第 2 章　综合型语言知识库系统

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，这已经成为学界的共识。本章首先简要介绍国内外有代表性的语言知识库系统；其次，详细介绍 ICL/PKU 已有的综合型语言知识库，包括数据资源和工具软件；最后，介绍综合型语言知识库系统的构建思想、总体建设规划及完成的主体功能模块。

2.1　国内外相关研究

语言知识库的建设涉及语言知识的整理、发现、形式化、规范化等工作，语言知识库的内容和知识表现形式是多样的。广义上，语言知识库包括词汇知识库、语料库、规则库及各种处理工具等。其中，词汇知识库主要描述词语的词法、句法或语义信息；语料库最简单的是生语料（raw text），或者是参照某种语言学理论对文本中隐含的语言现象进行显性标注的语料（annotated text）。表 2.1 列出了部分国外有代表性的词汇知识库和语料库的基本情况。

表 2.1 中的语言知识库项目有一些已经涵盖中文，如 Penn Treebank 和 Chinese Gigaword。国外建设知识库的理论和方法对国内的相关研究提供了借鉴。例如，山西大学参照 FrameNet，以汉语语料事实为依据构建了 Chinese FrameNet，包括框架库、句子库和词汇库 3 个部分的内容；北京大学基于 WordNet

框架，构建了《中文概念词典》（CCD，Chinese Concept Dictionary）。除此之外，知网及《同义词词林》等汉语语言知识库，是国内学者自主研发的成果。北京大学、清华大学、北京语言大学、北京邮电大学、中国科学院软件所、哈尔滨工业大学及中国台湾“中央研究院”等大学和研究机构在汉语语言知识库的建设方面都颇有成就。其中，北京大学开发的综合型语言知识库是在以汉语为核心的多语言知识库建设中最全面、最重要的研究成果[1]。

表 2.1　部分国外有代表性的词汇知识库和语料库的基本情况

类型	名称	开发单位	开发时间	内容	语种
词汇知识库	WordNet	美国普林斯顿大学	1985－	本体知识库、同义词集合、语义关系描述	英语
	EDR电子辞书	日本电子辞书研究所	1986－1994	语义分类、语义关系描述	日语、英语
	MindNet	美国微软公司	1993－	语义关系描述	英语
	EuroWordNet	Human Language Technology Sector	1996－1999	本体知识库、同义词集合、语义关系描述	欧洲语言（德语、西班牙语等）
	FrameNet	美国加州大学伯克利分校	1997－	语义框架、语义关系	英语
	VerbNet	美国科罗拉多大学	1998－2002	句法语义信息的动词词汇库	英语
	Integrated Linguistic Database	英国剑桥大学、爱丁堡大学等	1993－1996	语义分类、语义特征、语义角色与选择限制等	英语
语料库	English Gigaword	Linguistic Data Consortium	2003年发布	4种来源的新闻语料（《纽约时报》等）	英语
	Penn TreeBank	美国宾夕法尼亚大学	1992－	短语结构的句法树库	英文、中文、阿拉伯语等多种语言

[1] 此句评论引自 2007 年 2 月在北京举办的教育部组织的技术鉴定会的鉴定结论，详细内容见参考文献[52]。

续表

类型	名称	开发单位	开发时间	内容	语种
语料库	NomBank	美国纽约大学	2007年发布第1版	名词的配价信息	英文
	PropBank	美国宾夕法尼亚大学	2003－	论元谓词结构信息	英文、中文、阿拉伯语等多种语言
	Chinese Gigaword	Linguistic Data Consortium	2003年发布	新闻语料（北京新华社和中国台湾中央通讯社）	中文（简体、繁体）

2.2 综合型语言知识库的资源概况

ICL/PKU 自 1986 年起，一直致力于各种语言知识库的建设，目前已经拥有一系列质量上乘、内容丰富的语言知识库。这些知识库既包含词语、短语、句子、篇章等不同语言单位，又涉及汉语、英语等不同语言，并从词法、句法和语义等不同角度进行信息描述。时至今日，各种语言数据资源大致成熟齐备，都遵循同样的语法理论体系，存在内在的联系和协调的分工，因此实现各个知识库资源的有机融合已经成为必然趋势。

