

面向科普翻译的语料库建设与研究： 应用及展望

郭鸿杰* 卢 静

(上海财经大学外国语学院, 上海 200433)

[摘 要] 文章从语料库语言学的视角, 探讨了科普翻译的描述性研究范式、内容、应用及研究前景。文章简要梳理了国内外科普平行语料库的建设情况, 重点介绍了英汉科普平行语料库的几个特点, 此外, 还从科普语篇、翻译策略、翻译共性、翻译应用等四个方面具体探讨了语料库语言学的理论价值和实践意义, 并对与科普翻译相关的研究进行了展望。

[关键词] 语料库 科普翻译 应用 展望

[中图分类号] H315.9 **[文献标识码]** A **[DOI]** 10.19957/j.cnki.kpczpl.2022.03.002

一、引言

科学往往是涉及艰涩的概念、术语、知识、信息等的结构严谨的知识系统。把科学普及到社会大众 (popularization of science), 从而促成“公众理解科学” (public understanding of science), 再拓展为科学传播 (science communication), 由此来提升社会大众的科学素养 (scientific literacy), 是信息时代的一件大事。当今社会, 大部分科学文献是通过英语来完成的^[1]。因此, 要创作更多适合本土文化的科普作品, 我国的科学家和科普工作者仍任重道远。事实上, 从传播的本质来看, 科普亦与翻译有着内在的相似性和相通性。

近年来, 国内出版的科普译作数量可观, 然而其翻译质量鱼龙混杂, 不少英语科普原著常常以“遇人不淑”终场^[2]。特别要指出的是, 相比其他应用翻译研究, 科普翻译研究尚未真正形成体系, 仍处于极度“欠发达”阶段^[3], 这无疑是当代翻译理论研究的一大缺憾。

近年来, 语料库作为一种新的研究范式和研究手段广泛应用于语言学、翻译学、文学、传播学等诸多人文社科领域, 且取得了重要的研究成果。笔者认为, 语料库对与科普翻译相关的研究亦具有重要的理论意义和应用价值。基于语料库的科普翻译研究涉及科普翻译语料库的建设、

基金项目: 国家社科基金重大项目“围绕汉语的超大型多语汉外平行语料库集群研制与应用”(21ZD290)、上海哲社规划项目“两岸三地英汉财经双语语料库建设与应用研究”(2020BYY008)。

* 通信作者: 郭鸿杰, 上海财经大学外国语学院教授, 研究方向为语料库语言学、科普翻译、英汉语言对比与翻译等。
guohongjie@sufe.edu.cn。

科普翻译策略及其制约机制研究、科普话语特征研究，以及术语提取、机器翻译训练、科普翻译教学平台建设等应用研究。此外，基于语料库的研究范式在很大程度上弥补了科普翻译在定量研究方面的短板，推动了科普翻译研究从规约性研究范式向描述性研究范式的转变，拓宽了科普翻译的研究空间和疆界。特别值得一提的是，值此中国科技蓬勃发展、科技文化蔚然成风之际，推进中国科技“走出去”，向世界传播中国科技文明，是时代所向。基于英汉科普平行语料库的逆向检索功能亦有助于我们熟悉西方的科技话语模式，从而借帆出海，提升中国科技外译质量，推动中国科技国际化。

二、科普翻译语料库建设概览

廖七一指出，西方翻译研究的理论突破往往伴随着研究范式的变迁^[4]。数智时代下基于语料库的科普翻译研究呈现出强大的生命力。纵览相关文献，科普语料库建设已成规模，业已公开的科普平行语料库建设如表1所示^[5]。

