国家标准化指导性技术文件《信息技术 汉藏平行语料加工流程》（征求意见稿）编制说明

一、工作简况

（一）任务来源

国家标准化指导性技术文件《信息技术 汉藏平行语料加工流程》制定项目（登记编号为20263104-Z-469）由全国信息技术标准化技术委员会提出并归口，由全国信息技术标准化技术委员会登记立项。

本标准由西藏大学主办，起草单位包括：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。

本标准主要起草人：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。

标准起草组人员组成和分工见表1。

表1 起草组人员组成和职责分工表

序号	起草人姓名	所在单位	工作内容
1	尼玛扎西	西藏大学	统筹指导
2	于永斌	电子科技大学	协调、起草
3	头旦才让	青海师范大学	汉藏平行语料加工流程
4	陈晓研	中国电子技术标准化研究院	标准起草，过程管理
5	班玛宝	西藏大学	加工任务定义
6	王向向	电子科技大学	加工规则
7	仁增多杰	西藏大学	预处理
8	仁青东主	西藏大学	语料对齐
9	黄姗姗	中国电子技术标准化研究院	语料对齐
10	洛桑嘎登	西藏大学	语料审核
11	群诺	西藏大学	语料审核

12	范满平	电子科技大学	加工工具
13	高凡	电子科技大学	加工工具
14	官却才让	电子科技大学	加工任务评估试验
15	道吉扎西	西藏大学	加工任务评估试验
16	冯箫	电子科技大学	质量控制流程试验
17	张子玥	电子科技大学	质量控制流程试验
18	李臣波	电子科技大学	内容整合
19	郑智文	电子科技大学	内容整合
20	刘羽童	电子科技大学	内容规范化
21	李蕾	北京邮电大学	内容规范化

（二）制定背景

汉藏平行语料是开展跨语言信息处理、构建多语种人工智能系统的基础性资源，其质量水平和加工规范直接影响机器翻译、语言模型训练、语言资源服务等下游任务的性能与可用性。

藏语作为我国藏族人民的主要语言，其与汉语的平行语料加工规范的制定对于推动藏语机器翻译技术的发展具有重要意义。当前，汉藏双语语料在数据来源、格式规范、加工流程、质量评估等方面存在显著不一致，严重制约了藏语在信息技术领域的标准化与产业化发展。制定《信息技术 汉藏平行语料加工流程》国家标准化指导性技术文件，能够为汉藏机器翻译及相关应用提供统一的技术基础，满足藏族群众在教育、医疗、公共服务等领域的实际需求，推动藏族地区经济社会和信息产业的协同发展。

我国政府高度重视少数民族语言文字的保护与发展，出台了一系列相关法律法规和政策规划。例如，《中华人民共和国民族区域自治法》明确规定保障少数民族使用和发展本民族语言文字的权利；《国家通用语言文字法》强调促进少数民族语言文字的规范化、标准化和信息化；此外，《西藏自治区国民经济和社会发展第十五个五年规划纲要》中也提出要加强少数民族语言文字教育和保护传承，实施藏语言大模型建设应用工程。制定《信息技术 汉藏平行语料加工流程》国家标准化指导性技术文件，符合国家关于少数民族语言文字保护的政策导向，是落实相关法律法规的具体实践，有助于推动藏语的信息化进程，传承和弘扬藏族

优秀传统文化。

从全球来看，自然语言处理技术正加速向多语言、低资源场景延伸，人工智能与大数据的深度融合为跨语言模型训练与应用开辟了新路径。我国在双语技术领域已形成较为完整的产业链，市场规模快速增长。作为民族语言信息化的重要组成部分，汉藏平行语料加工技术在教育数字化、跨境智能服务、文化遗产保护等领域具有独特价值，尤其在“一带一路”沿线多语言互通、边疆数字治理等国家战略中，展现出不可替代的应用前景。

