翻译教学语料库规模化建设的大数据解决方案

赵政廷，柴明颉
（上海外国语大学 高级翻译学院，上海　200083）

摘要：翻译教学中基于平行文本的双语平行语料库建设，本质上是译者根据自身需求进行方案设计、语料收集和语料管理的过程。目前受限于现有翻译技术手段和经济条件，翻译教学中难以规模化建设此类语料库。通过以大数据自动化采集的方式，引入人工智能领域的 Python 技术与 PostgreSQL 数据库系统，来解决目前建立大规模教学语料库的难题。为了完整呈现规模化建库过程，借用政府公文类翻译的真实案例，详细描述了建库方案设计、语料收集和语料管理等操作步骤，分析了各步骤中存在的自动化难题以及新技术介入的契机。部分技术细节与源代码也在文中公开，以期为大规模建设教学语料库提供一个切实可行的解决方案。

关键词：翻译教学语料库；规模化建设；Python 应用；大数据
中图分类号：TP 18; H 059　　**文献标识码：**A　　**文章编号：**1009－895X(2024)05－0403－10
DOI:10.13256/j.cnki.jusst.sse.230614322

The Scalability Challenge in Translation Corpus Building and a Big Data Solution from an Interdisciplinary Perspective

ZHAO Zhengting，CHAI Mingjiong
（*Graduate Institute of Interpretation and Translation*, *Shanghai International Studies University*, *Shanghai 200083*, *China*）

Abstract: In translation teaching, students regularly prepare “parallel texts” for a new translation project and build these linguistic materials into database, or a reference corpus. This corpus requires a relative high volume of data in order to be effective due to no ideal solution provided by available product/service-oriented translation technology. This article aims at this scalability challenge and introduces an interdisciplinary approach by combining Python and PostgreSQL. The automation challenges in processes like corpus design, data collection and corpus management are addressed by applying these technologies. The entire large-scale corpus building process is presented in detail with relevant Python and PostgreSQL source code disclosed.

Keywords：translation teaching corpus；scalability challenge；Python programming；Big Data

在笔译教学中，学生常常会遇到一些新的项目，他们对所涉及的专业领域并不熟知，仅靠其项目经验和背景知识无法应对。此时，学生需要借助外部信息源，去获取相关领域的专用表达法和背景知识。

收稿日期：2023－06－14
基金项目：上海外国语大学校级规划项目（2021114015）；上海外国语大学专项课题（2022zxkt22）
作者简介：赵政廷，男，讲师。研究方向：翻译教学、翻译技术。E-mail：z.zt@live.com

对于专业笔译而言，特定领域中已公开发布或已完成翻译的文本是非常好的参考资料，许多学者都论述过这类资料，并将其称为“平行文本”[1-4]。平行文本具有以下优点：（1）针对性强，信息精确；（2）展现相关领域的语言风格；（3）包含大量专业术语；（4）体现专业写作规范；（5）包含专业领域背景知识。

过去，这些平行文本以纸质文档和存储在磁盘上的电子文档为主，获取这些信息需要极大的工作量，而一些较难获取的信息也成为译者跨专业领域工作的障碍[4]137。在如今的大数据时代，互联网上的信息每时每刻都在呈现稳定的数量级增长，人类主体的脑力劳动平台已转向互联网，工作范式也发生了变化。许多专业机构都会在自己的网站或权威平台上发布信息，译者因此可以通过互联网渠道获取大量有价值的平行文本。

但是，这些资源都是散布在不同网页上的，来源广，内容庞杂。译者即使通过搜索引擎等工具进行关键词定位，其访问、查询效率仍然非常低，频繁的查证也将大大增加认知负荷，影响翻译效率。因此，系统性收集语料是一种合理的做法，例如在教学过程中组织学生创建一定规模的集体共享语料库[3]67。朱一凡等将这种语料库定义为翻译教学中的自建语料库（DIY corpora，或称 ad hoc / disposable corpora），“是为了完成某个翻译任务临时自建的小型专门语料库”[5]。采取双语平行语料库的形式建库，这些语料就能应用于提升翻译能力方面的数据挖掘或者转换为 TMX 格式，直接运用于计算机辅助翻译[6]142。在翻译教学中使用双语平行语料库有以下优点：（1）有助于学生发现语言使用的复杂性和句子的型式；（2）学生审读真实篇章中的例证，比依赖语法书和教科书中那些孤零零的范例更可取；（3）对应语料库可以用作专家系统，把学习者的注意力转向成熟译者或专家译者所发现的典型问题的典型（或非典型）处理方法[7]。

一、翻译教学语料库建库的规模化难题

在 Wynne 提出的语料库建设流程中，语料库架构设计、硬件部署、专职人员雇佣、文本获取、文本标注、日常维护是建库的核心内容[8]，也是日常建设语料库过程中需要解决的几大核心问题。对于在翻译教学中自建语料库的工作来说，通常困扰人们的不是复杂的架构或繁琐的硬件配置，也不是专员雇佣或复杂的标注工作，而是如何规模化获取语料。语料库的建库，如果能利用大数据时代的信息优势，让语料形成规模效应，译者就能在语料库中挖掘到大量有价值的翻译规律。正如有学者指出，在可能的情况下，建库应最大量地采集高质量语料。一些企业（如华为）在垂直领域拥有一定规模以上的双语平行语料之后，就专注于开发相应的机器翻译[6]73。尽管已发表的语料库创建研究文献中未见库容规模的说明，但从莎士比亚戏剧英汉平行语料库（300 万字词）、两岸三地英汉科普历时平行语料库（2 200 万字词）、英汉医学平行语料库（1 000 万字词）、中国法律法规汉英平行语料库（2 200 万字词）、欧洲议会平行语料库（6 000 万词）、华为汉英平行语料库（1 亿句对）等数据库的库容规模来看，百万级字词是“大数据”挖掘的前提。同时，这些大型语料库建设需要技术支持，无法单纯依靠人工。

