

句本位语法的 中文信息处理 理论与实践

彭炜明 / 著

非外信

外语教学与研究出版社

句本位语法的 中文信息处理理论与实践

彭炜明 / 著

外语教学与研究出版社
北京

图书在版编目 (CIP) 数据

句本位语法的中文信息处理理论与实践 / 彭炜明著. — 北京: 外语教学与研究出版社, 2021.3

ISBN 978-7-5213-2521-8

I. ①句… II. ①彭… III. ①汉字信息处理—研究 IV. ①H127

中国版本图书馆 CIP 数据核字 (2021) 第 053917 号

出 版 人 徐建忠

项目负责 范维哲

责任编辑 鞠 慧

责任校对 向凤菲

装帧设计 姚 军

出版发行 外语教学与研究出版社

社 址 北京市西三环北路 19 号 (100089)

网 址 <http://www.fltrp.com>

印 刷 北京捷迅佳彩印刷有限公司

开 本 720×980 1/16

印 张 21

版 次 2021 年 3 月第 1 版 2021 年 3 月第 1 次印刷

书 号 ISBN 978-7-5213-2521-8

定 价 68.00 元

购书咨询: (010) 88819926 电子邮箱: club@fltrp.com

外研书店: <https://waiyants.tmall.com>

凡印刷、装订质量问题, 请联系我社印制部

联系电话: (010) 61207896 电子邮箱: zhijian@fltrp.com

凡侵权、盗版书籍线索, 请联系我社法律事务部

举报电话: (010) 88817519 电子邮箱: banquan@fltrp.com

物料号: 325210001



记载人类文明
沟通世界文化
www.fltrp.com

序

2009年彭炜明从信息科学学院中文信息处理专业，考入文学院汉语言文学专业攻读博士学位，他是我们学科招收的第六届交叉学科博士。进入互联网时代，汉字标准化、国际化成为国家软实力的重要因素，简繁并存、古今并有产生的多种矛盾，在进入计算机以后更为尖锐，完全在纸上研究汉字已经不能适应国家建设的需要，信息科学与语言文字两个学科结合起来培养文理交叉知识结构的复合型人才，是适应这种发展形势的急需。所以，炜明前面的几届交叉学科的博士，都是从《说文》学入手，以研究汉字信息处理问题为具体方向的。炜明进入文学院的那一年，汉字问题已经有所缓解，语言智能问题提到日程上来，于是他选择了句子层面的信息处理问题作为重点。应当说，这是一个很有远见的选择。人类的自然语言带上意义和思想以后，句子才是它的基础单位，字义、词义都要结构为句义，才能有深度的人机对话。机器翻译和机器理解问题，早已经不是新问题，瓶颈在于语言的人文性和意义的经验性是电脑很难适应的。处理汉语的句子还有一个比英语等语言更大的难点，就是汉语作为词根语，缺乏语法形态，到处都是无形的变化和只有内涵的“意合法”。这些，都给解决问题带来比字词处理更大的难度。但是，不论怎样，先把句子问题理顺了，有了适应计算机处理问题的机制，都是解决问题的先决条件。

说实话，从计算机科学技术过渡到语言文字，两个学科这样大的跨度，对我们师生已经是很大的挑战。炜明的博士论文《句本位语法数字化平台的建设和应用研究》，由字词处理过渡到处理句法，是更大的挑战。令人欣慰的是，炜明作为信息技术专业的硕士，在句法问题上跟我们对话，几乎没有遇到什么很大的隔阂。相较于多数纯理工科学生而言，炜明属于文理科成绩俱佳的优秀学生，基础语法不生疏，阅读能力属一流。他用了将近一个学期的时间，梳理了现代汉语多个语法体系，最终还是认为面对汉语的应用问题，传统语法最为适合信息处理。选择句本位语法，以黎氏语法为基础，是基于以下四个原因：第一，汉语的句子结构通过《马氏文通》吸收了拉丁语法的基本范式，经过很多大家的修补、改造，

