目 录

第一章　绪论

近年来，随着语料库研制方法与加工技术逐渐成熟，双语平行语料库的规模从过去几十万、百万词级别提升至千万甚至上亿词级别。语料的规模化效应使得语料库初步具备了大数据的典型特征，不仅对计算机加工和检索技术提出了新的要求，也为其在语言教学、语言研究和自然语言处理等领域的应用带来了挑战。从数据到大数据，两者存在本质上的差别，数据量的急速增长对语料库的数据采集、加工和检索及面向语言应用的研究和数据分析方式产生巨大影响，仅仅使用“手工方式”已无法满足处理海量数据的需求。本书旨在报告大数据背景下大规模平行语料库的加工、检索和应用问题研究方面取得的进展。

本章首先阐述平行语料库研究的背景与意义，其次介绍“大规模英汉平行语料库检索平台”的研制，最后介绍本书的主要研究内容与基本结构。

1.1 平行语料库研究的背景

1.1.1 国内外平行语料库研制的成果

平行语料库（parallel corpora）是由源语文本及其具有翻译关系的平行对应目标语文本构成的双语语料库（McEnery *et al*. 2005：40）。翻译对应的层次可分为词汇、语块、句子、段落四种级别。

自20世纪90年代初世界首个平行语料库加拿大议会会议录英-法平行语料库（the Canadian Hansard Corpus）在加拿大诞生以来，平行语料库的研制逐渐掀起热潮。其中，欧洲的平行语料库发展较为成熟，建成的平行语料库规模最大，比较有代表性的包括欧洲委员会联合研究中心研制的10亿词的JRC-Acquis多语种平行语料库和英国爱丁堡大学研制的5千万词的欧洲议会平行语料库（European Parliament Proceedings Parallel Corpus）。上述语料库虽然容量较大，但主要面向欧盟官方语言文字应用和自然语言处理，且语体分布并不均衡，多为官方文本，语料库的加工也仅实现双语对齐。英语-挪威语平行语料库（The English-Norwegian Parallel Corpus，ENPC）是较早建成的采样均衡的平行语料库，于1997年由挪威奥斯陆大学建成，库容约为260万词。该库已完成文本头标记和句子级别的对齐，且面向翻译研究的各类应用，建库理念较为成熟。按照ENPC的采样标准，英语-瑞典语双向平行语料库（The English-Swedish Parallel Corpus，ESPC）、英语-意大利语双向平行语料库（Corpus of English X Italian，CEXI）和英语-葡萄牙语双向平行语料库（English-Portuguese Parallel Corpus，EPPC）等多语对的平行语料库作为奥斯陆多语种语料库（Oslo Multilingual Corpus，OMC）项目的姊妹项目也在同期框架下相继展开了研制，其中，ESPC更是达到了2,800万词，CEXI也达到460万词（王克非、黄立波 2012）。

除此之外，一些双语平行语料库还进行了短语结构句法标注，比较有代表性的包括：宾州大学英-汉翻译树库（English Chinese Translation Treebank，ECTT）（Bies *et al*. 2007）、宾州大学阿拉伯语-英语平行对应树库（Arabic-English Parallel Aligned Treebank，AEPAT）（Grimes *et al*. 2010）、宾州大学汉-英平行对应树库（Chinese-English Parallel Aligned Treebank，CEPAT）（Li *et al*. 2010）和图尔库依存树库（Turku Dependency Treebank，TDT）（Haverinen *et al*. 2013）。还有一些建立在依存句法体系上的对应树库，例如布拉格依存树库（Prague Dependency Treebank，PDT）（Čmejrek *et al*. 2003）。这些树库建立之初的目的是作为机器翻译的训练语料库，并没有为开展语言与翻译研究作任何设计。