就目前教学中可供使用的翻译技术而言，如何规模化建设双语平行语料库是一大难题。目前的技术工具使用范式主要以软件产品的操作为主，而企业研发这类软件产品需要一定周期，很难回应一些新颖和前沿的用户需求，对于目前的大数据收集也无有效的解决方案。Quah 是较早触及语料库建设规模化难题的学者，她指出，如果要调动人力、物力建立大规模的语料库，会让译者花过多精力在手动收集劳动上，性价比较低。她还指出，尽管当时已有自动化采集互联网语料的想法，但仍处在研究阶段，难以指导实践[9]108。笔者在带领学生操作一些涉及外交内容的项目时，也曾尝试手工收集外交类双语文件，并用对齐软件生成 TMX 文件。在实践中，由于需要启动、切换不同的工具，同时耗费大量时间进行对齐操作，一个 5 人组成的语料团队每小时只能建立约 2 万字的语料库。尽管译者花费了不少精力建库，但库容规模仍然较小，往往查证不到有用信息，语料库的参考价值也会受影响。

随着人工智能、大数据技术的发展，大库容语料库建设势在必行。近年来，部分研究者开始利用自动化工具建立规模化的语料库，涉及语种不但包括英语，也包括意大利语[10]195、阿拉伯语[11] 等。在这些建库方案中，不少人借助了 Baroni 等开发的 BootCaT 工具[12]。正如上文所说，此类软件工具存在一定的学习成本，灵活度较低。另一些学者开始将 IT 领域的技术运用到语料库建设中，如：著名

的美国当代英语语料库（COCA）的建设过程中，作者 Davies 利用 VB.NET 结合关系数据库技术[13]；王华树主编的《翻译技术教程》曾于 2017 年展示过一项科技公司以 Python + Scrapy 框架完成的翻译语料自动化收集案例[14]。然而，这些技术解决方案较难复制，原因有二：一是经济方面的困难，正如管新潮等指出，目前建设大型、超大型的，甚至是无限大的语料库对经济成本有很高要求，对于预算有限的教学项目来说，建库工作无法诉诸商业采购，也无法聘请 IT 开发团队[6]160；二是技术方面的困难，上述两个案例中并未公开技术细节，大多数缺乏软件开发背景的翻译专业的师生，若找不到能够帮助自己跨过技术门槛的路线图或者详细的技术说明，就很难在实践中运用这些技术。

由上述分析可知，双语平行语料库建设的规模化问题是大数据技术发展阶段译者不断增长的需求与现有工具生产力之间的矛盾。要解决这种矛盾，就必须引入新的“生产工具”，并合理安排和应用这些工具，以提升译者的工作效率。

二、规模化建库的大数据方案可行性

基于 Davies 及《翻译技术教程》中的技术方案，同时结合 BootCaT 的功能模块，笔者尝试提出 Python 语言配合 PostgreSQL 数据库的建库方案，以此解决目前规模化建设语料库所面临的经济和技术难题。从部署成本和学习成本两方面看，该方案具有以下优势。

（一）部署成本

从硬件层面看，200 元左右的单片机或学生手中千元左右的赛扬笔记本已能胜任数据收集任务。对于数据库建设，由于涉及查询，性能要求会略高，但由于几乎不会遇到大型服务器所要处理的高并发场景，普通笔记本仍能胜任。

从软件层面看，Python 和 PostgreSQL 都由开源软件社区支持，只要遵守开源协议，学习和使用的过程都是免费的。

（二）学习成本

Python 是人工智能领域主流的编程语言，可以通过几行简洁的代码集成各类功能丰富的扩展库（Libraries），按照用户需求开发功能，具有简洁、易学、高效、开源、资源丰富、扩展性强六大优点[15]52。PostgreSQL 是一种功能强大的对象——关系型数据库系统（ORDBMS），是先进的数据库系统之一[16]5。它于 1986 年由加州大学伯克利分校开发，得到广泛的开源软件社区支持，最新的版本功能多样，且运行效率较高，可以在 1 秒钟内处理上百万条数据。而在同样的硬件环境下，将这些数据写在 Word、Excel 或 txt 文档中，操作则耗时极长。

对于教学中的规模化建库难题，Python 能够让语料采集、语料清洗和入库工作完全自动化，这样规模化的难题迎刃而解，译者也能将精力集中在翻译工作上。PostgreSQL 自身的操作命令较为简易，且可以通过 Python 中的 Psycopg 进行调用，因此可以较快掌握。而对于 Python 编程来说，缺乏软件开发背景的学生可能会畏惧。然而，在大数据和人工智能技术飞速发展的今天，这样的担忧是多余的。作为一种解释性语言（interpreted programming language），Python 更像是帮助人们进行编程的软件，简单易学，学习曲线平缓，国内外很多中小学已将其作为信息技术课程的一部分。Python 语言在设计阶段已经免去了各种繁琐的操作，用户无需处理复杂的底层硬件交互，只需使用简单的命令就能实现所需功能。试对比表 1 中 C++和 Python 的代码，用户如要计算机输出一行“Hello World!”，Python 只需使用“print”命令即能完成，相比目前主流的编程语言 C++和 Java，其写法更为简易、直观。

三、大数据方案的设计思路与实现方法

建库程序主要包括语料数据的收集、清洗和管理三个环节，分别由 Python 中的 Requests，BeautifulSoup 和 Psycopg 三个库完成，整体架构见图 1。

