语料库翻译学文库

语料库与Python应用

管新潮◎著

上海交通大學出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

语料库翻译学文库

受2015年度教育部人文社科研究规划基金项目（15YJA740009）的资助

语料库与Python应用

管新潮◎著

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

内容提要

本书以如何在语料库的教与学及其应用、语料库科研中习得 Python 能力的逻辑关系为线索，描述了 Python 的价值、意义和作用，并将内容组合成可有效助力于 Python 能力习得的三个层次。第一层次是掌握与语料库相关的基础性代码；第二层次是活学活用这些基础性代码；第三层次是以创新方式运用这些代码去解决与语料库相关的较为复杂的问题。Python 是语料文本处理的利器，需要在一定的理念指导下方可充分理解其在特定领域内所呈现的特征，而本书的首要目标就是帮助读者去运用这一"语言＋技术"理念，其次才是 Python 技术本身。

本书的适用读者是那些设想从语料库中挖掘出更多信息的文科生、文科教师或相关的研究人员。

图书在版编目(CIP)数据

语料库与 Python 应用/ 管新潮著. —上海：上海交通大学出版社，2018
(语料库翻译学文库)
ISBN 978－7－313－19748－1

Ⅰ. ①语… Ⅱ. ①管… Ⅲ. ①软件工具—程序设计—应用—语料库—研究 Ⅳ. ①H0－39

中国版本图书馆 CIP 数据核字(2018)第 160808 号

语料库与 Python 应用

著　　者：管新潮
出版发行：上海交通大学出版社　　地　　址：上海市番禺路 951 号
邮政编码：200030　　电　　话：021－64071208
出 版 人：谈　毅
印　　制：上海盛通时代印刷有限公司　　经　　销：全国新华书店
开　　本：710 mm×1000 mm　1/ 16　　印　　张：14
字　　数：236 千字
版　　次：2018 年 7 月第 1 版　　印　　次：2018 年 7 月第 1 次印刷
书　　号：ISBN 978－7－313－19748－1
定　　价：58.00 元

前 言

“语料库+”时代仿佛是人间四月天，此时此刻完成本书的写作，有如春风拂煦、绿色满园，好不惬意。其实，写作本书的目的就是设想在万物皆生长的时节里，为“语料库+”烹饪出一道别样的美食。品味其色香，思量其作用，希望以此能够造就出更多“语料库+”的机会，让人思考，给人希望，使人进步。在这个现代技术充斥着人们生活的社会，写作本书的初衷是想展示一种技术逻辑思维，让语言与技术能够实现更美好的结合，呈现出更多令人难以忘怀的语言技术机会。

在学界，语料库的发展的确让人看到了诸多的希望和机会，在业界也是同样的道理和境遇。记得有一位国外学者在一次国际会议上曾说：“Corpus data can go only so far on their own, but corpus techniques will continue to play a vital role in combination with a range of other approaches and methods.” 细看来，本书也是一次尝试，尝试着拓展相关的机会，尝试着让文科生在面对纯技术思维的业者时拥有更加笃定的话语表达权。故此设定本书的适用对象为文科生，而且还将 Python 所要研究或处理的对象限定为语料库，意在增强文科生对 Python 的好感，感受其为语料库研究和应用所能带来的利好。

本书作者曾在《语料库与翻译》一书中提出了若干问题，如语料库检索和分析工具欠缺、软件编码格式不一致、术语提取准确性不足、技术工具融合应用问题等。这一次写作本书，在一定程度上为这些问题给出了初步的答案。如相关工具欠缺问题，这为 Python 施展其功用带来了机会，因为既有语料库工具的功能随着工具的定型成为“有限”，而 Python 编程所能开发的功能却是无限的；又如术语提取问题，其所适用的是用于提取多连词的工具包，书中的案例为此提供了一个较好的解决方案；再如技术工具融合问题，这一点已准确

