《语料库实用指南》评介

杨晓军

（湖南科技大学外国语学院，湖南 湘潭 410201）

Lynne Bowker and Jennifer Pearson. 2002. *Working with Specialized Language：A Practical Guide to Using Corpora*. London: Routledge. xiv + 242pp

《语料库使用指南》主要介绍研究专业语言时使用语料库的种种原则和方法、语料库的设计，以及语料库的应用。该书提出了设计和创建专用语料库(LSP Corpora)的指导性原则，以及开发和利用语料库资源的实用方法。书末有附录和术语表(Glossary)各一个。附录归纳了各章节末尾所提供的语料库分析工具和语料资源，信息量大，实用性强。此外，书中共有图表包括语料库检索图示共 122 个，列举了大量的原始数据和真实案例分析。该书是翻译工作者、专业写作人员以及从事基于语料库方法研究专门用途语言教学和学习的人士的理想读本。读者通过 www.routledge.com/textbooks/0415236991/可获得有关此书的信息。

该书正文共有三大部分，分为 12 章节。第一部分：入门知识(Setting the Scene),包括引言、第一、二章，主要介绍如何使用本书、语料库和专门用途语言等基本概念，创建和使用语言语料库的原则和方法，重点解释了为什么要考虑应用语料库来帮助研究专门用途语言。第二部分：语料库的设计、创建和加工（Corpus design, compilation and processing），共五章，介绍基于语料库的研究，阐述如何设计、创建专用单语和多语语料库，解释为有助于专门用途语言的研究如何标注和对齐文本以及对语料进行加工的理由；讨论如何利用为不同语料库专门设计的分析工具研究语料库。第三部分：专门用途语言中基于语料库的应用研究(Corpus-based applications in LSP)，共五章，探讨专用语料库的一些具体运用，如提取术语、创建术语表、专业写作和翻译，以及如何使语料库分析工具更好地服务于专门用途语言的各项研究。

本书前两部分主要是理论介绍，最后一部分是应用研究。全书围绕如何使用语料库这个主题，首先对语料库及其分析工具等一些基本概念进行界定，如：语料库、通用参考语料库(general reference corpora)、专门用途语料库(special purpose corpora)、平行语料库(parallel corpora)、可比语料库(comparable corpora)、语料库检索(concordance)、形符(tokens)、类符(types)、通用语言(LGP)和专门用途语言(LSP)等。第一部分这些内容为随后两部分内容的理解铺平了道路。接着在第二部分中花五章的篇幅探讨语料库的设计标准、创建技巧和加工方法。在第三章中，作者比较了通用和专用语料库设计标准的异同，提出专用语料库设计标准应包括语料库的规模（从几千字到几百万字的语料库均有不同用途,越大并非总是越好）、文本摘录还是全文收录（后者为宜）、文本数量（最好是多个作者多样文本）、语体（以书面语为宜）、主题领域（依研究目的而定）、语篇类型（依研究主题而定）、作者身份（权威著作）、语言选择（源语文本或原文和译文文本）、出版时间（以近期出版物为主）以及版权（尽可能获得各文本使用权）等。第四章讨论语料库创建技巧，解答了从文本收集到语料库建立的不同阶段所遇到的各类问题，如：如何寻找文本、如何获得使用文本的权利以及如何检索和存储文本等，尤其说明了万维网（World Wide Web）和 CD-ROMs 作为建立语料库资源的潜在优势。第五章首先区分语料库的标注（markup）和赋码

(annotation)概念,然后介绍不同的标注和赋码系统以及它们在基于语料库的研究中的用途。第六章在比较平行语料库和可比语料库的基础上，重点探讨平行语料库的创建和应用，强调对双语文本进行预加工（pre-processing）和对齐处理（alignment）的重要性, 并用具体例子阐释了平行语料库在语言学习和翻译研究中的应用价值。第七章分析如何运用常用的语料库加工工具（如，WordSmith Tools—http://www.oup.com/elt/global/isbn/6890 和 MultiConcord—http://www.copycatch.freeserve.co.uk）对语料库中的文本进行词汇列表（word lists）和检索，以及如何解读所列词汇表和检索结果,并总结了词表、关键词表、词串（word clusters）以及检索行各自的功能。最后一部分集中探讨专门用途语言中基于语料库的应用研究，是全书的重中之中和新颖之处。该部分共分五章，第八章研究如何运用语料库及其加工工具来产生术语表，用具体例子说明术语选择、定义的确立和术语表的内容与设计，指出了术语表在翻译和专业写作中的作用。第九章概述了术语识别和术语提取的常用方法（主要分语言学方法和统计方法），并解释了如何运用这些技巧研究专门用途语言，还比较了术语提取常用工具（如，Multitrans—http://www.multicorpora.ca 和 Xero TermFinder—http://www.mkms.Xerox.com）的优缺点。第十章探讨专门用途语言语料库作为专业写作资源的潜在优势，既可写出特定语篇类型的文章，也可按特定文体要求进行写作。第十一章列举了许多例子来说明如何把双语和单语语料库看成翻译资源来帮助解决寻找恰当的翻译对等语、选择同义词、决定适当的文风和求证概念信息。第十二章简述语料库在研究专门用途语言中的其它用途，如，创建学习者语料库（learner corpora）、研究新词(neologisms)和隐语(metaphors)等。

