国家社会科学基金项目成果

基于双语语料库的汉英体转换规则的形式研究

瞿云华 著

科学出版社

国家社会科学基金项目成果

基于双语语料库的汉英体转换规则的形式研究

瞿云华　著

科学出版社

北京

内 容 简 介

本书采用计算语言学和语料库语言学的方法来研究“体”，提出了一套分类与语法意义基本统一的汉英对应的体系统（aspect system），构建了具有1万多对做了体标注的汉英双语平行语料库，提出了汉语体的派生原则，揭示了体派生的语用原因，构建了两套形式化的汉英体转换规则，从新的角度来研究、描述和解释汉语的体和英语的体及它们之间的转换关系。研究中还使用了6种机器学习方法来检验汉英转换规则的鲁棒性，以及统计建模的方法筛选体分类特征，把基于规则的理性主义方法和基于统计的经验主义方法结合起来，符合当前自然语言处理发展的总趋势。

本书的目标对象为语言学及应用语言学研究者、博士生、研究生，以及自然语言处理研究者、博士生、研究生等。

图书在版编目(CIP)数据

基于双语语料库的汉英体转换规则的形式研究 / 瞿云华著. —北京：科学出版社，2016.3

ISBN 978-7-03-043556-9

I. ①基… II. ①瞿… III. ①汉语–语料库–翻译–研究 ②英语–语料库–翻译–研究 IV. ①H159 ②H315.9

中国版本图书馆CIP数据核字(2015)第045966号

责任编辑：阎 莉 常春娥 / 责任校对：李 影
责任印制：肖 兴 / 封面设计：刘可红

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
科学出版社发行 各地新华书店经销
*
2016年3月第 一 版 开本：A5(890×1240)
2016年3月第一次印刷 印张：11 3/8
字数：380 000

定价：88.00元

(如有印装质量问题，我社负责调换)

作者简介

瞿云华，女，1961 年 7 月生，浙江杭州人，浙江大学外国语言文化与国际交流学院语言与翻译系教授，博士生导师，师从著名计算语言学家冯志伟，2007 年 7 月在中国传媒大学获计算语言学博士；2009 年 3 月~2009 年 9 月获世界大学联盟研究交流基金在英国利兹大学翻译中心任访问学者，2012 年 1 月~2013 年 1 月获国家公派访问学者（含博士后研究）项目在美国北亚利桑那大学语料库语言学研究中心任访问学者；主要从事计算语言学、语料库语言学、汉英时体对比、社会语言学等研究。

曾主持国家社会科学基金课题一项，参与省部级重点课题一项，在 SCI 检索的 *Journal of Zhejiang University Science-C*，SSCI 检索的 *Journal of Quantitative Linguistics*，一级期刊《浙江大学学报》（人文社会科学版）、核心期刊《外语研究》《外语学刊》《外语教学》、*Teaching English in China*，EI 检索的国际计算语言学会议等，发表论文近 20 篇，总计被引频次 142 次，其中一篇被中国人民大学复印资料全文转载，在科学出版社出版专著一部。

本书出版承蒙

浙江大学外国语言文化与国际交流学院

精品学术著作出版基金

与浙江大学董氏文史哲研究奖励基金资助

序　言

2008年瞿云华申报的国家哲学社会科学基金课题“基于双语语料库的汉英体转换规则的形式研究”得到立项，接着她就得到了国家留学基金与世界大学联盟研究基金的资助。她先后在英国利兹大学访学半年，在美国北亚利桑那大学访学一年，扩大了学术视野。借助于天时地利，她在6年的时间内，辗转于中国、英国和美国几所大学，一直专注地进行汉英体转换规则的形式研究，现在终于顺利地、圆满地完成了这个课题，写出了这本专著。

