国家重点基础研究计划973课题“文本内容理解的数据基础”

自然语言处理
与
自然语言理解

俞士汶
北京大学计算语言学研究所(ICL/PKU)
Email: yusw@pku.edu.cn
2009年3月11日，中信所(ISTIC)

主要内容

- 关于研究对象与目标
- 自然语言处理的主攻方向
- 综合型语言知识库概要
- 前进目标——自然语言理解
- 领域知识工程与领域知识库
- 结语与致谢

主要内容

- 关于研究对象与目标
- 自然语言处理的主攻方向
- 综合型语言知识库概要
- 前进目标——自然语言理解
- 领域知识工程与领域知识库
- 结语与致谢

学科定位

《学科分类与代码表》（中国国家标准 GB/T13745--92 ）

语言学

——应用语言学

——计算语言学（740.3550）

（三级学科“计算语言学”属于文科 ）

俞士汶 主编《计算语言学概论》，商务印书馆，2003年

计算机科学技术

——人工智能

——自然语言处理（520.2020）

（三级学科“自然语言处理”属于理工科）

两个三级学科的研究对象与内容基本相同，

新兴交叉学科在学科体系中已有一席之地。

（从北大ICL/PKU、北大软微学院语言信息工程系 到 教育部计算语言学重点实验室）

不同的术语与所指

计算语言学 （Computational Linguistics，CL）
自然语言处理 （Natural Language Processing，NLP）
　自然语言理解（Natural Language Understanding，NLU）
人类语言技术（Human Language Technology，HLT）
语言信息处理（Language Processing Technology，LPT）
（大致相同，又各有侧重。）

中文信息处理（Chinese Information Processing，CIP）
——汉字信息处理(Chinese Characters Processing)
——汉语信息处理
（Chinese Information Processing，CIP）

我们聚焦于汉语信息处理，实质是以汉语为核心的多语言信息处理。

“自然语言处理”在研究什么？

（1）机器翻译与机器辅助翻译（最早）
（2）信息检索（索引技术）与信息提取
（3）文本与知识管理（术语提取、分类、摘要、述评）
（4）人工系统的自然语言界面
（5）词典计算机辅助编纂
（6）面向语言本体研究与教学研究的应用

……

NLP是IT的任务子集，作为计算机处理的对象，发生了变化:

表现形式（字符串） —》词、句子、篇章
字符信息（数据集） —》语言信息（知识）

需要对相关的理论、方法与技术以及 “语言及其认知机制”有个概括的了解。

研究得怎么样？

看看机器翻译的水平，以Google Translate Beta为

- 2009年1月3日完成的翻译实例

（1）北京大学俞士汶教授应邀将于2009年3月到中国科学技术信息研究所进行学术交流。

（2）你得藏在一个你看得见他，可是他看不见你的地方。

（3）车臣武装分子和世界其他地区的恐怖分子是一丘之貉，应该合力打击他们。

（4）新一届测绘学名词审定委员会的主要特点是年青化，吸收了一些工作在教学、科研前沿的青年专家学者，充分发挥他们接触新知识多，对名词工作热情高、活力大的特长，同中老年专家共同做好新一届委员会的名词审定工作。

- 2009年1月13日完成的翻译实例

（5）胡六点横看成岭侧成峰，见仁见智。

（摘自《参考消息》2009年1月13日第10版台报社论）

人贵有自知之明，然而机器却什么都敢干。

难怪有人说规则翻译是傻子，统计翻译是疯子。

关于 “语言”

英国《新科学家》周刊 2005年4月9日 的文章

——生命进化的十大奇迹：脑（第3项）和语言（第4项）

脑常常被视作进化过程中的最高成就，因为它赋予了人类一些高级特征，例如 语言、智慧、意识。

语言是进化的终极发明。在令人类区别于动物的特征中，语言处于核心地位。语言也许称得上是人类的决定性特征之一。我们的祖先如何实现了语言从无到有的飞跃，这也许是科学史上最大的谜。语言是生物进化的最后一笔。这是因为语言令那些掌握了它的动物超越了纯生物的范畴。

语言系统是动物进化到人的两大标志之一。

关于“自然语言处理”

自然语言处理是数值计算机
在非数值领域最早的应用（MT，Turing试验）。
语言学对计算机科学也有重要贡献（Chomsky）。
自然语言理解又特别困难:
（1）依据对人类语言机制的认识
（2）语言既是对象，又是工具
（3）依据对当代计算机能力的认识
（4）依据NLP技术发展的历史经验
汉语理解研究和其他语言一样困难，
汉语信息处理技术又有特殊的课题。
以汉语为母语的学者还有其独特优势。

关于“计算语言学”

为自然语言处理提供理论模型、实现算法、工程方案。

语言模型：

实际问题太复杂，需要根据应用的需要，进行简化、变换，使其成为可计算的形式，这就是模型化。

上下文无关语法就是一种语言模型，便于分析和生成符合规则的句子，可以覆盖相当一部分自然语言句型。

向量空间模型可以刻画文本的主要特征，可用于信息检索、文本分类。

常用算法：

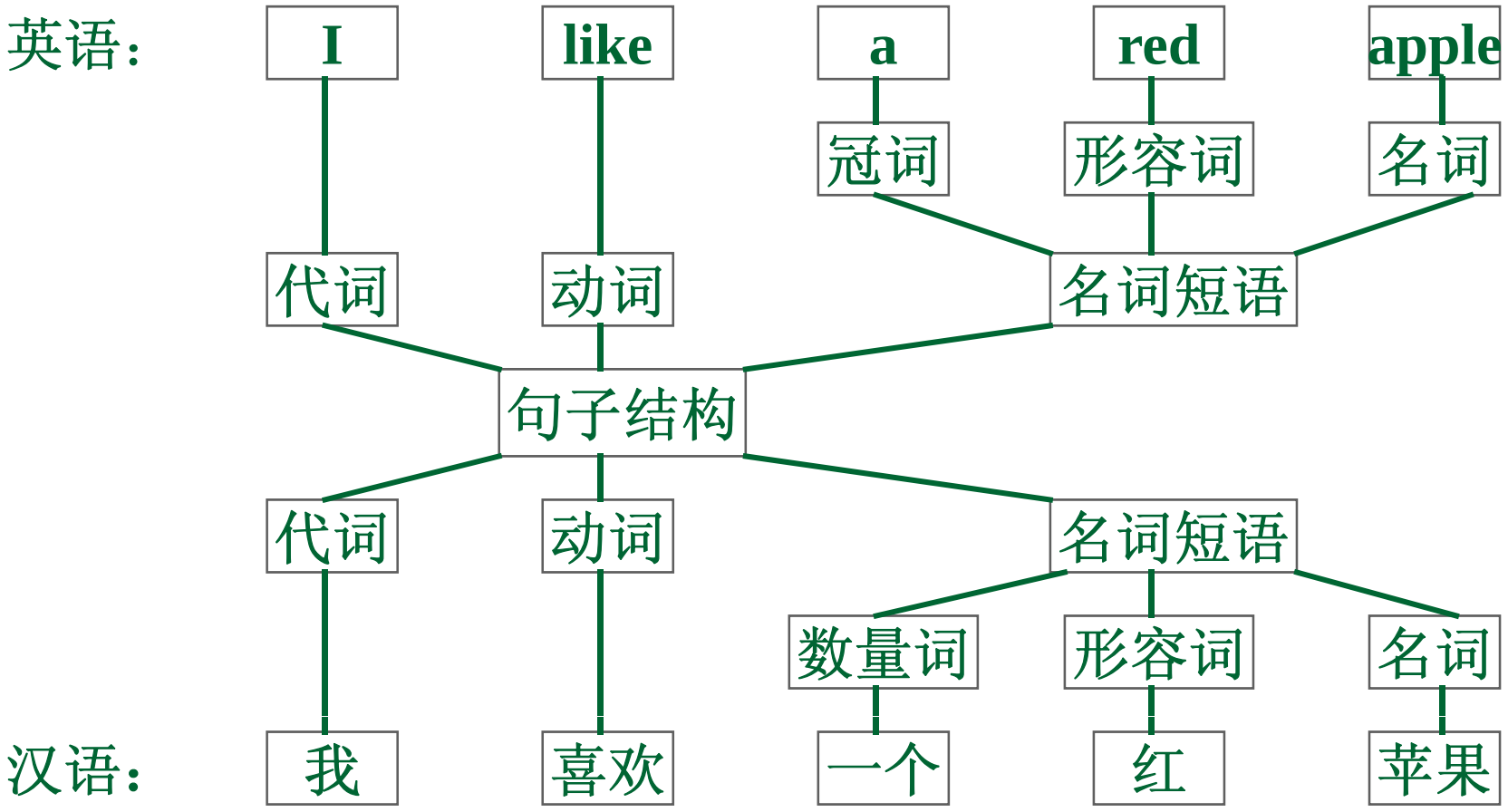
基于规则的方法（词法、句法、语义、语用）

基于统计的方法（原始语料－加工了的语料）

不同的应用需要以不同的语言单位作为研究对象，不同的应用也要采用不同的处理方法。不同类型的方法的结合可能提供最好的效果。实际系统常常兼收并蓄，博采众长。

NLP的典型任务——机器翻译

基于规则（上下文无关语法）的方法



基于统计的方法

对于给定的原文句子**F***，在译文语料库**{E}**中，搜索使得概率**P(E|F*)**达到最大值的句子**E***。

在**{E}** 中，求**E*=argmaxP(E|F*)**

=argmaxP(E)P(F*|E)

P(E)是目标语言的语言模型，任何一个句子**E**的出现概率。

P(F*|E)是两种语言的翻译模型。

问题归结于如何计算 **P(E)** 和 **P(F*|E)**。

$$P(E)=P(W_1W_2 \cdots W_n)$$

$$=P(W_1)P(W_2|W_1)\cdots P(W_n|W_1W_2\cdots W_{n-1})$$

进一步简化， $P(E)= P(W_1)P(W_2|W_1)\cdots P(W_n|W_{n-1})$

这就是二元语法模型，这些概率值可以在语料库中统计得到。

主要内容

- 关于研究对象与目标
- 自然语言处理的主攻方向
- 综合型语言知识库概要
- 前进目标——自然语言理解
- 领域知识工程与领域知识库
- 结语与致谢

自然语言理解的困难

实例之一

关于自动升降晾衣架的对话

妻子：“嘿，过了一年才坏。”

丈夫：“什么呀，才一年就坏了。”

丈夫理解了妻子的意思吗？

——虚词词义：才（数量词前后，意义不同）

——背景知识：保修期

——知识激活机制？

自然语言理解的困难

实例之二

关于“沙漠化”的文章

“几年前由于种植籽瓜有利可图，使大批的种植者就到过渡带来开垦，……。在这样的绿洲和沙漠过渡带开垦，极易造成风蚀。”

——<今日民航>2001年9月号

就/ 到/ 就到/ 到/ 到过/ 过/ 过渡/ 带/ 来/ 带来/

（未登录词的识别——知识背景——认知机制）

自然语言处理的主攻方向

自然语言理解研究特别困难。

退而求其次:

自然语言处理。

计算机处理自然语言的第一个障碍是

自然语言固有的歧义问题。

人能够利用多模态的知识和 语境信息。

消解歧义的能力远远超出计算机。

自然语言处理主攻方向——歧义消解

词语切分歧义：白天鹅

可能的切分：白天鹅/---白/ 天鹅/---白天/ 鹅/---白/ 天/ 鹅/

计算机程序可以按某种算法实现这种切分，给出一种或多种结果。对否？

白天鹅飞过来了——白/ 天鹅/ 飞/ 过来/ 了

白天鹅可以看家——白天/ 鹅/ 可以/ 看/ 家/

白天鹅在湖里游泳——白/ 天鹅/ ？白天/ 鹅/ ？

词性标注歧义：只——量词 q [zhi1] ？副词 d [zhi3]?

这只会测水温的鸭子

——这/ 只/ 会/ 测/ 水温/ 的/ 鸭子/ （切分无歧义）

——这/r 只/q 会/v 测/v 水温/n 的/u 鸭子/n ， 挺有用的

——这/r 只/d 会/v 测/v 水温/n 的/u 鸭子/n ， 没什么用

（意义决定词性，还是词性决定意义？）

主攻方向——歧义消解

读音相同的“连”也有不同的词性（意义）：

一个连有三个排——“连”是名词 n

我们兄弟心连心——“连”是动词 v

苹果可以连皮吃——“连”是介词 p

短语结构的歧义： **m + q + n + “的” + n**

三个大学的老师　　三/**m** 个/**q** 大学/**n** 的/**u** 老师/**n**

——**[[** 三/**m** 个/**q** 大学/**n]** 的/**u** 老师/**n]**

——**[** 三/**m** 个/**q [** 大学/**n** 的/**u** 老师/**n]]**

三所大学的老师——**[[** 三/**m** 所/**q** 大学/**n]** 的/**u** 老师/**n]**

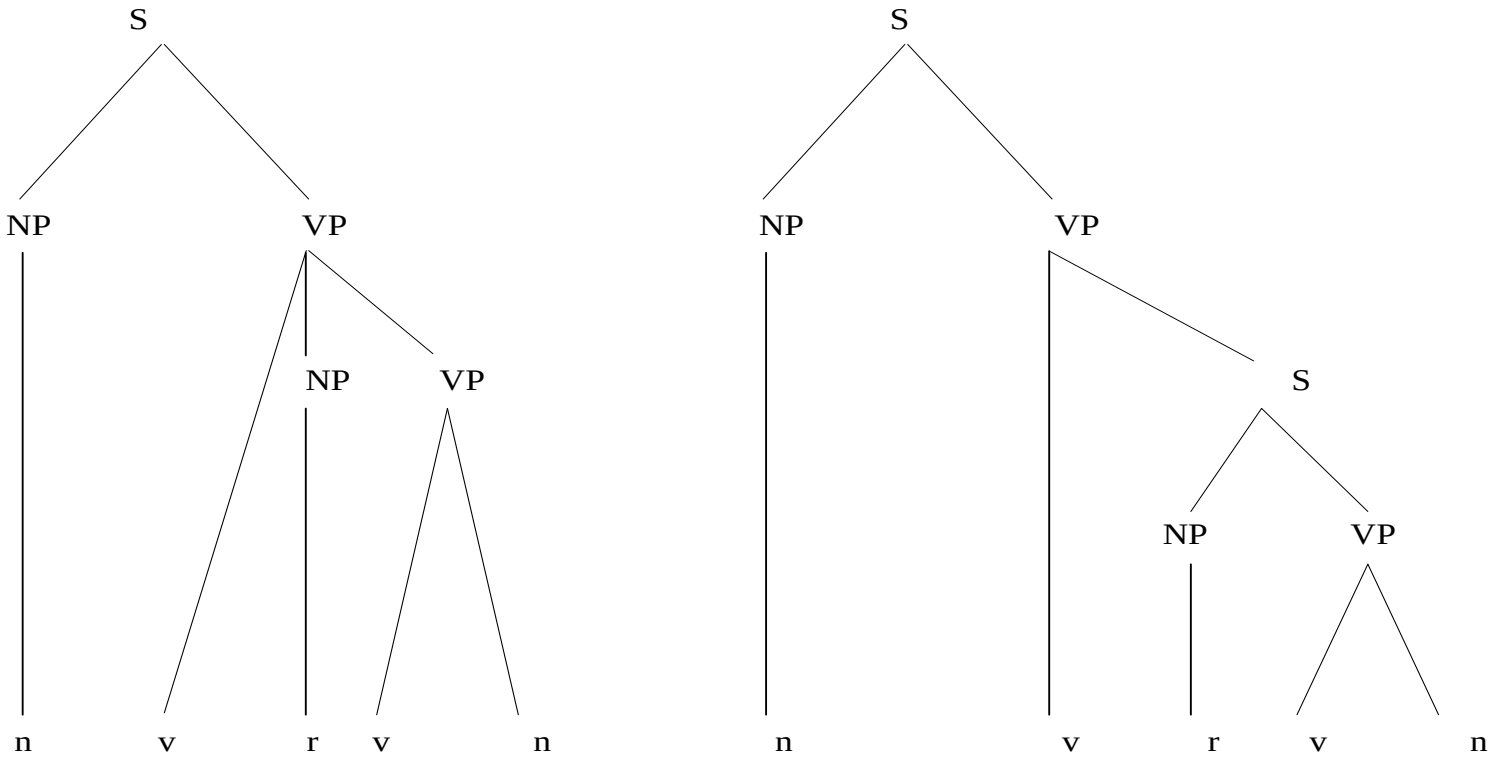
三位大学的老师——**[** 三/**m** 位/**q [** 大学/**n** 的/**u** 老师/**n]]**

句法结构的歧义

例1　　会员　选举　他　当　主席

例2　　学生　认为　他　是　校长

n　v　r　v　n　（切分、标注无歧义）



句法结构（树）不同

语义歧义以及依赖语境的歧义消解

汉语语义分析（切分、标注、句法分析都无歧义）

熊猫/n　吃/v　竹笋/n
学生/n　吃/v　食堂/n
民工/n　吃/v　大碗/n
老师/n　写/v　毛笔/n

汉语语义指向分析

写/v 好/a 了/u （文章）
写/v 累/a 了/u （老师）
写/v 秃/a 了/u （毛笔）

汉语语境分析

小张/n　打针/v　去/v　了/u （护士？病人？）

其他：长句与句号、逗号

新一届测绘学名词审定委员会的主要特点是年青化，吸收了一些工作在教学、科研前沿的青年专家学者，充分发挥他们接触新知识多，对名词工作热情高、活力大的特长，同中老年专家共同做好新一届委员会的名词审定工作。 （长句子，一逗到底）

其他：长句与句号、逗号 （Cont'd）

1. 你得藏在一个你看得见他，可是他看不见你的地方。（逗号断开了结构）

2. 车臣武装分子和世界其他地区的恐怖分子是一丘之貉，应该合力打击他们。（分句的主语省略，“他们”又指谁？）

其他：指代与省略 （Cont'd）

小明要求他爸爸给他弟弟买一件
他喜欢的衣服，他同意了。
（4个“他”，各指谁？）

重庆队得88分，客场负于台湾队2分。
（CBA, 台湾队和重庆队各得多少分？比赛地点？）

其他：时态、语态、语气

我在家里。(be)

我在家里看书。(in)

我在看书。 (-ing)

