中文信息处理与汉语研究
—— 现状和发展

詹卫东

北京大学中文系
北京大学汉语语言学研究中心
北京，100871
zwd@pku.edu.cn
http://ccl.pku.edu.cn/doubtfire/

提 纲

1) 中文信息处理研究的格局
2) 中文信息处理的现状和发展趋势
3) 语言知识资源的建设
4) 面向中文信息处理的汉语研究

一 中文信息处理研究的格局

- 信息的两个层次：
  符号层 —— 中文 / 汉语 / 汉字
  内容层 —— 符号所承载的意义

（信号 vs. 信息）

- 中文信息处理的两个层次：
  字符处理（输入、存储、输出等）
  内容处理（词语切分，词性标注，结构分析，意义理解，推理，翻译……等等）

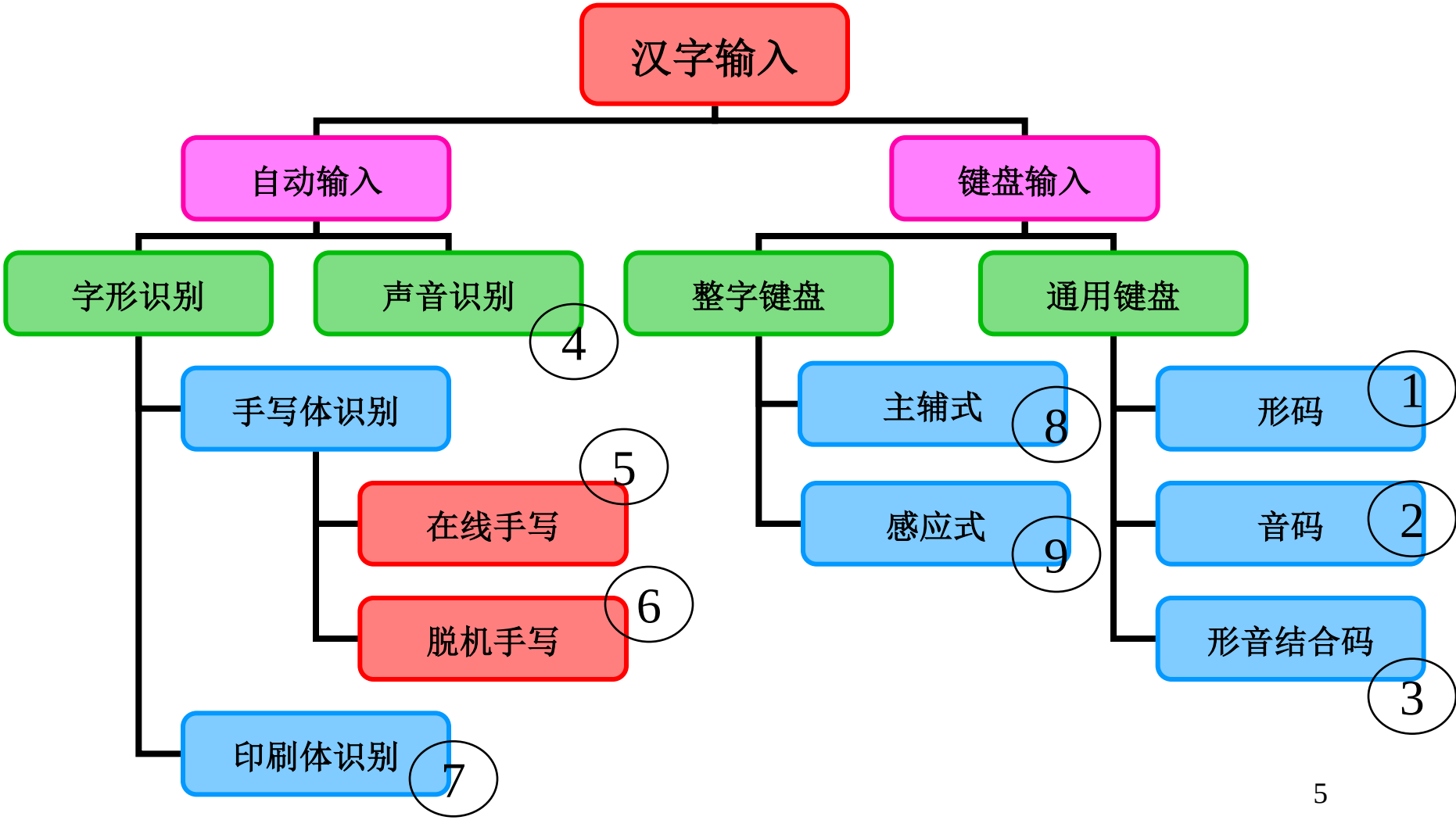
符号层的信息处理

- 拼音文字：小字符集 —— 比较容易
- 非拼音文字：大字符集 —— 难度很大

- 拉丁字母只有26个符号
- 斯拉夫字母只有33个符号
- 阿尔明尼亚字母只有38个符号
- 泰米尔字母只有36个符号