反映在 Python 的使用理念之中，即增强人们的逻辑思维能力。

因此，本书的特色有三：

- 一是聚焦于语料库。所有的代码、代码段或代码块均围绕语料库这一主题展开，所处理的对象是语料库研究过程需要面对的各种相关问题。
- 二是语言＋技术＋法律三位一体。这意味着案例的解释都包含了三方面的要素，而非从纯粹的技术角度出发。所选择的语料也多是法律文本，目的是为了呈现法律文本的语篇特征，即以法律文本通过技术研究获取其语言学方面的特征信息。
- 三是深入浅出，易学易用。本书强调“编程”并非是工科专业的“专利”，文科生同样可以学会编程，而且能够编写出更为细腻、更利于语言学处理的代码。再者，阅读本书至少可习得一种能力即读懂代码的能力，就像学会一门自然语言一样。

本书的写作过程其实就是一个教与学的过程。所编写的程序均以语料库语言学和翻译学领域的问题为处理对象，这种问题导向的思路可以更加贴近学生的所思所想。在实施过程中，先确定需要解决的语言学问题，明确之后再展开编程工作。而在完成代码编写后，又将代码直接交由学生进行测试，去检验相关问题，即相关案例的语言学意义是否真实，解决了什么样的语料库语言学和翻译学问题，相关代码可否进一步实施优化，等等。

依据这一过程，我们将本书内容编排为三个层次，这是本书的知识组合架构，也是学生学习 Python 编程的三个阶段（对应本书的上中下三篇）：

- 第一阶段是熟悉 Python 应用于语料库的基础性代码，试看 Python 能够解决哪些与语料库相关的基础性问题。这一阶段出现的代码就像是一些基本公式，而且是以人们所熟知的语言形式来表述的。记住：理解这些代码的含义和作用，而无需死记硬背。
- 第二阶段是以第一阶段所熟知的代码去解决真实的语料库问题，如语篇词汇特征、现有语料库工具所能解决的问题等。这一阶段旨在牢固掌握基础性代码的编程运用。
- 第三阶段是借用与创新并存，讲求代码的创造性应用，也就是说如何去解决语料库语言学和翻译学领域中未知的或需要优化解决的问题。相

关案例均源于语料库研究和应用实践：有的是为了获取更为可靠的数据而设置，有的则是在阅读相关论文后设置的，目的在于使案例更具语言学意义，等等。

中篇和下篇的案例所涉代码已经能够解决语料库研究和应用中的实际问题，但这也绝不意味着相关代码已没有需要继续优化的可能。这或许正是Python的魅力之所在，我们会努力发现能够更为有效解决问题的新工具包，使之前费力编写的代码可瞬间替换为一两行代码。Python的魅力还在于其应用对象不仅仅是语言学本体，还在于通过编程可以揭示诸如Alzheimer症所能体现出的语言学规律性问题。后者的意义已经远超语料库研究本身。

为了撰写本书，我们特意组建了一个临时团队，意在有的放失地解决论文写作过程中出现的各种问题，在此特别感谢他们的辛劳付出：上海交通大学的郭鸿杰老师，为解决语言学方面的问题提供了学术支撑；上海交通大学的金毅老师，为算法设计提供了技术逻辑支持；上海交通大学外国语学院2016级MTI的王天奇同学(4个代码案例)和2018级MTI的李建林同学(1个代码案例)，为代码的创造性应用编写了若干相关的代码案例。参与本书代码测试的同学除了前述两位，还有上海交大外国语学院和东南大学外国语学院的部分2017级MTI研究生。

作为一次尝试，每当编写完成一段可有效执行的代码之时，那完全就是一次可喜的体验，极具成就感。但多数情况下，还是需要测试再测试，才能完成既定任务。所以，不足之处在所难免，还敬请学界业界同仁不吝赐教，可以使本书所涵盖的知识能够得到进一步的升华。