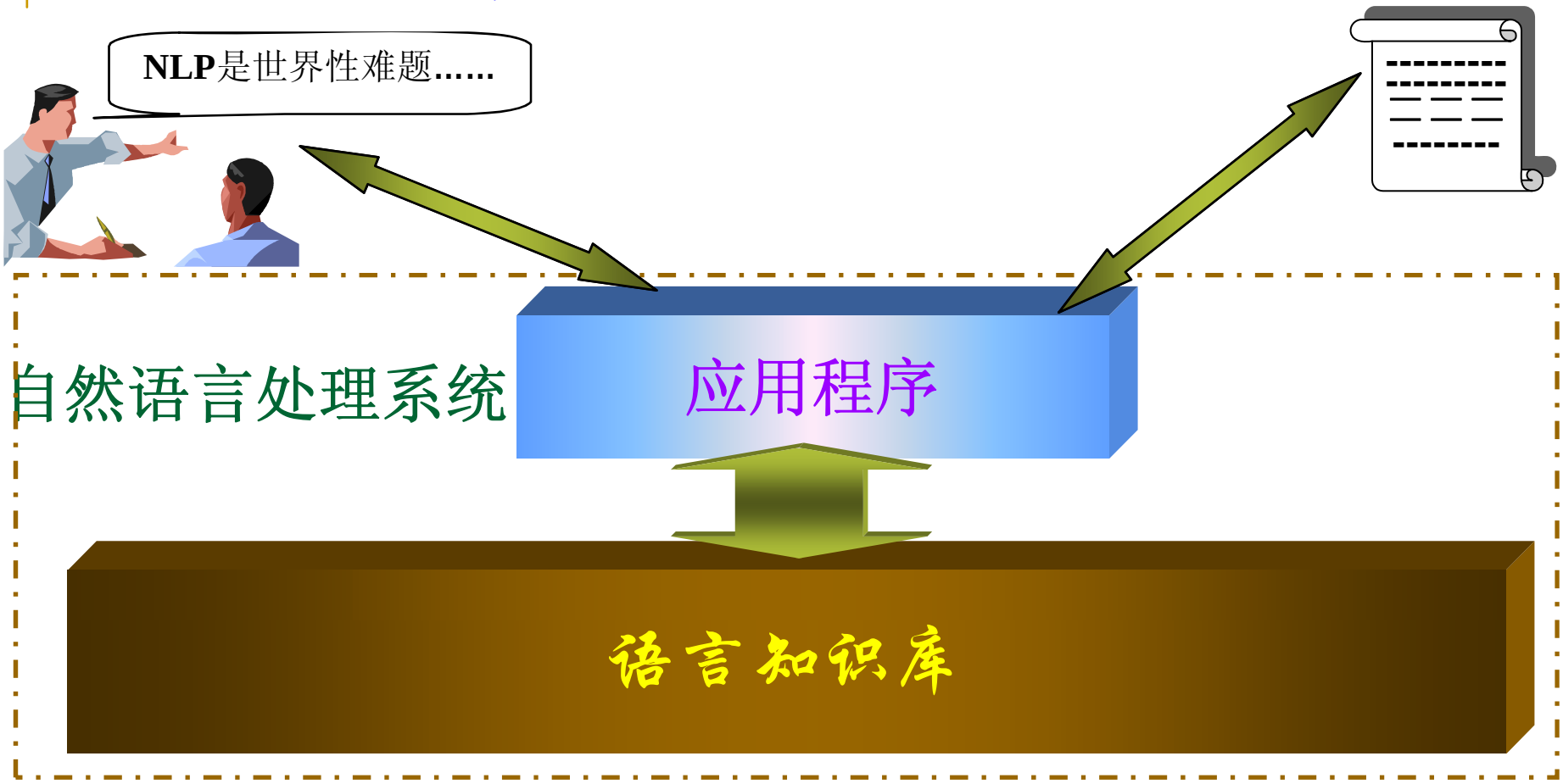
你在干什么？——看书。

你喜欢干什么？——看书。

如果我是你，我就去了。

如果我有时间，我就去。

关于一般的自然语言处理系统



语言知识库是自然语言处理系统不可或缺的组成部分，语言知识库的规模和质量在很大程度上决定了自然语言处理系统的成败。面向自然语言处理的语言知识库对语言本体研究和语言教学也有重要意义。

书面汉语特点及其对信息处理的影响

吕叔湘：“有了形态变化，语法分析就比较容易进行。没有严格的形态变化，在语法分析上就比较容易引起问题。”

汉语缺乏形态变化，缺乏形式标记，自动分析也就缺少可以把握的线索。汉语自动分析如果不比其他的语言更困难，至少不会比其他的语言更容易。

汉语信息处理尤其需要大规模的高质量的语言知识库的支持。

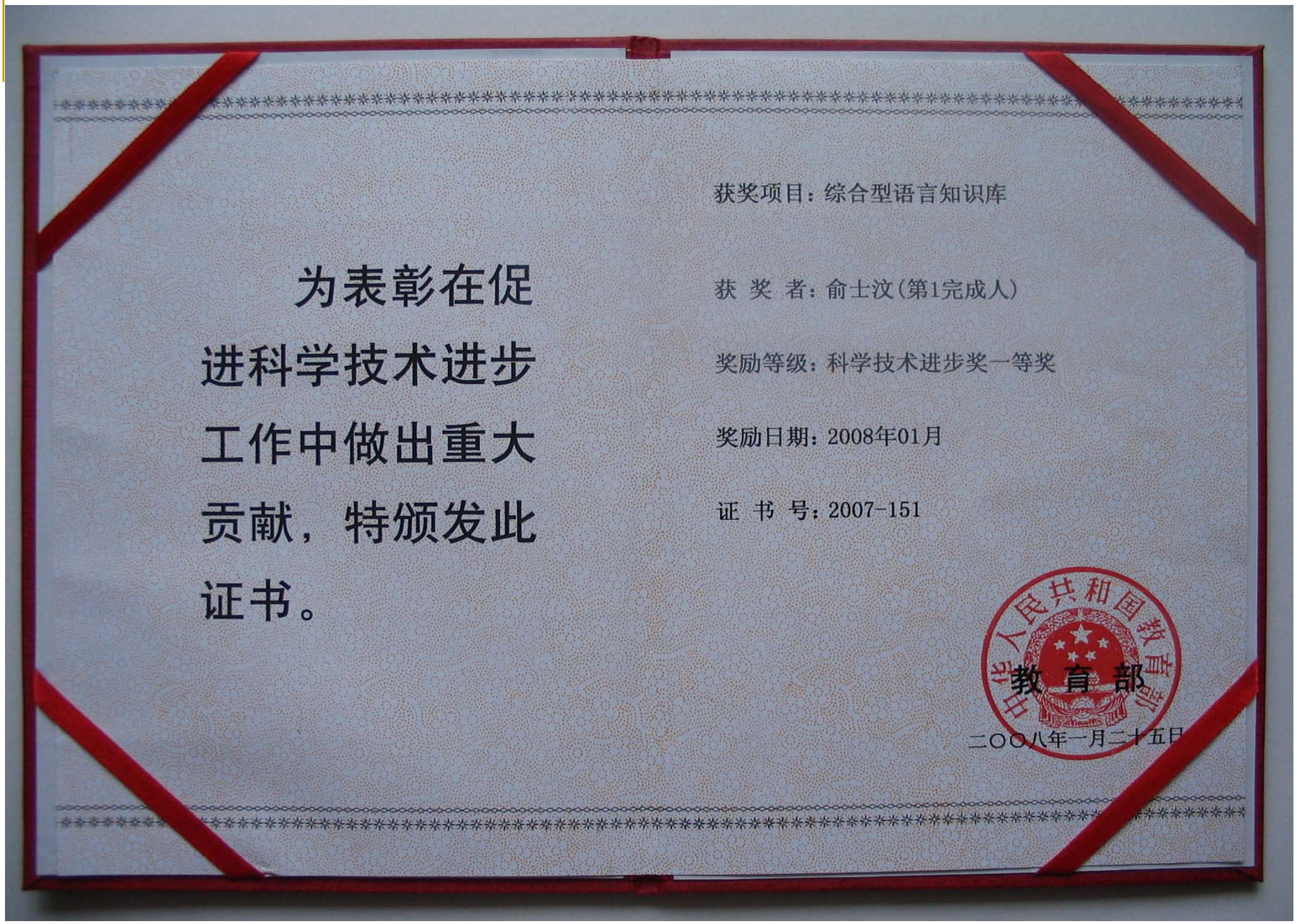
主要内容

- 关于研究对象与目标
- 自然语言处理的主攻方向
- 综合型语言知识库概要
- 前进目标——自然语言理解
- 领域知识工程与领域知识库
- 结语与致谢

北大开发的现代汉语语言知识库

（1）现代汉语语法信息词典（8万词语）
（2）面向汉英机器翻译的现代汉语语义词典（6万）
（3）面向跨语言文本处理的中英文概念词典 （10万概念）
（4）大规模现代汉语基本标注语料库（6000多万汉字）
（5）句子对齐的双语语料库 （英汉80万句对）
（6）多个专业领域的术语库 （35万英汉对照术语）
（7）现代汉语短语结构规则库（600条规则）
（8）用于语言知识库开发的各种工具软件

规模大、种类多、质量上乘，已产生广泛影响，仍期待发展与合作.


获奖项目：综合型语言知识库
为表彰在促
进科学技术进步
工作中做出重大
贡献，特颁发此
证书。
获 奖 者：俞士汶(第1完成人)
奖励等级：科学技术进步奖一等奖
奖励日期：2008年01月
证 书 号：2007-151
中华人民共和国教育部
二〇〇八年一月二十五日

《现代汉语语法信息词典》简介

《现代汉语语法信息词典》是一部面向语言信息处理的大型电子词典。它按照语法功能和意义相结合的准则收录了7.3万余词语。依照语法功能分布的原则，建立了词类体系，完成了这7.3万词语的归类。并在此基础上，分类描述每个词语的各种语法属性。

引自中国工程院编《20世纪我国重大工程技术成就》

之第二项汉字信息处理与印刷革命（暨南大学出版社2002年第一版31页）

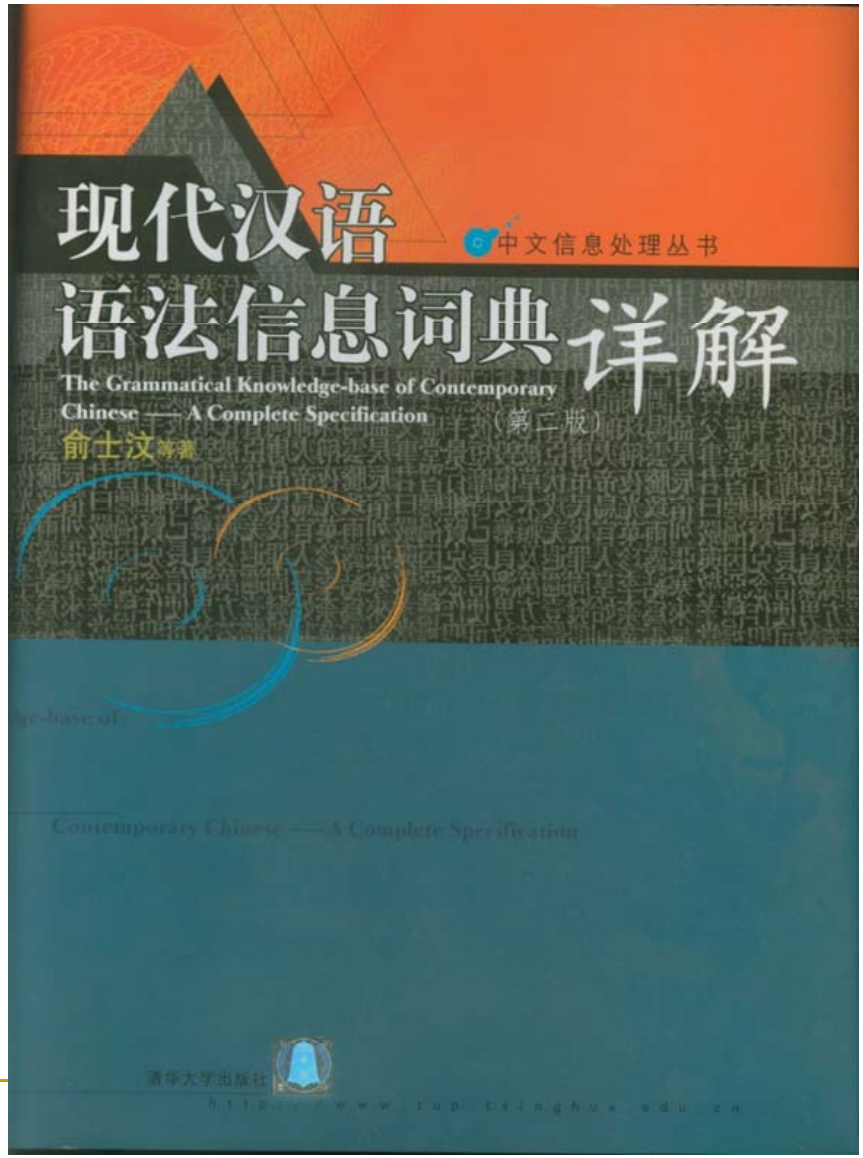
词典采用数据库文件格式。目前已扩充到8万词语。

清华大学出版社出版了介绍这部电子词典的专著

《现代汉语语法信息词典详解》第一版1998年，第二版2003年

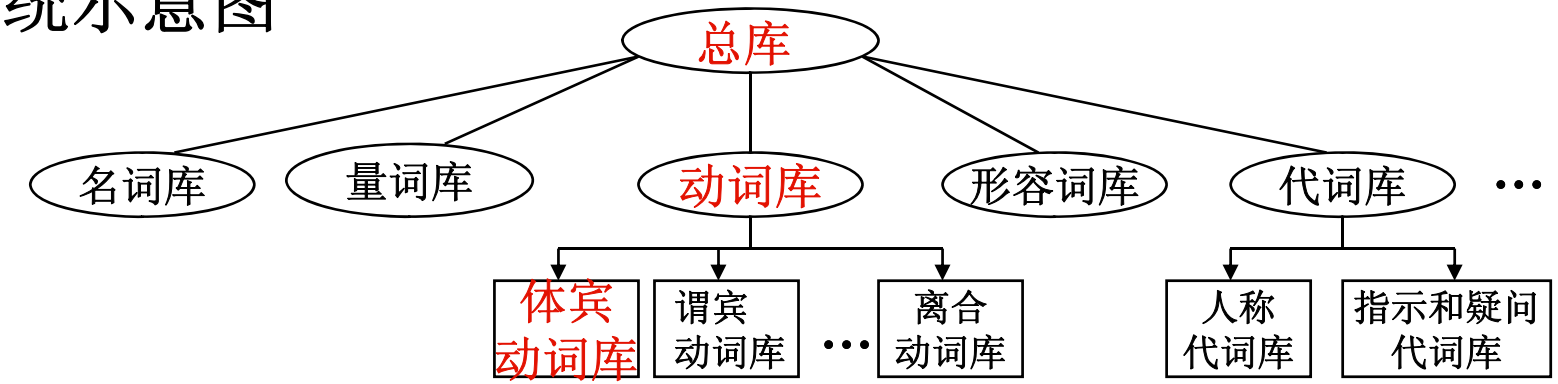
1998年曾获教育部科技进步二等奖

现代汉语语法信息词典详解



现代汉语语法信息词典

词类系统示意图



总共34个库文件，通过
“词语+词类+同形”连接，
构成上下位继承关系的树

《现代汉语语法信息词典》总库之样例

“词语＋ 词类＋ 同形” 是主关键项，希望增加 “频次”和“例句”

词语	词类	同形	拼音	注	频次	例句
挨	v	A	ai1	触，碰，靠近	119?	?
挨	v	B	ai2	遭受，忍受	?	?
安装	v		an1zhuang1		599	
保管	v	1	bao3guan3	保存	58?	?
保管	v	2	bao3guan3	担保	?	?
抄	v	A	chao1	照原稿写	104?	?
抄	v	B	chao1	走近道	?	?
地道	a		di4dao5	正宗	46	
地道	n		di4dao4		3	
叫	v	A1	jiao4	人或动物发出的较大声音	1591?	?
叫	v	A2	jiao4	呼唤，招呼；雇	?	?
叫	v	A3	jiao4	称为	?	?
叫	v	B	jiao4	使，让，命令	?	?

动词库样例（数据库文件主关键词还是“词语+v+同形”）

词语	同形	义项	系词	助动	趋向	体谓准	双宾	单作补	复数主	后名	很	着了过	重叠	离合	兼类
保存						体				可		着了过	ABAB		
成为			系			体									
得到						体准						了过			
告诉						体谓	双					了过			
协商						体谓			复	可		了			
加以						准									
冒险										可		过	VVO	离	a
去	A1	除掉				体						了过	VV		
去	A2	~上海			趋	体		可				了过	VV		
去	B	扮演				体						了过			
应	A	答应						可				了			
应	B	应该		助		谓									
支持	1	支撑				体						着了过			
支持	2	鼓励并帮助				体谓准					很	着了过	ABAB		
指挥						体谓				可		着了过	ABAB		n

体宾动词分库样例（主关键词仍是是“词语+v+同形”）

词语	同形	受事	格标1	与事	格标2	…	施事	备　　注
得到		受						得到可靠的数据
告诉		受	把	与				把好消息告诉他
去	A1	受	把					把苹果皮去了
去	A2	受						去封信/去香港
去	B	受						去白娘子
支持	1	受	把					把顶棚支持住
支持	2			与	对			对模范要支持
坐		受	把				施	前排坐嘉宾

现代汉语语义词典 CSD

机器翻译要求（更精细的）词义消歧

例1　　她的仪表很精密。

例2　　她的仪表很端庄。

例1和例2的句法结构完全一样，对“仪表”的词义消歧无贡献，只能根据与其搭配的形容词的“主体”的语义选择特性。

“精密（**precise**）”的　“主体”是“器具　（**instrument**）”，

“端庄（**decorous**）”的　“主体”是“品貌　（**appearance**）”。

现代汉语语义词典（含6万实词）

—《现代汉语语法信息词典》的扩充，面向机器翻译。

动词库部分信息样例（“义项码”是细化的词义信息）

词语	词类	同形	义项码	语义类	释义	英译	配价	主体	客体	与事
冲	v	A	1	创造	冲茶	make （tea）	2	人	固饮	
冲	v	A	2	促变	冲胶卷	develop (a film)	2	人	材料	
冲	v	A	3	促变	冲盘子	Rinse (the plate)	2	人	器皿	
冲	v	B		位移	冲锋	charge	1	动物		

中英文概念词典 CCD

汉外翻译既提出了词义消歧的需求，也是检验词义消歧的手段，不过这个手段并不是充分的。

“病毒” —— “virus”

（1）“生命体”（生物学领域）

（2）“恶意代码”（信息技术领域）

如果进行（跨语言）信息检索或信息提取，区分这两个概念是必要的。

中文概念词典（**CCD: Chinese Concept Dictionary**）从另一个视角组织词汇语义知识。面向（跨语言）信息 提取/检索 和 文本处理。

中英文概念词典 CCD

CCD 是一个双语WordNet

汉英双语概念对应

直接复用WordNet的理论、方法、技术，全球WordNet资源建设的组成部分。

概念：由同义词集(Synset)来表示，概念即同义词集

{教师、教员、老师、先生、导师、老板、师傅、孩子王、臭老九、…}

概念网络：概念之间多种语义关系

反义（Antonymy）、上下位（Hypernymy）、整体-部分（Holonymy）、
致使（Cause）、蕴涵 （Entailment）等

GKB，CSD主要反映词与词的组合关系，

CCD 则主要反映词义间的聚合关系。

中英文概念词典 CCD

面向（跨语言）信息 提取/检索

Offset	Synset	Csynset	Hypernym	Hyponym	Definition	Cdefinition
07632177	teacher instructor	教师 教员 老师 先生 导师 老板 孩子王 臭老九 …	07235322	07086332 07162304 07209465 07243767 07279659 07297622 07341176 07401098 07414251 07425180 07494025 07520938 07533674 07551404 07551581 07561151 07632624 07632736	a person whose occupation is teaching	以教学为职业的人