由“暂拟体系”和“系统提要”做了综合，能够作为教学语法使用，证明它是符合汉语也便于应用的。第二，处理句子问题，以句本位语法为基础，有一个完整的基本单位垫底，起步会顺利得多。第三，汉语是一个重意义的语言，处理意义和词类是句法最关键的问题。在传统语法的发展中，产生了两种处理问题的倾向：第一种尽量从形式上界定句法功能和词类，意义迁就功能，变得复杂；第二种看重意义的作用，界定词类和句法功能时，将意义因素加入其中，使意义问题相对比较简化。黎氏语法是后一种，更加适合汉语。第四，黎氏语法的图解法，对任何一个句子的结构都不回避，图解形式在功能的界定上又基本合乎逻辑，用线条的层次、方向区别了不同功能，也以线条的相连与相隔表明了句中词语意义关系的亲疏，这是利于计算机理解和表达的有利条件。事实说明，这个选择是成功的。

炜明的第二步工作是采集大量的图解材料，对黎氏语法有更深入的了解，同时对黎氏语法解决问题不彻底的地方做进一步的加工。这一阶段我看过他 11 版草稿，不断和他讨论，学习各种流派的语法学论著寻找适应方法，处理问题有所精进，逐步解决了以下几个问题：

(1) 进一步清理了句子的“主—谓—宾”主干，以主干为基础，排列了具有层次的句式。由于句式的概括性，使意义纷繁的语句，被有限的句式控制，从复杂变为简化。又因为句式有不同的层次，各自有包含关系，条分缕析，又进一步简化了句子的模式。

(2) 黎氏语法最复杂的是它与谓语相关的所谓补足语部分，这一部分所以杂乱，是因为缺乏从词类、关系和层次上进行区别性的分解。在吸取其他语法学家的方法进一步清理分解后，固定了不同补足语在图解中的位置，使这部分补足语变得清晰。

(3) 正确解释了黎氏“依句辨品”和“离句无品”的原则，将“类”和“品”与句子的关系重新做了更为合理的界定，定出了规则，将已有的图解一一做了微调适。在此基础上，把已经入句的、词类确定的语词做了标注，入库待用。

这三个问题，都是依靠计算机强大的存储、统计分析、数据挖掘乃至可视化表达等能力的持续迭代、不断优化才得以渐趋清晰和完善的。

解决了以上三个问题以后，所构建的句本位树库和词库，向着“功能—词类—词义”具有统一性、“结构形式—意义关系—图式表现”具有统一性这两个

目标，向前迈了一大步。

从论文答辩到现在，炜明以及句本位研究团队，加紧语料标注，树库规模得到了极大的扩充，遇到的各种微观问题不断修正，有了更为明晰的成果。此次专著出版，基于持续构建的工具平台、大规模标注实践的基础上，专门总结提炼理论原理和实践程序的科学性，相较于10年前的研究，又有很新的进展。

看完书稿，我也有很多感触：在信息时代，解决人机对话问题，是容不得说空话的。重新反思之前的研究，想得更现实一些，用电脑检测一下人脑，是推进汉语语法研究很重要的一项工作。在应用过程中，各种流派和方法需要综合，各具实力的研究团队力量需要整合，取长补短，共同攻坚，那种坚持一条道走到黑，各个自成一家、孤军奋战已经无法满足时代和应用的需要，经过实践的检验，根据汉语的特点，运用信息技术，从基础研究开始向前推进，中国的语言文字学才能为人工智能的发展起到应有的作用，而不会在信息科学发展的飞快速度下被边缘化。在此过程中，语言学各种流派的相互学习、彼此借鉴，应当成为必然的趋势。而以计算机为核心的信息技术所具有的巨大潜力、融合能力乃至据此引发的创新动力已经得以充分显现，但是，对具有个性和经验性的人文现象的处理，尚有巨大的可塑空间，需要在质量和数量上尽快突破。信息科学技术和人文科学学术有机结合起来，我们才能有更大的自主创新，信息技术才能为人类生活的幸福作出更大的、更直接的贡献。

