

第五届中国英语教学国际研讨会暨第一届中国应用语言学大会主旨发言

中国国家重点基础研究计划 973课题 “ 文本内容理解的数据基础 ”

计算语言学之基础研究成果及其应用

Foundational Studies in Computational Linguistics: Achievements and Applications

俞士汶

北京大学信息科学技术学院

北京大学计算语言学研究所

2007年5月19日，北京

主要内容

- 计算语言学与中文信息处理概要
- 汉语信息处理的主攻方向
- 综合型语言知识库及其应用潜力
- 研究中的课题
- 致谢

主要内容

- 计算语言学与中文信息处理概要
- 汉语信息处理的主攻方向
- 综合型语言知识库及其应用潜力
- 研究中的课题
- 致谢

计算语言学与中文信息处理概要

计算语言学 Computational Linguistics

应用语言学的分支学科

自然语言处理 Natural Language processing

人工智能的分支学科

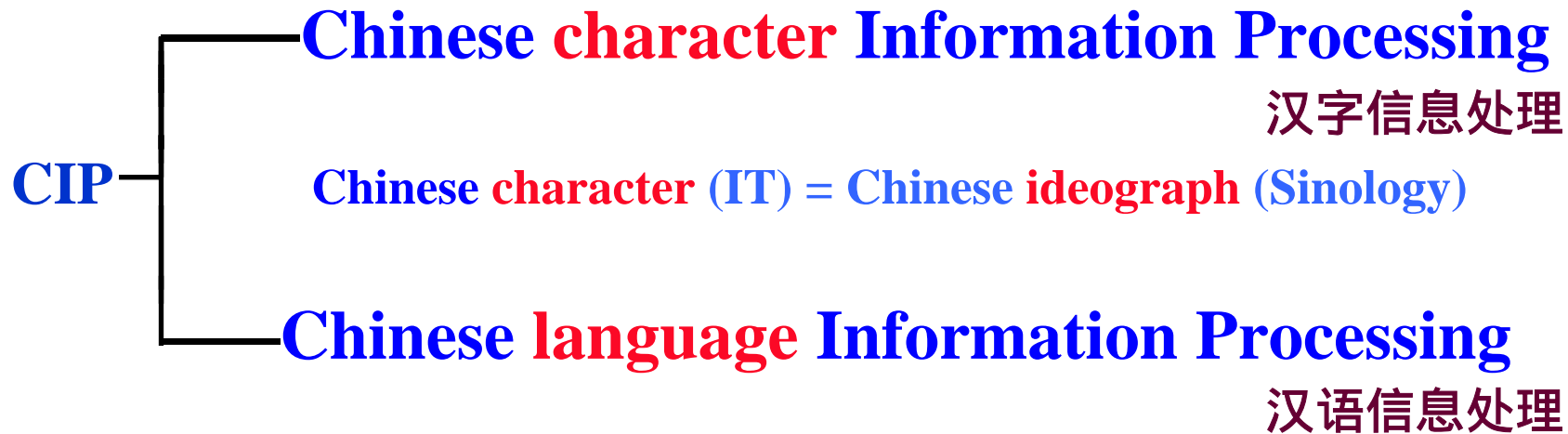
自然语言理解 Natural Language Understanding

自然语言处理的最高境界

ICL/PKU 以文科学科命名，设在理科的信息科学技术学院，正好显著地反映了文理交叉的特点。

计算语言学与中文信息处理概要

Chinese Information Processing (CIP) 中文信息处理



Language Information Processing 语言信息处理

(1) **NLP/ CIP**

(2) **Chinese-centered Multi-lingual Information Processing**

以汉语为核心的多语言信息处理

“ 汉语信息处理 ” 在研究什么？

实用系统：

- (1) 人工系统的自然语言界面（问答系统）
- (2) 机器翻译与机器辅助翻译
- (3) 信息检索、信息提取与搜索引擎
- (4) 文本管理（文本分类与聚类、文献摘要与述评、OCR后处理）
- (5) 词典计算机辅助编纂
- (6) 领域知识工程（术语提取、知识元数据库、百科全书编撰）
- (7) 语音接口技术（语音识别的后处理、语音合成的预处理）
- (8) 自然语言处理系统评测技术
- (9) 面向语言本体研究与语言教学的应用

……

汉语信息处理既立足于汉字信息处理，
又区别于汉字信息处理。

处理对象不再是单个的汉字或字符串，
而是语言学的单位：词、短语、句子乃至篇章、文档集合。
两者之间也有联系：拼音汉字转换、简繁转换、OCR 后处理、
文献检索、语音识别与合成等等。

*关于“汉语信息处理”的基本认识

自然语言处理是数值型计算机在非数值领域最早的应用。语言学对计算机技术的发展有重要贡献。但在“自然语言理解”的层次上，至今没有突破。

- (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。

不仅仅是技术问题。脑科学、认知科学



在技术层面上，汉语信息处理又有特殊的课题。

主要内容

- 计算语言学与中文信息处理概要
- 汉语信息处理的主攻方向
- 综合型语言知识库及其应用潜力
- 研究中的课题
- 致谢

自然语言（汉语）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。”

Wife:’ ’

丈夫：“什么呀，才一年就坏了。”

Husband:’ ’

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（汉语）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

汉语信息处理的主攻方向

自然语言理解研究特别困难，目前难以突破。
退而求其次：自然语言处理（汉语信息处理）。
计算机处理自然语言的第一个障碍是**歧义**问题。
人能够利用语言知识、语境信息、
背景知识消解歧义。
计算机进行机械式的分析，
面临的困难要大得多。
以下介绍信息处理所遇到的
汉语歧义的类型
和求解之道。



汉语信息处理主攻方向——歧义消解

词语切分问题：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/ ？

同形词辨析：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

汉语信息处理主攻方向——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

词义辨析：讲真话 / 讲卫生

短语结构的歧义： $m + q + n + \text{“的”} + n$

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]

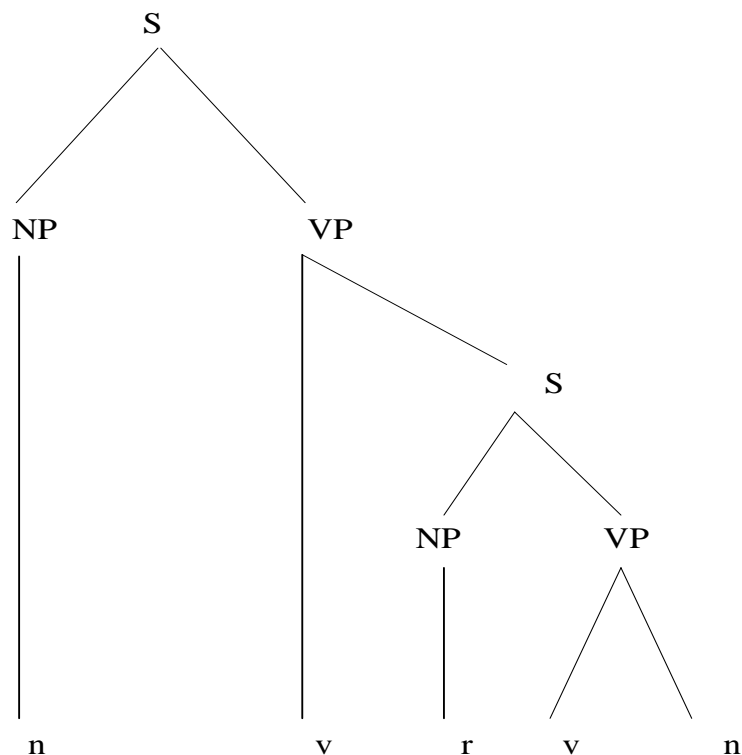
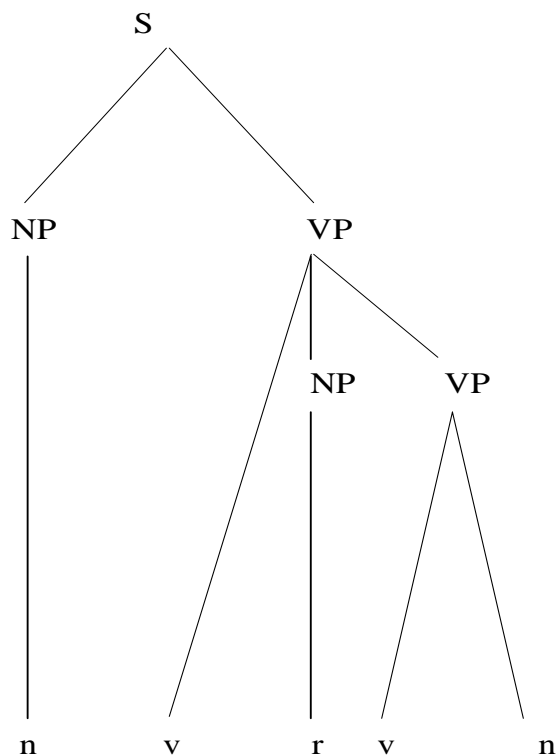
三位大学的老师——[三/m 位/q [大学/n 的/u 老师/n]]

