前言

本书是外研社“外语学科核心话题前沿研究文库·翻译学核心话题系列丛书·语料库翻译研究系列”中的一部，聚焦平行语料库的研制和应用，所选择话题涵盖平行语料库制作及双语语料在研究中的应用，涉及平行语料库设计、平行语料处理、语料的跨平台应用、语料检索、检索数据统计与分析，以及语料库支持的翻译研究和对比语言研究，等等。

按照一般分类（如Barrière 2016：105），双语语料库（bilingual corpora）包括平行语料库（parallel corpus）和类比语料库（comparable corpus）。本书探讨的重点则是平行语料库，但在语料库应用研究部分使用与平行语料库直接关联的类比语料库。

近二十年来，双语语料库尤其是平行语料库的创建和应用在中国蓬勃发展，基于语料库的翻译研究方兴未艾。同时也要看到，由于缺少广泛认可的规范，平行语料库建库方式多样，建库质量参差不齐，使用效果不是特别令人满意。库容大、代表性强、检索方便且内容丰富的平行语料库一向是语料库翻译学孜孜以求的目标，平行语料库的研制自然成为语料库翻译学的核心问题。本书的编写既是回应翻译和对比语言研究对平行语料库的现实需求，也是尝试通过普及基础知识突破语料库翻译学发展中的瓶颈问题：平行语料的制作和应用。为实

现这一目标，本书聚焦平行语料库的创建和应用，设置以下核心议题：在平行语料库的创建方面，核心话题为语料选择、处理、元数据使用和平行语料对齐、多语语料检索；在平行语料库应用方面，核心话题为平行语料库在翻译研究中的应用、基于平行语料的对比语言研究、平行语料在翻译教学中的应用。

本书力求简明、直接地向学习者或研究者介绍相对成熟的语料库建库理念、制作过程和应用工具，帮助他们初步了解经验性语言数据的分析和统计方法。本书辟有专章探讨语料库在翻译研究和对比语言研究中的实际应用，这对于研究者有效运用平行语料库从事翻译、教学和研究有启发意义，期待本书能为外国语言文学领域研究生和研究者从事外语研究提供有益的参考。

全书分为九章。第一章综述语料库和双语语料库的性质、特征，探讨双语语料库的主要用途(翻译研究、语言研究、翻译教学和翻译实践)，描述平行语料库研究的进展，提出有待解决和完善的问题。第二章探讨平行语料库的主要类型和结构，涉及平行语料库的发展状况、代表性平行语料库的构成特征以及抽样结构和对齐方式。第三章探讨研制平行语料库的主要方法和工具，涉及语料采集、语料处理、篇头元数据设置、平行语料的对齐、标注以及平行语料文件创建六个关键步骤。第四章聚焦平行语料检索，探讨多条件检索的实现方式，涉及普通检索、通配符检索和正则表达式检索。第五章讨论检索数据的统计分析。第六章谈平行语料的跨平台应用，重点探讨使用可扩展标记语言(XML/TMX)文本便于数据交换的特征，展示这类文本的跨平台应用问题，扩展平行语料服务于教学科研的范围。第七章介绍专用翻译语料库(也属双语语料库)的设计和创建，涉及学习者语料库、多模态翻译语料库和历时复合语料库的构成和创建方法。第八章通过案例分析展示双语语料数据在翻译研究、对比语言研究和翻译教学中的应用。第九章为结语。

出于篇幅考虑，在撰写过程中我们尽量考虑当下读者对相关软件

操作和使用的熟悉程度，对读者比较熟悉的应用工具只作粗略介绍，不再详细展示操作步骤。这样做还有一个原因：学习者可通过软件自带的帮助文件（Help）自行学习和了解程序运用的细节；鉴于大多数软件都有类似的帮助文件或者带有操作指南链接，读者可以通过指南了解软件使用信息。另外，本书介绍的应用软件多为免费软件，在介绍其用途和使用方式时尽量提供下载网址，读者可通过相关的链接自行下载。

