



团 体 标 准

T/TAC 7.1—2021

中国特色话语翻译 高端语料库建设 第 1 部分：基本要求

Translations of Chinese key terms and expressions—
Corpus construction—Part 1: Basic requirements

2022-04-01 发布

2022-05-01 实施



中国翻译协会 发布

目 次

前言	III
引言	IV
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 缩略语	2
5 基本原则	2
5.1 全面性	2
5.2 准确性	2
5.3 高端性	2
5.4 时效性	2
6 语料库建设总体流程	2
7 只有中文无对应译文的建库流程	3
7.1 基本流程	3
7.2 建设准备	3
7.3 语料预处理	7
7.4 语料翻译	7
7.5 标记入库	7
7.6 语料库管理与维护	7
8 已有中文和对应译文的建库流程	7
8.1 基本流程	7
8.2 建设准备	8
8.3 文档数字化	9
8.4 转化为可编辑格式	9
8.5 文档审校	9
8.6 标记入库	9
8.7 语料库管理与维护	9
9 利用已有语料库的建库流程	10
9.1 基本流程	10
9.2 建设准备	10
9.3 语料库内容审校	11
9.4 语料库格式转换	11
9.5 修改、补充标记	11
9.6 入库	11
9.7 语料库管理与维护	12
10 语料库质量审查与评价	12

前 言

本文件按照 GB/T 1.1—2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

本文件是 T/TAC 7《中国特色话语翻译 高端语料库建设》的第 1 部分。T/TAC 7 的以下部分已发布：

- 第 1 部分：基本要求；
- 第 2 部分：系统架构；
- 第 3 部分：抽样检验。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国翻译研究院提出。

本文件由中国翻译协会归口。

本文件起草单位：中国标准化研究院、当代中国与世界研究院、传神语联网网络科技有限公司、中国翻译协会、对外经济贸易大学、中国质量标准出版传媒有限公司、北京悦尔信息技术有限公司、新译信息科技(北京)有限公司、天津工商大学。

本文件主要起草人：王海涛、于运全、杨平、范大祺、魏洁、张彤、赵庆、闫栗丽、刘强、曹馨宇、吴晓蕊、崔启亮、刘晓东、蒙永业、刘劲松、赵超。

中国特色话语翻译 高端语料库建设

第 1 部分:基本要求

1 范围

本文件规定了中国特色话语翻译高端语料库(以下简称“语料库”)建设的基本原则、语料库建设总体流程、只有中文无对应译文的建库流程、已有中文和对应译文的建库流程、利用已有语料库的建库流程、语料库质量审查与评价等。

本文件适用于中国特色话语翻译高端语料库建设、管理、维护、应用等。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB 18030 信息技术 中文编码字符集

GB/T 40035—2021 双语平行语料加工服务基本要求

T/TAC 1—2016 翻译服务 笔译服务要求

T/TAC 7.2—2021 中国特色话语翻译 高端语料库建设 第 2 部分:系统架构

T/TAC 7.3—2021 中国特色话语翻译 高端语料库建设 第 3 部分:抽样检验

ZYF 001—2018 语料库通用技术规范

ISO 639(所有部分)语种名称代码(Codes for the representation of names of languages)

ISO 3166(所有部分)国家及其地区的名称表征代码(Codes for the representation of names of countries and their subdivisions)

3 术语和定义

GB/T 40035—2021、ZYF 001—2018 界定的以及下列术语和定义适用于本文件。

3.1

语料库 corpus

由依据一定方法收集、排列的语言文本资料所构成的电子数据库

[来源:ZYF 001—2018,3.2,有修改]