语言中“体”(aspect)一直是传统语言学本体研究关注的内容，积累了汗牛充栋的文献，取得了累累的成果。瞿云华原是我的博士研究生，熟悉计算语言学和语料库语言学的理论和方法。她本人的语法理论基础比较好，她把传统的语法理论研究与她在读博士期间学习的专业知识结合起来，另辟蹊径，采用计算语言学和语料库语言学的方法来研究“体”，提出了一套分类与语法意义基本统一的汉英对应的体系统(aspect system)，构建了具有1万多对做了体标注的汉英双语平行语料库，提出了汉语体的派生原则，构建了两套形式化的汉英体转换规则，从新的角度来研究和描述汉语的体和英语的体及它们之间的转换关系，取得了可喜的成果。

更加值得我们高兴的是，瞿云华在她的研究中还使用了决策树、随机森林算法、朴素贝叶斯分类算法、支持向量机、AdaBoost分类器算法、高阶的有权重的脊回归方法等6种机器学习方法来检验规则的鲁棒性，把机器学习方法运用到汉英转换规则的测试中，这样的机器学习方法实质上是一种统计方法。

所以，我认为，瞿云华的这本著作，是基于规则的理性主义方法与基于统计的经验主义方法结合的产物。

在大数据环境下，20世纪90年代在自然语言处理(natural language processing)中越来越多地采用基于统计的研究方法。这样的统计方法在21世纪进一步以惊人的步伐加快了发展速度。我认为，这样的加速发展在很大

程度上受到下面3种彼此协同趋势的推动。

首先是建立带标记语料库的趋势。在语言数据联盟(Linguistic Data Consortium)和其他相关机构的帮助下，自然语言处理的研究者可以获得口语和书面语的大规模的海量语料。在这些海量语料中还包括一些标注过的语料，如宾州树库(Penn Treebank)，布拉格依存树库(Prague Dependency Tree Bank)，宾州命题语料库(PropBank)，宾州话语树库(Penn Discourse Treebank)，修辞结构库(RST-Bank)和时态库(TimeBank)。这些语料库是带有句法、语义和语用等不同层次标记的标准文本语言资源，其中蕴藏着丰富的语言学知识。这些带标记的语言资源大大地推动了人们使用有监督的机器学习方法(supervised machine learning)来处理那些在传统上非常复杂的自动句法分析和自动语义分析等问题。这些语言资源也推动了有竞争性的评测机制的建立，评测的范围涉及句法自动分析、信息抽取、词义排歧、问答系统、自动文摘等诸多领域。

其次是统计机器学习的趋势。在大数据的环境下，对于机器学习日益增长的重视，导致了自然语言处理的研究者与统计机器学习的研究者更加频繁地交互，彼此之间互相切磋，互相影响。对于支持向量机技术、最大熵技术及与它们在形式上等价的多项逻辑回归、图式贝叶斯模型等技术的研究，都成为自然语言处理研究的重要内容。

再次是高性能计算机系统发展的趋势。在大数据环境下，高性能计算机系统的广泛应用，为机器学习系统的大规模训练和效能发挥提供了有利的条件，而这些在20世纪是难以想象的。

最近，大规模的无监督的机器学习方法(unsupervised machine learning)得到了重新关注。在机器翻译和文本主题模拟等领域中统计方法的进步，说明了除了使用带标注的语料库之外，也可以训练完全没有标注过的语料库来构建机器学习系统，这样的系统也可以得到有效的应用。由于建造可靠的带标注语料库要花费很高的成本，建造的难度很大，在很多问题中，这成为使用有监督的机器学习方法(supervised machine learning)的一个限制性因素。因此，这个趋势的进一步发展，将使人们有可能更多地使用无监督的机器学习方法来降低建造语料库的成本。