本标准旨在建立一套面向信息技术应用的汉藏平行语料加工规范，系统定义语料的采集、预处理、对齐、审核等全过程技术要求，涵盖语料格式检查、拼写校对、长度过滤、双语对齐、脱敏处理、质量审核等关键环节。标准还明确了加工人员能力要求、工具选择规范、加工环境配置以及质量控制机制，提出抽样审核比例不低于 10%、语料对齐准确率不低于 99%、句级对齐误差率不高于 3%等定量指标，确保加工结果的完整性、准确性与可用性。

本标准适用于藏语自然语言处理、机器翻译语料构建、语言资源服务、跨境数字政务等多种应用场景，亦可为面向藏语的语言智能系统开发与科研工作提供统一的技术支撑，推动藏语资源的规范化、结构化与共享化建设，为低资源语言技术发展提供切实可行的标准基础。

（三）起草过程

2025年3月：起草核心人员编制完成标准草案预研稿。

2025年4月：全国信息技术标准化技术委员会字符集与编码分技术委员会（SAC/TC28/SC2）协助修改完善标准草案，准备全套标准立项材料。

2025年4月9日：参加全国信息技术标准化技术委员会组织召开的立项内审会，修改后同意立项。会后，起草核心人员按内审意见完善标准草案。

2025年7月11日：经全国信息技术标准化技术委员会下属藏文信息技术国家标准工作组审查，同意以此方案提出立项。

2025年10月：按照申请标准立项的要求，起草核心人员编制完善标准草案，由中国电子技术标准化研究院提出立项。

2026年6月12日：全国信标委下达本标准制定计划。

2026年7月：西藏大学牵头成立本标准编制工作组，成员来自科研院所、高

校、用户等单位。起草组成员信息导入“标准智能研制系统”，分配系统账号。

2026年8月：编制形成工作组讨论稿（第1稿），经全国信息技术标准化技术委员会组织专家审查，同意开题。会后，根据审查专家意见，就涉及数据安全的技术内容征求全国网络安全标准化技术委员会SAC/TC 260有关专家建议。

2026年9月：由起草组以线上会议形式组织召开标准编制讨论会2次，对标准草案的范围、术语、流程图等内容进一步修改完善。起草组成员按照任务分工，通过“标准智能研制系统”对标准文本各章节协同编写，确保文本表述清晰、严谨，格式符合GB/T 1.1要求。

2026年9月底：形成征求意见稿初稿。

2026年10月14日：全国信标委组织召开内审会，审议通过标准征求意见稿初稿，并提出修改意见。

2026年10月至11月：由起草组以线上会议形式组织召开标准编制讨论会，根据内审意见，完善形成征求意见稿和征求意见稿编制说明。

二、国家标准编制原则、主要内容及其确定依据

（一）标准编制原则

本标准的编制紧密围绕我国少数民族语言文字信息处理的发展需求，遵循“科学合理、技术先进、经济可行、规范适用”的总体要求，具体遵循以下五项编制原则：

1、规范统一与可追溯原则

本标准严格按照 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编写》的规定进行起草。针对汉藏平行语料加工全过程，系统规范了数据采集、预处理、语料对齐、质量审核等关键环节的技术要求，统一了数据格式与元数据标准。同时，确立了量化的质量评估指标与试验方法，确保语料加工全流程有章可循、过程可追溯、结果可验证。

2、需求导向与工程实用原则

本标准坚持立足实际应用场景，紧密结合汉藏机器翻译、大语言模型预训练与监督微调数据构建、评测数据集建设及知识增强等前沿与主流应用需求。在技术路线选择上，注重人机协作的工程可行性，既明确了自动化对齐与清洗工具的

功能与兼容性指标，又规范了人工校正与质检的操作流程，切实保障语料资源具备较高的工程化可用性与产业落地性。

3、汉藏双语与专业保障原则

本标准充分考虑汉语与藏语在文字编码、语法结构、语义表达及断句规则等方面的天然差异，针对性地制定了多粒度（篇、段、句、短语）对齐规范及拼写校正、噪声剔除等前处理标准。同时，强化人-机-环境的专业化保障体系，明确了加工人员的双语理解与质控能力要求，规定了岗前培训及资质管理制度，从源头上保障汉藏语义对齐的专业度与精准性。

