Modern Linguistics 现代语言学, 2022, 10(1), 1-8
Published Online January 2022 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2022.101001

Hans汉斯

三江源国家公园汉英平行语料库的建设与应用构想

杨　洋，戴延红

青海大学基础部，青海　西宁

收稿日期：2021年11月23日；录用日期：2021年12月29日；发布日期：2022年1月5日

摘　要

三江源国家公园的正式设立是党中央、国务院统筹推进“五位一体”总体布局的重大战略决策。三江源国家公园英汉平行语料库建设是响应国家政策、顺应时代发展，以青海大学为依托，构建国家公园专用双语平行语料库为目的的一次尝试，对提升三江源国家公园宣传外译质量、保护和传承三江源原生生态、打造三江源地区优质旅游品牌，促进区域产业经济绿色发展具有积极意义。三江源国家公园英汉平行语料库建设方案包括建设意义、指导依据、总体设计、语料采集、语料电子化、语料抽样、文本清洁、语料分词、元信息标注、词性标注、句级平行对齐及其保存。通过自建微型语料库的应用案例展示，拟建语料库可应用于翻译实践、语言分析、国际传播等多领域研究。

关键词

三江源国家公园，平行语料库，建设与应用

Concept for the Construction and Application of Three-River-Source National Park Chinese-English Parallel Corpus

Yang Yang, Yanhong Dai

Department of General Education and Research, Qinghai University, Xining Qinghai

Received: Nov. 23rd, 2021; accepted: Dec. 29th, 2021; published: Jan. 5th, 2022

Abstract

The official establishment of the Three-River-Source National Park represents a major strategic

文章引用: 杨洋, 戴延红. 三江源国家公园汉英平行语料库的建设与应用构想[J]. 现代语言学, 2022, 10(1): 1-8.
DOI: 10.12677/ml.2022.101001

decision made by the Party Central Committee and the State Council to promote the "five in one" layout from an overall perspective. With the support of Qinghai University, the construction of Chinese-English parallel corpus of Three-River-Source National Park marks an attempt made to construct a dedicated bilingual parallel corpus in response to the national policies and the development of the times. It is essential to improve the translation quality of Three-River-Source National Park, preserve and sustain the original ecology of Sanjiangyuan, and build a high-quality tourism brand in the Sanjiangyuan region, which plays a significant role in promoting the eco-friendly development of regional industrial economy. The scheme of constructing the Chinese-English parallel corpus of Three-River-Source National Park involves the significance of the construction ,the basis of guidance, overall design, text collection, text electronization, texts sampling, text cleaning, texts tokenization, metadata tagging, POS tagging, sentence-to-sentence alignment and the preservation. Through the application case of the self-built micro corpus, the corpus is demonstrated as applicable in various fields such as translation practice, linguistic analysis, and international dissemination.

Keywords

Three-River-Source National Park, Parallel Corpus, Construction and Application

1. 引言

三江源国家公园，地处青藏高原腹地，包括长江源、黄河源、澜沧江源及可可西里国家级自然保护区等，总面积为 12.31 万平方千米，素有“中华水塔”“亚洲水塔”之称，是亚洲重要的生态安全屏障、高原生物资源库和全球最敏感的气候启动区之一，其生态保护价值对全国乃至世界范围都具有重大意义。2021 年 10 月，三江源国家公园正式列为我国第一批国家公园，对我国加快构建以国家公园为主体的自然保护地体系具有里程碑式的意义，也标志着我国秉着绿色可持续发展，坚持人与自然和谐共生的理念，在建设生态文明和美丽中国的进程中又向前迈出了坚实的一步。如何推进三江源国家公园生态价值、经济价值、人文价值的海内外宣传，提升其国际知名度和美誉度，也应是学界高度重视的课题。