在对各个数据资源的特点进行仔细分析的基础上，参考文献[60]提出了以词义为主轴建设综合型语言知识库（CLKB，Comprehensive Language Knowledge Base）系统的构想。本节主要介绍 CLKB 目前已有的主要数据资源及相关工具软件等[2]。

CLKB 包含的语言数据资源如下。

（1）《现代汉语语法信息词典》（简称《语法信息词典》或 GKB）。

（2）《面向汉英机器翻译的现代汉语语义词典》（简称《语义词典》或 CSD）。

（3）现代汉语多级标注语料库（简称多级标注语料库或 STC）。

[2] 综合型语言知识库所含资源不限于列出内容，本章只介绍和本书工作相关的数据资源和工具软件，更详细的内容见参考文献[62]。

（4）《面向跨语言文本处理的中文概念词典》（简称《中文概念词典》或CCD）。

辅助语言知识库建设的工具软件如下。

（1）语料库检索软件。

（2）《语法信息词典》管理平台。

（3）文本型语料库与结构化语料库之间的转换软件。

（4）分词及词性标注软件。

2.2.1 语言数据资源简介

1.《语法信息词典》（GKB，Grammatical Knowledge Base）

《语法信息词典》是语言知识库大厦的第一块基石，它是一部面向语言信息处理的大型电子词典，收录 8 万个词语，依据语法功能（优势）分布的原则，建立了面向信息处理的词类体系，又按类描述每个词语的详细的语法属性[59]。《语法信息词典》将现代汉语的词分为 18 个基本词类，基于文本自动处理的需要，也包含 7 类非词的语言成分，如表 2.2 所示。表 2.2 中的“词类代码”是用于计算机处理的小写英文字母，“词语实例”给出了该类的部分词语样例。

表 2.2 《语法信息词典》中的基本词类和非词的语言成分

类型	词类名称	词类代码	词语实例
基本词类	名词	n	地道、门、马、书、水、教师、祖国、心胸
	时间词	t	现在、明天、元旦、春天、清朝、宋代
	处所词	s	空中、低处、郊外
	方位词	f	上、下、左、右、前、后、东、西、里面、外头、中间
	数词	m	一、二、三、百、千、万、零、第一、许多、百万
	量词	q	个、位、张、把、匹、头、群、克、杯、片、种、些
	区别词	b	男、女、金、银、公共、微型、初级
	代词	r	我、你、他、我们、这里、哪儿、谁、怎么样
	动词	v	调配、来、去、讲、同意、能够、是、调查、编辑

续表

类型	词类名称	词类代码	词语实例
基本词类	形容词	a	地道、顺利、干净、好、红、大、温柔、美丽
	状态词	z	冰凉、雪白、金黄、泪汪汪、马马虎虎、空荡荡
	副词	d	非常、不、很、挺、都、刚刚、难道、忽然
	介词	p	把、被、对于、关于、以
	连词	c	和、与、或、虽然、但是、不但、而且
	助词	u	了、着、过、的、得、地
	语气词	y	吗、呢、吧、嘛、啦、呗
	拟声词	o	呜、啪、叮咚、哗啦
	叹词	e	啊、唉、喔、哎哟、嗯、呀
非词的语言成分	前接成分	h	阿、老、非、超、微
	后接成分	k	子、儿、性、员、器、者
	语素	g	民、衣、失、遥、郝
	非语素字	x	蜈、蚣、枇、杷、琵、琶
	成语	i	指鹿为马、南辕北辙、龙飞凤舞、滥竽充数、朝秦暮楚
	习用语	l	木头疙瘩、光杆司令、跑龙套、走后门
	简称略语	j	北大、人大、政协、三好

当代若干计算语言学语法理论都以采用复杂特征集的词汇知识表示和基于合一的分析算法为特征。《语法信息词典》在这些语法理论的启示下，采用在大致分类的基础上，以“属性-属性值”的形式详细描述词语的句法知识，采用成熟的关系数据库技术，将“属性-属性值”的描述形式转换成数据库二维表的字段与值。表 2.3 显示了《语法信息词典》中总库的样例和部分字段。其中，属性“词语”+“词类”+“同形”构成了《语法信息词典》的主关键项。其实，除详细的语法信息外，《语法信息词典》也包含丰富的词汇语义知识，如“同形”字段区分了词的粗粒度义项等。

