计算语言学
与语言科技
原文丛书

TREEBANKS
Building and Using Parsed Corpora

树　库
句法分析语料库的构建和使用

[法] Anne Abeillé　编
詹卫东　导读

北京大学出版社
PEKING UNIVERSITY PRESS

TREEBANKS:

Building and Using Parsed Corpora

树 库

——句法分析语料库的构建和使用

[法] Anne Abeillé 编

詹卫东 导读

阜阳师范学院图书馆 藏书

北京大学出版社
PEKING UNIVERSITY PRESS

著作权合同登记号　图字：01-2012-7369
图书在版编目(CIP)数据

树库：句法分析语料库的构建和使用 = Treebanks：building and using parsed corpora：英文 / (法)阿贝耶(Abeillé, A.)编. — 北京：北京大学出版社, 2014.12
（计算语言学与语言科技原文丛书）
ISBN 978-7-301-24952-9

Ⅰ. ①树…　Ⅱ. ①阿…　Ⅲ. ①计算语言学－语料库－文集－英文　Ⅳ. ①H087-53

中国版本图书馆CIP数据核字(2014)第233543号

Reprint from English language edition:
Treebanks
by A.Abeillé
Copyright © 2003 Springer Netherlands
Springer Netherlands is a part of Springer Science+Business Media
All Rights Reserved
This reprint has been authorized by Springer Science & Business Media for distribution in China Mainland only and not for export there from.

书　　名：树库——句法分析语料库的构建和使用
著作责任者：[法] Anne Abeillé　编
责任编辑：李　凌
标准书号：ISBN 978-7-301-24952-9/H · 3599
出版发行：北京大学出版社
地　　址：北京市海淀区成府路205号　100871
网　　址：http://www.pup.cn　　新浪官方微博：@北京大学出版社
电子信箱：zpup@pup.pku.edu.cn
电　　话：邮购部 62752015　发行部 62750672　编辑部 62753374　出版部 62754962
印 刷 者：北京大学印刷厂
经 销 者：新华书店
787毫米×1092毫米　16开本　28.5印张　511千字
2014年12月第1版　2014年12月第1次印刷
定　　价：72.00元

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究
举报电话：010-62752024　　电子信箱：fd@pup.pku.edu.cn

“计算语言学与语言科技原文丛书”由北京大学—香港理工大学汉语语言学研究中心、北京大学计算语言学研究所(由973课题“文本内容理解的数据基础”、863课题“大规模汉语语义基础资源库和知识库设计构建及工具平台”支持)和北京大学出版社合作推出

学术委员会
Academic Advisory Committee

主　任：

黄居仁(香港)

委　员：

Chris Manning (Stanford)
Harold Somers (Dublin)
Maarten de Rijke (Amsterdam)
Suzanne Stevenson (Toronto)
陈克健(台北)
冯志伟(北京)
李宇明(北京)
陆俭明(北京)
郭　锐(北京)
石定栩(香港)
苏克毅(台北)
孙茂松(北京)
王厚峰(北京)
王士元(香港)
俞士汶(北京)
松木裕治(奈良)
郑锦全(Urbana-Champaign)
邹嘉彦(香港)

此为试读,需要完整PDF请访问: www.ertongbook.com

编委会
Editorial Committee

主　编：
黄居仁教授(香港)

编　委：
顾曰国教授(北京)
姬东鸿教授(武汉)
陆　勤教授(香港)
苏新春教授(厦门)
夏　飞教授(Seattle)
薛念文教授(Waltham)
詹卫东教授(北京)
赵铁军教授(哈尔滨)
宗成庆研究员(北京)
黄萱菁教授(上海)
刘　群教授(Dublin)
蒙美玲教授(香港)
孙　栩研究员(北京)
徐飞玉教授(Saarbrücken)
曾淑娟副研究员(台北)
张凤珠编审(北京)
周　明研究员(北京)
常宝宝副教授(执行秘书)(北京)

丛书前言

“计算语言学与语言科技原文丛书”于2010年创立，2010 COLING国际计算语言学会议在北京举办之前出版了第一批图书。本丛书的出版象征着国内计算语言学研究与国际的接轨。国内学者正跻身于计算语言学的国际舞台：一些资深学者已在COLING两个最主要的国际会议/组织中获选并担任重要的领导职务；而积极参与这些重要的国际会议也已在年轻学者中蔚然成风，他们已可谓是会议的主流参与者之一。在这样的氛围中，希望本丛书第二批图书的出版，能让国内有心投入语言科技与计算语言学研究的学者们如虎添翼，在国际舞台上创新并引导议题！