3.2

中国特色话语 Chinese key terms and expressions

中国共产党和国家领导人论著、报告、讲话,以及中央人民政府的文献资料 and 各类文件的总称

3.3

语料对齐 text alignment

对齐

在源语文本和目的语文本具体单位之间建立的对应关系,可分为词汇、语块、语句、段落和篇章等层面的对齐

7 只有中文无对应译文的建库流程

7.1 基本流程

只有中文但无对应译文时,应使用必要的建库系统和软件,将翻译和建库结合进行。流程如图 1。

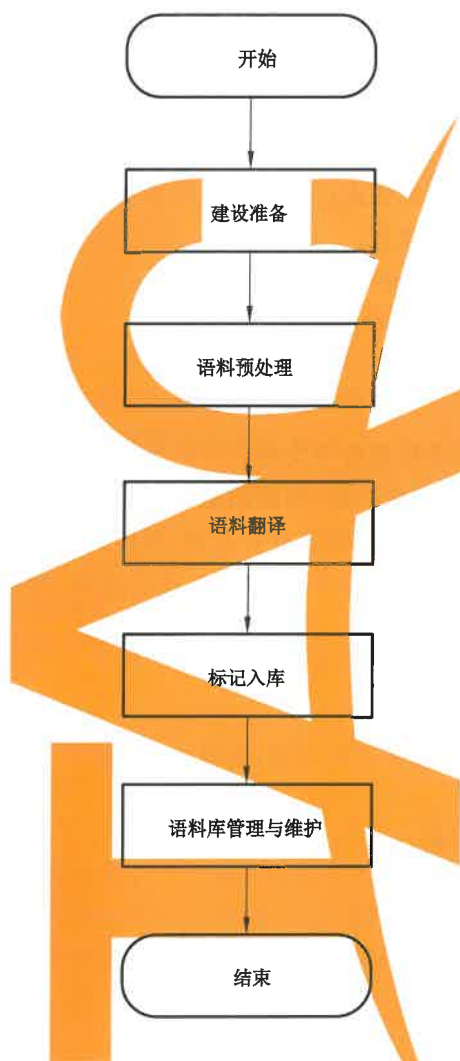


图 1 只有中文无对应英文的建库流程

当建设多语种(目标语言)语料库时,应建立多语种协调机制,确保对中文的理解保持一致。各语种翻译之间应建立沟通交流机制,以统一对重要词汇的理解和认识,避免不同语种译文之间出现偏差。当外文表达形式有分歧时,应以英译方案表达形式为准。

当对中文的理解有分歧且不能达成一致时,应由上级机构或项目委托方进行解释。

7.2 建设准备

7.2.1 总体要求

在对建库需求进行广泛和重点调查、分析的基础上,从需要和可行性两个方面提出建库方案等,由相关部门批准后,形成任务书(合同书、协议书),然后正式开始建库工作。

建库准备工作宜包括:

示例：

国务院 2019 年《政府工作报告》

d) 党和国家其他政治文献和资料；

示例：

《中国共产党历史》《总结党的历史经验 加强党的建设》

e) 党和国家领导人的论述，包括讲话、报告、题词、指示、批示、信件等。

示例：

《在庆祝中国共产党成立 100 周年大会上的讲话》

7.2.4.3 参考资料

参考资料主要是中国特色话语中外文对照资料等，包括但不限于以下类型：

- a) 含有中国特色话语文本的中外文对照资料；
- b) 相关中文资料。

7.2.4.4 术语表

术语表主要是词汇的规范译法等。术语应能进行溯源和标记。术语表包括但不限于以下类型：

- a) 中国外文局等权威机构制定或发布的规范译法术语表；
- b) 基于已有语料库和翻译项目归纳整理的术语表；
- c) 对多文档组成的大型项目，在译前阶段提取中文词汇所制作的术语表。

7.2.4.5 采集源文件

获取并录入语料库建设相关的译文、中外文资料等。

已有的可编辑电子文本的资料可直接使用，也可通过人工录入、扫描输入等方式采集或补充其他文本。

7.2.5 理解源文件并确定译文体例规范

7.2.5.1 参与中文起草

在条件允许的情况下，宜派遣有经验的审定稿人提前介入中文文本或外宣版本的起草和审定工作，以便从翻译专业化角度出发，依托自身跨文化优势，提出建设性意见及合理化建议，使得译文更加符合国外读者的阅读习惯，并帮助读者更好地理解中国特色话语文化内涵。

7.2.5.2 梳理中文

译员团队应梳理中文、把握内容框架、了解背景信息、准确理解相关概念，并邀请著作、文件或报告等的主要起草者或有关专家，针对中文文本的背景和内容，系统地梳理逻辑、讲解难点。

7.2.5.3 译前研讨

应在充分搜集近期党和国家领导人讲话、各部委权威解读等中文资料以及已有其他政治文献译文、近期海外主流媒体报道等外文资料的基础上充分研讨，把握中文源文的感情色彩，研讨确定恰当的词法句法，确定统领全篇的重要政治概念的译法和译文风格。

特别重大的政治概念应准备多种备选翻译方案，由上级机构或项目委托方进行确定。新概念、新范畴、新表述的翻译方案应得到充分论证。

7.2.5.4 确定译文体例规范

译文应明确标点、数字、表格、图示等体例，不同语种译文体例应符合该语种国家的表述习惯。

7.2.6.3 语料库结构

应根据语料库内容、建库目的、语种数量、应用需求等确定语料库组织和存储结构。语料库应结构清晰,便于扩展维护,便于检索和应用。

7.2.7 选取(或开发)建库软件

可利用语料库建设和管理系统完成语料库建设工作。语料库建设系统应具备译文录入、审校、更新、管理、检索等功能,可具备计算机辅助翻译、术语管理等功能。相关人员可利用该系统完成翻译、录入及管理工作。系统应自动记录工作流程信息。

