依存语法的理论与实践

DEPENDENCY GRAMMAR
from theory to practice

刘海涛 著

科学出版社
www.sciencep.com

依存语法的理论与实践

DEPENDENCY GRAMMAR

from theory to practice

刘海涛 著

科学出版社

北京

图书在版编目(CIP)数据

依存语法的理论与实践 / 刘海涛著. —北京：科学出版社，2009

ISBN 978-7-03-024866-4

I. 依… II. 刘… III. 数理语言学－研究 IV. H087

中国版本图书馆 CIP 数据核字(2009)第 104914 号

责任编辑：郝建华 阎 莉 / 责任校对：宣 慧

责任印制：赵德静 / 封面设计：开和工作室

联系电话：010-6403 3862 电子邮箱：haojianhua@mail.sciencep.com

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

中国科学院印刷厂印刷

科学出版社编务公司排版制作

科学出版社发行 各地新华书店经销

*

2009 年 6 月第 一 版 开本：A5(890×1240)

2009 年 6 月第一次印刷 印张：10 3/4

印数：1—2 000 字数：345 000

定价：48.00 元

（如有印装质量问题，我社负责调换〈科印〉）

冯志伟序

“依存关系”和“依存语法”一直是刘海涛博士多年来关心的问题，他广泛收集了国内外有关依存语法的著作细心研读。这些著作涉及多种不同的语言，除了汉语和英语的著作之外，还有德语、法语、俄语、荷兰语、日语等外语写成的著作。为了读懂这些著作，他还学习了德语、法语等外语，具备了一定的阅读多种语言原著的能力，获取了第一手的资料。由于他尽量从原著中获取资料，他关于依存语法的这些论述，自然就具有很强的说服力。

海涛并不满足于依存语法的理论研究，他还勇于实践，建立了一定规模的依存树库，使用计算机来验证他提出的基于配价模式的句法分析理论，这就使他的研究，不仅在理论上翔实可靠，在实践上也经得住检验。

我认为海涛的这本专著具有以下三方面的特色。

第一，系统地梳理了国内外依存语法理论研究的历史，提供了内容丰富、翔实可靠的大量资料，有助于我国学术界全面地理解依存语法这一重要的理论。

20 世纪 70 年代末期，我在法国格勒诺布尔理科医科大学应用数学研究所研制汉—法/英/日/俄/德多语言机器翻译系统 FAJRA 时，为了克服乔姆斯基(N. Chomsky)的短语结构语法在汉语自动分析中的困难，我的导师沃古瓦(B. Vauquois)教授和巴黎第七大学的法国朋友格罗斯(M. Gross)教授都建议我读一读泰尼埃(L. Tesnière)的《结构句法基础》(*Eléments de Syntaxe Structurale*)，这是我第一次接触“依存语法”。我用两个多月的时间认真阅读了该书的法文版，眼界大开。在我提出的中文信息 MMT 模型(多叉多标记树模型)中，我根据依存语法，确定了汉

语句法—语义分析的“行动元”(actants)和“状态元”(circonstants)，这些表示依存关系的标记成为 MMT 模型的多标记中的最重要的标记。我还根据“价”(valence)的概念，赋予汉语动词不同的“价”标记，通过动词的“价”来控制句子的句法—语义结构，并吸收依存语法中“支配”和“被支配”的理念，给每一个短语都标注出“中心词”(标记为 gov)，给每一个句子都标注出“轴心”(标记为 pivot)，明确了短语和句子中的“支配”和“被支配”关系，大大地提高了 MMT 模型的性能。从法国回来之后，我马上就在当时的《国外语言学》杂志上写文章介绍了泰尼埃的语法理论(冯志伟 1983)[1]，这是国内最早系统介绍依存语法的文章。

当时我只注意到动词的“价”，没有注意到名词和形容词的“价”。上世纪 80 年代我有机会到了德国，访问了曼海姆的德语研究所(IDS)，才知道德国语言学界除了研究动词的“价”之外，还研究了名词的“价”和形容词的“价”。德国语言学家早就出版了《德语动词配价词典》、《德语名词配价词典》和《德语形容词配价词典》，德语研究所的托依拜特(W. Teubert)在 1979 年出版的《名词的价》，就是他在海德堡大学的博士论文。这时，我才认识到不仅仅动词有“价”，名词和形容词也有“价”，我开始把依存关系看成是普遍存在于语言中的一种支配与被支配的关系，并且深深感到自己过去对于依存语法的学术发展的情况实在是了解得太少了。