目前互联网上的网站大都以 HTML5 的标准建立，并以 HTTP 为数据通信基础。Requests 库允许用户发送 HTTP 请求，无需进行 URL 添加查询字串、POST 数据表单编码等手工劳动。人们在分析网页的结构、定位语料位置后，就能使用这个库来请求数据。BeautifulSoup 为用户提供了多种文档对象模型（DOM）搜索方法，能够简易实现 XML 语言过滤功能，人们甚至不用正则表达式，就能从 HTML 网页中提取到文本文件[17]162。最后，Psycopg 能将这些文本中的语料数据建成 PostgreSQL 数据库，让语料库的安全性更高，访问效率也更高。

表 1 以屏幕输出“Hello World！”为例的 Python、C++、Java 代码对比

Tab. 1 Comparison of Python, C++, and Java code for “Hello World!” output

Python	C++	Java

print (“Hello, World!”)	#include <iostream>
	using namespace std;
	
	int main()
	{
	cout << “Hello, World!”;
	return 0;
	}
		public class HelloWorld {
		public static void main(String[] args){
		System.out.println(“Hello World!”);
		}
		}

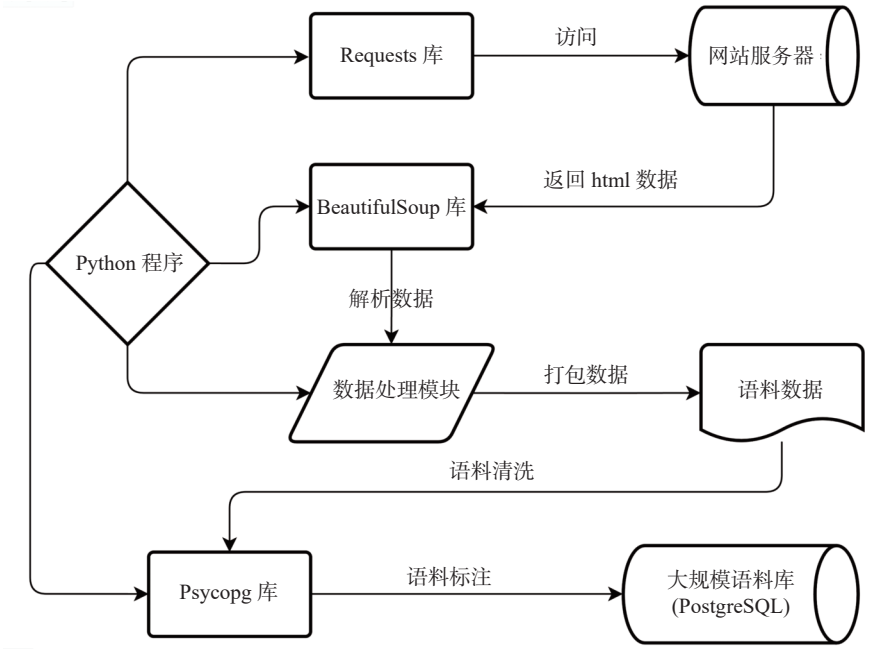


图 1 使用 Python 进行大规模在线语料收集的基本程序架构

Fig. 1 Basic architecture for large-scale online corpus collection by Python

此外，对于使用大规模收集语料的技术，我们必须遵守服务提供方的“服务条款”，例如：避免频繁访问服务提供方的服务器，以免影响服务稳定性；避免简单复制网站的数据并进行营利；避免影响网站的收入；等等。因此，在设计程序架构时，可以提前了解服务提供方有无提供应用程序接口（API）让研究人员以大数据的方式访问其网站，并优先利用 API 访问数据。在没有 API 的情况下，需要严格设定访问网站的时间间隔，限制短时间高密度的访问。同时，要对收集的数据设立管理机制，例如仅允许翻译团队在这些数据中挖掘翻译方面的规律，而不能随意复制和分发数据，以保护服务提供方的知识产权。对于涉及版权的语料，也要通过官方渠道获得授权，例如王惠等[18]在建设“英文财经报道中文翻译及注释语料库”时，即获得了《金融时报》的官方授权。

下文将以政府公文类中译英教学项目为例，分析大规模建库的过程，主要分为确立建库方案、语料采集、语料管理等流程。

（一）确立建库方案

学生在进行项目准备时，常常需要查阅平行文本。他们一般会访问外交部（fmprc.gov.cn）、商务部（mofcom.gov.cn）、中国翻译研究院（catl.org.cn）、中国网（china.com.cn）、新华网（xinhuanet.com）、中国日报（chinadaily.com.cn）等网站。这些网站的信息大都由国务院新闻办公室等国家部门发布，从术语的权威性、行文的专业性角度来说，对于新近接触项目的学生有极高的学习价值。我们不妨按照 Bowker 等[19]54 的“自建库清单”来设计语料采集方案（见表 2）。

此外，建库应采用“段落对齐”的方式进行，这主要出于三层考虑：

（1）通常采用的句段对齐语料库主要追求翻

表 2　基于 Bowker 等[19]54 的“自建语料库清单”的语料采集方案

Tab. 2　Corpus collection scheme based on Bowker et al. [19]54 “Self-built Corpus Checklist”

语料属性	详情
规模	初始30万字词，并能规模化扩展
篇目	至少20篇
翻译类型	笔译
语料主题	政府公文类
文本类型	政府公报、领导人发言、政策文件等
发布机构	国务院新闻办公室等权威机构
翻译语向	中译英
发布日期	近三年内

