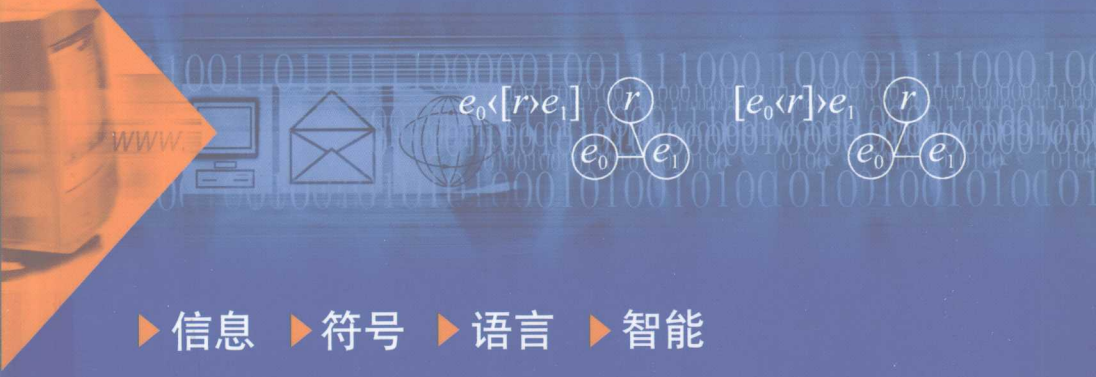
信息依存句法标注模型

Information Dependency Syntax Tagging Model

李良炎 著



信息 符号 语言 智能

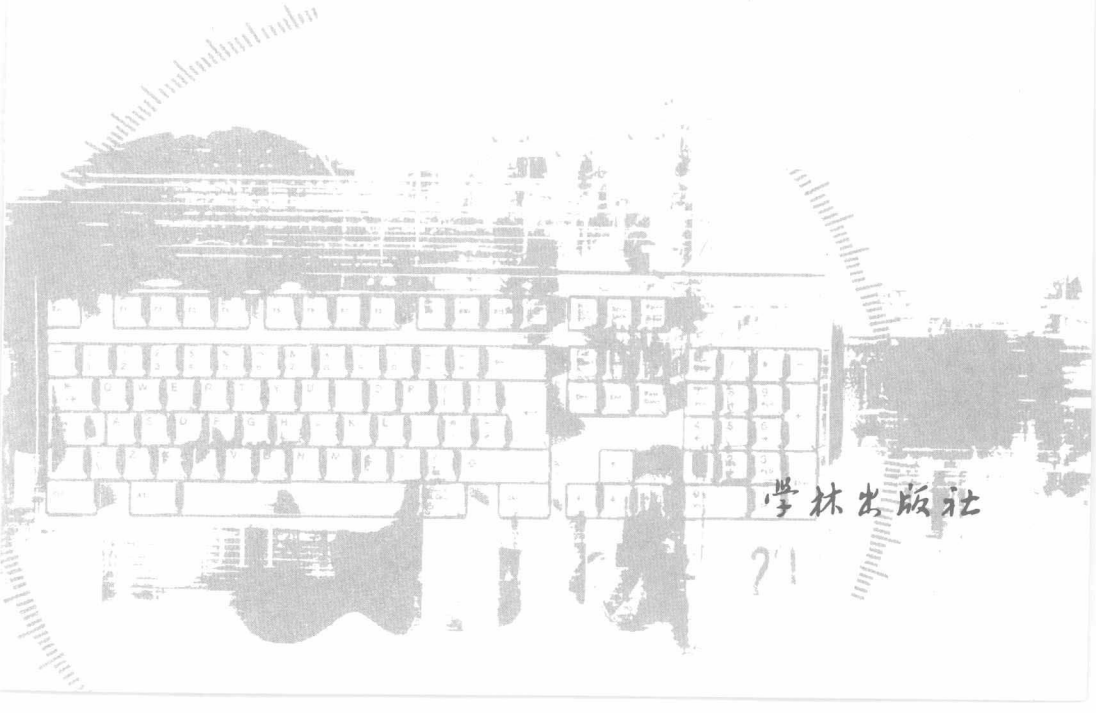
学林出版社

重庆市人文社会科学重点研究基地基金资助
重庆市教育委员会人文社会科学研究项目（07SK158）
重庆大学“985”人才引进项目（0903005104218）

信息依存句法标注模型

Information Dependency Syntax Tagging Model

李良炎 著



学林出版社

图书在版编目(CIP)数据

信息依存句法标注模型/ 李良炎著. —上海:学林出版社, 2009.2
ISBN 978-7-80730-730-3

Ⅰ. 信… Ⅱ. 李… Ⅲ. 汉语—句法—研究 Ⅳ. H146.3

中国版本图书馆 CIP 数据核字(2008)第 175090 号

重庆大学外国语学院学术文库

信息依存句法标注模型

作　　者——李良炎
责任编辑——李晓梅
封面设计——周剑峰
出　　版——上海世纪出版股份有限公司
学林出版社(上海钦州南路81号3楼)
电话:64515005　传真:64515005
发　　行——新华书店上海发行所
学林图书发行部(上海钦州南路81号1楼)
电话:64515012　传真:64844088
印　　刷——上海书刊印刷有限公司
开　　本——890×1240　1/32
印　　张——8.5
字　　数——22万
版　　次——2009年2月第1版
2009年2月第1次印刷
书　　号——ISBN 978-7-80730-730-3/G·207
定　　价——24.00元

(如发生印刷、装订质量问题,读者可向工厂调换。)

前　言

语言的复杂性在于语言与认识的关系。语言表达意义，意义是人对主客观世界的认识结果。主客观世界的复杂性决定了意义的复杂性，从而决定了语言的复杂性。语言本身又可以视为人的主客观世界中的一部分，因此语言研究是一种特殊的认识活动，是人对语言的认识。由此可见，语言离不开认识。