表 2.3 《语法信息词典》中总库的样例和部分字段

词语	词类	同形	拼音	注释	……
挨	v	A	ai1	触，碰，靠近	
挨	v	B	ai2	遭受，忍受	
保管	v	1	bao3guan3	保存	
保管	v	2	Bao3guan3	担保	
报告	n		bao4gao4	书面文件	
报告	v		bao4gao4	发表讲话	
别	d		bie2	不要	
别	v	A	bie2	分离	
别	v	B	bie2	附着或固定	
地道	a		di4dao5	正宗	
地道	n		di4dao4		
叫	v	A1	jiao4	人或动物发出的较大声音	
叫	v	A2	jiao4	呼唤，招呼；雇	
叫	v	A3	jiao4	称为	
叫	v	B	jiao4	使，让，命令	

以《语法信息词典》的分类为基础，需要按类详细描述每个词语的语法属性。以动词为例，动词的信息最丰富，因此在动词库中共确定了 46 个属性，内容涉及该动词前面能不能受副词“不、没、很”修饰，后面能不能带助词“着、了、过”，还包括它能带什么类型的宾语等。本书不再详细列举《语法信息词典》中的属性描述，如果读者希望详细了解每个属性字段的含义，请见参考文献[53]和[59]。

2.《语义词典》（CSD，Chinese Semantic Dictionary）

《语义词典》是一个面向机器翻译的大规模汉语语义知识库，收录 6.6 余万个实词，详细描述每个词语的义项、语义类及基于配价理论的语义搭配限制，可为计算机语义自动分析、词义消歧等任务提供强有力的支持[35]。

《语义词典》继承了《语法信息词典》的数据模式，依据词义理解的需要设

定多个不同的特征属性，依据属性值的不同即可辨别出不同的义项，而且《语义词典》描述的语义知识和《语法信息词典》描述的句法知识采用统一的描述形式，便于实现句法-语义接口（Syntax / Semantic interface）。《语义词典》完全继承了《语法信息词典》的“词语”、“词类”、“同形”这 3 个字段的信息，并增加了“义项”字段，“词语”+“词类”+“同形”+“义项”构成了《语义词典》的主关键项，“同形”和“义项”两个属性字段共同构成一个词语的意义编码。表 2.4 就是《语义词典》对动词“开”的不同义项的描述。

表 2.4　《语义词典》中的样例

词语	同形	义项	释义	语义类	子类框架	配价数	主体	客体	Word	特殊句法位置
开	1	1	打开	其他行为	[NP]	2	人		Open	
开	1	2	展开，分开	变化	[～]	1	～人		open out	
开	1	3	沸腾	变化	[～]	1	～人		Boil	
开	1	4	支付；开销	领属转移	[NP]	2	人	钱财	Discharge	
开	1	5	发动或操纵	其他行为	[NP]	2	人	交通工具|武器	Drive	
开	1	6	开办	创造	[NP]	2	人	建筑物	Open	
开	1	7	举行	社会活动	[NP]	2	人	事件	Hold	
开	1	8	写出	创造	[NP]	2	人	票据	Write	
开	1	9	开创（抽象事物）	创造	[NP]	2	人	抽象物	Open	
开	1	10	打开（电器）使运作	其他行为	[NP]	2	人	电器	turn on	
开	1	11	开辟	其他行为	[NP]	2	人	地理	open up	
开	1	12	开通	社会活动	[NP]	2	人	符号	open up	

续表

词语	同形	义项	释义	语义类	子类框架	配价数	主体	客体	Word	特殊句法位置
开	1	13	使开阔	其他行为	[NP]	2	人	心理特征	open up	
开	2	14	开始	其他行为	[NP]	2	人	事件	Begin	
开	3	15	用在动词或形容词后做补语							a+～|v+～|v+得+～|v+不+～

