《Introduction to Chinese Natural Language Processing（中文自然语言处理导论）》书评 (Review of *Introduction to Chinese Natural Language Processing*)

姜松
(Jiang, Song)
美国夏威夷大学
(University of Hawai'i at Mānoa, United States)
sjiang@hawaii.edu

Kam-Fai Wong, Wenjie Li, Ruifeng Xu, Zheng-sheng Zhang (2010). Introduction to Chinese Natural Language Processing. In Hirst Graeme (Series editor), Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. x+148 pp. Paperback ISBN: 978-1-59829-932-8. e-book ISBN: 978-1-59829-933-5.

《Introduction to Chinese Natural Language Processing（中文自然语言处理导论）》是由 Morgan & Claypool Publishers 出版的“人类语言技术综合讲座”（Synthesis Lectures on Human Language Technologies）系列中的一部。这套英文系列丛书立足于自然语言处理、计算语言学、信息检索、自然语言人机接口等与人类语言技术相关的学科，以综合讲座的形式介绍各个学科领域的发展概貌、着重突出各学科最新出现的重要的技术和方法及其在相关研究中的实际应用。人类自然语言计算机处理作为一个学科已有几十年的发展历史。随着计算机和互联网技术的成熟和普及，人类语言活动中各种语言材料的搜集、加工、存储和提取正变得更加迅捷与便利，进而带动了自然语言计算机处理学科的飞跃性发展。但从自然语言处理的整个学科的发展来看，这一领域的主要研究成果大都是在以英语为主的西方语言的基础上取得的。相比之下，中文的自然语言处理研究起步较晚，基础薄弱，且由于汉语与英语在类型学上的显著差异，面临着许多特殊的挑战。随着汉语地位在全球范围内的提升，特别是汉语在国际互联网和商业软件开发上使用比例的飞速增加，商业信息提取、企业与客户关系管理、机器翻译、自动文摘、汉语语音识别、语言学研究、辅助汉语教学系统的开发等领域越来越需要从汉语出发的自然语言处理技术的支持与跟进。在这样的背景下，《中文自然语言处理导论》一书的出版正是适应了日益增长的对汉语语言处理技术的需求，从汉语类型学的角度深化和扩展了普通自然语言处理的研究，为丰富和完善自然语言处理学科的理论与实践做出了一份重要贡献。

全书主体内容共分八章，另外包括一个列举与中文自然语言处理相关的语言学资源的附录、一个参考文献以及一份作者简介。全书八章主体内容大致可分为三个部分：基本概念（第一、第二章）、自动识别（第四、第五章）和汉语语词的语义特征（第六、第七、第八章）。

作为全书的导引，第一章首先在人类自然语言、普通语言学、计算语言学的大背景下，把自然语言处理定位为一门将语言理论转化为实际应用的技术，并将这一技术在语言处理过程中的实施平台定位在词法、句法和语义三个层面上。作者以汉语形态分析为基础，通过一系列具体的实例，列举出汉语的形态特征给汉语语言处理造成的困难和挑战：如汉语词的切分、词类标注、句法与语义歧义等。基于汉英两种语言在词语层面上的形态区别，以及由此决定的对两种语言处理的不同的技术要求，作者确立了汉语的形态分析在本书中的核心主导地位。

第二章从普通语言学的角度，介绍汉语的字、语素和词的概念，汉语词的形成过程以及汉语词的基本特征。作者首先提供了汉语的字、语素和词三个核心形态单位的语言学定义，并详尽论述了三者之间的区别与联系。作者指出字为汉语的书写单位，呈线性等距排列，不具备代表独立语素的功能。语素为最小的语义单位，通常为单音节，以单一汉字的形式出现，可经过语素组合形成词。词是介于语素与词组之间的语言单位，受到分布调控与词汇整体性的制约。在本章接下来的篇幅中，作者详细描写了汉语合成词的构词方式。作者首先根据词的音节数量，按双音节、三音节、四音节三类分别描写，然后讨论带有词缀（包括前缀、后缀以及动词后缀）和由重叠构成的合成词的构词特点，最后讨论了离合词的特征。通过以上详尽的分析，作者指出，汉语复杂多样的构词方式是造成分词困难的重要原因，是汉语自然语言处理中无法回避的关键所在。这一章为全书提供了一个必要的语言学基础，明确了汉语语言处理所面对的关键问题。

