中 文 信 息 学 报
第8卷 第3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.8 No.3

面向语料库标注的汉语依存体系的探讨

周 明 黄昌宁

(清华大学计算机科学系)

【摘要】实现大规模真实文本的处理，是信息化社会的迫切要求，也是国际计算语言学界的一个战略目标。目前一项迫在眉睫的任务是建立一套满足大规模真实文本处理的语言处理体系，包括分词的标准、词的分类体系、句法体系和语义体系。其中句法体系是核心环节。本文提出并论证了依存语法是合乎大规模真实文本处理要求的句法体系，并结合汉语的特点，研究了汉语的依存语法，划分了44种依存关系。最后简要讨论了依存语法的一些应用。

关键词: 汉语、依存语法、语料库语言学

一. 大规模真实文本处理迫切需要建立一套语言标注体系

近年来，国际计算语言学界围绕其战略目标及相应的理论、方法问题展开了热烈的讨论[1]，1990年8月在赫尔辛基举行的第13届国际计算语言学大会（即Coling'90)会前讲座的主题是:"处理大规模真实文本的理论、方法和工具"，明确地提出了计算语言学今后一个时期的战略目标。此外，1992年6月在蒙特利尔举行的第四届机器翻译的理论和方法国际会议（即TMI-92）的主题是:"机器翻译中的经验主义和唯理主义方法"，所谓唯理主义是指以生成语言学理论为基础的方法，所谓经验主义则是指以大规模语料库的分析为基础的方法。由此可见，基于语料库的语言研究（或称语料库语言学）已经成为计算语言学研究的重点。

为了实现"大规模真实文本处理"这一战略目标，各国学者均十分强调语料库的作用。这是因为，从"大规模"和"真实文本"两个角度去观察，语料库是最理想的语言知识资源。然而，要使语料库成为名副其实的语言知识库，就必须对库存语料进行从词法、句法、语义等各种层次上的加工。这一步骤使语料由"生"变"熟"，才能使知识获取成为可能。所以，语料加工的理论、方法和工具是目前学术界的关注焦点[2]。

以汉语为例，语料的加工由浅入深，按图1所示的流程进行:

图1中，设有分词标注、词性标注、依存标注、格关系标注等四个标注模块，语料的处理深度依次增加。注意，在语料标注的每一级，均应有一个知识获取模块，该模块从

①本文1993年6月22日收到

(1) 标注模块的人工操作（如遇到歧义时，人工排歧）中，获取知识，(2) 从标注过的语料中获取知识。所获取的知识，又反馈到标注模块，以提高标注的自动化程度和一致性。

