前　言

计算语言学是从多种角度研究如何通过计算机来模仿人类语言处理能力，并用这种能力解决语言交流问题的学科，它的终极目标是构造一个能懂人语、会说人话、可用自然语言进行交流的机器(刘海涛等 2005，Hausser 2001)。这个定义突出了计算语言学的两个特点：理论性和实践性。前者体现在为了模仿人的语言处理能力，我们必须对这种能力有深刻的认识，而且要把这种认识上升到一定的理论层面。如果这种认识不能用精确的方式表述出来，将会影响到最终目标的实现。后者说的是，计算语言学也应该能够解决实际问题，它是一种“应用驱动”的语言学研究。计算语言学的这种特性也使得技术现实对理论框架产生反作用和限制，说起来近乎完美的理论，如果现有的技术无法实现，那么也难以解决好实际问题。

关于计算语言学和语言学理论的关系问题，我们认为以下几点值得考虑：计算语言学需要语言学理论，这种理论不仅应该能够描述真实语料，而且也能用精确方法来表述；计算语言学有着高远的目标，这种目标虽然在可预见的将来可能难以完全实现，但这绝不意味着研究者可以忘记这种目标，而只满足于一种短视的灵巧做法；计算语言学家的任务不仅仅是构建一些语言信息处理的应用系统，他们也应该有能力从(语言学)理论的角度解释此类人造系统的行为；面向计算语言学的语言学理论是一种可以通过机器来验证的理论，如受技术所限，某些思想一时无法实现，可实现部分不但应能从理论上自圆其说，而且也应有足够的扩展能力。总之，为了让计算机能够处理人类语言，我们需要一套切实可行的(形式)语言学理论。但计算语言学需要的是面向应用的语言学理论，

也就是说这种理论不仅应该能够形式化，而且也应具备足以描写真实语料的能力。

Hudson（1990：3）总结了现代语言学的一些发展路向，如词汇主义（lexicalism）、整体主义（wholism）、关系主义（relationism）、单层次主义（mono-stratalismo）、实现主义（implementationism）等。无论其他的语言学家赞成与否，Hudson 所说的这些研究方向确实在很大程度上反映了现代语言学理论的主要发展路向。依存语法就是具有这些特质的语言学理论。

依存语法具有悠久的历史。从古印度的波你尼语法、欧洲中世纪的摩迪斯泰句法理论、阿拉伯的传统语法到世界许多国家的传统语法，或多或少地都含有依存关系的思想。就现代语言学的发展来看，如果人们将乔姆斯基 1957 年的理论和他最新的“最简方案”理论进行比较，也不难发现他向依存语法研究传统中某些思想靠近的趋势。学界流行的面向应用的语言学理论，如词汇功能语法（LFG）、头词驱动的短语结构语法（HPSG），也均显现出了这种趋势。在计算语言学界，目前最好的英语句法分析器，如 Collins（1999）、Charniak（2001）等，均采用了头词（Head）等概念[4]。被公认为最好的语言形式化理论之一的树邻接语法（TAG），近年来也多次展现了自己与依存语法的亲缘性。在基于机器学习的计算语言学研究中，语言资源是极其重要的。从 1993 年美国宾州大学英语树库问世以来，世界各国的（计算）语言学家掀起了“植树造林”的热潮。十几年的研究与实践表明，树库中的树种有从短语结构向依存结构转变的趋势[5]。

依存语法在（计算）语言学领域的兴起，可能归功于这种语言学理论：更有利于自然语言处理中的某些应用领域，更便于从句法层面到语义层

4 按照 Nugues（2006：11，244）的说法，尽管依存语法不如短语结构语法流行，但实践证明，它在句子分析方面却很有效。对依存语法的这种优点，计算语言学家早有认识。例如，在世界上第一部计算语言学教材中（Hays 1967），依存语法具有不亚于短语结构语法的地位。