- 缅甸字母只有52个符号
- 泰文字母只有44个符号
- 老挝字母只有27个符号
- 藏文字母只有35个符号
- 韩文字母只有24个符号
- 日文假名只有48个符号

- 汉字是一个大字符集
  - 《说文解字》（东汉）：9353字
  - 《玉篇》（南朝）收录16,917字
  - 《广韵》（宋代）收字26,194字
  - 《字汇》（明朝）收录33,197字
  - 《康熙字典》（清朝）收录47,043字
  - 《汉语大字典》（1992年）5.6万
  - 《中华字海》（ 1994年） 8.6万

符号层的信息处理


汉字输入
自动输入
键盘输入
字形识别
声音识别
整字键盘
通用键盘
4
手写体识别
主辅式
8
形码
1
5
在线手写
感应式
9
音码
2
6
脱机手写
形音结合码
3
印刷体识别
7

内容层的信息处理

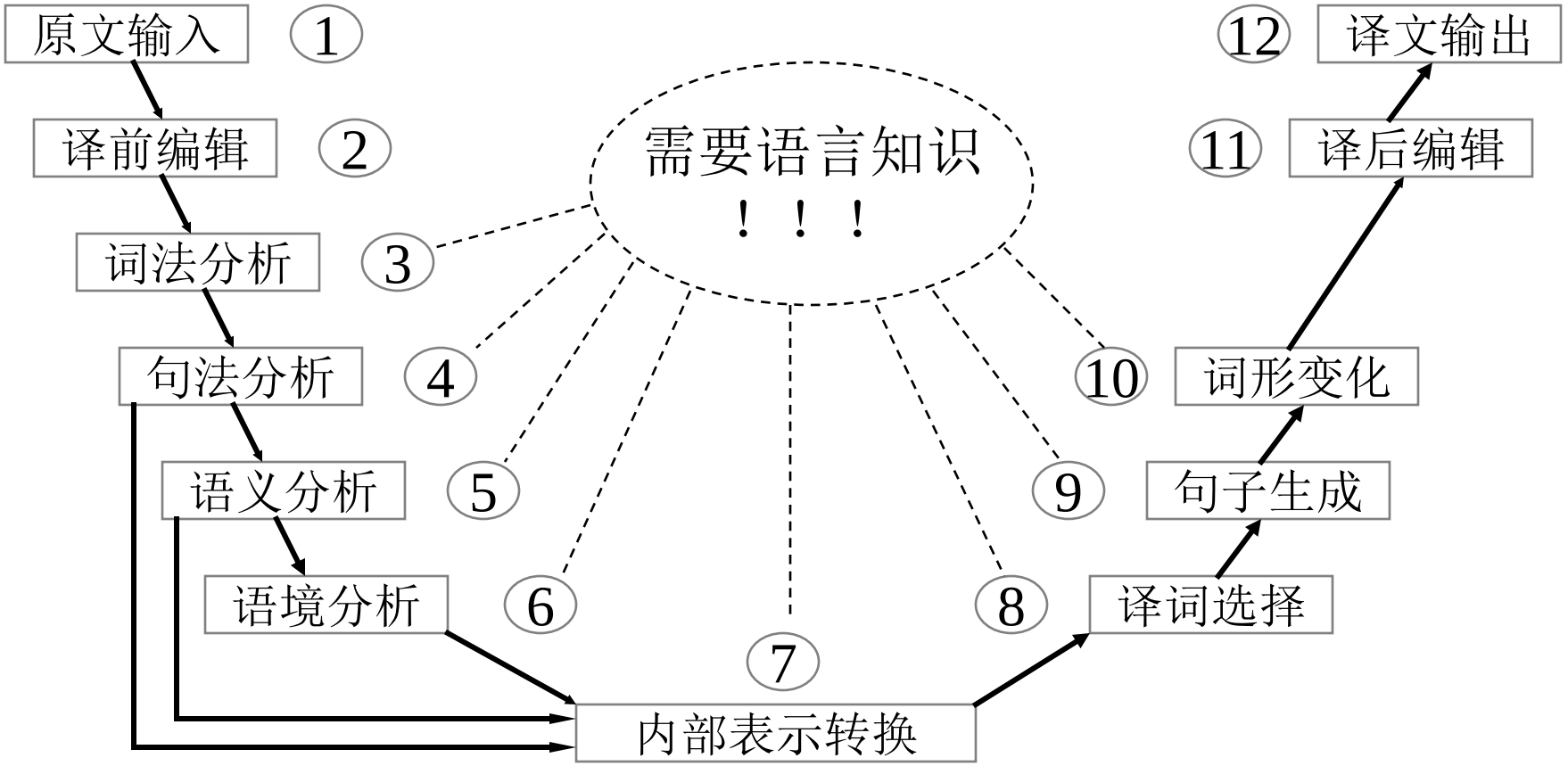
- 形态丰富的语言（inflecting language）：处理难
- 形态不丰富的语言（analytic language）：处理更难

