | 来 | 74 | .6584 |
| 17 | 也 | 74 | .6584 |
| 18 | 个 | 72 | .6406 |

语料库统计-检索语料

统计结果:

| 序号 | 字词 | 出现次数 | 频率 |
|----|----|------|--------|
| 1 | 的 | 550 | 4.9937 |
| 2 | 是 | 260 | 2.3846 |
| 3 | 了 | 265 | 2.3579 |
| 4 | 他 | 247 | 2.1977 |
| 5 | 不 | 218 | 1.9397 |
| 6 | 一 | 182 | 1.6194 |
| 7 | 有 | 120 | 1.0677 |
| 8 | 她 | 113 | 1.0054 |
| 9 | 们 | 107 | .952 |
| 10 | 她 | 106 | .9431 |
| 11 | 人 | 99 | .8809 |
| 12 | 我 | 94 | .8364 |
| 13 | 说 | 88 | .783 |
| 14 | 在 | 78 | .694 |
| 15 | 这 | 76 | .6762 |
| 16 | 来 | 74 | .6584 |
| 17 | 也 | 74 | .6584 |
| 18 | 个 | 72 | .6406 |



教育部语言文字应用研究所

INSTITUTE OF APPLIED LINGUISTICS, MINISTRY OF EDUCATION

WWW.VSRLC.COM

样例 句法树库的信息检索

树库查询工具

左子节点

语言

词类

/n

父节点（所在结构）

右子节点

词类

/n

查询

导出结果

是否查重

☐ 查找重复统计频率
 ☒ 不查重复完全列出结果

排列顺序

☒ 不排序
 ☐ 节点字母顺序
 ☐ 频率升序排列
 ☐ 频率降序排列

频率筛选

最小频率


0

| 父节点 | 左子节点 | 右子节点 | 示例 |
|------|------|-------|--------------------|
| DZNP | 语言/n | 声学/n | [DZ 语言/n 声学/n]NP |
| DZNP | 语言/n | 中枢/n | [DZ 语言/n 中枢/n]NP |
| LHNP | 语言/n | 文字/n | [LH 语言/n 文字/n]NP |
| DZNP | 语言/n | 工作者/n | [DZ 语言/n 工作者/n]NP |
| DZNP | 语言/n | 生活/n | [DZ 语言/n 生活/n]NP |
| DZNP | 语言/n | 科学/n | [DZ 语言/n 科学/n]NP |
| DZNP | 语言/n | 特点/n | [DZ 语言/n 特点/n]NP |
| DZNP | 语言/n | 成分/n | [DZ 语言/n 成分/n]NP |
| DZNP | 语言/n | 单位/n | [DZ 语言/n 单位/n]NP |
| DZNP | 语言/n | 形式/n | [DZ 语言/n 形式/n]NP |
| DZNP | 语言/n | 功能/n | [DZ 语言/n 功能/n]NP |
| DZNP | 语言/n | 成分/n | [DZ 语言/n 成分/n]NP |
| DZNP | 语言/n | 关系/n | [DZ 语言/n 关系/n]NP |
| DZNP | 语言/n | 材料/n | [DZ 语言/n 材料/n]NP |
| DZNP | 语言/n | 结构/n | [DZ 语言/n 结构/n]NP |
| DZNP | 语言/n | 情境/n | [DZ 语言/n 情境/n]NP |
| DZNP | 语言/n | 差异/n | [DZ 语言/n 差异/n]NP |
| DZNP | 语言/n | 灵活性/n | [DZ 语言/n 灵活性/n]NP |
| DZNP | 语言/n | 史/n | [DZ 语言/n 史/n]NP |
| FWNP | 语言/n | 中/nd | [FW 语言/n 中/nd]NP |
| DZNP | 语言/n | 系统/n | [DZ 语言/n 系统/n]NP |
| DZNP | 语言/n | 哲学/n | [DZ 语言/n 哲学/n]NP |
| DZNP | 语言/n | 现象/n | [DZ 语言/n 现象/n]NP |
| DZNP | 语言/n | 作风/n | [DZ 语言/n 作风/n]NP |
| DZNP | 语言/n | 灵物/n | [DZ 语言/n 灵物/n]NP |