译实践中的既有语料的重复使用率[6]72，丢失了语料的篇章结构和段落信息；

（2）译文可能存在省略、错位等现象[19]102，采取段落对齐可以非常简单地避免这些问题出现；

（3）许多中英对照网页的对齐方式也是段落对齐，采用相应段落对齐可以简化程序结构，而且后续也可以借助 Python 加载第三方扩展包进行分句。

（二）语料采集

中国网、中国翻译研究院等网站上均能找到双语材料的栏目页面，将这些页面地址输入到 url 变量中，然后使用 Chrome 浏览器内置的网页源码检查功能，用鼠标右键查看这些语料的链接地址（见图 2）。

可以发现，这些语料的链接地址皆位于类名称为 newsList 的列表 ul 之下，且文字背后的超链接为“a href”。对于这种树形结构的 HTML 内容，可以调用 BeautifulSoup 库中的 find_all 方法来定位信息，具体实现代码如下：

```
import requests
from bs4 import BeautifulSoup
str= ‘’
url = []
for u in url:
    web = requests.get(u)
        web_soup = BeautifulSoup(web.content, ‘lxml’)
        ul = web_soup.find_all(‘ul’, class_=‘newsList’)
        for i in ul:
            for link in i(‘a’):
                links = link.get(‘href’)
                str += (links.strip() + ‘\n’)
f = open(‘links.txt’, ‘w’)
f.write(str)
f.close()
```

在上面的代码中，我们可以将需要采集语料的双语文件列表链接赋予 url 这个变量，并用 for 循环对每个需要采集的页面进行迭代。然后，调用 BeautifulSoup 中的 find_all 和 get 方法，将所有包含语料链接的“a href”通过另一个 for 循环赋予变量 str，并写入 links.txt 这个文本文件中，供后续模块调用。

接下来，我们对需要采集语料的网页进行采样分析。图 3 显示了 Chrome 浏览器中待采集网页的源码信息。

可以看到，中英对照的网页是通过 HTML 语言中 tr 与 td 这样的表格结构排列，而具体的语料数据全部位于 td 之中。因此，可以继续使用 BeautifulSoup 中的 find_all 方法，来定位这些数据，而

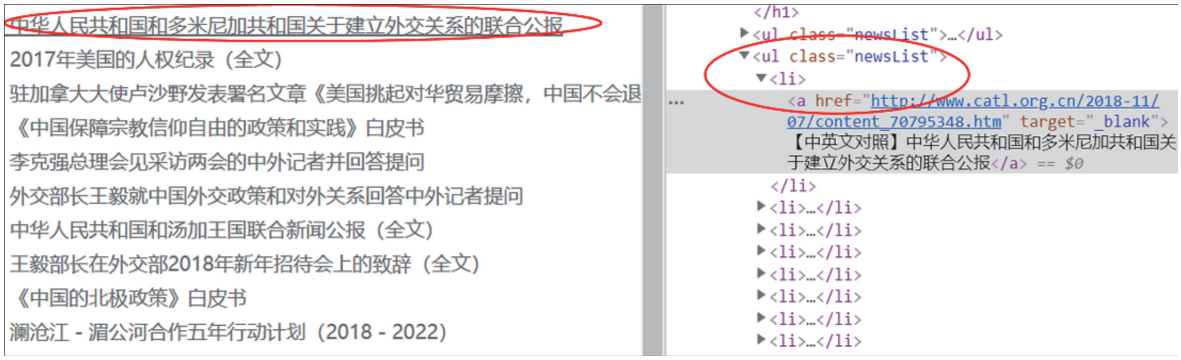


图 2　使用 Chrome 浏览器检查网页中包含语料链接的元素

Fig. 2　Inspecting elements containing corpus links with Chrome

图 3　使用 Chrome 浏览器检查待采集语料网页的元素

Fig. 3　Inspecting webpage elements for corpus collection with Chrome

其他诸如标题栏、导航栏、广告位、友情链接等信息则会被过滤。具体代码如下：

```
with open('links.txt', 'r') as f1:
    url = f1.readlines()
for links in url:
    web_r = requests.get(links.strip())
    list1 = []
    web_soup = BeautifulSoup(web_r.content, 'lxml')
    for link in web_soup.find_all('td'):
        text = link.text
        list1.append(text)
```

首先打开前面的 txt 文件，使用 readlines 方法将文件内容转化为列表变量 url，然后通过 for 循环让程序根据每一条链接地址进行访问，并将访问到的 HTML 网页传递给变量 web_soup，然后通过 BeautifulSoup 的 find_all 方法找到表格 td 中的全部语料数据，用.text 方法给 td 中提取的数据降噪，得到包含纯文本的变量 text。最后创建一个空列表 list1，把每一次迭代的 text 内容添加到 list1 的尾部。

至此，网页中语料数据都收集到了 list1 这个变量中，我们可以按照上一段代码的结尾部分将这个 list1 直接写入 txt 文档作为原始语料数据，通过 Psycopg 模块导入 PostgreSQL 数据库，限于篇幅，此处不再赘述。

前文提到，在没有 API 的情况下收集语料，要防止程序访问频率过高而增加服务器负荷，因此要设定每次收集访问的时间间隔，需要在上面的程序第一个 for 循环的下方添加这样两行代码：

```
import time
time.sleep(30)
```

这样就能通过 Python 内置的 time.sleep 方法实现访问间隔（括号内的数值单位为“秒”）。

（三）语料管理

在搭建 PostgreSQL 结构时，我们可以设计两张表，一张收录语料的信息（txt_info），另一张收录语料的具体内容（txt_content），每张表的具体内容及关系如图 4 所示。