汉语	英语
老师都来了	All professors came here.
张老师都来了	Even Professor Zhang came here.
编辑工作很难	Editing is very difficult.
如何当好编辑	How to become a good editor

内容层的信息处理

原文

译文



机器翻译全过程

内容层处理对符号层处理的反作用

拼音串（无声调）	xue xi dian nao ji shu

内容层处理对符号层处理的反作用

拼音串（无声调）	xue xi dian nao ji shu	
候选字串	雪 系 点 脑 机 树	共有 14×98×41×15 ×167×68＝95.8 亿种可能性
	学 洗 电 闹 给 述	
	学 西 颠 挠 记 书	
	……	

内容层处理对符号层处理的反作用

拼音串（无声调）	xue xi dian nao ji shu	
候选字串	雪 系 点 脑 机 树	共有 14×98×41×15 ×167×68＝95.8 亿种可能性
	学 洗 电 闹 给 述	
	学 西 颠 挠 记 书	
	……	
候选词串	学习 电脑 级数	共有2×1×7＝14 种可能性
	血洗 电脑 奇数	
	血洗 电脑 基数	
	……	

内容层处理对符号层处理的反作用

拼音串（无声调）	xue xi dian nao ji shu	
候选字串	雪 系 点 脑 机 树	共有 14×98×41×15 ×167×68＝95.8 亿种可能性
	学 洗 电 闹 给 述	
	学 西 颠 挠 记 书	
	……	
候选词串	学习 电脑 级数	共有2×1×7＝14 种可能性
	血洗 电脑 奇数	
	血洗 电脑 基数	
	……	
正确文字串	学习电脑技术	

二 中文信息处理的现状和发展趋势

- 现状
  符号层的处理成果已经得到广泛应用；
  中文输入/字库/字处理软件/排版/……

  内容层的处理目前在词语识别和词性标注方面已经取得重要进展，句子结构分析和语义分析方面仍有待探索

系统演示

- 北京大学现代汉语分词/词性标注/句法分析系统（孙斌、刘群、常宝宝、詹卫东等）

- http://www.icl.pku.edu.cn/nlp-tools/segtagtest.htm
  （北大计算语言所网上分词、标注、注音系统）

中文信息处理的发展趋势

- 发展趋势
  信息产品的多样化
  网络的迅速发展

信息家电，内容计算，……

积累更多基础资源，
开发更多应用系统。
内容层的处理将受到越来越多的重视

三 语言知识资源的建设

- 现代汉语语法信息词典
- 基于配价理论的现代汉语语义词典
- 现代汉语短语结构信息库
- 2700万字现代汉语分词与词性标注语料库
- 句子对齐的汉英双语语料库
- 现代汉语树库
- 现代汉语短语结构规则库