自21世纪初以来，国内业已兴建了大批多用途双语/多语平行语料库。

比较有代表性的如面向自然语言处理的北京大学计算语言学研究所的汉英平行语料库和哈尔滨工业大学信息检索研究室的汉英双语语料库。还有专门用途语料库，如针对语言研究与翻译领域的燕山大学的《红楼梦》中英文平行语料库（刘泽权等 2008）、上海交通大学的莎士比亚戏剧英汉平行语料库（胡开宝、邹颂兵 2009）、香港理工大学的新型双语旅游语料库（李德超、王克非 2010）和西安外国语大学的中国现当代小说汉英平行语料库（黄立波 2013）；又如针对口译领域的上海交通大学的汉英会议口译语料库（胡开宝、陶庆 2010）和北京外国语大学的中国口译学习者语料库（张威 2015）。以上这些语料库均实现了句级对齐，有的还进行了词法和浅层句法的标注，口译语料库还手工标注了副语言等特征，使得语料库的应用价值大大增加。此外，北京外国语大学中国外语与教育研究中心兴建了包含3千万词的通用汉英对应语料库（黄立波、朱志瑜 2013：46），该库采样均衡，包括翻译文本库、百科语料库、专科语料库和对译语句库四部分，并实现了句级对齐和词性标注，是目前国内通用性最强、规模最大、加工深度最高的联机平行语料库。

自20世纪90年代以来，平行语料库的发展呈现出爆发式的增长趋势，门类上呈现出专门、通用库共存，口、笔译库同步，面向自然语言和语言研究共生的趋势。继单语语料库之后，双语平行语料库已逐渐成为面向语言研究、译学研究、翻译教学、词典编纂和自然语言处理等的基础设施。

1.1.2 平行语料库研制的方法

在平行语料库的研制中，“搜集双语对应语料本身不是目的，更重要的是后续的对齐加工以及利用”（王克非、熊文新 2011：31）。语料的标记与标注的设计与方案决定了语料库的基本构架和语料形态，直接影响研究者对语料库的使用（李文中 2012：336），平行语料库加工的程度甚至影响课题的选择（王克非、黄立波 2008：10）。检索是从语料库中批量提取、观察、统计和分析数据，是研究语言现象的重要途径（梁茂成等 2010：57）。快速、有效和精准地实施检索，准确查找出符合研究目的所需的语言信息，是保证语料库研究充分性和可靠性的关键（陈功 2011：10，13）。相较于单语库，双语平行语料库的加工要求实现翻译单位的对齐，而检索

要求实现双语语对的提取，因而平行语料库加工和检索的难度远高于单语语料库（梁茂成、许家金 2012：37）。

语料库加工就是对语料库中的各类语言学单位附加语言学与语境信息的过程（McEnery *et al*. 2005：25），包括语言学标注和元信息标记两个基本层面。加工既是保证语料库可重用性的关键，也是语料库重要价值的体现（Leech 1997：2；McEnery 2003：454-455）。

利用自动语言分析技术可有效提高语料库标注的效率，避免人工处理中由于疏忽而造成的各类问题。在计算语言学中，按照处理深度的不同，语言分析技术可划分为浅层分析和深层分析两类（刘挺、马金山 2009：100）。浅层分析技术主要用于处理语言基本单位，例如中文分词、词性标注、词形还原和命名实体识别等。深层分析技术是对语句和语篇等较大语言单位的语法、语义甚至语用信息的处理，例如双语对齐、句法分析、情感分析和语义理解等。

当前，语言的浅层分析技术已经成熟，被广泛应用在语料库基本加工处理中并取得显著效果，例如，面向中文分词和词性标注的ICTCLAS[1]、CKIP[2]，面向英文词性标注的CLAWS[3]系统，面向多语言的词性标注系统Stanford Parser[4]、LTP[5]，这些系统融合使用了基于规则和基于机器学习模型的处理方法，机器学习模型包括隐马尔可夫模型（Hidden Markov Model，HMM）、条件随机场模型（Conditional Random Field，CRF）和最大熵模型（Maximum Entropy Model，MEM）等，在部分应用场景中，针对英语的词性标注的错误率约为1.5%，另有3.3%的歧义无法正确识别（Leech *et al*. 1994：625），而针对汉语的词性标注精确度更是高达98.38%（刘群等 2004：1427）。