人对主客观世界的**认识**可以如此描述：认识**主体**借助认识**工具**按照认识**方法**处理认识**对象**获得认识**结果**。认识是由四种认识因素（主体、工具、方法、对象）共同构成的活动，认识结果是这一活动的产物，被多种认识因素共同决定，任何一种认识因素的改变必然会导致认识结果或大或小的差异。认识具有主观能动性，是认识主体对认识对象的选择性反映。显然，认识结果与认识对象不能等同。从这个意义上讲，认识不可能也不应该去被动地还原认识对象，而应**从符合主体目的性出发，力求简单有效地描述和预测认识对象**。借用模型的概念，所谓认识结果就是认识对象的**模型**(model)，而所谓认识就是建立认识对象的模型，简称**建模**(modeling)。这是一种实用主义认识观。

模型一般分为**心理模型**(psychological model)、**数学模型**(mathematical model)和**物理模型**(physical model)。心理模型是认识对象在人认识中的定性关系，是数学模型的基础；数学模型是认识对象在人认识中的定量关系，是物理模型的基础；物理模型是人借助特定材料和工具按照认识对象的数学模型实现的物质结构。传统意义上的建模主要指建立数学模型和物理模型，普遍意义上的建模还包括建立心理模型。人的认识能力是有限的，表现

在：在特定的时空条件下，人不能建立任意认识对象的心理模型，也不能建立任意心理模型的数学模型，也不能建立任意数学模型的物理模型。由于具有明确的实用主义特点，建模在理工科领域大行其道，在文科领域也逐渐受到青睐。人类将二进制数学模型成功实现为电子管、晶体管等物理模型，并开发出越来越复杂和先进的计算机软件和硬件，从而进入信息时代。20世纪以来一些主要或次要的语言理论都或多或少应用了数学模型，特别是一些面向语言计算的语言理论。

随着计算机技术的飞速发展，人们对计算机自动或辅助处理语言信息的需求越来越大。但对计算机而言，凡是不能建立数学模型的信息是无法处理的。传统语言理论往往只在心理模型层面定性研究，无法满足这一需要。因此有必要引入数学模型研究语言，称为**语言数学模型**，简称**语言模型**(language model)。**统计语言模型**(statistical language model)就是一个成功的例子。但统计语言模型的性能取决于训练语料的规模和质量。目前，由于语料的不断积累和计算机技术的不断进步，语料规模已不成问题，语料中包含语言知识的数量和质量才是关键。

计算机的语言知识主要来源于人。将语料中包含的语言知识标注出来，有助于计算机获得更丰富、更有价值的语言知识，提高语言信息处理水平，这就是**语料标注**(corpus tagging)。经过标注的语料还可以用于语言学研究、语言教学、语言测试、词典编撰等诸多理论研究和实践应用领域，越来越受到人们重视，并逐渐形成一门新兴学科——**语料库语言学**(corpus linguistics)。