2. 语料库的定义及研究现状

拟建的语料库属于汉英双语平行专用语料库。王克非认为语料库是运用计算机技术，按照一定的语言学原则，根据特定的语言研究目的而大规模收集并储存在计算机中的真实语料，这些语料经过一定程度的标注，便于检索，可用于描述研究和实证研究[1]。管新潮认为平行语料库是指源语与译语平行对齐的语料库，可进行在词组、语块、句子、段落和语篇层面上的对齐；专用语料库是指语料仅涉及某一专门语域或语体的语料库[2]。

近年来，国内外学者在专用平行语料库的建设中做了许多尝试，例如德国学者 Kenny 负责的德英文学文本平行语料库，上海交通大学胡开宝主持建设的莎士比亚戏剧英汉平行语料库、郭鸿杰主持设计的英汉科普平行语料库，燕山大学刘泽权主持构建的《红楼梦》中英文平行语料库，香港理工大学和北京外国语大学联合建设的旅游资源双语语料库、西安外国语大学黄立波主持研制的中国现当代小说汉英平行语料库等。除依托大学等学术机构创设的大型专用平行语料库之外，小型专用平行语料库的研制也备

受关注。这些语料库通常由研究者个人自行研制。以 CNKI 为例，以“平行语料库建设”为主题，可检索到 2002 年至 2021 年已发表 357 篇文献，其中，外国语言文学(215 篇)，中国语言文学(100 篇)以及计算机软件及计算机应用(50 篇)学科的发文量最多，涉及了政治话语、科技术语、典籍外译、旅游外宣、英语教学、词典编撰、体育赛事等众多领域。可见，专用平行语料库的建设和应用在国内已蓬勃发展，受到了学界的广泛认可和肯定，为大数据时代的多领域学术研究提供了实证分析支持。但目前，基于语言学和传播学视角的三江源研究和国家公园研究凤毛麟角，三江源国家公园英汉平行语料库的建设尚属空白。

3. 三江源国家公园汉英平行语料库的建设方案

(一) 语料库建设的意义和依据

拟建的语料库是一个真实的英汉双语语料资源集成，语料库建成后可提供在线检索系统，可为语言学、国际传播、生态教育等领域的研究提供翔实范例和数据支撑，有利于提高相关文本的英文翻译效率和质量，保护和传承三江源原生生态，促进区域特色产业经济的绿色发展等，对三江源地区的政治、经济、文化、生态等方面具有积极意义。

本语料库建设主要依据中国标准研究中心编制的《建立术语语料库的一般原则和方法》和中国翻译协会发布的《语料库通用技术规范》。同时，以青海大学双一流学科建设以及语言学、生态学、环境科学、环境生态工程、计算机科学等专业人才队伍为依托，为本语料库的各子库建设提供学科分类和技术指导。

(二) 语料库建设的总体设计

本语料库建设的主要目的是建设示范性的三江源国家公园汉英平行语料库，基于该库开展语料库语言学、汉英翻译理论研究，开发三江源国家公园在线机助翻译业务，助力三江源国家公园的国际传播事业发展。基于此目的，语料库创建流程总设计如下：① 语料采集；② 语料电子化和格式统一；③ 语料抽样；④ 语料清洗和脱敏；⑤ 语料分词；⑥ 语料标注；⑦ 语料句级平行对齐；⑧ 语料保存。建成后的语料库库容可达 100 万字，语料按时间跨度建设 2015 年~2021 年七个年份字库，按学科内容建设政治、经济、人文、生态、科技五个子库。同时，建设过程中还关注语料来源的代表性和可靠性、软件工具的易操作性和兼容性、以及多人创建语料库标准的统一性等设计参数。

(三) 语料采集

语料采集以书面语料和口语语料为主，书面语料包括人工输入、扫描输入以及现有电子文本，口语语料包括音频和视频材料等的获取和转写。为保证语料库质量，要遵循语料采集的真实性、准确性、代表性、一致性和电子化五个原则，确保采集到合适的语料。如果创建语料库是应用于译学研究时，须考虑以下因素：学术价值或影响力；语料可及性和时间，当所创建的语料库应用于翻译实践时，选取的语料以多采用归化策略翻译的译文为主[3]。因此，在采集译文语料时，还需特别重视译文的质量，只有将高质量译文应用在机器翻译或记忆翻译，才能保证后续翻译实践和产出译文的准确性和规范性。其中，高质量译文是指完全转化了原文的含义、表述简洁、易于理解，符合译文语言表达习惯、在资深译员校审时无需做任何修改的文本。以采集三江源国家公园生态语料库为例，可从以下来源进行筛选(见表 1)。语料来源除新闻报道、学术文献的汉语原文及英译译文，也可是各级政府发布、权威学术机构或是知名出版社出版的音像制品，如纪录片等。