相较于浅层分析技术，深层分析技术挖掘语言中更为抽象和隐含的信息，应用深层分析技术对语料库进行加工可为语料库进一步“增值”。在

1　参见 http://ictclas.nlpir.org/ 检索日期：2023年8月20日

2　参见 http://ckipsvr.iis.sinica.edu.tw/ 检索日期：2023年8月20日

3　参见 http://ucrel.lancs.ac.uk/claws/ 检索日期：2023年8月20日

4　参见 http://nlp.stanford.edu/software/lex-parser.shtml 检索日期：2023年8月20日

5　参见 http://www.ltp-cloud.com/ 检索日期：2023年8月20日

双语对齐方面，自1995年以来，国内外学界对双语句对齐的研究也取得了重大进展，采用的主要方法包括：基于句子长度的方法、基于词典的方法、基于句子长度和词典的混合方法（黄俊红等 2007）。这些方法可实际应用在多个语种的平行语料库研制上，例如，梁茂成、许家金（2012）提出了同时保留元信息、段落信息的基于文本长度与基于词典相融合的方法，开发了图形化的英汉双语文本自动句级对齐模块（CETA），自动对齐准确率高，在对220万字的非文学类文本的测试中，句子对齐的准确率达到了95%以上（梁茂成、许家金 2012：42）。句法分析一直以来都是语言深层分析技术关注的核心问题之一，然而，受限于语言形式、语义的多样性，这类技术距离实际应用仍然存在较大的差距（刘挺、马金山 2009：100）。因而，目前在平行语料库的深度加工研究领域，特别是针对句法结构的复杂标注还必须依靠手工进行（王克非等 2004：22）。但手工句法标注费时费力，为保证标注质量，对参加标注的人员素质要求非常严格，因而手工标注在面向大规模平行语料库的建设中缺乏实际可操作性。

自2010年以来，自然语言处理技术迅速发展，现有的句法分析技术也取得进步。目前，用于自动句法分析的语法体系包括：短语结构语法（phrase structure grammar，PSG）、依存语法（dependency grammar，DG）和组合范畴语法（combinatory categorial grammar，CCG）。其中，短语结构语法和依存语法得到广泛应用，关于二者的自动分析技术也层出不穷。就短语结构语法的精度而言，以宾州英语树库（Penn Treebank，PTB）作为训练集，使用词汇化概率上下文无关文法（Lexicalized Probabilistic Context-Free Grammar，Lexicalized PCFG）模型对英语文本的短语结构进行句法分析，精度可达到90%（Bikel & Chiang 2000；Levy & Manning 2003），而以宾州汉语树库（Chinese Treebank，CTB）作为训练集，使用最大似然性估计概率上下文无关文法（Maximum Likelihood Estimation PCFG，MLE PCFG）模型对汉语文本进行句法分析，精度达到81.1%（Levy & Manning 2003）。依存语法利用神经网络进行自主学习，对宾州树库英语本族语语料的分析精度高达90.7%，对汉语本族语语料的分析精度也达到82.4%。

当前，基于统计的自然语言处理技术和基于人工神经网络的深度学

习是自动语言分析技术的主流。现有技术在语言的浅层、深层加工中的精度和适用性已得到极大提升，大规模平行语料库标注的人工成本可极大缩减。如何结合平行语料库深度加工的特点应用上述技术，设计开发面向语料库加工的自动方法和技术，是大数据背景下语料库研制亟待解决的问题。

1.1.3 平行语料库的应用

平行语料库是关于翻译的语料库，其应用离不开翻译与语言运用这个核心话题。

首先，将描述翻译理论与平行语料库结合起来，是平行语料库的主要研究途径。1993年，Baker首次将语料库引入翻译研究标志着基于语料库的翻译研究（corpus-based translation studies，CTS）这一研究范式的确立（黄立波、王克非 2011）。当前，基于语料库的翻译研究大多以平行语料库和翻译类比语料库作为数据基础，以翻译语言特征、译者风格、翻译规范、翻译教学和口译等作为研究对象和内容，以单语语内类比和双语语际对比作为观察模式，对类符-形符比、词汇密度、词类分布、平均句长、句对应类型、叙事结构等宏观的语言特征展开考察（Baker 1995，2000）。此外，对诸如选择性“that”、“把”字句、“被”字句、连接成分、特定词类等微观语言特征展开考察也是基于平行语料库的研究的重要趋势（Olohan & Baker 2000；柯飞 2003；秦洪武、王克非等 2004；黄立波 2007；胡显耀、曾佳 2010）。上述宏观和微观两个维度的研究探索了翻译的共性与本质，为语料库翻译学这一研究范式奠定了基础（王克非 2006；黄立波、王克非 2011；胡开宝、毛鹏飞 2012；黄立波、朱志瑜 2012）。值得注意的是，已有研究将平行语料库应用于翻译词典的编纂（刘泽权、张丹丹 2012），这类建立在真实的大规模平行语料库上的翻译词典更加可靠，适合译者和学习者在实践中应用。

