福建省社会科学规划项目（2010B153）成果

A CCAT-BASED STUDY OF E-COMMERCE TRANSLATION

# 基于CAT及语料库技术的电子商务翻译研究

王朝晖　余　军◎著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

福建省社会科学规划项目（2010B153
本书获厦门理工学院学术专著出版基金(

A CCAT-BASED STUDY OF E-COMMERCE TRANSLATION

# 基于CAT及语料库技术的电子商务翻译研究

王朝晖　余　军◎著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

**图书在版编目(CIP)数据**

基于CAT及语料库技术的电子商务翻译研究/王朝晖，余军著. —厦门：厦门大学出版社，2016.10

ISBN 978-7-5615-6296-3

Ⅰ.①基… Ⅱ.①王…②余… Ⅲ.①电子商务-翻译-语料库-研究 Ⅳ.①F713.36②H08

中国版本图书馆CIP数据核字(2016)第249859号

**出 版 人** 蒋东明
**责任编辑** 王扬帆 冀 钦
**装帧设计** 蒋卓群
**责任印制** 许克华

**出版发行** 厦门大学出版社
**社 址** 厦门市软件园二期望海路39号
**邮政编码** 361008
**总 编 办** 0592-2182177 0592-2181406(传真)
**营销中心** 0592-2184458 0592-2181365
**网 址** http://www.xmupress.com
**邮 箱** xmupress@126.com
**印 刷** 厦门市明亮彩印有限公司

**开本** 720mm×1000mm 1/16
**印张** 20.75
**字数** 362千字
**版次** 2016年10月第1版
**印次** 2016年10月第1次印刷
**定价** 70.00元

# 前 言

语料库翻译学研究发轫于 1993 年莫娜·贝克(Mona Baker)所发表的 *Corpus Linguistics and Translation Studies: Implications and Applications* 一文,迄今已 20 余年。期间成果丰硕,已成为目前翻译研究的主流,但也存在一些问题,如专门用途双语语料库构建较少,应用研究相对薄弱,对语言服务产业缺乏关注等等。语料库翻译学研究亟需开拓一条理论与应用结合、研究与产业接轨的新路径,以推动学科的深入发展。

在一带一路、互联网+以及我国大力发展跨境电子商务的背景下,本书以国内外学者极少涉及的电子商务翻译为研究对象,系统分析了语料库翻译学及电子商务翻译研究现状,将语料库与计算机辅助翻译(CAT)两大新技术结合,提出了语料库及计算机辅助的翻译(CCAT)这一新的研究范式,并充分论证了其理据、架构及应用。该范式将语料库与 CAT 融合,并应用于电子商务翻译的质量评估及实践操作。

在 CCAT 理论架构的基础上,本书提出了电子商务翻译的"信、达、效"原则及标准;阐述了电子商务双语语料库的研制及其应用;利用 CCAT 平台,以电商网站为案例,展开了基于实证的电子商务翻译质量评估;倡导了基于 CCAT 的电商网站人工辅助机器翻译模式。这些探讨,对于电子商务网站的建设,电子商务翻译质量的提高,都具有较强的实用价值。

技术发掘及技术推广是本书关注的一个重要部分。无论是语料自动采集技术,句子自动对齐技术,还是文本加工及处理,本书都力图对其有所突破和创新,如网络矿工在双语数据自动采集中的应用,将极大地提高双语语料的采集效率,推动大型专门用途双语语料库的建设。本书系统整理了双语语料库的建库技术,配以丰富的例证和图示,予以逐步讲解,详尽描述。在技术推广方面,做了较为有益的工作。

在本书成稿之际,我们要感谢所有关心、支持本书写作和出版的机构和个人。本书获厦门理工学院学术专著出版基金优先资助,厦门理工学院外国语学院张跃军院长、覃庆辉副院长也为本书的出版给予了资助,在此一并致谢!