4、前瞻创新与可扩展原则

本标准在满足传统信息处理技术的基础上，前瞻性地契合人工智能与新型大模型技术对高质量数据集的建设需求。标准架构设计兼顾多源异构数据（如公文资料、教材出版物、书面文本等）的兼容性，并保留了多层级加工的技术扩展接口，确保标准在技术演进和未来应用场景拓展中保持持续的适用性与领先性。

5、安全合规与风险防控原则

本标准高度重视数据安全与合规管理。在前期准备与任务执行阶段，明确规定了语料采集的合法性要求，建立了严格的数据清洗脱敏、访问控制、账号权限管理及实时监控日志机制，保障汉藏双语语料在采集、分发、加工及交付全生命周期中的信息安全与合规使用。

（二）标准主要内容

本标准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。本标准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。

本标准的主要技术内容涵盖汉藏平行语料加工的全流程，包括流程设计、前期准备和任务执行，具体如下：

（1）汉藏平行语料加工流程

本标准规定了汉藏双语平行语料加工的全过程技术要求，涵盖数据准备、预

处理、语料对齐、语料审核和结果导出与归档等关键环节，旨在构建面向信息处理任务的标准化语料加工流程，提升双语语料资源的规范化程度与工程化可用性。

（2）汉藏平行语料加工前期准备要求

本标准明确了语料采集的数据类型范围、格式规范及使用合规性要求，覆盖书面文本、公文资料、教材出版物等主要来源形式。针对语料质量保障，规定了清洗、去重、脱敏、拼写校正、噪声剔除、长度过滤等前处理环节的操作标准，确保输入数据满足结构完整、语义清晰和信息安全的要求。

本标准系统规范了汉藏语料对齐的基本流程和操作方法，包括双语数据导入、自动断句、句子对齐、人工校正和对齐结果导出等操作步骤，并对所使用工具的功能特性、使用效率、兼容性及数据格式提出了相应技术指标要求，以保障加工过程的标准化和可追溯性。

（3）加工任务执行过程的要求

- 加工任务创建要求：包括任务包创建和任务账户创建两个环节。任务包创建需对数据按类型、内容或难度合理分组，设置质量检查点和明确标注说明；任务账户创建需根据角色配置账户权限，制定账户使用规则，确保数据安全和加工规范性。

- 加工任务分发要求：需根据任务要求选择合适的分发方式（如主动领取或系统自动分配），制定优化策略平衡质量与效率，并采取安全保障措施包括建立安全管理制度、实施日志记录与实时监控，确保分发过程的安全合规。

除上述主要内容外，本标准另设资料性附录“标注人员管理要求”，明确了标注人员的选拔范围、培训要求及能力档案管理，确保人员素质达标。

（三）标准主要内容的确定依据

本标准主要技术内容的选择与指标的设定，基于我国少数民族语言文字信息处理的工程实践、大模型时代的数据应用需求，以及现行国家标准体系的协调衔接，具体确定依据如下：

1、汉藏平行语料加工流程设计依据

本标准规定的数据采集、预处理、语料对齐、质量审核全流程，主要依据**工程化语料库建设的标准化实践**与**复杂数据流水线架构**确定。流程的设计旨在打通

从原始非结构化文本到高品质标注数据的全链路，解决传统少数民族语言语料加工中存在的流程割裂、标准不一、交付物规范缺失等问题，确保加工全过程符合信息技术规范化与工程化实施要求。

2、加工前期准备要求的确定依据

依据国家法律法规关于知识产权保护、数据安全及个人信息保护（如数据脱敏、合规采集）的相关要求，结合公文、出版物、教材等多源异构文本的结构特征，设定了文本清洗、去重、拼写校正及噪声剔除的操作细则，确保输入数据的语义清晰度与合法合规性。