其次，随着基于统计的自然语言处理技术和基于人工神经网络的深度学习技术的发展，平行语料库是提升机器翻译系统精确度和可用性的重要途径。早在20世纪80年代末，美国的IBM公司就提出了基于噪声信道模型的统计机器翻译系统（Brown *et al.* 1990），通过从平行语料库中提取词

汇对译信息来获取翻译中的规则，并根据这些规则构造自动翻译系统，随后发展出基于短语的最大熵模型（Och & Ney 2002）、基于形式化句法模型和经过句法标注的平行语料库的统计翻译模型（Wu 1997）。Hinton *et al*.（2006）发表的题为“A fast learning algorithm for deep belief nets”的文章打开了人工神经网络之门，随后美国的BBN公司（Devlin *et al*. 2014）和Google公司（Sutskever *et al*. 2014）相继提出了基于人工神经网络和平行语料库的机器翻译模型。

由此可见，平行语料库的研制，尤其是大规模平行语料库的研制对于语言研究和自然语言处理研究至关重要，甚至成为重要研究的关键环节，具有广阔的应用前景。

大规模平行语料库的研制与应用已成为当前亟待探索的课题。大规模平行语料库之所以称为“大数据”，不仅是双语数据规模的拓展，也是加工深度和信息抽取深度的拓展。本书旨在为大数据背景下平行语料库的加工、检索和应用提供先导性研究，以期为未来大规模、深加工和多用途的双语或者以汉语为中心的多语平行语料库的研制与应用提供借鉴，本书在理论与应用层面都具有重要意义。

本书在理论层面：(1）将系统考察现有自然语言处理方法在大规模英汉平行语料库开发中的可靠性和适用性，为今后其他语言对的平行语料库研制的理论、方法、技术和实践积累先导性经验；(2）将介绍北京外国语大学自2016年以来开展的关于自动词法、句法标注和语料库检索的方法、技术和软件平台的研究，研究成果将提高平行语料库的加工效率，进一步降低人工成本；(3）将报告应用大规模平行语料库开展翻译研究的部分成果，这些成果将为语料库翻译研究提供大数据背景下的新的观察视角，为基于大数据的语料库翻译研究进一步积累经验。

本书在应用层面：(1）由作者设计和开发的软件、工具可为其他语料库的研制提供借鉴；(2）对自动标注和检索方法信度的研究将为双语语料库加工和提升自然语言处理的精度提供借鉴。

1.2 大规模英汉平行语料库的设计

“大规模英汉平行语料库”的研制框架主要包括语料库的设计、加工与检索三个部分，下文分别对其进行介绍。

1.2.1 语料收集

“大规模英汉平行语料库”的总体设计规模为1亿字/词以上，包括非文学、文学和综合类三个组成部分。其中，非文学部分包括科技、科普、社科、财经、口译、时政新闻和法律7个子库，文学部分包含文学库和历时复合库。该语料库具备如下性质：从结构分布上，可分为平行库和双语库；从用途上，可分为通用库和专门库；从时间上，可分为共时库和历时库；从语体上，可分为笔译和口译。该库的语料以书面语为主、口语为辅，体裁上分为文学、新闻、政论、科技和应用文5个大类。语料包括英译汉和汉译英两个翻译方向，前者约占总库容的2/3，后者约占1/3，具体的语料规模和各部分库容如表1.1所示。

表1.1 “大规模英汉平行语料库”的架构与规模

文类	子库名称	字/词数	合计
非文学	科技库	8,586,410	67,943,621
	科普库	8,299,427	
	社科库	9,981,951	
	财经库	21,762,931	
	口译库	956,614	
	时政新闻库	8,076,805	
	法律库	10,279,483	
文学	文学库	11,395,489	22,926,985
	历时复合库	11,531,496	
综合类	综合库	12,895,686	12,895,686
总计		103,766,292	

