ICS 35.240.01
CCS L 72

中华人民共和国国家标准化指导性技术文件

GB/Z XXXXX—XXXX

信息技术　汉藏平行语料加工流程

Information technology—Chinese-Tibetan parallel corpus processing procedure

（本草案完成时间：2026 年 11 月 18 日，征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX-XX-XX 发布

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前言 III
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
4 缩略语 3
5 汉藏平行语料加工流程 4
6 汉藏平行语料加工前期准备要求 5
6.1 待加工数据的确定 5
6.2 汉藏平行语料加工任务 5
6.3 加工规则 6
6.4 加工需求变更 8
6.5 加工环境要求 8
7 加工任务执行过程的要求 10
7.1 加工任务创建要求 10
7.2 加工任务分发要求 10
附录A （资料性）加工人员管理要求 12
A.1 加工人员范围 12
A.2 加工人员培训 12
A.3 加工人员能力档案 12
参考文献 13

前　言

本文件为规范类指导性技术文件。

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。

本文件由全国信息技术标准化技术委员会（SAC/TC 28）提出并归口。

本文件起草单位：西藏大学、电子科技大学、中国电子技术标准化研究院、青海师范大学、北京邮电大学。

本文件主要起草人：尼玛扎西、于永斌、头旦才让、陈晓研、班玛宝、王向向、仁增多杰、仁青东主、黄姗姗、洛桑嘎登、群诺、范满平、高凡、官却才让、道吉扎西、冯箫、张子玥、李臣波、郑智文、刘羽童、李蕾。

信息技术　汉藏平行语料加工流程

1　范围

本文件规定了汉藏平行语料加工的流程、前期准备、任务执行、质量控制等要求。

本文件适用于汉语和藏语平行语料在篇章、段落、句子、短语或其他适宜粒度上的数据准备、预处理、对齐、审核、质量控制、导出、归档等相关加工活动，并应用于机器翻译、大语言模型训练数据准备、监督微调数据构建、评测数据集建设以及知识增强等自然语言处理场景。

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/T 13000—2025　信息技术 通用编码字符集（UCS）

3　术语和定义

下列术语和定义适用于本文件。

3.1

文本 text

以字符、符号、词、短语、段落、句子、表格或其他字符排列形成的数据，用于表达意义，其解释基本上取决于读者对于某种自然语言或人工语言的知识。

[来源：GB/T 40035—2021，3.1]

3.2

语料 corpus

语言材料或资料。

[来源：GB/T 40035—2021，3.2]

3.3

汉藏平行语料 Chinese-Tibetan parallel corpus

由汉语和藏语构成，并在篇章、段落、句子、短语或其他级别建立对应关系的双语语料（3.2）集合。根据翻译方向可分为汉藏平行语料和藏汉平行语料。

3.4

原文 source language text

源语言文本（3.1）。

[来源：GB/T 40035—2021，3.4]

3.5

译文 target language text

目标语言文本（3.1）。

[来源：GB/T 40035—2021，3.5]

3.6

元数据 metadata

关于数据的内容、质量、状况和其他特性的描述性数据。

[来源：GB/T 35295—2017，2.2.7]

3.7

光学字符识别 optical character recognition；OCR

采用光学识别设备对文件中的字符进行识别的过程。

3.8

翻译存储交换 translation memory exchange；TMX

翻译存储交换的标准格式。

[来源：GB/T 40035—2021，3.10，有修改]

3.9

纠正 correction

为消除已发现的不合格内容所采取的措施。

[来源：GB/T 19000—2016，3.12.3]

3.10

去标识化 de-identification

去除可确认个人或组织身份的数据与数据主体之间联系的过程。

[来源：ISO 25237:2017，3.20，有修改]

3.11

个人信息 personal information

如果公开或者滥用会造成潜在危害的信息。

[来源：GB/T 40035—2021，3.15]