句子结构的歧义

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

中文中常有长句子，一逗到底。例：

“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。”

形式上的一句话包含100多个汉字。

除第一个分句外，后面的分句都没有主语。

其他：长句与句号、逗号

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

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。

汉语信息处理还有其他障碍

隐喻

幽默

夸张

双关

影射

.....

2006年11月

“中国中文信息学会二十五周年学术会议”

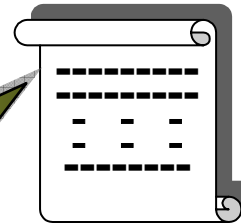
俞士汶报告：

“文学语言与自然语言理解研究”



关于一般的自然语言处理系统

NLP是世界性难题.....



自然语言处理系统

应用程序

语言知识库

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

- 语言单位不清晰：语素、词、短语、句子
- 词缺乏形态变化：词类多功能与词的兼类
- 虚词：词形与实词无区别、隐现不定
- 句子与句法结构：
 - 嵌套不需要附加成分
 - 主谓结构作谓语
 - 不完整，缺省主语
- 时态、语态和语气的表现缺乏形式标记
- 形式和意义之间的对应关系复杂
(吃香蕉、吃筷子、吃食堂)

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。

(1) 《现代汉语语法信息词典》 GKB

“《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）

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
词典采用数据库文件格式。目前已扩充到8万词语。

ERSHI SHIJI
WOGUO ZHONGDA
GONGCHENG JISHU
CHENGJIU



20世纪
我国重大工程技术成就

中国工程院 编
常平 主编

暨南大学出版社
Jinan University Press

序言：工程技术百年颂 宋健 (1)

成就说明及大事记

一、两弹一星 (12)

二、汉字信息处理与印刷革命 (29)

三、石油 (42)

四、农作物增产技术 (51)

五、传染病防治 (65)

六、电气化 (76)

七、大江大河治理和开发 (83)

八、铁路 (96)

九、船舶 (109)

十、钢铁 (117)

十一、计划生育 (127)

基于汉字属性的汉字信息处理软件的研发。

1995年首次建立了一个含14万个记录的现代汉语语素数据库，实现了汉字构词知识的数据化。而后又开展了基于语素数据库的“汉语词的构造”研究，进一步扩大了数据库。

《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结合的准则收录了7.3万余词语。依据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。

语自动分词研究

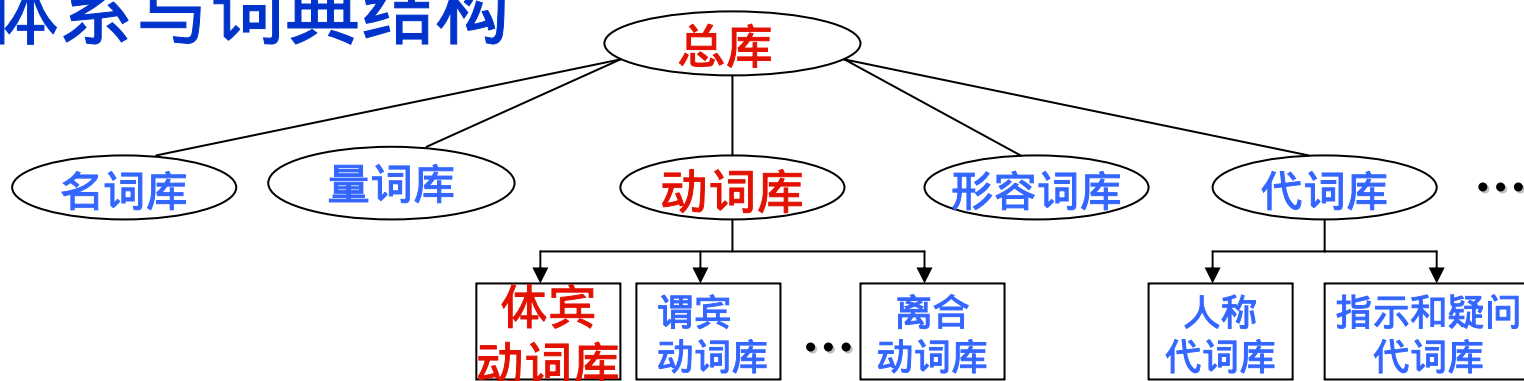
由于现代汉语的词与词之间没有像西方拼音文字那样留下空格，又没有统一的对“词”的定义，因此制定汉语

现代汉语语法信息词典详解



现代汉语语法信息词典 GKB

词类体系与词典结构



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。

采用1980年代当时先进的数据库技术，编制机器词典。
技术路线、总体设计、规格说明已经受了历史的考验。

《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

词语	词类	同形	拼音	注
挨	v	A	ai 1	触, 碰, 靠近		
挨	v	B	ai 2	遭受, 忍受		
安装	v		an1zhuang1			
保管	v	1	bao3guan3	保存		
保管	v	2	bao3guan3	担保		
抄	v	A	chao1	照原稿写		
抄	v	B	chao1	走近道		
地道	a		di 4dao5	正宗		
地道	n		di 4dao4			
叫	v	A1	j i a o 4	人或动物发出的较大声音		
叫	v	A2	j i a o 4	呼唤, 招呼; 雇		
叫	v	A3	j i a o 4	称为		
叫	v	B	j i a o 4	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“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

(2) 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对其“主体”语义的选择特性。

“**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		

(3) 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）

在海量信息的环境中，要提高（跨语言）信息检索、信息提取的查准率，区分这两个概念是必要的。

中英文概念词典（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。

概念由同义词集合(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、...} 就是一个概念。

“先生”表示的概念之一

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

“先生”表示的另外两个概念

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)

Generator and Browser [CCD]

File History Noun Verb Adjective Adverb HypoTree Node GetInfo DoMyJob Help

Total 6084 Present 5680 POSs Noun Verb ADJ ADV

- 继承人 后任 后尘 接班人 后嗣 接任者 继任者
- 表侄女 表侄子
- 兄弟姐妹 同胞
 - 半血亲者 同母异父者 同父异母者
 - 四胞胎
 - 五胞胎
 - 三胞胎
 - 双胞胎 孪生子
- 配偶 伴侣 侣伴 夫妻 比翼鸟 终身伴侣 佳偶 佳侣 结发夫妻 那口子
 - 重婚者 二婚者
 - 丈夫 先生 夫君 夫婿 爱人 老公 郎君 驸马 驸马爷
 - 新婚男子 新郎
 - 绿帽子男人 乌龟 绿头龟
 - 居家男人 已婚男人 有妻室者
 - 家庭主夫 家庭主男 家庭煮夫
 - 新婚者 新婚夫妇 新人 蜜月新人
 - 新妇 新娘 新嫁娘 新媳妇儿
 - 新郎 马夫 新郎官
 - 重婚者 多配偶者 一妻多夫者 一夫多妻者
 - 太太 妻子 老婆 内助 内室 妇人 妻室 婆姨 爱妻 内

[husband][hubby][married_man]

丈夫 先生 夫君 夫婿 爱人 老公 郎君 驸马 驸马爷

a married man; a woman's partner in marriage

已婚男子; 婚姻中女性一方的伴侣

[]

丈夫和妻子应平等相待

Enter search word and press return.