在 txt_info 中，包括 id、序号、日期、文本类别、标题原文、标题译文、来源等信息，分别对应

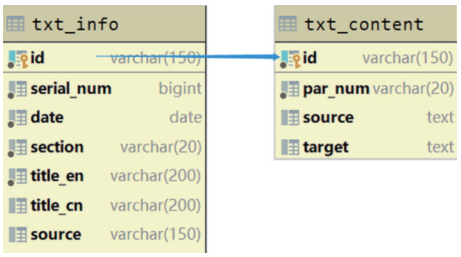


图 4　PostgreSQL 数据结构图

Fig. 4　PostgreSQL data structure diagram

varchar，bigint，date 等数据类型；而 txt_content 中，则包括 id、段落编号、原文、译文等数据类型，其中 id 是一个外键，它与 txt_info 数据表中的 id 形成对应关系。在进行数据请求时，可以根据相应的文本信息访问到具体的语料数据。我们可以在 Python 中调用 Psycopg 模块，创建出图 4 中的数据库结构，代码如下：

```
import psycopg2
cur = conn.cursor()
cur.execute(
        ‘CREATE TABLE txt_info (’
        ‘id varchar (150) not null primary key,’
        ‘serial_num bigserial not null,’
        ‘date date not null,’
        ‘section varchar(20) not null,’
        ‘title_orgin varchar(200) not null,’
        ‘title_trans varchar (200),’
        ‘source varchar(150)’
        ‘)’
    )
```

此外，我们可以用相同的方法，对数据库进行批量数据写入，代码如下：

```
conn = psycopg2.connect(host=“127.0.0.1”,user=“user”,password=“password”,database=“database”)
cur=conn.cursor()
sql= ‘insert into txt_info(id, date, section, title_origin, title_trans,source) values (%s, %s, %s, %s, %s, %s)’
cur.executemany(sql, data_str)
conn.commit()
conn.close()
```

这段代码中，host 是默认本机地址，可根据数据库位置进行调整；user 和 password 字段中应填入相应的数据库用户名和密码，database 则是数据库名称。下方的 data_str，是以批量方式插入到数据库中的变量，一般可以设计成长度为 6 的列表，对应上方 insert 命令中的格式，从前一节的变量 list1 中读取数据。插入完成以后，可以用 SELECT 命令查看表中的数据（见图 5）：

```
SELECT * FROM txt_info
```

可以看到，与语料相关的信息都井然有序地插入到了相应的位置，库容结构一目了然。同样，可以使用相同的命令查看语料内容表中的数据（图 6）：

```
SELECT * FROM txt_content
```

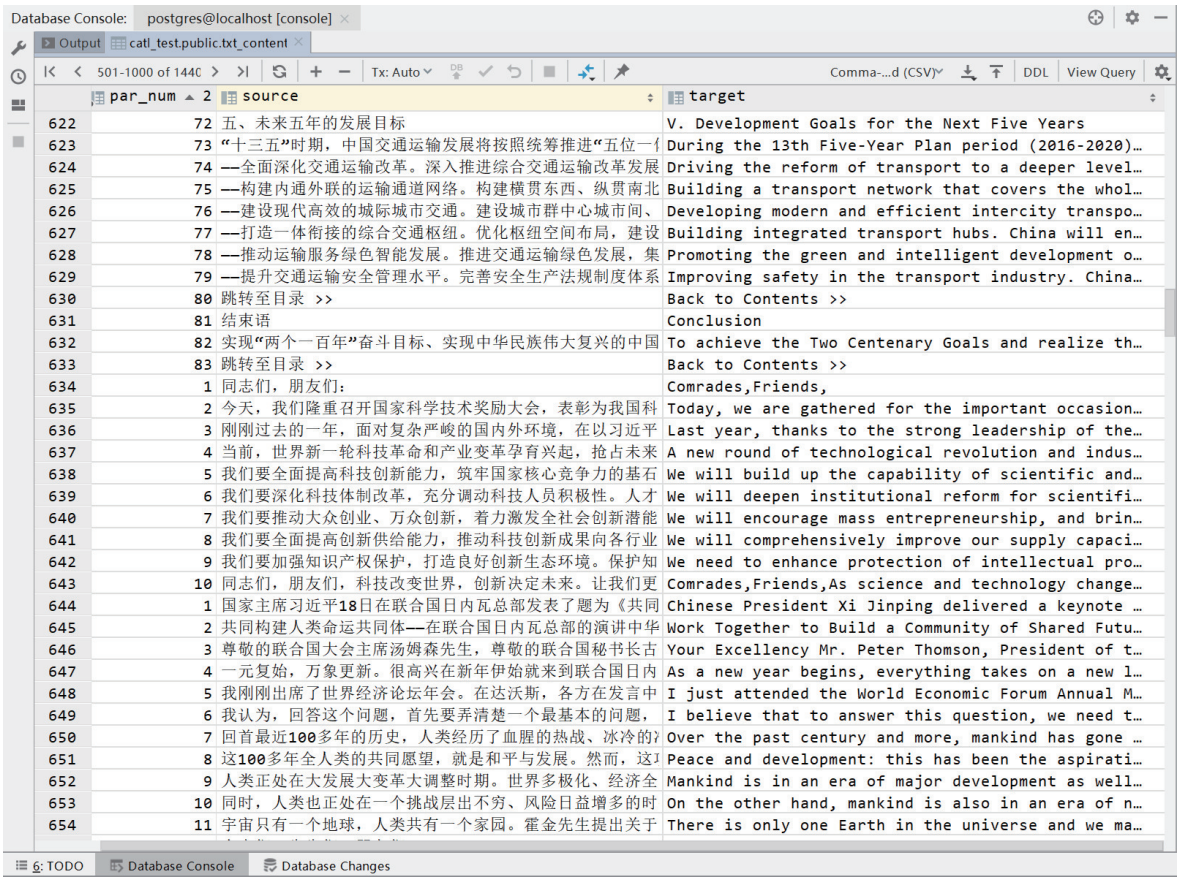
在图 6 表中可以看到，尽管语料是按照序列号从小到大排列的，但数据 par_num（段落编号）中的信息能够完整还原原文的篇章结构，最大程度地保留语料的信息。至此，双语平行语料库的建设基本完成。