作为第二章的深化和延续，第三章在自然语言处理框架下具体勾勒出直接造成技术处理困难的汉语语词的语言学个性与文本特征。作者将这一章所涉及到的影响汉语机器分析的汉语特征归纳为汉字、文本和语言学特征三类。与汉字有关的影响技术处理的特征包括：汉字总量的不确定性、繁简转化、异体字、方言用字、汉字编码的多样化。与文本有关的特征包括：排印格式和标点符号。与语言学特征相关的包括：缺乏语法和词性的形态标记、同音异义词与同形异义词、歧义以及以缩略语、专用人名地名、音译词、地域变体以及风格变体为代表的未登录词。第二和第三章对困扰机器处理的汉语语言与文本特征的分析为以后各章有关技术性处理对策的讨论作好了准备。

第四章集中探讨汉语的分词问题。分词被认为是汉语语言处理的第一步，分词的精确度直接关系到分析结果的优劣。作者首先通过对比英汉两种语言在分词上的差异，指出汉语句子无标记的线性字符排列决定了汉语分词必须解决确认字词顺序、加注词的分界的问题，并以此为出发点，为汉语分词归纳出一个技术性的定义。作者随后利用具体实例，说明汉语分词过程中存在的两大挑战：对歧义和未登录词的处理。针对这两大挑战，作者将现有的汉语分词方法归纳为两大类：以字为

基础的方法和以词为基础的方法，详细比较、讨论了这两类方法在计算上的差别以及各自的优缺点，并着重介绍了能有效解决分词中歧义问题的两种计算方法：词典法与统计法。此外，作者还在这一章中，介绍了现有的三个大型语料库所采用的汉语分词的评估标准。这三个语料库分别为：北京大学中文系现代汉语语料库、台湾中研院现代汉语平衡语料库和宾夕法南亚大学汉语树形结构语料库。最后，作者对一些免费的汉语分词工具进行了简要的介绍。

第五章讨论的是未登录词的识别问题。作者认为在汉语分词过程中出现的未登录词的识别问题主要是由于所需处理的未登录词尚未加入到作为分词依据的分词词典中，导致分词程序无法找到针对未登录词的分词依据，因而无法作出有效的分词判断。汉语新词的不断出现也是造成未登录词问题的一方面的原因。汉语新词的产生主要是以语素合成和词缀附加的方式实现的。未登录词主要是指那些指代人名、地名以及机构等的专用名词、特定范畴内的技术性名词以及缩略语等。作者在这一章中讨论了识别各类未登录词的计算方法，并着重介绍了对人名、组织机构名称和地名的识别方法。识别未登录词的形式依据主要包括：常用名称的内部结构、名称中的常用字以及文本信息等。

在接下来的第六至第八章，作者将讨论的重心从汉语语词的结构层面转移到语义层面。对于缺乏形态标记的汉语来说，比起形式结构，语义在语言理解过程中所起的作用似乎更为关键。正因为如此，不少语言学家将汉语归为语义型语言。不考虑语义的因素，许多汉语机器处理的任务都不可能获得满意的结果。第六章开篇介绍了与词汇语义学相关的基本概念，如：义元、多义、同义、反义、上位、下位、整体、部分、专指等，并对接下几章将涉及到的语义框架、连用语、动词配价（施事、受事、工具等）进行了说明。这一章的主要篇幅集中对三种有代表性的汉语分类辞典进行介绍和评价。这三种辞典是：《同义词词林》、知网（HowNet）和《中英文概念词典》。《同义词词林》是中国第一部纸质现代汉语义类词典，它以三层等级树形结构模式勾画出所收常用词的语义关系，由多所大学输入汉语电脑词库后，被广泛应用在写作、翻译和自然语言处理上。知网是一个以汉语和英语的词语所代表的概念为描述对象，揭示概念与概念之间以及概念所具有的属性之间的关系为内容的常识知识库。作者通过与英文 WordNet 的对比，总结出知网的独到之处：以义元为分析单位、以图形结构描写语义关系、以英汉词汇概念的对比为分析基础。《中英文概念词典》是由北京大学计算语言所开发的英汉双语词汇概念知识库。它采用 WordNet 的结构布局，在保证与 WordNet 兼容的同时，对算法和功能进行了改进。这些改进包括更细化的名词分类、更精密的关系描写、真实语料库的支持以及具有统计学意义的量化处理等。

第七章主要论述了与汉语连用语（collocation）相关的基本概念，包括定义、特征、类别以及语料来源。针对自然语言研究领域对于连用语同现搭配这一概念的不同理解和争议，作者首先列举了一系列以英语为基础的连用语的不同定义，指出这些不同的定义源于定义者对连用语不同特点的关注。通过比对，作者进一步总结出汉语连用语与英语连用语在宏观层面上的显著不同：即汉字串的连续性、汉语字