New_BrotherNode
New_Child_Node
Del_CurNode_One
Del_CurNode_All
CutOut_CurNodes
Copy_CurNodes
PasteAsBrothers
PasteAsChildren

老婆 爱人

(3) 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）

(4) 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/v 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

《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。



(5) 汉英双语平行语料库 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>埃内斯特·海明威</AUTHOR>
<CH_TITLE>谁钟为谁而鸣</CH_TITLE>
</TEXT_HEAD>
<TEXT_BODY>
<p id="1">
<a id="1" no="1">
<cs id="1">
<CH_TITLE>谁钟为谁而鸣</CH_TITLE>
</cs>
</a>
</p>
<p id="2">
<a id="2" no="1">
<cs 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法炸毁一座桥梁。</cs>
</a>
<a id="2" no="1">
<cs id="2">一位中立的左向与女塞万莫将他带到当地游击队的营地。</cs>
</a>
<a id="4" no="1">
<cs 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。</cs>
</a>
<a id="5" no="2">
<cs 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。</cs>
<cs id="5">皮拉尔这个女人相貌丑陋，却比她丈夫勇敢得多。</cs>
</a>
<a id="6" no="1">
<cs 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。</cs>
</a>
<a id="7" no="1">
<cs id="7">乔丹获悉，她最近刚被游击队从法西斯分子手中救出，曾被遭法西斯分子蹂躏，身体和精神仍处于恢复之中。</cs>
</a>
</p>
<p id="9">
<a id="9" no="1">
```

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


(6) 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	彼得·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 Houston	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 gyms	体育场、馆		