Database Console: postgres@localhost [console]
Output | catl_test.public.txt_info
20 rows | Tx: Auto | Comma-...d (CSV) | DDL | View Query

	date	section	title_cn	title_en	source
1	2015-12-31	政府公文	国家主席习近平发表二〇一六年新年贺词	President Xi Jinping’s New Year Message (2016)	中国网
2	2016-03-05	政府公文	关于2015年中央和地方预算执行情况与2016年中央…	Report on the Execution of the Central and Local…	中国网
3	2016-03-05	政府公文	李克强2016年政府工作报告	Premier Li Keqiang Delivers the Annual Governmen…	中国翻译研究院
4	2016-06-21	政府公文	习近平在乌兹别克斯坦媒体发表署名文章《谱写中乌…	Chinese President's Signed Article on Uzbek News…	中国网
5	2016-09-04	政府公文	二十国集团领导人杭州峰会公报	G20 Leaders' Communique, Hangzhou Summit	中国网
6	2016-10-17	政府公文	李克强在中国－葡语国家经贸合作论坛第五届部长级…	Li Keqiang's speech at Forum for Economic and Tr…	中国网
7	2016-11-17	政府公文	习近平在秘鲁媒体发表署名文章《共圆百年发展梦	Chinese president's signed article in Peruvian n…	中国网
8	2016-12-27	政府公文	《2016中国的航天》白皮书	China's Space Activities in 2016	国务院新闻办公室
9	2016-12-29	政府公文	《中国交通运输发展》白皮书	Development of China's Transport	国务院新闻办公室
10	2017-01-09	政府公文	李克强在国家科学技术奖励大会上的讲话	Chinese Premier Li Keqiang Addresses an Annual C…	中国网
11	2017-01-18	政府公文	共同构建人类命运共同体——在联合国日内瓦总部的演…	Work Together to Build a Community of Shared Fut…	中国翻译研究院
12	2017-03-05	政府公文	关于2016年中央和地方预算执行情况与2017年中央…	Report on the Execution of the Central and Local…	新华社
13	2017-03-08	政府公文	全国人民代表大会常务委员会工作报告	Report on the Work of the Standing Committee of …	中国网
14	2017-05-15	政府公文	“一带一路”国际合作高峰论坛成果清单	List of Deliverables of Belt and Road Forum	新华社
15	2017-07-05	政府公文	习近平在德国媒体发表署名文章《为了一个更加美好…	Chinese president's signed article on German med…	新华社
16	2017-11-06	政府公文	决胜全面建成小康社会夺取新时代中国特色社会主义…	Secure a Decisive Victory in Building a Moderate…	新华社
17	2017-12-31	政府公文	国家主席习近平发表二〇一八年新年贺词	President Xi Jinping’s New Year Message (2018)	中国网
18	2018-06-10	政府公文	上海合作组织成员国元首理事会青岛宣言	Qingdao Declaration of the Council of Heads of S…	中国日报网
19	2018-06-28	政府公文	《中国与世界贸易组织》白皮书	China and the World Trade Organization	国务院新闻办公室
20	2018-07-16	政府公文	第二十次中国欧盟领导人会晤联合声明	Joint statement of the 20th China-EU Summit	中国网

图 5　入库后的语料信息表 txt_info

Fig. 5　SQL table txt_info containing corpus information



Database Console: postgres@localhost [console]