计算语言学（Computational Linguistics，CL）在语言科学与信息科学的研究中扮演着关键性的角色。语言学理论寻求对语言现象进行规律性的预测，做出完整的解释，计算语言学正好为这两点提供了验证与应用的大好机会。作为语言学、信息科学乃至于心理学与认知科学结合的交叉学科，计算语言学更为语言学基础研究与福国利民应用研究的接轨提供了绝佳界面。事实上，计算语言学与人类语言科技（Human Language Technology，HLT）可以视为体用两面，不可切分。

计算语言学研究的滥觞，其实源于上世纪五六十年代的机器翻译研究，中文计算语言学的研究也几乎同步开始。在美国伯克利加州大学研究室，王士元、邹嘉彦、C.Y. Dougherty等人1960年已开始研究中英、中俄机器翻译。他们的研究是与世界最尖端的科技同步的。国内中俄翻译研究也不遑多让，大约在20世纪50年代中期便已开始。可惜的是，这些中文方面早期机器翻译研究，由于硬件与软件的限制，未能有效传承下来。中文计算语言学研究比较系统的发展始于1986年。这一年，海峡两岸不约而同地分别成立了两个致力于建立中文计算语言学基础架构的研究群：北京大学的计算语言学研究所，在朱德熙先生倡议下成立，随后一段时间由陆俭明、俞士汶主持；而台北“中研院”的中文词知识库小组，由谢清俊创立，陈克健主持，黄居仁1987年回去后加入。

中文计算语言学的研究，近30年来已累积了相当可观的成绩。计算语言学的重要研究领域与议题中都能看到中文方面的相关研究成果，华人计算语言学学者也渐渐在国际学术界崭露头角。随着世界经济转向知识密集型

产业，跨语言跨文化沟通与知识整合是知识产业的关键瓶颈，语言科技的发展成为国际主流指日可待。在这个有利发展的大环境下，我们期待看到，中文计算语言学与华人计算语言学学者的成绩，百尺竿头更进一步，中文方面的研究可以进入计算语言学的学术核心，能够产生有能力引导议题并掌控研究方向的大师。

回顾国内的计算语言学发展，计算机科学的贡献多于语言学的贡献。这个现象，在理论与概率模型整合研究的趋势中，不免令人忧心。语言学的贡献弱，或许可以部分归咎于英文研究专著在国内不易取得；而比较容易取得的期刊或会议论文，在篇幅的限制下，又往往无法对理论做深入完整的铺陈，从而导致国内的年轻学者长于运算而拙于理据。因此，在期待大师与引领世界研究潮流两个方向，藉由英文专书来巩固研究理据，进而开拓研究视野，是非常重要的一步。

“计算语言学与语言科技原文丛书”的引进，就是在上述背景下促成的。个人忝为剑桥大学出版社“自然语言处理研究”（Studies in Natural Language Processing，SNLP）系列的主编，对于将此系列中较重要的几本书引入国内，责无旁贷。第二批出版的原文书，除了剑桥大学出版社的图书外，还有施普林格出版公司（Springer）语言科技系列中的几本书，以进一步拓展领域涵盖面。引进原书，原样出版，是容易的，然而若要真正搭建知识的桥梁，使国内学者与学生不仅能开拓研究视野，更能将原文著作的理论精髓应用于中文研究，则实在不易。因此，本系列除了原书出版外，每本书我们都邀请了一位专家撰写中文导读。这些导读可以说是本系列的精华、重点，使每本书比剑桥和施普林格的原本增加了不少附加价值。

每篇中文导读都包括三个重要的组成部分。第一部分是全书内容概要的介绍。导读专家都是长年浸淫于该领域的学者，他们能提纲挈领，并提供相关研究背景。因此，通过阅读导读，读者更易掌握并吸收该书的重要内容。第二部分是中文相关研究。原文著作不见得会提到相关的中文研究，由导读专家补充介绍，搭起理论与中文相关应用的桥梁，更能引导读者找到在这个议题进入中文研究的最佳切入点，让中文相关研究的开拓者的成绩更能发扬光大。第三部分重点在于补充原书出版后该领域研究的新发展。现代科技发展迅速，任何经典著作出版后，几乎马上就有新的相关研究。因此，在理论架构的脉络中，加上新近发展，能使读者更贴切地掌握研究脉动。全书的内容摘要通常以文字叙述，而中文相关研究及最新研究发展则分别以文字叙述和延伸阅读书目的方式呈现。延伸阅读书目，可以使读者很快上手，进入相关研究领域，也是本系列的重要设计之一。