感谢魏志成教授在学术上给予我们的帮助和激励。

本书第三、六、八、九章由余军撰写，约8万字；其他章节由王朝晖撰写，约28万字；全书由王朝晖统稿。

我们希望本书能引起更多人对电子商务翻译研究的关注。由于水平有限，不当之处在所难免，诚挚地希望广大读者批评指正。

王朝晖　余军

2016年1月于厦门

# 目 录

## 上篇 绪论

## 中篇 电子商务双语语料库的构建研究

# 下篇 电子商务翻译应用研究

# 上篇　绪论

本篇概述了国内语料库翻译学的研究情况。虽然语料库翻译学研究取得了较为丰硕的成果，但也存在一些问题。双语语料库的研制偏于通用类，专门用途类明显不足。而且颇为遗憾的是，已建成的语料库都没有输入CAT，或者说，尚未有人提出来应该这么做。究其原因，仍在于对CAT的一些偏见，认为CAT只适合科技翻译，不适合文学翻译，以文学翻译语料为主的语料库，便顺理成章地与CAT毫无关系了。而事实并非如此。不论是科技翻译，还是文学翻译，CAT都用处甚大。或者说，已经构建的语料库，皆可在CAT中得到利用。

CAT的历史不长，目前主要在翻译行业使用，进入学术界和大学课程之中时日尚短，仅数年而已。翻译行业使用CAT的译员、公司，并未想过将其积累的翻译记忆库做成语料库，以供学界研究之用，他们或是认为学界对此兴趣不大，或是出于为客户保密需要。学界也未提出此方面的需求，主因是学界的关注点在文学翻译领域，接触CAT者不多，即便是了解CAT者，也并未思考过CAT与语料库融合的问题。

基于上述分析，本篇倡导CAT与语料库合二为一的理念，以及CCAT这一翻译研究和翻译实践模式，通过对商务翻译研究的概述以及对电子商务翻译的初步探析，提出电子商务翻译的多元理论框架以及“信、达、效”的电子商务翻译原则和标准。

在上述讨论的基础上，我们尝试架构CCAT平台下电子商务翻译的HAMT模式。

# 第一章　语料库翻译学概述

## 1.1　引言

自1993年莫娜·贝克发表《语料库语言学和翻译研究：启示与应用》(*Corpus Linguistics and Translation Studies: Implications and Applications*)(1993)一文，开启语料库翻译学(corpus-based translation studies，简称CTS)的滥觞以来，语料库翻译学研究已日益发展成为翻译研究的重要分支和显学之一。“传统的翻译研究是以原语文本为参照，以忠实程度为取向，主要探讨译文与原文之间的关系或对应关系。语料库翻译学基本上就是语料库语言学加描写翻译研究。这也可以说是一种新的研究范式。”(王克非，2012：5)近几年来，这一新的研究范式蓬勃发展，呈现一片繁荣景象。

与传统翻译学相比，语料库翻译学注重实证研究(empirical research)，在统计分析(statistical analysis)的基础上，阐明翻译本质、翻译过程、翻译策略及翻译活动的制衡因素，从而有效弥补传统翻译学研究的先天缺陷和不足(胡开宝，2011：1)。经过20余年的发展，语料库翻译学在研究领域、研究方法、技术手段等方面，都得到拓深和发展，呈现出各种研究趋势和视角。

本书将提出并探究一个新的研究视角——CCAT，即corpus and computer-assisted translation，以电子商务翻译为研究对象，将语料库与计算机辅助翻译技术相互结合，探讨语料库及计算机辅助翻译(computer-assisted translation，简称CAT)二者融合成CCAT的理据、方法及应用。有关CCAT的探讨对于应用翻译研究具有很大的实用价值，对于语料库翻译学与语言服务行业接轨也将产生深远的影响。

## 1.2 语料库与语料库翻译学

语料库——corpus(复数形式 corpora)一词,来自拉丁语,意为身体或躯干(body),后来指某一主题文字形式的汇编、全集。对语料库的定义,目前尚无统一界定,不同研究者的定义各有差异。麦克内里(McEnery)和威尔逊(Wilson)(1996)对语料库给出了三个层次的定义:

(1)任意文本库;

(2)可以机读的文本库;

(3)可以机读的一定量的文本库,其取样可在最大程度上代表一种语言或变体。

肯尼(Kenny)(2001:22)认为,语料库是"依照某种原则方式所收集的大量文本总汇"。王克非(2012:9)的定义则是,语料库"指运用计算机技术,按照一定的语言学原则,根据特定的语言研究目的而大规模收集并贮存在计算机中的真实语料,这些语料经过一定程度的标注,便于检索,可应用于描述研究与实证研究"。