Output | catl_test.public.txt_content

501-1000 of 1440 | Tx: Auto | Comma-...d (CSV) | DDL | View Query

	par_num ▲ 2	source	target
622	72	五、未来五年的发展目标	V. Development Goals for the Next Five Years
623	73	“十三五”时期，中国交通运输发展将按照统筹推进“五位一	During the 13th Five-Year Plan period (2016-2020)…
624	74	——全面深化交通运输改革。深入推进综合交通运输改革发展	Driving the reform of transport to a deeper level…
625	75	——构建内通外联的运输通道网络。构建横贯东西、纵贯南北	Building a transport network that covers the whol…
626	76	——建设现代高效的城际城市交通。建设城市群中心城市间、	Developing modern and efficient intercity transpo…
627	77	——打造一体衔接的综合交通枢纽。优化枢纽空间布局，建设	Building integrated transport hubs. China will en…
628	78	——推动运输服务绿色智能发展。推进交通运输绿色发展，集	Promoting the green and intelligent development o…
629	79	——提升交通运输安全管理水平。完善安全生产法规制度体系	Improving safety in the transport industry. China…
630	80	跳转至目录 >>	Back to Contents >>
631	81	结束语	Conclusion
632	82	实现“两个一百年”奋斗目标、实现中华民族伟大复兴的中国	To achieve the Two Centenary Goals and realize th…
633	83	跳转至目录 >>	Back to Contents >>
634	1	同志们，朋友们：	Comrades,Friends,
635	2	今天，我们隆重召开国家科学技术奖励大会，表彰为我国科	Today, we are gathered for the important occasion…
636	3	刚刚过去的一年，面对复杂严峻的国内外环境，在以习近平	Last year, thanks to the strong leadership of the…
637	4	当前，世界新一轮科技革命和产业变革孕育兴起，抢占未来	A new round of technological revolution and indus…
638	5	我们要全面提高科技创新能力，筑牢国家核心竞争力的基石	We will build up the capability of scientific and…
639	6	我们要深化科技体制改革，充分调动科技人员积极性。人才	We will deepen institutional reform for scientifi…
640	7	我们要推动大众创业、万众创新，着力激发全社会创新潜能	We will encourage mass entrepreneurship, and brin…
641	8	我们要全面提高创新供给能力，推动科技创新成果向各行业	We will comprehensively improve our supply capaci…
642	9	我们要加强知识产权保护，打造良好创新生态环境。保护知	We need to enhance protection of intellectual pro…
643	10	同志们，朋友们，科技改变世界，创新决定未来。让我们更	Comrades,Friends,As science and technology change…
644	1	国家主席习近平18日在联合国日内瓦总部发表了题为《共同	Chinese President Xi Jinping delivered a keynote …
645	2	共同构建人类命运共同体——在联合国日内瓦总部的演讲中华	Work Together to Build a Community of Shared Futu…
646	3	尊敬的联合国大会主席汤姆森先生，尊敬的联合国秘书长古	Your Excellency Mr. Peter Thomson, President of t…
647	4	一元复始，万象更新。很高兴在新年伊始就来到联合国日内	As a new year begins, everything takes on a new l…
648	5	我刚刚出席了世界经济论坛年会。在达沃斯，各方在发言中	I just attended the World Economic Forum Annual M…
649	6	我认为，回答这个问题，首先要弄清楚一个最基本的问题，	I believe that to answer this question, we need t…
650	7	回首最近100多年的历史，人类经历了血腥的热战、冰冷的	Over the past century and more, mankind has gone …
651	8	这100多年全人类的共同愿望，就是和平与发展。然而，这	Peace and development: this has been the aspirati…
652	9	人类正处在大发展大变革大调整时期。世界多极化、经济全	Mankind is in an era of major development as well…
653	10	同时，人类也正处在一个挑战层出不穷、风险日益增多的时	On the other hand, mankind is also in an era of n…
654	11	宇宙只有一个地球，人类共有一个家园。霍金先生提出关于	There is only one Earth in the universe and we ma…

6: TODO | Database Console | Database Changes

图 6　入库后的语料内容表 txt_content

Fig. 6　SQL table txt_content containing corpus content

（四）方案成效与拓展应用

完成建库工作后，笔者利用 Python 脚本对此次语料采集实验进行了字数统计，见图 7。

```
28821 words in 【中英文对照】2016年中央和地方预算报告（全文）.txt
29417 words in 【中英文对照】2017年中央和地方预算报告（全文）.txt
9554 words in 【中英文对照】“一带一路”国际合作高峰论坛成果清单（全文）.txt
18393 words in 【中英文对照】《2016中国的航天》白皮书.txt
17768 words in 【中英文对照】《中国与世界贸易组织》白皮书.txt
16521 words in 【中英文对照】《中国交通运输发展》白皮书.txt
13211 words in 【中英文对照】上海合作组织成员国元首理事会青岛宣言（全文）.txt
5033 words in 【中英文对照】习近平在乌兹别克斯坦媒体发表署名文章《谱写中乌友好新华章》.txt
4211 words in 【中英文对照】习近平在德国媒体发表署名文章《为了一个更加美好的世界》.txt
3351 words in 【中英文对照】习近平在秘鲁媒体发表署名文章《共圆百年发展梦　同谱合作新华章》.txt
10407 words in 【中英文对照】习近平重要讲话双语实录：在联合国日内瓦总部的演讲.txt
20181 words in 【中英文对照】二十国集团领导人杭州峰会公报（全文）.txt
54555 words in 【中英文对照】党的十九大报告双语全文.txt
24305 words in 【中英文对照】全国人大常委会工作报告（2017）（全文）.txt
2389 words in 【中英文对照】国家主席习近平二〇一六年新年贺词.txt
2865 words in 【中英文对照】国家主席习近平发表二〇一八年新年贺词.txt
35955 words in 【中英文对照】李克强2016年政府工作报告（全文）.txt
7464 words in 【中英文对照】李克强在中国—葡语国家经贸合作论坛开幕式上的主旨演讲（全文）.txt
3816 words in 【中英文对照】李克强在国家科学技术奖励大会上的讲话（全文）.txt
14193 words in 【中英文对照】第二十次中国欧盟领导人会晤联合声明.txt

Total word count: 322410
```

图 7　经过 Python 收集的语料字数统计

Fig. 7　Collected corpus word count information

通过以上方法，程序在几分钟的时间内收集了32万余字的中英双语平行语料，效率远高于人工收集。当然，这只是一次小规模的建库实验。由于原先依靠手工的语料采集、清洗和入库工作彻底自动

化，我们完全可以继续利用 Python 强大的扩展特性，将上文中已编写完的功能模块如同乐高积木一样排列组合，例如，在程序头部加入批量导入网络资源的功能，那么这个程序收集语料的规模将能够扩展至千万甚至数亿字数。对于掌握这项技术的学生来说，他们只需要在建库前的设计工作上花费一些时间。例如，按照图 2 和图 3 的分析步骤定位需要获取的语料，并在图片下方相应的代码中进行替换，就能完成任何互联网上的语料收集任务。