王宁

2021年3月5日

于北师大

序

应邀为彭炜明博士的专著《句本位语法的中文信息处理理论与实践》写序，想到需要了解一下当前自然语言处理技术的现状。于是，先用一小段中文对一个在线机器翻译系统进行测试，同若干年前相比，英文译文质量确实有了明显的改进。不过，当输入一首自制的打油诗《Email 赞》

Email 赞

伊妹传书超雁飞，
鸟语花艳音乐美；
待到网络送香来，
虚拟世界人亦醉。

时，得到的译文

Email like

Yi Mei's biography surpasses Yanfei's,
The beautiful music of birds and flowers;
When the network sends incense,
People in the virtual world are also drunk.

却难以让人满意。尽管诗的题目有所提示，但系统显然不知道“伊妹”就是Email，也不知道“鸿雁传书”这个典故。当然，人如果没有这些知识，也不会理解诗的第一句的意思。在诗中加上两个“注”：

Email 赞

伊妹^①传书超雁飞^②，
鸟语花艳音乐美；
待到网络送香来，

虚拟世界人亦醉。

注①伊妹即 Email, 见《现代汉语词典》(第7版), 1541页。

②这里化用了典故“鸿雁传书”。

这样, 人便知道“伊妹”就是 Email, 不再是生词; 若又知道典故“鸿雁传书”, 就能理解这句诗了, 不知道这个典故的人也会去查成语词典。再次让系统翻译这首带“注”的诗, 得到的译文是

Email like

Yi Mei's biography surpasses Yan Fei's,
The beautiful music of birds and flowers;
When the network sends incense,
People in the virtual world are also drunk.

Note ① Yimei is email, see modern Chinese dictionary, 7th Edition, page 1541.

② Here, the allusion “Hongyan Zhuanshu” is used.

由此发现了系统存在的问题。系统无视“注”的存在和作用, 还是认为自己什么都可以译出来, 至于译得对不对, 那就不管了。还有, 第2句“鸟语花艳音乐美”的句法结构是3个主谓结构的并列, 译文折射出系统不能识别这样的结构, 原因或许是缺乏相关的语言学知识吧。

从这个简单的例子, 可以看出自然语言处理所表现的人工智能与人类智慧的差别。孔子曰: “知之为知之, 不知为不知, 是知也。”人明白自己有所知、有所不知, 这是人类智慧的特征之一。自然语言处理是数值计算机在非数值领域的最早应用, 但在自然语言理解的层面上, 可以说至今尚未能取得突破性进展。关键原因是到目前为止处理自然语言的计算模型与算法, 与人脑的语言机制大相径庭。即便是当今独领风骚的深度学习技术所基于的人工神经网络, 虽然号称模拟人类大脑, 其实与人脑机制相去甚远, 人类对自身的语言机制还是不甚了了。要想登上自然语言理解这座人工智能的高峰, 还有很长的路要走, 应当了解人类的大脑是如何理解自然语言的, 需要人工智能、脑科学、认知科学、语言学(包括神经语言学、计算语言学)等领域的专家协同攻关。彭伟明博士正是在中文信息处理这个交叉学科领域辛勤耕耘, 收获了这部专著, 为奔腾向前的中文信息处理

长河奉献了一朵绚丽的浪花。

这部专著包含诸多创新成果，其研究思路亦颇具借鉴价值。择其要者，举以下3点。

首先是形式化的句式结构体系。以黎锦熙先生《新著国语文法》的图解法为基础，吸收多种语法理论之精粹，建立了“基本句式→扩展句式→复杂句式”的句式结构体系。该体系不仅有可视化的图解表达，而且体现了关于句法结构层次性的独到见解。比如，在句法树中设置动态词节点，明确地让词法结构与句法结构区分开来。再有，对句式结构的层次深度、分析层次等进行了严密的论证和推导，这些理论分析工作为其后来实现基于模式匹配的自动句法分析奠定了基础。