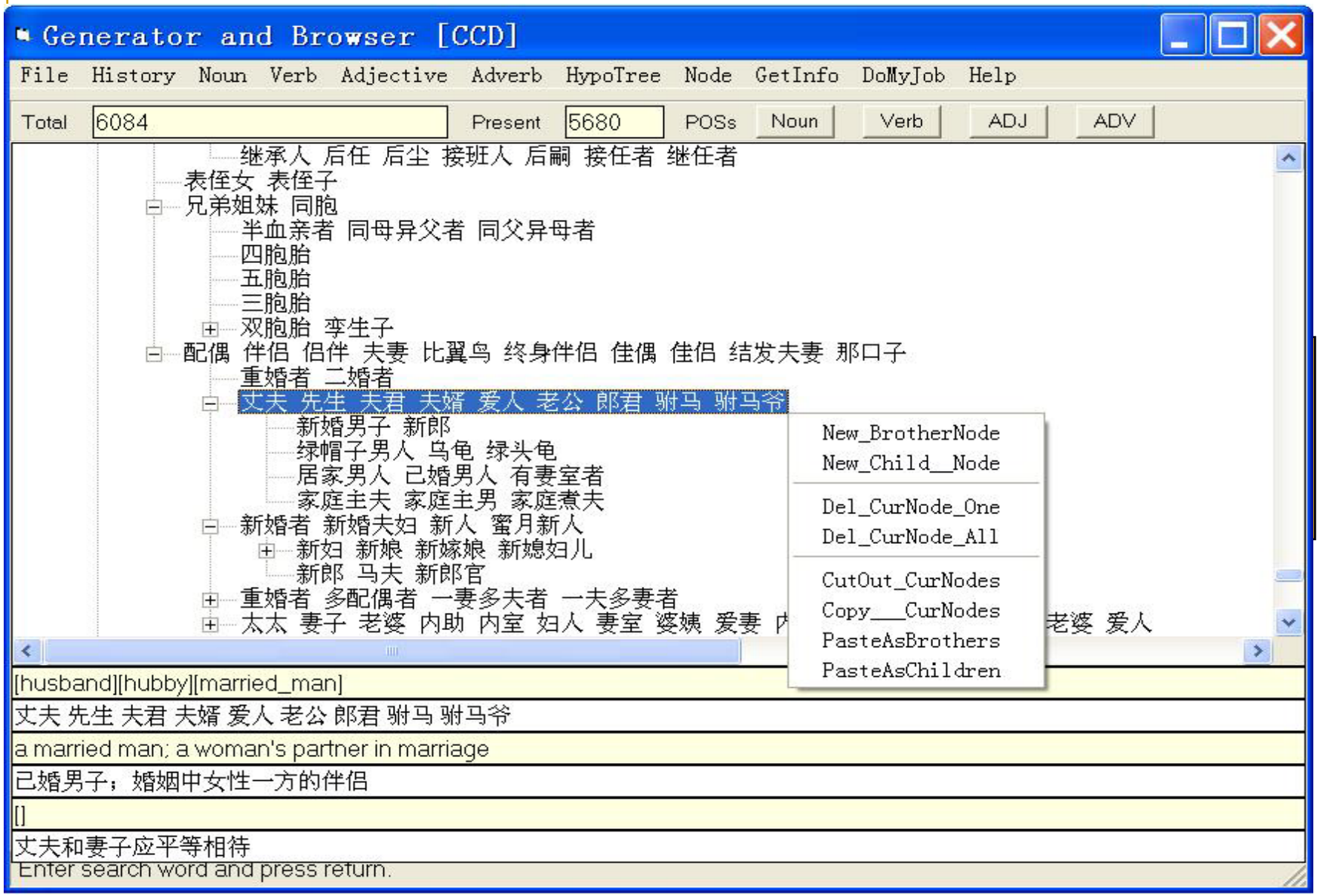
中文概念词典 CCD

面向（跨语言）信息 提取/检索

Offset	Synset	Csynset	Hypernym	Hyponym	Definition	Cdefinition
07331418	husband hubby married_man	丈夫 先生 夫君 夫婿 爱人 老公 郎君 驸马 驸马爷	07602853	07109482 07195968 07255726 07328008	a married man; a woman's partner in marriage	已婚男子; 婚姻中女性一方的伴侣

Offset	Synset	Csynset	Hypernym	Hyponym	Definition	Cdefinition
07414666	Mister Mr.	先生 师傅 同志 大哥 老兄 老弟	07391044		a form of address for a man	对男子的一种称呼

中英文概念词典可视化表示（树之节点——同义词集合）



中英文概念词典 CCD

CCD 不仅仅是双语 WordNet

它反映汉语的特点，面向中文信息处理的需求。

（1）对概念、概念关系有调整和发展

汉语有“叔父，伯父，姑父，姨夫，舅父”，英语中没有分别对应的概念，CCD 的解决办法是让这些概念对应英语中的“uncle”。

汉语中有“笔”这个概念，英语中没有，只有“pen, pencil, …”

设立“虚概念节点”（writing tool）

（2）增添汉语特有的特征属性

褒贬义、汉语反义词的音节限定特征（暗-亮，黑暗-明亮）

（3）增添词义分析必要的组合关系

搭配信息（锻炼身体，锻炼意志，*锻炼道德）

大规模现代汉语基本标注语料库

原始语料

例1：此类编著内容抄自别人的多，多到被人告到了法庭。

例2：炮兵学院原来围墙残缺，周边群众进城，习惯抄近道。

加工后的语料

例1：此类/r 编著/vn 内容/n 抄/v 自/p 别人/r 的/u 多/a ，/w 多/a 到/v 被/p 人/n 告/v 到/v 了/u 法庭/n 。/w

例2：炮兵/n 学院/n 原来/d 围墙/n 残缺/v ，/w 周边/n 群众/n 进城/v ，/w 习惯/v 抄/v 近道/n 。/w

词典中的语言知识（静态、显性、不确定）

与语料库中的语言知识（动态、隐性、确定）

实现语料库基本标注使词汇知识、词性知识显性化

知识显性化的目的之一便于实现机器学习（Learning from Data）

北大语料加工中的规范

重要性——大型语言工程不可或缺
科学性——词组本位语法体系
实践性——指导实践，接受检验，加以修订
适用性——标记集的慎重选择
（两套标记集，先后发表，接受广泛的检验）
稳定性——一定时期内相对稳定

《北京大学现代汉语语料库基本加工规范》*中文信息学报*，2002. No.5，pp.49-64；No.6，pp.58-65
2007年获第四届中国科协期刊优秀学术论文奖

《北大语料库加工规范：切分·词性标注·注音》新加坡：*汉语与语言计算学报*，*2003.* No.2，pp.121-158

台湾中研院语言学研究所黄居仁研究员2006年8月在第三届学生计算语言学研讨会（沈阳）作“语言学理论与分析在计算语言学中的应用”之特邀报告：“因此，北大的整套语料库标记系统，就是一个语言学理论。有了这个认识，任何自然语言处理，当然必须建立在好的语言学理论上。”

获奖证书

俞士汶、段慧明、朱学锋、孙斌同志：

你们撰写的论文“北京大学现代汉语语料库基本加工规范”，被评为第四届中国科协期刊优秀学术论文，特发此证，以资表彰。

中国科学技术协会

2007年1月

现代汉语基本标注语料库的继续发展

- 有/v 人/n 嫌/v 脏/a ，/w 提出/v 用/v 水/n 冲/v 一/m 冲/v 。 /w
- 待/p 我/r 再/d 去/v 冲/v 胶卷/n 时/Ng ， /w 他/r 见/v 了/u 面/n 就/d 像/p 老/a 熟人/n 一样/u ，/w 闲谈/v 中/f 问/v 了/u 我/r 的/u 职业/n 。/w
- 丁/nr 玉珍/nr 把/p 冲/v 好/a 的/u 照片/n 交给/v 了/u 孔/nr 玲/nr 。/w
- 一/m 只/q 大/a 鸟/n 直/d 冲/v 云霄/n
- １９９５年/t 洪水/n 冲/v 倒/v 了/u 他/r 家/n 在/p 村子/n 里/f 的/u ３/m 间/q 土屋/n ， /w 也/d 没有/v 能力/n 翻盖/v 。/w
- 经/p 风暴/n 一/d 冲/v ，/w 经济/n 结构/n 的/u 深层/b 毛病/n 加速/v 暴露/v ，/w 提早/d 进入/v 了/u 调整期/n 。/w

粗粒度词义标注（同形）的实例——基于GKB

- 有/v 人/n 嫌/v 脏/a ，/w 提出/v 用/v 水/n 冲/v!A 一/m 冲/v!A 。 /w
- 待/p 我/r 再/d 去/v 冲/v!A 胶卷/n 时/Ng ， /w 他/r 见/v 了/u 面/n 就/d 像/p 老/a 熟人/n 一样/u ，/w 闲谈/v 中/f 问/v 了/u 我/r 的/u 职业/n 。/w
- 丁/nr 玉珍/nr 把/p 冲/v!A 好/a 的/u 照片/n 交给/v 了/u 孔/nr 玲/nr 。/w

- 一/m 只/q 大/a 鸟/n 直/d 冲/v!B 云霄/n
- １９９５年/t 洪水/n 冲/v!B 倒/v 了/u 他/r 家/n 在/p 村子/n 里/f 的/u ３/m 间/q 土屋/n ， /w 也/d 没有/v 能力/n 翻盖/v 。/w （自然力也可以是“冲/v”的施事）
- 经/p 风暴/n 一/d 冲/v!B ，/w 经济/n 结构/n 的/u 深层/b 毛病/n 加速/v 暴露/v ，/w 提早/d 进入/v 了/u 调整期/n 。/w （隐喻，用自然力“风暴”比喻“金融危机”）

细粒度词义标注（义项）的实例——基于CSD

- 有/v 人/n 嫌/v 脏/a ，/w 提出/v 用/v 水/n 冲/v!A－3 一/m 冲/v!A－3 。 /w
- 待/p 我/r 再/d 去/v 冲/v!A－2 胶卷/n 时/Ng ， /w 他/r 见/v 了/u 面/n 就/d 像/p 老/a 熟人/n 一样/u ，/w 闲谈/v 中/f 问/v 了/u 我/r 的/u 职业/n 。/w
- 丁/nr 玉珍/nr 把/p 冲/v!A－2 好/a 的/u 照片/n 交给/v 了/u 孔/nr 玲/nr 。/w

- 一/m 只/q 大/a 鸟/n 直/d 冲/v!B 云霄/n
- １９９５年/t 洪水/n 冲/v!B 倒/v 了/u 他/r 家/n 在/p 村子/n 里/f 的/u ３/m 间/q 土屋/n ， /w 也/d 没有/v 能力/n 翻盖/v 。/w （自然力也可以是“冲/v”的施事）
- 经/p 风暴/n 一/d 冲/v!B ，/w 经济/n 结构/n 的/u 深层/b 毛病/n 加速/v 暴露/v ，/w 提早/d 进入/v 了/u 调整期/n 。/w （隐喻，用自然力“风暴”比喻“金融危机”）

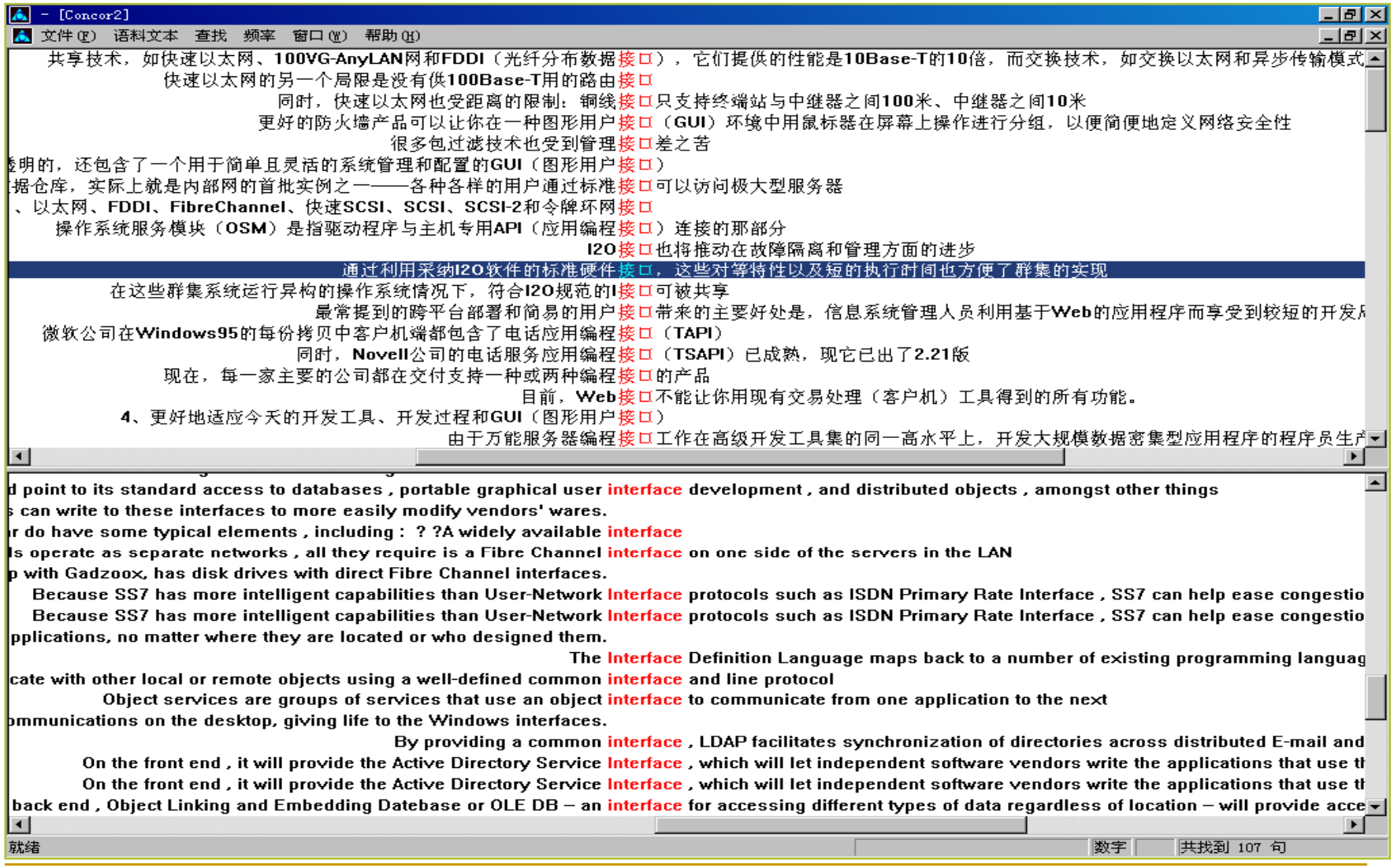
双语平行语料库 BAC

- 由篇章到句子级对齐，英汉80万句对，日汉约3万句对。用途广泛。
- 样例：XML 标记文件。也有纯文本文件。
- 系统的流程，深入的加工——相关句列（Concordance）检索

```
<?xml version="1.0" encoding="gb2312" ?>
- <TEXT>
  - <TEXT_HEAD>
      <AUTHOR>欧内斯特·海明威</AUTHOR>
      <CH_TITLE>丧钟为谁而鸣</CH_TITLE>
    </TEXT_HEAD>
  - <TEXT_BODY>
    - <p id="1">
      - <a id="1" no="1">
        - <s id="1">
            <CH_TITLE>丧钟为谁而鸣</CH_TITLE>
          </s>
        </a>
      </p>
    - <p id="2">
      - <a id="2" no="1">
          <s id="1">罗伯特·乔丹是一位在西班牙内战中帮助共和国游击队作战的美国人，他被派往法西斯后方去炸毁一座桥梁。</s>
        </a>
      - <a id="3" no="1">
          <s id="2">一位牢靠的老向导安塞尔莫将他带到当地游击队的营地。</s>
        </a>
      - <a id="4" no="1">
          <s id="3">游击队队长帕布洛，曾经是个出色的战士，现在却已开始丧失勇气。</s>
        </a>
      - <a id="5" no="2">
          <s id="4">乔丹见到的游击队其他成员有友善厚道但缺乏能力的吉普赛人拉斐尔和帕布洛的妻子皮拉尔。</s>
          <s id="5">皮拉尔这个女人相貌虽丑，却比她丈夫勇敢得多。</s>
        </a>
      - <a id="6" no="1">
          <s id="6">尤其给乔丹留下深刻印象的是玛丽亚，她是个害羞的年轻姑娘，头发剪得短短的。</s>
        </a>
      - <a id="7" no="1">
          <s id="7">乔丹获悉，她最近刚被游击队从法西斯分子手中救出，曾惨遭法西斯分子蹂躏，身体和精神仍处于恢复之中。</s>
        </a>
      </p>
    - <p id="3">
      - <a id="8" no="1">
```

```
<?xml version="1.0" encoding="gb2312" ?>
- <TEXT>
  - <TEXT_HEAD>
      <MODE>书面语</MODE>
      <FIELD>政治</FIELD>
      <STYLE>文学</STYLE>
      <PERIOD>Present-day English</PERIOD>
      <AUTHOR>Ernest Hemingway</AUTHOR>
      <EN_TITLE>For Whom the Bell Tolls</EN_TITLE>
    </TEXT_HEAD>
  - <TEXT_BODY>
    - <p id="1">
      - <a id="1" no="1">
        - <s id="1">
            <EN_TITLE>For Whom the Bell Tolls</EN_TITLE>
          </s>
        </a>
      </p>
    - <p id="2">
      - <a id="2" no="1">
          <s id="1">Robert Jordan, an American fighting for the Republicans in the Spanish Civil War, is sent behind Fascist lines to destroy a bridge.</s>
        </a>
      - <a id="3" no="1">
          <s id="2">Anselmo, an old and trustworthy guide, takes him to a local guerilla camp.</s>
        </a>
      - <a id="4" no="1">
          <s id="3">Its leader is Pablo, a distinguished soldier who has begun to lose his nerve.</s>
        </a>
      - <a id="5" no="1">
          <s id="4">The other members of the group whom Jordan meets are the gipsy Rafael, amiable but feckless, and Pablo's wife Pilar, an ugly woman who is far braver than her husband.</s>
        </a>
      - <a id="6" no="1">
          <s id="5">Jordan is particularly struck by Maria, a shy young girl with a cropped head.</s>
        </a>
      - <a id="7" no="1">
          <s id="6">He learns that she has recently been rescued from the Fascists and is still recovering from the ill
```