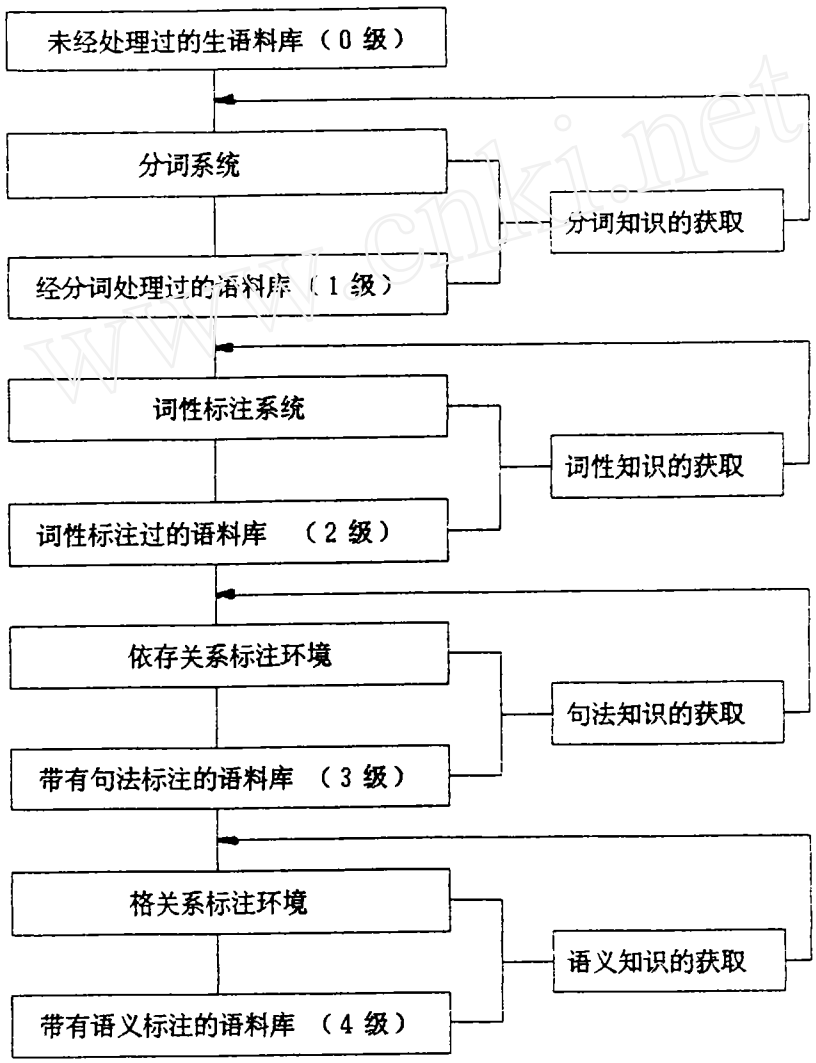


图 1 语料库的加工和知识的获取

对汉语来讲，未经分词处理的原始语料（即"生"语料）只能用来进行字频等简单统计，如果要生成词表有或进行词频统计，就必须给语料加上分词标记。进而为了实现词性自动标注，就需要进一步给分过词的语料加上词性标记，以便能够获得任意两个相邻标记的同现频率，以支持基于统计的词性标注方法[3][4][5]，或者获取确定词性的规则等等。

如果进一步，想从语料库中抽取句法知识，包括词汇之间的修饰关系、搭配关系，支配与被支配关系等，就必须给语料库加上句法标记，这时，语料库就成为一个树库，所谓树库就是一组句子的集合，其中，每个句子均被标记了句法成分。

如果树库覆盖了较大的语言范围，那么从中抽取的语言规律的覆盖面更宽，因而树库

可以支持许多研究。例如，英国 Lancaster 大学用了五年时间，采用"骨架分析法"标注了一个二百万词的语料的句法关系（选自 LONDONLUND 语料库），目的之一就是获取概率化短语结构句法分析器的概率数字[6]。此外树库还可以用来测试已有的句法分析器的效能。可以说建立一个大规模的树库是关系到语料库语言学发展的重要一环。

显而易见，无论在词汇、词类、句法、语义的任何层次上，要进行标注，必须面对大规模真实文本的实际情况，确立一套相应的标注体系作为规范。针对汉语来讲，这些标注体系包括：(1) 分词规范；(2) 词类划分标准；(3) 适宜的语法模型；(4) 适宜的语义表示方法。其中语法体系是核心环节，也是本文要重点讨论的内容。

考虑到语料库语言学的需要，我们认为选择一个合适的语法体系要满足以下 6 方面的要求：

(1) 应能做到真实、客观、全面地描写语法现象，而不是简单的就事论事；

(2) 应保证标注体系的权威性，从而也保证了各个研究单位间资源的共享；

(3) 要最大地节省所占的空间；

(4) 可操作性强，以方便人工进行标注；

(5) 利于将要进行的知识获取；

(6) 易于转换成其它语法体系的表示，并能方便地生成语义表示。

通过考察现有的语法体系，如短语结构文法、词汇功能文法、功能合一文法、依存语法等，我们认为依存语法比较利于满足以上的要求，针对汉语的特点，设立了 44 种依存关系，从而初步确立了一套汉语句法标注体系，最后我们讨论了该标注体系的一些具体应用。

二. 依存语法适宜作为语料库语言学研究的语法体系

2.1 依存语法介绍

依存语法是法国著名语言学家 Tesniere. L,在其所著的《结构句法基础》(1959 年) 提出的[7]。他主张主要动词作为一个句子的中心，支配其它成分，而它本身不受任何其它成分支配。此后 1970 年，Robinson J.J.提出了依存关系的四大公理，为依存语法奠定了基础，这四条公理是：

(1) 一个句子中只有一个成分是独立的；

(2) 其它成分直接依存于某一成分；

(3) 任何一个成分都不能依存于两个或两个以上的成分；

(4) 如果 A 成分直接依存于 B 成分，而 C 成分在句子中位于 A 和 B 之间的话，那么，C 或者直接依存于 A，或者直接依存于 B，或者直接依存于 A 和 B 之间的某一成分。

在此之后，Anderson, J.A. Hudson, R.A., Melcuk, I.A., 冯玉德美相继发表许多文章，使之逐步走向实用。

依存语法描述的是句子中词与词之间直接的句法关系。这种句法关系是有方向的，通常是一个词支配另一个词，或者说，一个词受另一个词支配，这种支配和被支配关系体现了词在句中的关系。