丛书2010年出版第一批图书，现在出版第二批图书，必须感谢许多同行的付出。在规划出版的漫长过程中，北大计算语言学研究所的俞士汶老师及常宝宝老师一直无私无悔地支持。而香港理工大学的挹注，北大－理大汉语语言学研究中心石定栩、郭锐等几位的支持，使得整个系列能够顺利出版。此外，还要感谢北大出版社王飙主任、杜若明编审及李凌编辑，他们认同我们的宗旨，落实了丛书的出版工作。最后，感谢丛书的国内编委，特别是此次担任导读主笔的各位，正是他们脑力与心血的付出，才替读者们搭建了进入学术殿堂的台阶。

丛书主编　黄居仁

谨志于香港，红磡

二零一四年九月

导　读

詹卫东

1　学科背景及本书的定位

树库(treebank)属于深加工语料库,是语料库语言学和自然语言处理(NLP)技术发展到相对成熟阶段的产物。宽泛而言,语言研究一直以来都离不开“语料”。但从“语料”到现代意义的“语料库”,是从二十世纪五六十年代伴随着电子计算机的应用才开始的,其发展轨迹及趋势有几个明显特点:(1)语料库规模不断扩大,类型不断多样化;(2)标注信息不断丰富;(3)应用范围不断拓宽。这些特点是跟过去半个世纪整个信息社会大环境的飞速变化和NLP技术的进步分不开的。计算机存储能力和互联网的加速发展,使得电子化的大规模的自然语言资源越来越容易获得。从上世纪六十年代起步时的百万词级规模到八九十年代的上亿词级规模,再到今天语料库的规模已不再是人们关心语料库的重点,不难感受到这种惊人的扩容速度。与此同时,语料也从原始形态的生语料库发展到经过多级标注(annotation)的所谓熟语料库。标注的信息从一般的词语形态信息、词类信息等很快发展到了标注句法结构、句法功能、语义角色信息等等。标注词类信息的语料库跟原始语料一样仍然保持着一维串性结构,而标注了句法结构、句法功能信息的语料库则因描述了词语(以及词组)之间的层级组合关系,成为二维的树状结构(tree structure),因此这样的语料库就被称为树库。像树库这样的带标语料库的发展还明显得力于NLP技术本身发展的推动。这一方面是NLP技术的发展需要有树库这样的深加工语料库提供数据支持。另一方面则是由于NLP技术的进步反过来大大提高了树库加工的效率,减低了人工成本,使得树库加工成为切实可行的一项工作。从二十世纪九十年代开始,NLP的主流技术从基于规则的方法纷纷转向基于统计的方法,在这样的背景下,来自真实语料的语言统计数据逐渐取代以往由人工归纳的语言学专家知识,成为NLP应用系统所依赖的主要知识源。在词类标注、句法分析、机器翻译等许多NLP技术的相关评测中,基于统计方法的系统都取得了更胜一筹的成绩,从而吸引了更多的研究人员来推进这种数据驱动型NLP技术的研究。尽管构建树库是

成本相对比较高的语言工程，但受到英语树库的成功鼓舞，从二十世纪九十年代中后期开始，其他语种也陆续启动了树库加工项目。随着机器学习技术在NLP领域应用热潮的不断升温，树库的研究和应用也受到越来越多的重视，不但涉及的语种已经扩展到几十个，而且句法标注所依据的理论体系也由生成语法的短语结构语法发展到中心语驱动短语结构语法（HPSG）、依存语法（dependency grammar）、词汇功能语法（LFG）等等多种理论框架并存的局面（有的树库甚至是把短语结构跟依存关系的标注融合到一块进行标注）。本书出版于2003年，距离上世纪九十年代初英语树库问世已有10年。尽管如编者在导言中所说的，树库作为语言资源的一种新形式，本书的多数篇幅是在讨论如何加工树库，关于如何使用树库的篇幅相对较少，但仍然可以说本书内容基本反映了这10年间树库研究的整体面貌，是树库研究发展到一定阶段的一个比较全面的总结，起到了承前启后的作用。