平行相关句列 Parallel Concordance



多个专业领域的术语库

- 信息科学技术领域术语库

  中英文对照，条目约15万对

- 体育、商务、餐饮、旅游领域术语库

领域		汉英版 （术语对）	英汉版（术语对）
体育	术语	37,832	36,640
	缩略语	1,305	1,232
	专名	3,302	3,304
商务		107,962	118,498
餐饮		17,969	22,555
旅游		25,501	28,711

体育术语 1

	A	B	C	D	E	F	G
1	ProperName-en	ProperName-cn	Event-en	Event-cn	Country-en	Country-ch	note
2	Allan Budi Kusuma	魏仁芳	Badminton	羽毛球	Indonesia	印度尼西亚	
3	Arbi	阿尔比	Badminton	羽毛球	Indonesia	印度尼西亚	
4	Budi Santoso	布迪·桑托索	Badminton	羽毛球	Indonesia	印度尼西亚	
5	Camilla Martin	卡米拉·马丁	Badminton	羽毛球	Denmark	丹麦	
6	Candra Wijaya	陈甲亮	Badminton	羽毛球	Indonesia	印度尼西亚	
7	Cheah Soon Kit	谢顺吉	Badminton	羽毛球	Malaysia	马来西亚	
8	Chris Hunt	克里斯·亨特	Badminton	羽毛球	Britain	英国	
9	Chung Jae-hee	郑在喜	Badminton	羽毛球	South Korea	韩国	
10	Gade Christensen	盖得·克里斯滕森	Badminton	羽毛球	Denmark	丹麦	
11	Hidayat	西达亚特	Badminton	羽毛球	Indonesia	印度尼西亚	
12	Hoyer Larsen	霍耶·拉尔森	Badminton	羽毛球	Denmark	丹麦	
13	Jens Eriksen	简斯·埃里克森	Badminton	羽毛球	Denmark	丹麦	
14	Jesper Larsen	杰斯珀·拉尔森	Badminton	羽毛球	Denmark	丹麦	
15	Lee Dong-soo	李东秀	Badminton	羽毛球	South Korea	韩国	
16	Mainaky	迈纳基	Badminton	羽毛球	Indonesia	印度尼西亚	
17	Peter Rasmussen	彼特·拉斯姆森	Badminton	羽毛球	Denmark	丹麦	
18	Ra Kyung-min	罗景民	Badminton	羽毛球	South Korea	韩国	
19	Simon Archer	西蒙·阿切尔	Badminton	羽毛球	Britain	英国	
20	Subagia	苏巴吉亚	Badminton	羽毛球	Indonesia	印度尼西亚	
21	Susi Susanti	王莲香	Badminton	羽毛球	Indonesia	印度尼西亚	
22	Tony Gunawan	吴俊明	Badminton	羽毛球	Indonesia	印度尼西亚	
23	Yong Hock-kin	杨景福	Badminton	羽毛球	Malaysia	马来西亚	
24	Yoo Young-sun	柳镛成	Badminton	羽毛球	South Korea	韩国	
25	Abdul-Jabbar	阿布杜·贾巴尔	Basketball	篮球	USA	美国	NBA
26	A.C.Green	A.C.格林	Basketball	篮球	USA	美国	NBA
27	Allan Huston	阿兰·休斯顿	Basketball	篮球	USA	美国	NBA
28	Alonzo Mourning	阿兰左·莫宁	Basketball	篮球	USA	美国	NBA
29	Anthony Mason	安东尼·梅森	Basketball	篮球	USA	美国	NBA
30	Antonio Mcdyess	安东尼奥·麦克戴斯	Basketball	篮球	USA	美国	NBA
31	Archibald	阿奇巴尔德	Basketball	篮球	USA	美国	NBA
32	Bill Russell	比尔·拉塞尔	Basketball	篮球	USA	美国	NBA
33	Brian Grant	布里安·格兰特	Basketball	篮球	USA	美国	NBA
34	Buck Williams	别克·威廉姆斯	Basketball	篮球	USA	美国	NBA
35	Causwell	考斯威尔	Basketball	篮球	USA	美国	NBA
36	Charles Barkley	查尔斯·巴克利	Basketball	篮球	USA	美国	NBA
37	Chris Mullin	克里斯·穆林	Basketball	篮球	USA	美国	NBA
38	Chris Webber	克里斯·韦伯	Basketball	篮球	USA	美国	NBA

体育术语2

1	term-en	term-cn	cat1-en	cat1-cn	cat2-en	cat2-cn	cat3-en	cat3-cn
2	Appeal Committee	仲裁委员会	physical culture	体育	Olympic games	奥林匹克运动会		
3	Executive Board	执行委员会	physical culture	体育	Olympic games	奥林匹克运动会		
4	International Olympic Committee	国际奥林匹克委员会	physical culture	体育	Olympic games	奥林匹克运动会		
5	IOC Session	国际奥委会会议	physical culture	体育	Olympic games	奥林匹克运动会		
6	Olympic champion	奥林匹克冠军	physical culture	体育	Olympic games	奥林匹克运动会		
7	Olympic city	奥运会城	physical culture	体育	Olympic games	奥林匹克运动会		
8	baseball field	棒球场	physical culture	体育	stadiums and gyr	体育场、馆		
9	basketball court	篮球场	physical culture	体育	stadiums and gyr	体育场、馆		
10	bleachers	露天看台	physical culture	体育	stadiums and gyr	体育场、馆		
11	box	专席	physical culture	体育	stadiums and gyr	体育场、馆		
12	bulletin board	公告牌	physical culture	体育	stadiums and gyr	体育场、馆		
13	centre pole	中心旗杆	physical culture	体育	stadiums and gyr	体育场、馆		
14	competition arena; ground; court	比赛场地	physical culture	体育	stadiums and gyr	体育场、馆		
15	covered gymnasium	室内运动场	physical culture	体育	stadiums and gyr	体育场、馆		
16	electrically controlled movable st	电动式活动看台	physical culture	体育	stadiums and gyr	体育场、馆		
17	air (/aviation/flying) sports	航空运动	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
18	ancient sports	古典运动	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
19	athletics	田径运动	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
20	autumn sports	秋季运动	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
21	ball games	球类运动	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
22	bob sleighing	滑橇	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
23	dragon boat race	龙舟竞渡	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
24	modern pentathlon	现代五项运动	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
25	modern sports	现代运动	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
26	yak race	牦牛赛	physical culture	体育	physical culture a	体育运动	kinds of sport	体育运动种类
27	back number	运动衣背后的号码	physical culture	体育	physical culture a	体育运动	sports wear	运动服装
28	badge	运动衣上的队标	physical culture	体育	physical culture a	体育运动	sports wear	运动服装
29	blazer	艳色运动上衣	physical culture	体育	physical culture a	体育运动	sports wear	运动服装
30	flannel trousers	法兰绒运动裤	physical culture	体育	physical culture a	体育运动	sports wear	运动服装
31	girl's slacks	女运动裤	physical culture	体育	physical culture a	体育运动	sports wear	运动服装
32	gym outfit	成套运动服	physical culture	体育	physical culture a	体育运动	sports wear	运动服装
33	jockstrap	松紧运动内裤	physical culture	体育	physical culture a	体育运动	sports wear	运动服装
34	moccasin	软底运动鞋	physical culture	体育	physical culture a	体育运动	sports wear	运动服装

（7）现代汉语短语结构规则库

（1）汉语中短语（词组）的地位

（2）短语分类体系，重点是功能分类，与词类体系一致。

（3）短语结构描述：在面向计算机时，笼统地谈“动宾结构”是不够的，需要更明确地指出哪个子类的或具有什么属性的动词和哪个子类的或具有什么属性的名词能构成什么样的短语，这个短语的特性如何，它继承了构成成分的哪些属性，丢失了哪些属性，又派生了哪些新的属性 。

（7）现代汉语短语结构规则库

（4）短语结构数据库（675条规则）

名称代码	表达式	自粘	功用	粘组	线层	结构	中心	例
zzaap	a(状=“可”)+a	自	谓	粘	线	状中	后a	绝对可靠
zwaap	a+a	自			线	主谓		谦虚好
aaccp	a	自	谓		线			多
aaccp	a+<aaccp>	自	谓	粘	层	词串		多快好省
dzccp	aa+n	自	体	粘	线	定中	n	新衣服

（5）数据库记录到产生式规则（扩充的上下文无关语法）的转换

zzaap::= a(状=“可”)+a;自粘=‘自’；功用=‘谓’；粘组=‘粘’；线层=‘线’；
结构=‘状中’；中心=‘后a’ /*绝对可靠*/

（6）与 GBK 适配，可扩展到与 CSD 适配。

（8）用于语言知识库开发的各种工具软件

- 《现代汉语语法信息词典》管理软件
- 汉语词语切分及词性标注软件
- 汉语词语注音软件
- 可视化中文概念词典辅助开发软件
- 基本标注语料库查询软件
- 双语语料库辅助开发工具集
- 科技术语辅助自动提取软件
- （粗/细 粒度）词义自动消歧与辅助校对软件
- 文本型语料库—结构化语料库转换软件
- …… ……

语言数据资源建设的基本经验

（1）规模与质量：规模足够大，质量是生命线

（2）基础与应用：面向应用，遵循基础研究规律

（3）工程与学术：阶段性成果与长期求同辨异

（4）专家与工具：发挥各自优势

（5）语言知识：以词汇为本，以句法知识为基础，向语义深入

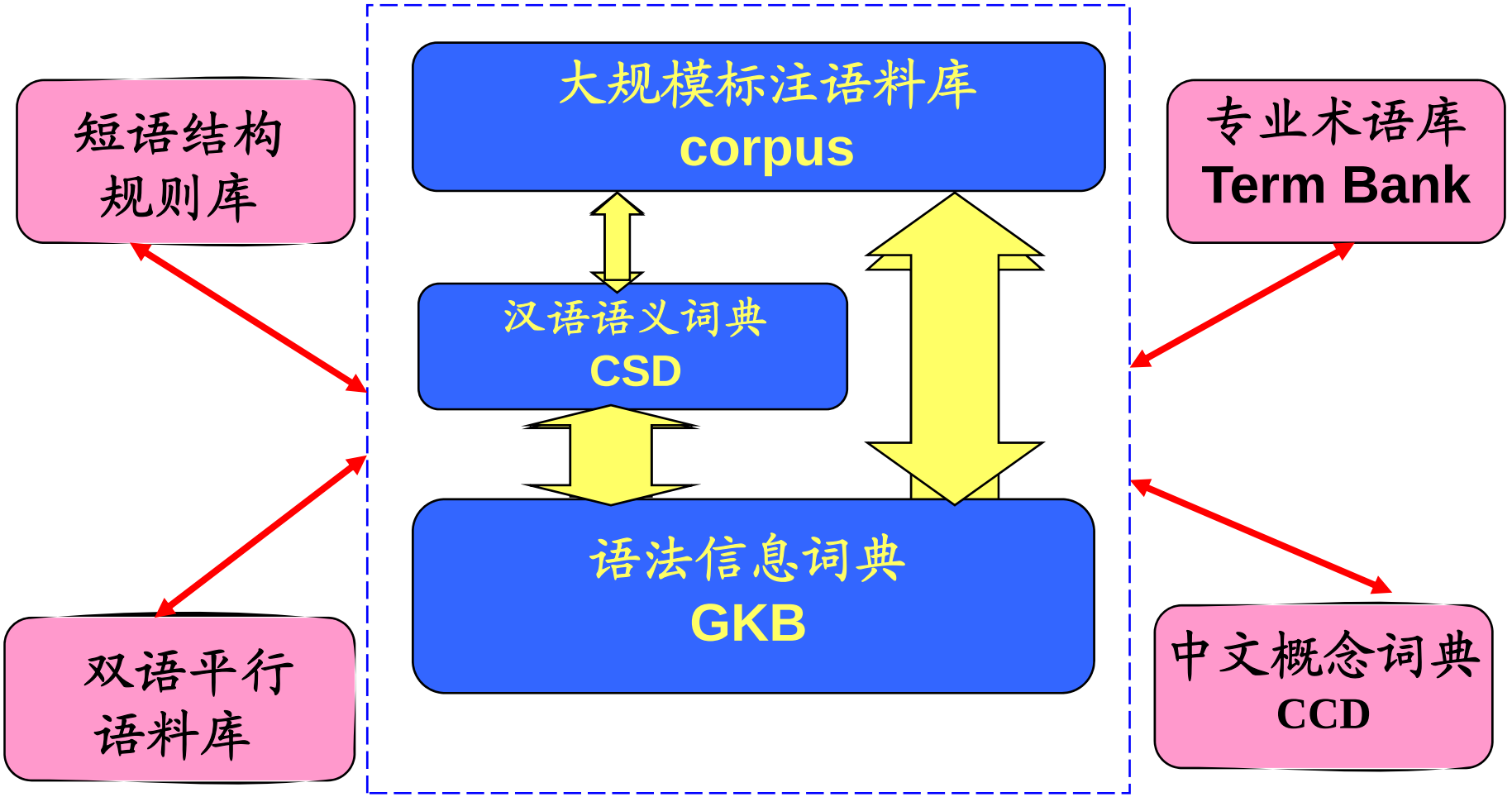
（6）知识表达：借鉴理论成果，语言知识及其表述形式独立于信息处理系统和实现算法，结构化与非结构化权宜采用。

（7）理论指导：基于规则的方法和基于统计的方法并举；反过来，又促进了这两种方法的发展。

（8）人才培养：人才与科研成果同步增长

（9）相关问题：汉语与多语，常识与专业，现代与古代，应用检验，知识产权，------

各种语言数据资源之相互支撑



综合型语言知识库的建设

现状：

尽管各语言数据资源之间有内在的逻辑上的紧密联系，但在物理上彼此却是孤立存在的，它们之间还不能实现“准确的”和“便捷的”交叉存取。

目标：

（1）支持各成分数据资源之间便捷的准确的交叉参照，方便用户（包括人和机器）从结构各不相同的多种语言资源获取丰富的语言知识。

（2）提供统一的应用程序接口（API）和风格一致的友好的用户界面（UI）。

（3）提供数据挖掘工具，发展机器学习机制，支持知识发现，充分展现综合型语言知识库的价值和作用。

（4）提供知识传播和信息服务的机制，既做到知识共享，让知识库发挥最大的效益，同时对知识产权又有妥善处理。

语言数据资源的集成方案

在各种语言数据资源中，
《现代汉语语法信息词典》和“标注语料库”最具基础性。
首先，集成**GKB**和“标注语料库”，作为“综合型语言知识库”的主体部分。然后，再把其他类型的语言资源集成进来。
从系统功能的角度考虑，首先实现各成分数据资源之间便捷的准确的交叉参照，然后再增加其他功能。