第二是实现词法、句法一体化的句式结构自动分析器。词法分析和句法分析是串行级联还是一体化，这又是中文文本自动分析的一个两难问题。本书自动分析器总体上采取一体化方法，即词法分析不消歧或少消歧，歧义交由句法语义模块消除。关于自动分析器的评价标准也很有特色：总体准确率不再是唯一的甚至主要的评价指标，而更加看重自评价能力，即系统对分析结果的可解释性。

最后是大规模的树库和知识库。在理论研究的基础上进一步构建大规模句本位语法树库，并实现树库和词汇、词法、句法、语义知识库的联动。语言知识资源建设需要理论与实践相结合，语料库和知识库相联通，如此构建的语言知识资源才能为计算模型的构建提供支持，实际推动中文信息处理技术的发展，并长久地葆有生命力。语言知识资源建设的过程与句式结构体系的建立也是一个相互促进的过程。

这是一部另辟蹊径、独立探索的佳作。基于传统句本位语法理论取得的研究成果表明，在进军自然语言理解的征途中，语言学研究，包括计算语言学研究，也是一支重要的方面军。即便在深度学习占据技术主流的当下，也应当坚持多学科深度交叉融合的研究风范，这就需要有多学科交叉融合的人才。彭炜明博士大学本科和硕士阶段学习计算机专业，博士阶段在王宁教授指导下专攻汉语言文字学，接着又在北大计算语言学研究所从事博士后研究。20余年寒窗苦读，彭炜明博士不仅知识广博，养成了独立从事科学研究的能力，而且具备了良好的多学科交叉融合的知识结构。这正是彭炜明博士取得创新性成果的前提条件。彭炜明博士在明确攻关方向之后，能够长期坚持，锲而不舍，十年磨一剑。语言知识资

源建设是一项大型语言工程，团队的支持也是取得优秀成果的关键因素之一。在博士后合作研究期间，我了解到彭炜明博士有很好的团队协作精神。天时地利人和，彭炜明博士的佳作终于问世了，谨此表示衷心的祝贺。

衷心期望中文信息处理学界涌现出更多的优秀人才，涌现出更多的创新成果。

俞士汶

2021 年 1 月 28 日

于北京大学

编写说明

句本位语法是汉语语法研究早期的一个流派，以“句子成分分析”的析句方法和“依句辨品”的词类观点著称。本书基于上述理论观点开展汉语的计算机自动析句研究，是汉语传统语法理论与现代中文信息处理深度交叉的学术专著。书中引言部分界定术语“句本位语法”的广、狭二义，概述全书主要内容。第2章综述与广义句本位语法相关或相通的主要语法体系，辨析异同。第3章对黎氏语法的图解法进行形式化改造，设计实现一种面向中文信息处理的新型形式化语法结构体系——句式结构。第4、5章分句法和词法两部分展开句式结构析句规范和语例说明，为汉语语法的系统描写。第6章阐述汉语句式结构析句的理论框架，属于计算语言学性质的理论探讨。第7章初步实现句式结构自动分析器。第8章利用自动分析器支撑树库构建，以树库数据迭代更新知识库，形成树库和分析器的技术闭环。第9章基于树库为相关教学领域提供数据和技术支持，为技术的应用落地。第10章总结句本位语法和句式结构在中文信息处理领域的价值，展望后续研究的发展图景。

本书是国家自然科学基金面上项目“中华传统文化经典作品专题学习模型的研究与构建”（编号：61877004）、国家自然科学基金青年项目“基于图式理论的中学生阅读能力提升模型构建”（编号：62007004）的研究成果，可以作为中文信息处理、自然语言处理、计算语言学、现代汉语语法等专业科研人员的阅读材料，也可供相关专业的本科生和研究生阅读参考。