由此可见，基于统计的经验主义方法已经成为当前自然语言处理研究的主流。

在这样的情况下，一些使用基于统计的经验主义方法取得成功的学者开

始头脑发热，贬低基于规则的理性主义方法。IBM 公司的 Fred Jelinek 是一位使用统计方法研究语音识别与合成的著名学者，他在统计自然语言处理研究中取得的成绩是人所共知的，我们都很佩服他的成就。可是，他却看不起使用规则方法研究自然语言处理的人。他于 1988 年 12 月 7 日在自然语言处理评测讨论会上的发言中曾经说过："每当一个语言学家离开我们的研究组，语音识别率就提高一步。"（Anytime a linguist leaves the group the recognition rate goes up.）根据一些参加这个会议的人回忆，当时 Jelinek 的原话更为尖刻，他说："每当我解雇一个语言学家，语音识别系统的性能就会改善一些。"（Every time I fire a linguist the performance of the recognizer improves.）Jelinek 的这些话，把基于规则的自然语言处理研究贬低到了一无是处的程度，把从事基于规则的自然语言处理研究的语言学家，贬低到了一钱不值的程度，他对于基于规则的自然语言处理，采取了嗤之以鼻的态度。①

Jelinek 的这种言论不无偏颇；我们了解的情况与他的言论并不一致。

在机器翻译的研究中我们发现，每当一个有坚实语言学素养的语言学家加入我们的研究团队，给我们提供可靠的、形式化的短语规则和句法规则，机器翻译的译文的忠实度和流畅度就会提高一步，而每当一个有坚实语言学素养的语言学家离开我们的研究团队，机器翻译的译文质量就会明显地降低。

事实证明，语言学家是自然语言处理研究不可缺少的，关键在于这样的语言学家应当是有语言学素养的高水平的语言学家，语言学家的素养越高，他们提供的语言学规则越是科学，对于自然语言处理的研究越有帮助。我们决不可轻视语言学规则在自然语言处理中的重要作用。

因此，我不同意 Jelinek 的主张。我们认为，在自然语言处理研究中，应当把基于规则的方法和基于统计的方法结合起来，把语言学、数学和计算机科学紧密地结合在一起，取长补短，相得益彰。

我们高兴地看到，目前在基于统计的方法中，研究者开始更加自觉地引入语言学信息。

首先，在统计机器翻译中提出了基于短语的统计机器翻译模型，这种模型把语言学中的短语（phrases）作为翻译的原子单元。在短语翻译表中，短

① M. Palmer and T. Finin, workshop on the evaluation of natural language processing systems, Computational Linguistics, 16(3), 175-181, 1990.

语之间是一一映射的，也可能存在调序。短语翻译表可以从词对齐中通过机器学习而自动地得到，与词对齐一致的所有短语偶对都被添加到短语翻译表中。在扩展原始的翻译模型时，还引入了额外的模型组件，这些组件包括双向翻译概率、词汇化加权、词惩罚和短语惩罚。

其次，为了在基于统计的方法中引入语言学信息，在统计机器翻译中，还提出了整合语言学知识的问题，其中包括利用句法标注的语言学信息来提高统计机器翻译的质量；在基于短语的统计机器翻译中，融入字母翻译、词汇翻译和句子结构等语言学知识。如果源语言和目标语言在词序方面差别明显，还可以使用基于句法的方法来调序。当处理句法树的重构时，可以使用子结点调序限制来降低计算的复杂性，也可以使用重排序(re-ranking)方法，在挑选最佳翻译时利用语言的句法特征，检查输入和输出的一致性，等等。

由此可见，在基于统计的方法中引入语言学信息，可以弥补统计方法的不足，使基于统计的方法如虎添翼。在大数据环境下，把基于统计的方法与基于规则的方法紧密地结合起来，是自然语言处理研究取得成功的关键。

2012 年 6 月，《纽约时报》披露了“谷歌大脑”(Google Brain)项目。这个项目用 16 000 个 CPU Core 的并行计算平台，训练一种称为“深度神经网络”(Deep Neural Networks, DNN)的机器学习模型。在神经网络中，该模型把基于规则的方法和基于统计的方法在更深的程度上结合起来。目前深度神经网络在语音识别和图像识别等领域已经获得了巨大的成功。