3．多级标注语料库（STC，word-Sense Tagging Corpus）

ICL/PKU 从 1992 年开始进行汉语语料库的多级加工研究。第一步就是对原始语料进行切分和词性标注，加工《人民日报》语料，形成了大规模现代汉语基本标注语料库，规模达到 6000 万字[57]。基本标注语料库中的人名、地名及团体机构名等命名实体，都有相应标记予以标识。此外，以《语法信息词典》和《语义词典》为参考，在基本标注语料中加注不同粒度词义信息，就形成了多级标注语料库。

语料中词的每次出现都有语境，词义、句法功能、语义角色都是确定的，但却是隐性的。语料加工就是使文本中隐含的信息显性化，加工越深，显性化的知识就越多。例（2.1）和例（2.2）显示了多级标注语料库的标注情况，其中的命名实体用“[”和“]”标识，并给以“nt”的标记，其中的“活动”、“问题”和“使”等多义词的词义信息也予以区分，并分别标注了它每次出现在上下文中的词义。例句中词性标记说明可参见附录 A，详细的分词规范和解释可见参考文献[57]。

（2.1）中国/ns　积极/ad　参与/v　[亚太经合/j　组织/n]nt　的/u　**活动/vn!2-1**　，/w　参加/v　了/u　东盟/ns　—/w　中/j　日/j　韩/j　和/c　中国/ns

—/w 东盟/ns 首脑/n 非正式/b 会晤/vn 。/w 这些/r 外交/n **活动/vn!2-1** ，/w 符合/v 和平/a 与/c 发展/v 的/u 时代/n 主题/n ，/w 顺应/v 世界/n 走向/v 多极化/vn 的/u 趋势/n ，/w 对于/p 促进/v 国际/n 社会/n 的/u 友好/a 合作/vn 和/c 共同/b 发展/vn 作出/v 了/u 积极/a 的/u 贡献/n 。/w

（2.2）温/nr 家宝/nr 在/p 农民/n 家中/s 详细/ad 询问/v 他们/r 生产/vn 生活/vn 情况/n ，/w 同/p 干部/n 群众/n 一起/s 研究/v 扶贫/v 开发/v 的/u 路子/n 。/w 他/r 说/v ，/w 必须/d 坚持/v 开发/v 扶贫/v 的/u 方针/n ，/w 通过/p 发展/v 经济/n 解决/v 贫困/a 人口/n 的/u 温饱/n **问题/n!0-1** 。/w 要/v 把/p 农业/n 生产/vn 尤其/d 是/v 粮食/n 生产/vn **放/v!1-8** 在/p 第一/m 位/q ，/w 首先/d 解决/v 群众/n 吃饭/v **问题/n!0-1** 。/w 同时/c ，/w 面向/v 市场/a 需求/n ，/w 充分/ad 利用/v 当地/s 资源/n ，/w 积极/ad 发展/v 多种/m 经营/vn ，/w 增加/v 农民/n 收入/n 。/w 温/nr 家宝/nr 考察/v 了/u 农田水利/l 建设/vn 工地/n ，/w 他/r 说/v ，/w 要/v 大/d 搞/v 农田/n 基本建设/l ，/w 植树造林/l ，/w 治水改土/l ，/w 改善/v 生产/vn 条件/n 和/c 生态/n 环境/n ，/w **使/v!2** 脱贫/v 建立/v 在/p 比较/d 坚实/a 的/u 物质/n 技术/n 基础/n 之上/f 。/w

4.《中文概念词典》（CCD，Chinese Concept Dictionary）

《中文概念词典》是基于 WordNet 框架下的现代汉语概念词典，描述了概念之间丰富的语义关系，包括组合分布信息和聚合关系信息[22, 48]。《中文概念词典》呈现给用户的基本数据结构是若干棵树，树上的每个节点都是一个概念，而概念则用同义词集（Synset）来表示。《中文概念词典》描述了概念间的上下位关系（hyponymy）、反义关系（antonymy）、整体-部分关系（meronymy）、蕴涵关系（entailment）、致使关系（cause）等，所有这些信息附加在树结构之上，构成了更复杂的网结构，这种网结构体现了概念之间的关系和约束。目前，《中文概念词典》已包含十万多个概念的描述和汉英双语概念的对应。以词语“先生”为例，表 2.5 给出了“先生”反映的不同概念，或者说“先生”的不同义

项分别用同义词集合表示，处于“树”上的不同节点中。

表 2.5 《中文概念词典》中的词语样例

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 ……	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
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
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