表1 国内外主要科普平行语料库一览表

语料库	源语	目标语	库容	源语 / 目标语对应类型
英语—德语科普平行语料库 ^[6]	英语	德语	500000	一对一
英语—阿拉伯语《环球科学》平行语料库 ^[7]	英语	阿拉伯语	287306	一对一
英语—意大利语科普平行语料库 ^[8]	英语	意大利语	21879	一对一
英语—法语科普平行语料库 ^[9]	英语	法语	355677	一对一
英语—希腊语科普翻译语料库 ^[10]	英语	希腊语	262338	一对一
英语—希腊语科普平行语料库 ^[11]	英语	希腊语	272595	一对一
英语—汉语科普翻译语料库 ^[12]	英语	汉语	1787077	一对二
英汉科普平行语料库 ^[13]	英语	汉语	9645089	一对一 / 一对多

总体上，国内外知名科普语料库因建库目的不同，规模参差不齐，从数万词到近千万词不等，语料大多为单译本。与之相应，学界也开展了基于语料库的科普翻译研究。下文将简要

介绍一下笔者正在主持建设的英汉科普平行语料库（English-Chinese Parallel Corpus of Popular Science，简称 ECPCPS）。

第一，ECPCPS 主要收集了 20 世纪 80 年代以来源自科普书籍和科普杂志的英汉双语语料，其中包括 2002 年以来已评选 11 届的吴大猷科普译著奖中的英译中作品，如《费马最后定理》（上海译文出版社 1998 年版）、《魔鬼出没的世界》（吉林人民出版社 1998 年版）、《大崩坏：人类社会的明天》（时报文化出版有限公司 2006 年版）、《人类大历史：从野兽到扮演上帝》（天下文化出版公司 2014 年版）等，以及《中国近代科学的文化史》（上海古籍出版社 2009 年版）、《万物简史》（接力出版社 2005 年版）、《数字生存》（海南出版社 1997 年版）等其他具有影响力的科普双语文本，还包括《自然》（*Nature*）、《科学新闻》（*Science News*）、《新科学家》（*New Scientist*）、《科学美国人》（*Scientific American*）等报纸杂志上登载的科普文章及汉语译文。目前，科普语料库仍在扩容，现已超过千万字/词，

包括中国大陆、中国台湾两个子库，有上千个英汉科普对应文本，保证每个文本的语篇尽量完整，且单个英汉对应语篇的字/词数上限约为 45000。

第二，对 ECPCPS 语料库进行统计，结果显示源语单词和目标语汉字的比例是 1 : 1.64。表 2 为我们所统计到的几个大型平行语料库的英语单词、汉字数量对比结果^[13]。

表 2 平行语料库英语单词、汉字数量对比

语料库	库容	英汉平均词字数比
英汉双语平行语料库英译汉部分（非文学） ^[14]	270 万	1.72
英汉医学平行语料库 ^[15]	400 万	1.72
英汉科普平行语料库 ^[13]	964.5 万	1.64
英汉财经平行语料库 ^[16]	1000 万 +	1.72

第三，语料库加工采用了句对齐标准，以英语源语句子为参照，分割标记为句号、分号、问号、叹号等。若英语句子为完整的语义单元，碰到破折号、冒号等也进行了断句处理。王克非提到，“句子仍不失为翻译的一个主要转换单位，特别是除文学汉译英之外的另三类翻译，其 1:1 的句对齐比例达到 80% 以上……英译汉 1:1 的语句对应高于汉译英，主要原因是汉语译者翻译时多参照原文的句式和标点，特别是在比较严肃的文本中。”^[14] 根据王克非的统计结果，文学类和非文学类英译汉的句对齐比例分别为 81.9% 和 84.7%^[14]。而我们的科普平行语料库统计出 1:1 的句对齐语料约为 84%，数据基本接近。特别值得一提的是，鉴于科普翻译具有较强的应用性，翻译过程中会出现不少改写、编译、创译等，这对于句对齐语料加工是一项重要挑战。

第四，ECPCPS 主要收集了一对一类型的文本。鉴于某些出名的科普作品存在多个译本的情况，如中国大陆版和中国台湾版，ECPCPS 中也收集了同一地域或不同时期的多个译本。这不仅有利于基于平行语料库的语言对比研究，而且有助于揭示翻译模式以及影响翻译策略的底层机制和动因^[17]。