基于汉语（SVO 为主）与藏语（SOV 为主）在语序、句法及分词/切词机制上的天然结构差异，确定了支持篇章、段落、句子及短语的多粒度对齐体系；根据自动化语料处理工具的通用研发性能与适配实践，提出了工具在兼容性、批处理效率及数据导出格式上的量化技术指标。

根据汉藏双语平行语料加工对专业语言能力的极高要求，参照翻译服务及数据标注行业的从业资质规范，确立了加工人员的语言理解、工具操作与质控能力的准入标准及岗前培训制度；硬件与安全环境配置要求则依据信息系统安全等级保护的基本规范制定。

3、加工任务执行过程要求的确定依据

依据敏捷数据标注项目管理与多角色协作（如标注员、审核员、项目经理）的最佳实践，确定了按难度/类型进行数据分组和设置质量检查点的规范，以及基于角色的访问控制的账户权限管理机制，确保任务分配的科学性与数据资产安全性。同时，结合云端/局域网部署环境下语料分发的安全风险，依据信息安全日志审计规范，建立了主动领取/自动分配的动态平衡策略，以及包含全生命周期操作日志记录与实时监控的安全保障要求。

4、质量评估方法的确定依据

基于机器翻译（如 BLEU/TER 评估对语料质量的敏感度）及大语言模型（如 SFT 指令对齐数据对准确率的要求）的技术特点，结合统计抽样检验理论（如 GB/T 2828 抽样检验程序），设定了抽样审核比例、对齐准确率及误差率等核心量化指标。同时，依据“自动化检测工具+专家人工复核”的双重验证机制确定试验方法。自动化检测负责格式规范、断句对齐率及字符异常的快速扫描，人工

审核负责深层语义一致性与翻译质量判定，两者协同保障语料成果的准确性、一致性与工程可用性。

三、试验、验证情况分析

（一）试验情况

为确保本标准所规定的试验方法、质量控制流程及评估指标在实际工程中的科学性、有效性与可操作性，标准起草组联合相关科研院所、高校及专业数据加工企业，组建了专项测试小组，开展了汉藏平行语料加工试验验证工作。具体验证情况如下：

1、试验验证总体设计

试验验证采用了“主观评价方法、信息化方法与仲裁方法”相结合的综合验证策略：

（1）测试数据集：抽取了涵盖公文资料、教材出版物及新闻文本等典型场景的汉藏双语语料共计 10 万句对（含篇章、段落及句子级别），作为验证样本。

（2）**测试环境与工具：**搭建了标准的汉藏双语平行语料加工系统，配置了自动化断句工具、高精度符号匹配检测模块及质量审计日志系统。

2、加工任务评估试验验证

（1）可行性及规则审查测试：加工管理团队对任务的时间资源匹配度及加工规则进行了书面评估，验证了评估记录表单的规范性。

（2）预加工与小规模抽样测试：在 10 万句对样本中抽取 2000 句对（占比 2%）进行小规模预加工测试。测试结果表明，通过“预加工—人工验收反馈—规则微调”闭环，成功在正式任务启动前识别并修正了 3 处关于藏语特殊符号与断句规则的歧义条款，有效验证了预加工验收确认机制对规避大规模加工风险的实用价值。

3、质量控制流程试验验证

（1）常规质量检查点验证：系统成功对无效提交、文本截断、格式错位等结构性错误进行了 100% 自动识别并输出检测报告，大幅提升了基础质检效率。同时，进行了主观评价与仲裁。质检员对错字、多字、少字及语义微瑕进行了人工核查。在系统与人工判定存在不一致（占比约 1.5%）的样本中，通过执行“人工复检仲裁机制”，最终确保了语义对齐准确率达到 98.5% 以上，证明了仲裁

机制的有效性。

（2）特殊质量检查点验证：采用高级符号匹配工具对藏语特有标点、噪声符号与预设规则的一致性进行自动化比对；同步由高级双语专家进行终审抽检。验证表明，“工具扫描+专家终审”的协同模式可将藏语复杂噪声加工的隐蔽错误率降低 80% 以上。