资源演示

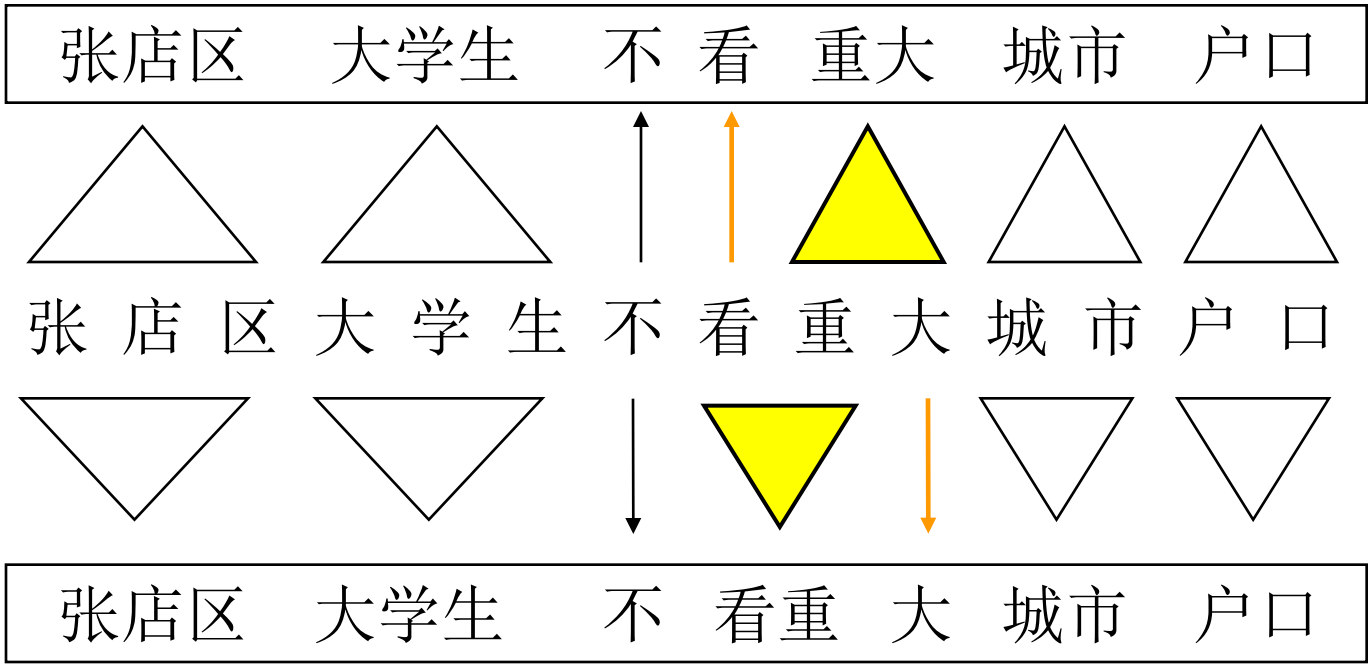
- 现代汉语语义词典（詹卫东、王惠等）http://ccl.pku.edu.cn
- 汉英平行语料库（常宝宝、柏晓静等）
- 现代汉语树库（詹卫东、常宝宝等）