全书探讨的问题几乎涉及语料库创建和应用领域的各个方面。全书共有图表 122 个，演示了语料库加工和分析工具的所有功能，具有实践指导意义和实用参考价值。书后附录及术语表可称为基于语料库研究领域的“百宝箱”。附录中既收录了大部分语料库加工和分析工具软件，并附有功能解说和联系方式及其网址等信息，又收录了有关术语提取工具主要产品和常用搜索引擎的名称和网址。此外，还收录有三大资源库：高效网页搜索指南、创建专门用途语言语料库可用资源（覆盖商业/经济、计算/技术、科学、法律和医药等领域）和多语语料库信息资源（介绍了全球与语料库建设相关的六大主要工程及其网址，如 ELRA, ELSNET, LDC 和 INTERSECT 等）。术语表收列了本书所用术语共 54 个，除有定义外，还附上了每个术语所出现的页码。值得一提的是，每章后的阅读书目和全书参考文献包括了语料库研究领域里的主要文献信息。

由于作者是将此书作为教材来编写，阅读对象也包括从未接触过语料库语言学的初学者，加之书后附有相关术语表，所以它既是一本指南，也是一本入门书，读起来不费力。如果读者已经有基于语料库的研究经验，则可跳过第一、二部分，直接研读第三部分，即应用研究部分。若想了解有关专门用途语言语料库创建知识，则又可回头查阅相关章节。此外，作者所介绍的解决语料库建设中所遇问题的方法十分有效。比如，在谈到平行语料库段落对齐时常遇到因翻译等因素造成双语文本段落数不一致的情况，作者提出了三种方法来可供查找原因和解决段落对齐问题：一是查找是否因硬回车（hard return）造成的;二是可能因为原文中的一段被翻译成了两段，或者相反，那么最好的解决办法是将段落合并以确保段落一一对齐，但为了确保文本的完整性和翻译的效果，还得在段落合并处加上文本标记所没有的特殊符号（如££）。此外，有时可能出现这样的情况：原文中某一段根本就没翻译，或者译文中某一段根本就找不到对应原文。作者对前一种情况解决的办法是在译文中创建一段注明“未译段落”，对后一种情况解决的办法是在原文中创建一段注明“译文新增段落”即可（Bowker and Pearson 2002：97）。当然，有的软件也可采用段落对空的办法。另一个实用性很强的例子是作者提出如何利用电子邮件说服版权持有人允许在语料库中使用其文本的方法（详见 Bowker and Pearson 2002：61）。

本书有些局限性，如所涉语种较少，仅谈到少数欧洲语言如英语、法语，而如今语料库发展趋势是从欧洲大语种过渡到了欧洲小语种，并延伸到了亚洲各语种和其它地区的一些少数语种。由于作者对亚洲语言和语料库建设不了解，全书未提及中国所建成的最早的两个专门用途语言语料库—the JDEST Corpus(Shanghai Jiaotong Daxue Corpus of English For Science and Technology，Yang 1985；上海交通大学科技英语语料库) 和 the GPEC (Guangzhou Petroleum English Corpus，Zhu 1989；广州石油英语语料库)。而在探讨平行语料库的加工程序中，作者也忽略了汉语等语言书写时并非像英语那样词与词之间是分开的事实（对汉语语料加工时，首先就得进行分词，国家语言文字工作委员会和北京大学计算研究所为此联合制定了《现代汉语分词规范》），书中对分词这一程序只字未提。

参考文献

[1] Bowker ,Lynne and Jennifer Pearson. *Working with Specialized Language: A Practical Guide to Using Corpora*［M］. London: Routledge, 2002.

[2] Pearson, Jennifer. *Terms in Context*［M］. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1998.

[3] Leech, Geoffrey. Teaching and Learning Corpora: a Convergence［A］in Anne Wichmann, Steven Fligestone, Tony McEnery and Gerry Knowles(Eds). *Teaching and Language Corpora*［C］. London: Longman Limited, 1997.

Lynne Bowker and Jennifer Pearson：Working with Specialized Language：A Practical Guide to Using Corpora . 2002

YANG Xiao-jun

(School of Foreign Languages, Hunan University of Science & Technology, Xiangtan, 411201,China)

收稿日期: 2004-7-10;　本刊修订稿: 2004-7-20

作者简介: 杨晓军，湖南科技大学外语学院副教授，北京外国语大学中国外语教育研究中心在读博士生；研究方向为英汉语比较与翻译、语料库语言学。