3.12

汉藏平行语料加工方 Chinese-Tibetan parallel corpus processer

承担汉藏平行语料（3.3）加工任务的人员或机构。

3.13

汉藏平行语料需求方 Chinese-Tibetan parallel corpus data user

提出汉藏平行语料（3.3）加工需求的人员或机构。

3.14

汉藏平行语料加工管理方 Chinese-Tibetan parallel corpus processing administrator

管理汉藏平行语料（3.3）加工任务评估、分发、交付、验收以及质量把控的人员或机构。

3.15

加工工具 processing tool

用于执行汉藏平行语料（3.3）采集、清洗、编码转换、断句、对齐、审核、导出及质量控制等环节的软件工具或辅助系统。

3.16

加工任务说明 processing task description

汉藏平行语料需求方（3.13）提供的，用于明确加工对象、加工粒度、质量要求、交付格式、应用用途和验收指标的正式文档。

注：平行语料加工任务说明通常包括对所需加工任务的描述、加工流程与方法、示例格式、质量验收方法及验收指标等内容。

3.17

汉藏双向句对 Chinese-Tibetan bidirectional sentence pair

汉语文本（3.1）与藏语文本之间建立句子级语义对应关系，可实现汉藏和藏汉两个方向使用的平行文本二元组。

3.18

藏文断句单元 Tibetan sentence segmentation unit

在藏文文本（3.1）处理中，依据语言规则、标点形式或加工规则划分得到的最小句子级处理单元。

3.19

藏文编码归一化 Tibetan encoding normalization

按照统一字符编码和表示规则，对藏文文本（3.1）中的字符表示形式、编码形式或兼容写法进行规范化处理的过程。

3.20

多粒度对齐 multi-granularity alignment

在篇章、段落、句子、短语或其他不同层级上建立汉语与藏语对应关系的加工方式。

3.21

术语一致性 terminology consistency

在汉藏平行语料（3.3）加工过程中，同一术语或固定表达在相同或相近语境下保持稳定对应关系的属性。

3.22

训练样本 training sample

用于机器翻译、大语言模型训练或监督微调的数据单元，可由文本（3.1）、标签、元数据（3.6）及其结构化表示组成。

3.23

评测样本 evaluation sample

用于模型性能评估、质量验证或基准测试的数据单元。

3.24

集合隔离 dataset split isolation

对训练集、验证集、测试集或评测集进行独立划分并防止交叉重叠的数据组织要求。

3.25

模型适配格式 model-adapted format

适用于机器翻译系统、大语言模型训练、微调、评测或知识增强应用要求的数据组织与导出格式。

4 缩略语

下列缩略语适用于本文件。

JSON：JavaScript 对象表示法（JavaScript Object Notation）

OCR：光学字符识别（Optical Character Recognition）

TMX：翻译存储交换（Translation Memory Exchange）

TXT：文本文件（Text File）

UTF-8：8位通用字符集/Unicode转换格式（Universal Character Set/Unicode Transformation Format-8bit）

5 汉藏平行语料加工流程

汉藏平行语料的加工流程包括数据准备、预处理、语料对齐、语料审核和结果导出与归档五个环节，见图1。数据准备环节主要包括待加工数据整理、加工任务确认、语种方向确认、元数据范围确定及数据安全检查等内容，用于为后续加工提供基础条件。预处理环节主要包括格式检查、清洗、藏文编码归一化、去重、去标识化、长度过滤与截断、模式过滤、语种方向校验、元数据检查和集合隔离与数据污染检查等处理步骤，用于保障语料的规范性和基础质量。语料对齐环节包括汉藏双语语料导入、藏文断句、长句切分与结构调整、多粒度对齐、汉藏双向句对校核、术语一致性校验和人工修正等步骤，通常通过语料对齐工具结合人工辅助完成，用于保障语义对齐的准确性和结构的一致性。语料审核环节包括抽样检查、人工审核和结果记录，以实现对整体语料加工质量的有效控制。结果导出与归档环节包括标准格式导出、格式转换、模型适配格式导出、元数据补全、质量报告生成和结果归档，用于适配不同应用场景下的数据使用要求。