教育部语言文字应用研究所

INSTITUTE OF APPLIED LINGUISTICS MINISTRY OF EDUCATION
WWW.IAL.CN

样例 基于互联网的语料库例句检索

地址  <http://219.238.40.213:8080/QRslt.srf>

  转到

检索结果: 语言 n

共 3587 句 第 1 / 144 页

| | |
|----|---|
| 1 | 试题/n 分/v 语言/n 基础/n 和/c 短文/n 翻译/v 两/m 部分/n , /w 份量/n 约/v 各/r 占/v |
| 2 | 不同/a 层次/n 、 /w 不同/a 行业/n 的/u 人/n 交往/v , /w 学习/v 专业/n 语言/n 、 /w 广泛/a 了解/v 生活/n 、 /w 历史/n 、 /w 文学/n 、 /w 民俗/n 等/u |
| 3 | 图片/n 从/p 语言/n 技巧/n 、 /w 服务/v 态度/n 、 /w 业务/n 知识/n 、 /w 乘客/n 心理/n 等/u |
| 4 | 词语/n 来/vd 表达/v 概念/n , /w 使/v 术语/n 符合/v 本/r 民族/n 语言/n 的/u 语法/n 规则/n 和/c 语言/n 习惯/n 。 /w |
| 5 | , /w 使/v 术语/n 符合/v 本/r 民族/n 语言/n 的/u 语法/n 规则/n 和/c 语言/n 习惯/n 。 /w |
| 6 | , /w 这样/r 搞/v 有/v 没有/v 必要/a ? /w 须知/n 我们/r 是/vl 中国/ns 语言/n 文学/n 系/n 啊/u ! /w |
| 7 | 着重/v 介绍/v 中西/j 文艺/n 理论/n 的/u 异同/n , /w 用/p 简明/a 的/u 语言/n 描绘/v 出/vd 中国/ns 古代/nt 文论/n 的/u 体系/n , /w 用/p |
| 8 | 的/u 任何/r 一/m 篇/q 作品/n , /w 你/r 都/d 能/vu 感到/v 她/r 的/u 语言/n 是/vl 那般/r 淳朴/a 率真/a 、 /w 典雅/a 秀/a 巧/a , /w 在/p 一/m 种/q |
| 9 | 干部/n 考虑到/v 小/h 潘/nhf 母亲/n 有病/v , /w 语言/n 不通/v , /w 转车/v 、 /w 食宿/n 困难/a , /w 就/d 到/v 团里/nl 给/p |
| 10 | 马燕红/nh 准备/v 先/d 在/p 东京/ns 的/u 一/m 所/n 语言/n 学校/n 补习/v 一/m 年/nt 日语/n 和/c 英语/n , /w 然后/c 到/v 东京/ns |
| 11 | 马燕红/nh 准备/v 先/d 在/p 东京/ns 的/u 一/m 所/n 语言/n 学校/n 补习/v 一/m 年/nt 日语/n 和/c 英语/n , /w 然后/c 到/v 东京/ns 女子/n 体育/n 大学/n 攻读/v 四/m 年/nt 的/u 体育/n 专业/n 。 /w |
| 12 | , /w 充分/a 运用/v 动态/n 、 /w 造型/v 、 /w 音乐/n 、 /w 语言/n 等/u 多/a 结构/n 及/c 多层次/n 的/u 视觉/n 形象/n 来/vd 完成/v 的/u |
| 13 | 诗/n , /w 那/r 朴实/a 、 /w 无华/v 的/u 画面/n 就/d 象/p 凝练/a 的/u 语言/n , /w 默默/a 含情/v 地/u 咏叹/v 着/u 北方/nd 的/u 山水/n 和/c 田园/n |