三、语料库在科普翻译中的应用

如前文所述，语料库为科普翻译研究提供了一个新视角，引起了科普翻译研究范式的变化，拓展了科普翻译研究的深度和广度。下文将从科普语言特征、翻译共性、翻译策略、翻译应

用四个方面介绍基于语料库的科普翻译研究现状和未来前景。

首先，译语语料库海量的数据有助于高效准确地获取一些语言特征的计量结果，在

宏观层面，包括词单（word list）、关键词单（keyword list）、词频分布（frequency profile）、词频谱（frequency spectra）、平均词长（mean word length）、词串（cluster）、词覆盖率（coverage）、词汇密度（density）、平均句长（mean sentence length）等；在微观层面，可以分析主题语气词、量词、固定习语、句型、语用、隐喻、篇章等内容。科普读物为吸引大众读者，常常比科技文本写得更为生动有趣，通俗易懂，从而让读者享受这种知性的乐趣。通过语料库可以容易地捕捉到这些语言特征。比如，“摹声词”在科普读物中的出现频率是十分高的^[18]，略举几例如下：

They can snap, whistle, hum, vibrate, boom, and whine.

（羽毛）能发出啪嚓声、哨笛声、嗡嗡声、颤动声、隆隆声与刺耳的尖锐声。（《羽的奇迹》）

Tigers did not purr at all but instead emitted “a peculiar short snuffle, accompanied by the closure of the eyelids” when happy.

老虎完全不会发出呼噜声，不过开心的时候，会用鼻子发出“一种特别的短嗤声，然后阖上双眼”。（《动物也疯狂》）

Sooner or later, there will be real human hardware, great whirring, clicking cabinets intelligent enough to read magazines and vote, able to think rings around the rest of us.

迟早有一天，会出现真正与人一样的硬件，出现一些嗡嗡叫、噼里啪啦响的聪明的大盒子，能读杂志，能参加选举，脑瓜转得极快，快得我

们没法比。(《细胞生命的礼赞》)

大自然的神秘奇妙时时让我们叹为观止。

诸如“摹声词”之类的语言，生动贴切，能够帮助读者享受到阅读之趣，感受到语言之美。徐彬、郭红梅也提出，阅读、翻译当代科普书籍，我们会越来越感觉许多一流的科普作品也是一流的科学散文作品^[3]。

其次，翻译语言早就引起了学者的研究兴趣。学界一般持两种观点。一种观点认为翻译文本是一种可预测的语言变体 (variety)，这归因于受翻译为媒介的间接语言接触的影响。这种变体常被称为“第三码” (third code)^[19]。莫娜·贝克 (Mona Baker) 提到，在分析译本语言时完全可以把源语抛开进行分析，并在此基础上提出了翻译共性假设 (translation universals)，即相对于源语和目标语原创语言，译文具有显化 (explicitation)、简化 (simplification)、消歧 (disambiguation)、规范化 (normalization) 等特征^[20]。另外一种观点认为翻译语言偏离目标语规则，被标签为“翻译腔”，这是一种消极的语言观。事实上，基于语料库的翻译语言计量分析结果有助于客观地呈现这种“第三码”的共性特征，而非一种“蝴蝶标本式”的感性认识。譬如，分析翻译语言特征的一个重要参数为词覆盖率，即在词频表中按次序选择一定数量的单词，计算这些单词在总语料中所占的比例^[21]。ECPCPS 中前 50 个常用词的覆盖率统计结果显示，中国大陆译本均高于中国台湾译本。由此推测，中国台湾译本用词更富于变化。在此基础上，我们将基于译本的相关数据与汉语原创文本比较，结果见