以上语言数据资源涉及的语言知识及其表述形式独立于特定的语言信息处理系统和实现算法，这种设计理念使得知识库的内容便于用户理解和运用，并得以广泛传播。

2.2.2 工具软件简介

1. 语料库检索软件

语料库检索软件是一个针对基本标注语料库和生语料开发的检索系统。该系统使用了一种基于词语和标记的混合全文索引方法[34]，大大提高了检索效率，用户可以用词语和词性的各种组合作为条件进行检索，方便研究者们快速高效地检索和分析标注语料，从而发现大数据下的语言规律和语言现象。该软件功能主要包括关键词查询、搭配分析、关键词的例句提取等。图 2.1 显示了该软件中检索远距离搭配的一个例子，检索语料中包含“就 * * * 问题”的句子（其中*表示任意的切分单位），并且可以设定检索词之间的距离，该例中限定为 4，即两个词语之间最多包含 3 个切分单位。

图 2.1 语料库检索软件的远距离搭配检索界面

2.《语法信息词典》管理平台

《语法信息词典》管理平台是针对《语法信息词典》开发的集建设、管理、检查、使用为一体的多功能平台[42]。该平台在《语法信息词典》的扩充过程中发挥了很大的作用，保证了词典中不同层级数据库的一致性。图 2.2 显示了《语法信息词典》管理平台的操作界面。

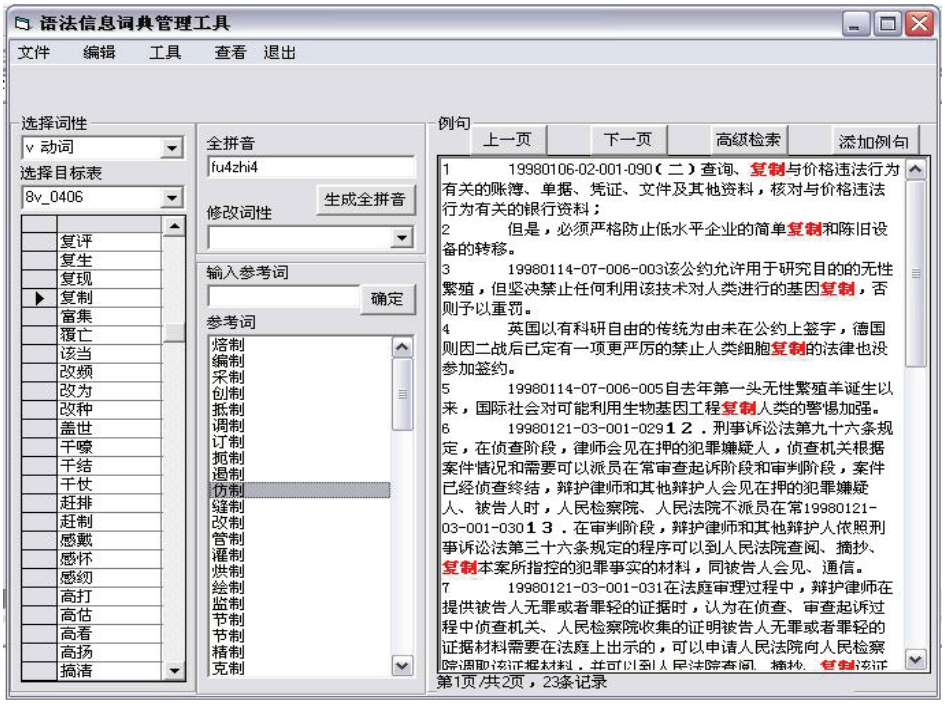

图 2.2 《语法信息词典》管理平台的操作界面

3．文本型语料库与结构化语料库之间的转换软件

文本型语料库与结构化语料库之间的转换软件可以将线性顺序的文本转换成结构化的关系型数据库，基于关系型数据库可以方便、快速地完成各种统计及进一步的知识挖掘；此外，该工具软件也为基本标注语料库和《语法信息词典》的集成提供技术支持[74]。