在汉藏平行语料加工任务开展前，应由汉藏平行语料需求方、加工管理方和加工方共同明确汉文文本和藏文文本的加工级别、语种方向、采集元数据范围、对齐要求及质量要求，并评估其可行性。加工过程中应把控语料质量，使汉藏平行语料加工活动安全、有序开展，并及时做好数据备份、结果留存和质量追溯。

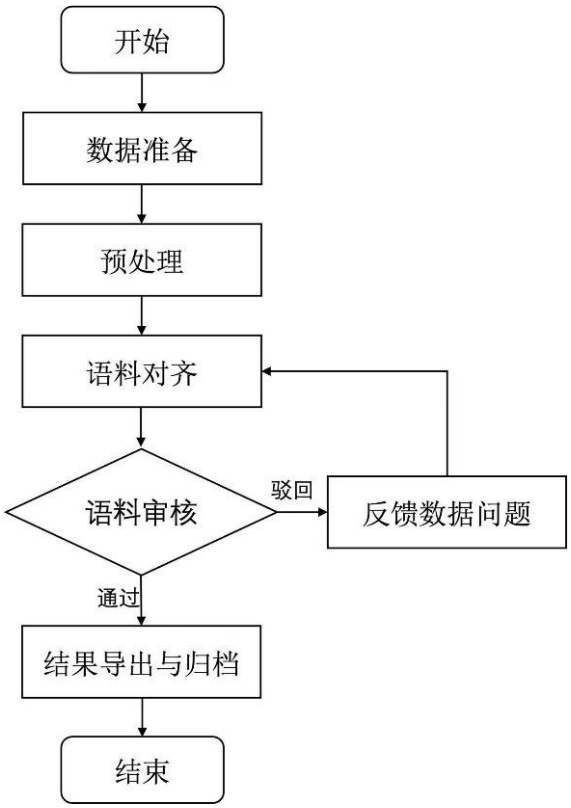


图 1 汉藏平行语料加工流程

6 汉藏平行语料加工前期准备要求

6.1 待加工数据的确定

汉藏平行语料需求方、加工管理方和加工方对于待加工数据确定的要求如下。

a) 待加工数据来源与范围确定：汉藏平行语料需求方应明确待加工数据的来源、内容范围、语种类别和使用边界。待加工数据来源宜包括古籍文献、现代图书、新闻文本、政务公文、法律文书、教材资料、权威网站发布内容及其他合法合规获取的汉藏双语文本资源。

b) 待加工数据整理与分类：数据加工前，汉藏平行语料需求方应对待加工数据进行核查、整理和分类。分类维度宜包括语料来源、主题领域、文本体裁、数字化状态、加工难度和应用场景等，并依据分类结果形成加工批次或任务包，为后续加工提供基础。

c) 语种方向与加工粒度确定：应根据加工任务需要，明确待加工数据属于汉藏方向、藏汉方向或双向加工任务，并确定加工粒度。加工粒度可包括篇章级、段落级、句子级、短语级或其他级别。对于不同方向和不同粒度的加工任务，应分别提出对应的对齐要求和质量要求。

d) 元数据与应用用途确定：应根据加工目标确定待采集和保留的元数据范围。元数据宜包括语料来源、形成时间、语言方向、领域类别、文本类型、处理状态、质量等级和版本信息等。对于面向机器翻译、大语言模型训练、监督微调、评测或知识增强应用的数据，还应明确其用途属性和数据组织要求。

e) 数据安全与可用性确定：汉藏平行语料需求方、加工管理方及加工方应在加工前核验待加工数据的来源合法性、授权情况和使用边界，识别数据类别、级别以及是否涉及个人信息、敏感个人信息、重要数据等需重点保护的数据，并检查数据的完整性和可用性。对需重点保护的数据，应根据加工目的和安全风险采取访问控制、加密、去标识化、匿名化等适当措施。

6.2 汉藏平行语料加工任务

6.2.1 加工内容

汉藏双语语料应由需求方提供。语料来源主要包括古籍文献、现代图书、新闻文本、政务公文、法律文书、教材资料、权威网站发布内容及其他合法合规获取的汉藏双语文本资源。

汉藏平行语料的加工应优先选择数字化后的汉藏双语语料，尚未数字化的语料可通过扫描、拍照等手段，再采用光学字符识别（OCR）的方式转换成数字化形式，或直接通过键盘录入。通过光学字符识别（OCR）或键盘录入的汉藏双语语料应增加校对环节，以提升内容的准确性与质量。

