中 国 翻 译 协 会 标 准

T/TAC x—xxxx

语料库通用技术规范

General specifications for corpus

（征求意见稿）

2018-xx-xx 发布　　2019-xx-xx 实施

中国翻译协会　发　布

前　　言

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。

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》。

本规范起草单位：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。

本规范主要起草人：胡开宝、杨平、罗慧芳、张雪涛、陈圣权、吴永波、谢凝、彭成超、许文胜、李爱军、梁红丽、王海涛、王海波、李洁、潘轶岑、何征宇、刘四元、蔡方仁、俞敬松、高志军、张威、程乐、严志军、黎昌抱、易绵竹、毕玉德、郭庆、管新潮、田绪军、李婵、李晓倩、胡昂、任才淇等。

本规范按照 GB/T 1.1—2009 给出的规则起草。

本规范由中国翻译协会提出并归口。

目　　录

1. 适用范围

本标准侧重于描述并规定语料库的建设与加工、管理与维护、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。

通过实施本标准，语料库提供方可以证明其语料符合语料库标准，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。

2. 规范性引用文件

下列文件对本文件的应用是必不可少的。凡标注日期的引用文件，仅标注日期的版本适用于本文件。凡不标注日期的引用文件，其最新版本（包括所有的修改）适用于本文件。

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》。

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。

3.6 语料库设计

语料库（3.2）建设者对**语料**（3.1）的规模、领域、体裁、语种、语料的加工程度以及语料的应用领域等相关参数进行设定。

3.7 语料采集

将收集到的不同介质**语料**（3.1）转写为可机读的格式或直接利用现有的电子文本。

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。

3.10 脱敏

对**语料**（3.1）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。当涉及客户安全数据或者一些商业性敏感数据时，在不违反系统规则的条件下，对真实数据进行改造。

例：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。

3.11 语料预处理

在加工**语料**（3.1）之前进行**语料清洗**（3.8）和**脱敏**（3.10）等技术处理。

3.12 分词

将连续的字符序列切分成一个个单独的词。

3.13 标注语言

将文本以及文本相关的信息结合起来，展现出关于文档结构和数据处理细节的计算机编码。

3.14 篇头信息标注

篇头信息标注说明整篇**语料**（3.1）样本的属性。

例：语体、领域、标题、作者、作者性别、出版时间、来源出处和出版社等。

3.15 篇体信息标注

对文本内部各种语言学属性的标注，包括词性标注、句法标注、语义标注、语用标注和语音标注等。

3.16 语块

具有完整的意义且高频出现的大于单个词汇的语言现象，包括短语、词语搭

配、习语等。语块不仅包括连续性短语结构，如“in the end”，也包括不连续的句子框架，如“不但……而且……”、“not only...but also…”，还包括一些完整的句子，如“How do you do?”。

3.17 语料对齐

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。

3.18 正则表达式

对包括普通字符和特殊字符在内的字符串进行描述的一种逻辑公式。用事先定义好的一些特定字符以及这些特定字符的组合，组成一个“规则字符串”，用来描述在搜索文本时要匹配的一个或多个字符串，可应用于对**语料**（3.1）的加工、检索等不同阶段。

3.19 生语料

未经任何技术处理的自然**语料**（3.1）。

3.20 粗加工语料

经**语料清洗**（3.8）或**语料预处理**（3.11）后能够进行基本检索和数据提取的**语料**（3.1）。

3.21 副语言特征

以视觉、听觉、嗅觉、味觉、触觉等感知为信息载体的符号系统，如韵律特征（语调、重音等）、突发性特征（说话时的笑声、哭泣声等）、次要发音（圆唇化音、鼻化音等）以及面部表情、视觉接触、体态、手势、谈话时双方的距离等。

3.22 精加工语料

根据特定**语料库**（3.2）建设目的，采用机器或人工手段进行**语料标注**（3.9）的**语料**（3.1）。这些标注包括语音标注、词性标注、句法标注、语义标注以及错

误标注等**篇体信息标注**（3.15）。口语语料的标注还包括**副语言特征**（3.21）标注。

3.23 历时语料库

收录不同时间周期**语料**（3.1）的**语料库**（3.2）。

3.24 共时语料库

收录相同时间周期**语料**（3.1）的**语料库**（3.2）。

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）。

3.27 通用语料库

收录代表语言整体的**语料**（3.1）的**语料库**（3.2）。

3.28 专门语料库

收录代表某一语言的专门语体或专门领域**语料**（3.1）的**语料库**（3.2）。

3.29 口语语料库

收录口语**语料**（3.1）的**语料库**（3.2）。

3.30 书面语语料库

收录书面语**语料（3.1）**的**语料库**（3.2）。

3.31 平行语料库

收录某一语言文本和与之对应的翻译文本的**语料库**（3.2）。

3.32 单语语料库

收录一种语言**语料**（3.1）的**语料库**（3.2）。

3.33 口译语料库

根据口译音、视频材料制作的**语料库**（3.2），可分为**单语语料库（3.32）**和**平行语料库（3.31）**。

3.34 笔译语料库

收录书面翻译**语料**（3.1）的**语料库**（3.2），可分为**单语语料库（3.32）**和**平行语料库（3.31）**。

3.35 多语语料库

收录两种或两种以上具有翻译关系的**语料**（3.1）的**语料库**（3.2）。

3.36 可比语料库

设计和结构上能保证对语料进行不同层面比较的**语料库**（3.2）。可比语料库可分为单语、双语或多语可比语料库。单语可比语料库是指收录具有可比性的一种语言文本的**语料库**（3.2）。双语可比语料库或多语可比语料库指收录具有可比性但不存在翻译关系的两种或两种以上语言文本的**语料库**（3.2），如收录汉语军事新闻话语和英语军事新闻话语的语料库。