| 14 | 陈莲笙/nh 道/n 长/n 还/d 告诉/v 记者/n , /w 因/c 受/v 各地/nl 语言/n 和/c 民间音乐/n 的/u 影响/v , /w 各地/nl 宫观/n 的/u 斋/n 醮/x 音乐/n |
| 15 | 编辑/n 贯穿/v 全/a 剧/n , /w 每/r 集/n 故事/n 独立/v 成章/v , /w 语言/n 诙谐/a 幽默/a 的/u 系列/n 室内剧/n 。 /w "/w |
| 16 | 科学/n 方法/n 以/p 经验/n 证实/v 或/c 确认/v 为/vl 根基/n , /w 以/p 语言/n 分析/v 为/vl 特色/n , /w 以/p 对/p 科学/n 知识/n 的/u 逻辑/n 重构/v |
| 17 | 就是说/c , /w "/w 存在/v "/w 离不开/v 本源/n 意义/n 上/nd 的/u "/w 语言/n "/w , /w "/w 存在/v "/w 在/p "/w 语言/n "/w 中/nd 。 /w |
| 18 | 本源/n 意义/n 上/nd 的/u "/w 语言/n "/w , /w "/w 存在/v "/w 在/p "/w 语言/n "/w 中/nd 。 /w |
| 19 | 在/p 这/r 种/q 最/d 原始/a 的/u 语言/n 中/nd , /w 不仅/c 保存/v 了/u "/w 事实/n "/w , /w 而且/c 保存/v 了/u |
| 20 | " /w 或许/d "/w 、 /w "/w 将来/nt 或许/d 会/vu "/w 等/u 猜测/v 性/k 语言/n , /w 没有/v 任何/r 实质性/n 证据/n , /w 没有/v 任何/r 确凿/a 的/u |
| 21 | 一/m 年/nt 的/u 努力/v , /w 小仲祺/nh 从/p 一/m 字/n 不识/v , /w 语言/n 表达/v 有/v 障碍/n , /w 变成/v 了/u 会/vu 写/v 几十/m 个/q 字/n , /w |
| 22 | 代表/n 、 /w 胡平平/nh 等/u 33/m 位/q 代表/n 建议/v 对/p 我国/n 语言/n 文字/n 进行/v 立法/v 的/u 议案/n (/w 第/h 13/m 号/n 、 /w 35/m |
| 23 | 长期以来/i , /w 由于/c 缺少/v 相应/a 的/u 法律/n 、 /w 法规/n , /w 语言/n 文字/n 应用/v 不/d 规范/a 、 /w 不/d 标准/a 的/u 现象/n 十分/d 严重/a |
| 24 | 、 /w 不/d 标准/a 的/u 现象/n 十分/d 严重/a , /w 亟/d 需要/v 把/p 语言/n 文字/n 工作/n 纳入/v 法制化/v 轨道/n 。 /w |
| 25 | 才/d 正式/a 启动/v , /w 随即/d 组织/v 了/u 社会/n 历史/n 、 /w 民族/n 语言/n 两/m 个/q 学科/n 的/u 研究/v 人员/n 组成/v 考察/v 队/n 。 /w |
| 26 | 卷/n 》 /w 、 /w 《 /w 门巴/ns 、 /w 珞巴/ns 、 /w |

www.china-language.gov.cn

样例 语料库全文检索

语料库全文检索 (CorpusFSearch)

全文检索

半导体

检索

设置检索选项

检索规则说明

【检索结果摘要】 共计 814 行

生成摘要

切换视图

<25> FTP_FH0070301 《缘分》节录 韩蕊丽 收获文学杂志社
……他笑完以后，就又聚精会神地去拨弄**半导体**了。一个显得很高级很洋气的玩意儿，显然不是国产货。由此我还知道他是华侨。当今，连“业余华侨”都“洋”得不行不行了了，……