本书是我在平行语料库研究领域近二十年的学习心得和经验总结。新千年之初，蒙王克非师知遇，我幸得机会在北外潜心深造；更为幸运的是，我在王克非师鼓励、导引下迈入语料库翻译学这片新奇、广阔的天地。有他导引、鼓励和提携，想不作为都难。平行语料库制作和研究从来都需要协同和合作，我要感谢多年来在语料库翻译学领域一同努力的同事和同行，她们先后加入团队，有夏云教授(参与第一章撰写)、王青副教授、黄万丽老师、徐欣副教授、孔蕾教授。我也很幸运，我的努力耕作从来都不是踽踽独行，始终有众多师友鼓励、督促、分享、指点、解疑、答惑，他们是胡开宝教授、张威教授、黄立波教授、朱一凡教授、胡显耀教授、刘泽权教授、刘鼎甲博士、赵秋荣博士、庞双子博士，等等。当然，在语料库语言学研究不断探索的征途中，我还要感谢慷慨分享不吝指教的同行、学者和朋友。在我读博期间，许家金教授带我进了语料库百花园，那里的语料库工具目不暇接；梁茂成教授推荐的数据处理和分析程序让人爱不释手。我还要特别感谢我的博士生周霞，她在紧张学习之余对本书做了大量校对、修改和补充工作。同时，我要感谢外研社段长城主任、董一书女士和李晓雨女士，她们的鞭策、鼓励和严肃认真的工作态度提示我多行思考，少走弯路，强一份责任，少一分敷衍，在此我衷心感谢她们。我要感谢的人应是一个长长的单子，恕我因篇幅所限，不在此一一列出。

由于作者知识所限，相关论述和介绍难免有漏误；而且，由于本人

并非软件工程技术专业人员，表述时难免欠专业性，若因技术问题妨碍阅读和理解，敬请谅解并不吝指正，我们也会根据需要及时修改和补充。

秦洪武

曲阜师范大学

2020年6月

第一章 双语语料库：基本概念和应用

本书探讨双语语料库。一般认为（如Barrière 2016：105），双语语料库至少包含平行语料库和类比语料库。本书探讨的内容涉及这两类语料库的使用，但重点是平行语料库的创建和应用。双语语料库建设在中国已有二十年的历史，它在翻译与语言研究领域的应用价值已得到学界肯定。但是双语语料库的建设与研究仍有不少问题待解决，新型语料库的开发与研制依旧是关键问题，语料库在翻译研究中的应用潜力有待进一步开发（王克非、黄立波 2012）。鉴于此，本书设置的核心话题涉及平行语料库的构成与应用、平行语料库的样本选取、平行语料库的研制、语料检索和数据统计，以及语料的跨平台应用，等等。

本章描述语料库的定义性特征和双语语料库，尤其是平行语料库的构成特征，概述双语语料库在翻译研究、语言研究、翻译实践和翻译教学中的主要用途。

1.1 语料库

Kenny（2001：22）将语料库定义为“依照某种原则方式所收集的大量文本的总汇”。早期的语料库主要通过人工收集，规模小、应用范围窄，

主要用于词典编纂、语法研究、方言研究以及语言习得研究。而现代意义上的语料库是指“按照一定的语言学原则，根据特定语言研究目的，运用计算机技术大规模收集多种文本语料的电子语料库”（王克非 2012）。语料库主要有以下特点：1）规模大。随着信息处理技术飞速发展，当代语料库的规模空前提升，库容达几亿词乃至几十亿词的语料库并不罕见。2）代表性。按照语料总体建立的抽样结构能保证样本选取代表研究对象（特定语言或语言变体）的特征。3）以电子形式保存，可自动赋码和标注，便于自动检索、查询和统计，支持语言描述和实证研究。

语料库建设的质量对于基于语料库的语言研究和翻译研究至关重要。任何基于语料库的研究均以语料库建设为前提，而且，语料库的设计及选材直接影响基于语料库的研究工作，研究结果与语料库的建设质量紧密相关（Sinclair 1991）。语料库有多种分类方式。根据语料涉及的语种，可分为单语语料库、平行语料库和多语语料库；按照用途可分为通用语料库和专门用途语料库；按照语料文本产生的时间，可分为共时语料库和历时语料库；而根据言语产生的途径和模态，又可分为书面语语料库和口语语料库。

语料库的兴起带来了新的语言研究范式，促成语料库语言学和语料库翻译学的产生。通过语料库，我们可以观察先前没有意识到或仅仅隐约觉察到的语言模式（Johansson 2007：1）。随着语料标注、分析等加工技术的发展，以及检索工具的不断更新，语料库在规模、多样性和使用方面发展迅猛，迅速成为语言研究的普遍资源（黄昌宁、李涓子 2002：2）。语料库作为研究方法已经广泛应用于语言习得、词典编纂、对比语言学、语义学、语用学、话语分析、语言接触以及口笔译等研究领域。

1.2 双语语料库

双语语料库包括两种语言的文本，两种语言之间可以是类比关系，即不同语言中交际功能相似的具有可比性的原创语篇，如双语类比语料库英国书面语语料库（FLOB）和兰卡斯特现代汉语语料库（LCMC）这两个抽样结构和大小基本相同的平衡语料库可以成为类比语料库。同样，两种语言之间也可以是对译关系（即原创语言文本和与之对应的目标语文本，即平行语料库）。当然，双语语料库还包括翻译语料库（translational corpora），这类语料库目标语一致，但源语不同，只要求篇章上对应收录，不涉及句级对应关系（王克非等 2004：7）。