按标注层次分，语料标注主要包括**词法标注**(lexical tagging)、**句法标注**(syntax tagging)和**语篇标注**(discourse tagging)。目前，相对句法标注，词法标注有更成熟的规范、准确率更高的技术和更大的标注规模。句法标注的主要困难在于，没有一个真正成熟的语法或语义标注模型。句法结构很难统一描述，现有的句法理论

此为试读，需要完整PDF请访问：www.ertongbook.com

还不完善，难以制定统一规范，标注主观性很大，自动标注准确率比较低。因此，句法标注成了语料标注的瓶颈。由于句法知识在语言知识中具有十分重要的地位，有理由相信：如果有了大规模、高质量的句法标注语料库，基于语料库的各种研究和应用有可能在现有水平上产生质的飞跃。因此，对句法标注模型的研究是当务之急。语料库语言学属于交叉学科，句法标注模型是语料库语言学的基础理论，又与语言学的句法理论密切相关。一方面可以借鉴现有句法理论；另一方面，也可以从语料库语言学的角度研究句法，提出新的句法标注模型。本书立足于在分析比较现有主要句法理论的基础上提出新的面向语料库语言学的句法标注模型——**信息依存句法标注模型**(Information Dependency Syntax Tagging Model, IDSTM)，主要包括四部分：问题的提出、语言模型研究、语义解释策略研究、IDSTM 参数与实现。

第1章为问题的提出。构建了句法标注一般模型(STGM)作为参照，比较分析了短语结构语法标注模型(PSGTM)和依存语法标注模型(DGTM)，并基于认知语法(CG)的有关理论提出了改进 DGTM 的思路。

第2、3、4章为语言模型研究。基于意义、符号、语言三者之间的逻辑联系，先后提出了信息依存模型(IDM)、信息依存符号模型(IDSM)、信息依存语言模型(IDLM)。IDM 给出了信息依存(ID)这一核心概念，是本书的基石。IDLM 为 IDSTM 奠定了理论基础。

第5、6章为语义解释策略研究。通过对语义解释根本难点和基本策略的分析，提出了语义解释综合策略(SICS)，进一步对这种策略所依赖的智能模型进行了研究，提出了信息依存智能模型(IDIM)。IDIM 给出了包括人、机器人、软件人在内的一切智能体的形式化模型，使综合化的语义计算成为可能，为 IDSTM 的语义标注提供了解决方案。

第7章为 IDSTM 参数与实现。以 STGM 为参照，以 PSGTM 和 DGTM 为比较，分析了 IDSTM 的参数特征，提出了 IDSTM 的实现步骤，并重点研究了 IDSTM 的实现工具——信息依存标记语言(IDML)和对象模型标记语言(OMML)。IDML 基于 IDM 提出，OMML 基于 IDML 和 IDIM 提出。OMML 为大规模语料标注提供了简洁而强大的工具。

英国哲学家培根说过："知识就是力量。"语料标注本质上是对语言知识的描述，只有更加丰富、准确的语言知识才能真正提高语言研究和语言应用各个领域的水平。这是语料标注的真正价值所在，也是本书的出发点。同时，语言知识的描述难度是相当大的，主客观世界的复杂性、主体认识的复杂性、语言中大量存在的不确定性是语料标注困难的根源。但德国哲学家黑格尔说过："凡是现实的东西都是合乎理性的。"只要以理性为工具就能得到对现实的合理解释，从而得到符合主体目的性的有用知识。与语言有关的一切复杂性和不确定性都应当最终体现为语言知识的简洁形式。因为力求简洁，这正是主体认识的基本特征。

本书是笔者历经三载余潜心撰写而成，个中甘苦如鱼饮水，冷暖自知。遥想2004年准备博士答辩期间夜读《西游记》，曾赋《西游梦》七绝一首以抒怀：

经年碌碌西去险，
见性归来山海平。
拂却征尘浑忘忆，
拈花微笑月华明。

而今完成本书之余，重温《西游记》，仍有感慨，再赋《西游梦》七绝一首以抒怀：

有字本从无字吟，
八十一难见真经。
长风送我今朝返，
万水千山总是情。

在此，感谢我读硕士期间的导师赵伶俐教授，感谢我读博士期间的导师陈廷槐教授、吴中福教授、何中市教授，是他们让我懂得了为人冶学的道理！感谢我的父母和妻子给我的极大支持！谨以此书献给我的恩师、学友和家人，特别献给我已经开始懂事的女儿！