为保证语料收集的多样性和平衡性，语料按新闻、文学、科技、政论四大类体裁采样，在大类内部力求小类均衡，如在新闻大类兼顾典型报道、综合报道、述评性报道和批评性报道；在文学大类兼顾小说、诗歌、散文和戏剧。在大类内部分类建设下属子库。

Table 1. Selected sources of corpus
表 1. 可选择的语料来源

分类	举例
各级党政官方网站	中国政府网；青海省人民政府网；三江源国家公园管理局网
中国广播电视	中文国际频道；青海广播电视局
中国主流媒体网站	人民网；新华网；央视网；中国网；中国日报网；国际在线；
学术权威或专业机构	社会科学文献出版社出版的《三江源生态保护研究报告》和《青海生态文明建设报告》；国家林业和草原局国家公园管理办公室编写的《中国国家公园》；中国科学技术出版社出版的《三江源国家公园解说手册》
知名出版社	商务印书馆出版的《心随星海皈自然－三江源国家公园黄河源区环境解说》

通常，语料采集好之后，要进行电子化和格式转化处理。一般将 WORD、EXCEL、HTML、XML、PDF 文档转换为 TXT 文档，并保存为 ANSI、UTF-8 或 Unicode 等文本编码。其中，音像语料需进行录音，通过电脑文字转写和人工校对，生成高质量 TXT 文档。对于纸质图书的语料，需将其先扫描为 PDF 文档，转换为 WORD 文档后进行其他操作。对于纯文字排版、文字与图形混排以及扫描版的 PDF 文档，均可通过 ABBYY FineReader 12 软件实现 PDF 与 WORD 文档格式转换。

(四) 语料加工

采集后的语料通常要进行语料预处理、语料分词和标注、语料对齐的加工程序。

1) 语料预处理

语料预处理是指对采集的语料进行抽样，降噪，脱敏等操作，为下一步的文本处理做好前期准备。语料抽样(sampling)。在正式收录文本前，要进行语料的随机抽样或聚类抽样，以检查语料是否存在乱码、空行或嘈杂信息、无关文本混杂的标点符号或字号等。语料清洗(clean)是指消除语料中多余的字符或影响语料对齐的字符、公式、图表等，以提高语料库统计分析效用。目前语料清洗的工具种类繁多、功能强大。其中 WORD、EmEditor 软件被广泛使用。本语料库的清洗将采用 EmEditor 软件处理，它可同时打开多个文件进行相同操作，为文件数量巨大的语料加工处理提供极大便利。语料脱敏(desensitization) 是指数据中某些敏感信息通过设定规则进行数据的变形，用以保护这些敏感数据。当涉及安全数据或商业性敏感数据时，在不违反系统规则的条件下，对真实数据进行改造，如身份证号、手机号、卡号、客户名称等信息都需要进行数据脱敏。