语言数据资源的集成方案

语法信息词典 **GKB**

和

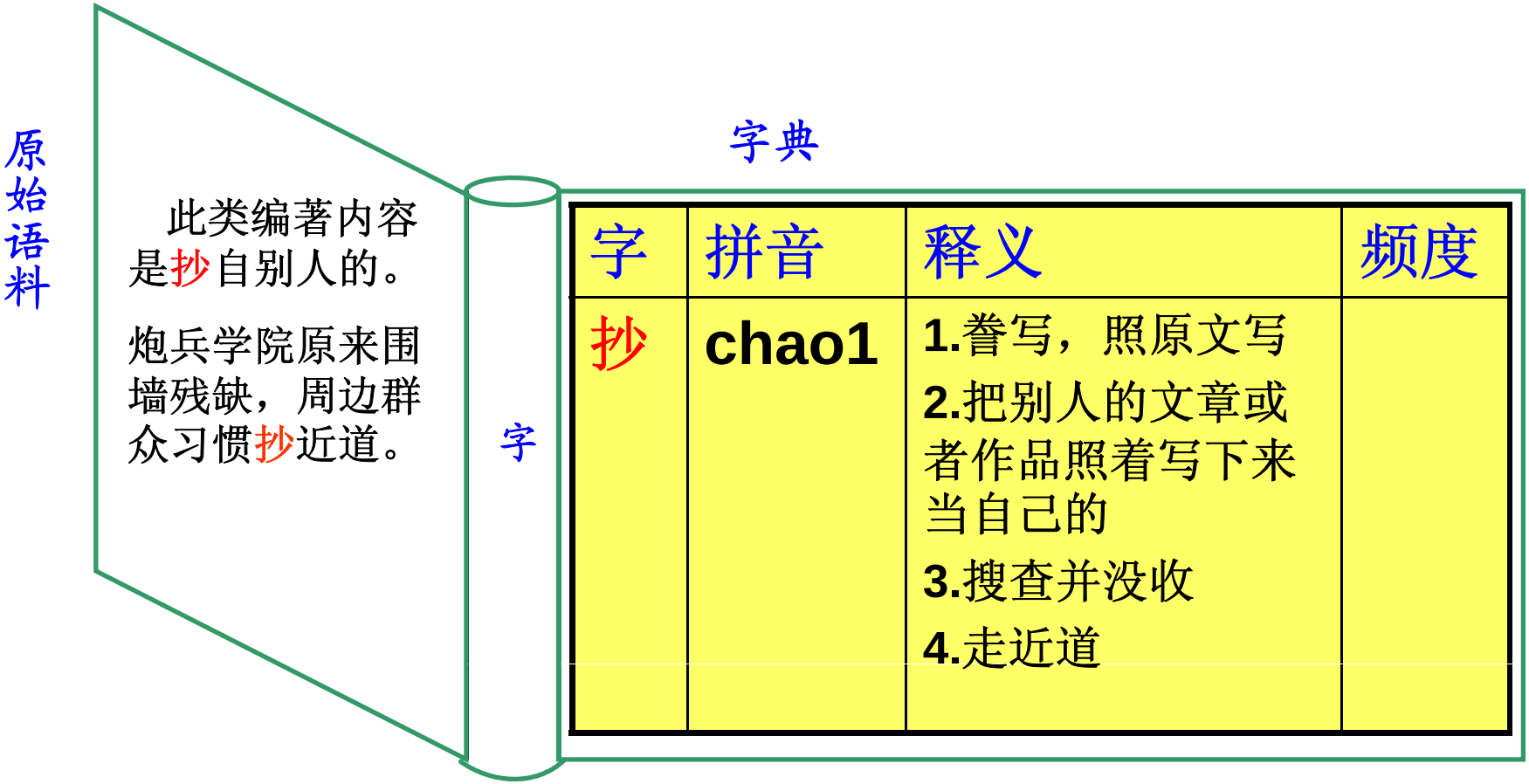
汉语语义词典 **CSD**

都是二维表，可以想象为一个平面。

同样，把 “标注语料库”也想象为一个平面。

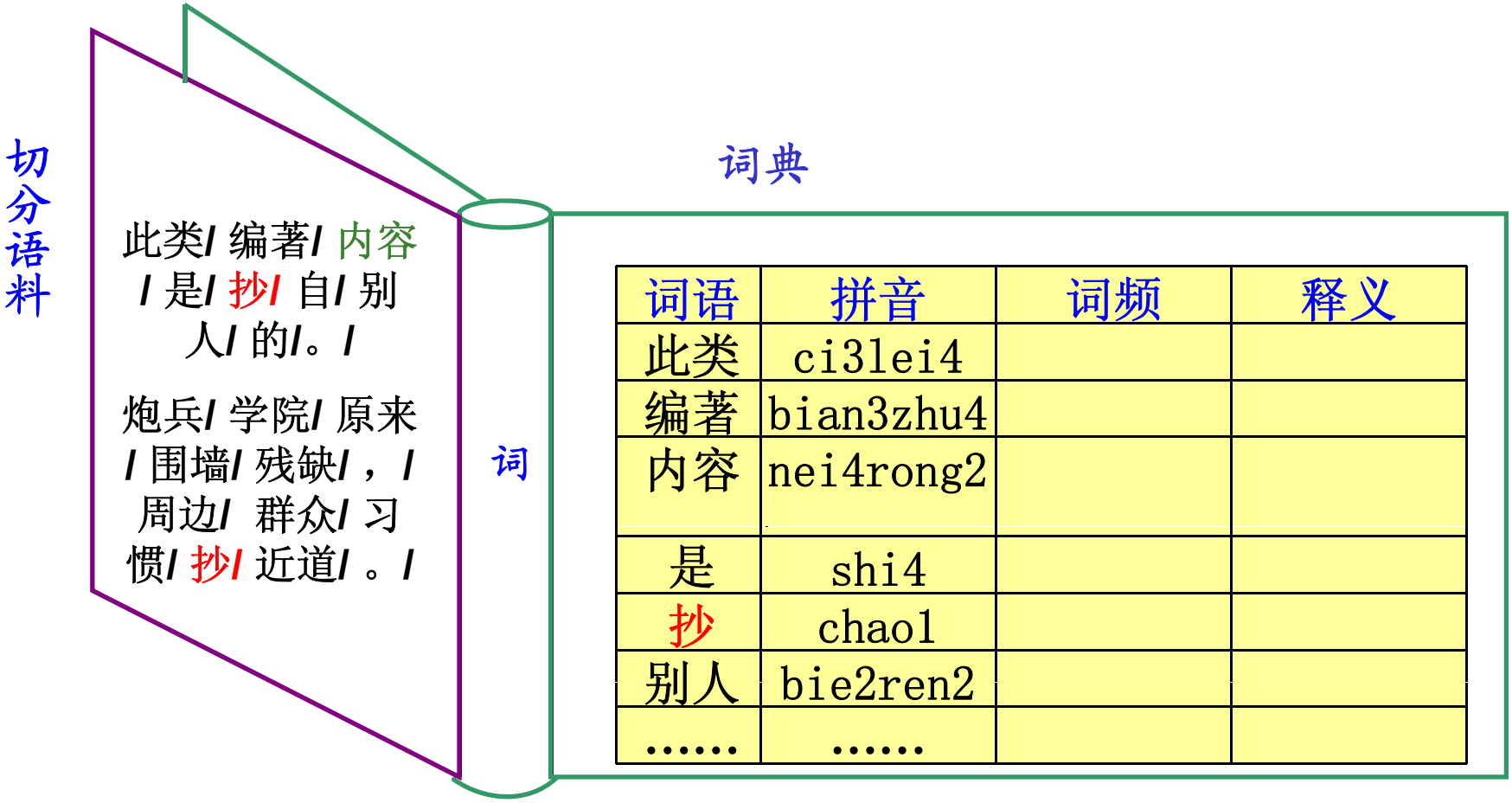
两个平面通过一根“轴”可以连接起来。

集成原始语料库和字典
——以“字” 为主轴



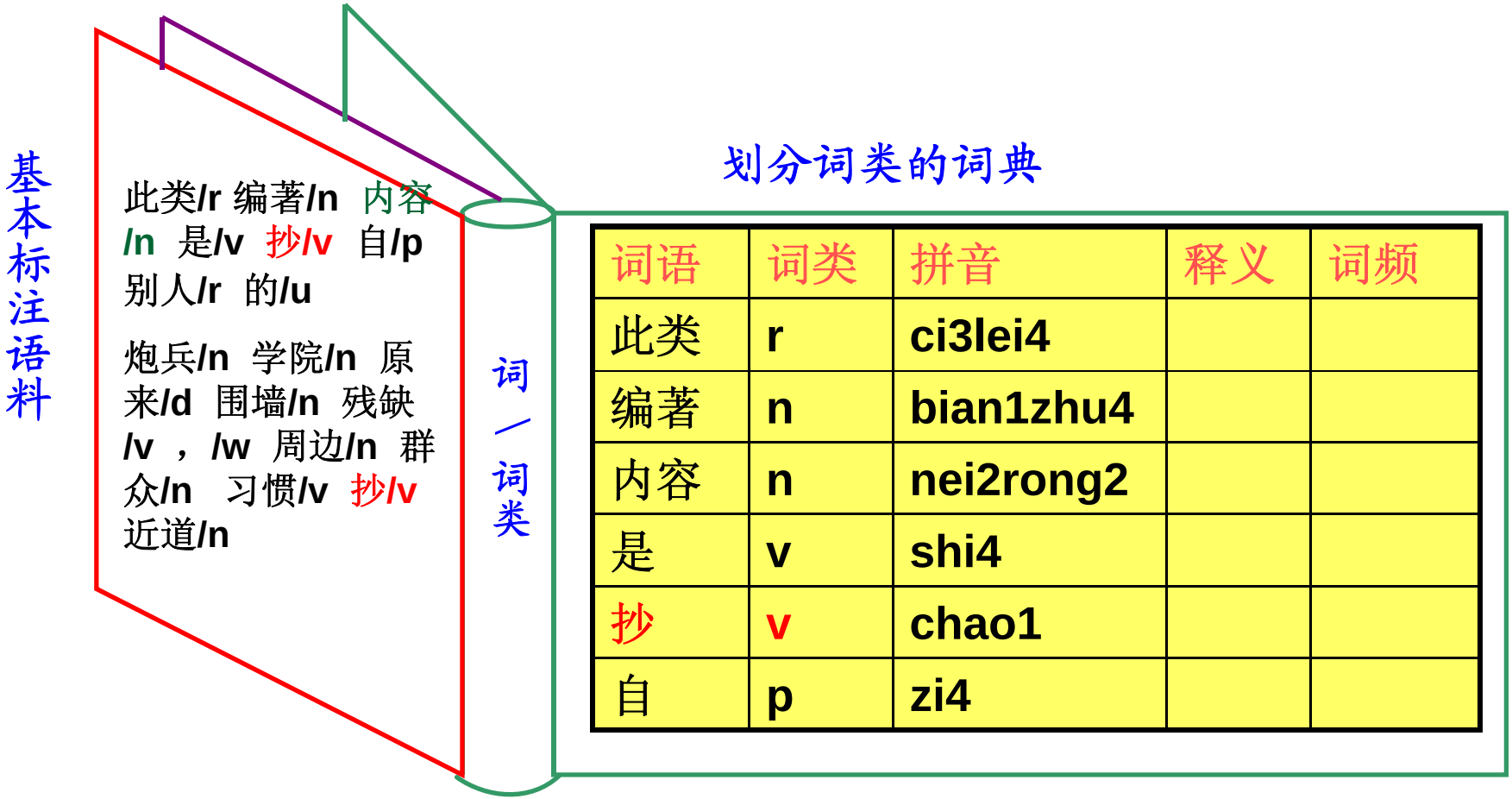
字	拼音	释义	频度
抄	chao1	1.誊写，照原文写 2.把别人的文章或者作品照着写下来当自己的 3.搜查并没收 4.走近道	

集成切分语料库和词典
——以“词” 为主轴



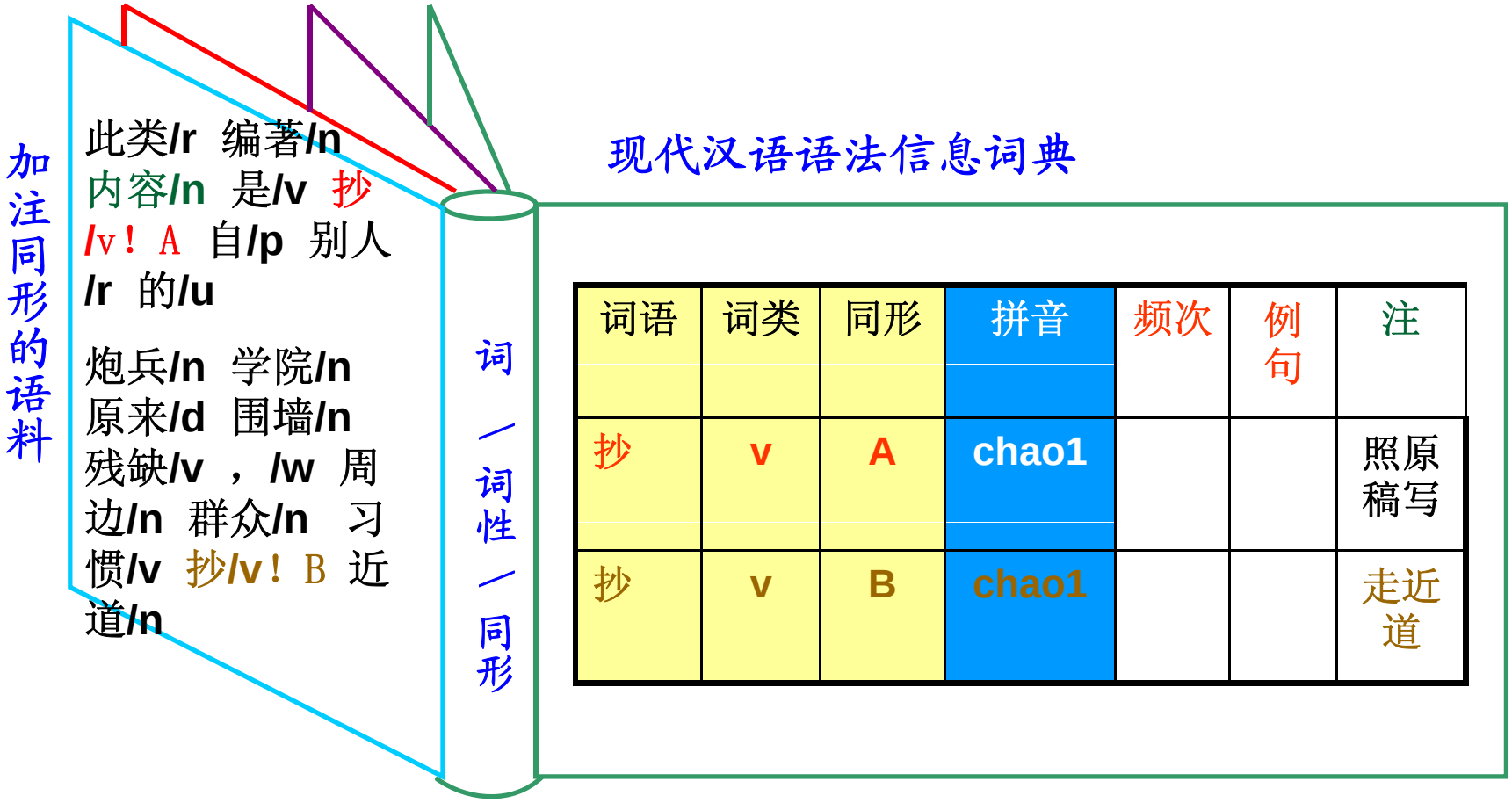
词语	拼音	词频	释义
此类	ci3lei4		
编著	bian3zhu4		
内容	nei4rong2		
是	shi4		
抄	chao1		
别人	bie2ren2		
……	……		

集成基本标注语料库和划分词类的词典
——以“词语”＋“词性”为轴



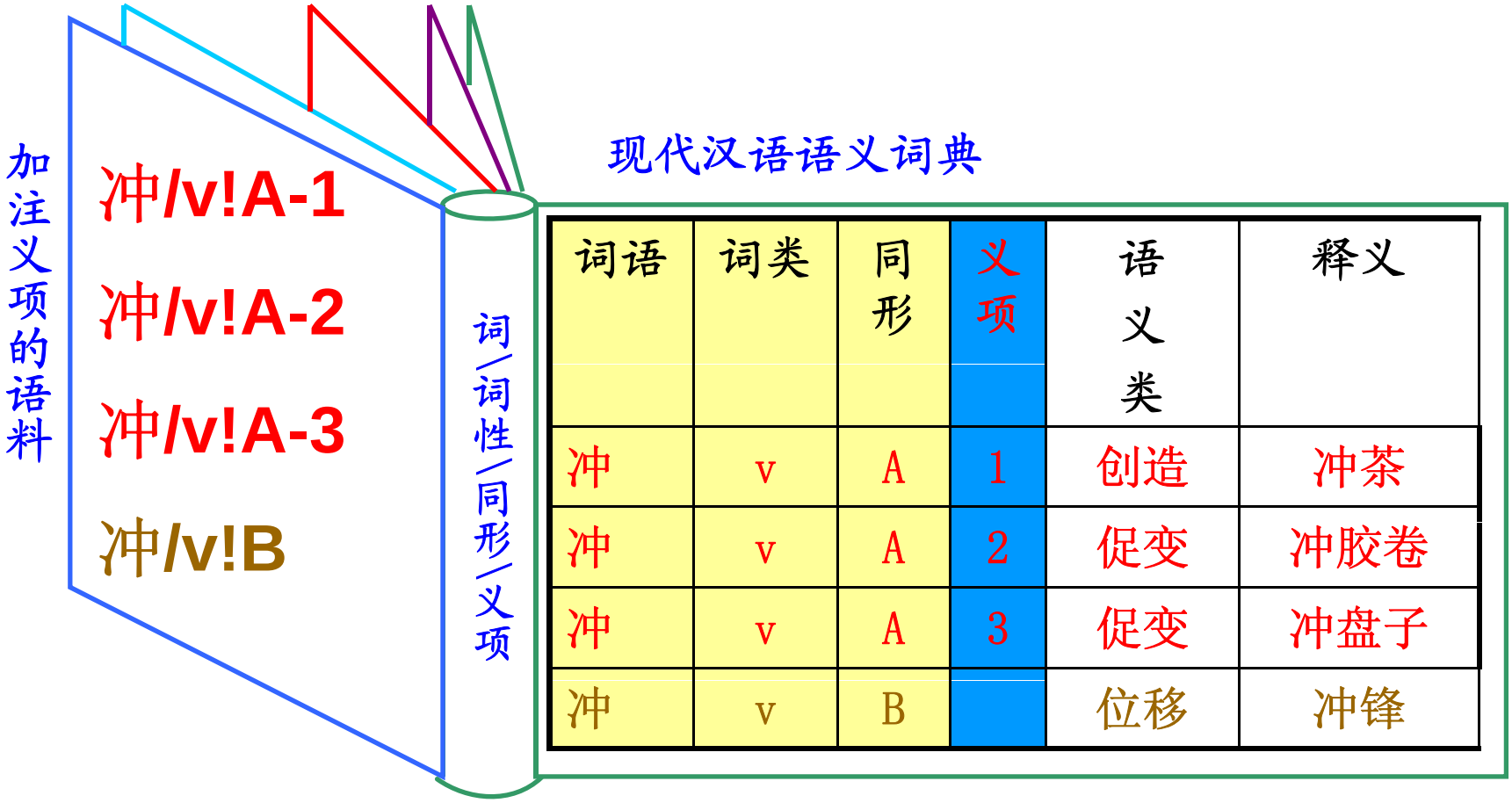
词语	词类	拼音	释义	词频
此类	r	ci3lei4		
编著	n	bian1zhu4		
内容	n	nei2rong2		
是	v	shi4		
抄	v	chao1		
自	p	zi4		

以粗粒度词义为主轴的集成方案



词语	词类	同形	拼音	频次	例句	注
抄	v	A	chao1			照原稿写
抄	v	B	chao1			走近道

以细粒度词义为主轴的集成方案



词语	词类	同形	义项	语义类	释义
冲	v	A	1	创造	冲茶
冲	v	A	2	促变	冲胶卷
冲	v	A	3	促变	冲盘子
冲	v	B		位移	冲锋

《语法信息词典》和“同形”标注语料库之集成方案

——文本型语料库与结构化语料库之双向转换

1998年《人民日报》基本标注语料库的文本形式样例如下所示:

19981201-01-002-001/m　圆满/ad　结束/v　对/p　俄罗斯/ns　和/c　日本/ns　的/u　访问/vn

19981201-01-002-002/m　江/nr　泽民/nr　主席/n　回到/v　北京/ns　朱/nr　镕基/nr　胡/nr　锦涛/nr　等/u　前往/v　[人民/n　大会堂/n]ns　迎接/v

（分别是199年12月1日第一版第二篇文章的第一段和第二段）

不同结构数据资源集成方案的实现技术：

《人民日报》基本标注语料库的结构化表示（关系数据库文件）

切分单位	长	年	月	日	版	篇	段	句	位
19981201-01-002-001/m	21	1998	12	01	01	02	001	01	00
圆满/ad	07	1998	12	01	01	02	001	01	01
结束/v	06	1998	12	01	01	02	001	01	02
对/p	04	1998	12	01	01	02	001	01	03
俄罗斯/ns	09	1998	12	01	01	02	001	01	04
和/c	04	1998	12	01	01	02	001	01	05
日本/ns	07	1998	12	01	01	02	001	01	06
的/u	04	1998	12	01	01	02	001	01	07
访问/vn	07	1998	12	01	01	02	001	01	08
19981201-01-002-002/m	21	1998	12	01	01	02	002	01	00
江/nr	05	1998	12	01	01	02	002	01	01
泽民/nr	07	1998	12	01	01	02	002	01	02
……									

可以同GKB，CSD连接 (JOIN)，一体化，便捷存取。
（同上页比，增加了“同形”不空的例子）

切分单位	标记	词语	词类	同形	……
19981201-01-002-001/m					
圆满/ad	ad	圆满	a		
结束/v	v	结束	v		
对/p	p	对	p		
俄罗斯/ns	ns	俄罗斯	n		
和/c	c	和	c		
日本/ns	ns	日本	n		
的/u	u	的	u		
访问/vn	vn	访问	v		
19981201-01-002-002/m					
江/nr	nr	江	Ng		
泽民/nr	nr	泽民			
……	……	……	……		
抄	v!A	抄	v	A	
抄	v!B	抄	v	B	
……	……	……	……	……	……