目 录

第1章 引言 / 1

- 1.1 句本位语法 / 1
- 1.2 中文信息处理与形式语法 / 2
- 1.3 本书内容与组织框架 / 5

第2章 句本位语法的历史回顾 / 7

- 2.1 汉语语法体系 / 7
- 2.2 黎氏语法体系 / 9
 - 2.2.1 语法体系概述 / 10
 - 2.2.2 图解法要点补述 / 18
 - 2.2.3 析句方法评述 / 26
- 2.3 《暂拟系统》和《系统提要》 / 34
- 2.4 其他各家的语法体系 / 38
 - 2.4.1 史存直《语法新编》 / 38
 - 2.4.2 陈望道《文法简论》 / 39
 - 2.4.3 张志公《汉语语法常识》 / 42
 - 2.4.4 吕叔湘《现代汉语八百词》 / 43
 - 2.4.5 邢福义《汉语语法学》 / 50
 - 2.4.6 黄伯荣、廖序东《现代汉语》 / 52
 - 2.4.7 北大中文系《现代汉语》 / 54

第3章 图解法的形式化——句式结构 / 58

- 3.1 设计原则 / 58
- 3.2 句法结构 / 59

- 3.2.1 基本句式 / 60
- 3.2.2 扩展句式 / 61
- 3.2.3 复杂句式 / 62
- 3.2.4 特殊句式 / 67
- 3.3 词法结构 / 68
 - 3.3.1 词库词: 词类和义项 / 68
 - 3.3.2 动态词: 结构模式 / 71

第4章 句式结构的语法体系(句法) / 74

- 4.1 句子成分 / 74
 - 4.1.1 主语和谓语 / 74
 - 4.1.2 宾语 / 76
 - 4.1.3 补语 / 79
 - 4.1.4 定语和状语 / 82
 - 4.1.5 独立语 / 85
- 4.2 复式结构 / 86
 - 4.2.1 NP复式结构 / 86
 - 4.2.2 VP复式结构 / 89
- 4.3 句式 / 108
 - 4.3.1 主谓谓语句 / 108
 - 4.3.2 省谓句 / 111
 - 4.3.3 存现句 / 114
 - 4.3.4 双宾语句 / 117
 - 4.3.5 数量对应句 / 120
 - 4.3.6 比较句 / 121
 - 4.3.7 强调句 / 124
 - 4.3.8 无主句 / 126
 - 4.3.9 独词句 / 126

第5章 句式结构的语法体系（词法） / 128

- 5.1 基本概念 / 130
 - 5.1.1 语素、词素、词、短语 / 130
 - 5.1.2 词根/词缀、自由语素/不自由语素 / 130
 - 5.1.3 单纯词、合成词 / 131
 - 5.1.4 构词、构形 / 131
 - 5.1.5 熟语、缩略词和离合词 / 132
- 5.2 名词 / 133
 - 5.2.1 时间词 / 139
 - 5.2.2 方位词 / 140
- 5.3 动词 / 141
- 5.4 形容词 / 146
- 5.5 副词 / 148
- 5.6 数词、量词、数量词 / 150
- 5.7 代词 / 153
- 5.8 介词 / 154
- 5.9 连词 / 155
- 5.10 助词 / 158
- 5.11 叹词、拟声词 / 160

第6章 句式结构析句的理论框架 / 161

- 6.1 析句目标——建构形式化句义 / 161
- 6.2 析句原则——形式和意义统一的“活看”法 / 163
 - 6.2.1 突显主干，中心自由 / 164
 - 6.2.2 允许转类，规范省略 / 165
 - 6.2.3 小句为纲，兼顾章法 / 167
- 6.3 析句理念——以有限驾驭无限 / 169
 - 6.3.1 层次深度和句式结构表达式 / 170
 - 6.3.2 分析层次和单层结构表达式 / 172

- 6.3.3 句式结构特征的层次约束 / 175
- 6.4 析句思路——自顶向下逐层匹配 / 176
 - 6.4.1 模式的初始建立 / 177
 - 6.4.2 模式分层和规则化 / 179
 - 6.4.3 规则的例外处理 / 181
- 6.5 析句策略——基于语言知识工程 / 182