我们把储存具有翻译关系句对的双语语料库称作平行语料库。例如，英汉平行语料库收录的是英语原文和汉语译文，或者是汉语原文和英语译文。根据收录语料所涉语种的数量，可分为双语平行语料库和多语平行语料库；根据翻译的语向，平行语料库可以是单向的，也可以是双向的。例如，北京外国语大学的“通用汉英平行语料库”（General Chinese-English Parallel Corpus，简称GCEPC）同时收录英汉翻译语料和汉英翻译语料（王克非 2004）。平行语料库区别于其他语料库最典型的特征是语料之间的平行对齐，即源语文本和目的语文本之间在词汇、语句和段落等层面存在对应关系或翻译关系（胡开宝 2011：34），对齐程度和对齐方式可视研究的具体需要而定。例如，用于研究不同译本翻译风格的语料库可以在一个原文本和多个译本之间进行平行对齐，而用于计算机辅助应用文体翻译的语料库通常在句子层面实现“一对一”对齐即可。平行语料是跨语言研究的重要资源，可广泛应用于双语对比和翻译转换研究，也可用于双语词典编纂和翻译实践。近年来，通过平行语料库的研制与应用，国内学者在英汉语言接触及语言互动影响方面的研究已取得显著进展。

根据双语语料库的建库目的，可大致分为研究型和应用型语料库两大类。前者主要用于语言对比与翻译研究，后者主要用于双语词典研编、计

算机辅助翻译以及自然语言处理等。用于研究的双语语料库尤其是平行语料库在研制时必须有明确的研究目标，该目标决定着语料库的性质、语料构成、语料规模、取样标准以及加工深度等。而应用型平行语料的句对通常是脱离上下文、打乱次序的孤立的句子，英译汉与汉译英语料夹杂，不易区分翻译方向（王克非 2012），本书在编写时兼顾了这两类对齐方式的应用和实现方式，将篇头数据设置、句对齐和段对齐的实现方式等列入介绍和探讨的范围。

1.3 双语语料库的应用

双语语料库既包含源语–译语的一对一平行语料库，也包含一种源语–多种译本的平行语料库，多种译本本身就能构成类比语料。进一步说，双语语料库还包含翻译语料和目标语原创语料构成的类比语料。双语语料库在翻译研究和语言研究中的独特作用已经得到学界认可。基于双语语料库的语言研究涉及翻译转换特征研究、翻译文体研究、语言接触及语言演化、语言对比研究、词典编纂、翻译教学及实践，等等。

1.3.1 双语语料库与翻译研究

1）翻译转换特征研究

既然语料库收集的对译文本是语际转换的经验性知识，研究者就有望从中观察和学习源语–目标语的对等关系和转换规律。翻译转换不仅涉及词语层面和修辞层面，还包括存在句、无主句、被动句、把字句、省略句等在内的各种结构（王克非 2008）。基于语料库观察和分析词汇、句子、修辞、语篇等层面的翻译转换，并将观察和研究结果置于特定语境中加以解释，可为翻译研究提供其他研究方法无法提供的数据支持。

基于语料库的翻译转换研究起步较早，研究成果可观。例如，Øveras

（1998）以英语–挪威语平行语料库（ENPC）为语料来源，考察翻译转换中的显化与隐化现象。Kenny（2001）借助“德语–英语文学文本平行语料库”（GEPCOLT），探讨德语源文本中创造性复合词及搭配在英译文中的范化情况及译者的创造性。秦洪武、王克非（2004）基于北外“通用汉英平行语料库”，对so…that结构和它的汉语对应结构进行描写和分析。胡开宝（2008）对《哈姆雷特》两个译本中概念功能、人际功能以及语篇功能信息的显化进行考察。黄立波（2008）运用平行语料库探讨人称代词主语在文学和非文学翻译中使用的频次和转换类型，研究发现，在语际转换中人称代词主语的使用带有源语迁移特征，语内类比显化突出。刘泽权等（2008，2009）基于《红楼梦》多译本语料库考察《红楼梦》习语和叙事标记语的再现以及译者在翻译策略使用上表现出来的规律性。