管新潮

2018年4月21日于上海

目　录

中篇　基础性代码的组合使用

第1章 绪 论

1.1 语料库与 Python

1.1.1 语料库的若干维度

现今社会，语料库的作用已不言而喻，无论是学校教学科研还是相关的社会生产实践都离不开它。语料库规模大小不一。一般而言，社会生产实践所使用的语料库其规模要远远大于学界所使用的。如华为公司的一亿句对平行语料库，从平均字词统计看，其规模可达几十亿字词，这是学界语料库难易企及的规模。从严格意义上说，学校教学科研所使用的语料库可能都属于封闭语料库，有着较为严格的语料库边界限制，不仅是规模上的限制，还有语料内容等方面的限制。这种限制有时较为宽松，如平衡语料库，但有时却是极其严格的，如针对知识产权这一法律子体系的研究，可将语料文本对象限定为著作权法、商标法、专利法、反不正当竞争法文本。因此，为提升学习和实践应用的有效性，本书所使用的语料库对象均指封闭语料库。

语料库还有单、双语和多语之分。单语语料库仅指由单语语料文本构成的语料库，而双语语料库的语言为两种，多语语料库是指三种或以上语言构成的语料库。双语语料库可分为双语可比语料库和双语平行语料库，前者的两种语料文本之间不存在翻译关系，但两者在内容上有着极大的关联性，后者的两种语料文本之间存在翻译关系，可在语料文本之间实现句级、语块级甚至是词汇级的平行对等。双语平行语料库还有单向和双向之分，即仅有从一种语言到另一种语言的翻译关系的语料库为单向双语平行语料库，有着双向翻译关系的语料库则为双向双语平行语料库。另外，两个单语语料库也可构成单

语可比语料库，其前提是两个单语语料库均为同一种语言语料。多语语料库既可以是多个单语可比语料库，也可以是一种源语加上多个译语或者是多个双语平行语料库，等等。例如，欧洲议会平行语料库就是一个多语平行语料库，其语种涵盖欧盟各国的语言。又如，可将《中华人民共和国著作权法》英译文、《德国著作权法》英译文和《美国版权法》原文三者构成一个单语可比语料库，研究其中的大陆法系、普通法系、中国特色社会主义法系三者在著作权法/版权法方面的法律用语区别。

在互联网使用已极为广泛的当下，可以说语料库的创建已几乎无难处。按创建难易度计，单语语料库最为容易，双语平行语料库较难，最难的当属多语之间均实现平行对应的语料库或者是一对多的平行语料库。双语平行语料库的创建难易度也与文本类型有关，如文学类和法律条款类的对齐难易有时竟有天壤之别。从现有的语料库创建技术看，语料对齐是关键，而且语料对齐还有句级对齐或词汇级对齐之分。创建语料库的关键还在于为何目的创建语料库，即创建能够达成应用目的的语料库即可，所需关注的是语料的代表性、权威性、系统性。

学界对语料库的研究与应用有着高度的关注，相比之下，语料库的质量问题似乎并未引起学界的足够重视。业界虽对语料库质量有所关注，但苦于受限于技术瓶颈而无突破性进展。小规模语料库可以在人工介入之下提高质量，但大规模语料库想要实现语料库质量提升则绝无这种人工介入的可能，这呼唤着技术的进步。但语料库的质量问题也绝非仅仅是纯技术问题，问题的解决或质量的提升均需要语言学和翻译学知识融入其中。

语料库自诞生之日起就离不开相关技术工具的助力应用。现有的技术也为语料库的研究与应用提供了不少的选择，如 WordSmith 或 AntConc 等单语应用工具以及 ParaConc 等的双语应用工具，又如 Xbench 等句对质量保证工具，再如汉语的分词工具，等等。请注意，语料库技术工具的开发和应用必须考虑到语言学知识本身，未顾及语言学本体知识的技术工具不可能成为强有力的语料库工具。因技术等原因，现有的涉及语料库开发、研究与应用的相关参数或工具手段似乎无法满足人们对语料库所寄予的厚望，这又极大地阻碍了语料库研究与应用的发展。