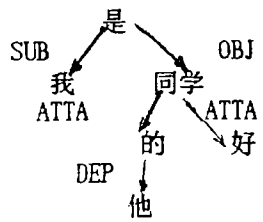
利用依存语法进行句法分析得到的是一棵依存树，简称 DRT (Depe ndency Relation Tree)，它与短语结构语法得到的句法结构树，简称为 PST (Phrase Structure Tree)有两方面的区别：首先在 DRT 中不含非终结结点，只有由具体词构成的终结结点；而 PST 中既含终结结点，又含非终结结点（如 NP，VP 等）。其次，从分枝上看，DRT 的父子关系表示相应的两个词之间的关系，且 DRT 是无序的，即这种关系只反映支配者和被支配者的关系，与相对位置关系较弱。而 PST 上的分枝表示子结点是父结点的组成部分，那些子结点是有序的，因此 DRT 偏重于关系结构，而 PST 则侧重于组成结构。

下面给出依存语法的几个基本概念。

(1) 词与词之间存在的支配和被支配的关系称作依存关系。

(2) 若句中任意两词 v，w 存在依存关系，设为 r，且 v 支配 w，则 v 是 w 的主词，记做 v>>w。w 是 v 的从词，记做 w<<v。r 称做 w 的向上依存关系，而 r 也被称为 v 的向下依存关系。

(3) 依存网络 G：一个边上带权的有向图 (V，E)，其中，V 是顶点人集合，任意 v∈V，对应着句中的一个词，E 是边的集合，如果词汇 v，w 之间存在依存关系，设为 r，而且 v>>w，则在 G 中，存在一条从 v 到 w 的有向边 e，e 上的标记为 r。



"是"是全句的中心词，具有两个向下依存关系，SUB 和 OBJ，"我"和"同学"的主词是"是"。其中，SUB--主语关系，OBJ--谓语关系，ATTA--定语关系，DEP--"的"字结构。

图 2　依存分析树(共 6 个结点，5 条边)

2.2　依存语法体现了当代语言学研究的主要倾向

目前流行的文法，如管辖一约束理论，扩充短语结构语法理论，定子句语法，词汇功能语法，功能合一语法都具有如下几个共性[8]：

(1) 它们均考虑了依存关系，也都使用了"中心词"或"支配者"的概念；

(2) 它们均重视句法的功能方面，重视句法角色的表达；

(3) 它们均认为句法应由词汇限制，而把大量句法信息放于词汇描写中；

(4) 它们均基于合一原则建立分析算法，在句法体系中，复杂特征集起到很重要作用。

而依存语法恰好体现了这些特性。

在依存语法中，树是用(a,b,r)的三元组表示的，其中 a、b 为词汇单元，r 是 a 和 b 之间的有向弧，它由 a 指向 b，表示 a 和 b 之间的直接关系，其中 a 是支配者，b 是依存者。依存语法认为这种不对称体现了自然语言的实际情况。只有通过这种不对称，才能有效地表达自然语言表达式的结构。但是短语结构却难以直接表示这种直接成分之间的不对称关系。此外，短语结构无法表达成分结构的中心词及其作用，因此依存性是表达句法结构的必要手段。

针对原有的依存语法没有表达复杂特征集，没有引入合一运算的不足，Hellwig 于 1986 年提出"依存合一语法"[9]，扩充并完善了依存语法。三元组(a,b,r)中，元素 a 和 b

均可采用复杂特征描写，每一特征用属性一值对表示。如“The cat like fish”，可以表示如下：

(ILLOCUTION: assertion: clse type <1>
    (PREDICATE: like:verb fin <1> num <1> per <1>
(SUBJECT: cat; noun num <1> per <3>
    (DETERMINER: the:dete))
(OBJECT: fish: noun)));

上述表达式一个优点是将功能、词汇、句法特征集成在一起表达，如果在整 LFG 框架中，就是将 f-结构和 c-结构合成在一起了。显然，在短语结构文法中不易做到这一点。

在基本词典中，存贮了词汇的复杂特征描写，如：

CAT → (*: cat: noun. num <1> per <3>);

likes → (*: like: verb num <1> per <3>);

在配价词典中，则表示述语动词的配价关系，如

(*: like: verb fin <1> num <1> per <3>
    (SUBJECT: −: noun num <1> per <3> adj <1>)
    (OBJECT: −: noun seq <2>));

这样句法与词汇集成在一起。由于 DUG 把复杂句法分解为原子句法关系，使得配价易于描述，相互之间牵连也较小，语法融于词汇，而词汇结构却很简明。在分析过程中，先查基本词典，得到句中每一词的词汇和句法特性；(2) 查配价词典，得到单元之间的组合能力；(3) 由自底向上地利用槽-填充机制得到句子句法关系。

2.3 满足大规模文本的标注

(1) 表示简洁

依存语法生成的句法树不含非终结符，一个具有 n 个词的句子之句法树只有 n 个结点和 (n−1) 条边。而利用短语结构类文法得到的句法树由于含有非终结符，结点数大大超过 n，大大超过 (n−1) 条边。例如：“我是他的好同学”的依存语法表示见图 2 所示，而短语结构文法的表示如图 3 所示：