4．分词及词性标注软件

分词及词性标注软件基于《语法信息词典》，采用有效的语言模型和算法

模型，将分词和词性标注结合起来，利用丰富的词类信息对分词决策提供帮助，并且在标注过程中又反过来对分词结果进行检验、调整。对于汉语命名实体的识别，利用规则与统计的方法建立了切分/标注/NE（命名实体）识别一体化的模型[31]。

所谓“工欲善其事，必先利其器”，各种工具软件在语言资源的建设过程中发挥了至关重要的作用，不仅加快了资源开发的速度，也在很大程度上保证了资源的质量。

2.3 系统集成方案

2.2 节介绍了 ICL/PKU 目前已有的语言数据资源的结构特点及相关的工具软件。这些语言数据资源虽然是在不同时间、面向不同应用需求开发的，但是它们都遵循同样的语法理论体系，存在内在的联系和协调的分工，已经形成了比较完整的体系[52]。目前，各种语言数据资源基本上是独立存在的，应当把这些语言数据资源集成起来，形成一个综合型的语言知识库系统。

建成这样的综合型语言知识库系统不可能一蹴而就，需要分步实现，开始规模不宜过大，成分数据资源的类型不宜过多。在各种语言数据资源中，《语法信息词典》和基本标注语料库最具基础性。《语法信息词典》是综合型语言知识库大厦的第一块基石，多级标注语料库是在《语法信息词典》的基础上发展的，而《语义词典》是《语法信息词典》的自然扩充，它更细致地描述词语的语义信息。经过仔细的思考和规划，参考文献[60]明确提出了以词义为主轴将现有的语言数据资源集成为综合型语言知识库系统的方案，首先集成《语法信息词典》、多级标注语料库及《语义词典》三大基础资源，形成综合型语言知识库系统的主体部分。这个方案解决了集成不同类型知识库的最大难题，可以实现句法知识和语义知识的整合。

要实现三大语言数据资源的无缝整合，必须填补它们之间的“缝隙”。考察《语法信息词典》与基本标注语料库的缝隙，存在于 3 个方面：语言单位不同、词性标记集不同及词的信息表达方式不同。首先，词典中登录项是语

法词，而语料库中的切分单位是句法词，二者并不完全等价，不能建立起一一对应；其次，基本标注语料库的词性标记与《语法信息词典》的词类代码存在多对一的关系；最后，词典中的信息是显性的和确定的，真实文本的信息是隐性的和不确定的。

对于这些“缝隙”，根据不同的应用需求，可以采取相应措施予以填补。例如，对于语法词和句法词的不同，参考文献[64]提出了“部件词”作为过渡，无论是语法词的集合还是句法词的集合，都可以看作是由“部件词”和“非部件词”两个部分构成的，基于“部件词”频次可以构建更符合人们认知的常用词库。对于词类代码与词性标记的不一致，词典的词类代码是一个静态的分类，而语料库的词性标记则动态地反映了词语在实际使用时的情况，如“动词 v”分化出“名动词 vn”和“副动词 vd”等，我们通过构造对照表（见附录 A）进行转换。

对于综合型语言知识库系统，按照以词义为主轴集成现有语言数据资源的指导思想，首先需要解决的是填补词的信息表达方式方面的“缝隙”，解决方法是在基本标注语料库中增加“同形”信息。2.2.1 节说明了《语法信息词典》的关键项是“词语”+“词类”+“同形”，基本标注语料库只有“词语”和“词类”，因此在基本标注语料库上增加“同形”信息就可以实现《语法信息词典》和基本标注语料库的连接，这正是以“词语”+“词类”+“同形”为主轴的含义。《语义词典》的关键项是“词语”+“词类”+“同形”+“义项”，基于同样的原理和技术，在同形标注语料库上增加“义项”，就可以实现与《语义词典》的连接。上述过程实际上就是多级标注语料库的加工流程，如图 2.3 所示。