2　内容提要

本书正文共21章，正文之前有一篇导言（introduction）。导言是本书编者对全书内容的概括介绍。21章中有的是专为本书撰写的，有的则是由发表在一些相关会议上的论文改写的。21章内容分为两大部分：第一部分从第1章到第15章，讲如何构建树库；第二部分从第16章到第21章，讲如何使用树库。

第1章到第4章介绍英语树库的构建，内容分别是对美国宾州树库整体情况的介绍，对近20年英语树库构建工作的思考，BOE英语语料库的词汇形态标注、句法标注以及后续的句法功能标注，ICE-GB（国际英语语料库—英国部分）树库的句法结构校对方法。第5章和第6章介绍德语树库的构建，分别是德语新闻报纸语料库的句法标注、德语新闻组语料库（USENET）的错误类型标注。第7章和第8章是两种斯拉夫语族语言树库的构建，第7章介绍捷克语树库的构建，第8章介绍基于HPSG的波兰语句法测试语料库。第9章到第12章是四种罗曼语族语言树库的构建，第9章介绍西班牙语树库的开发，第10章介绍法语树库的构建，第11章介绍意大利语句法—语义树库的构建，第12章介绍一个中世纪葡萄牙语树库的创建。第13章到第15章是其他语种树库的构建情况介绍，第13章介绍台北“中研院”Sinica中文树库，第14章介绍日语树库，第15章介绍土耳其语树库。

第16章介绍树库标注的编码形式。第17章和第18章讨论如何利用树库进行句法分析评测，第17章介绍如何构建专门的供句法分析评测用的树库，第18章介绍用标注依存关系的树库来评价MINIPAR句法分析器性能的实

此为试读，需要完整PDF请访问：www.ertongbook.com

验。第19章到第21章讨论的是从树库中抽取语法知识的问题，第19章介绍直接将树库看作是一部概率语法来进行句法分析的实验，第20章介绍从树库和HPSG规则中抽取词汇化概率树语法的方法，第21章介绍从树库资源生成词汇功能语法（LFG）功能结构（F-structure）标注语料库的方法。

3 章节内容详细介绍

导 言

标注语料库相对于普通的未标注语料库，对自然语言处理和语言学研究的价值更大。本书的21篇文章都是讨论标注语料库特别是句法结构标注语料库的，涉及的问题包括：(1)如何选择语料库进行标注？(2)选择标注什么内容？(3)手工标注还是自动标注？用什么辅助工具？采用什么标注格式？(4)如何检索标注语料库？(5)从标注语料库中可以抽取什么语法知识，与从未标注语料库中抽取知识相比有何优越性？(6)如何利用标注语料库对NLP工具，比如句法分析器（parser）或语法检查软件（grammar checker）进行评估？

第1章 宾州树库概述

从1989年到1996年，历时8年，美国宾州大学树库（The Penn Treebank）建成，包括约700万词的带词性标记语料库，300万词的句法结构标注语料库（树库），200万词的谓词—论元结构标注语料库和160万词的非流利口语转写带语音标记语料库。本章介绍了宾州树库的三个标注规范：词性标记规范，句法结构标记规范和非流利口语语音标记规范。此外对语料库加工的方法也做了介绍，词性标注、句法结构标注、非流利口语语音标注都采用的是自动标注和人工校对相结合的方式。自动词性标注先后采用过PARTS词性标注程序（Church 1988）[①]和基于转换的错误驱动的词性标注程序（Brill 1993）[②]。句法结构标注采用的是Fidditch句法分析器（Hindle 1989）[③]。非流利口语语音标注采用的工具是简单的Perl脚本程序，先利用程序将非句子成分自动加上标记，之后再利用一个跟句法结构标注类似的图形界面程序，可

① Kenneth W.Church, 1988, A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, In *Proceedings of the Second Conference on Applied Natural Language Processing, 26th Annual Meeting of the Association for Computational Linguistics.*

② Eric Brill, 1993, A Corpu-based Approach to Language Learning, Ph.D Dissertation, University of Pennsylvania.