7.3 语料预处理

在对所收集语料进行加工之前,应进行语料清洗、文献资料查重等处理。

语料清洗主要是将源文件转换为统一的编码格式,过滤各种排版符号、删除乱码,将全角字符转换为半角字符,将繁体字转换为简体字,移除与语料库建设无关的符号和内容等。

7.4 语料翻译

语料翻译是语料库建设的核心,是保障语料库质量的重要环节,应确保充足的时间分配。译员团队应具有丰富的工作经验。

语料翻译具体步骤及要求应符合附录 A 规定。

7.5 标记入库

语料源文和译文应进行对齐,对齐层级可为篇章、段落、句子、短语、词汇级别等,宜按词汇、短语、句子、段落、篇章顺序进行对齐选择。

语料标记时,应根据语料的类型选择适用的标签。语料种类可参考 7.2.4,不同种类的语料宜分类入库。

语料标记工作应通过人工、办公软件或语料库建设系统进行。语料标记应保持标签、格式统一。

语料库应按统一格式进行存储和管理。语料库格式可参考 GB/T 40035—2021。

7.6 语料库管理与维护

宜根据设计要求,安排专人使用特定的工具定期进行语料库的管理与维护,确保语料库安全和保密性。应及时调整、改进和补充相关中国特色话语的内容,并保留完整的管理、维护与更新文档。

语料库安全应符合 GB/T 40035—2021 中第 8 章的规定。

8 已有中文和对应译文的建库流程

8.1 基本流程

应将译文经数字化处理、格式转换、审核后标记入库,具体流程见图 2。

8.2.2 分析建库需求

对语料库建设要求、应用需求等相关事项进行分析,主要包括以下内容:

- a) 分析源文及译文的范围、来源、采集方式;
- b) 分析置标语言、元数据,设计语料库存储结构和格式;
- c) 分析数据处理、输入、输出的要求;
- d) 确定工作团队结构;
- e) 选取(或开发)软件及工具。

需求分析应详细、具体,并形成分析报告等文件。

按照需求分析,提出对建库系统功能、性能的要求,包括硬件要求、软件要求、质量要求等,明确待开发系统的目标。分析现有条件,提出待购置或开发软、硬件纲目表,提出对原有系统进行改造与扩充的建议方案,完成软、硬件要求说明书。语料库建设系统选择、建设应符合 T/TAC 7.2—2021 的规定。

8.2.3 制定工作计划

应按照 7.2.3 规定执行。

8.2.4 选取和收集源文和译文

选取和收集建库所需的中国特色话语文献资料源文、对应译文、参考资料、术语表等。

8.2.5 确定置标语言、元数据方案及语料库结构

置标语言的选取、元数据方案及语料库结构设计应按照 7.2.6 规定执行。

8.2.6 选取(或开发)建库软件

建库软件的选取和开发应按照 7.2.7 规定执行。

8.3 文档数字化

如果源文或译文以纸质形式存在,应首先将其通过人工录入、扫描等方式录入计算机,形成电子文档。

8.4 转换为可编辑格式

如果源文或译文为 PDF 或图片形式,应通过人工录入、OCR 识别等方式转换为可进行文字编辑的文件格式。

8.5 文档审校

应按照 7.4 的规定,由具有中级及以上资格的译员对译文内容进行全面审校。

8.6 标记入库

应按照 7.5 规定执行。

8.7 语料库管理与维护

应按照 7.6 规定执行。

- d) 确定置标语言、元数据方案及目标语料库结构；
- e) 当置标语言不同、元数据方案与已有库不同时，确定转换方案；
- f) 选取(或开发)建库软件；
- g) 其他。

9.2.2 分析建库需求

对语料库建设要求、应用需求等相关事项进行分析，主要包括以下内容：

- a) 分析已有的可用语料库的来源、获取方式；
- b) 分析置标语言、元数据，设计语料库存储结构和格式；
- c) 分析置标语言、元数据转换方案及目标语料库结构；
- d) 分析数据处理、输入、输出的要求；
- e) 确定工作团队结构；
- f) 选取(或开发)软件及工具。

需求分析应详细、具体，并形成必要的工作文件。

按照需求分析，提出对建库系统功能、性能的要求，包括硬件要求、软件要求、质量要求等，明确待开发系统的目标。分析现有条件，提出待购置或开发软、硬件纲目表，提出对原有系统进行改造与扩充的建议方案，完成软、硬件要求说明书。语料库建设系统选择、建设应符合 T/TAC 7.2—2021 的规定。