四 面向中文信息处理的语言学研究

- 充分重视各个层次上的语言歧义研究
- 拓展语言现象的研究面
- 强调研究结果的可操作性，推动语言知识的形式化、系统化和规模化

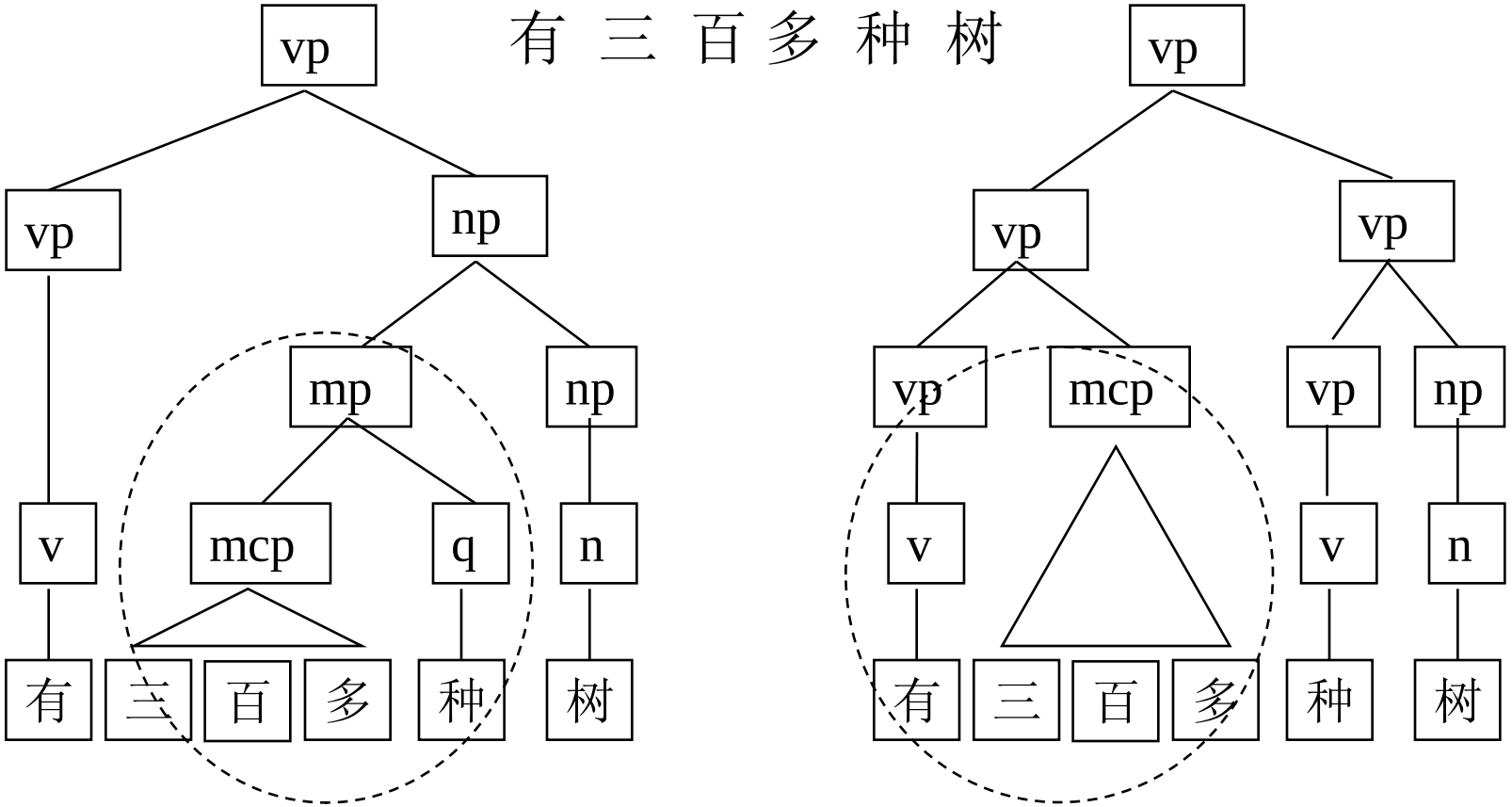
加强语言知识库的工程建设，为中文信息处理（内容层的处理）积累更多基础资源

歧义示例



歧义示例（续）

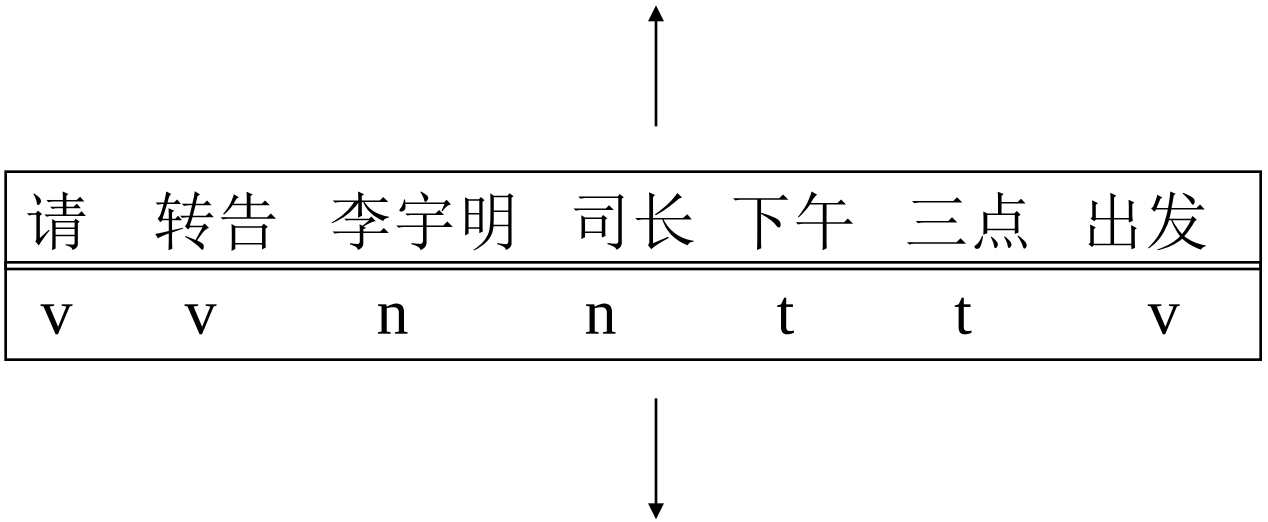
有 三 百 多 种 树



有	三	百	多	种	树
v	m	m	m	q/v	n

歧义示例（续）

请转告李宇明司长　　下午三点出发



请转告李宇明　　司长下午三点出发

结 语

要让计算机“理解”一个句子，实际上要解决下面两个核心问题：

（1）一个句子的结构和意义是什么？

（2）如何得到一个句子的结构和意义？

第一个问题是“What”的问题，这是理论语言学关心的问题；

第二个问题是“How”的问题，这是计算语言学关心的问题，
也就是面向中文信息处理的语言研究需要关心的问题。

参考文献

慈林林 鲁元魁，1999，《中文信息处理新技术展望》，《计算机世界》1999年第44期“产品与技术”版“专题报道”。

刘梦松，1998，《中文信息处理软件概述》，《计算机世界》1998年第26期“技术专题”版。

许嘉璐，2002，《现状和设想 —— 试论中文信息处理与现代汉语研究》，《中国语文》2000年第6期。

俞士汶，朱学锋，2002，《关于汉语信息处理的认识及其研究方略》，《语言文字应用》2002年第3期。

俞士汶，朱学锋，王惠，2001，《<现代汉语语法信息词典>的新进展》，《中文信息学报》2001年第1期。

詹卫东，常宝宝，俞士汶，2002,《机器翻译与语言研究》，《语言科学》2002年第1期（创刊号）。

詹卫东，2000，《 80年代以来汉语信息处理研究述评》，《当代语言学》 2000年第2期。

张华平，2003，《中文信息处理技术发展简史》，http://www.nlp.org.cn（中文信息处理开放平台网站）

国内外重要的语言知识资源举例

- WordNet，http://www.cogsci.princeton.edu/~wn/
- FrameNet，http://www.icsi.berkeley.edu/~framenet/
- HowNet，http://www.keenage.com/
- 台湾中研院词库、现代汉语平衡语料库 http://www.sinica.edu.tw/ftms-bin/kiwi.sh

谢 谢

请大家批评指正

欢迎访问

http://ccl.pku.edu.cn

http://icl.pku.edu.cn