综合型语言知识库的利用

基于语言数据资源的知识挖掘研究及其成果

词汇知识

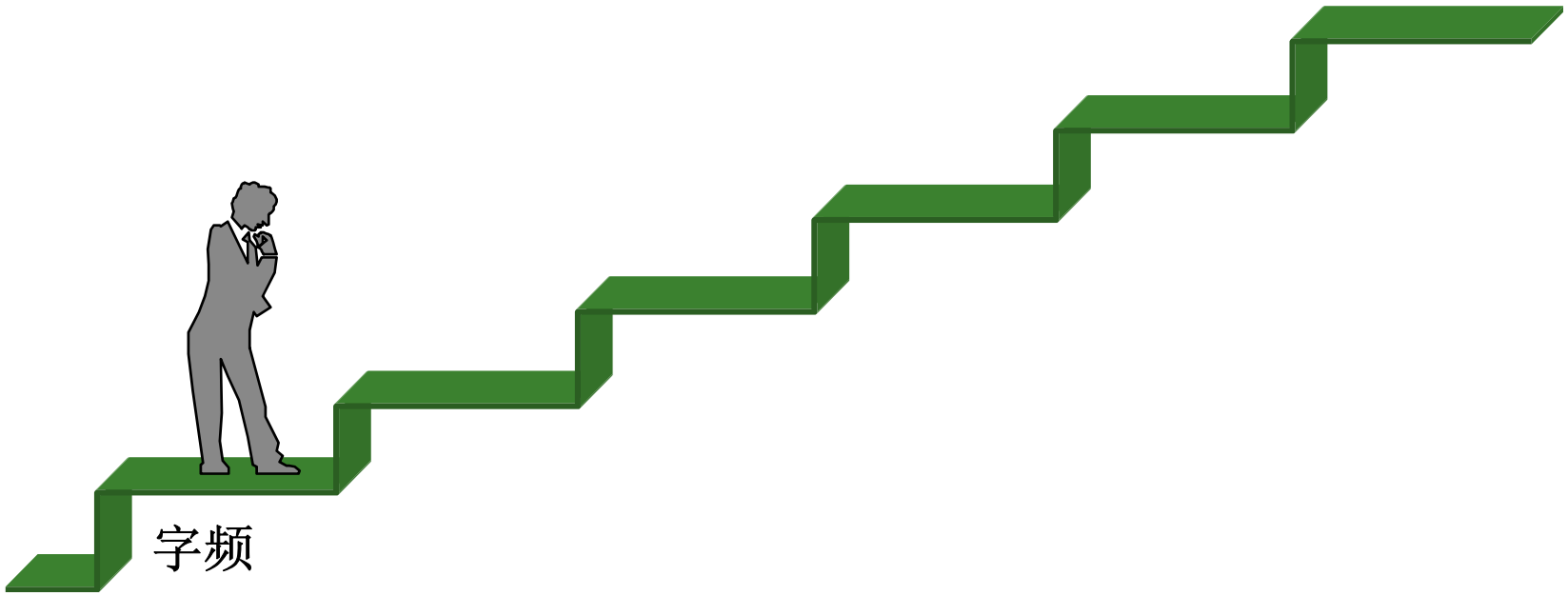
（1）词频、带词性的词频、词义频度统计

（2）词的分布均匀度（语文词的获取）

（3）兼类词的分布概率

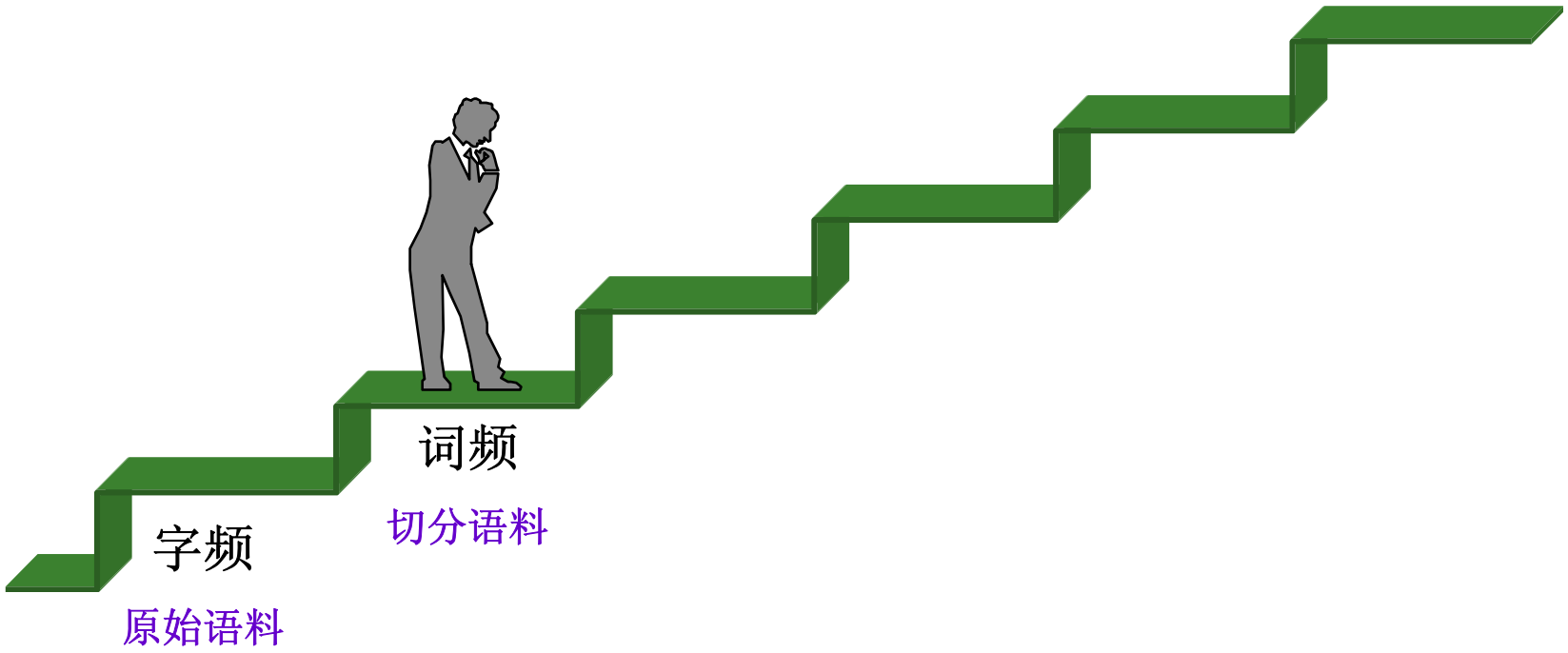
汉语词汇计量研究之发展历程

（伴随着语料库加工的逐步深入）



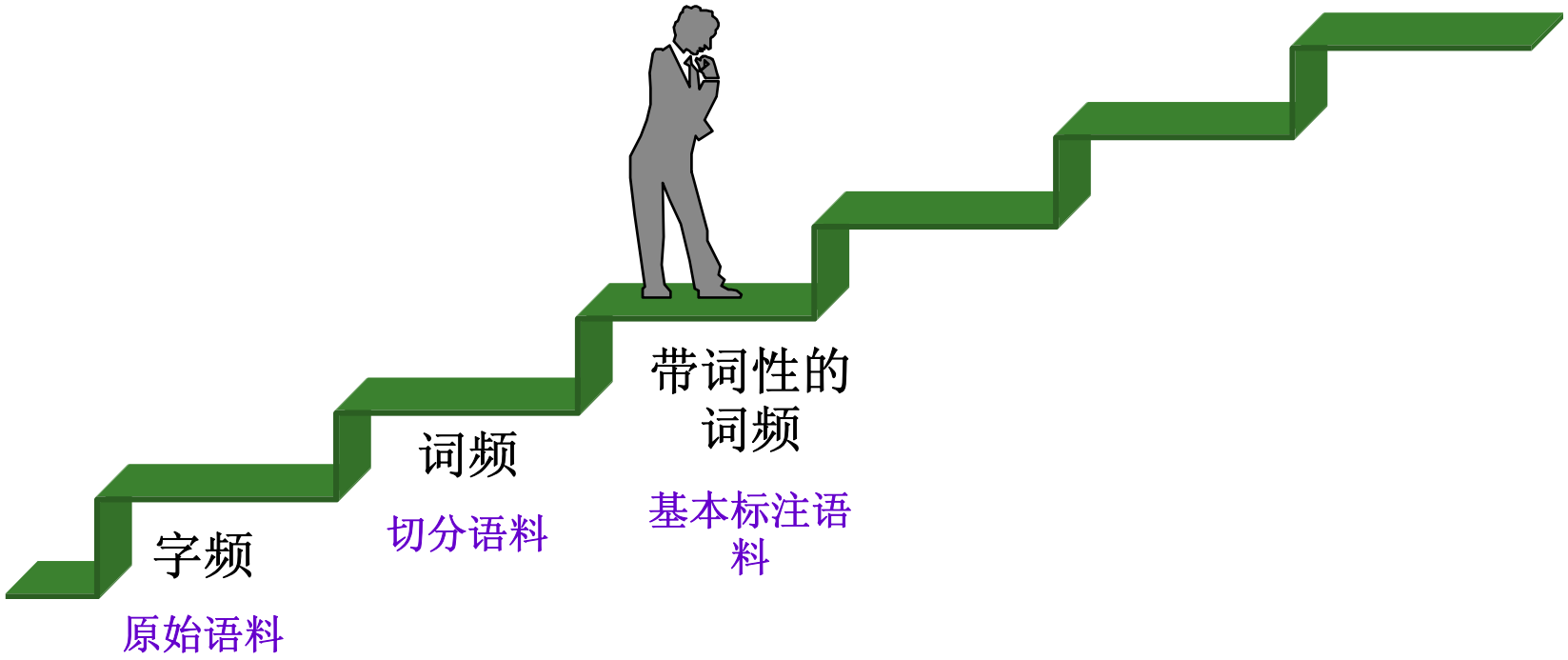
汉语词汇计量研究之发展历程

（伴随着语料库加工的逐步深入）

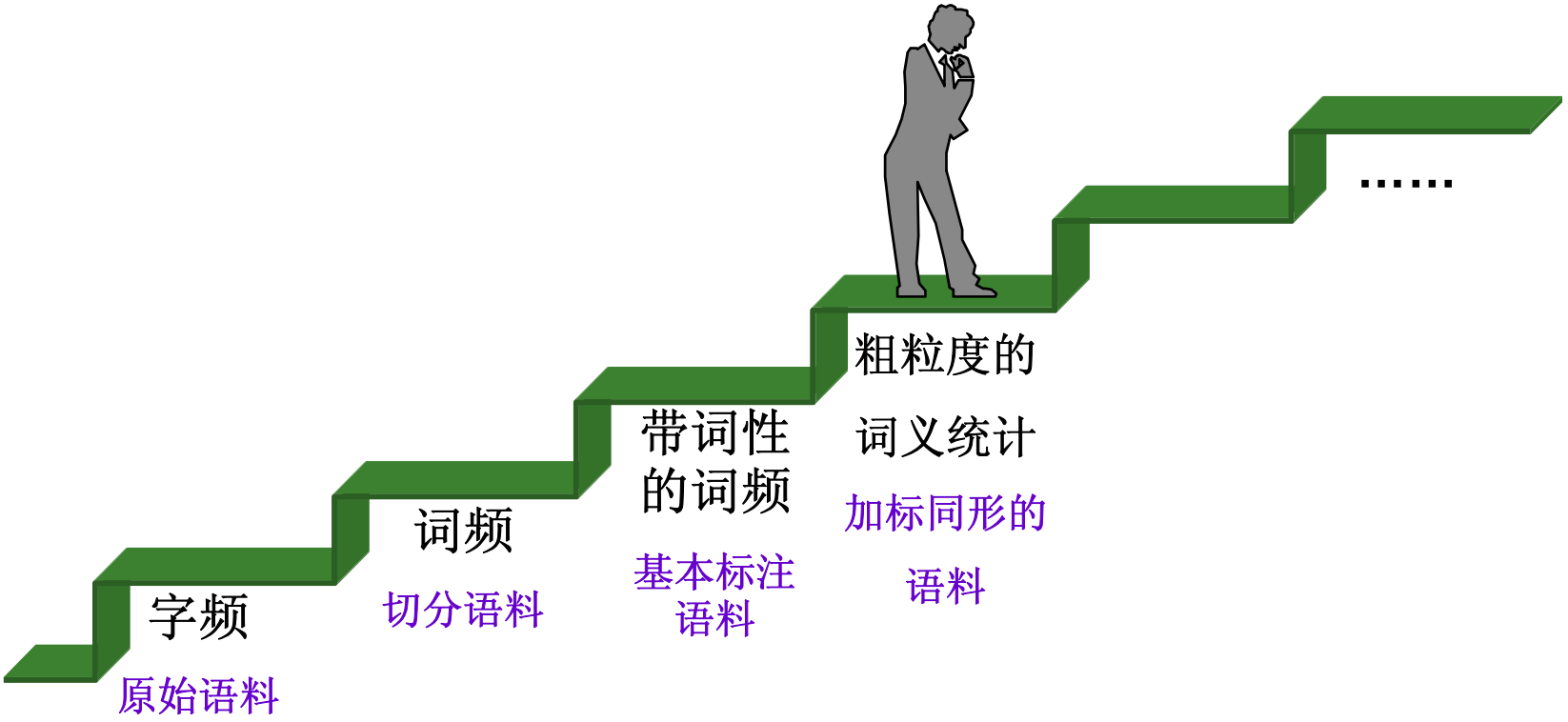


汉语词汇计量研究之发展历程

（伴随着语料库加工的逐步深入）

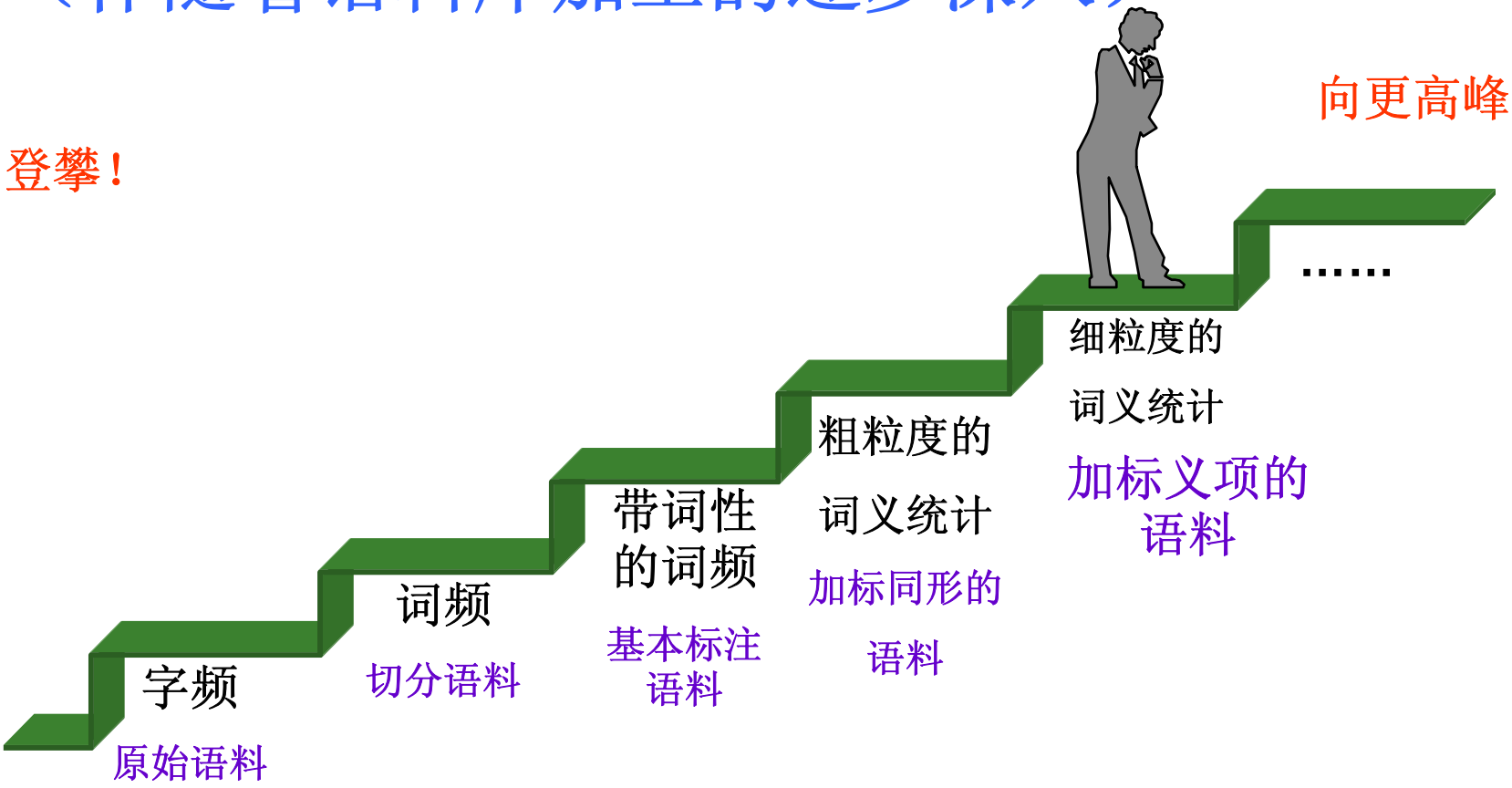


汉语词汇计量研究之发展历程
（伴随着语料库加工的逐步深入）



汉语词汇计量研究之发展历程

（伴随着语料库加工的逐步深入）



基于语言数据资源的知识挖掘研究及其成果

词汇知识

（1）词频、带词性的词频、词义频度统计

（2）词的分布均匀度（语文词的获取）

（3）兼类词的分布概率

词	词性1：概率	词性2：概率	词性3：概率	词性4：概率
把	p：0.96	q：0.03	v：0.01	m：0.00
被	p：1.00	Ng：0.00		
并	c：0.85	d：0.14	c：0.01	
次	q：1.00	Bg：0.00		
从	p：1.00	Vg：0.00		
大	a：0.92	d：0.08	v：0.00 ？	
到	v：0.80	p：0.20		
得	u：0.76	v：0.24	e：0.00	
等	u：0.98	v：0.02	q：0.00	
地	u：0.89	n：0.11		
对	p：0.98	v：0.01	q：0.01	a：0.00
就	d：0.87	p：0.13	c：0.00	
以	p：0.84	c：0.11	？ j：0.05	
由	p：1.00	v：0.00		
在	p：0.95	d：0.02	v：0.02	