2012 年 11 月，微软在中国天津的一次活动上公开演示了一个全自动的同声翻译系统，讲演者用英文演讲，后台的计算机自动完成英语语音识别、英汉机器翻译和汉语语音合成等自然语言处理过程，把演讲者说的英文翻译成汉语普通话由计算机流畅地讲出来。据这位演讲者透露，后面支撑的关键技术也是深度神经网络，这种深度神经网络就是一种深度学习(Deep Learning, DL)模型，同样也是基于规则的方法和基于统计的方法在更深层度上的结合。

瞿云华的这本著作，在建立双语并行语料库的基础之上，应用统计方法来进行汉语和英语体转换规则的形式化研究，用机器学习方法来测试汉语和英语的体转换规则，把基于规则的理性主义方法和基于统计的经验主义方法结合起来，符合当前自然语言处理发展的总趋势，这样的研究无疑有着光辉的前景。

希望该书的出版，能够进一步推动基于规则方法和基于统计方法在更深程度上的结合，使中国的自然语言处理研究开出更加绚丽的花朵，结出更加丰硕的果实。

冯志伟

2015 年清明于杭州

前　言

本书所研究的“体”即通常所说的体，也就是 Smith(1991)所称谓的视点体(viewpoint aspect)，与动词相关，是一种狭义的体，属于广义的体(体貌，aspect)范畴。它的下位概念有体标记(“着、在、正在、了”等)和谓语动词、少量的具有谓词性的名词、形容词的组合表示的未完整体、进行体、完整体、完成体。但不包括使用专门语法手段(虚化的趋向成分“起来、下来、下去”、补语性的“完、好、过、到、得、着”、动词重叠“说说”、复叠“说说笑笑”)描述动作基本阶段的阶段体(起始体、延续体、完结体、结果体、短时体、反复体)[①]，也不包括根据参照说话时间来定位事件外部时间意义的时制(过去时、现在时、将来时)。

体是一种对事件内部时间的观察部位的表现，通俗地说是一种对动作是在进行、持续之中，还是已经完成之描述。对这三种状态的描述分别对应了体的三种主要类型：进行体、未完整体和完整体。汉英两种语言分属不同的语系，对体有截然不同的表示形式和系统。汉语的体形式由体标记“着、在、正在、在、了、过等”+谓语动词组成。英语的体形式有 be+V-ing 和 have+V-ed 等。

当今，中国的经济、文化、科学技术正日益受到世界的关注，大量宣传介绍中国情况的中文资料需要翻译成英语。然而，在汉译英时汉语的一些常用体和英语体存在着一对多的映射现象：一种汉语体可以翻译为多种英语的体。比如，由“在＋V”组成的进行体可以翻译为英语的进行体、完成体、完成进行体等。因而，初学翻译者和英语学习者往往感到无规律可循，不知如何准确地将汉语体转换为英语的体。商品化的机器翻译系统在体翻译时也常常出现各种错误，因此，急需体转换规则，以提高翻译的正确率。

体研究历来是语言研究的难点和热点，汉英体的对应研究却很少有人问津，转换规则的研究更是鲜为人知。综观汉英体学界、体对比研究界、汉英翻译研究界、计算语言学界、语料库语言学界、机器翻译研究界，汉英体研究存在着以下三个研究难点。

---

① 详见 1.3.3 中有关陈前瑞对阶段体划分的观点。

一是目前汉英体没有一套公认且行之有效的对应系统，可以作为描写和分析对比体现象的框架。现有的那些研究采用的汉语体系统，不适合进行汉英体的对比研究。它们的分类标准和英语体的分类标准不相同，造成有的汉语体名称和英语体名称相同，但意义却大相径庭。因此，这样的分类体系无法和英语的体系进行对比，更谈不上总结转换规则。

二是这些研究大多没有采用语料库。语料库能为汉英体的对比找到丰富的实例，并有助于找到体转换的规律。

三是这些研究没有关注体之外的附加成分对体意义的影响。附加成分可以使原有体派生为其他体类型，继而翻译成其他体类型。这些附加成分是汉英体转换的语义条件，即转换规则的条件。