<26> FTP_FH50010301 《割掉了鼻子的大象》节录 迟叔昌 中国少年儿童出版社
……有些幻想是实现了，凭我们自己的两只手。举例来说吧，我们各人做了一对只有手表大的**半导体**收音机。冬天把它安在猫皮耳罩上，带着倒是挺舒服，不但能听广播，还管预防耳朵……

<27> FTP_FH50010801 《一次有趣的旅行》节录 鲁克 江苏人民出版社
……给大家看的。”“这是用什么东西做成的？”我用手摸了摸，问他。“用**半导体**材料制成的。”何教授说。“什么叫**半导体**呢？”我又问。何教授……
……”我用手摸了摸，问他。“用**半导体**材料制成的。”何教授说。“什么叫**半导体**呢？”我又问。何教授告诉我：“象铜、铁等能传电的东西，叫做‘导体’……
……说完，我就插嘴说：“我知道啦，把一段传电的铜和一段不传电的橡皮接在一块儿，就成了**半导体**了。”“这不能成为**半导体**。”何教授笑了笑，打断了我的话。“**半导体**是……
……一段传电的铜和一段不传电的橡皮接在一块儿，就成了**半导体**了。”“这不能成为**半导体**。”何教授笑了笑，打断了我的话。“**半导体**是这样的一种东西，它有许多特性，它……
……就成了**半导体**了。”“这不能成为**半导体**。”何教授笑了笑，打断了我的话。“**半导体**是这样的一种东西，它有许多特性，它在低温下不传电，但温度加高的时候，就能传……
……传电，但温度加高的时候，就能传电了。”“举个例子说说，究竟什么样的东西叫**半导体**呢？”“你一定做过矿石收音机吧？”何教授说，“它里面有一颗米粒般大……
……矿石收音机吧？”何教授说，“它里面有一颗米粒般大的矿石（通常用閃鋅矿），就是一种**半导体**。”“你这个收音机也是用这种矿石做的吗？”“不，这个收音机……
……还告诉我：“过去收音机里的真空管是用玻璃做的，一受震动就容易损坏。我这个收音机用**半导体**代替了真空管，它不但大大地缩小了体积，而且不容易摔破震坏，可以用上几十年。……
……大地缩小了体积，而且不容易摔破震坏，可以用上几十年。”“这真是奇妙的事情，**半导体**还可以做成这么小巧的收音机！**半导体**还能做什么呢？我还想问问何教授。这时我偶然俯……
……坏，可以用上几十年。”“这真是奇妙的事情，**半导体**还可以做成这么小巧的收音机！**半导体**还能做什么呢？我还想问问何教授。这时我偶然俯身往下一瞧，忽见地面上一片光亮，……
……着屋顶的瓦问：“这是玻璃瓦吧？”何教授说：“不，这不是玻璃瓦。这种瓦是用**半导体**材料做成的。”我真不懂，为什么用这种**半导体**来做瓦。何教授告诉我：“这种瓦，用处很大，这是一种特制的‘光电池’……
……“不，这不是玻璃瓦。这种瓦是用**半导体**材料做成的。”我真不懂，为什么用这种**半导体**来做瓦。何教授告诉我：“这种瓦，用处很大，这是一种特制的‘光电池’……

<28> FTP_FH50021301 《储—电子技术的粮食》节录 陈念貽 天津科学技术出版社
……究的目标。十几年来，电子学的许多新发展，特别是电子计算机，自动化等技术，都需要用**半导体**材料制成的元件，用储还能制造袖珍式的功率强大的收音机……。煤中除了碳……
……储，就能得到四氧化锗，精制后水解制成氧化锗，再用氢还原便得到金属锗。用作**半导体**的锗，必须十分纯净，含锗要在八个九，即百分之九十九点九九九以上，将锗……
……达到十个九的锗。这种纯锗，进一步制成单晶，就可制造出电子计算机，雷达和**半导体**收音机上所需要的元件了。锗不仅可由烟灰提取，而且可由铅锌矿……