第7章 句式结构自动分析器初探 / 185

- 7.1 自动析句的评价标准 / 185
- 7.2 正则表达式及其扩展 / 189
 - 7.2.1 基于字符的通用正则表达式 / 189
 - 7.2.2 基于词语特征的扩展正则表达式 / 193
- 7.3 小句的单层结构分析 / 194
 - 7.3.1 算法框架 / 194
 - 7.3.2 词法分析 / 195
 - 7.3.3 句式匹配 / 196
 - 7.3.4 结果排序 / 198
 - 7.3.5 实验分析 / 199
- 7.4 整句的完全结构分析 / 200
 - 7.4.1 算法框架 / 200
 - 7.4.2 标点句分析 / 201
 - 7.4.3 小句归并 / 202
 - 7.4.4 实验分析 / 205

第8章 句本位语法树库的构建实践 / 206

- 8.1 标注工具 / 206
- 8.2 标注模式 / 209
- 8.3 工程管理 / 212
 - 8.3.1 语料入库与分类管理 / 212

8.3.2 标注效率与质量控制 / 214

8.4 知识库构建与更新 / 215

8.4.1 词汇知识库 / 216

8.4.2 词法模式库 / 218

8.4.3 句法规则库 / 219

8.4.4 框架语义库 / 220

第9章 句本位语法树库的教学应用 / 222

9.1 面向国际汉语教学的应用 / 222

9.1.1 交互式出题 / 223

9.1.2 可读性评估 / 229

9.2 面向中小学语文教学的应用 / 233

9.2.1 文言文阅读理解 / 233

9.2.2 语篇连贯性分析 / 235

9.3 面向计算语言学教学的应用 / 237

第10章 结语和展望 / 242

参考文献 / 244

附录 / 246

附录一：词法模式库（前100条高概率模式） / 250

附录二：句法规则库（前22条高频规则） / 254

附录三：图解树库示例（鲁迅《孔乙己》） / 255

附录四：图解析句标注的反馈 / 315

后记 / 318

第1章 引言

1.1 句本位语法

在汉语的语言学科中，作为独立分支的汉语语法学肇始于 1898 年马建忠的《马氏文通》。在这迄今并不算长的 120 余年历史中，涌现出了诸多语法学派和语法理论体系。其中，以黎锦熙为代表的黎氏语法学派及其句本位语法理论显然在汉语语法学史上留下了浓墨重彩的一笔。不过，曾经盛极一时的句本位语法在今天的语法学界已经少有问津者。语法学界对黎氏句本位语法的评价也是毁誉不一，有人赞其代表作《新著国语文法》为“现代汉语语法开创性和奠基性的著作”（胡明扬，2002），也有不少论著将其视为“模仿西洋文法”的典型。

要客观地看待这些评价，批判地继承和发扬句本位语法中的优秀思想，应当首先对“句本位语法”做出狭义和广义的两重定义。因为历史上以“句本位”自称的语法体系不止黎氏一家（当然影响最大的还是黎氏语法），而不称“句本位”的语法体系中也有许多其实具备或部分具备句本位语法特征。以黎氏语法为例，句本位语法的主要特征如下：

（1）句法分析采用“句子成分分析法”，句法研究以句子结构格局的归纳为指归；

（2）词类分析服从于句法分析，即著名的“依句辨品，离句无品”说；

（3）成分分析以逻辑语义为主，有各种所谓“变式”和“省略”的分析；

（4）句子成分中设“补足语”（或称“足语”），作为动词的连带成分之一；

（5）“补语”与“状语”合为一类，即副词性的附加语，简称“副附”；

（6）注重语法教学法，操作上通常采用“图解法”表示语法分析结果。

所谓“句子结构格局”是指将句子按成分划分成几个部分后形成的排列模式，教学上通常称作“句式”或“句型”。真实的句子无穷无尽，要将其归纳为有限的句式，就必须明确成分分析的层次。事实上，任何句法分析方法都要讲究层次，而句子成分分析法的层次有其鲜明的句本位特色——体系设计中蕴含着—一个预设的“结构原型”，比如：几大句子成分首先分为“主干成分”和“附加成分”