由表1.1可见，除口译库外，“大规模英汉平行语料库”的各子库库容从8百多万到2千多万不等，大部分语料库的规模在1千万上下，库容相对接近，因此分布具备均衡性，各子库之间具备可比性。非文学和文学的库容比约为3:1。语料库中的文档将按照1.2.2小节中表1.2的分类框架进行元信息标记。

1.2.2 语料的加工与检索

“大规模英汉平行语料库”的加工包括元信息的设计与标记、双语句级对齐、词法与句法标注三个部分。其中，元信息的设计依照语料的来源、年代、语域、体裁等分类，有关体裁的分类见表1.2。此外，还包括语料库中英文名称、样本的发生年代、语体、翻译方向、原始数据篇名、作者、译者等文本的基本信息。平行语料库中的英汉对译文均在句子级别进行对齐，并针对英语进行词法标注，包括英文分词、词形还原和词性赋码，针对汉语则进行中文分词和词性赋码，方便用户从语料库中抽取词汇和语法信息。鉴于句法标注比较困难，该语料库将尝试进行部分文本的句法分析和句法标注。

表1.2 “大规模英汉平行语料库”的分类框架

文类	体裁
非文学	新闻报道、社论、评论、政府文献、法律文献、技艺、贸易、娱乐、财经、哲学、社会科学、科技文本、科普文本、杂类（书信、宗教、风俗、旅游等）
文学	纯文学、传记、散文、普通小说、神怪小说、侦探小说、科幻小说、武侠小说、戏剧、传奇、幽默、儿童文学

“大规模英汉平行语料库”的研制还将设计、开发配套的检索软件系统，并实现以下三个目的：(1) 实施元信息检索，便于用户根据自己的设定从语料库中提取具有特殊语境的文本；(2) 实现双语双向检索、简单检索、复杂检索、词形还原和有条件检索；(3) 实施较精确的词汇、语块检索，并呈现搭配信息。

1.3 主要内容与基本结构

笔者将在本书之后的章节中对面向1亿词级别的“大规模英汉平行语料库”的加工、检索和应用分别展开讨论。

第一，本书介绍面向双语规模化加工的词法标注和双语对齐的基本方

法、标注规范、采用的标注集、标注流程和标注工具的设计，并根据语料标注的实践，对比本书设计的标注软件和其他标注工具，并分析其在实际标注中的信度。第二，本书讨论了面向英汉平行语料库句法标注的句法理论、句法模型、标注集、存储模型、数据可视化和分析方法与技术，并对句法标注的信度展开分析。第三，本书介绍了面向1亿词级别英汉平行语料库检索的“大规模英汉平行语料库检索平台 1.1”的研制，帮助研究人员操作大规模平行语料库，并开展面向研究与教学的语料检索分析工作。平台不仅具备单语检索、双语检索、词汇搭配分析和历时检索等模块，还提供元信息过滤、词形还原、模糊检索和复杂模式检索等功能，可快速准确地从语料库中抽取信息。第四，本书介绍了基于大规模英汉平行语料库的应用研究的案例，考察科技文体中英语被动结构的语言特征及其在汉语中的对译形式，试图从结构形式、句法、语义和文体四个维度揭示被动结构由英语向汉语的翻译转换过程中呈现的特征和趋势。

本书的第一章回顾了平行语料库研究的背景和大规模英汉平行语料库的设计。第二章探讨面向大规模平行语料库研制的加工、检索技术。第三章通过知识图谱和可视化方法梳理国际平行语料库研究的现状与发展趋势。第四章综述句法标注语料库的研制与应用。第五章探讨句法标注在英汉语言研究中的信度分析。第六章报告大数据背景下大规模英汉平行语料库的加工方法与笔者开发的工具。第七章介绍大数据背景下大规模英汉平行语料库的检索方法。第八章通过研究案例报告科技文体中英语被动结构的语言特征及其汉语对译形式。第九章对影响英语被动结构汉译形式的显著因素展开多变量分析。第十章以文体作为考察变量，对英语被动结构的汉译方式展开多文体对比研究。第十一章以《国富论》不同时期的汉译本为例，探讨翻译汉语名词短语复杂性特征的历时变化。