3.37 静态语料库

由所选**语料**（3.1）构成的固定规模的**语料库**（3.2）。

3.38 动态语料库

为考察某些语言变化而建设的不断更新的开放性**语料库**（3.2）。

3.39 本族语者语料库

收录本族语者所产出**语料**（3.1）的**语料库**（3.2）。

3.40 学习者语料库

收录语言学习者所产出**语料**（3.1）的**语料库**（3.2）。

3.41 单模态语料库

收录音频、视频或文字材料之一种模态**语料**（3.1）的**语料库**（3.2）。

3.42 多模态语料库

收录音频、视频和文字材料等**语料**（3.1），并采用多模态方式加工、检索和统计的**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。

具体流程如图 1 所示：

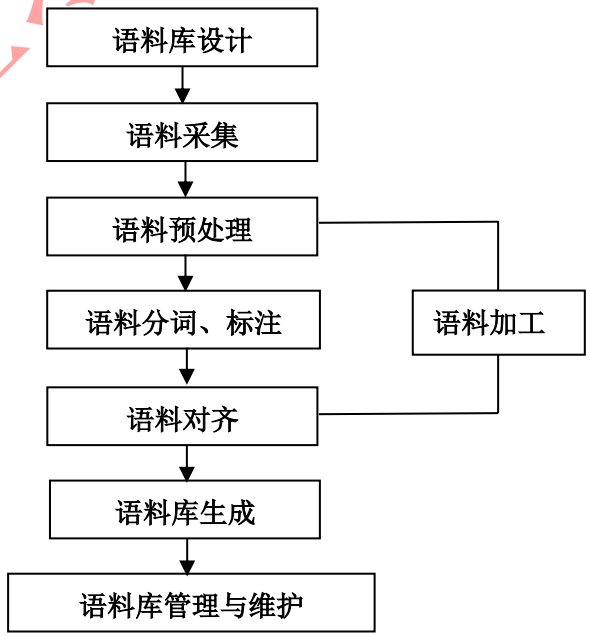

图 1：语料库建设流程图

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。

4.2.2 语料采集要求

（1）真实性

注：语料库收录的语料应为真实语言环境下使用的语料。

（2）准确性

（3）代表性

注：语料应最大限度代表具体语言应用的实际或某一语言变体。

（4）电子化

（5）一致性

注：语料的格式应统一，保持一致。

4.3 语料预处理

在对所采集语料进行加工之前，应进行语料查重、语料清洗和语料脱敏等方面的技术处理。

（1）语料查重

注：检查所采集语料与已有语料是否有重复，避免重复加工。

（2）语料清洗

（3）语料脱敏

4.4 语料标注

4.4.1 语料标注内容

语料标注包括篇头信息标注和篇体信息标注。

4.4.2 标注语言的要求

（1）通用性

语料库应采用适合于语料库软件的标注语言，或适合于编程语言的应用。

（2）简洁性

语料库应功能完备、简单易用，适用于软件检索和扩充增容。

（3）兼容性

语料库应适用于不同语料库检索软件和平台，允许跨平台进行语料的交换与共享。

4.5 语料对齐

双语语料的对齐应依据以下原则进行：

（1）以源语文本为基准，以句子为基本单位，尽量实现双语语句之间的一一对齐；

（2）句号、问号和感叹号等标点符号可视作语句的标记。

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

6.1 语料库描述

字段	说明	备注
名称	用于交易的语料库的名称，应突出语料库内容和属性等。	
语言对	语料库涵盖的语种。语言对描述应依据 ISO 639 规定的语言代码与 ISO 3166 规定的国家（区域）代码的组合。	例如 en-US 表示美式英语、en-BR 表示英式英语、zh-CN 表示简体中文等。
领域	语料所属的专业领域。语料所属领域可由语料库提供方自定义。可包含一个或多个标签。	例如“IT”（IT 相关的文本）、“法律”（法律相关的文本）、“IT-法律”（与 IT 相关的法律文本）。
库容	语料库所含语料的句对总数或形符总数。	
对齐单位	语料对齐的级别，包括： 1.词汇对齐，即一个词汇对一个翻译对应词； 2.语块对齐，即一个语块对一个翻译对应语块； 3.句对齐，包括一对一对齐、一对多对齐、一对零对齐和多对一对齐等； 4.段落对齐，即一段原文对一段译文； 5.篇章对齐，即一篇原文对一篇译文。	建议采用句对齐。
格式	语料库采用的数据存储方式。语料库格式包括： 1.TXT（纯文本格式）； 2.CSV（逗号分隔格式）； 3.TSV（制表符分隔格式）； 4.TMX（一种符合 TMX 1.4b 规范的 xml 文件格式）。	无论 CSV、TSV 或 TMX，均建议采用 UTF-8 编码。
语料来源	语料库所含语料的来源，包括： 1.翻译，即通过翻译活动产生；	