基于语言数据资源的知识挖掘研究及其成果

（1）带词性的词频统计

（2）词的分布均匀度（语文词的获取）

（3）兼类词的分布概率

（4）动词、形容词向名词漂移现象的考察

圆满/ad　结束/v　对/p　俄罗斯/ns　和/c
日本/ns　的/u　访问/vn

“访问”是名词性短语的中心词：名词？动词？

汉语学界存在不同学派：

朱德熙先生的“名动词”说。

也有主张标注为名词的。

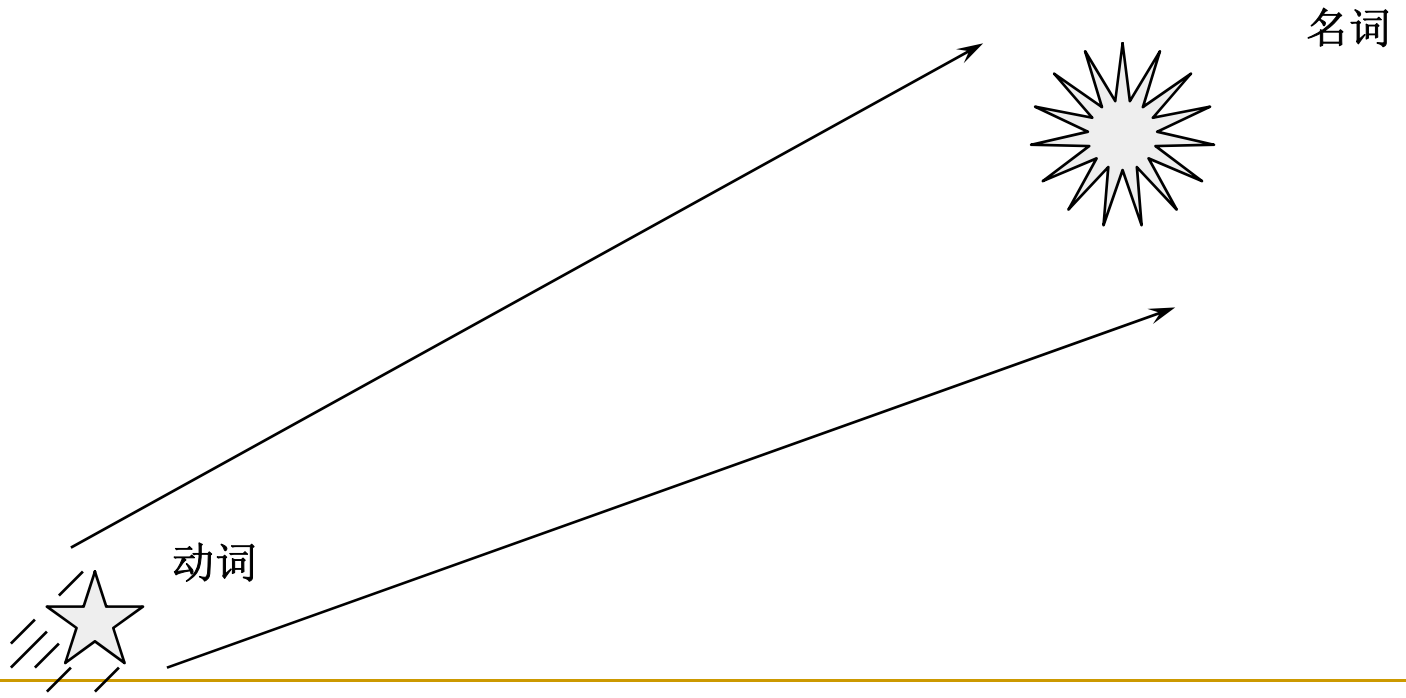
胡明扬先生的“漂移说”。

北大语料库基于朱先生观点标注为　vn

（现在可以对这类词的动态现象进行统计研究）

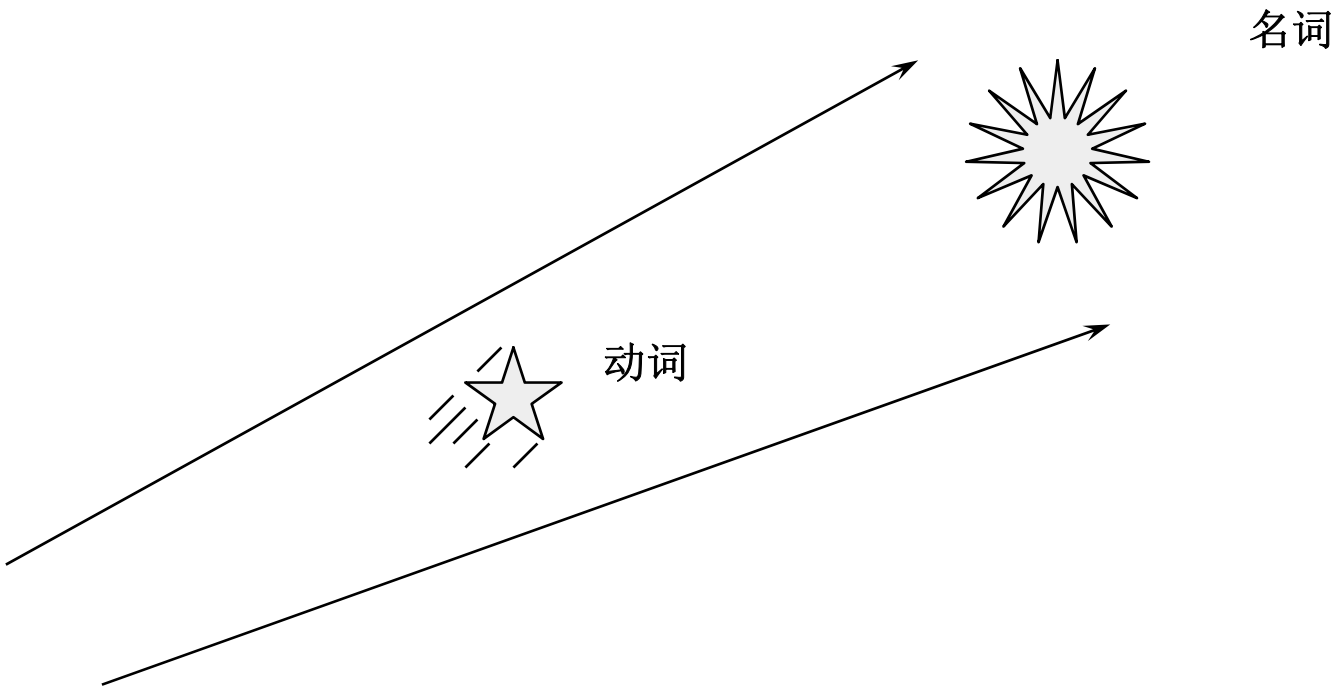
若P(v)>> P(vn)，则该动词向名词彼岸的漂移刚刚开始。

- 提高：　P(v)=0.89,　P(vn)=0.11
- 会见：　P(v)=0.94,　P(vn)=0.06
- 出版：　P(v)=0.73,　P(vn)=0.27
- 处理：　P(v)=0.68,　P(vn)=0.32



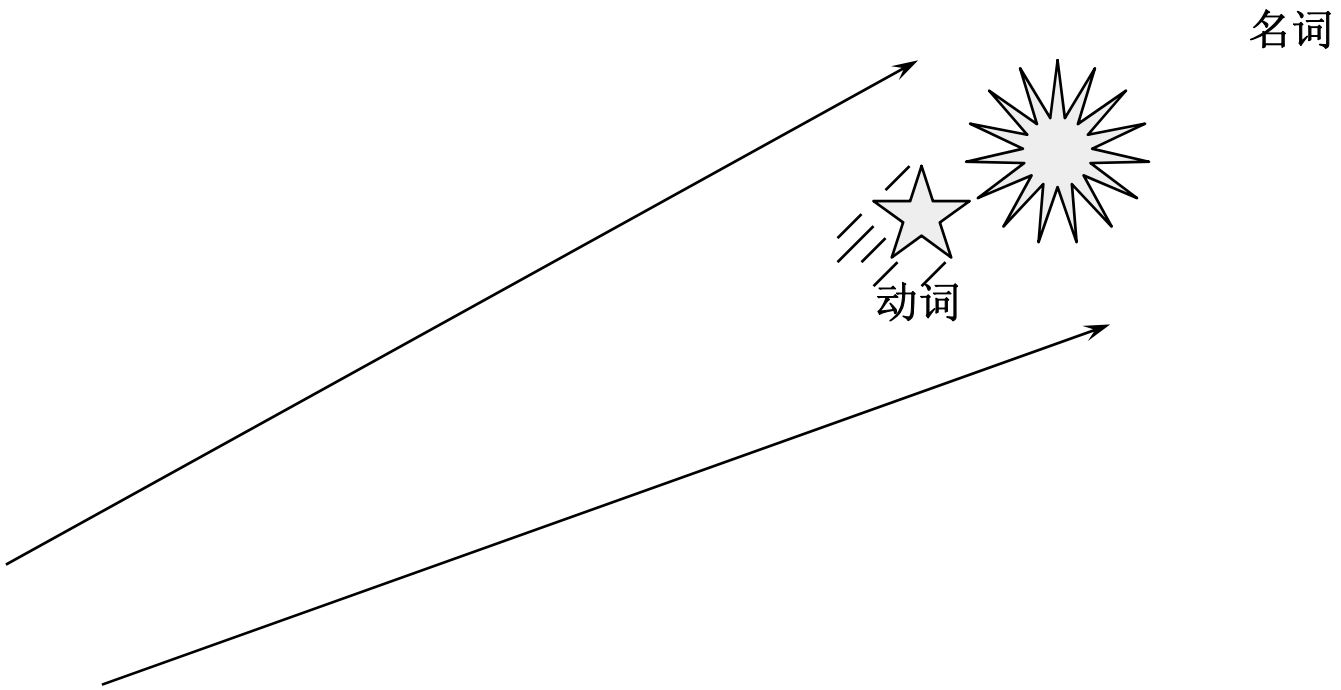
若P(v) ≈ P(vn)，则该动词处于向名词漂移的过程中。

- 改革：　P(v)=0.46,　P(vn)=0.54
- 发展：　P(v)=0.48,　P(vn)=0.52
- 选举：　P(v)=0.50,　P(vn)=0.50
- 认可：　P(v)=0.52,　P(vn)=0.48



若P(v)<< P(vn)，则该动词已漂移到接近名词的彼岸。

- 处分：　　P(v)=0.22,　P(vn)=0.78
- 教育：　　P(v)=0.13,　P(vn)=0.87
- 挫折：　　P(v)=0.25,　P(vn)=0.75



基于语言数据资源的知识挖掘研究及其成果

（1）带词性的词频统计

（2）词的分布均匀度（语文词的获取）

（3）兼类词的分布概率

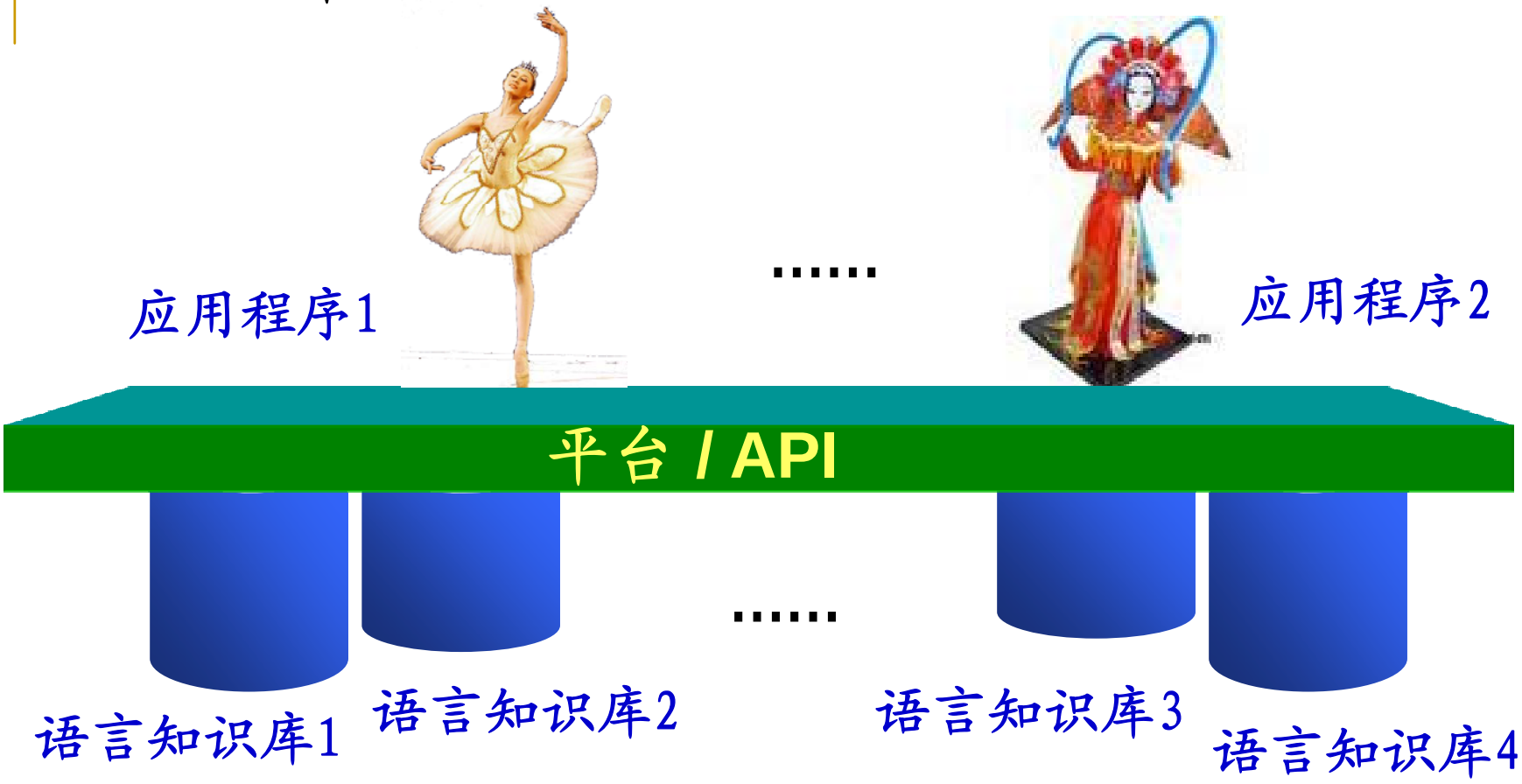
（4）动词、形容词向名词漂移现象的考察

（5）词的语法属性的概率值

若干动词某些属性的“可否”值与统计次数的参照

动 词		副 词										助 词					
词语	频度	不		没		很		在		正在		着		了		过 U	
吃	3125	98		24		0		7	在	2		25	着	56	了	85	过
到	3058	2045		9		0		0		0		0		2454	了	122	过
发	2261	26		8		0		0	在	0		5	着	194	了	12	过
发展	21044	27		0		0		72	在	16		20	着	161	了	1	过
进	4429	38		7		0		1		2		15		584	了	0	过
进行	19826	32		1		0		70	在	193		79	着	3572	了	77	过
睡	289	7		7		0		0	在	0		9	着	5	了	9	过
说	34354	96		29		0		12	在	0		96	着	141	了	262	过
听	1667	44		4		0		2	在	0		74	着	210	了	21	过
为	26838	76		0	否	0		0		0		0		2		0	
想	4340	262		19		38	很	31	在	0		89	着	38	了	34	过
要	41148	364		8		0		0	在	0		0	着	2	了	2	过
有	60910	0	否	240		573		0		9		27	着	2739	了	316	过
走	7616	70		2		0		6	在	7		16	着	242	了	2	过

语言知识库与应用系统



在语言知识库搭建的平台上可以上演
威武雄壮生动活泼的应用系统的剧目

主要内容

- 关于研究对象与目标
- 自然语言处理的主攻方向
- 综合型语言知识库概要
- 前进目标——自然语言理解
- 领域知识工程与领域知识库
- 结语与致谢

前进目标——自然语言理解

- 尽管目前我们处于自然语言处理的初级阶段，但我们要记着是在向自然语言理解的最高境界登攀。
- 自然语言理解研究需要评测手段。图灵测试仍在进行。2008年10月12日第18届图灵测试在英国雷丁大学进行。测试结果显示：6个人机对话程序中的每一个程序至少欺骗了12名裁判中的一名裁判，最聪明的程序Elbot欺骗了3名，获得铜牌（金牌、银牌空缺）。现在还只是针对文本，今后还要处理音频及视频信息。
- 目前的自然语言处理研究的主攻方向集中在歧义消解（本质是分类问题）？即便作为热点的情感计算仍没超出歧义消解的范围？
- 消解了歧义是否实现了理解？ 隐喻、影射、双关、夸张、幽默、拟人 以及遣词造句的技巧对对自然语言理解研究提出了挑战——超出歧义范围？

  实例（双关）： “您的健康是天大的事——天大药业”

  “您 的 健康 是 天 大 的 事”

  “您 的 健康 是 天大 的 事”

前进目标——自然语言理解

1 对隐喻的认识

语句级隐喻　　张三简直就是一头狮子　/　李四才是老狐狸

篇章级的隐喻：打起黄莺儿，莫叫枝上啼。啼时惊妾梦，不得到辽西。

隐喻的机器识别与理解超出了歧义消解的范围？

一般的，消歧是“同中辨异”，隐喻的识别与理解则是“异中求同”。

2 如何识别与理解隐喻

发现本体与喻体之间的冲突，寻找喻底（共同属性）。在我们的知识体系ontology中，“狮子”和“狐狸”本不是“人”的上位概念，不合逻辑，但通过包含喻底的实例

狮子勇猛地扑向猎物。

狐狸再狡猾，街上还有卖狐狸皮的。

则可以识别上述隐喻表达并理解到 “张三是勇猛的”，“李四是狡猾的”。

河北有个老太太吃土块——不合常识，但并不是隐喻。

男人都是动物——符合逻辑，可能是隐喻，也可能不是。新的歧义现象。

隐喻与歧义纠结在一起，要根据更大的语境才能判断，才能实现理解。

对自然语言理解形成另一个挑战？

隐喻自动处理是自然语言理解征途上必须逾越的障碍？

隐喻自动处理的实际应用

（1）翻译（机器翻译、机器辅助翻译）

翻译可以作为检验隐喻识别与理解的一个指标。

铁榔头：**iron hammer**？**iron fist**？

翻译还涉及文化问题。

（2）检索

—— 提高准确率

区分词语的本义和隐喻义（网页索引与查询都需要）。

过滤掉不相关的隐喻用法的文献或网页，可以提高本义检索的准确率（如：“航班起飞时间”，“起飞跑道距离”等等，排除“经济起飞”、“东方美女歌坛起飞”）。

—— 提高召回率

对于本体（目标域）不出现的情况，可以通过喻体（源域）分析出关于本体的描述，提高召回率。如常用“祖国的花朵”比喻“儿童”，其中并不出现“儿童”。当搜索“儿童”时，
包含“祖国的花朵”的文献或网页往往也是相关的。

实际应用还要综合考虑投入-产出性能比等多种因素。

前进目标——自然语言理解

ICL/PKU 关于隐喻的研究

- 王治敏博士：名词性短语隐喻的识别（知识海洋）
- 贾玉祥：动词性隐喻的识别（汽车喝汽油）
- 贾玉祥：一类隐喻句的自动理解（女人是水）

隐喻自动处理同样需要隐喻知识库的支持

用于识别的名词性隐喻知识库

用于识别的动词优先选择语义类之知识库

用于一类隐喻句理解的名词显著特征知识库

用于识别的名词性隐喻知识库

经常作为“源域”的名词库是知识库的组成部分

主线	种子	支柱	钥匙	雨露	阴云	摇篮	阳光	烟水
漩涡	序幕	星	峡谷	舞台	微波	土壤	隧道	曙色
食粮	森林	热浪	桥梁	旗帜	喷泉	纽带	泥潭	泥坑
门槛	脉络	脉搏	绿洲	路线	龙头	灵魂	浪潮	精髓
枷锁	脊梁	激流	基石	火炬	火花	火	灰尘	花
河流	气息	航船	海洋	瑰宝	光线	光辉	光	痼疾