5 Witkam（2005：93）把这种现象称为“Francaj arboj revivas, usonaj sekiĝas”（法国的树正在复活，美国的树正在枯萎）。

面的转换，更适宜于处理自由语序的语言，具有更好的心理现实性，更易于构造基于机器学习的高精度句法分析程序等。

这些事实说明，依存语法既古老，又年轻，是一种可以解决语言分析问题的实用的语法理论。依存语法研究不但有益于计算语言学，也有助于一般的语言学研究。遗憾的是，有关依存语法理论的研究，特别是系统性的研究却不多见。造成这种局面的原因可能在于依存语法是一种开放性的理论，这种开放性给文献收集、整理和研究带来了极大的困难；二是依存语法的主要文献是用德语、法语等写成的，这又为研究增添了语言障碍；三是缺少适宜的依存语法形式化手段来研究依存语法的形式化问题；四是理论语言学研究与计算语言学实践的脱节，语言理论的研究者搞不懂计算语言学中的基本方法和手段，计算语言学研究者又忽视语言学理论的建设。所有这些困难和问题使得我们很难看到结合依存语法、配价理论和计算语言学应用的系统性研究。

本书的主要目的是，在充分了解前人有关依存关系、配价理论、依存形式化和依存句法分析方法的基础上，归纳出依存语法和配价理论的一般原理和方法，提出一套较完整的基于配价模式的依存语法分析框架，并用实验来证明这一框架的可行性。与此同时，我们也力图用本书提出的理论架构作为主线，将相关领域的主要研究成果串在一起，形成一部配价理论和依存语法研究的简史。

为了让国内读者更好地了解依存语法的一些基本思想和方法，本书在介绍其他学者的观点时，尽可能采用“引”而非“述”的方式，目的是为了更好地表现原义，减少误读率。在写作过程中，我们尽可能采用第一手的文献，所引外文资料一般均由作者自译。在计算语言学方法方面，本书对基于规则的方法和基于统计的方法都给予了足够的重视。理论求高、应用求实，是本书的基本方针。

除前言和结语外，本书共分八章，前五章构成了历史与理论部分，贯穿其中的主线是我们提出的基于配价模式的依存语法分析模型。每一

章都把泰尼埃结构句法理论中的有关内容放在参照位置上，以体现我们用信息时代语言观诠释泰尼埃理论的愿望。计算语言学的特殊性也要求本书不仅应提出理论，还应能在机器上验证所提出的理论。因此，本书的后三章是用依存语法分析汉语的实践部分。在这一部分，我们不仅采用实现了本书提出的基于配价模式的汉语句法分析，也用几种流行的基于规则的依存句法分析器进行了汉语句法分析实验，目的是为了更好地理解汉语依存句法分析的可行性与特殊性。我们也采用依存树库作为机器学习的资源，对汉语进行了基于统计的句法分析研究，此种研究不但有助于明确语言学家在这一领域的作用，而且也对影响依存句法分析精度的因素有了更多的了解。最后一章以依存树库为基础，对汉语作了一些定量分析，此种分析不仅开辟了语言(汉语)定量研究的一条新路，也为“概率配价模式”的获得提供了手段。

笔者从20年前开始系统搜集研究依存语法的文献，对于世界各国学者在此领域的研究均有一定的了解，与国际依存语法研究领域的主要学者保持着经常的学术联系。近年来，依存语法研究在国内外呈不断增长之态势，有关应用也逐渐广泛，但遗憾的是，国内仍没有系统性的专著问世。为满足教研之需要，我将自己这些年来学习研究依存语法的一些体会整理成书。希望本书不但能起到抛砖引玉的作用，也有助于加深大家对依存语法的了解和研究。

本书内容繁杂，涉及领域广泛。为了让读者对依存语法有一个较完整的了解，我们在书中使用了大量世界各国学者用多种语言发表的文献，但这也增加了出错的可能。笔者虽已尽心尽力，仍难保没有错漏。不当之处，请大家不吝赐教。

刘海涛

2009-04-15