9.2.3 制定工作计划

应按照 7.2.3 规定执行。

9.2.4 选取和收集已有可用的语料库

选取和收集已有的中国特色话语翻译语料库中的数据，并分析所收集库的结构。

9.2.5 确定置标语言、元数据方案及目标语料库结构

置标语言的选取、元数据方案及目标语料库结构设计应按照 7.2.6 规定执行。

9.2.6 选取(或开发)建库软件

建库软件的选取和开发应按照 7.2.7 规定执行。

9.3 语料库内容审校

对收集的语料库，应按照 7.4 的规定，由具有中级及以上资格的译员对语料的内容进行审校。

9.4 语料库格式转换

若所收集语料库的结构、格式与当前在建语料库的要求不符，应按照当前在建语料库结构和格式要求，对收集而来的语料库结构和格式进行转换，转换为当前在建库所允许的形式。必要时，可对所收集的语料库进行简单加工处理。

9.5 修改、补充标记

按照当前在建语料库的置标语言和元数据规定，对收集而来的语料库中的数据标注进行补充、删除、修改、转换等处理，使之满足当前在建库的要求。

9.6 入库

将按照 9.5 处理后的数据收录于当前在建库中。

附录 A
(规范性)
翻译阶段要求

A.1 翻译阶段流程

应预留足够的时间以确保翻译工作质量。语料翻译阶段具体流程见图 A.1。

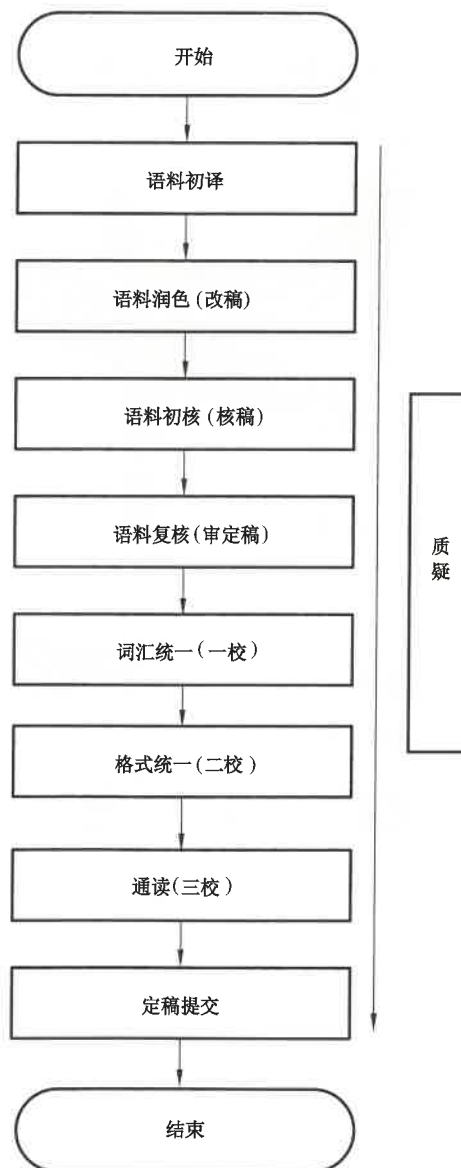


图 A.1 语料翻译流程

作应做到“首查”和“二查”：

- a) “首查”指查阅、运用以往的特色话语对外翻译资料；
- b) “二查”指对源文中出现的新概念、新范畴、新表述，要查询外文译法、凿实中文准确含义。

对初译环节的具体要求如下。

- a) 应无漏译、无常识性错误。
- b) 资料查证过程中首查已有中国话语对外翻译资料，对于之前的译法应尽量源文引用；对于同一概念范畴表述存在不同译法的，应引用时间点最近的译文。
- c) 文本中首次出现的新概念、新范畴、新表述，提交译文至润色环节时，应同时提交中文查证资料和外文查证资料。如外文中文无相关表述，应依据中文查证资料提出初步翻译方案，连同中文查证资料一并提交；如未查询到中文官方解释性文件，应提出初步翻译方案并做显著标记。

A.4 润色(改稿)

润色环节的主要目的是修改不符合外文表达的译文，减少中式外语，提高译文可读性。同时，通过译文反推、确保中文理解的准确性。润色环节应由外籍专家负责，部分工作宜有中级译员参与。

