中国翻译协会标准

T/TAC 1-2018

语料库通用技术规范

General specifications for corpus

编　制　说　明

《语料库通用技术规范》标准起草组

2018-10-10

一、任务来源

中国翻译协会是包括翻译与本地化服务、语言教学与培训、语言技术工具开发、语言相关咨询业务在内的语言服务行业的全国性组织。制定语言服务规范，推动行业有序健康发展，是中国翻译协会的工作内容之一。为顺应国家支持大力发展团体标准的有利形势，中国翻译协会发起制定本规范。本规范是中国翻译协会发布的第四个在我国语言服务领域实施的标准。

近年来，随着人工智能、语言服务、学术研究和语言教学的发展，语料库交易活动日益活跃。为了更好地规范语料库交易市场，推进语料库在人工智能和语言服务、学术研究及其他相关领域中的应用，特编制《语料库通用技术规范》，并计划于 2018 年内予以发布，并同时开始推广应用。

二、目的和意义

随着中国的外商投资和对外直接投资进入新的阶段，以及国家“一带一路”倡议、经济文化“走出去”的进一步实施，语言服务特别是翻译成为了各项政策落地的关键点和瓶颈，传统的方式已经不能满足语言服务需求的迅猛增长，而基于人工智能神经网络技术的机器翻译正在快速发展，创新的人机融合语言服务模式正在悄然兴起，而语料库的研究、建设、交易和共享是其中的关键。建立一个既能服务于学术研究又能服务于语言服务和机器翻译的语料库通用技术规范将起到以下作用：（1）响应“一带一路”倡议和经济文化“走出去”等国家大政方针，服务于国传、外宣、文化、经贸、安全等重大垂直领域，推进国家话语权和语言软实力建设；（2）制定和落实语料库行业标准，可以在此基础上建立自主、可控的以中文核心的语料共享和交易平台；（3）倡导“安全语料大数据”，在充分尊重版权的前提下，以互联网思维和人工智能助力语言服务模式的变革。

本规范部分内容源于国内和国际标准，但在具体方面补充了不少上述标准所没有涵盖的内容。包括：中国国家标准 GB 13715《信息处理用现代汉语分词规范》；国际标准 ISO 639 *Codes for the Representation of Names of Languages*（《语种名称代码》）；国际标准 ISO 3166 *Terminology Bulletin - Country Names*（《国家名称用语公报》）；中国国家标准 GB/T 19682-2005《翻译服务译文质量要求》；国际公约 *Berne Convention for the Protection of Literary and Artistic Works*《保护文学艺术作品伯尔尼公约》等。这些补充体现在：（1）本规范将用于研究类的语料库和用于语言服务及机器翻译的语料库进行了较为完整的分类，并就其共性和不同之处做出了相应的规范，如：对库容和翻译质量的不同规范；（2）本规范搜集整理了各起草单位特别是作为语料使用方提供的方案，针对用于语言服务和机器翻译的双语语料，设计了质量评价方式、评价维度和评价标准，从而使语料交易和共享有了明确的指导依据；（3）基于各起草单位提供的参数，提出了在双语语料评价中语料翻译质量、语料对齐质量和语料数据质量的权重系数范围，其中语料翻译质量的权重系数不得低于 0.7。

本规范的推广应用将对中国方兴未艾发展的语言服务行业起到重要的引导和规范作用，同时有利于中国的语言服务企业和世界同行接轨，促进中国整个语言服务行业的健康发展。

三、规范起草制定原则

本规范按照 GB/T 1.1-2009《标准化工作导则 第 1 部分：标准的结构和编写》给出的规则起草。并符合中国翻译协会提出的“由易到难、急用先行、适度领先”的团体标准建设原则。

四、本规范起草制定过程

中国翻译协会于 2018 年初批准本规范起草方案，委托北京中译天凯教育咨询有限公司组织编写，上海交通大学胡开宝教授主持，并由上海交通大学、北京大学、北京外国语大学、中国人民解放军外国语学院、同济大学、东南大学、南京师范大学、浙江大学、浙江财经大学、中国社会科学研究院、中国标准化研究院、中译语通科技股份有限公司、阿里巴巴（中国）网络技术有限公司、华为技术有限公司、传神语联网网络科技股份有限公司、北京中译天凯教育服务有限公司、成都优译信息技术股份有限公司、杭州中语科技有限公司、苏州联跃科技有限公司等机构的专家共同起草。

本规范起草工作组按照国家标准的制定程序起草本规范。规范起草工作组的全体成员召开了多次研讨会，认真讨论研究了所有条款内容，对术语和条款反复推敲，形成目前的征求意见稿。