9	basketball court	篮球场	physical culture	体育	stadiums and gyms	体育场、馆		
10	bleachers	露天看台	physical culture	体育	stadiums and gyms	体育场、馆		
11	box	专席	physical culture	体育	stadiums and gyms	体育场、馆		
12	bulletin board	公告牌	physical culture	体育	stadiums and gyms	体育场、馆		
13	centre pole	中心旗杆	physical culture	体育	stadiums and gyms	体育场、馆		
14	competition arena; ground; court	比赛场地	physical culture	体育	stadiums and gyms	体育场、馆		
15	covered gymnasium	室内运动场	physical culture	体育	stadiums and gyms	体育场、馆		
16	electrically controlled movable stand	电动式活动看台	physical culture	体育	stadiums and gyms	体育场、馆		
17	air (/aviation/flying) sports	航空运动	physical culture	体育	physical culture	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	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	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	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	体育运动	kinds of sport	体育运动种类
22	bob sleighing	滑雪	physical culture	体育	physical culture	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	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	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	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	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	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	体育运动	sports wear	运动服装
29	blazer	艳色运动上衣	physical culture	体育	physical culture	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	体育运动	sports wear	运动服装
31	girl's slacks	女运动裤	physical culture	体育	physical culture	体育运动	sports wear	运动服装
32	gym outfit	成套运动服	physical culture	体育	physical culture	体育运动	sports wear	运动服装
33	jockstrap	松紧运动内裤	physical culture	体育	physical culture	体育运动	sports wear	运动服装
34	moccasin	软底运动鞋	physical culture	体育	physical culture	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，

主要内容

- 计算语言学与中文信息处理概要
- 汉语信息处理的主攻方向
- 综合型语言知识库及其应用潜力
- 研究中的课题
- 致谢

研究中的课题

国内的热点课题（难免挂一漏万）

- (1) 基础与应用：应用系统占主体
- (2) 知识与统计：统计方法是主流
- (3) 以汉语为核心的多语言信息处理
- (4) 认知机制与多元感知信息的融合

.....

我的同事和我本人正在研究的课题

- (1) 综合型语言知识库系统的建设
- (2) 文本内容理解的进展
- (3) 信息处理应用系统的开发
- (4) 语言教学中应用的探讨

.....

综合型语言知识库系统的建设

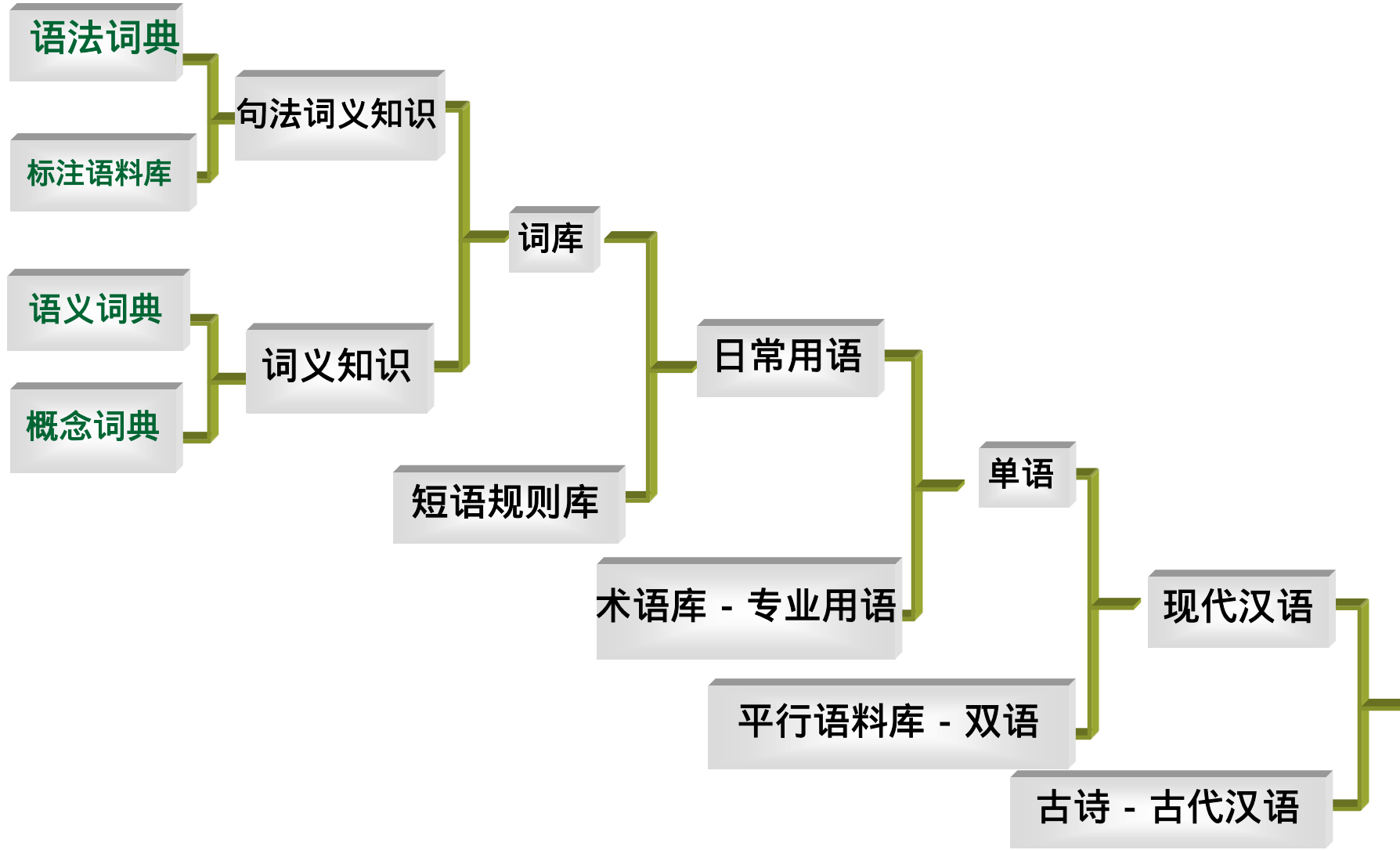
现状：

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。

北大现有语言数据资源之内在联系



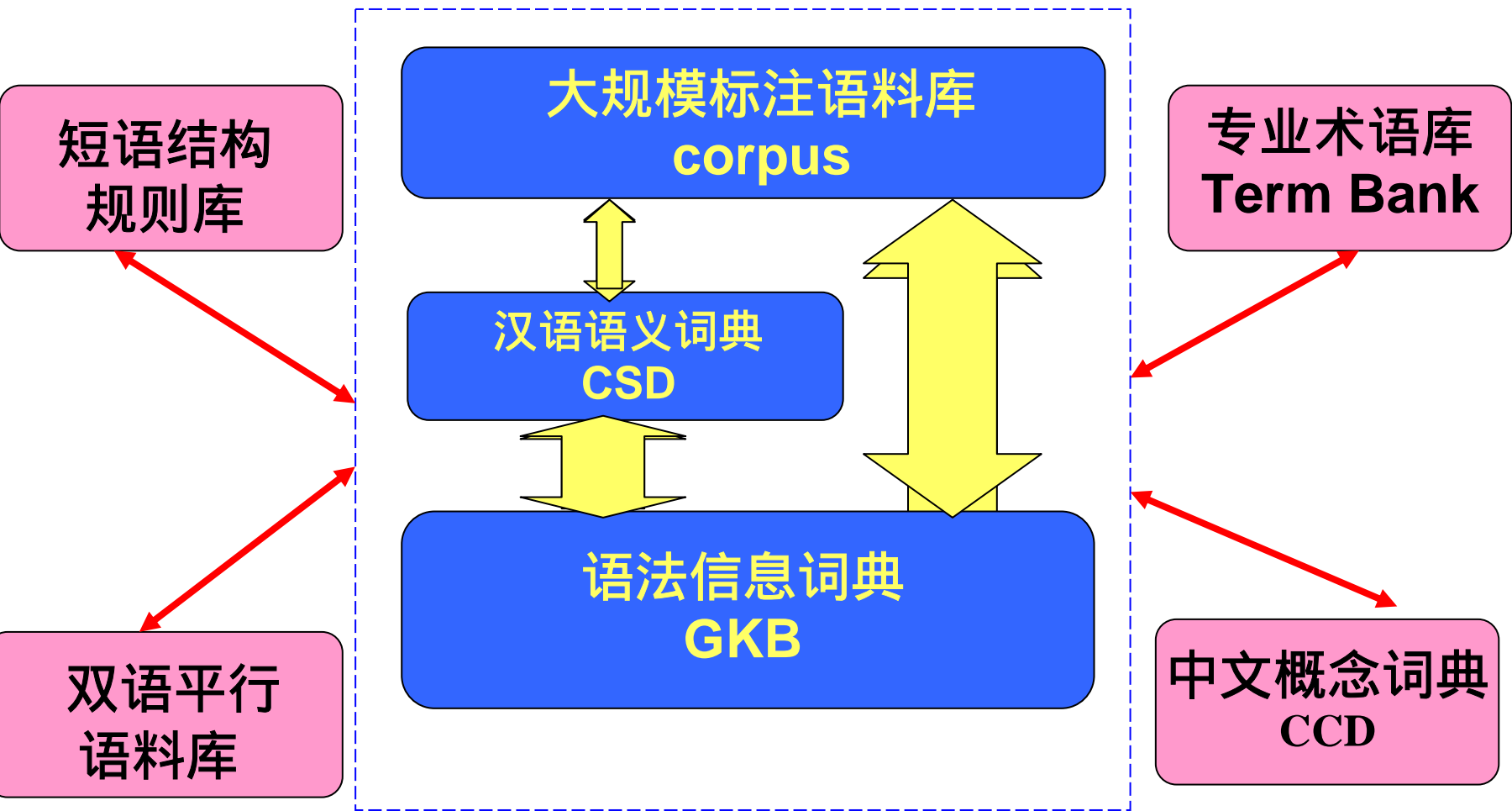
语言数据资源的集成方案

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。

各种语言数据资源之相互支撑



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。

基本标注标注语料库之样例

原始语料

例句1：二是此类编著内容抄自别人的多，多到被人告到了法庭；

例句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

例句1：二/m 是/v 此类/r 编著/v 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

例句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