图 1。其中，原创文本的数据参照了彭临桂有关两岸小说译本词汇覆盖率的数据^[22]。

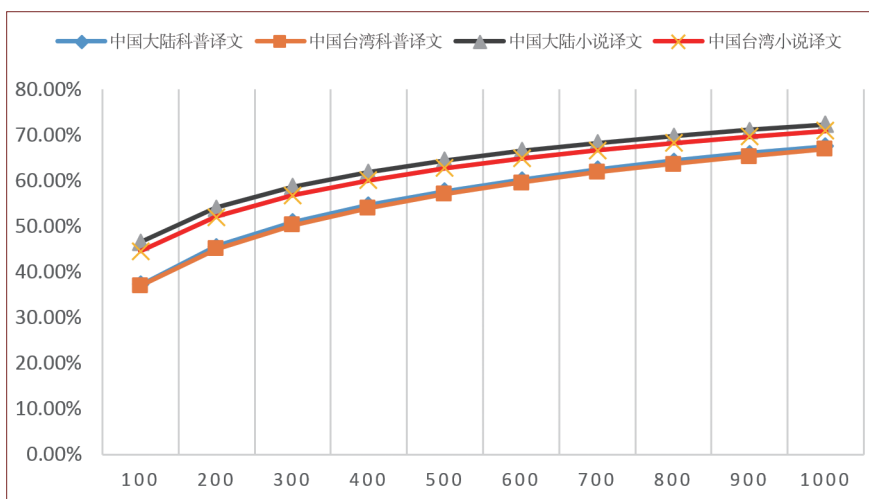


图 1 两岸科普和小说文体译本词汇覆盖率差异

依上图，4 条折线呈现出三个分布趋势：第一，无论是中国大陆译本，还是中国台湾译本，小说译本的词汇覆盖率显著高于科普译本，可见科普文本的词汇变化更大；第二，词汇覆盖率的地域差异在小说文本中差异更加显著，科普文本的词汇覆盖率特征比较稳定；第三，相似之处是，两种文体词汇覆盖率的差异在前 300 个常用词最显著，随着词频的减少，差距越来越小。

再其次，基于语料库的研究还有助于从语素、词语、习语、隐喻等多个层面分析科普翻译的策略和技巧。比如，一词多译是科普中常见的翻译策略。海量的自然语言例证能更清楚地显示该词或词语丰富的语义特征。下文将以语料库中出现的 development 一词为例来说明：

Also, the rise or origination of anything by natural development, as distinguished from its production by a specific act.

或者任何事物经由自然演变的增长或初生，有别于经由特定作为而产生。(《从达尔文到爱因斯坦》)

One of the principal benefits of the development

of human intelligence is our ability to understand the true nature and import of dreams.

理解梦一般的生活实质和含蓄的意思了解人类智力进化是很有好处的。((《伊甸飞龙》))

On the other hand, the sun of Naples might be conducive to learning something about the biochemistry of the embryonic development of marine animals.

另一方面,那不勒斯的阳光倒可能有助于学习海洋动物胚胎发育生物化学。((《双螺旋》))

These plants are threatened by lumbering, grazing, and development.

这些植物受到砍伐、放牧和开发的威胁。((《花朵的秘密生命》))

The development of a flower is one of the things we understand least about plants.

花的成长是我们对植物最不了解的部分之一。((《花朵的秘密生命》))

Drug development will change in two dramatic ways.