（3）数据归档与可追溯性验证：所有的质检日志、错误记录表及修正历史均成功通过信息化系统实现无缝归档与全流程溯源，验证了数据链条的完整性与安全性。

4、试验结论

试验结果表明，本标准规定的加工流程能够有效保障汉藏平行语料加工结果的准确性。标准中的各项质量控制要求能够真实、客观地评估汉藏平行语料的加工质量，具有极高的工程可操作性与行业推广价值。

（二）预期的经济效益

本标准的实施将为汉藏双语技术研发与产业化应用提供核心支撑，推动机器翻译系统、跨语言信息检索工具、双语教育平台等相关产业的快速发展。这不仅能够激活民族地区数字经济潜力，还可为涉藏企业及技术服务机构提供标准化数据服务，降低多语言产品开发成本，提升其在跨境贸易、文化旅游等领域的市场竞争力。

（三）预期的社会效益

本标准的制定和推广将显著提升汉藏双语技术的实用性与覆盖面，使藏族群众更便捷地获取双语政务服务、教育资源及数字化内容，促进民族地区数字包容性发展。同时，通过规范化的语料加工，汉藏双语的跨文化传播与教学将更加精准高效，有助于增强各民族间的文化互信与社会凝聚力，为铸牢中华民族共同体意识提供技术赋能。

四、与国际、国外同类标准技术内容的对比情况

本标准技术内容为我国独创，无同类国际标准、国外标准。

当前，国际标准化体系中尚未形成针对汉藏双语平行语料加工的专项技术规范。国际上，英语、法语等高资源语言对已建立成熟的语料加工标准（如欧盟的

Europarl 平行语料库标注框架、ISO 24622 多语言资源管理规范），但这些标准主要面向语法结构相近的印欧语系语言，难以适配汉藏双语间显著的语法差异（如藏语的黏着语特性与汉语的孤立语结构）及文化专有项的等效对齐需求。现有研究多集中于学术机构零散实践，缺乏可推广的标准化加工框架，制约藏语自然语言处理、机器翻译等领域的研究与产业化应用。本标准的制定将填补该领域国际标准空白，为全球藏语语言资源建设提供中国方案。

五、以国际标准为基础的起草情况

无同类的国际、国外标准，本标准未采用国际、国外标准。

六、与有关法律、行政法规及相关标准的关系

本标准未替代或废止现行相关标准，与现行相关法律、法规、规章和强制性国家标准协调一致。

本标准规范性引用了 GB/T 13000-2025《信息技术 通用编码字符集（UCS）》，规定汉藏双语平行语料对齐工具应支持 GB/T 13000-2025 藏文区的字符显示与编辑，以保障生成的语料库中的藏文字符在跨系统、跨国界交换传输中的一致性。

本标准在起草过程中还参考了以下我国国家标准和国际标准：

- GB/T 19000—2016　质量管理体系 基础和术语
- GB/T 31219.2—2014　图书馆馆藏资源数字化加工规范 第 2 部分：文本资源
- GB/T 40035—2021　双语平行语料加工服务基本要求
- GB/T 42755—2023　人工智能 面向机器学习的数据标注规程
- ISO 25237:2017　Health informatics — Pseudonymization

七、重大分歧意见的处理经过和依据

无。

八、涉及专利的有关说明

本标准未涉及专利、版权和著作权等知识产权问题。

九、实施国家标准的要求以及措施建议

建议本标准作为指导性技术文件发布，自发布之日起 6 个月实施。在标准实施中，建议及时在行业内及相关单位开展标准的宣贯，使相关单位能够学习并应用本标准，并结合实际应用要求贯彻实施标准。

十、其他应当说明的事项

无。

国家标准化指导性技术文件

《信息技术 汉藏平行语料加工流程》编制工作组

2026-11-18