③ Donald Hindle, 1989, Acquiring Disambiguation Rules from Text, In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguisitics.*

以让标注人员方便地手工加注语音标记。作为全世界最早开始的树库加工计划，宾州树库的加工规范、加工方法都带有示范的意义。

第2章　树库构建20年的思考

通过对过去20年参与树库加工工作的反思，本章作者提出，真实语料所反映的结构事实跟当代理论语言学秉持的一些假设是相左的。自然语言很难说是一个被定义得很完美的系统（well-defined system）。短语结构的分析也并不总是有“正确答案”。从树库中揭示的语言事实是有树库之前没有认识到的。作者批评以乔姆斯基为代表的理论语言学者缺乏对短语结构进行系统分类的兴趣。一些理论语言学者认为语言远远没有表面上的那么复杂，语言的底层结构可以用有限的规则去定义。这种具有遗传特性的“心理语言机制”是适用于所有自然语言的共性。从语言的“深层结构”这方面来说，英语和捷克语的复杂程度几乎不会超过像Pascal或Java这样的程序设计语言。但是树库加工的实践经验表明这种理论假设与事实不符。语言中的规则不是法律或化学意义上的规则，它是灵活可变的。语言的使用者可以改变甚至不遵守规则。即使是小规模的树库语料，都可以通过结构的频次统计，得出与以往的理论语言学研究看法不同的结论。比如通过统计SUSANNE[①]树库中“主语—及物动词—宾语”和“主语—不及物动词”结构的频次发现，后者并不是跟前者地位同等的常用结构（而这却是一般语法书上通行的看法）。在真实语料中，如果谓语动词后面没有带宾语的话，也总是带有其他句法成分。再比如通过统计CHRISTINE[②]口语树库可以发现语言结构的复杂性跟年龄的关系更密切，即随着人的年龄增长，所使用语言结构的复杂程度逐渐提高，而不是像之前的研究所声称的那样：结构复杂性跟人所处的社会阶层密切相关。

第3章　BOE英语语料库及后续加工

BOE（Bank of English）英语语料库是由Collins出版社资助的国际性英语语料库加工项目。从1993年到1995年，BOE语料库中有2亿词语料做了词形分析和句法标注。词形分析采用的是一个双层词汇形态分析器，句法标注采用的是一个基于英语约束语法（ENGCG）的浅层句法分析器。整个项目的应用目标是将BOE标注语料库用于COBUILD英语词典第二版的编纂。

---

① SUSANNE是“Surface and Underlying Structural Analysis of Natural English”（自然英语的表层和基础结构分析）的首字母缩写（参考http://www.grsampson.net/RSue.html）。

② 有关CHRISTINE口语树库名称的来历请参考http://www.grsampson.net/RChristine.html。

1995年BOE句法标注的任务结束之后，研究人员对ENGCG进行了改进，设计了新的基于规则的功能依存语法分析器(FDG)，FDG考虑了更多的复杂语法现象(如非投射结构、长距离依赖、省略结构和同形成分省略等)，比ENGCG的语法规则覆盖率更广，在错误率没有增加的情况下，FDG产生的分析结果中歧义数量仅是ENGCG的四分之一。小规模的测试显示，FDG比当时最好的基于统计的句法分析器的效果更好。

第4章 完成句法分析语料库：从单文本遍历式校对到跨文本定向式校对

本章介绍国际英语语料库中英国部分(ICE-GB)的100万词语料库加工成树库的过程，并对人工校对的方法进行了深入讨论。传统的人工校对是对每个文件中的每个句子逐句进行校对(我们称之为单文本遍历式校对)，这种方式的缺点是工作量非常大，同时容易造成前后不一致的问题。本章提出了一种基于结构的校对方法，即对跨文本的树库语料进行检索得到同类标注问题的实例，然后由校对人员来做统一的修改(我们称之为跨文本定向式校对)，这样同类问题的处理方式前后一致的。不过，这样虽然可以带来更好的一致性，但相应的辅助工具实现难度更大，多人同时校对时还需要解决管理问题，此外程序还会出现误报的假性错误；传统的单文本遍历式校对方式则没有这些问题。