近年来，我国出现了研究配价理论和依存语法的热潮，在具体的细节研究方面取得了很好的成果，但是，在系统性和整体性方面还显得很不足。

海涛在这本书中，根据大量的历史资料，为我们勾画出了依存语法和依存关系的观念发展的历史脉络，告诉了我们如下鲜为人知的重要历史事实：

1 当时我把 Tesnière 翻译成特斯尼耶尔，不妥，按照法语人名的规范读音，最好翻译为泰尼埃。

早在12世纪，语言学家赫利亚斯(Petrus Helias)在他的著作中就提出了“动词中心说”，他认为，动词要求的句子成分的数量是不同的，动词的必有成分一般是指名词性的，这些成分是构造一个 perfectio constructionis 所必需的。这种“动词中心说”指出了动词对于句子成分的要求，已经隐含了“配价”的理念。

德国普通语言学家梅纳尔(Johann Werner Meiner)在1781年的著作里就明确将谓语(动词)分为：一价动词，二价动词和三价动词，只不过他没有直接使用“价”这个词，而是用了一个德语词“seitig-unselbständig”，但是其实质已基本无异于现代人定义动词“配价”的说法了。

1934年，奥地利语言学家卡尔·比勒(Karl Bühler)在其《语言理论》中说，“每种语言中都存在着选择亲缘性；副词寻找自己的动词，别的词也是如此。换言之，某一词类中的词在自己周围辟开一个或几个空位，这些空位必须由其他类型的词来填补。”卡尔·比勒关于“空位”的见解，揭示了“配价”的本质。虽然他没有使用过“配价”这个词，但是国外研究配价理论的学者普遍将卡尔·比勒看做是配价理论研究的先驱。

1948年，苏联语言学家科茨年松(Kacnel'son)首次提出“配价”这个术语。他说，“在每一种语言中，完整有效的具体化的词不是简单的词，而是带有具体句法潜力的词，这种潜力使得词只能在严格限定的方式下应用，语言中语法关系的发展阶段预定了这种方式。词在句中以一定的方式出现以及与其他词组合的这种特性，我们可以称之为句法配价。”

科茨年松特别强调“配价”的“潜在性”。他认为，明显的语法范畴、功能和关系是“通过句法形态来表现的”，而在词的句法组配和语义中隐含了潜在的语法范畴、功能和关系。他还说，“语法如同一座冰山，绝大部分是在水下的。”

1949年，荷兰语言学家格罗特(A. W. de Groot)在他用荷兰语出版的《结构句法》(*Structurale Syntaxis*)一书中也使用了“配价”这一概念。格罗特在他的书中写道，“与其他词类相比，某些词类的运用可能性受到

限制，即词类具有不同的句法配价。配价是被其他词所限定或限定其他词的可能性或不可能性。”他在句法研究中使用了“valentie”和“syntactische valentie”这两个术语。

这些鲜为人知的历史资料零星地隐藏在国外文献的汪洋大海之中，互联网上也不容易搜索到，获取它们犹如大海捞针，而且，这些文献涉及德文、俄文和荷兰文等不同的外文，要读懂它们需要丰富的外语知识，特别是格罗特的《结构句法》(*Structurale Syntaxis*)一书，尽管过去我曾有所耳闻，但由于是荷兰语写的，一直不敢问津，海涛在外国朋友的帮助下，弄清楚了荷兰文的原意，使得我们有机会了解到这本重要文献的内容。海涛做的这些钩沉探源的工作是非常有意义的。

我一再对我的博士研究生们说，“治史须读原著”，鼓励他们尽量阅读外文原著。我们绝对不能仅靠翻译成中文的材料来研究国外语言学的历史，因为可能会出现翻译错误，因错就错；我们更不能仅靠道听途说的资料来研究国外语言学的历史，因为可能会以讹传讹，谬种流传。因此，阅读原汁原味的外文原著是非常重要的，特别是对于研究语言学的博士生来说，我认为是必不可少的。既然是博士，就不能像其他人那样只依靠中文译文来进行研究，既然是语言学的博士，就应当在外文水平方面胜人一筹。海涛坚持了“治史须读原著”这个原则，严格要求自己，进行了不懈的探索，终于揭开了语言学历史上这些鲜为人知的事实的神秘面纱，这是令人高兴的。

