

# 焱炎英汉平行语料库的创建

北京外国语大学 徐秀玲 许家金

焱炎英汉平行语料库 (Yiyan English-Chinese Parallel Corpus, 以下简称焱炎语料库) 是按布朗语料库 (Brown Corpus) 模式创建的百万词级的平衡英汉平行语料库。焱炎语料库由北京外国语大学许家金教授统筹设计。语料的采集、整理、对齐工作主要由徐秀玲博士、何煜婷老师、聂平俊老师、陈天歆同学、杜非凡同学完成。在语料库的收集整理过程中得到熊文新教授、刘燕博士等人的大力协助。

焱炎语料库按照布朗语料库的建库模式创建, 含新闻、通用、学术、小说4种体裁, 并可细分为15个子类。该库共包含500对英汉平行文本<sup>1</sup>, 每对文本包含约2,000词的英语原文及其对应的汉语译文。该库总规模约260万字词, 其中英语原文1,005,249词, 汉语译文1,625,701字<sup>2</sup>。各子类所收文本情况详见表1。

表1 焱炎语料库构成情况

体裁类型	子体裁代码	子体裁类型	文本数量	英语原文词数	汉语译文字数
新闻	A	新闻报道	44	88,284	149,388
	B	社论	27	54,181	91,713
	C	报刊评论	17	34,022	64,173
通用	D	宗教	17	34,038	52,861
	E	日常技艺及消遣爱好	36	71,894	111,100
	F	通俗读物	48	96,089	158,599
	G	传记、回忆录等	75	150,309	245,582
	H	政府或机构公文及文宣	30	60,721	92,407
学术	J	学术	80	160,984	260,663

(待续)

(续表)

体裁 类型	子体裁 代码	子体裁 类型	文本 数量	英语原文 词数	汉语译文 字数
小说	K	一般小说	50	100,739	159,330
	L	侦探小说	12	24,101	37,021
	M	科幻小说	12	24,317	38,838
	N	历险悬疑小说	13	26,443	42,394
	P	言情小说	30	60,976	94,352
	R	幽默	9	18,151	27,280
合计			500	1,005,249	1,625,701

新闻、通用、小说语料全部采集自“译言网”(http://www.yeeyan.org)的英汉双语文本。学术语料主要来自正式出版的汉语学术译著及对应的英语原著。大多数文本的产出时间为2010年前后。

燧炎语料库所有文本均通过双语对齐工具进行自动句对齐,并逐句进行人工校对。语料库分别保存为英汉语独立存储的txt纯文本格式以及tmx翻译记忆库交换格式,以方便平行语料库分析工具和计算机辅助翻译工具检索分析和利用。

该语料库取名为“燧炎通用英汉平行语料库”。从语音上,“燧炎”与“译言”谐音,表示库中所收为翻译语言;这一命名也是对“译言网”作为主要语料来源的鸣谢。从造字形态上,“燧炎”二字也暗示该语料库平行对齐的特点,以及倡导“众人拾柴火焰高”的众源翻译(crowd-sourced translation)之意。

不同于其他平行语料库多取名家名译的特点,燧炎语料库注重收集现实翻译世界中的普及型翻译语料。这有助于我们贴近翻译活动的常态。同时该库也可与经典译文库进行对照,考察资深专业译者与普通译者的翻译异同。

燧炎语料库可合可分。它既可以用于原文-译文的转换策略和对应关系研究,又可与原创汉语语料库(如ToRCH2009、ToRCH2014、ToRCH2019、LCMC、The UCLA Corpus of Written Chinese等)进行对比,考察翻译汉语的译语特征。燧炎库的英文库本身就是一个完整的布朗家族语料库英语原创库。其汉语译文也可单独成库,还可与ZCTC等汉语译文库组合使用。燧炎库中的新闻、通用、学术、小说也可析出为独立的体裁库。

### 注释

- 1 一些文本由多个短文本组合而成(如新闻报道第一个文本A01由A01A和A01B组成),短文本单独存储,因此焱炎语料库发布版本中实际包含799对文本。
- 2 用于统计英文词数的正则表达式为:[A-Za-z0-9-]+,用于统计汉语字数的正则表达式为:[\u4e00-\u9fa5][a-zA-Za-zA-Z0-9-\.\% ]+。

### 获取方式

感兴趣的读者可访问北京外国语大学语料库语言学团队网站<http://corpus.bfsu.edu.cn/info/1082/1693.htm>,免费下载该语料库。

### 声明

该语料库只可作学术研究之用,不得用于任何形式的商业活动。

**通讯地址:** 100089 北京市 北京外国语大学高级翻译学院(徐秀玲)

100089 北京市 北京外国语大学国家语言能力发展研究中心(许家金)