1.1.2 语料库的技术实现

目前可供各种语料库类型使用的技术工具已有较多的选择。既然如此，为何还要引入 Python 呢？其理由有二：一是有时针对某一任务的工具过多，

最好可以组合成一种工具；二是现有的工具有时无法提供某些使用功能，无法进一步拓展语料库的应用。其实，这两种不足均可通过 Python 编程加以弥补。其一，采用 Python 的好处是避免了因不同工具对语料文本可能会产生的统计上的差异性或不一致性。其二，Python 有着近乎是无限的拓展功能，可以从语料库中挖掘到出乎意料的数据信息。

Python 的应用并不排斥既有语料库工具的使用，问题的关键是如何才能有效组合使用相关工具。为此必须考虑到不同工具之间的衔接性、数据的一致性、输出结果的可视化效果等因素。工具之间的衔接性不能受人为的熟悉程度所影响，而在于使用相关工具的内在逻辑性；数据的一致性在于经不同工具处理后的数据的语言学意义是否受到影响；可视化呈现效果会对数据的语言学解读或阐释产生直接影响。

本书旨在教会读者使用 Python 工具，故将 Python 在语料库应用中的能力习得分为三个层次：一是掌握 Python 在语料库应用中所需的基础性代码，二是将基础性代码组合应用于简单的语料库实践，三是解决语料库应用中较为复杂的实践问题。本书的适用对象包括没有编程基础的学生和语料库实践者，因此针对第一层次所涉内容均有详尽的代码解释，旨在对具体代码的语言学含义进行说明。针对第二层次所涉内容，更多是对算法设计的说明和解释，强调基础性代码的灵活应用，而第三层次所涉内容则着力于代码的创新性，意在解决新问题，可能是意想不到的问题的解决方案。

1.2 本书概要

本书共分为 3 篇共 8 章。第 1 章“绪论”自成一章，述及语料库的若干维度以及学习本书可以习得的三层次能力和其他各章的概要。其他各章分为三篇，上篇包含从第 2 章至第 5 章的内容，涉及能力习得的基础；中篇包含第 6 章和第 7 章，述及基础能力的习得；下篇为第 8 章，涵盖能力的提升以及解决方案的发现。

第 2 章为“语料文本的读取及其运行结果的输出”，围绕各种读取语料文本的不同方式和输出不同格式的运行结果而展开，这是运用 Python 进行编程实践的前后两个阶段即数据的来源和数据的结果输出。第 3 章为“语料库应用的基础性代码”，这是本书的关键基础部分，所汇总的代码均为语料库研究和应用所需的相关代码，涉及停用词的使用、文本降噪、文本语言学处理、词频

排序、检索统计等。第 4 章为“数据可视化”，旨在更为直观地提供最后的运行结果。第 5 章为“代码运行错误分析”，汇总了本书所涉相关代码运行时可能遇到的各种问题，旨在提升编写代码过程中解决问题的能力。

第 6 章为“算法、代码与编程”，对编程活动中可能涉及的规律性知识进行了总结归纳，旨在提升语料库编程实践的可实现性。第 7 章为“基础性代码的语料库组合应用”，是中篇所述各种代码的语料库实践应用，这是习得语料库编程能力的中间必经过程。

第 8 章为“Python 的语料库拓展应用”，呈现了 7 个语料库研究应用案例，每个案例均可直接服务于语料库的研究与应用。这是本书所述及的习得语料库编程能力的高阶，在习得本阶段所明确的能力后即可展开突破性的语料库编程实践。

第 7 章和第 8 章所涵盖的近 16 种编程代码均源自本书作者及其团队的语料库研究和应用实践，或是研究需要，或是应用实践，或是论文读后有感，其应用具有很强的针对性。源于研究和应用之需要的代码，其服务目标就是语料库本身，意在提升语料库的价值。

本书使用 Anaconda2 实现各项编程工作。Anaconda2 的下载网址为：https://www.anaconda.com/download。

上篇
语料文本的基础性代码