本书的研究针对上述三个难题，提出了一套分类和语法意义基本统一的汉英对应的体系统，构建了一个具有 1 万多对含有汉语体标记的汉英对应的平行语料库，提出了一个汉语体派生原则，揭示了汉语显性体标记与动词和显性附加成分、时间状语之间的语义关系，并以此为基础构建了两套形式化的在文学作品和白皮书语料中测得准确率在 77% 以上的汉英体转换规则。

本书提出的汉英体系统是在对陈前瑞(2003)、Comrie(1976)、Smith(1997)、Olsen(1997)体系统的综合改进基础上构建而成的，具有汉英两种体系统的分类。在这套体系统中，汉语体分为未完整体、进行体、完整体和完成体。英语体分为一般未完整体、进行体、一般完整体、完成体和完成进行体。而且分类和语法意义基本统一和平行，解决了以往汉英体系统中找不到体类型对应的难题，符合汉英两种语言事实，适用于汉英体的转换规律研究。

本书提出的汉语体派生原则，解释了上述汉英规则的转换机理是汉语体发生了派生现象。派生现象的动因是汉语体的语句中含有一些附加的时间状语，这些附加时间状语的语义强度超越了基本体的语义，使之发生体变异，成为派生体。上述转换规则中的时间状语的语义特征，其实质上就是汉语体派生的语义条件，也就是规则转换的深层原因。

本书提出的形式化的汉英体转换规则描述了在汉英翻译中一种汉语体可以翻译为几种英语体的现象。汉英体转换规则是多组能够根据语义条件进行一一对应转换的规则。每一组规则都由“时间状语+动词+体标记”这样的

结构组成。本书以体标记“着”和“了”为例[①]，进行汉英体转换规则的描述，因而，转换规则包含了这两个体标记的所有四组规则，包含了这两个体标记与各种语义类型的动词及各类时间状语的组合。

在这几组规则中，这两种体都可以根据不同的语义条件转换为多种英语体。每一组语义特征都对应了一组从汉英体对应语料库中直接抽取的时间词、时间副词和频率副词实例。使用者只要根据具体的时间词、时间副词、频率副词及对应的动词类型、基本体类型就能求得对应的英语体类型。

本书采用了冯志伟(1983)提出的多叉多标记树模型(Multiple-branched and Multiple-labeled Tree, MMT)和特征合一(Unification)方法，运用了Prolog(Programming in Logic)逻辑语言含有的PATR语法规则，在机器上实现了形式化“着”规则和“了”规则例句。

这些形式化的上下文无关语法(Context Free Grammar, CFG)结构树，不仅包含了原有的句法信息、词汇信息，而且包含了体类型信息、动词和时间状语词例的语义特征。经过MMT扩展的CFG结构树的树叶节点上分别标记了动词、体标记和时间状语的语义特征和词汇信息。

CFG结构树通过自下而上的合一过程，不仅将树叶上的动词、体标记与时间状语词例的词汇信息合一成了句法信息，并经过层层传递到最上层的树杈节点，而且也将词汇的语义特征经过层层合并和传递把“动词+体标记”的基本体类型和语义特征，以及“动词+体标记+时间状语”的派生体类型和语义特征，经由树杈节点传递到了最上层的树杈节点上；显示了VP结构是如何从“动词+体标记”组合成基本体，又从基本体派生为派生体的过程；反映了体标记与动词、时间状语及对应的目标英语体类型之间的语义关系。

本书采用了六种机器学习算法，即决策树、随机森林算法、朴素贝叶斯分类算法、支持向量机、AdaBoost分类器算法、高阶的有权重的脊回归方法，证明了“着”规则和“了”规则的鲁棒性，准确率分别达到了77.8%和77.7%以上，AUC曲线下平均面积达到了75%以上[②]。运用统计方法建模将“着”测试特征集从原始的40组删减到了14组，删除了“了”测试特征集中的相冲突特征2个，进一步论证和筛选了测试特征。