2) 语料分词和标注

根据特定语料库建设目的，语料预处理之后得到的粗加工语料还须进行后续的精细加工。以语料分词和语料标注为主。语料分词(tokenization)，是指将一连串字符转换成相互分离、容易识别的形符的过程。在文本采集过程中，由于文本的来源和格式存在差异，须进行语料分词处理，处理后可避免检索困难、统计错误等问题[4]。本语料库建设中将采用 SegmentAnt 软件进行汉英文本的自动分词。语料标注(tagging)包括篇头信息标注和篇体信息标注。篇头信息标注是指元信息标注(metadata tagging)，是标注关于文本的信息，如文本说明(如文件序号、文本分类、版权声明)、文献信息(如作者、时间、标题、来源)、文本结构(如篇章、段落、句子)等。元信息标注可为后续语料库检索和分析提供查询条件和依据。篇体信息标注是指根据上下文信息进行语料的标记，涉及词性标注(POS tagging)、语法分析(grammatical parsing)、篇章照应标注(anaphoric annotation)和语义标注(semantic tagging)等。其中，涉及语义标注越多，标注质量就越难保证[5]。虽然人工标注能识别机器难以判断的信息，可深入到语义或语用层面，但这需要花费大量人力和时间进行校对和修正。因此，本库建设中将重点采用目前较成熟的词性自动标注方式。使用

MyTxtSegTagTool 软件进行汉语词性标注，使用 CLAWS 软件进行英文标注。这两个标注软件标注的质量准确率高，可满足一般研究要求。标注语言须考虑到通用性、简洁性和兼容性三个要素。

3) 语料对齐

语料对齐(alignment)是指在源语文本和目的语文本具体单位之间建立相互对应的关系，可分为词汇、语块、语句、段落和篇章等层面对齐。就译学语料库而言，中英文的语料对齐单位以完整的句子为主。本语料库将用于计算机辅助翻译，因此本语料库将采用语句对齐方式。可使用的对齐工具有 ParaConc 软件的 View Corpus Aligment 功能和 ABBYY Aligner 软件。需要注意的是，做对齐处理时，要遵循两个标准：一，原文与译文以句级一一对应为主，允许出现一对多或多对一的情况；二，分句一般以句号、分号、问号、感叹号为标记，要考虑句法逻辑的完整性。语料文本实现自动对齐后需进行人工校对(见图 1)。