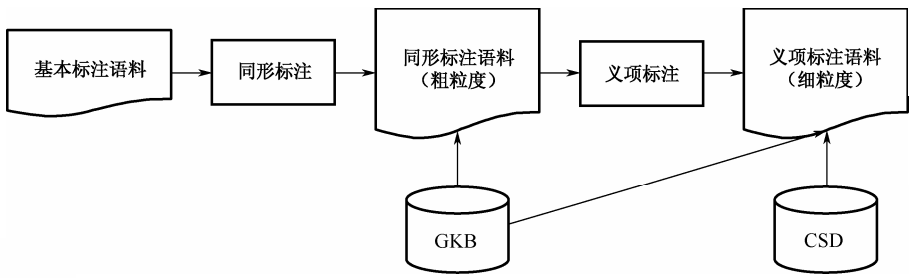

图 2.3　多级标注语料库的加工流程

2.4 系统功能

综合型语言知识库系统的主体功能模块由 3 个部分组成，分别是语言加工模块、知识检索模块、知识挖掘模块，每个模块提供不同的服务，如图 2.4 所示。建成的综合型语言知识库系统可以为语言本体研究者及一般用户提供全方位的、精准的知识查询和服务，同时也为自然语言处理应用系统的开发提供支持。下面将详细介绍各个模块的主要功能。

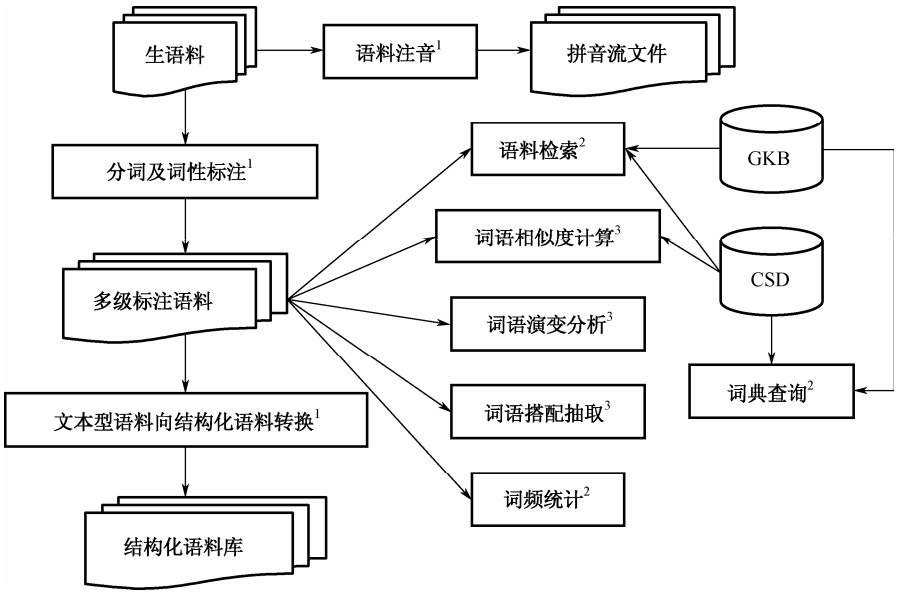

图 2.4 综合型语言知识库系统功能示意图[3]

2.4.1 语言加工模块

该模块是一个基本的语言资源加工模块，可以对生语料进行不同层次的加工和处理，为后续应用提供相应的资源。该模块包括分词及词性标注、语料注音、文本型语料向结构化语料转换等基本功能。其中，语料注音软件，输入是

[3] 其中，1 表示语言加工模块，2 表示知识检索模块，3 表示知识挖掘模块。

生语料（文本文件），输出提供两种结果：一种是拼音流文件，将文字转换成拼音，结果文件中不包含汉字；另外一种是标注流文件，保留分词结果，并在词语后面加注拼音。该功能可以为语音合成（TTS，Text To Speech）系统提供文本的预处理。其余两个功能的介绍见 2.2.2 节。

2.4.2 知识检索模块

该模块为用户呈现知识库中显性的、直接的语言知识，主要包括 3 个功能。

第一，语料检索功能，提供多种查询条件下对语料的检索，并以 KWIC（Key Word In Context）方式显示检索结果。此外，提供语料检索结果的同时，该模块还提供交叉参照功能，该功能实现了三大基础资源之间便捷、准确的对应和参照，方便用户从数据结构各不相同的多种语言资源获取丰富的语言知识。根据语料库标注深度的不同（同形或义项），可以准确地定位到《语法信息词典》和《语义词典》中的相应记录，获取不同层次的词汇信息。图 2.5 显示了语料检索功能的一个示意界面，检索词是“中国”，图中左下部分显示从语料库中得到的例句，右边部分分别显示检索词在两部词典中的详细属性信息。