词使用的灵活性和以实词为主要对象的连用语提取特点，并据此给出了一个针对汉语的，较为严密的定义：连用语是一个由两个或两个以上词语组合而成的，具有句法和语义关联，能够重复出现并符合使用习惯的表达结构（p. 98）。以这一定义为出发点，在接下来的篇幅中，作者从定性、定量、类别 (成语、结合度等) 以及语料来源等方面详细论述了汉语连用语的具体特性，为下一章讨论汉语连用语的自动提取方法和运算规则奠定了基础。

第八章重点介绍汉语连用语的自动提取方法。作者首先在本章导言部分指出汉语连用语的自动提取过程实际上是连用语提取技术在汉语自动分词和词类标记上的具体应用。根据主要的区别特征和目标搜寻策略的不同，作者将当今通行的自动提取技术划分为以下三种：统计提取、句法提取和语义提取。统计提取以目标关键词为切入点，将在关键词周边限定范围内出现的字词列为候选连用语，进而依据关键词与其周边字词组合的统计显著性确定连用语。句法提取利用关键词与连用语的组合必须合乎句法规范这一要求，根据预先设定的句法规则，通过句法剖析程序（parser）对目标关键词与同现词的组合进行过滤，然后再依照统计显著性确定连用语。语义提取利用语义限定测试来确定连用语。语义限定测试主要包括同义词测试和翻译测试两种。同义词测试利用同义词替换的有限性，以排除的方式提取。如果在语料库中某一关键词的同义组合出现频率超过一定的限度，其连用语的合法性将会被质疑并可据此予以排除。翻译测试利用候选连用语的非组合性特点进行提取。如果一个词语的组合不能逐字翻译成另一种语言，与其搭配的字词则被认为不具有组合性，因而可据此确立该组合的连用词地位。针对汉语连用语的语法特性，作者又特别介绍了一种综合性的提取方法：类别提取。这一方法主要是针对不同类型的连用语而设计的。它综合了以上三种基本的提取方式，包含六种不同的运算程序。提取过程中，首先由程序确定被检测的搭配的类型，而后根据检测结果，选取匹配的运算方式进行连用语提取。作者指出，跟前面三种单一的基本提取方法相比，类别提取更适合汉语的特点，更能获得理想的提取结果。

作为一部汉语自然语言处理导论性质的专著，本书从汉语的实际出发，选取汉语的形态分析作为全书的切入点，针对汉语自然语言处理中存在的难题，按照字词切分、词类标记、未登录词识别、语义分析和连用语自动提取编排章节，展开论述，对现有的处理方案进行了详尽的归纳、总结和评估，既反映了当前汉语处理技术的概貌，又突出了学科的重点成就。

本书结构完整、条理清晰、论述缜密、循序渐进。特别是作者利用汉语描写词汇学和语义学的研究成果，通过与英语的对比，突出汉语中制约技术处理的语言特征，为讨论技术性处理方案作出铺垫的行文方式，将汉语语言学与汉语计算处理自然地联系起来，增加了本书的目的性和可读性。

作为英文系列丛书中的一部，本书的出版有着十分特殊的意义。它将汉语自然语言研究的成果以专著的形式介绍给英语世界，为整个人类语言技术领域提供了一个汉语案例，也为英语世界的同行了解汉语自然语言处理的现状和成果打开了一个窗口、为进一步的技术交流奠定了基础。

本书设定的阅读对象为已经具备初级语言处理知识的读者，因此对从事自然语言处理、计算语言学、信息检索、机器翻译、网络文字处理等工作的专业人员有着重要的参考价值。不仅如此，作为一部论述清晰、内容翔实的学科导论，本书对于自然语言处理、计算机与应用、汉语应用语言学、作为第二语言的汉语教学等专业的学生来说也可以作为理想的入门教科书来使用。此外，本书对有志于或正在从事计算机辅助汉语教学研究的汉语教育工作者也具有一定的指导意义。计算机辅助教材编写、现行课本的分析研究、汉语水平等级大纲的制定、汉语教学词频统计、汉语中介语语料库的建立等，都离不开汉语自然语言处理技术的支持。对于不具备专业的语言处理知识的大多数汉语教育工作者来说，本书不仅能为他们提供一个了解相关技术知识的途径，而且能够使他们从计算分析的角度获得对汉语特征的新认识，提高教学应用程序的开发与评估能力，因而值得在此推介。然而，对于对外汉语教育者来说，如果本书能在介绍和评价一些汉语处理工具的同时，适当增加有关具体操作的说明，将更能满足他们的需求。另外，本书个别地方在排印和文字拼写上还存在一些疏漏，有待再版时校正。