对于语料的后续应用，有多种方式，如批量导出 TMX 并使用 TMX Editor 进行查证（见图 8），而以 PostgreSQL 为技术后台的语料库进一步高效、轻松驾驭大型语料库。当数据库增加到百万甚至千万的数量级时，我们仍然可以对其进行高效的访问和便捷的管理，不会出现传统文本文件或表格文件操作性能急转直下的情况。我们可以通过 Python 内置的其他模块如 Django 或 Flask，建设一个简易的查询界面，供学生在项目实践时查证。

图 8　语料导出为 TMX 后利用 TMX Editor 进行查证

Fig. 8　Corpus term search task demo by TMX editor

本案例借助大数据解决方案，完成了翻译项目中的自建语料库搭建工作，并解决了建库的规模化难题。对于小规模翻译团队的实践而言，实现这种规模化升级并非不可能，只要我们懂得使用何种工具来解决问题。从技术上说，无论是翻译专业的学生还是职业译者，掌握这种技术并不存在太大的障碍。正如管新潮在《语料库与 Python 应用》一书开头提到，Python 编程并非理工专业的专利，文科背景人士也能掌握[20]2。他同时还总结了 Python 语料库能力三层次：掌握基础性代码；简单语料库实践；以创新方式解决复杂问题。对于本文中的大数据建库工作，其应用难度大概位于第二层次。只要学生或译员有过简单的 Python 语料库实践经历，并对翻译任务的领域特点有所把握，就能顺利实现规模化的建库工作，给翻译团队带来极具价值的大数据语料。

四、结语

翻译职业教育的人才培养模式是否能够跟得上互联网、大数据时代的步伐，关乎整个语言服务行业的生存与发展[21]。本文的案例反映了人工智能领域的技术与翻译技术进行融合，用以解决教学语料库规模化建设的可行性难题。笔者通过在翻译教学

过程中引入人工智能领域的Python技术和PostgreSQL数据库系统，帮助学生利用大数据时代的数据优势，寻找和发现有利于翻译工作的语言规律。同时，这也为解决双语平行语料库建设的规模化难题开辟了一条新的途径。Python和PostgreSQL作为人工智能技术引入教学，能够帮助学生改变以操作某种既定的软件产品为解决问题的思路，发挥主观能动性，真正享受到技术工具带来的高效和便捷。这相当于一种“元技术手段”，在工具的设计、编程、开发阶段就把翻译工作者的需求结合进去，让翻译技术工具更有针对性和通用性。基于这种技术建设起来的大规模翻译教学语料库，可以成为开启语料库应用“大数据之门”的钥匙。通过这些语料库，教师能够提升翻译教学的效果，做到真正以真实语言规律为参照，并为国家培养符合现代化建设需求的语言服务人才。

参考文献：

[1] WILSS W. Knowledge and Skills in Translator Behavior[M]. Amsterdam: John Benjamins Publishing Company, 1996.

[2] KUSSMAUL P. Training the Translator[M]. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1995.

[3] COLINA S. Translation Teaching: From Research to the Classroom[M]. New York/San Francisco: McGraw-Hill Humanities, 2003.

[4] GILE D. Basic Concepts and Models for Interpreter and Translator Training[M]. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2009.

[5] 朱一凡, 王金波. 语料库翻译教学的理论与实践 [J]. 北方工业大学学报, 2015, 27（4）：65－69.

[6] 管新潮, 陶友兰. 语料库与翻译 [M]. 上海: 复旦大学出版社, 2017.

[7] 王克非, 秦洪武. 论平行语料库在翻译教学中的应用 [J]. 外语教学与研究, 2015, 47（5）：763－772.

[8] WYNNE M. Building corpora[EB/OL]. (2008-01-24)[2019-08-10] https://mailman.uib.no/public/corpora/2008-January/005805.html.

[9] QUAH C K. Translation studies and translation technology[M]//QUAH C K. Translation and Technology. London: Palgrave Macmillan, 2006.

[10] ZANETTIN F. Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies[M]. London: Routledge, 2014.

[11] ATWELL E. Using the web to model modern and Qur'anic Arabic[C]//MCENERY T, HARDIE A, YOUNIS N. Arabic Corpus Linguistics. Edinburgh: Edinburgh University Press, 2019: 100－119.

[12] BARONI M, BERNARDINI S. BootCaT: bootstrapping corpora and terms from the web[C]//Proceedings of the Fourth International Conference on Language Resources and Evaluation. Lisbon: ELRA, 2004: 1313-1316.

[13] DAVIES M. The 385+ million word *Corpus of Contemporary American English* (1990－2008+): design, architecture, and linguistic insights[J]. International Journal of Corpus Linguistics, 2009, 14（2）：159－190.

[14] 王华树. 翻译技术教程 [M]. 上海: 上海外语音像出版社, 2017.

[15] SARKAR D. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing[M]. 2nd ed. Bangalore: Apress, 2019.

[16] SCHONIG H. Mastering PostgreSQL 11: Expert Techniques to Build Scalable, Reliable, and Fault-Tolerant Database Applications[M]. 2nd ed. Birmingham: Packt Publishing, 2018.

[17] DALE K. Data Visualization with Python and JavaScript[M]. Beijing: O'Reilly Media, Inc. , 2016.

[18] 王惠, 朱纯深. 翻译教学语料库的标注及应用——“英文财经报道中文翻译及注释语料库”介绍 [J]. 外语教学与研究, 2012, 44（2）：246－255.

[19] BOWKER L, PEARSON J. Working with Specialized Language: A Practical Guide to Using Corpora[M]. London: Routledge, 2002.

[20] 管新潮. 语料库与Python应用 [M]. 上海: 上海交通大学出版社, 2018.

[21] 柴明颎. 互联网大数据的语言服务——从AlphaGo说起 [J]. 东方翻译, 2016（3）：4－9.

（责编：朱渭波）