可见，借助于双语语料库提供的经验数据，翻译研究更有能力探索和发掘语际转换的语别特征和跨语言规律（王克非 2008）。

2）翻译文体研究

利用语料库分析译作文体风格、译者风格和译文文体特征，过程中产生的量化数据让分析具有相对的客观性，而不必过度依赖直觉和主观感受。在这方面，黄立波（2014：1）就是基于平行语料库，通过语际对比和语内对比探讨翻译文本中的文体特征。这类研究借助平行语料主要从语言形式上观察翻译文本表现出的特征，探讨译者在翻译域中特有的行为方式，即“译者的声音”（translator’s voice）（Hermans 1996：27），研究对象主要涉及小说、戏剧等文学文本翻译；研究路径涵盖译者翻译风格考察，对同一原作多译本翻译风格比较，翻译风格的性别差异研究，以及基于大规模文学文本语料的译文特征研究，等等；描写对象包括标点、词汇、搭配、修辞方式、句长、句式结构、叙事结构等方面，基本涵盖了语言描述可能涉及的所有层面。

例如，刘泽权、闫继苗（2010）考察《红楼梦》英文语料库中的报道动词及其三种英译，尝试探讨译者在翻译高频使用的报道标示语“（某人）

道”时表现出的风格和运用的策略。徐欣（2010）基于多译本语料库对比分析《傲慢与偏见》的三个译本，探讨译本在用词丰富程度、固化表达式、平均句长等方面表现出的差异。有的研究关注词汇使用表现出的译者风格，如萧乾与金隄在翻译与创作上表现出的用词倾向差异（王青、秦洪武 2011）。还有研究尝试提出翻译文学的描写参数，如黄立波（2014）基于“中国现当代小说汉英平行语料库”（Chinese-English Parallel Corpus of Modern and Contemporary Chinese Novels，简称CEPCOCN），对《骆驼祥子》多译本以及葛浩文、张爱玲等的译者风格进行考察，将翻译文体研究分为统计文体（包括类符/形符比、平均词长、句长、词汇密度、关键词、词类与句子分布等参数）、叙事文体（包括话语表达、指示语、情态、及物性、前景化等参数）和语言文体（以短语、搭配、语义韵、方言等为切入点考察翻译文本），建立了基于语料库的翻译文体学分析框架。作出相似尝试的还有任晓霏等（2014，2016），他们根据翻译文本的文体特征、作者风格、翻译目的、翻译性质、译者风格、读者反应等要素及其相互关系，确定了戏剧、儿童文学、小说、散文、典籍文论、传记等体裁的语料库翻译文体描写参数，构建语料库翻译文体学研究体系，并以主要参数为检索项，对经典作品进行基于语料库的多译本翻译风格比较分析。

1.3.2 双语语料库与语言研究

双语语料库中用于语言比较研究的一般为类比语料库，当然也可以使用平行语料库。从既有的研究看，双语语料库支持以下几个方面的语言研究。

1）语言接触与语言演化研究

翻译作为一种间接语言接触手段，与目标语演化关系密切，这种变化会在词素、词语、搭配、句式、句长、语篇等层面上反映出来。以英汉语言接触为例，自20世纪以来，受西洋语法影响，汉语经历了巨大变

化。例如，特定词性和句式的使用频率发生变化，句法形式严密化，分句位置发生变化等（王力 1943；秦洪武 2001；朱一凡 2011）。一般认为，触发汉语演化的一个重要因素是翻译。要考察这一演化过程，仅凭借研究者的自然语感和有限的素材显然不够，而借助大规模历时平行语料库对各年代语言样本进行比较则可以弥补这一不足，为语言演化、翻译以及翻译语言与目标语间的互动研究提供客观的描写和分析依据（王克非 2016）。近年来，已有研究开始关注基于历时语料库的英汉语言接触与汉语演化研究（秦洪武 2014，2015；秦洪武、王玉 2014；夏云 2013，2014；朱一凡 2011），研究内容涉及句段长、句长、词性分布等宏观语言特征，以及叠用词、结构容量、话语标记等微观语言特征（参见秦洪武、夏云 2017）。

2）语言对比研究

随着大规模双语语料库尤其是平行语料库的开发与应用，基于双语语料库的语言对比研究正在逐渐成长为语言研究的新领域，双向平行语料库在这一新领域所起的作用不可或缺。研究者可以对语料库中不同语言的原文部分进行比较分析，也可以结合原文和译文进行对比研究。例如，“英语–挪威语平行语料库”（ENPC）收录了英语原文与挪威语译文，以及挪威语原文与英语译文。该库已被用于英语与挪威语的对比研究，研究内容包括英语与挪威语中显现结构（presentative constructions）对比、语序和信息结构对比、词汇对比、文本的跨语言类比等不同的层面；同样，基于“英语–瑞典语双向平行语料库”（English-Swedish Parallel Corpus，简称ESPC）也有大量的对比语言研究成果问世（见王克非、黄立波 2012）。自北京外国语大学创建“通用汉英平行语料库”以来，国内学界也在持续发表英汉语对比方面的量化研究，研究内容包括英汉语“体”的分布、致使结构、把字句、被动结构等方面（见王克非 2004）。