药物的研制工作将在两方面彻底改观。((《细胞叛逆者:癌症的起源》))

由此,不同主题内容下的代表性例句,为分析一词多译现象提供了重要参照。此外,语料库还会提供影响一词多译的其他语言因素或非语言因素。

一般说来,人们在理解抽象艰深的新理论或新概念时,常常会提取已存的认知基模做映射(mapping)。此类以一种具体熟悉的概念结构来构造另一种陌生抽象概念的现象被称作隐喻(metaphor)。事实上,为吸引读者注意并增进对新概念的理解,科普文本经常使用隐喻。隐喻对形成科学观念的重要性亦逐渐受到重视,隐喻翻

译研究也成了科普翻译的重要话题。比如,以科普文本中英文植物词或植物结构习语及其汉译为例,基于 ECPCPS 的研究结果显示,英汉植物词语义的异同可大致分为重叠、错位和空缺三种情况,在此基础上,其翻译策略可归纳为直译、意译、替换三种手段等,详见以下各例:

He recalled a child in Memphis who was an excellent student, got influenza, and became 'a vegetable'.

他回想起孟菲斯的一个孩子,曾是一名优秀的学生,患上流感之后却变成了“植物人”。((《大流感》,直译法)

Health inspectors were looking for cases among civilians "to nip the epidemic in the bud".

卫生检查员正在寻找平民病例以便“将流行病扼杀在摇篮中”。((《大流感》,替换法)

But numbers do not fall ripe into our laps, someone has to find and fetch them; far easier, some feel, not to bother.

有用处的数字绝不会凭空而降,一定要有人去发现、获得,但是有些人认为,别自找麻烦会比较好。((《如何用数字唬人》,意译法)

英汉科普平行语料库不仅为我们提供了有关英译汉策略的数据支撑,而且其丰富的科普英语语料亦对汉译英实践提供了重要参考依据。譬如,以汉语成语的英译为例,中文为母语的译者可能十分熟悉汉语成语,但不一定有能力将其译成地道的英文;而英文为母语的译者往往只能在成语字典找到直译、历史典故或是冗长的解释^[23]。如果从回译(back translation)的视角来看,英汉平行语料库的逆向搜索功能则在某种程度能够弥补这种翻译的缺憾。比如,以“生死攸关”这个成语为例,我们通过检索

汉语译文，会发现其对应的英语表达方式灵活多样，结果见表3所示。

“生死攸关”的英语对应形式包括单词、习

表3 部分科普作品中与“生死攸关”相关的英语表达

英语源语	出处
a life or death matter	<i>Death Penalty in the U. S.</i>
a matter of life and death	<i>Pale Blue Dot</i>
a matter of life and death	<i>The Artful Universe</i>
crucial	<i>The New World of Mr Tompkins</i>
life-and-death	<i>Gift From The Sea</i>
	<i>Out of Their Minds: The Lives and</i>
life-critical	<i>Discoveries of 15 Great Computer</i>
	<i>Scientists</i>
life-threatening	<i>One Renegade Cell: How Cancer</i>
	<i>Begins</i>
to be fatal	<i>Darwin's Dreampond</i>

语、短语、复合词等，而且语法功能也不完全能对号入座，包括名词（短语）、表语性形容词、描绘性形容词等。显而易见，英语表达灵活，追求变化，不拘泥于某一固定结构。所以，双语平行语料以及英语源语语料无疑对提高中译英质量具有一定的启发意义和参考价值。

最后，从翻译应用来看，科普翻译语料库的价值主要体现在机助翻译、机器翻译、检索

平台三个方面。加工好的句对齐语料除了用于构建平行语料库或检索平台外，还可以用作翻译记忆库（translation memory），协助人机翻译。ECPCPS语料库主要分为五大子库，即自然科学库、生命医学库、地球环保库、技术发明库、科技教育库。每个子库下又分为若干小类。

例如，自然科学包括化学能源、数学统计、物理机械、宇宙航空、信息智能等，这样可以确保语料内容丰富，包罗万象，从而在借助Trados等机辅翻译工具进行人机协同翻译时，就可以根据翻译的题材调取各个主题内容下的翻译记忆库。因此，无论从主题相关性，还是储存的高质量句对齐语料来看，都会大大提升翻译效率。