请注意：例句1和例句2的“抄”只标注为动词 v ，
未区分“同形”（粗粒度的义项），
与《语法信息词典》之间存在“缝隙（gap）”，
不知道例句中的“抄”对应词典中的哪个记录。

《语法信息词典》和语料库相互参照的准确定位

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

此类/r 编辑/v 内容/n 是/v 抄/v!A 自/p 别人/r 的/u

炮兵/n 学院/n 原来/n 围墙/n 残缺/v ,/w 周边/n 群众/n 进城/v ,/w 习惯/v 抄/v!B 近道/n 。/w

语料中未标“同形”，不能找到正确的信息

例：GKB 和 同形标注语料库 之 集成

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

此类/r 编著/v 内容/n 是/v 抄/v!A 自/p 别人/r 的/u

炮兵/n 学院/n 凉来/d 围墙/n 残缺/v ,/w 周边/n 群众/n 进城/v ,/w 习惯/v 抄/v!B 近道/n 。/w

以“词语”+“同形”+“同形”为轴，进行集成

以粗粒度词义为主轴的集成方案

加注同形的语料

抄/v!A

抄/v!B

词-词性-同形

现代汉语语法信息词典

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

集成原始语料库和字典 ——以“字”为主轴

原始语料

此类编著内容是抄自别人的。

炮兵学院原来围墙残缺，周边群众习惯抄近道。

字

字典

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

切分语料

此类/ 编著/ 内容
/ 是/ 抄/ 自/ 别
人/ 的。 /

炮兵/ 学院/ 原来
/ 围墙/ 残缺/ , /
周边/ 群众/ 习
惯/ 抄/ 近道/ 。 /

词

词典

词语	拼音	词频	释义
此类	ci 3l ei 4		
编著	bi an3zhu4		
内容	nei 4rong2		
是	shi 4		
抄	chao1		
别人	bi e2ren2		
.....		

集成基本标注语料库和划分词类的词典 ——以“词语”+“词性”为轴

基本标注语料

此类/r 编著/n 内
容/n 是/v 抄/v
自/p 别人/r 的
/u 炮兵/n 学院
/n 原来/d 围墙
/n 残缺/v , /w
周边/n 群众/n
习惯/v 抄/v 近
道/n

词—词性

划分词类的词典

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

加注同形的语料

抄/v!A

抄/v!B

词-词性-同形

现代汉语语法信息词典

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

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		

粗粒度词义标注（同形）的实例