汉藏平行语料加工任务应明确以下内容。

a) 应用目标：包括但不限于机器翻译、自然语言处理、大语言模型训练数据准备、监督微调数据构建、评测数据集建设和知识增强应用。

b) 语种方向：包括汉藏方向、藏汉方向或双向加工任务。

c) 加工粒度：包括篇章级、段落级、句子级、短语级或其他级别。

d) 加工单元：包括汉藏双向句对、多粒度对齐单元及其元数据单元。

e) 数据类型：包括训练样本、评测样本、术语资源、参考对齐样本或其他应用所需数据类型。

f) 交付要求：包括数据格式、元数据范围、质量指标、审核要求和交付方式。

6.2.2 加工结果

6.2.2.1 完整性

在符合汉藏平行语料需求方数据处理要求的前提下，加工方的结果应使藏文、汉文语料以及对应元数据完整，无信息丢失。

6.2.2.2 准确性

在符合汉藏平行语料需求方处理要求的前提下，加工方的结果应达到藏文与汉文的对应关系准确，藏文、汉文以及元数据均应经过校对，不应出现语料对齐错误。

6.2.2.3 可用性

汉藏平行语料加工方的结果应符合以下要求：

a) 能被语料检索、管理和生产工具解析；
b) 无乱码、多余标签等无效信息；
c) 无格式混乱或汉藏语料不对应情况；
d) 无需求方未要求的冗余信息。

6.2.2.4 格式要求

汉藏平行语料加工方的结果应符合需求方给出的格式要求，加工结果的数据格式应包括翻译存储交换（TMX）、TXT、JSON等，并符合以下要求：

a) 翻译存储交换（TMX）文件符合翻译记忆库交换要求，包含留存版本号、编码格式、制作语料的工具名称、制作时间、汉藏双语语言编码等元数据信息；
b) TXT 文件采用 UTF-8 编码格式；
c) JSON 文件采用 UTF-8 编码格式，字段定义清晰，键值结构完整，符合后续加工、管理和应用要求；
d) 面向模型训练、监督微调、评测或知识增强的数据文件，根据应用任务要求保留必要的输入、输出、标签、来源和版本等信息。

6.3 加工规则

6.3.1 预处理

在开展语料对齐前，按以下要求对语料进行预处理。

a) 格式检查：对扫描、拍摄、导入或转换形成的数据进行格式识别和确认，检查文件格式、字符编码、文本结构及图像清晰度是否符合后续加工要求。
b) 清洗：对汉藏语料中的乱码、异常字符、无关标记、重复标签、广告信息及其他干扰内容进行清除。藏文部分宜采用藏文拼写检查工具和自动校对工具进行规范化处理，汉文部分宜采用成熟校对工具进行纠错处理。
c) 藏文编码归一化：对藏文文本中的字符表示形式、编码形式和兼容写法进行统一处理，保证不同来源数据在字符编码和显示方式上的一致性。
d) 去重：对待加工语料与既有语料、不同来源语料以及不同任务批次中的数据进行比对和查重，避免重复加工和重复收录。
e) 去标识化：对经识别需重点保护的信息，根据数据类型、加工目的和安全风险采取适当的安全保护措施。涉及个人信息的，根据需要采取去标识化等处理措施。
f) 长度过滤与截断：根据加工任务和应用目标设置语料最大长度和最小长度。对过长文本宜按规则进行切分或截断；对过短、严重不完整或明显不适于加工的文本宜予以过滤。

g） 模式过滤：对不符合语料加工要求的文本模式、内容噪声严重的文本、语言混杂且难以校正的文本以及无实际语义对应关系的文本进行过滤。

h） 语种方向校验：对汉藏语料的源语言、目标语言及双向标识进行检查，防止语种方向标注错误、语言混放或方向信息缺失。

i） 元数据检查：对语料来源、形成时间、领域类别、文本类型、处理状态、版本信息等元数据进行检查，使元数据与文本内容对应一致。

j） 集合隔离与数据污染检查：对于面向机器翻译、大语言模型训练、监督微调、评测或知识增强的数据，根据用途对训练集、验证集、测试集或评测集进行独立划分，并检查不同集合之间是否存在重复或不当混用。