第二，全面地讨论了依存语法的形式化方法和句法分析算法，提出了“概率配价模式”，这个模式不仅可以定性地描述依存树中的支配和被支配关系，而且，还可以定量地计算依存关系的强度。

海涛并不满足于对历史事实的考察，他考察历史的目的是吸取前人的学术成果，“古为今用”，建立现代汉语的依存句法，因此，他更多的工作是研究依存语法的理论。

他讨论了依存语法的形式化方法和句法分析算法，并根据依存语法

的基本原则，深入地研究了汉语依存树库中众多树形图节点之间的支配和被支配的依存关系，构建了现代汉语配价模式；为了表示这种依存关系的不同强度，他在模型中引入了概率成分，提出了“概率配价模式”，在配价模式图中，用粗细不同的线条来直观地表示依存关系的强度，这样，就可以从定性和定量两个方面来描述树形图中的依存关系。在这本书里，海涛也简单介绍了如何通过依存树库来提取配价模式的方法。

在国内外依存语法研究中，大多数都只关注依存语法的定性研究，还没有明确地提出“概率”的概念，“概率配价模式”是海涛的创新，这个模式使我们有可能对依存关系进行定量的研究，这是海涛对于依存语法研究的新贡献。关于“概率配价模式”，我和海涛在 2007 年写了一篇论文，发表在《语言科学》上，有兴趣的读者可以阅读，以便加深对于本书的理解。

第三，通过一定规模的依存树库，检验了作者提出的基于配价模式的句法分析理论，并对汉语进行了初步的定量分析，发现汉语的依存距离远远大于英语、德语和日语的依存距离，这一发现有助于推动语言复杂网络研究的进展。

海涛构建了包含 13 个词类和 34 种依存关系的现代汉语依存句法，提出了汉语的树库格式，标注了一个包含两万词次的、实验性的汉语依存树库，同时还使用了哈尔滨工业大学的依存树库，在这样的基础上来检验他的理论。他采用了 XDK 和 MaltParser 等软件对汉语进行自动句法分析实验，并且在实验中适当地调整对某些语言现象的处理方法和标注的精细程度，把计算机的自动句法分析与语言学家的语言知识结合起来，有效地改善了句法分析的效果。

海涛的研究还发现，汉语的依存距离为 2.81，远远大于英语和日语的依存距离(英语为 1.386，日语为 1.43)。这是海涛的一个重要发现，而这样的发现是使用其他的语言研究方法难以做到的，这从另一个侧面说明了依存语法的长处。

依存距离的研究有助于把语言作为一种复杂网络来进行研究。目前关于复杂网络的研究结果表明，语言网络是一个无标度的复杂网络(scale free complex network)，是一个“小世界”(small world)的复杂网络。采用复杂网络来研究语言，有助于把语言与其他的复杂网络相比较，可以提高语言研究的普适程度，这方面的研究的价值是不言自明的。在这一方面，本书也有一定的探索。希望海涛能继续进行这种研究，把语言研究与复杂网络研究结合起来，做出更多的创新。

我和海涛认识已经将近30年了，他原来是学习自动化的，他的本职工作是从事企业信息化的高级工程师，是一个语言学的业余爱好者，与我常有书信往来，利用业余时间探讨语言学的各种问题。进入21世纪以后，他出于对语言研究的热爱，毅然改变了原来的专业方向，潜心投入清苦而艰巨的语言研究中，成为一个专业的(计算)语言学家。

在当今经济大潮下，很多人都忙于赚钱，并以此为乐，海涛凭着他的技术水平和外文功底，在这一方面也不乏机会。但他却选择语言学作为他的努力方向，宁愿与我等这样收入菲薄的语言学家为伍，下决心坐冷板凳，以探索学问作为自己的乐趣。这种品德，是值得我们学习的。

海涛有很好的自然科学和工程技术的基础，又有广博的语言学知识，对于语言研究充满了热情，这正是从事计算语言学研究的极好条件。本书是他出版的第一本专著，我对他表示热烈的祝贺，特作此序，算是我阅读此书的粗浅体会。希望他不断努力，今后做出更多的创新，出版更多的著作。

冯志伟

2009年5月7日于德国海德堡

Foreword

Richard Hudson[2]

This book deserves a prominent place in the growing international literature on dependency grammar and computational linguistics. The nature of syntactic structure is one of the most disputed questions in linguistics because science and tradition are so hard to separate in one of the most fundamental disputes.