由于笔者水平有限，本书难免存在不少错误和不足，恳请读者批评指正！

李良炎
2008年4月9日

目　录

前　言…… I
1　句法标注模型(STM) …… 1
1.1　句法标注一般模型(STGM) …… 1
1.2　短语结构语法标注模型(PSGTM) …… 5
1.3　依存语法标注模型(DGTM) …… 8
1.4　改进 DGTM …… 11
1.5　整体改进思路…… 16
1.6　小结…… 18
2　信息依存模型(IDM) …… 19
2.1　意义与图式…… 20
2.2　事物与关系…… 23
2.3　信息…… 26
2.4　信息依存模型…… 29
2.5　信息依存图…… 32
2.6　知识信息与经验…… 37
2.7　小结…… 43
3　信息依存符号模型(IDSM) …… 44
3.1　媒介与媒义…… 45
3.2　媒映与媒符…… 47
3.3　记符与符号…… 51
3.4　记符系统与符号系统…… 54
3.5　信息依存符号模型…… 55
3.6　符用过程及其预设与优化…… 62

3.7 小结 …… 72
4 信息依存语言模型(IDLM) …… 73
4.1 语言 …… 73
4.2 语言单位 …… 77
4.3 标准句素 …… 80
4.4 衍生句素 …… 83
4.5 复杂句 …… 97
4.6 语用优化 …… 101
4.7 语用不确定性 …… 109
4.8 小结 …… 120
5 语义解释综合策略(SICS) …… 122
5.1 语义范畴相对性 …… 123
5.2 语义解释基本策略 …… 127
5.3 自然语言解释策略 …… 130
5.4 人工语言解释策略 …… 134
5.5 语义解释综合策略 …… 138
5.6 基于智能模型的语义解释 …… 146
5.7 小结 …… 153
6 信息依存智能模型(IDIM) …… 155
6.1 意识 …… 156
6.2 意识结构 …… 160
6.3 意识要素 …… 163
6.4 智能模型 …… 172
6.5 智能模型实现 …… 177
6.6 主体生存与学习 …… 184
6.7 小结 …… 195
7 信息依存句法标注模型(IDSTM) …… 197
7.1 IDSTM 的参数特征 …… 197

7.2 IDSTM 的实现步骤 …… 203
7.3 信息依存标记语言(IDML) …… 207
7.4 对象模型标记语言(OMML) …… 216
7.5 一个 IDSTM 的 OMML 简单示例 …… 225
7.6 小结 …… 229
8 总结与展望 …… 230

参考文献…… 232
术语索引…… 237
符号索引…… 254

此为试读,需要完整PDF请访问: www.ertongbook.com

图 表 目 录

图 1-1 PSGTM 示例 …… 7
图 1-2 DGTM 示例 …… 11
图 1-3 句 1-4 的 DGTM …… 13
图 1-4 句 1-4 的改进 DGTM …… 15
图 2-1 苹果与桌子 …… 21
图 2-2 苹果与桌子的空间关系 …… 26
图 2-3 前向信息与后向信息的 IDG …… 32
图 2-4 组合信息 IDG 的构造方法 …… 33
图 2-5 组合信息 IDG …… 33
图 2-6 完整信息 IDG 的初步构造 …… 34
图 2-7 完整信息 IDG 的三种改造方法 …… 35
图 2-8 复杂信息 IDG 中的信息接口过载 …… 36
图 2-9 复杂信息 IDG 的改造 …… 36
图 2-10 含多个延迟联通的 IDG …… 37
图 3-1 媒介及其相关意义的映射 …… 49
图 3-2 记号、信号及其意义的纵向组合与横向映射 …… 53
图 3-3 记映结构简化为记映序列 …… 58
图 3-4 基于 S'DM 的 SDM 构造 …… 61
图 3-5 SDM 投影为 SSM …… 62
图 3-6 符用过程 …… 63
图 4-1 $h_0\langle[h_r\rangle h_1]$的位移句素 …… 92
图 4-2 $[h_0\langle h_r]\rangle h_1$的位移句素 …… 93
图 4-3 合并句素的位移句素 …… 93