6.3.2 语料对齐

语料对齐流程包含以下步骤。

a） 导入：将经预处理后的汉藏语料导入对齐工具，并明确语种方向、加工粒度、任务类型和元数据范围。

b） 藏文断句：依据藏文语言规则、标点形式或加工规则，对藏文文本进行句子边界识别和切分；必要时可结合人工方式对自动断句结果进行修正。

c） 长句切分与结构调整：对于句法结构复杂、语义跨度较大或双语对应关系不清晰的长句，应按语义完整性和对齐需要进行合理切分、合并或重组。

d） 多粒度对齐：根据加工任务要求，在篇章级、段落级、句子级、短语级或其他级别上建立汉语与藏语之间的对应关系。通过自动对齐工具结合人工校正方式实现汉藏语料对齐。对齐过程中应重点处理一对一、一对多、多对一及局部缺省等情况，使语义对应关系清晰、结构关系合理。

e） 汉藏双向句对校核：对于双向加工任务，应对汉藏方向和藏汉方向的句对关系进行交叉检查，保证双向句对的一致性和可复用性。

f） 术语一致性校验：对法律、公文、教材、政务、新闻等领域中的专名、固定表达和高频术语进行一致性检查，避免同一术语在相同或相近语境下出现不必要的多种对应形式。

g） 人工修正：对自动对齐结果中存在的断句错误、错位对齐、语义不对应、术语不统一等问题进行人工修正，并记录必要的修正信息。

6.3.3 语料审核

加工后的汉藏平行语料按以下要求进行质量审核。

a） 抽样检查：随机抽取不少于加工结果总量 10%的语料进行人工审核。

b） 人工审核：根据汉藏平行语料需求方、加工管理方和加工方前期共同确认的交付要求，参照汉藏平行语料需求方提供的示例语料进行核对，抽样数据的准确率应达到 99%以上。

c） 结果记录：审核记录宜包括语料批次及版本、审核人员、审核日期、发现的问题及对应语料标识和处理结果，以便于将审核意见定位到具体语料，用于后续修改和质量追溯。

6.3.4 结果导出与归档

语料加工结果按以下要求进行导出与归档。

a） 标准格式导出：将经过预处理、对齐、审核后的双语语料转换为国际或行业通用的标准结构化文本格式。

b) 格式转换：针对不同的应用场景，将对齐后的双语文本转换成便于高效读取的表格或纯文本对格式。
c) 模型适配格式导出：适配当前主流的神经网络机器翻译、大语言模型微调、评测或知识增强所需的特定数据结构，导出的数据应字段完整、结构清晰、可直接用于后续处理。
d) 元数据补全：为语料库附上标识，确保数据的可溯源性和学术严谨性。
e) 质量报告生成：对最终交付的语料库进行客观的评估，生成可视化的质量报告。
f) 结果归档：建立安全、规范的语料存储与版本管理机制，防止数据丢失或污染。

6.4 加工需求变更

汉藏平行语料需求方提出需求变更时，应按以下流程。

a) 需求变更通知：需求方及时以书面形式通知语料加工管理方，明确说明变更的具体内容和原因。
b) 评审变更：
   1) 变更评审：加工管理方对需求变更进行详细审核，评估变更对加工任务的影响，包括需求方加工规范、流程可行性、数据安全以及资源需求的变化；
   2) 可行性分析：确认变更后的需求是否可行，并评估所需的调整和额外工作量。
c) 更新加工任务：
   1) 任务调整：根据审核结果更新语料加工任务内容，包括更新需求方加工规范、调整时间安排及资源配置；
   2) 文档更新：及时更新相关的需求方加工规范和文档，使参与加工任务的人员能掌握最新的需求与加工流程。
d) 重新进入评估阶段：
   1) 预加工测试：在更新后的需求方加工规范基础上进行小规模的预加工测试，验证变更后加工任务的有效性与可操作性；
   2) 结果验收：将预加工测试结果提交需求方验收，根据反馈意见及时进行必要的调整。
e) 正式执行：需求方确认更新后的加工任务无误后，正式启动加工需求变更后的任务执行。