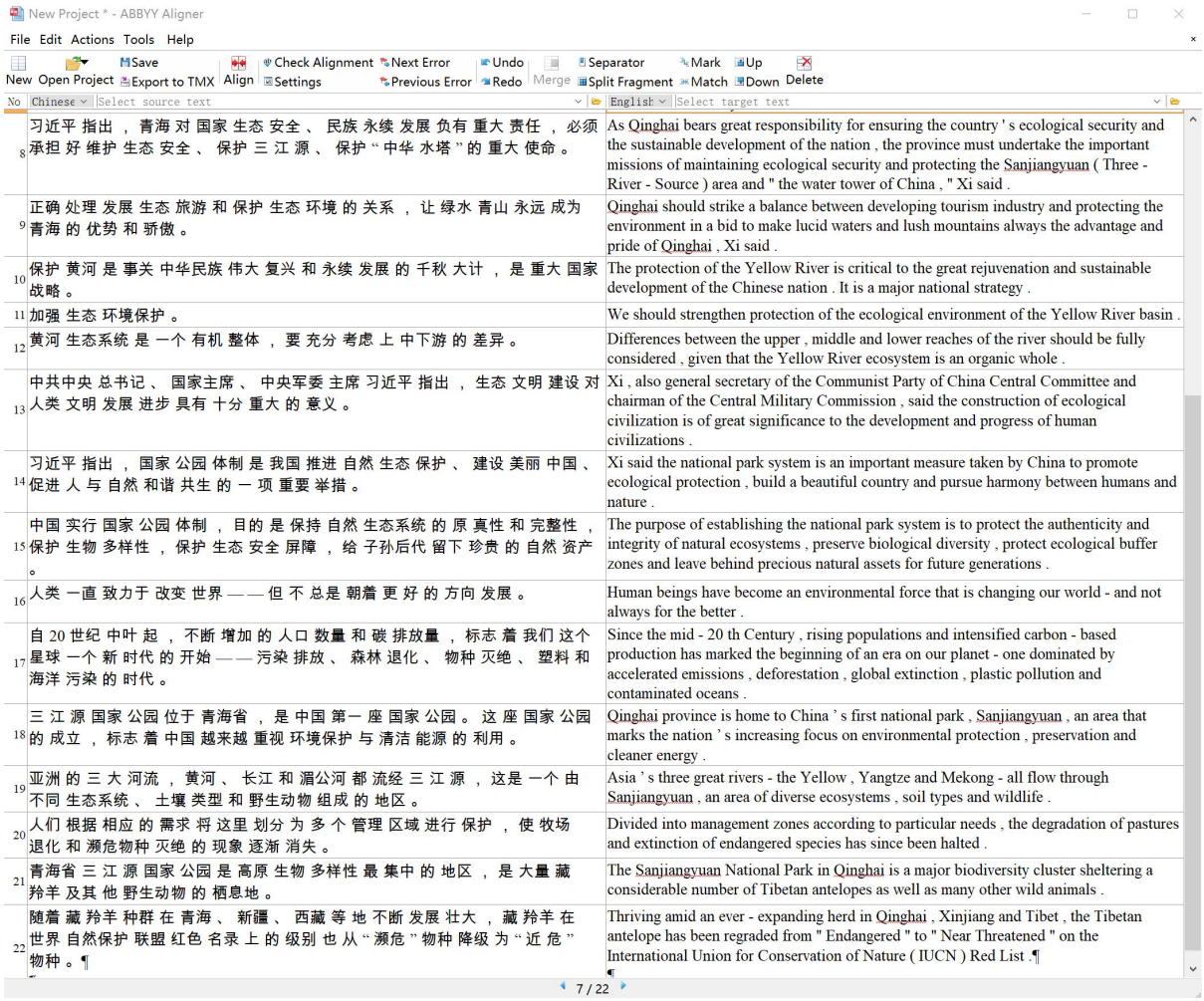

Figure 1. Sentence-to-sentence alignment in self-built micro parallel corpus
图 1. 自建微型语料库的句级对齐

(五) 语料保存

在全部语料对齐处理结束后，要对语料进行保存。本语料库将保存为 TMX 格式和 XML 可扩展标记语言。TMX 格式是计算机辅助翻译中翻译记忆库使用的文件保存格式，可将语料库直接应用于翻译检索和匹配，从而提高翻译效率。采用 XML 文件保存语料库能够反映语料的元信息、词性信息和句对齐信

息，提高查询效率，利于后期数据的挖掘分析和管理维护。

4. 三江源国家公园汉英平行语料库的预期应用

(一) 翻译相关领域应用

三江源国家公园汉英平行语料库的构建是促进翻译地方化研究的开拓性尝试。它包含了客观、详实的真实双语素材，使翻译语料变得多样化，为对外宣传研究提供了丰富的实践例句，有效推动翻译实践的深入发展。以笔者制作的中国日报三江源双语报道微型语料库为例(见图 2)，搜索“绿水青山”一词，可直接找到带有搜索词的整句翻译。也就是说，利用 ParaConc 软件，可在汉英平行语料库中直接搜索到单个词汇、短语，以及句子的一一对应翻译结果，译者只要审查通过就可以接受翻译结果，为译者进行专业术语及长难句的翻译提供了现实参考。