人工智能时代下数智技术应用日益广泛是科普翻译无法回避的现实，机器翻译已经承担了译者以前大量重复枯燥的劳动。因此，机器学习或深度学习模型的应用，将为科普翻译提供另一种新视野，即呈现出更为细腻的数字信息，进而帮助识别大数据科普语料库框架下科普翻译的语义内涵。而基于高质量精准翻译语料，借助于深度神经网络机器翻译模型，可以训练机器翻译的深度和精准度。

目前，我们已经根据英汉科普平行语料库，开发了网络共享检索平台（SUFE-Corpus），为科普翻译爱好者或者译者提供浏览、检索、统计等各项功能，深化智能化、专业性、共享型资源建设，如图2所示。



图2 SUFE-Corpus 英汉科普翻译检索平台

此外,科技术语是科普语篇进行叙述和描写的重要手段^[24]。在技术层面,语料库通过提取科普术语,建立术语库,规范术语译名,有助于训练机器翻译,推动双语科普术语库的构建,将实现术语查询、归类、对照、统计等功能;在语言层面上,基于对比短语学的理论框架^[25],分析双语术语在构词理据、形式结构、功能关系、搭配句法、隐喻认知层面的异同,以及通用词汇和科技术语的转换机制;在翻译层面上,探讨术语翻译策略选择的国际化与民族化、术语译名的规范化和本地化等。

在翻译教育背景下,语料库建设亦与翻译教学存在天然契合。众所周知,可比语料库(comparable corpus)已经应用于翻译教学。相比之下,平行语料库应用于翻译培训的潜力尚未开发。事实上,越来越多的学者提出,平行语料库可以应用于开发翻译教学案例库,辅助教材编写、词典编纂等,从而解决资源短板和时效瓶颈等问题。另外,基于语料库的定量研究结果对于翻译质量评估亦具有重要的借鉴意义。

四、结语

数智时代下的双语语料库建设在数字人文基础建设中大有作为。构建一个动态性、多维度、多层次的科普翻译语料库有助于把科普翻译置于一个大历史背景中去观照,从而有助于准确把握科普翻译与时代背景、意识形态、地域文化、译者主体等社会因素之间的互动关系。这不仅有助于科普翻译学科体系的建设,而且也有助于激发科普翻译研究的多学科交叉与多元化突破。基于语料库的科普翻译研究方兴未艾,将来可以在以下几个方面继续探索,如研发语料标注系统(如翻译策略的标注、句法系统的标注等)、术语抽取、机器翻译训练的效率,科普翻译在个别语言中呈现出何种异质性,在跨语言中又呈现何种同一性,原创语言和翻译语言的隐喻性表达差异,如何将可比语料库和平行语料库结合并更有效地应用于“以译者为中心”的翻译教学模式,以及基于语料库的翻译质量评估等方面。此外,科普翻译语言会对科普原创语言乃至现代汉语的词汇、构词,甚至句法带来什么样的影响等话题,也仍有十分广阔的研究空间。

参考文献

- [1] SHAKAS H. Translation Quality Assessment of Popular Science Articles: Corpus Study of the Scientific American and its Arabic Version[J]. Trans-kom, 2009, 2(1): 42-62.
- [2] 潘震泽. 自然科学书籍的翻译[R]. 2001 年“翻译工作坊”研讨会, 2001.
- [3] 徐彬, 郭红梅. 科普翻译的挑战[J]. 上海翻译, 2010(4): 45-49.
- [4] 廖七一. 翻译研究的趋势与中国译学的现代化[J]. 中国外语, 2006(2): 6-8.
- [5] 郭鸿杰, 周芹芹, 管新潮. 英汉科普平行语料库的创建和研究[J]. 外语与翻译, 2015(04): 38-43, 4.
- [6] HOUSE J. Covert Translation: Popular Science[EB/OL]. (2011-06-30) [2022-06-26]. <http://hdl.handle.net/11022/0000-0007-BFF1-2>.
- [7] MERAKCHI K, ROGERS M. The Translation of Culturally Bound Metaphors in the Genre of Popular Science Article: A Corpus-based Case Study from Scientific American Translated into Arabic[J]. Intercultural Pragmatics, 2013, 10(2): 341-372.
- [8] MUSACCHIO M T, ZORZI V. Scientific Controversies and Popular Science in Translation Rewriting, Transediting or Transcreation? [J]. Lingue e Linguaggi, 2019, 29, 481-507.
- [9] PEARSON J. Using Parallel Texts in the Translator Training Environment[C]//ZANETTIN F, BERNARDINI S, STEWART D. Corpora in Translator Education. Manchester: Routledge, 2003.