第5章 德语新闻报纸语料库的句法标注

本章介绍了德语新闻报纸语料库(NEGRA)的句法标注。NEGRA包含超过2万个句子(35万词)。句法标注框架采用的是上下文无关的短语结构和依存语法结构的混合体，在标注句中短语结构类型(如NP、VP等)的同时，还标注了短语之间的依存关系(如主语、宾语等)。跟一般上下文无关文法不同的是，NEGRA的句法树标注允许树分支交叉，这是跟德语词序灵活的特点相适应的，在德语实际语料中存在着大量的“未完成”结构(incomplete structure)。允许分支交叉，以及倾向采用更为扁平的多分支树结构(相对于二分支树结构)来进行标注，是NEGRA的特点。NEGRA的标注过程大致分为3个阶段：(1)预处理；(2)基于图形界面的人机交互式标注；(3)通过比较确定最终标注结果，即每个句子都有两个人分别独立进行标注，然后进行比较，再决定最终标注结果。在第2个环节的交互式标注中，NEGRA利用基于统计的词性标注软件和句法分析软件，对每个句子的多种分析结果给出概率评分，并高亮显示程序认为不可靠的分析结果。在计算机界面上呈现分析结果时采用分段逐步推进的方式。这些手段都显著地提高了标注效率。

第6章　德语新闻组语料库的错误类型标注

德国政府资助的FLAG项目旨在开发德语受限语言(controlled language)和语法检查技术。为此,项目组收集了互联网新闻组(USENET)电子邮件语料约12万句。在确定了一个含16种错误类型的标记集后,先在纸上由人工对这12万句进行了标注,即把句中的各类错误(如形态错误、句法结构错误、句法语义选择错误、拼写错误等等)加上标记,共标记了约6万句。然后在计算机标注工具[①]的辅助下,将纸上标注的6万句中的14,492句录入计算机并核验是否符合标注要求,形成了一个带错误类型标记的新闻组语料库。具体标注信息包括每句的错误数、错误位置、错误范围等等。对错误类型的初步统计显示,真实电子邮件语料中的错误83%是纯粹的拼写错误,语法错误的比重不高,仅为16%。FLAG的项目实践表明,先在纸上快速标注真实语料中的错误,再利用计算机标注工具制成最终的带错误类型标记的语料库,是行之有效的语料库标注方法。

第7章　布拉格依存树库的构建:一个三层标注方案

布拉格依存树库(PDT)分三个层级对捷克语语料进行标注。第一层是形态标注;第二层是表层句法标注,采用依存语法架构;第三层是语义标注,仍然采用依存树的表达形式,理论框架则是所谓的功能生成描述(functional generative description)。PDT的语料来源是捷克语国家语料库(CNC)。语料分布为:普通报纸语料占60%,经济新闻和分析占20%,科学杂志语料占20%。项目计划从1996年到2004年,标注第一层的语料规模要达到180万词,标注语义信息的语料规模要达到100万词。本章成稿时,项目完成了约三分之二的标注任务。所有三层标注均采用SGML语言作为格式规范。第一层形态标注对句子中的每个实例都标注两个信息:词形(lemma)和一组形态范畴特征值(MTag)。第二层表层句法标注要给出句子的依存关系树。在列举了依存语法树需要遵循的基本原则后,本章还介绍了PDT对依存语法理论不易处理的一些特殊语法现象(如"并列结构")的标注策略。在积累了一定量的表层句法标注语料后,就可以借助Collins的词汇化概率语法分析器进行训练然后对未标注语料进行依存关系标注,正确率可以达到80%。第三层语义标注要在句子的依存语法树基础上继续在树节点标签上标注更多的语义和情态范畴属性。PDT设定了约40个语义功能项(比如行动者/承受者、效果、使因等),此外还有时态、数、性、比较级等语法情态属性项的标注。

---

① 在评估了DiET和Annotate两个均支持图形界面的标注工具的特性后,项目组选择了前者作为辅助工具。

第8章 一个标注HPSG特征结构的波兰语测试语料库

本章介绍了一个波兰语书面语句子测试集:BRG(波兰语直译即语法分析数据库)。作为句法分析测试句集,BRG不仅包含合语法的句子(193句),还包含不合语法的句子(147句)。句子并不是来自真实语料,而是根据语言学评测分析的需要人为设计的。句子按照复杂程度分为基本句型、复杂句型、特别复杂句型三个类别。每个句子都以人工方式标注了是否正确(其中不合语法的句子都有对应的合语法的句子),包含哪些语法现象(所有语法现象可以在一个索引表中查到,共计264种),以及采用HPSG的特征结构框图(AVM)标记句中的语法成分及语法属性。在BRG的图形界面上,句子可以呈现其短语结构树形式以及AVM特征结构。BRG可以评测波兰语形式语法的覆盖率,其设计出发点是针对基于HPSG的语法框架,不过已经实现的BRG也可用于评估基于其他形式化语法框架(如定子句语法、依存语法)的波兰语语法系统。