用于识别的名词性隐喻知识库

标注了隐喻短语的语料库也是知识库的组成部分

- 19980123-11-003-008/m 握/v 着/u 这样/r 的/u 手/n ，/w 能/v 不/d 倍感/v 温暖/a 和/c 鼓舞/a ？/w 像/p 他们/r 这个/r 年龄/n 的/u 小伙子/n ，/w 有/v 多少/r 还/d 依偎/v 在/p 父母/n 的/u 身旁/s ，/w 享受/v 着/u 温馨/a 的/u 抚爱/vn ；/w 有/v 多少/r 正/d 坐/v 在/p 宽敞/a 明亮/a 的/u 教室/n ，/w 遨游/v 于/p 〈知识/n 的/u 海洋/n〉。/w 而/c 他们/r ，/w 却/d 走/v 进/v 了/u 绿色/n 的/u 军营/n ，/w 来到/v 了/u 北大荒/ns ，/w 用/p 青春/n 和/c 血汗/n ，/w 为/p 军队/n 为/p 国家/n 创造/v 巨大/a 的/u 财富/n 。/w

- 19980303-05-003-003/m 与会/vn 同志/n 认真/ad 回顾/v １９９７年/t 全军/n 文艺/n 工作/vn 后/f 认为/v ，/w 军队/n 文艺/n 创作/vn 的/u 根本/a 任务/n 是/v 贯彻/v 党/n 的/u 十五大/j 精神/n ，/w 高举/v 邓小平理论/n 伟大/a 旗帜/n ，/w 站/v 在/p 〈时代/n 高度/n〉，/w 积极/ad 弘扬/v 主旋律/n ，/w 为/p 部队/n 和/c 社会/n 提供/v 更/d 多/a 更/d 好/a 的/u 〈精神/n 食粮/n〉；/w 加强/v 文艺/n 创作/vn 要/v 从/p 深入/v 生活/vn 抓起/v ，/w 把握/v 〈时代/n 脉搏/n〉，/w 使/v 作品/n 更/d 具有/v 时代/n 〈生活/vn 气息/n〉；/w 要/v 进一步/d 树立/v 精品/n 意识/n ，/w 把/p 主要/b 精力/n 放/v 到/v 军事/n 题材/n 创作/vn 上/f ，/w 在/p 艺术/n 上/f 进行/v 新/a 的/u 探索/vn 和/c 表现/vn ，/w 开创/v 崭新/b 的/u 多样化/vn 创作/vn 局面/n ；/w 全面/ad 提高/v 文艺/n 队伍/n 的/u 整体/n 素质/n ，/w 使/v 我们/r 的/u 队伍/n 成为/v 一/m 支/q 政治/n 强/a 、/w 业务/n 精/a 、/w 作风/n 硬/a 、/w 特别/d 能/v 战斗/v 的/u 队伍/n 。/w

用于识别的名词性隐喻知识库
——源域与目标域对应的知识库之样例

源域词语	搭配词	目标域词语	标记词	隐喻类型	CCD映射	同形	评价
海洋		知识/时间/人生/歌/花/舞/笑声/灯/市场/经济/光/科学/知识/彩旗/大赛/ 生活	的		是		
大战		非典/洪水/挂历/羊绒衫/鳗苗/储蓄/利率/啤酒/价格/贸易/点球			是		
道路	健康	谈判/社会主义/和平/人生/革命/资本主义/致富	的		是		
堤坝		旧观念/心灵/ 理智/心/感情	的		是		
地狱		人间					消极

用于识别的名词性隐喻知识库

——源域（“海洋”）与目标域基于CCD的概念映射之样例

词语	词性	频次	offset	type	csynset	cdescription	cnote
知识		16	03_00013243	心理特征 精神特征	学识 学问 知识 学力 见闻 认知 认识	通过感觉学习和推理得到的心理成果；	从书本上可以获得许多知识；
时间		2	03_00015594	抽象概念 抽象观念 抽象概括	光阴 岁月 时间 工夫 年光 时光 时日 时期 期间 辰光 蹉跎岁月 春光	事物存在或继续的期间，从过去到现在再到未来	我们应该珍惜时间；
人生		3	28_10868596	time	人生 一生 一辈子	某人的一个特定生活阶段的一段时间	X

飘流/v 在/p 时间/n 的/u 海洋/n 里/f 人生/n 的/u 海洋/n 浩淼无涯/l

前进目标——自然语言理解

- 以上讨论只限于文本信息。仅在文本范围内，能否实现自然语言理解？很久以来一直在思考这个问题，以往也发表过只言片语，尚未形成清晰、完整的认识。
- 注意到：图灵测试今后还要处理音频及视频信息。
- 多模态信息的融合与交互作用

  ——顾曰国教授建立了记录实际场景的现场即席话语多模态语料库（包括话语活动的音频、视频文本及其转写的文字）。
- 关注脑科学、认知科学的进展

  ——语言学、认知语言学对自然语言理解研究的价值

  ——隐喻与人类的认知机制密切相关

  需要多学科的通力合作、协同攻关。

前进目标——自然语言理解

关于语言知识库的规划—— 向深度、广度发展

（1）建设综合型语言知识库系统，

开发概率型汉语句法语义词典。

填平补齐，提高质量，扩大规模，不惜浴火重生。

（2）筹划多模态语言知识库

—— 寄希望于同行的伙伴

（3）探索应用的新领域

—— 例如：文本信息隐藏与数字水印

主要内容

- 关于研究对象与目标
- 自然语言处理的主攻方向
- 综合型语言知识库概要
- 前进目标：自然语言理解前进
- 领域知识工程与领域知识库
- 结语与致谢

领域知识工程

- 领域知识工程的价值
- 领域知识工程的基本思路

  ——利用自然语言处理技术（计算术语学以及文本知识挖掘等），通过计算机辅助的方式从多种信息来源中发现科技术语以及与术语相关的概念知识，进而在此基础上，构建各个学科领域的领域知识库。

面向领域知识工程的计算术语学

计算术语学 研究内容:

- 术语自动提取
- 术语表达的概念的关系的自动提取
- 术语（概念）定义的自动获取
- 领域知识本体（架构）的自动构建

领域知识工程与领域知识库

领域知识工程：建立各个学科的领域知识库，实现面向智能信息处理和全社会信息服务、知识服务。

- 领域术语库
  - 信息科学技术领域术语库：包含计算机技术、通信技术、多媒体技术、自动控制技术、视频技术、遥感技术等多个专业领域
  - 面向数字奥运的体育、商务、餐饮、旅游领域术语库
- 领域知识库（雏形）
  - 计算机硬件领域知识库
  - 现代医学领域心血管疾病知识库

应用系统（一）

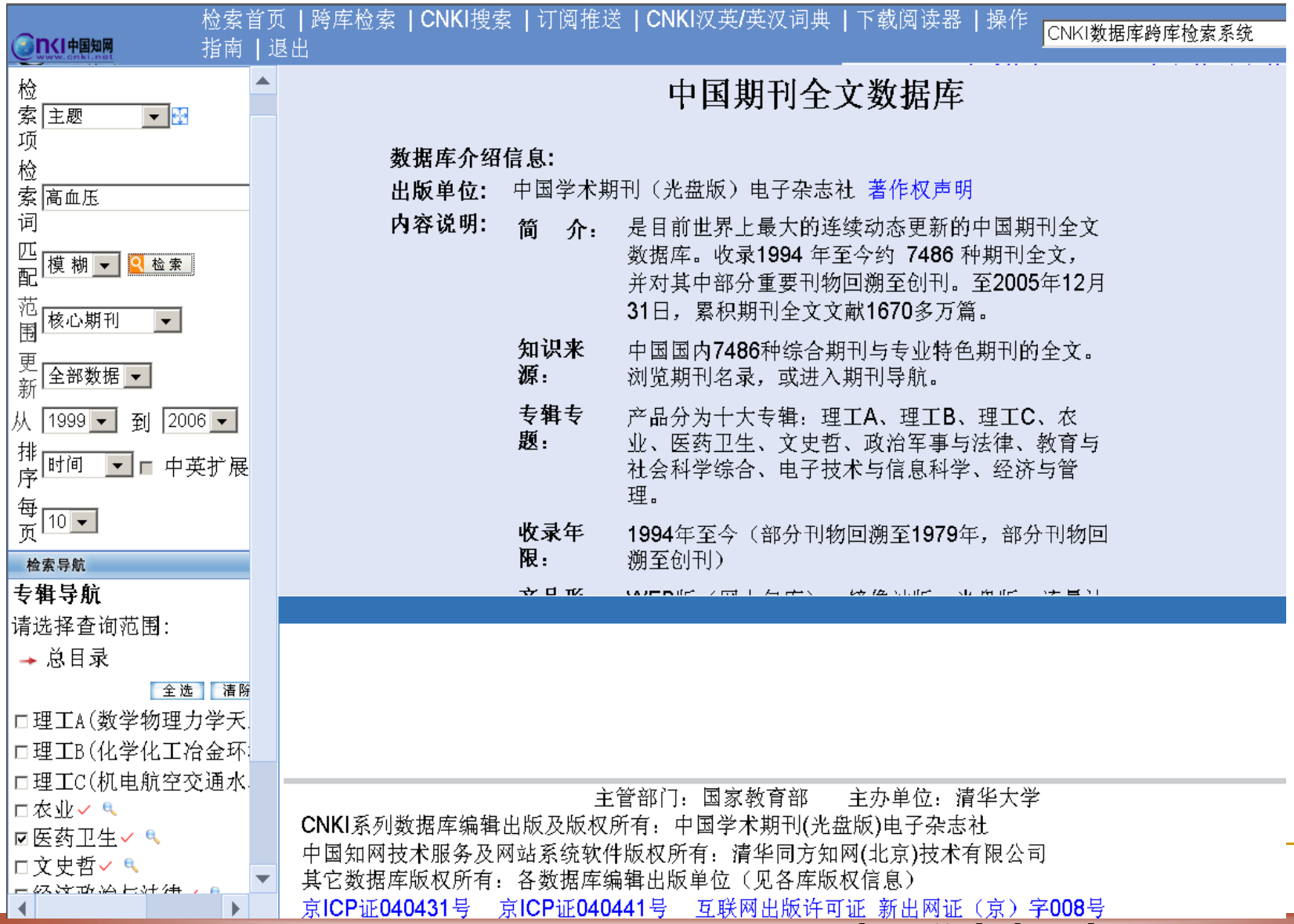
面向百科全书编纂的语言分析及知识更新平台

- 探索在网络时代编纂百科全书的全新模式
  - 利用先进的计算机网络技术以及北京大学计算语言学研究所在汉语信息处理领域长期积累的各种语言分析和处理工具，针对《全书》编撰和再版的工作需要，定制一系列智能化和人机互助、动态开放的语言分析和知识更新的软件工具，为百科全书各个专业领域的专家服务，为百科全书的编纂和再版提供便利。
- 利用自然语言处理技术建立知识元数据库的前期探索
  - 从充分开发和利用百科全书资源的角度出发，利用自然语言处理技术和语言分析工具分析百科全书资源，发现其中的知识点以及知识点之间的内在关联，将大量的、不断出现的知识点结构化地组织和关联起来，建立实验型知识元数据库原型系统，构建百科知识导航系统。

应用系统（二）
知识元数据库及基础平台

- 总体目标：基于百科全书，建立知识元数据库及其基础平台。面向社会、面向不同知识层次的人群，提供知识服务，推进科普知识的教育，提高我们民族的知识水平和素质 。
- 建立各个学科领域的知识元基础数据库
- 为社会提供全方位和多样化的知识服务
- 支持各种类型的智能化信息处理
- 为我国的信息化建设和文化建设提供知识基础设施

检索“高血压”

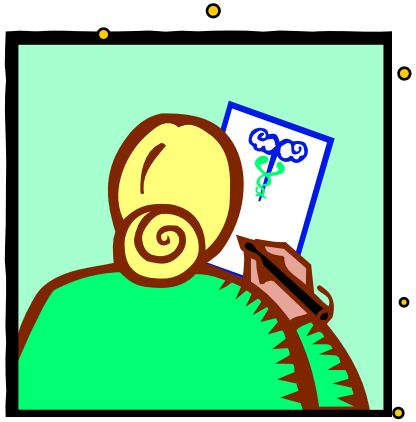

检索首页 | 跨库检索 | CNKI搜索 | 订阅推送 | CNKI汉英/英汉词典 | 下载阅读器 | 操作指南 | 退出
CNKI数据库跨库检索系统
检索项 主题
检索词 高血压
匹配 模糊 检索
范围 核心期刊
更新 全部数据
从 1999 到 2006
排序 时间 中英扩展
每页 10
检索导航
专辑导航
请选择查询范围:
总目录
全选 清除
理工A(数学物理力学天
理工B(化学化工冶金环
理工C(机电航空交通水
农业
医药卫生
文史哲
中国期刊全文数据库
数据库介绍信息:
出版单位: 中国学术期刊（光盘版）电子杂志社 著作权声明
内容说明:
简 介: 是目前世界上最大的连续动态更新的中国期刊全文数据库。收录1994 年至今约 7486 种期刊全文，并对其中部分重要刊物回溯至创刊。至2005年12月31日，累积期刊全文文献1670多万篇。
知识来源: 中国国内7486种综合期刊与专业特色期刊的全文。浏览期刊名录，或进入期刊导航。
专辑专题: 产品分为十大专辑：理工A、理工B、理工C、农业、医药卫生、文史哲、政治军事与法律、教育与社会科学综合、电子技术与信息科学、经济与管理。
收录年限: 1994年至今（部分刊物回溯至1979年，部分刊物回溯至创刊）
主管部门：国家教育部 主办单位：清华大学
CNKI系列数据库编辑出版及版权所有：中国学术期刊(光盘版)电子杂志社
中国知网技术服务及网站系统软件版权所有：清华同方知网(北京)技术有限公司
其它数据库版权所有：各数据库编辑出版单位（见各库版权信息）
京ICP证040431号 京ICP证040441号 互联网出版许可证 新出网证（京）字008号

检索结果：5245条记录

什么是高血压?

高血压的症状有哪些?

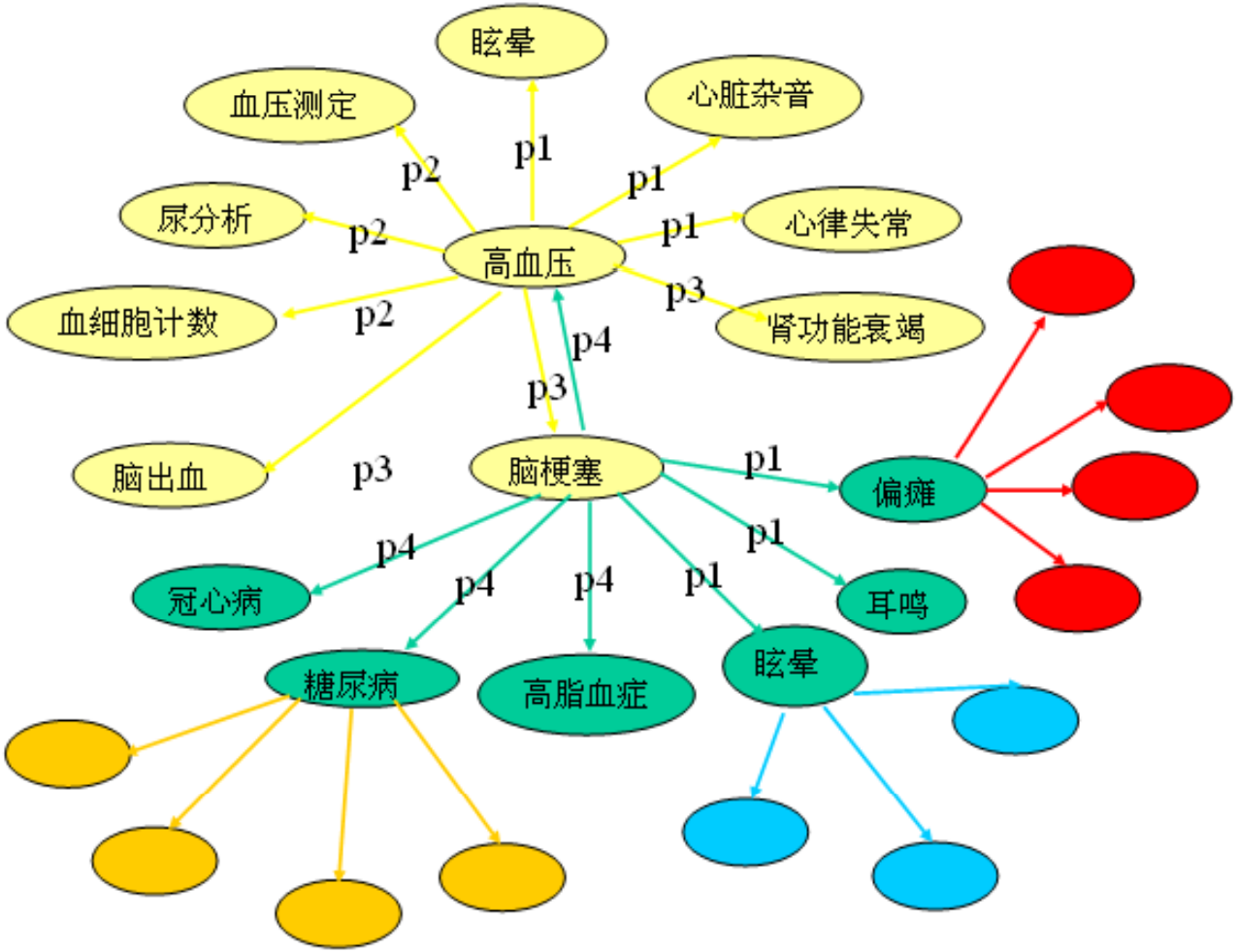
确诊高血压要做哪些检查?



怎样预防高血压?

…… ?

知识元数据库——以知识元（概念）为结点，建立概念之间的关联，构建网状的知识元数据库。



领域知识工程前程远大

信息需要转化为知识!

由信息服务转化为知识服务!

自然语言处理技术大有用武之地!

“国家知识资源数据库工程”

——最近一次论证于2009年2月8日西苑宾馆进行

主要内容

- 关于研究对象与目标
- 自然语言处理的主攻方向
- 综合型语言知识库概要
- 向自然语言理解前进
- 领域知识工程与领域知识库
- 结语与致谢

结语与致谢

座右铭：“路漫漫其修远兮，吾将上下而求索”。

中国科学技术信息（情报）研究所对中国NLP早有贡献：MT与Dialog系统。

长期以来得益于 ISTIC 的支援与帮助。

大学同班同学多人在中信所任职。

频繁的学术交流：王惠临教授、刘耀博士。

得到赵俊杰副主任的邀请，今天获有机会来这里与大家交流。十分荣幸。

谢谢大家。

欢迎大家访问　北大计算语言学研究所www.icl.pku.edu.cn
北京大学软件与微电子学院语言信息工程系
www.ss.pku.edu.cn