- [10] PAPADOUDI D. Conceptual Metaphor in English Popular Technology and Greek Translation[D]. Manchester: The University of Manchester, 2010.
- [11] MATSIRA M. Extracting Terms from an English–Greek Popular Science Parallel Corpus for Translation Teaching Purposes[D]. Birmingham: University of Birmingham, 2008.
- [12] CHEN W. Explication through the Use of Connectives in Translated Chinese: A Corpus-based Study[D]. Manchester: The University of Manchester, 2006.
- [13] 郭鸿杰等. 英汉科普平行语料库 [DB]. 上海: 上海财经大学, 2018.
- [14] 王克非. 英汉 / 汉英语句对应的语料库考察 [J]. 外语教学与研究, 2003(6): 410–416, 481.
- [15] 管新潮, 胡开宝, 张冠男. 英汉医学平行语料库的创建与初始应用研究 [J]. 当代外语研究, 2011(9): 36–41, 61.
- [16] 郭鸿杰等. 英汉财经平行语料库 [DB]. 上海: 上海财经大学, 2022.
- [17] LAVIOSA S. Corpus-based Translation Studies[M]. Amsterdam & New York: Rodopi, 2002.
- [18] 郭鸿杰, 宋丹. 基于语料库的英汉对比与翻译 [M]. 上海: 复旦大学出版社, 2020.
- [19] FRAWLEY W. Prolegomenon to a Theory of Translation[C]//FRAWLEY W. Translation: Literary, Linguistic and Philosophical Perspectives. London & Toronto: Associated University Presses, 1984.
- [20] BAKER M. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research [J]. Target, 1995, 7(2): 223–245.
- [21] 杨惠中. 语料库语言学导论 [M]. 上海: 上海外语教育出版社, 2002.
- [22] 彭临桂. 语料库翻译研究: 两岸小说译文语言分析 [D]. 台北: 台湾师范大学翻译研究所, 2009.
- [23] PELLATT V, LIU E T. Thinking Chinese Translation: A Course in Translation Method: Chinese to English [M]. New York: Routledge, 2010.
- [24] 杨信彰. 英语科技语篇和科普语篇中的词汇语法 [J]. 外语教学, 2011(4): 18–21, 100.
- [25] 卫乃兴. 基于语料库的对比短语学研究 [J]. 外国语, 2011(4): 32–42.

(编辑 / 姚利芬 齐钰 朱彦霏)

Corpus Building and Corpus-based Studies in the Translation of Popular Science: Application and Prospect

Guo Hongjie Lu Jing

(School of Foreign Studies, Shanghai University of Finance and Economics, Shanghai 200433)

Abstract: From the perspective of corpus linguistics, this study intends to address the descriptive paradigm, research topics, application, and prospect of translation of popular science. It sketches the well-known parallel corpora of popular science with a focus on features of ECPCPS (English–Chinese parallel corpora of popular science). Furthermore, it discusses the theoretical and practical significance of corpus linguistics as applied to the analysis of discourse features of popular science, translation strategies, translation universals, and its application, and provides an outlook on research related to popular science translation.

Keywords: corpus; translation of popular science; application; prospect

CLC Numbers: H315.9 **Document Code:** A **DOI:** 10.19957/j.cnki.kpczpl.2022.03.002