3）平行语料库与词典研编

大型双语平行语料库内含充足的词汇、语法、语义和语用信息，在翻译研究、对比语言研究和双语词典编纂中的作用举足轻重。自20世纪60年代以来，为便于提取例句并获得使用频率信息，国际知名的研究机构就着手建设便于计算机处理和读取的语料库。李德俊（2006）认为，利用检索工具可以从较大规模的平行语料库中检索到具有翻译关系的数量可观的句对。平行语料库不仅可以辅助查找具有翻译关系的对等词，而且还可提供大量实例来解释词义、说明用法。编纂者通过观察对译句为特定词目确定较为客观的释义，或选择并呈现高频对应词。总之，编纂双语词典时，平行语料可用于义项辨识与排序、词典配例、新词新义的收集和处理、词典立目、文化局限词的处理以及词汇搭配研究等方面。

1.3.3 平行语料库与翻译教学

双语语料库在翻译教学中具有独特的应用价值。国内外已有研究着手从多个角度探讨双语语料库尤其是平行语料库在翻译教学中的可能用途（Awal *et al*. 2014；Pearson 2003；Zanettin 1998；秦洪武、王克非 2007；史汝波等 2009；王克非等 2004；王克非等 2007；王克非、秦洪武 2015）。根据现有的研究，通过观察双语句对，译者可以查找特定表达的译法，有针对性地掌握翻译技巧以解决翻译问题，提高翻译质量；其次，能够通过语料呈现，直接观察源语和译语语言系统的规则，提高语言意识。根据王克非、秦洪武（2007），语料呈现大致有语料呈现–刺激、学生浏览语料、学生报告、展示心得及发现、师生共同讨论五个步骤（见8.5.1）。

此外，还可以将学习者的翻译文本整理成电子版，创建学习者翻译语料库。通过对照自己的译文和专业译者的译文，学习者可以借此自我评析，反思不足，提高翻译能力。

总之，将平行语料库应用于教学，有利于创造自主学习的环境，使学

习者对翻译技巧的观察更加客观，解释更加充分，对两种语言的特征及其差异的认识也更加敏锐（Bernardini 2004b）。

1.3.4 平行语料库与翻译实践

平行语料库对于翻译实践的应用价值主要体现在两个方面：（1）基于翻译记忆库的计算机辅助翻译实践；（2）基于大规模句对库的机器翻译平台开发。借助SDL Trados、ABBYY Aligner、TMXmall、OmegaT等工具，可以将平行语料库转换成TXM格式的翻译记忆库。在翻译过程中，以双语句对形式存在的记忆库可以为机器翻译系统所直接使用，也就是说，当拟翻译的句子与记忆库中特定句子达到设定的匹配值时，就会呈现候选译文。在利用双语语料生成记忆并用于翻译实践方面，管新潮、陶友兰（2017）有较为详细的介绍。

随着机器翻译技术的发展和翻译质量的不断提升，机器翻译开始融入计算机辅助翻译系统，成为机辅翻译（Computer-Aided Translation，CAT）的重要支持资源。自20世纪90年代开始，研究者尝试基于大规模平行语料研制机器翻译系统，使用统计手段和对应的词表直接实施不同语言之间的翻译。从设计思路上看，机器翻译可粗分为基于规则（Rule-Based）和基于语料库（Corpus-Based）两个大类。所谓基于规则，即使用词典和规则库构成的知识源；所谓基于语料库，则使用由切分、赋码的语料所构成的知识源。一般认为，两者结合是解决机器翻译的有效途径。目前，微软、谷歌、百度、金山、网易、科大讯飞等大公司都已经开发出具有较高产出质量的机器翻译模型。

1.4 小结

本章指出，作为语料库的一种，双语语料库尤其是平行语料库虽晚于

其他类型的语料库，但它在语言服务和语言研究中的基础数据作用不可或缺。平行语料库提供多层面、多维度的经验性数据支持，其独特的功能正在翻译研究、翻译教学、对比语言研究和翻译实践中日益凸现出来；在大数据和人工智能时代，高质量的平行语料正在成为机器学习的宝贵资源，它的有效应用还将提升人工智能应对跨文化、跨语言交流的能力。总之，在平行语料基础数据资源建设方面，外语学科将作为数据提供者继续发挥其独特作用。