语料库的不同定义,反映了语料库发展的不同阶段[①]。早期(18—20 世纪初)的语料库并非电子文本,而是手工收集的语料,成果主要用于语法研究和词典编纂,这类语料库对应麦克内里和威尔逊对语料库第一个层次的定义。第二阶段(20 世纪 50—80 年代)的语料库已由手工收集向计算机处理及分析过渡,初步具有简单的标注,规模突破了千万词的容量,这类语料库对应麦克内里和威尔逊对语料库第二、三层次的定义。第三阶段(20 世纪 90 年代至今)的语料库利用了先进的计算机技术,规模、类型、标注、应用等方面都有了飞跃发展,出现了规模超大、类型各异、深度加工、应用更广的各种语料库,这类语料库对应王克非的定义。

基于语料库的翻译研究始于第三阶段。之前,"翻译研究经历了语文研究、语言学研究、文化研究、哲学研究和认知研究五个范式",但"最为突出的问题是定性研究和定量研究相脱节,理论研究与语言转换的实践相脱节,缺乏客观的量化标准和评估模式"(王克非,2012:12-13)。双语语料库的出现,使基于语料库的翻译研究成为可能,从而弥补了上述不足。基于语料库的语言学研究

① 有关语料库三个发展阶段的阐述,详见王克非所著《语料库翻译学探索》第 9 页。

通称为语料库语言学，因此，基于语料库的翻译研究可以称为语料库翻译学（王克非，2012：4）。语料库翻译学“是指以语料库为基础，以真实的双语语料或翻译语料为研究对象，以数据统计和理论分析为研究方法，依据语言学、文学和文化理论及翻译学理论，系统分析翻译本质、翻译过程和翻译现象等内容的研究”（胡开宝，2011：1）。

双语语料库主要包括翻译语料库（translational corpus）、对应语料库（parallel corpus）和类比语料库[①]（comparable corpus），各有其特点和用途（王克非等，2004）。双语语料库一方面推动了语料库翻译学的快速发展，另一方面由于在研制及应用上存在各种不足和问题，也成为语料库翻译学发展需要突破的瓶颈。

## 1.3　语料库的类别

语料库种类繁多，没有统一的分类，按照不同标准和分类方法可划分为不同类型。从表达形式上可分为口语语料库（spoken corpus）与笔语语料库（written corpus）；从时间跨度上可分为共时语料库（synchronic corpus）与历时语料库（diachronic corpus）；从语言类型可分为原创语料库（original corpus）和翻译语料库（translational corpus）；从选材方式上可分为抽样型语料库（sample corpus）和监控型语料库（monitor corpus）；按照语言种类可分为单语语料库（monolingual corpus）、双语语料库（bilingual corpus）及多语语料库（multilingual corpus）；按照其用途可分为通用语料库（general corpus）与专门用途语料库（specialized corpus）；从模态角度可分为单模态语料库（monomodal corpus）与多模态语料库（multimodal corpus）；按是否加工可分为生语料库（raw corpus）与标注语料库（annotated corpus）；按文本属性可分为同质型语料库（homogeneous corpus）和异质型语料库（heterogeneous corpus）；按结构划分可分为平衡型语料库（balance corpus）和随机型语料库（random structure corpus）。此外，根据文本格式的类型可分为 TXT 型和 XML 型开放式语料库，及需专用软件运行的封闭式语料库；就语料规模而言可分为小型、中型和大型语料库；根据流通途径可分为商用语料库及自建语料库；根据应用平台可分为单机版语料库和在线语料库。