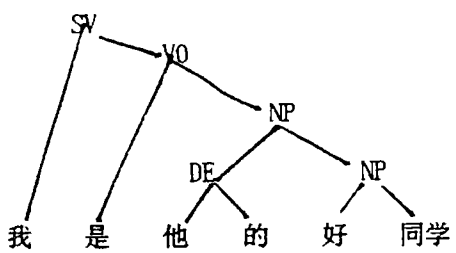


图 3 短语结构表示 (共 10 个结点,9 条边)

从以上可以看出，用依存语法标注句法树的结点数和边数要少得多，也就是说，使用依存语法可以节省分析树所占的存储空间。这对大规模语料的句法标注是一个极为重要的因素。

(2) 由于依存语法注重句子中词与词之间的关系，所以词汇知识、词汇之间的句法关系的获取较为直接。

(3) 注重语言成分之间的外部联系，强调了各成分之间存在的功能关系，所以较容易将依存关系影射为相应的语义表示，方便了今后要进行的语义分析。

例如，对"我是他的好同学"，图 2 给出了该句的依存句法树，图 4 是对应的语义网络表示，可以看出，二者之间存在着一种较为简单的对应关系，即：每一依存关系只对应一种语义关系，每一语义关系只对应一种依存关系。

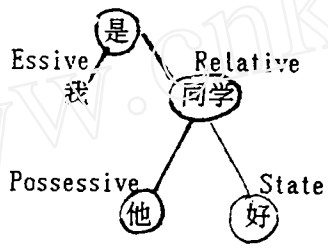


图 4　图一所对应的语义网络

(4) 采取中心词驱动，突出了中心词在句法语义上的中心作用，为深化分析创造了条件。而在短语结构语法中，主要是结构驱动，中心词的作用不明显。

三. 汉语依存语法体系

国内外对汉语依存语法体系的研究很少，主要有 (1) 中国社会科学院语言研究所李维等，他们为荷兰 DLT 机器翻译系统设计过一套汉语依存体系，当时，设计了 36 种依存关系；(2) 清华大学计算机科学系黄昌宁、苑春法等曾设立了 32 种依存关系；(3) 清华大学计算机科学系黄昌宁、吴升曾设计了　种依存关系[9]；(4) 我们曾设计了 106 种依存关系[10][11]。那么，在汉语中，到底应该设立多少种依存关系？

要想把汉语的复杂的依存关系描述清楚，需要建立许多依存关系。例如按照谓语成份，可以把主语分为系动词的主语、及物动词的主语、…等等，甚至还可以更细。这样做虽然可以全面、细致地描述汉语的复杂现象，满足了完备性、准确性的要求，但是由于依存关系集如果过于庞大，处理起来代价太高，又不满足经济性的原则。此外，对大规模的真实文本来讲，会加重统计数据的稀疏，另外会使得分析器的覆盖面窄，鲁棒性下降。因此依存关系的划分，必须要在描写的深度、精度与分析器的覆盖面的宽度上做一个必要的折衷，使依存体系既能完整、全面地描写汉语的语言现象，有利于大规模真实文本的处理。

在这一点上，我们深有体会，开始，我们曾经设立了 106 种依存关系[10][11]，划分得很细，例如，把主语、宾语、状语、定语、补语等关系均按照主词和从词的语法类型加以细分。如定语关系的细分：

1　atta-com　一般定语

2　atta-dep　"的"字结构作定语

3 atta-quap 数量词结构作定语
4 atta-pp 介词结构作定语
5 atta-cpp 方位结构作定语
6 atta-dpp 搭配结构作定语
7 atta-dig 数词结构作定语

这样，虽然描写得很细致，但是也带来一些问题。

(1) 类分得太细，导致在标注时，要仔细推敲，降低了操作性，影响了标注的效率，另外，由于每个人对如此细致的分类的理解经常会有差异，即使同一个人在不同时刻的理解也有偏差，所以，不可避免地带来了标注的严重不一致性。

(2) 在语料库的规模一定时，类分的越细，带来的统计数据的稀疏问题越严重。

(3) 类分得太细，分析器的分析正确率会大受影响，分析器的适应面和鲁棒性也受到很大影响。

但是，如果类划分得过粗，比如，仅划分“主、谓、宾、定、状、补”等几种依存关系，又无法描写汉语中一些常见的句法关系。例如：数+量，名+“们”等。所以，依存关系的划分，要做必要的折衷。