6.5 加工环境要求

6.5.1 选择加工工具

6.5.1.1 通则

汉藏平行语料加工依赖专用工具和通用工具两类软件系统。专用工具主要是指针对藏语特有的语言特征设计的工具，如藏文编码转换工具、藏文拼写检查工具、藏文长句分割工具等；通用工具则是指针对汉藏语料库共性需求的工具，如文本格式转换器、自动校对工具、文字识别工具、汉藏语料对齐工具等。

汉藏平行语料加工方应选择符合6.5.1.2给出的功能要求的语料对齐工具，并宜根据6.5.1.3给出的评价方法对语料对齐工具的效率进行评价，选择适宜的语料对齐工具完成语料加工任务。

6.5.1.2 功能要求

6.5.1.2.1 故障局部化能力

语料对齐工具的局部功能出现问题时，不应影响其他功能的正常使用。

6.5.1.2.2 数据自动保存和恢复功能

语料对齐工具能对数据自动保存和恢复，避免异常情况下的数据丢失。

6.5.1.2.3 显示功能

语料对齐工具的页面应提供直观的汉文显示，具有显示与编辑GB/T 13000—2025藏文区字符的功能，并提供用户友好的交互体验，有助于加工人员快速掌握工具的使用方法。

6.5.1.2.4 操作功能

语料对齐工具应具备对齐汉藏双语文本所需的以下操作功能：

a) 文字编辑：在允许文字输入的内容标识区域，能对文字进行修改、删除和添加等操作；
b) 合并：能将分布在两行的文本合并成一行；
c) 拆分：能将一行文本切分成两行；
d) 位置调整：能将文本上移、下移、插入行以及删除行；
e) 回退：能回退至上一步操作，没有上一步时，停留在当前操作；
f) 对齐：文本调整完成后，能执行段落或句子级别的对齐；
g) 导出：对齐完成后，能导出对齐的汉藏双语文本；
h) 保存：能将对齐过程中的文本进行保存。

6.5.1.2.5 帮助系统

语料对齐工具应提供：

a) 网络环境或非网络环境下的功能帮助文档，并与实际工具功能保持一致；
b) 基本操作指南与示例，便于用户快速熟悉工具操作；
c) 错误提示与定位功能，帮助用户快速解决操作问题。

6.5.1.2.6 兼容能力

语料对齐工具的兼容能力要求如下：

a) 服务器端语料对齐工具应适配主流浏览器，避免依赖特定浏览器或操作系统的脚本插件；
b) 服务器端工具应适应不同浏览器与分辨率，明确推荐的浏览器及最佳分辨率，使页面布局和元素显示完整；
c) 本地端工具应提供完整的安装文档，说明所适配的操作系统、配置需求及常见问题解决方案。

6.5.1.3 效率评价

对于语料对齐工具效率的评价，主要包括以下方面：

a) 响应时间：
   1) 工具启动时间；
   2) 自动对齐、拆分、合并、保存等基本操作的响应时间；
   3) 恢复作业时间：关闭后再次打开工具时，能快速定位上次作业位置的时间。
b) 便捷度：
   1) 能使用快捷键操作；
   2) 提供右键调出选单的功能。

6.5.2 选择加工场地

加工场地的选择应依据汉藏平行语料加工任务的数据安全、保密要求及加工效率考虑。

加工场地应具备以下条件：

a) 具备稳定的网络与电力供应等加工设备运行所需的基本条件；
b) 配备汉藏平行语料加工必需的软硬件设施，包括计算机系统、汉藏文字识别系统、汉藏词句对齐系统、藏文编码转换系统、藏文拼写检查系统、自动校对系统等；
c) 具备完善的数据安全管理措施，包括但不限于访问权限控制、数据备份和恢复机制、数据去标识化和安全存储的实施；
d) 符合国家相关法律法规及行业规范，具备适合长期数据加工的环境条件，保障加工人员工作效率与加工成果的质量和安全。