① 也称可比语料库。

各种类型的语料库还可进一步划分，如双语语料库这一类别就包括口译语料库、笔译语料库和手语语料库[①]；不论从何种角度划分，“研究目的与方法是语料库类型的根本决定因素”（王克非，2012：10）。例如，以学习者为研究对象而构建的语料库被称为学习者语料库（learner corpus），其中包括学习者口语语料库、学习者写作语料库、学习者笔译语料库及学习者口译语料库；双语语料库又可以分为类比语料库和对应语料库，前者侧重于不同语言间同类文本的比较，后者侧重于句级对齐的翻译文本的分析。除研究目的和方法外，语料库的设计思路和技术手段也决定了语料库的类型。目前已经构建的大多数语料库都是单模态的，只包含文本形式，包含音频、图片、视频的多模态语料库目前尚不多见。这是由于多模态语料库的设计思路复杂，技术要求很高，其构建及应用研究刚刚起步。虽然口译语料库、手语语料库，甚至笔译语料库，都可以构建为多模态的形式，但实际上限于设计思路和技术手段，此类语料库目前较为罕见。

就语料库翻译学研究而言，双语对应语料库是其根本，发挥的作用最大，也最重要。目前构建的双语对应语料库以通用语料库为主，专门用途语料库较为少见，其构建及应用研究在语料库翻译学这一学科领域是一个薄弱环节，有待突破。

## 1.4 双语语料库的建设与加工

双语语料库的建设，如果严格按照业界规范操作的话，一般而言是一个极其复杂、费时费力的工程，涉及语料收集、整理、加工、对齐、标注、验收等诸多环节，步骤繁多，有一定技术门槛，需要投入较大的人力物力，这是国内外双语语料库构建数量不多的主要原因。而专门用途双语语料库，由于语料收集比通用语料库更难，相关经验更少，建设完成的就更少了。

双语语料库的研制，根据建库目的和建库类型，在语料类型、信息采集、语料标注等方面所采用的标准和方案会各有不同，但通常都涉及语料收集、文本整理、双语对齐、语料验收等环节，在这些环节中，可以采用一些通用技术，提高语料库建库的效率和速度。但就目前所见的文献来看，双语语料库研制这

① 手语是一种特殊的符号系统，与原语对齐的手语翻译语料库可被视为一种特殊类型的双语语料库。

些环节公开的技术不是过时了，效率较低，就是语焉不详，令人不得其门而入。因此，双语语料库建设与加工的技术开发与推广是目前极为迫切及重要的任务。不但要研究发掘新的技术和手段，并且要广为传播，不能秘技自珍。关键技术如果只是掌握在少数人手中，是不利于学科建设和发展的。我们在旧有技术的基础上，重新思考，积极创新，发掘了双语语料库研制的一些关键技术，如大规模语料的采集，双语语料的自动对齐，对齐语料的校对，标注效率的提高等，并且提出了一些可行易懂的方案。本书将通过实际案例相关流程的详细演示说明，让更多的研究者了解并掌握双语语料库研制的关键技术，壮大双语语料库的构建队伍。我们建议的一些具体操作方案见表 1-1：

**表 1-1　双语语料库建库操作方案比较**

| 双语语料库建库各环节 | 现有方案 | 推荐方案 |
| --- | --- | --- |
| 纸质文本电子化 | 扫描成图片后使用 Abbyy FineReader 软件识别为电子文本，识别率视扫描质量及文本内容而定，一般准确率在 90% 以上。<br>存在问题：扫描需投入的成本较大；识别后的校对工作量较大。 | 尽量通过网络搜集电子版，无电子版的可考虑舍弃，使用同类有电子文本的材料替代。 |
| 网页文本[①]采集 | 人工采集为主。<br>存在问题：工作量大。 | 主要利用数据采集软件，如网络矿工，批量采集网络文本。 |
| 文本处理 | 人工为主。<br>存在问题：工作量大。 | 利用相关软件及正则表达式批量处理。 |
| 双语对齐 | ParaConc，Emeditor 及 Trados 的 WinAlign 模块等。<br>存在问题：ParaConc 和 Emeditor 不具备自动对齐功能；WinAlign 自动对齐功能较强，但软件易用性略差，中文兼容方面存在一些问题，不及推荐方案易用、效率高。 | 使用 Abbyy Aligner 及雪人翻译软件的对齐模块自动对齐双语语料。 |
| 对齐语料的校对 | 人工为主。<br>存在问题：人工投入大，且易漏掉错误。 | 正则表达式批量检查对齐语料。 |

① 网页文本是双语语料库的重要语料来源，之前我们不是忽视该类文本的采集，就是主要采用人工手段收集该类文本。