An ancient tradition in Europe and the Middle East gives priority to the word as the basic unit of syntax, which means that syntax is primarily a matter of defining the relations between individual words—what have come to be called "dependencies". For instance, in the sentence "Small children often cry", the syntactician identifies just three dependencies that relate *small* to *children*, *children* to *cry*, and *often* to *cry*; once these dependencies have been identified, and the words and dependencies have been classified, nothing more remains to be said about the sentence's structure.

A much more recent tradition started with Leonard Bloomfield and the American structural linguists in the early twentieth century, and has come to dominate syntactic theory. In this tradition, the structure of a sentence consists of a more or less elaborate hierarchy of "phrases" in which the word has no particular priority. In "phrase-

---

2 Fellow of the British Academy. Emeritus Professor, University College London. Founder of Word Grammar.

structure grammar", in contrast with "dependency grammar", the four words of our example are combined with at least three phrases (*small children, often cry* and *small children often cry*) and possibly more—for example, *cry* would typically be classified not only as a word but also as a one-word phrase.

Unfortunately for scientific progress, this tradition was built from scratch, with very little reference to the existing dependency theory, and continues to ignore the dependency alternative. The result is that the very foundations of the scientific study of syntax are unstable, with an unresolved conflict between phrase structure and dependency structure. The main influence on syntactic theory is not debate and research, but geography. Linguists trained in America adopt phrase structure, while the more independent syntacticians of Europe favour dependency theory. This cannot be good for our discipline.

This background explains why a European dependency grammarian like me is pleased to see dependency theory being so ably developed by Haitao Liu outside the traditional "battle-field" of Europe and America, in the People's Republic of China. His dependency analyses of Chinese are a particularly welcome contribution to dependency theory. However, what is most exciting about his work is the way in which he has applied dependency analysis to large corpora in different languages, something which is possible nowadays thanks to the use of computers.

A corpus of naturally occurring sentences is the ultimate test of any theory of language precisely because it shows how important it is, in theorizing about language, to go beyond mere grammar. For instance, Liu reports that his Chinese corpus contains a very similar pro-

proportion of nouns to the proportion that I reported some years ago for several English corpora: about 41%. This is, indeed, an extraordinary finding; but it demands an explanation. Why should this figure emerge from such different corpora? One thing is clear: the explanation cannot lie only in grammar. To understand usage, we need a much broader range of theories: not only linguistic theories of grammar, vocabulary and genre, but also psychological theories of working memory. Liu's studies address many of these questions, though it is surely too soon to expect satisfying answers to many of them.

Perhaps the most interesting topic discussed in this book is the statistical measure of syntactic difficulty called "dependency distance". This measures the load which a word places on working memory, on the reasonable assumption that a word is kept active in working memory until all its outstanding dependencies have been satisfied. Returning to our earlier example, "Small children often cry", most of the words are very easy to process because their dependencies are satisfied by the next word; for instance, *small* needs a "parent" word, but this is immediately provided by *children*; and the same is true of *often*, which depends on the next word *cry*. But *children* is slightly harder because it is the subject of *cry*, from which it is separated by *often*. This increased load is still trivially easy for adult English speakers, but as the dependency distance between *children* and *cry* increases, the difficulty increases, and most English speakers struggle with really long subjects such as "Small children with anxious parents who keep trying to get them to smile and be happy even when they have tummy ache or when they are teething often cry".

Earlier work on dependency distance in languages such as Eng-

lish suggest that the limitations of working memory keep the average dependency distance quite low, and one would expect the same to be true in other languages. But Liu has found evidence for considerable variation among languages. In particular, he reports that the average dependency distance in Chinese is at least twice as great as that in English. This is an extraordinarily important finding which should stimulate a great deal of productive research. Do other corpora in English and Chinese show the same differences? If they do, why are the effects of working memory so different in the two languages? Is it because Chinese words are easier to hold in memory, so that more words can be kept active? Or is it because Chinese speakers have less limited working memories? I, for one, look forward very much to the light that Liu's future work will certainly cast on these fascinating questions.