图 4-4 $h_0\langle[h_r\rangle h_1]$的位移句素简化 …… 95
图 4-5 $[h_0\langle h_r]\rangle h_1$的位移句素简化 …… 96
图 4-6 合并句素的位移句素简化 …… 96
图 4-7 句素接口过载 …… 99
图 4-8 句素组合前的位移复杂句示例 …… 101
图 4-9 句素组合后的位移复杂句示例 …… 101
图 5-1 语义再现 …… 128
图 5-2 语言解释 …… 129
图 5-3 简单智能模型的计算机实现 …… 149
图 6-1 意识结构 …… 162
图 6-2 实体运动的时间与空间测量 …… 178
图 6-3 神经元结构示意图 …… 183
图 6-4 易卦符号模型 …… 191
图 6-5 以土为主体的五行符号模型 …… 193
图 7-1 IDSTM 示例 …… 202
图 7-2 IDSTM 动态并行实现 …… 206

表 1-1 PSGTM 与 DGTM 的参数异同 …… 5
表 7-1 IDSTM 的参数 …… 197

1 句法标注模型(STM)

句法标注 (Syntax Tagging, ST) 是以一定的语法理论为指导，将句法结构形式化，便于计算机处理。**短语结构语法** (Phrase Structure Grammar, PSG) 和**依存语法** (Dependency Grammar, DG) 是现有句法标注的两种基本理论。**句法标注模型** (Syntax Tagging Model, STM) 是句法标注的形式化描述。基于 PSG 的句法标注模型称为**短语结构语法标注模型** (PSG-based Tagging Model, PSGTM)，基于 DG 的句法标注模型称为**依存语法标注模型** (DG-based Tagging Model, DGTM)。本章提出**句法标注一般模型** (Syntax Tagging General Model, STGM)，统一描述 PSGTM 与 DGTM，并分析两种模型的参数异同和各自局限。在此基础上，结合**认知语法** (Cognitive Grammar, CG) 的相关理论，提出 DGTM 的改进思路。

1.1 句法标注一般模型 (STGM)

句法标注通常是在词法标注之后进行的。词法标注包括分词、词结构标注、词性标注、词义标注等。词性标注和词义标注统称为**词类标注**(lexical category tagging)①。分词和词结构标注相对句法标注的独立性较强，词类标注则与句法标注的关系十分密切。“目前许多句法赋码系统以词类赋码系统的输出为输入……这在一定程度上隔离了词语和句法的关系”②(笔者注：赋码即

① 本书特别界定，词类标注可以是词性标注或词义标注。

② 杨惠中. 语料库语言学导论. 上海: 上海外语教育出版社、2002: 151

标注）。因此，有必要将词类标注纳入句法标注整体考虑①。

输入一个句子，进行分词处理后就可以进行句法标注了。预期得到一个带词类标注的，词语之间具有纵向聚合关系或横向配合关系的句法结构，一般可以形式化为树结构。PSGTM 侧重于词语之间的纵向聚合关系，DGTM 侧重于词语之间的横向配合关系，然而其句法标注基本过程异曲同工。可以用 STGM 统一描述 PSGTM 与 DGTM，并用 STGM 的不同参数描述其差异，从而更好地理解 PSGTM 和 DGTM 的异同，探索合理的改进思路。

STGM 的描述②如下：

输入：词串 $W=\{w_1, \cdots, w_m\}$，概念集 $C=\{c_1, \cdots, c_n\}$，知识库 $K=\{k_1, \cdots, k_l\}=R\cup E$，**中心原则** (Head Principle, HP) $R_{\text{HEAD}} \in R$。R 为规则库，E 为实例库。m 为词数，n 为概念数，l 为知识库容量。

标注：以人工标注辅以计算机校验方式，或计算机标注辅以人工校验方式，根据 W、C、K 构造句法结构网络 G。

输出：句法结构网络 $G=\{V, D, C^V, C^W, C^D\}$。结点集 $V=\{v_1, \cdots, v_i\}=W\cup C^V$，边集 $D=\{d_1, \cdots, d_j\}$，词类标注 $C^W=\{c_{w1}, \cdots, c_{wm}\}$，关系标注 $C^D=\{c_{d1}, \cdots, c_{dj}\}$，$C^V \in C$，$C^W \in C$，$C^D \in C$。$i$ 为结点数，j 为边数，m 为词数。

词串 W 是对句子进行分词处理的结果。词是能够独立运用且具有语义的最小语言单位。概念集 C 是根据特定的句法理论预设的一套语法或语义概念体系。知识库 K 是用以限定词语之间、概念之间或词语与概念之间关系的知识，是句法标注的根本依据，包括规则库 R 和实例库 E。规则库 R 是根据特定的句法