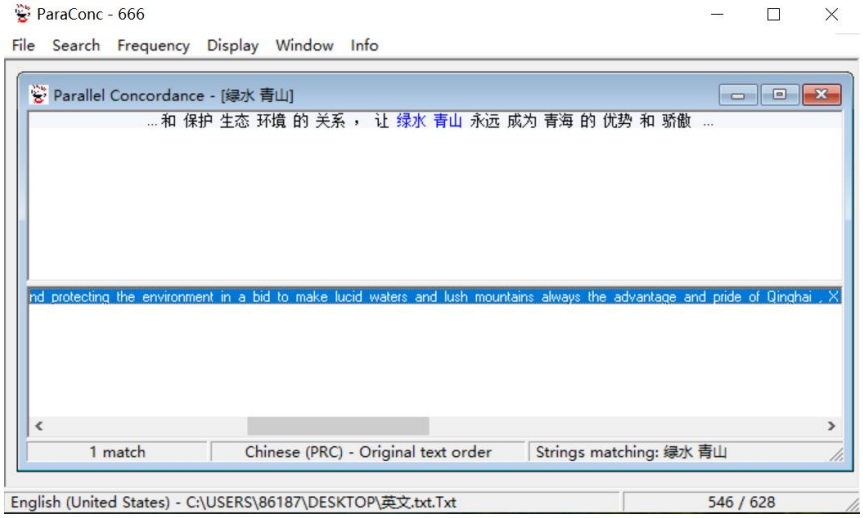

Figure 2. “Lucid waters and lush mountains” retrieval in self-built micro parallel corpus
图 2. 微型平行语料库中“绿水青山”检索演示

建设好的平行语料库可转为翻译记忆库，导入 SDL Trados、memoQ 等计算机辅助翻译系统，通过百分比匹配结果，识别翻译质量。对于关联显著的记忆库，达到 100%匹配的结果，就可视情况略去校对过程，大大节省人工校对间，极大提高了三江源国家公园宣传的外译效率和质量。

除此之外，本库也可为翻译教学、培养高质量翻译人才服务。在翻译教学过程中，师生可非常便利的获得丰富的语料和经验数据，通过语料的检索和提取，可快速了解术语、文体特征、翻译策略等，提高翻译语言的精准度和规范性。同时，教师也可引导学生利用语料库实现自主学习，鼓励学生自建小型平行语料库，为同类题材翻译提供经验参考，促进学生翻译水平提升。

(二) 语言学领域研究的应用

近年来，国内基于语料库、语料库驱动的语言研究得到了极大的发展，已成为学界进行语言实证研究的主要方式。研究者可通过语料库词频统计、词汇搭配强度分析、查询索引等工具，研究语言构建的本质特征和话语者认知方式等。例如，在笔者自建的微型语料库中，可利用 ParaConc 软件对平行语料库进行双语词频统计(见图 3)，或通过 AntConc 进行词汇搭配分析(见图 4)，通过语料库工具可研究中外媒体建构三江源国家公园传播话语方式、中外媒体的认知方式和情感态度以及中外媒体如何建构三江源国家公园形象等课题，为三江源国家公园的海内外传播事业，提升其国际知名度和美誉度提供切实可靠的

数据支持。

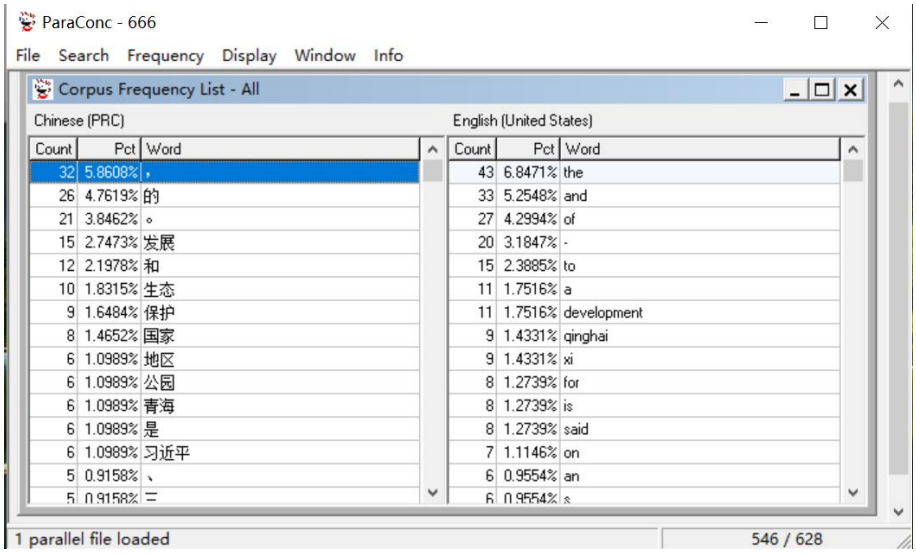

Figure 3. High-frequency words used in self-built micro parallel corpus
图 3. 微型平行语料库词频统计演示