- 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（同形）的实例

- 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”）

以细粒度词义为主轴的集成方案

加注义项的语料

冲/v!A-1

冲/v!A-2

冲/v!A-3

冲/v!B

词—词性—同形—义项

现代汉语语义词典

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1999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<语法信息词典>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

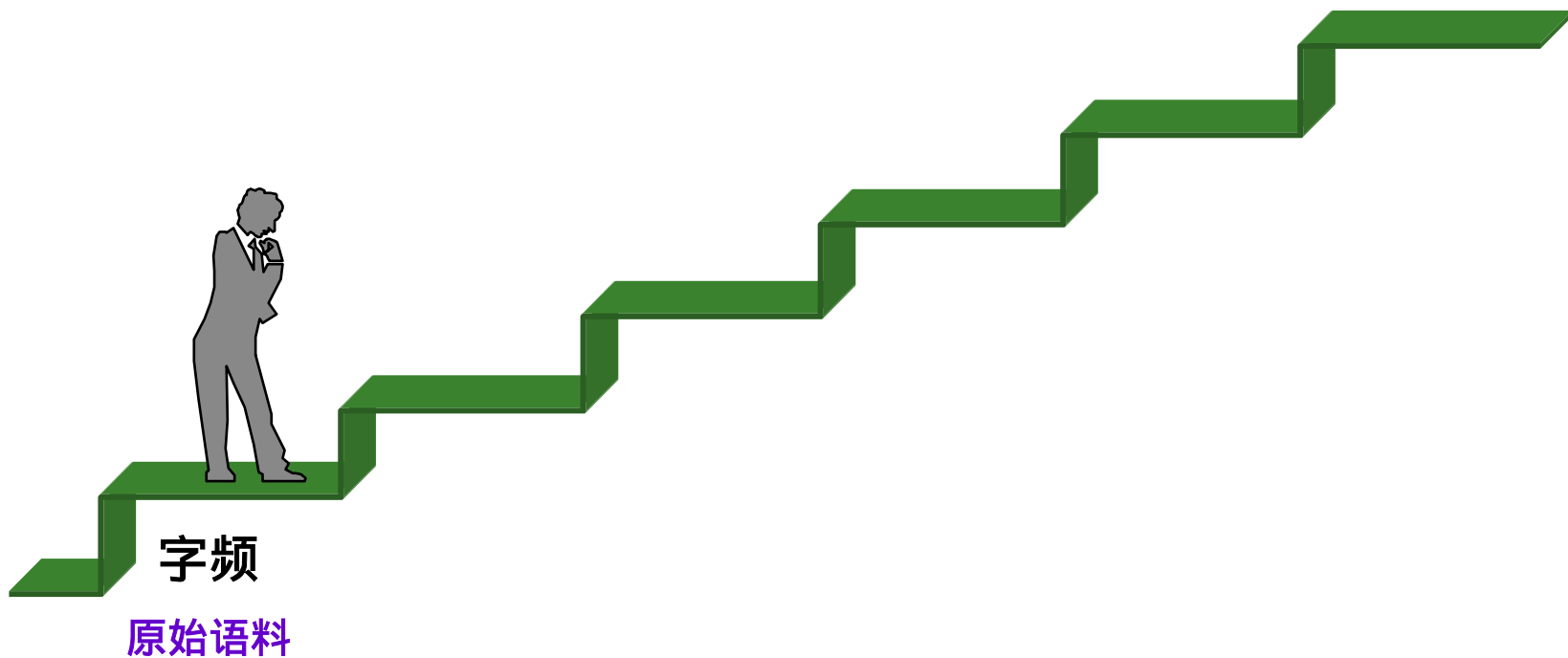
基于语言数据资源的知识挖掘研究及其成果

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

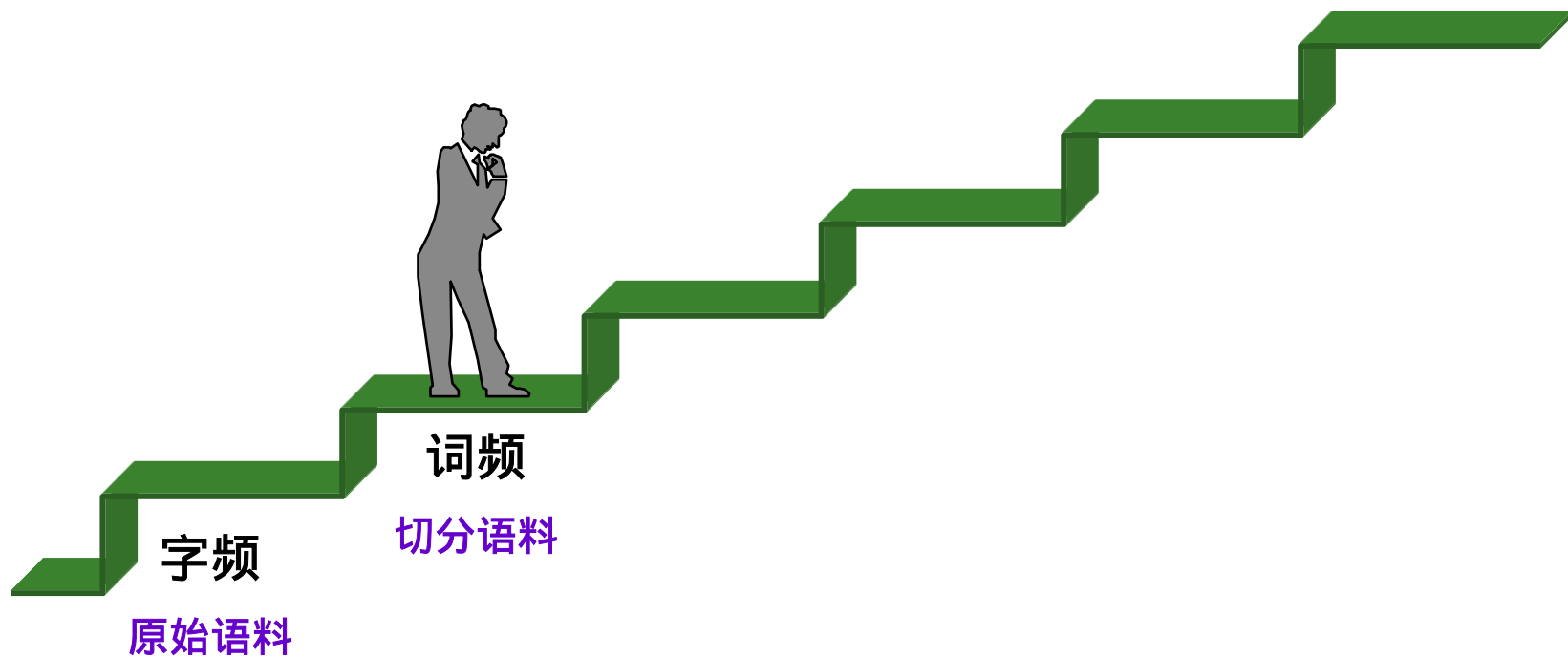
汉语词汇计量研究之发展历程

(伴随着语料库加工的逐步深入)



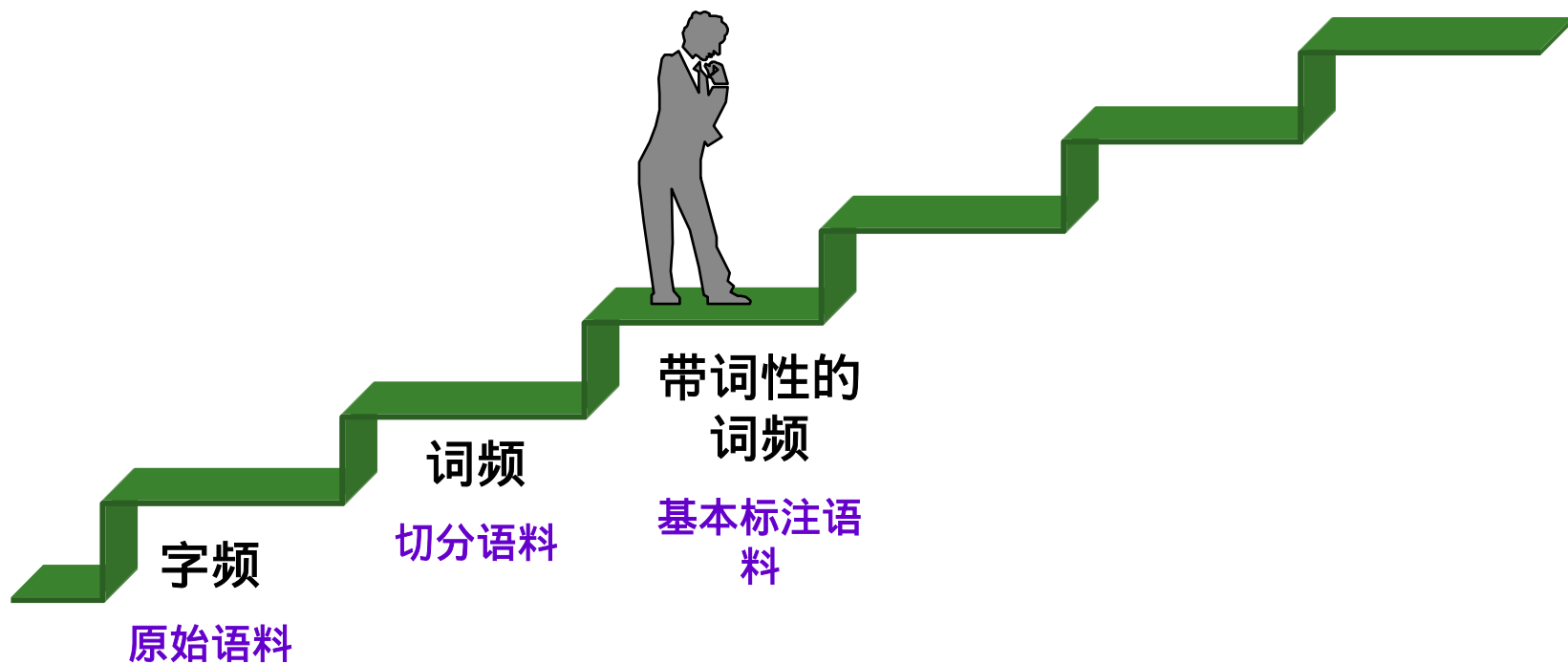
汉语词汇计量研究之发展历程

(伴随着语料库加工的逐步深入)



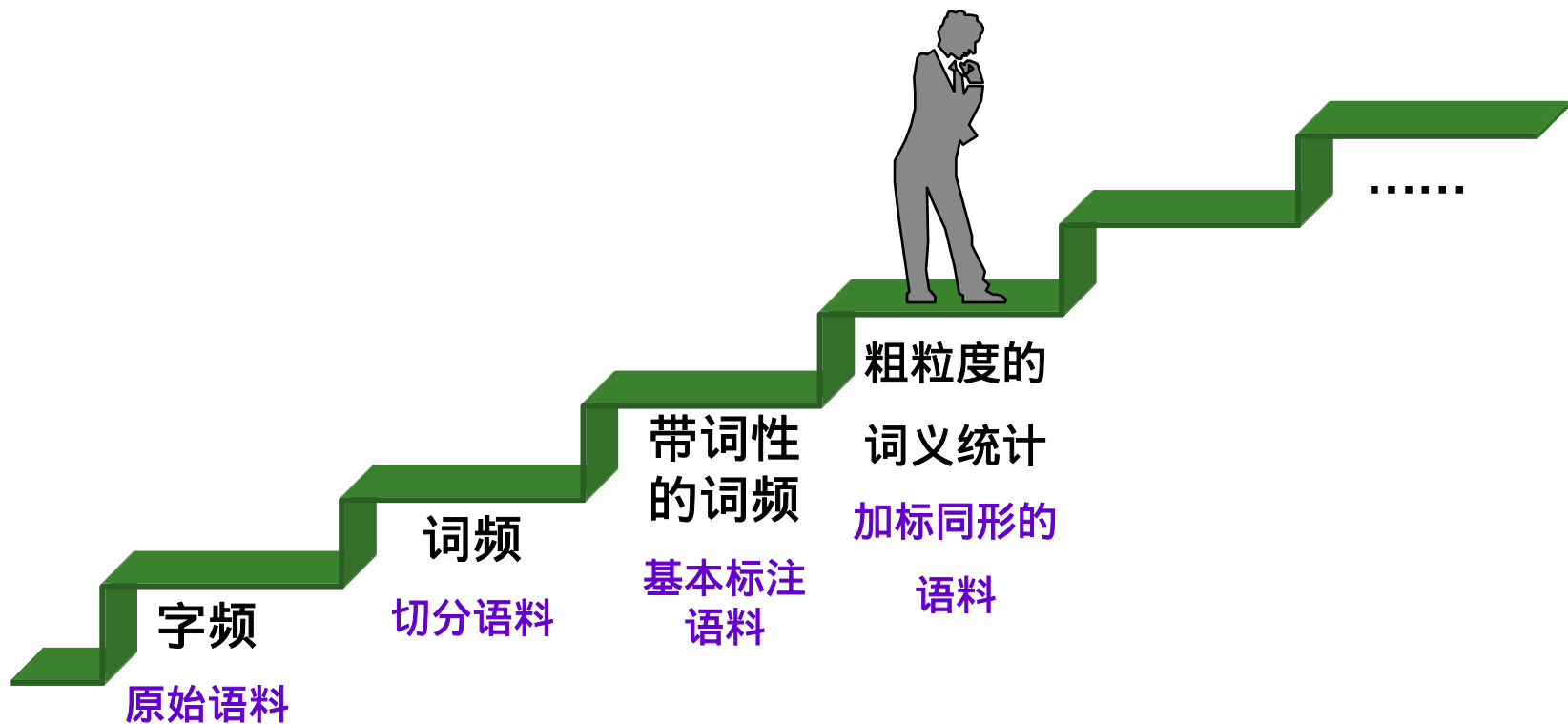
汉语词汇计量研究之发展历程

(伴随着语料库加工的逐步深入)



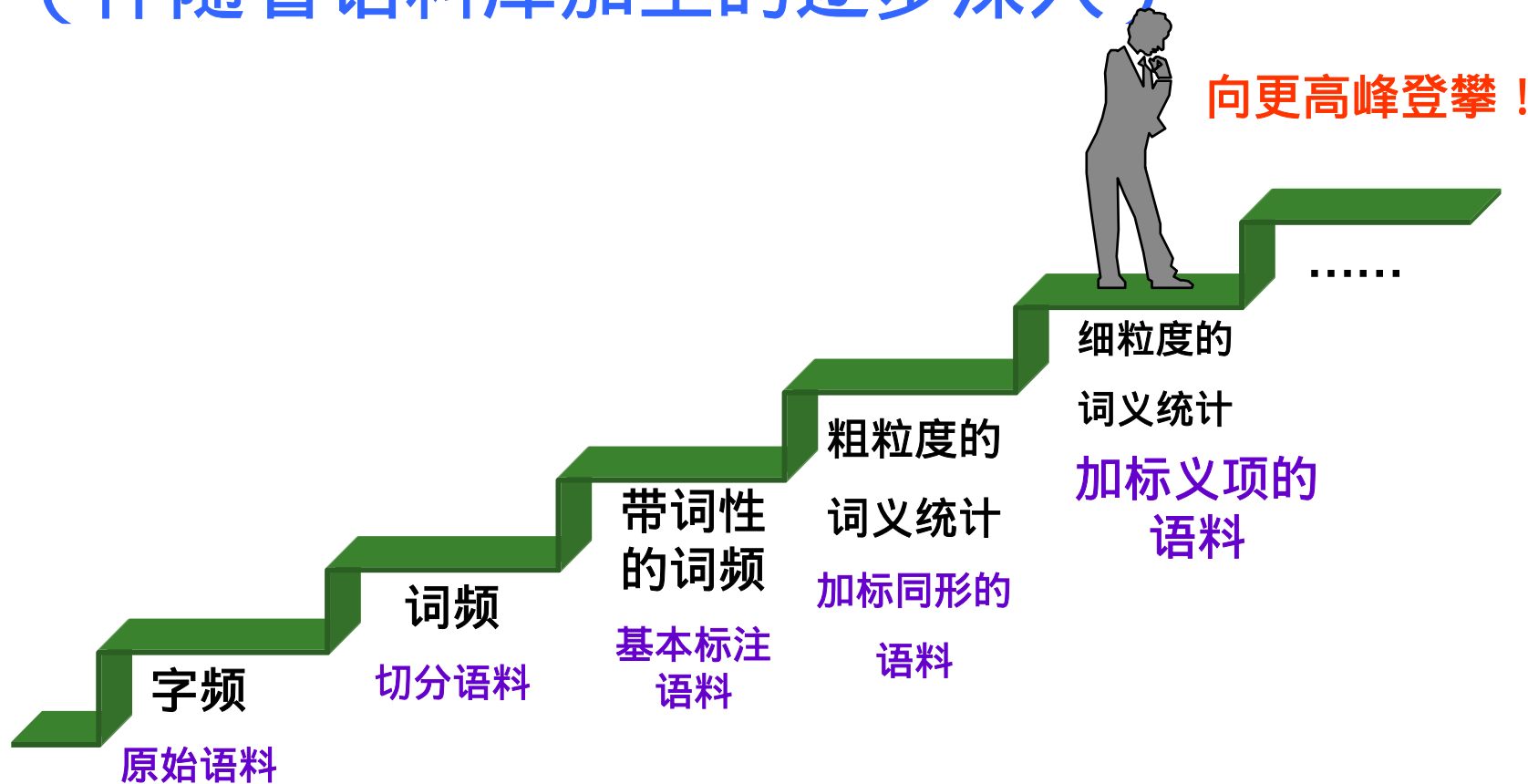
汉语词汇计量研究之发展历程

(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。

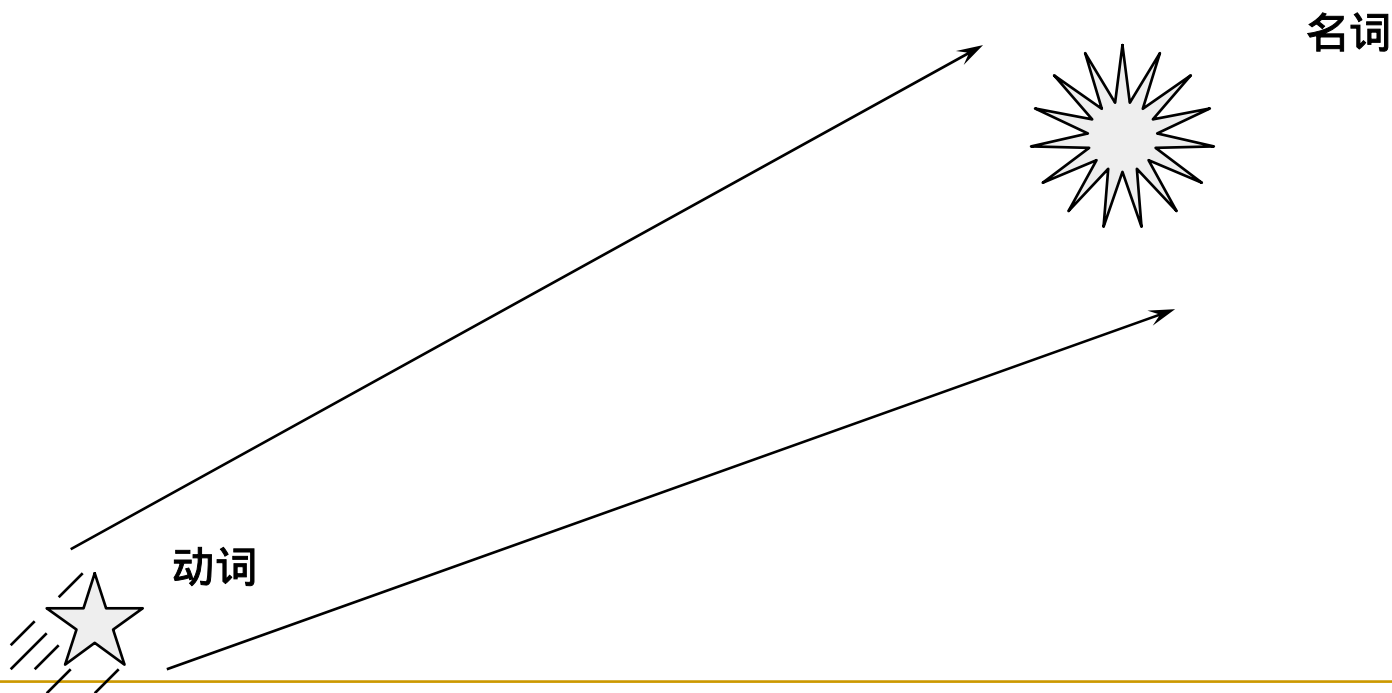
胡明扬先生的“漂移说”。

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)

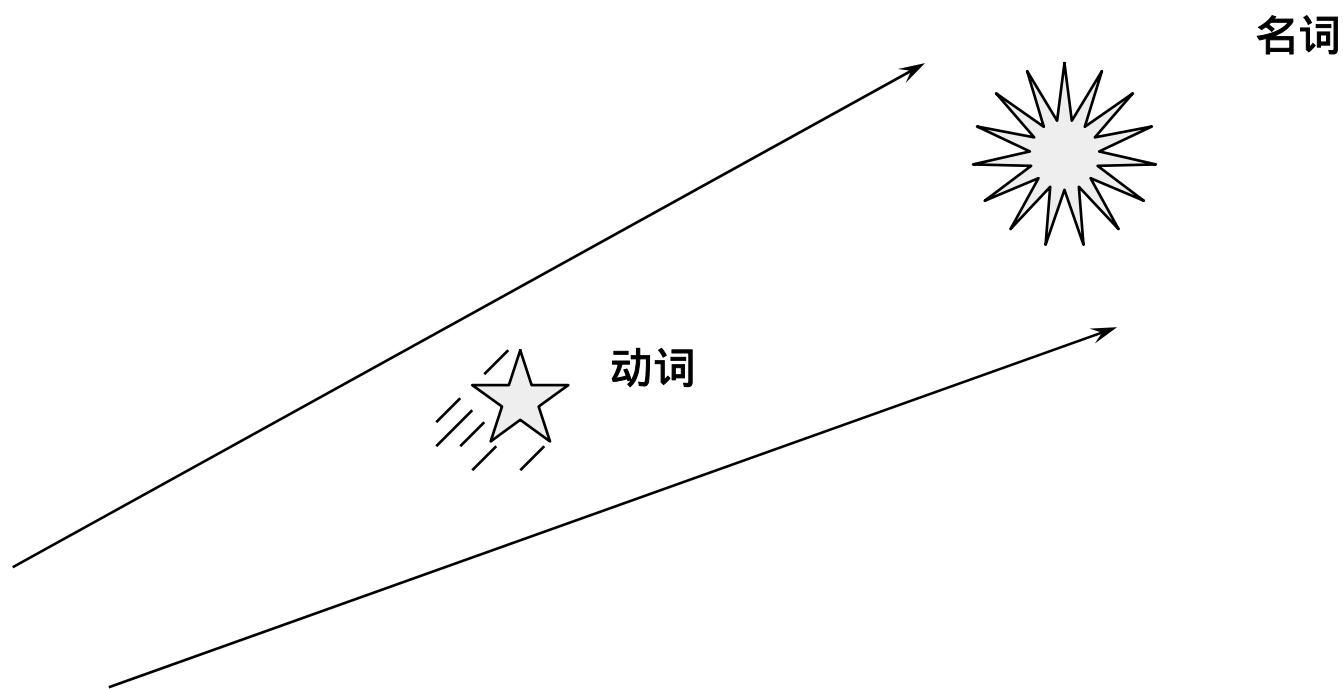
若 $P(v) \gg P(vn)$ ，则该动词向名词彼岸的漂移刚刚开始。

- 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： $P(v)=0.68$, $P(vn)=0.32$



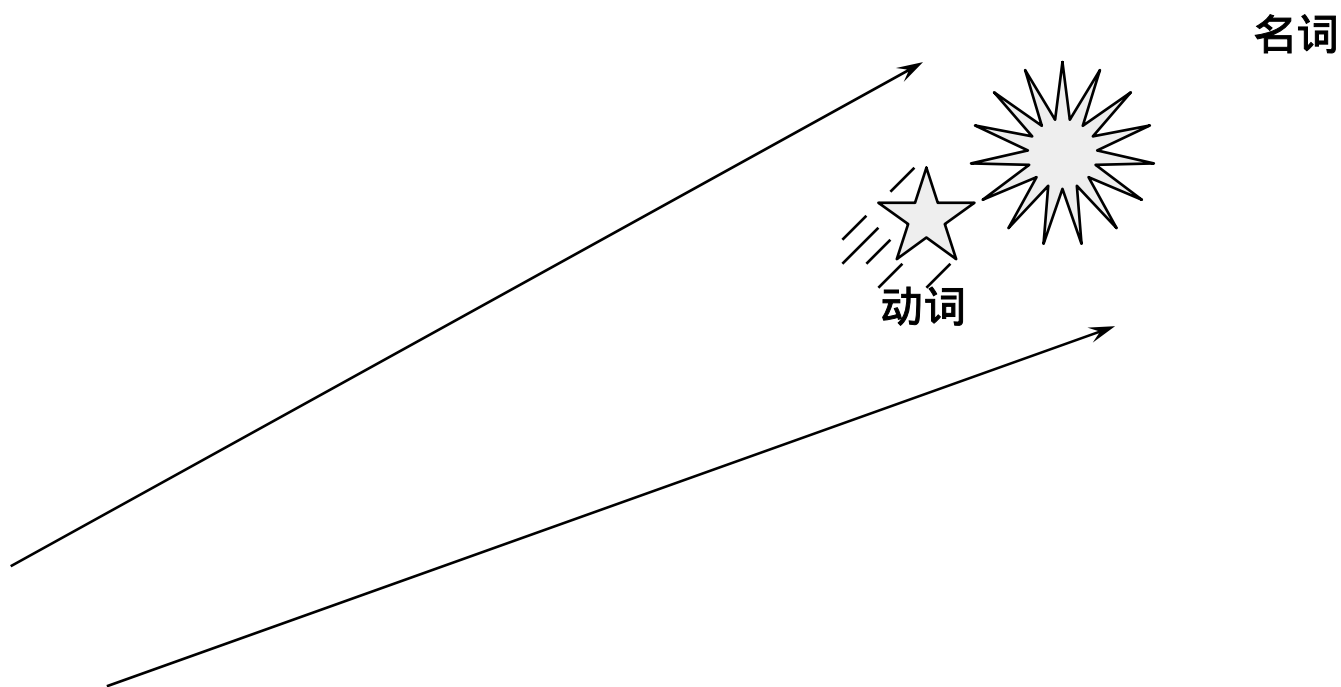
若 $P(v) > P(vn)$ ，则该动词处于向名词漂移的过程中。

- 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 $P(v) \ll P(vn)$ ，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

概率语法信息词典的研制

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