本文在以上研究基础上，结合汉语料库标注实例，重新做了划分，包括合并相近的依存关系等。目前设计了 44 类，现将分类体系说明如下：

(1) 谓语 GOV

谓语是全句的中心词，它一般就是动词，或名词，少数情况下也可以是名词。它没有支配词，为统一起见，专设一个虚拟词“*”支配它，“*”与谓语词之间的关系为 GOV。

(2) 主谓 SUBJ

在汉语句子中，位于谓语之前，作为谓语的主体的词称为主语。作主语的词可以是名词、代词、动词、“的”字结构、数词、数量结构、联合结构等。主语与谓语之间的关系是 SUBJ。

主词类型	从词类型	举例
V	N	中国 人民 站 起来 了。
V	R	我们 要 建立 新 中国。
A	V	学习 语言 重要。
A	Q	一 个 就 行。
V	USDE	做 工 的 走 了。
V	M	十 是 五 的 二 倍。
T	T	今天 星期三。
N	N	这 张 桌子 三条腿。
USDE	N	天 漆黑 漆黑 的。
V	CP	小王 和 小张 是 学生。

(3) 宾语 OBJ

在汉语句子中，位于谓语之后，作为谓语的动作对象的词称为宾语。作宾语的词可以是名词、代词、动词、“的”字结构、数词、数量结构等。宾语与谓语之间的关系是 OBJ。

主词类型	从词类型	举例
V	N	我 爱 祖国。
V	V	这 孩子 不 爱 读 书。
V	Q	我 要 五 个。
V	DE	这 本 书 不 是 我 写 的。
V	A	我 喜欢 勇敢。

(4) 定语 ATTA

名词短语中，如有修饰成份，则修饰成分位于名词之前。修饰成分与名词之间的依存关系为 ATTA。名词是支配词。修饰成分可以是“的”字结构、介词短语、形容词或名词。例如：

主词类型	从词类型	举例
N	DE	我 的 祖国，金色 的 海。
	A	好 人；聪明 孩子。
	N	学生 宿舍。
	V	学习 计划。
	Q	三 个 人。
	R	这些 家伙。
	R	哪些 人。
	R	我们 国家。
	R	什么 东西。

(5) 状语 ADVA

状语是修饰动词的成分，一般位于动词之前，这些成分可以是“地”字结构、介词短语、方位词短语、副词、形容词及表示时间、地点、处所的名词。这些词与谓语之间的依存关系是 ADVA。例如：

主词类型	从词类型	举例
V	USDI	紧张 地 工作。
V	P	他 在 上海 上学。
V	F	台 上 坐 着 主席团。
A	D	它 仍 很 健康。
V	A	勤奋 学习，英勇 斗争。
V	T	我 1960 年 出生。
V	V	笑 着 说。
A	Q	这 次 完 了。
V	P	为 人民 奋斗 终身。

(6) 补语 COMP

补语是对中心词做补充说明的成分，往往表示结果、趋向、可能、程序及时间等。这些成分可以是"得"字结构、介词短语、方位词短语、副词、趋向词形容词及表示时间、地点、处所的名词。这些词和中心词之间的依存关系是 COMP。例如:

主词类型	从词类型	举例
V	DE3	一 句 话 逗 得 大家 都 乐 了。
V	Q	看 一 下，输 了 三 次。
	VC	做 完 了 作业。
	A	洗 干净 手。
	P	努力 学习 以 报效 祖国。

(7) 第二宾语（又称间接宾语）OBJ2

在句子中，如果谓语是动词且动作对象有二个，一个对象为直接宾语，它与谓语的关系仍沿用 OBJ 表示，另一个对象为间接宾语，它们谓语的关系用 OBJ2 表示。

主词类型	从词类型	举例
V	R	我 送 你 一 本 书。

(8) 兼语 PIVT

有一些词既是谓语的宾语，又充当下一层次的主语，这些成分称为兼语。兼语与谓语之间的关系记做 PIVT。

主词类型	从词类型	举例
V	R	我们 选 他 当 班长。

(9) 兼语补语 SOC

当句中出现兼语时，以兼语为主体的动词与全句谓语之间的依存关系记做 SOC。

主词类型	从词类型	举例
V	V	我们 选 他 当 班长。
V	A	我们 喜欢 他 老实。

(10) 主题 TOP

当主谓结构作谓语时，我们将大主语称为主题，记做 TOP，而将主谓结构中的主语部分称为句子主语。例如:

主词类型	从词类型	举例
V	R	我 头 疼。
A	N	这 本 书 价值 不 大。
A	R	这 一 位 身体 健康。
V	N	鸡 我 吃 了。

(11) 同位语 EPA

在名词短语中，做中心词的名词与同位词之间的依存关系为 EPA，其中同位词为从属词。

主词类型	从词类型	举例
N	N	北京，中国 的 政治 中心。

(12) 连动 VA

当句子的中心词是动词时，如果其后还有其它动词表示一连串发生的动作，则中心词与这些词之间的依存关系为 VA，中心词为支配词。

主词类型	从词类型	举例
V	V	他 推门 进来 通知 我们。
V	V	我 有 句 话 要 说。

(13) 并列关系 COOR

包含几种情况：(1) 动词与动词并列，(2) 形容词与形容词并列，(3) 名词与名词并列；设 A1，A2…An 为并列成份，则将 A1 定为中心词，其它词与 A1 的依存关系均是 COOR，例如：

主词类型	从词类型	举例
V	V	宣传 教育 为 主。
A	A	美丽 鲜艳 的 花朵。
N	N	工厂 学校 要 做 好 卫生。

(14) 句尾语气 ZYQA

语气助词，“了”、“啊”、“吗”、“呢”、“哪”、“来”、“去”……等在句尾表示各种语气，他们和谓语的关系记做 ZYQA。例如：

主词类型	从词类型	举例
V	Y	他 原来 是 大学 毕业 呀。

(15) 标点符号 MARK

标点符号与其所属层次的中心词之间的关系，记做 MARK。

(1) 北京、上海、和 天津 是 中国 的 三个 直辖市。

(2) 他 原来 是 大学 毕业 呀！

(3) 他 从 北京 来 了。

(16) 时态 TENS

在汉语中，表示时态的助词有“着”、“了”、“过”。这些词在句中一般紧接着述语词。这些词和谓词之间的依存关系是 TENS。例如：

主词类型	从词类型	举例
V	UT	我 拿 着 书。

(17) 复数 PLU

名词与表示复数“们”之间的依存关系为 PLU，其中名词是支配词。

主词类型	从词类型	举例
N	们	工人 们

(18) 限量 LA

量词与修饰它的数词或定冠词或另一个量词之间的依存关系是 LA。此时量词为支配词。例如:

主词类型	从词类型	举例
Q	M	一 本 书。
Q	R	这 本 书。

(19) 连数 SA

数词之间依存关系是 SA。这些数词可以是基数词、准数词、疑问数词和特殊数词“半”。位置在前的词支配位置在后的词。

主词类型	从词类型	举例
M	M	一 万 零 五 千 公里 二 点 六 三 又 五 分 之 三 百 分 之 五 点 六

(20) “第”字结构

数词和序数词如“第”之间的依存关系为 DIC。数词为支配词。例如:

第 一 万 零 五 千 公里

(21) 介词的宾语 POBJ

在介词短语中，介词为支配词，它与其支配对象之间的依存关系为 POBJ，例如:

在 北京 有 很多 名胜古迹。

(22) 方位结构的宾语 FOBJ

在方位结构中，方位词为支配词，它与其支配对象之间的依存关系为 FOBJ，例如:

都市 里 的 村庄

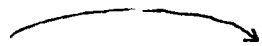
(23) 框式结构的宾语 KOBJ

在框式结构中，框式结构的尾词与框内中心词的关系为 KOBJ，框式结构的尾词为支配。例如：

在 这 篇 作品 中

(24) 框式结构的尾词 KWEI

框式结构的尾词与首词之间的依存关系为 KWEI，其中首词为支配词。例如：



在 这 篇 作品 中

(25) "的"字结构 DEP

"的"字结构中，"的"字之前的成分与"的"之间的关系为 DEP，其中"的"是支配词。

主词类型	从词类型	举例
USDE	V	我 买 的 书
USDE	A	鲜艳 的 花朵
USDE	F	桌子 上 的 茶杯
USDE	N	中国 人民 的 革命。
USDE	P	关于 时事 的 报告。

(26) "地"字结构 DIP

"地"字结构中，"地"字之前成分与"地"之间的依存关系为 DIP，其中"地"是支配词。

主词类型	从词类型	举例
USDI	D	认真 地 读书。
USDI	V	说 不 出 地 难受。
USDI	N	应该 历史 地 看待 问题。
USDI	O	扑通 扑通 地 乱跳。

(27) "得"字结构 DEIP

"得"字结构中，"得"字之后的成分与"得"之间的依存关系为 DEIP，其中"得"是支配词。例如：

主词类型	从词类型	举例
USDF	A	干 得 好。
	V	活 得 没 意思。
	VC	干 得 完。
	VC	说 得 出来。
	V	弄得 他 不知 如何 是 好。
	很	好 得 很。

(28)、(29) 并列连词（如“和”）LINKL 和 LINKR

一个句子中，当两个词由连词联结共同做句子的某种成分时，则前一词和连词之间的关系为 LNKL，前词为支配词。而该连词和后一词之间的依存关系为 LINKR，连词为支配词。例如：