理查德·哈德森序[3]

本书理应在日益增多的依存理论和计算语言学的国际性文献中占有一席之地。句法结构是语言学研究中最受争议的问题之一，因为在这样的争论中科学和传统很难被割裂开来。

在欧洲和中东，有一个古老的学术传统，认为词是句法的基本单位，这意味着句法的任务就是确定词与词之间的关系，即“依存关系”。例如，在句子“Small children often cry”中，句法学家仅需理清三组依存关系：small 与 children，children 与 cry，often 与 cry。一旦这些依存关系被确定下来，词与依存关系得以分类，关于句子结构的分析也就完成了。

20 世纪初期，布龙菲尔德和美国结构主义语言学家开创了一种新的传统，并开始主导句法理论的发展。在这种传统中，句子的结构或多或少是由“短语”间精细的层级关系组成的，词本身并没有什么特权。与“依存语法”相比，上述例子的“短语结构语法”分析，至少是由三个短语构成的（small children、often cry 以及 Small children often cry），也有可能更多，例如，cry 不仅是一个词，也可能被划分为一个短语。

这样的句法研究方法是白手起家的，几乎与已有的依存理论没有什么联系，并继续对依存方法持忽视的态度，这是科学发展之不幸。这也导致句法研究的科学基础不稳定，导致了短语结构和依存结构之间不可调和的矛盾。句法理论的主要问题不是争论和研究，而是地域分割。美国语言学家采用的是短语结构语法，许多欧洲的句法学家则倾向于依存理论。这不利于我们学科的发展。

以上背景解释了为什么像我这样的欧洲依存语法学家如此欣赏刘海

3 英国学术院院士，伦敦大学学院荣休教授，词语法理论的创始人。

涛，在欧洲和美国的“战场”之外，在中华人民共和国，所做的依存理论研究。他对于汉语依存句法的研究对依存理论的发展是一个巨大的贡献。令人尤为振奋的是他将依存理论应用于不同语言语料库的方法，这也是如今依存句法得以结合计算机使用的优势所在。

自然语言的语料库是精确测试语言理论的最佳方法，因为这样的方法显示了建立关于语言的而不仅仅是语法的理论的重要性。例如，刘海涛的研究显示他的汉语语料库中包含的名词比例与多年前我所研究过的多个英语语料库中的名词比例相似：约为 41%，这是一个非同寻常的重大发现，但同时这也需要一个合理的解释。为什么从不同的语料库中能够得出几乎相同的比例数据？显而易见，这个问题的答案不可能仅通过研究语法就能得到。我们需要更多的理论：不仅需要有关语法、词汇和语体的语言学理论，也需要工作记忆的心理学理论。刘海涛对这些问题做了很多研究，尽管这些研究仍不能使所有问题都得到满意的回答。

这本书中最有趣的地方也许是对“依存距离”统计测度的讨论。依存距离可以测量一个词在工作记忆中的负荷，我们可以将此理解为，一个词在它所有的依存关系被实现之前，一直被存储在工作记忆中。回到之前我们讨论的例子，“Small children often cry”，这个句子中的大部分的词都很容易处理，因为他们的依存关系都可以通过相邻的词来实现；例如，small 需要一个“支配”词，children 可以将其实现；often 的依存关系也可以通过 cry 来实现。但是 children 一词相对困难一些，因为它是 cry 的主语，但中间被 often 隔开。增加的这点负荷对于说英语的成年人来说并不难，大多数说英语者的困难在于为下面这样的句子寻求主语“Small children with anxious parents who keep trying to get them to smile and be happy even when they have tummy ache or when they are teething often cry”。

对英语等语言依存距离的早期研究表明，工作记忆的限制使得依存距离的平均值都非常低，并且有学者认为其他语言也将得到同样的结果。

然而，刘海涛的研究发现不同语言之间的依存距离是有明显差异的。尤其是汉语的平均依存距离至少是英语的两倍，这是一个极为重要的发现，理当激发更多的后续研究。英语和汉语的其他语料库是否也有同样的差异？如果答案是肯定的，为什么两种语言的工作记忆会如此不同？是否因为汉语词语更容易记忆，所以一次可以激活更多的词语？或者是因为说中国话的人有更大的工作记忆呢？我非常期待刘海涛将来的研究能一一揭开这些问题的神秘面纱。

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，