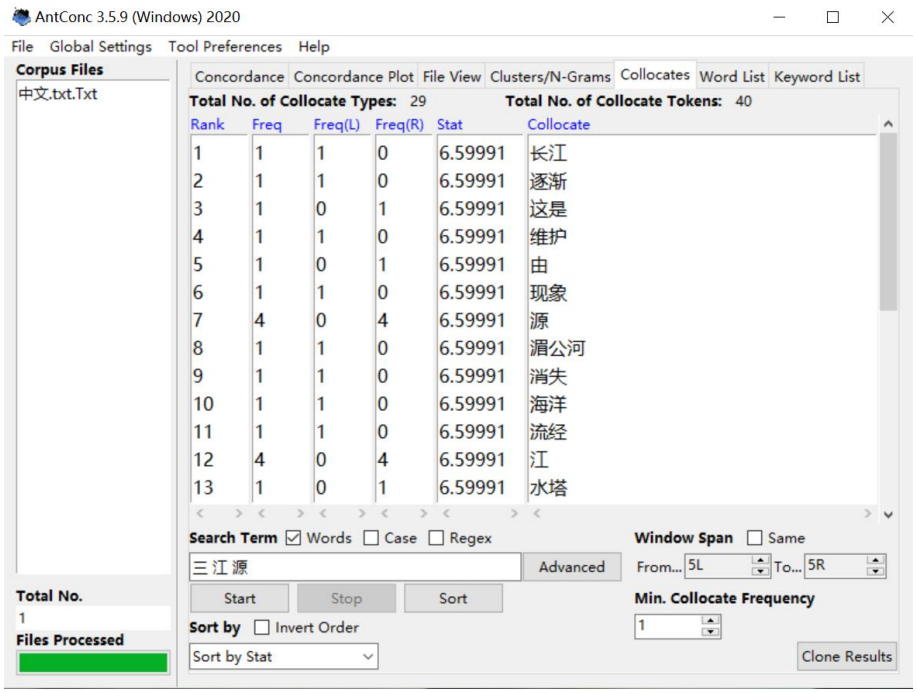

Figure 4. Collocability used in self-built micro parallel corpus
图 4. 微型平行语料库词汇搭配统计演示

5. 结语

通过语料采集、语料抽样、语料清洗和脱敏、语料分词、词性标注、句级平行对齐等环节建成的三江源国家公园汉英平行语料库，将开创以国家公园为专题的专用双语语料库建设先例，助力三江源国家

公园的对外宣传和国际传播、三江源原生生态的保护和传承，对后续三江源地区的学术研究、资源开发提供详实有力的参考范例和数据支撑。

目前，三江源国家公园汉英平行语料库建设还存在一些困难。例如，相关的中英文对照文本十分有限；本语料库的建设兼顾语料的时间跨度和平衡性因素，因此，在建库过程中需要加大双语语料的收集力度，需要投入大量的人力和精力。笔者希望能借助拟建的三江源国家公园汉英平行语料库，让更多的语言工作者投入到三江源国家公园的外译、宣传工作中，让更多的学者和科技工作者普及环保知识、保护和传承三江源原生生态，让更多民众能增强环保意识。只有坚持生态优先、绿色发展、人与自然和谐共生理念才能为区域经济的绿色发展和可持续发展做出更大的贡献。

基金项目

国家社会科学基金项目“三江源国家公园宣传教育功能研究”(18BSH085)。

参考文献

[1] 王克非. 中国英汉平行语料库的设计与研制[J]. 中国外语, 2012, 9(6): 23-27.

[2] 管新潮, 陶友兰. 语料库与翻译[M]. 上海: 复旦大学出版社, 2017: 8.

[3] 胡开宝. 语料库翻译学[M]. 上海: 上海交通大学出版社, 2011: 45-46.

[4] 梁茂成, 李文中, 许家金. 语料库应用教程[M]. 北京: 外语教学与研究出版社, 2010: 44-48.

[5] 秦洪武. 双语语料库的研制和应用[M]. 北京: 外语教学与研究出版社, 2021: 57-61.