<29> FTP_FH50023201 《到宇宙太空中去开发资源》节录 石土 天津科学技术出版社
……等特殊条件，制造出许多地面上所不能造出来的产品，例如高纯度的光通信纤维，高质量的**半导体**单晶，高激光效率的玻璃，特殊的合金等等。这些设想是有根据的，部分已经在环绕……

<30> FTP_FH50029501 《不睡觉的女婿》节录 肖建亨 四川少年儿童出版社
……那报纸是匆匆忙忙盖上去的。那台仪器，老教师前几天就注意到了：那里面全是密密麻麻的**半导体**管子，那五颜六色的电线也是一捆一捆的。“好啊，我看你怎么回答我。”……

<31> FTP_FH50030601 《腐蚀》节录 叶永烈 江苏人民出版社
……子。从外表看，这是一顶高帽子。可是真不简单，就在这顶高帽里，藏着2 0万个**半导体**电子管。一个电子管只有一颗火柴头那么大。那一天，没头脑看爸爸忙得不……
……机里的真空管，你看见过没有？”“看见过，可不是这个样子。”“这是**半导体**做的呀！又小巧，又省电，又坚固，又经用，又……你知道吗，有一种金属叫做锗？……
……呀！“不是大家都叫他没头脑吗？所以我想了个办法，给他安一个电脑。看吧，一个小小的**半导体**电子管，就抵得上一个脑细胞。”“可是脑细胞还要小得多呀！”……

<32> FTP_JC10010431 《中国历史》（初中全3册）人民教育出版社历史室 人民教育出版社
……在广大科研人员的努力下，1962年，我国的新兴科学技术：原子能技术、喷气技术、**半导体**、电子计算机技术等都从无到有的建立起来了。本世纪6 0—7 0年代，世……

<33> FTP_JL20000403 《固体物理学》（上册）绪论 第1,2,4章节录 方俊鑫 陆栋 上海科学技术出版社
……2），配位数为8。除口外，T 1 B r、T 1 I 等皆属此类。离子结合成分较大的**半导体**材料Z n S等，是由两种各为面心立方结构的离子沿空间对角线位移 $\frac{1}{4}$ 长度套构……

<34> FTP_JL20000404 《固体物理学》（上册）绪论 第1,2,4章节录 方俊鑫 陆栋 上海科学技术出版社
……地改善晶体的某种性能，常常有控制地在晶体中引进某类外来原子，形成替位式杂质。这在**半导体**的制备过程中是习以为常的。我们知道，锗、硅单晶体是四价的原子**半导体**，在纯粹……
……位式杂质。这在**半导体**的制备过程中是习以为常的。我们知道，锗、硅单晶体是四价的原子**半导体**，在纯粹的情况下，它们的半导体性质并不很灵敏。如果在高纯的锗、硅单晶体中有……
……为受主杂质。施主杂质和受主杂质在锗、硅中所构成的导电类型是不同的，这些在本书下册**半导体**一章中再行讨论。实验证明，I I I B族元素（硼、铝、镓、铟等）和V B……