图 2.5 语料库到词典的交叉参照示意界面

第二，词频统计功能，它的输入是文本型语料，语料可以是基本标注、同形标注及义项标注 3 个层次，对其进行处理转换成结构化语料库，在处理的同时，完全保留切分单位的标注信息，并将切分单位及其标注信息作为数据库的关键字，依次建立词条，统计各个词条的频次并存储在数据库中。

第三，词典查询功能，提供对《语法信息词典》和《语义词典》的查询和浏览，获取满足多种查询条件的词语的各种属性信息。

2.4.3 知识挖掘模块

该模块综合运用各种资源为用户获取隐性的、深层的语言知识，包括 3 个功能：词语搭配抽取、词语演变分析及词语相似度计算。

（1）词语搭配抽取功能提供 4 种假设检验的方法抽取搭配，分别是 Likelihood Interval Method、Likelihood Ratio Test、μTest、χ^2 Test。用户可以设定目标词（target word）及词性，在指定的语料中抽取目标词的搭配词，并对结果进行比较和分析。

（2）一个词语的语义环境（表现为上下文的词语）随着时间的变化，是逐渐改变的，有些词语可能不再出现（旧的语义消失），也有些新词语会出现（新的语义生成）。词语演变分析功能通过计算一个词语在不同时段的上下文之间的相关度，为用户提供词语的演变情况。该功能需要时间跨度较长的大规模语料的支持。

（3）词语距离有两类常见的计算方法：一类是仅根据某种世界知识（Ontology）或分类体系（Taxonomy）来计算，这里利用《语义词典》分类体系来计算语义相似度；另外一类除了依照分类体系外，还需要大规模的语料库的统计信息，这里借助多级标注语料库进行统计。基于这两类方法，该模块提供 5 种词语相似度计算功能[4]。

[4] 关于这 5 种计算词语相似度的方法，本书不作过多介绍，可见参考文献[81]在 5.3.2 节中对词义相似度计算的详细介绍。

2.5 本章小结

本章主要介绍了综合型语言知识库的资源概况、系统集成方案及系统功能等，它可以为深层次的语言知识发现提供有力的软件支持。以知识检索模块中的词频统计功能为例，我们按照月份处理语料，分别建立基本词频表，这些作为原始数据，可以对基本词频表进行各种操作和查询，例如，执行合并操作，生成更大时间跨度的词频表，或者指定查询条件，从词频表中获取特定的词频信息等。这样的处理方法不但可以方便、快速地完成各种统计，而且保证了统计结果的准确性。例如，为词语分布均匀度的计算提供基本参数[155]。在同形标注语料库上，可以统计《语法信息词典》中登录项的频次，得到粗粒度的词义分布情况。同理，在义项标注语料库上，可以进一步统计得到细粒度的词义分布情况。这些都是汉语词汇计量研究逐步深入的成果，为基于统计的自然语言处理方法提供必要的概率信息。此外，就本项研究工作而言，综合型语言知识库系统的相关功能模块在构建《概率型现代汉语常用词汇知识库》的过程中也发挥了重要作用。

需要特别说明的是，在综合型语言知识库系统的诸多功能模块中，如切分及词性标注、语料检索、语料注音、词语搭配抽取、词语演变分析等，都利用了 ICL/PKU 积累的相关软件代码，这样不仅节省了工作量，缩短了开发周期，而且也是对前人工作的继承和发扬。此外，按照系统总体规划和设计方案，对于没有相关积累的功能模块，如词频统计、词语相似度计算、语料库到词典的交叉参照及整个系统的集成工作，是我们后续开发完成的。

目前，综合型语言知识库系统还在发展中，在已有的基础上，我们还将逐步实现其他资源的集成。例如，基于 WordNet 构建的《中文概念词典》，如前所述，它的结构不同于《语法信息词典》和《语义词典》，它以同义词集（Synset）作为词典的基本单元，每个同义词集中的词语都承载了确定的概念，其词义是确定的。因此，如果实现《语义词典》到同义词集的映射关系，那么《语法信息词典》、《语义词典》、《中文概念词典》和语料库都集成到一起