对润色环节的具体要求如下。

- a) 中级译员应按照初译标准核实初级译员提交的中外文参考资料，审读、修改初级译员提交的译文，使其符合初译标准。
- b) 初译环节提交的质疑问题，如在本阶段收到反馈，应由中级译员依据中文解答提出相应翻译方案并提交给外籍专家。
- c) 外籍专家应依据中外文查证资料，核实译文用词是否准确、精炼；如对中文理解有异议的，应提交质疑至下一环节，以便中方译员进一步查证；对于译词的使用有较大异议的，应醒目标注修改意见，并连同相关外文查证资料提交至下一环节。
- d) 外籍专家应修改译文语法，保证译文逻辑合理、行文顺畅。
- e) 外籍专家应以第一读者视角通读译文，根据语感调整段落句式，提升译文可读性。
- f) 如有必要，应进行多次润色。

A.5 初核(核稿)

初核环节的主要目的是处理外籍专家提出的质疑，并对照中文源文审读译文，排查政治性错误及常识性错误。初核环节宜由中级译员负责。译员应保持高度的政治敏感性和警惕性。

对初核环节的具体要求如下。

- a) 应针对润色环节外籍专家出的关于中文理解的质疑，进一步查证中文资料并提出修改意见。
- b) 应结合初译和润色环节提供的中外文参考资料，对照源文逐字逐句审读译文，排除政治性错误及常识性错误；应对译文表达与源文内容有出入的部分予以标记并提出修改意见；应重点排查涉党、涉外、涉宗教、涉主权、涉港澳台等内容。
- c) 应汇总、整理初核环节中修改的译词或短语，形成词汇统一表，并提交至下一环节。

A.6 复核(审定稿)

复核的主要目的是通过资料查证和组内讨论，确保对中文的准确理解，并形成一致的翻译方案。复核环节应由高级译员负责，并宜有中级译员和外籍专家参与，以确保译法的权威性和准确性。应保障该环节工作时长充足，并对译文的准确性进行确认。

对于目标语国家语言体系中并不存在的中国特色话语表述,高级译员应充分了解源文表达目的和受众认知需求,结合质疑材料、上级意见以及外籍专家修改意见,对译文进行最后检查、调整,确保译文无疏漏、无错误。

A.11 质疑

质疑是贯穿语料翻译各环节的重要工作。质疑工作包括三种方式:

- a) 将中文写作团队对新概念、新范畴、新表述的解释及时传达给翻译团队,帮助译员更加准确地理解源文,拟定恰当的翻译方案;
- b) 对于翻译团队内部意见分歧较大的翻译方案,将备选翻译方案及其中文解释一并反馈至中文写作团队,由中文主要起草者提出参考意见,协助确定最终翻译方案;
- c) 在多语种语料库建设项目中,围绕各语言组提出的质疑展开内部讨论,促进同一语系译文互相参考,并向其他语言组提示原文的逻辑关系,提升各语种译文的准确性。

质疑可由初级译员负责,应确保翻译团队提出的问题能够准确无误地传递给中文写作团队,并在获得解答后及时反馈至相关译员。

A.12 其他

对于重要著作的翻译,翻译团队还应根据出版要求,负责红样、清样等专业环节具体工作。

A.13 译后整理

整理译后资料的主要工作是制作中外文对照资料,以备其他项目使用。

若采用计算机辅助翻译技术,译后阶段的资料整理应遵循以下要求:

- a) 将项目实施过程中形成的术语添加到已有术语库中,并将中外文查证资料作为参考依据一并录入,以备其他项目使用;
- b) 将中文源文及译文终稿转换为计算机辅助软件可利用的中外文对照资料,以备其他项目适用。

中国翻译协会
团体标准
中国特色话语翻译 高端语料库建设
第1部分:基本要求
T/TAC 7.1—2021

*

中国标准出版社出版发行
北京市朝阳区和平里西街甲2号(100029)
北京市西城区三里河北街16号(100045)

网址 www.spc.net.cn

总编室:(010)68533533 发行中心:(010)51780238

读者服务部:(010)68523946

中国标准出版社秦皇岛印刷厂印刷
各地新华书店经销

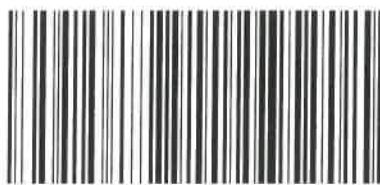
*

开本 880×1230 1/16 印张 1.75 字数 48 千字
2022年3月第一版 2022年3月第一次印刷

*

书号: 155066·5-4236 定价 42.00 元

如有印装差错 由本社发行中心调换
版权专有 侵权必究
举报电话:(010)68510107



T/TAC 7.1—2021



码上扫一扫 正版服务到