文本内容理解的进展

(1) 综合型语言知识库系统的研制

(2) 语言知识库的补足

 广义虚词知识库

 成语知识库

 缩略语知识库

 隐喻知识库

(3) 对隐喻等语言现象的思考与处理

(4) 认知模型、算法、机器学习的探讨

(3) 对隐喻等语言现象的思考与处理

隐喻(Metaphor): 修辞 认知

隐喻的层级划分：词汇级、语句级和篇章级。

词汇级——“山顶”、“桌子腿”、“无底洞”、“瓜分”、“露马脚”、“指桑骂槐”
(求解依赖于词汇知识库，没有特殊之处)

语句级——“知识的海洋”、“幻想是诗人的翅膀”、“激情燃烧的岁月”
铁榔头又立了新功

The iron hammer sets up the outstanding service again today.
iron fist

(求解要有新招，如：隐喻知识库，逻辑推理，机器学习)

篇章级——《春怨》(唐诗)

打起黄莺儿，莫教枝上啼。
啼时惊妾梦，不得到辽西。

(年轻妻子思念在远方浴血奋战的亲人，魂牵梦绕。机器理解尚在探索中)

(3) 对隐喻等语言现象的思考与处理

2003年9月——2006年6月

博士生王治敏完成博士论文

“汉语名词短语隐喻识别研究”

同期姜柄圭博士论文也分析了科技专著中的隐喻

2006年1月——2007年6月

博士后曲维光前后衔接，深入研究，今年申请

“汉语隐喻理解关键技术研究”自然科学基金课题

2006年9月——2010年6月

博士生贾玉祥拓展研究范围：语句层次和篇章层次的隐喻形态，动词隐喻研究，集成名词隐喻研究成果。同时探索在（深度）信息检索等相关领域的应用。

典型应用 系统开发



应用程序1

.....

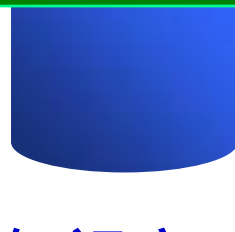


应用程序2

平台 / API



.....



语言知识库1

语言知识库2

语言知识库3

语言知识库4

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

信息处理应用系统开发

- (1) 面向中文专著的机器辅助翻译系统
- (2) 基于语料库的双语词典辅助编纂平台
- (3) 百科知识管理与服务平台
- (4) 命名实体的网页提取与服务系统
- (5) 海量新闻语料的主题标引系统
- (6) 机器翻译及自然语言处理系统的评价

语言教学中应用的探讨

- (1) 教学模式的发展
- (2) 汉语国际推广之需求
- (3) 面向人与面向机器的语言教学
- (4) 教学资源的整备

主要内容

- 计算语言学与中文信息处理概要
- 汉语信息处理的主攻方向
- 综合型语言知识库及其应用潜力
- 研究中的课题
- 致谢

致 谢

衷心感谢第一届中国应用语言学大会的邀请。

让我有机会在这里和大家交流。

衷心感谢北大英语系高一虹教授的推荐。

感谢王雷老师把我的文章译成英文。

衷心感谢在座各位，敬请指教。

欢迎访问

北京大学计算语言学研究所 www.icl.pku.edu.cn

北大软微学院语言信息工程系 www.ss.pku.edu.cn