我 和 你 都 是 工人。

(30) 从句 CSUB

若一个主从关系的复句含多个分句时，整个句子的谓词与各分句谓词之间的依存关系为 CSUB。当一个并列关系的复句含多个分句时，前一个分句的中心词做整个句子的中心词；并且前一个分句的中心词与各个分句的中心词之间的依存关系也为 CSUB。

因为 我们 学习 努力，所以 取得 了 好 成绩。

风 停 了，雨 住 了。

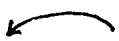
(31) 关联词 GLC

若一个复句含若干分句，每一分句的中心词与该分句的关联词之间的依存关系为 GLC。例如：

因为 我们 学习 努力，所以 取得 了 好 成绩。

(32) “每”字结构 MEI

“每字结构中”，“每”与后面的词之间依存关系记为 MEI，后面的词为支配词。例如：

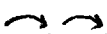


每 一 个 中国 人 都 要 有 民族 自豪感。

(33) 词的重叠形式 DUP

出现 AA、ABB、AABB、ABAB、ABA 重叠时，位置居前者支配位置居后者，二者之间的依存关系记为 PUP。例如：

欢 欢 喜 喜 (AABB)



亮 闪 闪 (ABB)

闪 闪 亮 (AAB)

敲打 敲打 (AA)

(34) 句尾语气词“等”，也好：DYH

句尾语气词“等”、“也好”与中心词之间的关系是 DYH。例如：

这里 有 大米、白面 等 粮食

(35)、(36) “之”字结构 ZHIL、ZHIR

一个句子中，当两个词由“之”字联结共同做句子的某种成分时，则前一词和“之”之间的关系为 ZHIL，前词为支配词．而“之”和后一词之间的依存关系为 ZHIR，“之”为支配词．例如：

这　是　其中　的　难点　之　一

(37) 被、受 BEI

表示被动的“被”、“受”等词与其后谓词性成分之间依存关系是 BEI，其中谓词为中心词．例：

他　被　我　狠狠　打　了　一　顿。

他　给　人　打　了。

(38) 词的前缀、后缀 SUFF

电影　儿

阿　爸

(39) 宾语提前 BAC

利用“把”、“将”等将宾语提前时，“把”与后面中心词之间的关系为 BAC，“把”是主词。

把　衣服　洗　了。

将　革命　进行　到底。

(40) 助词…一样，…似的与前面的成分之间关系为 CO，“一样”、“似的”为中心词。

风　一样　的　飞　走　了

(41) 介词结构的主语 PSUB，介词为主词。

我们　讨论　了　计算机　在　图书馆　的　应用。

(42) 形容词附加语 ADJ

如“长江　五千　公里长”，“长”和其前成分的中心词“公里”之间的依存关系是“ADJ”，其中，“长”是支配词。

(43) 数量词、数词的附加语 DIG

如“一米　多　厚”，“多”与其前的量词“米”的依存关系为 DIG，量词为支配词，又如

“十　多　个”“多”与其前的数词“十”的依存关系也为 DIG，数词为支配词。

(44) 插入语 INST，插入语与全句谓语之间关系为 INST，谓语为中心词

需要说明的是，在标注真实文本过程中，肯定会遇到新的语言现象，所以依存关系会不断有所补充。

四. 依存语法在语料库语言学研究中的应用

4.1 语料库的句法标注

前面已经介绍，依存语法表达简洁，适合于语料库的句法标注，下面给出标注样本：

句子：我是他的好同学。

标注为：

1	我	2	SUB
2	是	0	GOV
3	他	4	DE4
4	的	6	ATTA
5	好	6	ATTA
6	同学	2	OBJ
7	同学	2	MARK

可见，在标注时，只需给出每个词的支配词的序号，以及相应的依存关系，就可以表示一棵句法树。为了提高句法标注的效率，我们专门设计了一个交互式的人机互助的标注环境，机器会自动学习标注的句法知识，随着标注的句子数目的增加，机器积累的知识不断增加，到一定程度时，许多句子，机器已能自动标注，当出现不能解决的歧义时，向人发出询问，人干预后，机器便继续标注。这样做，不但减轻了标注的工作量，而且保证了标注的一致性。这一标注环境本身实际上就是一个具备自学习功能的依存句法分析器。有关这一方面的内容，将另文加以介绍。

4.2 句法知识的获取

针对每一词 Wi，可从标注过的语料库中获取如下三类知识：

(1) 搭配模式：某词在一个句子中和它的主词之间的关系。

(2) 传递模式：又称垂直同现约束，指某词在一个句子中和主、从词之间在同时存在的依存关系（对）的组合，该模式反映了语言的层次性和递归性，体现了一个句子的句法结构在纵向扩展演化的规律。