7 加工任务执行过程的要求

7.1 加工任务创建要求

7.1.1 任务包创建要求

任务包创建过程的要求如下。

a) 数据分组：汉藏平行语料需求方应将数据按类型、内容或加工难度合理分组，检查每个任务包中的数据是否具有一致性，以提高加工的准确性和效率。
b) 质量控制：任务包创建时应考虑加工质量的监控措施，设置质量检查点和反馈机制，以检验语料加工的准确性和一致性。
c) 任务说明：为每个任务包提供明确的加工说明，以指导加工人员进行正确的操作，减少加工错误。

7.1.2 任务账户创建要求

任务账户创建过程的要求如下。

a) 账户配置：根据加工人员的角色和职责创建账户，包括加工人员、审核人员和管理员，检查每个账户是否具备适当的权限和功能。
b) 权限管理：为不同角色分配相应的权限，包括数据访问、加工编辑和质量审核权限，以保护数据安全。
c) 账户规则：制定账户使用规则，包括登录安全、操作规范和数据保护要求。

7.2 加工任务分发要求

7.2.1 加工管理方应根据加工任务的具体要求、环境和工具选择合适的任务分发方式。

常见的方式包括：

a) 主动领取：加工人员自选任务；
b) 系统自动分发：系统根据任务需求和人员状态自动分配。

7.2.2 加工管理方应制定优化策略，平衡加工质量和速度。

示例：针对任务复杂度设置合理的优先。

7.2.3 加工任务分发应采取以下保障措施：

a) 安全管理制度：制定并实施加工任务分发的安全管理制度，检查分发过程的安全性和合规性，包括数据传输安全、访问控制和权限管理等方面；

b) 日志记录与监控：

1) 明确任务分发日志的内容，包括分发时间、任务分配情况和操作记录；

2) 实施实时监控，以便及时发现并解决可能的安全问题或异常情况。

附 录 A
（资料性）
加工人员管理要求

A.1 加工人员范围

汉藏平行语料加工方应根据汉藏平行语料加工任务的具体需求，明确加工人员范围与资质要求。
汉藏平行语料加工人员应具备以下能力:

a) 语言能力：具备较强的藏语与汉语阅读和理解能力，能快速阅读汉藏语料的原文和译文，准确识别语言表达和语义；
b) 专业能力：具备汉藏语料文本处理的专业知识与技能，掌握语料加工流程，能熟练使用藏文拼写检查系统、汉藏词句对齐工具等相关软件；
c) 技术能力：能操作加工所需的软件与设备，熟练掌握汉藏语料加工系统的应用，具备排查和解决加工过程中出现的技术问题的能力。

A.2 加工人员培训

汉藏平行语料加工方应制定并实施针对性的加工人员培训计划，培训内容应包括但不限于：

a) 汉藏语料加工任务的具体流程；
b) 汉藏语料加工工具和软件的使用方法，包括汉藏语料对齐工具、拼写检查工具及自动校对工具的操作方法；
c) 数据处理的流程和常见问题解决方案；
d) 数据安全、去标识化处理和隐私保护相关的法律法规与操作规范。

只有通过培训考核合格的加工人员，方可参与实际加工任务。

A.3 加工人员能力档案

汉藏平行语料加工方应建立详细的加工人员能力档案，记录加工人员的培训经历、考核成绩、任务参与情况及加工质量评价。承担当项任务结束后，应及时更新加工人员的能力档案，用于后续任务的人员选择、任务分配，以及质量监控和持续改进。

参 考 文 献

[1] GB/T 19000—2016 质量管理体系 基础和术语

[2] GB/T 35295—2017 信息技术 大数据 术语

[3] GB/T 40035—2021 双语平行语料加工服务基本要求

[4] GB/T 42755—2023 人工智能 面向机器学习的数据标注规程

[5] ISO 25237:2017 Health informatics — Pseudonymization