---

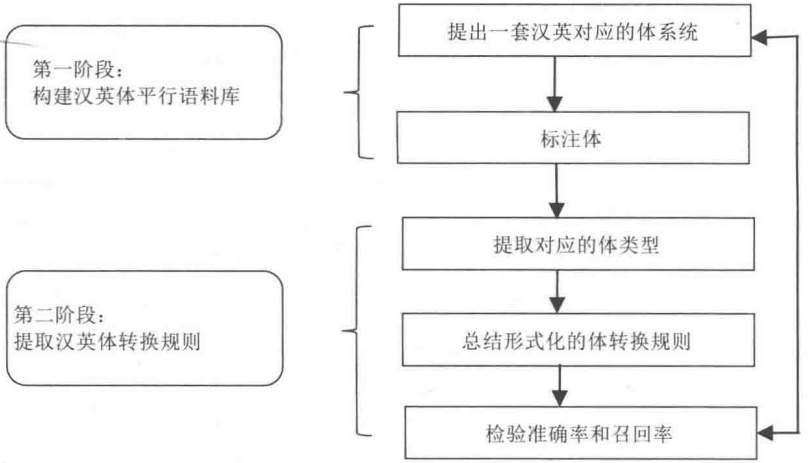
① 由于本书所抽取的进行体标记“在+V” “正+V” “正在+V”“S+呢”标记数量分别只有211、55、43、83，无法运用机器学习进行测试。因而，本书的转换规则的描述、形式化和机器学习测试均以“着”和“了”两个体标记为例。

② AUC曲线下面积是一种比召回率更有效和流行的检测准确率的方法，因此，本书补充测试了AUC曲线下的面积。

综上所述，本书的研究以一个汉英平行语料库为基础，采用统计分析、描述、解释语言事实、定量分析和定性研究相结合等研究方法，运用了语料库语言学、体理论、计算语言学、人工智能方法、统计学中的虚拟变量、线形回归建模方法等多学科知识，是一个跨学科的研究。特别是采用统计方法进一步测试是据笔者所知，迄今为止统计方法首次在语言特征筛选方面的应用。可以说，这也是本书有别于其他传统的体研究和机器翻译研究之处。

本书研究采用了双语语料库，可以得出传统的对比语言学研究方法所无法得出的结论，从而使我们对自然语言中汉语和英语的各种复杂的体映射现象获得更为深刻、全面的认识。本书提出的一套形式化的汉英体转换规则，可为机器翻译研究者在处理这一翻译难题时提供基础性的语义分析，使他们能在此基础上做进一步的深入研究，还可为英语学习者、初学翻译者及汉语作为第二语言学习者提供汉英体转换，以及汉语体和时间状语搭配的使用模板，因而，对于提高对外汉语教学质量、英语教学质量和汉英翻译水平，具有实际应用价值。本规则只是一种探索性的尝试，如果要应用到机器翻译上还有很长的路要走，还需要自然语言处理者们的不懈努力。

本书的研究采用了一个大型汉英双语平行语料库。该语料库含 89 万个汉字和 62 万个英语词，由 18 部中国政府颁发的白皮书和 5 部当代中国文学名著的汉英文本组成。整个研究工作步骤划分为两大阶段如下所示：



第一阶段：构建汉英体平行语料库

(1)提出一套汉英对应的体系统。首先梳理前贤的研究成果，提出一套分类和语法意义基本统一的汉英对应的体系统。结果是在双语语料库中，运用该体系统，除一些英语的非体结构以外，汉语体都能找到对应的英语体形式。

(2)标注各类体形式。确定具体的检索词，从大型的双语语料库中抽取汉语体和英语的对应语句，并对抽取结果进行人工体标注。

第二阶段：提炼汉英体转换规则

(1)提取对应的体类型。从已标注的汉英体平行语料库中抽取汉英体对应的实例。

(2)总结形式化的汉英体转换规则。描述和解释为什么在汉英翻译中会有一种汉语体翻译为多种英语体的现象，确定汉语体在转换为各种英语体时的语义条件，建立一一对应的汉英体转换规则，运用 CFG 文法和 MMT 模型使之形式化，并采用 CTT(Copenhagen Tree Tracer，哥本哈根句法树跟踪显示程序)软件中内置的 Prolog 逻辑编程语言中的 PATR 语法规则编程，实现在机器上自动形式化汉英体转换规则。