(3) 配价模式：配价模式又称水平同现约束，指某词在一个句子中和所有从词之间存在的依存关系（对）的组合，该模式反映了一个句子的句法结构在横向扩展演化的规律，规定了某词在句子中出现的可能环境。

这三种知识可以具体词与词之间的关系，以词汇项来索引的，称为低层知识库。在低级知识库基础上，可进一步归结出语义-句法类的搭配模式，传递模式和配价模式，所形成的知识库称为高层知识库。

此外，还可得到两种统计数据，一是任意两种依存关系的同现概率 $P(R_i|R_j)$,,亦即依存关系的概率转移矩阵，另一个是在给定依存关系 R_1 下词汇 W_k 出现的条件概率 $P(W_k|R_i)$。统计数据将用于依存关系的自动标注。

4.3 依存句法分析的过程

运用依存句法模型，运用上面获取的知识，结合统计方法和依存语法的四个公理，就可以进行依存分析[11]。主要步骤是：

(1) 输入一个句子；

(2) 自动分词；

(3) 用统计方法标注词性；

(4) 应用规则捆绑邻接词；

(5) 在知识库中查找每一词三种知识；

(6) 建立依存网络；

(7) 通过统计方法标注每一词依存关系，简化依存网络；

(8) 应用规则作进一步的简化；

(9) 对依存网络中的树进行评价并输出。

此外，我们正在研究依存关系到格关系的转换，依存关系到短语结构的转换，有关内容将另文叙述。

参考文献

[1] 黄昌宁，关于大规模真实文本的谈话，第三届中文信息处理国际学术会议论文集，1992 年 10 月，北京。

[2] 黄昌宁，关于处理大规模真实文本的谈话，语言文字应用，1993 年第二期。

[3] De Marcken, C.G., Parsing the LOB Corpus, Proceedings of the 28th Annual Meeting of the ACL, 6–9 June 1990, Pittsburgh PA.. pp. 243–251.

[4] Meteer, M., Schwartz, R. and Weischedel, R. POST: Using Probabilities in Language Processing, Proceedings of IJCAI'91, August 1991, Australia, pp.960–965.

[5] 白栓虎，汉语自动词性标注系统的研究与实现，清华大学计算机系硕士论文，1992,3。

[6] Ceofferey Leech and Roger Garside, Running a grammar factory: The Production of Syntactically analysed corpora or "treebanks", from "English Computer Corpora" PP.15–32, Mouton de Gruyter, 1991.

[7] 冯志伟，特思尼耶尔从属关系语法，国外语言学，1983 年第一期。

[8] Peter Hellwig, Dependency Unification Grammar, Proceedings of Coling'86. 1986, Bonn.

[9] 吴升，基于语料库的汉语句法分析的研究与实现，清华大学计算机系硕士论文，1992 年。

[10] 张敏，语料库，知识获取与汉语依存分析，清华大学计算机系硕士论文，1993.3。

[11] 周明、黄昌宁、张敏、白栓虎、吴升，"统计与规则并举的汉语句法分析模型"计算机研究与发展，待发表。

附录：汉语词性分类

nf	姓氏	cm	中置连词
np	专有名词	cb	后置连词
ng	普通名词	usde	"的"
t	时间词	uszh	"之"
s	处所词	ussi	"似的"
f	方位词	usdi	"地"
vg	一般动词	usdf	"得"
va	助动词	ussu	"所"
vc	补动词	ussb	"不"
vi	系动词	ut	时态助词
vh	动词"有"	ur	其它动词
vy	动词"是"	y	语气词
vv	来去连动	o	象声词
a	形容词	e	叹词
z	状态词	h	前缀
b	区别词	k	后缀
m	数词	i	成语
q	量词	j	简称语
r	代词	l	习用语
p	介词	x	其他
d	副词	p	标点
cf	前置连词		

Approach to the Chinese Dependency Formalism For the Tagging of Corpus

Zhou Ming　Huang Changning

(Dept. of Computer Science, Tsinghua University Beijing, 100084, P.R.China)

Abstract

As a strategic target of the global computational linguistic circle, the large scale authentic text processing becomes more and more significant for the modern informative society. To meet with the requirement of large scale corpus processing such as tagging, knowledge acquition and automatic analysis, an language processing formalism must be set up as soon as possible. The formalism includes the word segmentation standard, the part of speech scheme, the syntactic formalism and semantic formalism. The syntatic formalism is the key part among them. In this paper, it is presented that Dependency grammar is a suitable syntactic formalism for large scale authentic text processing, and the Dependency grammar for Chinese is specially studied, and 44 kinds of dependency relations are defined. Finally, some application of the Dependency grammar to Chinese language processing are briefly discussed.

Key Words: Chinese, Dependency Grammar, Corpus Linguistics