五、本规范主要内容及条文说明

通过实施本规范，语料库提供方可以证明其语料符合语料库规范，能够满足基本的语料使用要求；语料库使用方可以判断语料库是否能够用于某特定用途。本规范的核心内容除“术语和定义”部分用于界定语料和语料库的相关概念之外，其余各章先后涉及语料库的建设与加工、管理与维护，以及交易与共享。其他关于语料库的规范将在之后以系列标准的形式予以发布。

以下是本规范主要内容：

3. 术语与定义

为便于使用，以下列出主要术语和定义。

3.1 语料

语言材料或语言应用的样本。

3.2 语料库

由依据一定抽样方法收集的自然出现的**语料**（3.1）所构成的电子数据库，是按照一定

目的和方法进行选择并有序排列的数据汇集。

3.3 形符

语料库（3.2）中出现的所有词形，如 go、goes、went、going 视为 4 个英语词形。

3.4 句对

一个完整的语句（通常以句号、问号和感叹号等为语句标记）和与之内容对应的译文视为一个句对，句对可以是 n 对 n 的，这里的 n 为非负整数。

3.5 库容

语料库（3.2）的容量，即语料库的大小。面向学术研究的**语料库**（3.2）通常以**形符**（3.3）总数为单位来计算库容；面向人工智能和语言服务的**语料库**（3.2）通常以**句对**（3.4）为单位来计算库容。

3.8 语料清洗

使用软件消除**语料**（3.1）中的乱码、多余回车、空格、空行等杂质。

3.9 语料标注

采用人工或计算机自动方式对**语料**（3.1）样本的属性或特征进行描述。

3.25 标注语料库

经过标注处理的**语料库**（3.2）。

3.26 非标注语料库

未经标注处理的**语料库**（3.2）。

4. 建设与加工

4.1 建设流程

语料库建设流程一般应包括语料库设计、语料采集、语料预处理、分词、标注、对齐、语料库生成、管理与维护等步骤。其中，语料库设计、语料采集、语料预处理、语料库生成、管理与维护等为语料库建设的基本流程，分词、标注和对齐等为可以选择进行的流程。

4.2 语料采集

4.2.1 语料采集方法

书面语料的采集主要包括人工输入、扫描输入以及现有电子文本的利用。口语语料的采集包括音频和视频材料等的获取和转写。

4.6 语料库生成

生成后的语料库应满足以下要求：

（1）由加工后的语料构成；

（2）可直接用于语料检索和数据分析；

（3）提供关于语料库用途和库容以及语料的来源、领域和时间跨度等方面的信息；

（4） 语料版权必须清晰，不应存在版权纠纷。

5. 管理与维护

5.1 语料的分类

根据语料加工程度可分为生语料、粗加工语料和精加工语料。

根据语料对齐单位，双语语料可分为篇章对齐语料、段落对齐语料、句对齐语料、语块对齐语料和词汇对齐语料。

5.2 语料库的分类

（1）按语料选取的时间，可分为历时语料库和共时语料库；

（2）按语料的加工程度，可分为标注语料库和非标注语料库；

（3）按语料库代表的领域，可分为通用语料库和专门语料库；

（4）按语言传播媒介，可分为口语语料库和书面语语料库，或笔译语料库和口译语料库；

（5）按语料库中的语种，可分为单语语料库和多语语料库，多语语料库又可分为可比语料库和平行语料库；

（6）按语料库的动态更新程度划分，可分为静态语料库和动态语料库；

（7）按语言产出者的身份，可分为本族语者语料库和学习者语料库；

（8） 按语料保存的信息模态，可分为单模态语料库和多模态语料库。

6. 交易与共享

6.2 语料库评价

6.2.1 评价内容

（1）整体评价：语料库库容、语料范围、类型以及语料库的应用领域；

（2）性能评价：语料库的应用效果以及对用户需求的满足程度；

（3）语料评价：语料获取难易度、语料加工程度、语料质量和语料应用前景；

（4）效益评价：语料库为人工智能、语言服务和学术研究等方面提供服务所获得的社会效益和经济效益。

6.3 语料库交易

6.3.1 交付方式

语料库提供方应说明语料库产品交付给购买方的方式，主要方式包括：

（1）文件，即语料库提供方将整个语料库文件交付给购买方；

（2）API，即语料提供方提供 API 接口供购买方调用。

6.3.2 价格

语料库的交易价格由交易双方协商确定。建议根据以下因素进行综合考虑：

（1）语言对；

（2）领域；

（3）库容；

（4）对齐单位；

（5）格式；

（6）数据来源；

（7）评价结果；

（8）版权类型；

（9）脱敏程度；

（10）购买方免责声明、交付方式。