<35> FTP_JL20000401 《固体物理学》（上册）绪论 第1,2,4章节录 方俊鑫 陆栋 上海科学技术出版社
……理，指出了导体与绝缘体的区别，并断定有一类固体，它们的导电性质介于两者之间，叫做**半导体**。四十年代末、五十年代初，以锗、硅为代表的**半导体**单晶的出现并以此制……
……电性质介于两者之间，叫做**半导体**。四十年代末、五十年代初，以锗、硅为代表的**半导体**单晶的出现并以此制成了晶体三极管，进而产生了**半导体**物理，这标志着固体物理学……
……、五十年代初，以锗、硅为代表的**半导体**单晶的出现并以此制成了晶体三极管，进而产生了**半导体**物理，这标志着固体物理学发展过程的又一飞跃。在**半导体**物理的带动下，固体物理……
……制成了晶体三极管，进而产生了**半导体**物理，这标志着固体物理学发展过程的又一飞跃。在**半导体**物理的带动下，固体物理获得了大发展。**半导体**器件以及其他固体器件的发展，特别……
……标志着固体物理学发展过程的又一飞跃。在**半导体**物理的带动下，固体物理获得了大发展。**半导体**器件以及其他固体器件的发展，特别是尔后集成电路的发展使无线电电子技术、计算……
……地提出新的要求，促使人们利用固体内部电子运动的复杂规律，制造出许多新的元件，例如**半导体**元件、铁氧体元件、磁膜、磁泡、铁电体元件、超导体元件等。集成电路就是从这里……
……发展，使得液氮的供应为广泛地在低温下探索固体内部的复杂规律提供了有利条件，从而使**半导体**、超导体、磁性材料、顺磁共振、核磁共振等研究工作更加深入。电子谱仪和超高真……
……是如此，例如，由此出现了固体非线性光学的领域，又例如，激光和低温相结合，用以研究**半导体**中的激子，出现了电子—空穴液的新型量子液。固体中的激光散……

<36> FTP_JL20000603 《热学》绪论 第3,6,7,9章节录 李椿 章立源 钱尚武 高等教育出版社
……中的一部分内能转化为电磁能，另一部分在冷水处放热，冷水就是一个低温热源。利用两种**半导体**形成回路所制成的小型发电机，就是利用了这种温差电原理。经验告诉我们……

<37> FTP_JM20004212 《化学》（高中全一册）人民教育出版社化学室 人民教育出版社
……的增大而逐渐升高，它们的密度随着核电荷数的增大而逐渐加大。此外，硫不能导电，**硒**是**半导体**，而**碲**却能够导电。氧、硫、**硒**、**碲**的单质的化学性质也随着核电荷数的增……

<38> FTP_JM20004220 《化学》（高中全一册）人民教育出版社化学室 人民教育出版社
……元素的化合物进行研究，有助于制造新品种的农药。又如，在金属与非金属的分界线附近寻找**半导体**材料，在过渡元素中寻找催化剂以及耐高温、耐腐蚀的合金材料等。元素……

<39> FTP_JA20001903 《中国体制改革哲学探索》节录 庄福龄 中国人民大学出版社

语料库应用软件

- 语料切分和标注软件
- 树库标注软件
- 语料库校对加工软件
- 语料检索工具软件
- 语料库查询与统计工具软件
- 基于互联网的语料库例句检索
- 语料库全文检索软件



语料库的管理

- 国家语委语料库由国家语委委托语言文字应用研究所负责建设和管理
- 国家语委语料库可以提供的服务
 - 语料库使用权许可
 - 科研等非赢利性目的
 - 产品开发等赢利性目的
 - 检索、查询、统计等数据服务
 - 软件开发等其他服务



语料库提供服务的方式

- 通过签订使用权许可协议，提供语料库的有偿使用
 - 科研用途
 - 商业用途
- 为从事汉语教学与研究的科研人员、教师、学生提供无偿的检索统计等服务
 - 语料检索、例句提取、字词频率统计等
 - 检索的数量、范围有一定限制
- 为教育部、国家语委的相关科研项目提供数据支持
 - 语料库、标注语料库、相关软件等
- 已有国内外众多高校、公司和科研院所使用国家语委语料库进行科学研究和产品开发



相关资料

□ 国家语委语料库相关材料可以从网上下载

中国语言文字网

www.china-language.gov.cn



谢 谢 ！

肖航 exiaohang@163.com



教育部语言文字应用研究所

INSTITUTE OF APPLIED LINGUISTICS MINISTRY OF EDUCATION

WWW.IALC.EDU