(3)检验规则的准确率和召回率。运用专门的计算机程序建立训练集和测试集。通过机器学习算法在训练集获取语言知识，在测试集上自动检验规则的准确率和召回率，并再反馈测试结果，修正规则，进一步测试，直至得到足够的准确率和召回率。采用统计方法建模优化特征，为以后进一步的测试研究提供有足够准确率的特征。①

总之，汉英体平行语料库为映射研究提供汉英体对应的语言事实，语言学理论为从这些对应事实中归纳规则和语义解释提供理论支持。

依据上述研究阶段的划分，本书的结构主要分为四部分。第一部分是前言，介绍本书的研究背景、主要研究成果、研究意义、研究方法及全书的框架结构。第二部分(第一篇)介绍了体概念、汉英对应的体系统、汉英对应的体语料库、汉英体转换规则和汉语体派生原则。第三部分(第二篇)介绍了如何运用 CFG 语法、特征合一、MMT 方法与 Prolog 逻辑语言内置的 PATR 语法规则合用进行编程，在机器上实现“着”和“了”规则的形式化，并展示了所有汉英体转换规则例句剖析的结果截屏。第四部分(第

① 采用统计方法建模优化特征，为以后进一步的测试研究提供有足够准确率的特征——这一研究任务不在国家哲学社会课题申请书的计划之中，是本书的额外成果。

三篇)描述了如何运用机器学习算法测试“着”规则和“了”规则，以及运用统计方法建模优化特征的方法和结果。

由于时间仓促和作者的研究局限，本书难免有疏漏之处，敬请同行专家批评指正和读者谅解。

瞿云华

2016 年 1 月

符号对照表

@nucleu：“@”表示视点的位置，语法体的视点落在核心阶段

[+dur]：[+durative][+持续]

[+dyn]：[+dynamic][+动态]

[+imperfective]：[+未完整体]

[+pastel]：[+past telic][+过去终结]

[+perfective]：[完整体]

[+prestel]：[+present telic] [+现在终结]

[+spectim]：[+specific time][+时点]

[+stat]：[+state][+状态]

[+telic]：[+终结]

[-dur]：[-durative][-持续]

[-durative]：[-持续]

[-dynamic]：[-动态]

[ET∩RT]@coda：表示事件时间和参照时间相交于终结阶段

[ET∩RT] @nucleus：表示事件时间和参照时间相交于核心阶段

[nucleus<coda]ET：表示事件时间(ET)是由核心和终结两部分组成的，并且核心位于终结之前。

[-pastel]：[-过去终结]

[-telic]：[-终结]

“———”：情状的原因阶段

“＋＋＋”：情状的结果阶段

1-state：内含一个阶段的情状。这个内含的阶段可以是情状的原因阶段“———”，也可是情状的结果阶段“＋＋＋”。

2-state：内含两个阶段的情状，包括情状的原因阶段“———”和情状的结果阶段“＋＋＋”。

<CULM>：终结点情状在语料库中的标记

<PFCT>：完成体在语料库中的标记

Active V+ZHE：活动动词+着

A_A：accumulated-adverb 累积副词

A_D：adjective-directive adverb 形容词＋趋向副词

adj.：adjective 形容词

adv.：adverb 副词

Au：auxiliary 助动词

AUC：area under the receiver operating characteristic curve 曲线下面积，经常用于统计 ROC 曲线的面积；

basp：basic aspect 基本体

C：参数

C4.5：决策树算法中最流行的一种分类树，其核心算法是 ID3 算法。

C< RT：说话时间前于参照时间

CFG：Context Free Grammar 上下