① 本书特别界定，句法标注包含词类标注。

② 李良炎，何中市. 句法标注的一般模型与参数分析. 计算机科学，34(11): 189-192

理论制定的句法标注规则。实例库 E 是已经完成句法标注的句子实例集。中心原则 R_{HEAD} 规定句法结构的层次关系。例如，DG 根据**谓语中心原则** (predicate head principle) 规定，主语、宾语、状语等成分从属于谓语成分。

人工标注即由人根据对规则库 R 的理解辅以个人语感进行句法标注，适于初期小规模处理。人工校验是由人对句法标注结果进行检查和修正。计算机标注即**句法分析** (Syntax Analysis, SA)，由计算机根据规则库 R 或实例库 E 自动进行句法标注，适于后期大规模处理。计算机校验是由计算机根据规则库 R 和实例库 E 对句法标注结果进行检查，发现可能存在的错误并向人报告。SA 目前已有相对成熟的技术，可采取基于 R 的**规则技术** (Rule-based Technique, RT) 、基于 E 的**统计技术** (Statistic-based Technique, ST) 、基于 E 的**实例技术** (Example-based Technique, ET) 或**综合技术** (Comprehensive Technique, CT) 来实现。RT 需要规则化的专家知识，ST 需要以 E 为训练集计算统计模型，ET 需要计算实例的相似度，CT 可以采取 RT、ST、ET 中两种或三种技术的综合。由于 SA 的处理结果在处理大规模语料时不可能绝对正确，为了保证质量，必须辅以人工校验。但人工标注或人工校验也难以保证完美无缺，还存在效率低的缺点，通常只能抽样校验。因此在大规模语料标注时，一般还是采取计算机校验来保证个体前后标注一致性和团体标注整体一致性。

在句法结构网络 G 中：结点集 V 包括词语结点集 W 和概念结点集 C^V，W 是由词语构成的结点集，C^V 是由概念构成的结点集；边集 D 包括由结点集 V 构成的关系 $\{<v1_{d1}, v2_{d1}>, \cdots, <v1_{dj}, v2_{dj}>\}$；词类标注 C^W 是词集 W 中每个词所属概念构成的集合；关系标注 C^D 是边集 D 中每条边对应的结点关系所属概念构成的集合；C^V、C^W、C^D 均来自概念集 C。

一个具体的句法标注模型可以用特定参数的 STGM 来描

述。STGM 的基本参数包括：

(1) 概念集 C：C 的元素个数为 $|C|$，是句法知识粒度的一个粗糙量化指标。相对来说，$|C|$越大，句法知识粒度越精细。

(2) 知识库 K、R、E：K 的元素个数为 $|K|$，是知识库完备性的一个粗糙量化指标。相对来说，$|K|$越大，知识库越完备。从构成上讲，当 E 为空时 K 为纯规则库，当 R 为空时 K 为纯实例库，当 E、R 均不为空时 K 为混合库。

(3) 句法分析 SA：可以是规则技术 RT、统计技术 ST、实例技术 ET、综合技术 CT。

(4) 概念结点集 C^V：当 C^V为空，结点集 V 只包括词语。当 C^V不为空，结点集 V 既包括词语，又包括概念。

(5) 词类标注 C^W：当 C^W不为空，存在词类标注，否则不存在词类标注。由于词类标注是句法标注中不可缺少的一环，当 C^W为空时，词类标注可能隐含于其它标注结果中。例如：在 PSGTM 中，与词语结点有直接关系的概念结点隐含了词类标注（图1-1）。

(6) 关系标注 C^D：当 C^D不为空，存在关系标注，否则不存在关系标注。关系标注并不是句法标注中不可缺少的一环，但当 C^D为空时，也可能隐含于其它标注结果中。例如：在 PSGTM 中，与词语结点没有直接关系的概念结点隐含了关系标注（图1-1）。

(7) 边集 D：D 规定 G 的网络拓扑结构。由 V、D 构成的网络通常是树结构，因为这种结构具有层次性，比较好地刻画了语言的层次特征。但理论上 G 还可以是各种图结构。目前的句法标注语料库大多采取树结构，称为**树库** (tree bank) ①②③。

---

① 周强. 汉语句法树库标注体系. 中文信息学报, 2004, 18(4): 1-8

② M. Marcus, B. Santorini, et al. Building a Large Annotated Corpus of English: the Penn Treebank. Computational Linguistics, 1993, 19(2): 313-330

③ 台湾“中央研究院”语言所中文句结构树资料库. http://turing.iis.sinica.edu.tw/treesearch/

此为试读，需要完整PDF请访问：www.ertongbook.com