第9章 西班牙语树库的开发

本章介绍为构建西班牙语新闻语料树库所开发的规范和工具。该树库包含来自新闻报纸和杂志的1500个句子,共计22,695个词。这1500句中一部分是孤立的单个句子,没有上下文;另一部分是在自然段落中的,有上下文的句子。树库标注规范涉及的问题包括语言单位的识别以及信息标记方式(西班牙语树库参照了宾州树库的标记模式)。每个树节点标签由"语类范畴"(CAT)、"语言成分字符串"(string)以及后面带有的一组特征(feature)描述组成。除一般性的通用标记框架外,本章还介绍了西班牙语树库中一些特殊结构(比如"Se"字结构)的标记方式。西班牙语树库的开发借助了项目组自己开发的工具和一些公开资源,分为两类:(1)标注工具,包括形态句法标注系统和语块识别软件(chunker);(2)查错工具,包括图形化的树结构编辑工具、特征核查工具、短语结构规则生成工具等。

第10章 构建法语树库

本章介绍了一个法语树库加工的情况。该树库来自法语新闻语料,100万词规模。加工过程分为词汇层标注和句法结构层标注两个阶段。词汇层加工流程为:首先对原始文本进行断句处理,然后进行词性标注、形态标注、未登录词标注、合成词标注等,词汇信息标注的多个环节都在自动处理的基础上进行了人工干预,并将得到的正确标注结果记录到词库中,这样可提升词库对后续语料的处理能力。句法结构层加工流程为:对经过词汇信息标注

处理的文本，首先基于手工规则进行浅层分析，然后对得到的初步结果进行人工校对，再基于手工规则和法语配价词典进行功能标注，得到树库文本，经过人工校验后得到最终版本。本章还较为详细地介绍了法语树库的语法功能标记集，并讨论了一些具体结构类型的标注方式（比如非连续成分结构、并列结构等等）。

第11章　构建意大利语句法—语义树库

本章介绍了一个意大利语句法—语义树库（Italian Syntactic-Semantic Treebank，ISST）。ISST包含四个层级的标注信息：形态句法层、短语结构层、语法功能关系层、词汇语义层。语料规模达30万词，其中21万词为平衡语料，9万词为金融领域语料。形态句法层的标记集从16个基本词类标记扩展到31个子类标记，再加上形态句法特征标记，共236个标记。短语结构成分（共22类）的标注跟语法功能关系（共9种关系）的标注是分别独立进行的。短语结构成分先由浅层句法分析器自动识别，然后由人工修改。语法功能标注则采用依存关系模型，描写词语之间的依存关系。词汇语义层标注主要针对名、动、形以及多词复合表达式等实义性语言单位。具体标记包括三种信息：（1）从意大利语词网（ItalyWordNet）中获取的词的义项；（2）词的比喻义、惯用法、旧词新意、专名等；（3）标注操作相关信息（如记录标注者信息）。除标注内容外，本章还介绍了用于标注的软件工具，以及将ISST用于意—英机器翻译系统测试其效能的情况。

第12章　构建中世纪葡萄牙语树库

本章是对古代语言进行词法形态分析、词性标注和句法结构标注的一次尝试。通过使用针对当代葡萄牙语开发的词法分析工具、词性标注工具和层叠式浅层句法分析工具，以及相关的语言学资源（主要是词法规则知识库LKB），作者对中世纪葡萄牙语（主要是应用文文本，如遗嘱、赠品清单、法律文本等）进行了词法分析、词性标注和浅层句法结构的自动标注。本章的探索表明，在同一个语言不同时期的变体之间，或者不同语言之间，如果其相似度足够高，就可以利用已经标注好的树库资源，从中获取相关语言知识，来帮助加工新的树库语料。

第13章　SINICA中文树库

Sinica中文树库是跟宾州中文树库差不多同时开始构建的最早的中文树库之一。Sinica中文树库在设计时主要考虑了三个问题：最大资源共享、最小

此为试读，需要完整PDF请访问：www.ertongbook.com