	2.互联网，即从互联网上收集得到； 3.第三方，即从第三方获得，比如客户提供、从他人购买等； 4.未知，即无法确定数据来源； 5.国际机构共享； 6.上述五种来源的组合，例如“翻译+互联网”，表示部分语料由翻译产生，部分来自互联网。	
版权类型	根据《保护文学艺术作品伯尔尼公约》，语料库的版权情况包括： 1.完全版权，即原文和译文均取得版权； 2.部分版权，即仅部分内容取得版权； 3.无版权，即可以明确原文或译文均无版权； 4.版权未知，即无法确定是否有原文或译文版权。	
脱敏程度	语料库产品的脱敏程度包括： 1.完全脱敏，即语料库中的数据不包含任何敏感信息； 2.部分脱敏，即仅对语料库中的部分数据进行了脱敏处理； 3.未脱敏，即语料库中的语料数据未经过任何脱敏处理。	
购买方免责声明	语料库提供方应声明其提供的语料库因版权或隐私问题给语料库购买方带来的一切法律责任，均由语料库提供方承担。	

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。

6.2.2 语料评价维度

语料评价维度主要包括四个维度，即语料获取难易度、语料加工程度、语料

质量和语料应用前景。

语料获取难易度取决于语料来源的不同。语料加工程度是指语料的标注以及对齐的单位。

语料质量因语料具体用途的不同而不同。

语料应用前景是指语料适用的范围和具体领域以及语料的兼容性和适用性。

6.2.3 评价方法

采用分类抽样方法对语料质量进行评价，包括自动评价和人工评价，其中人工评价又包含专家评价和用户评价两种评价方法。

6.2.3.1 自动评价

采用自动化评价方法进行评价。

6.2.3.2 人工评价

通过人工方式进行评价，分为专家评价和用户评价。

（1）专家评价

依据相关的技术指标，对语料库的设计、建设过程以及语料库类型、用途、性能和语料质量等方面进行评价。

（2）用户评价

用户对语料库的功能、性能、可靠性和适用性等进行测试与评价。

6.2.4 评价流程

（1）评价组织的建立

评价组织应由语料库相关领域专家和用户代表组成。

（2）拟定评价计划

评价计划包括评价的目的、方法、参评人员、评价流程以及评价结果的应用等。

（3）评价细则制定

根据语料库评价的原则和方法，制定评价标准和具体细则等文件。

（4）评价的实施

根据语料库评价的标准和方法，组织专家和用户对语料库进行抽样、测试与评价。

（5）评价报告的形成

根据专家和用户对语料库质量的评价结果，形成语料库评价报告。

（6）评价结果的应用

根据语料库评价报告，语料库建设人员可对语料库进行调整、改进和补充，以提高其性能与效益。

6.2.5 双语语料评价标准

语料质量由翻译质量、对齐质量和数据质量共同决定，其中翻译质量的权重f(1)最高，可设置在 0.7 以上；对齐质量权重 f(2)及数据质量权重 f(3)可根据实际情况设置在 0~0.2 之间，设置权重时应注意 f(1)+f(2)+f(3)三项之和为 1，语料质量计算公式如下：语料质量得分 = 翻译质量* f(1)+ 对齐质量* f(2)+ 数据质量* f(3)。

语料质量可根据其得分情况划分为以下五个等级。详见表 1：

表 1　语料质量得分标准

语料质量等级	得分
A	90 - 100
B	80 - 89
C	60 - 79
D	40 - 59
E	0 - 39

（1）翻译质量依据“翻译服务译文质量要求”［源自：GB/T 19682-2005］）中的译文综合差错率进行评价。

根据译文综合差错率将翻译质量划分为五个等级，不同等级对应不同得分。详见表 2：

表 2　语料翻译质量打分标准

翻译质量等级	译文综合差错率	得分
A	0‰ - 0.5‰	90 - 100
B	0.5‰ - 1‰	80 - 89
C	1‰ - 1.5‰	70 - 79
D	1.5‰ - 5‰	60 - 69

	>5‰	0 - 59

（2）对齐质量由双语句对的匹配程度（原文与译文在语义上匹配）以及对齐单位一致性（语料实际对齐单位与“对齐单位”字段的描述一致）决定，可分为五个等级。不同等级对应不同得分。详见表 3：

表 3　语料对齐质量打分标准

对齐质量等级	评判标准	得分
A	原文与译文的对应率及对齐单位合规率的平均值介于 90% - 100%之间。	90 - 100
B	原文与译文的对应率及对齐单位合规率的平均值介于 80% - 89%之间。	80 - 89
C	原文与译文的对应率及对齐单位合规率的平均值介于 60% - 79%之间。	70 - 79
D	原文与译文的对应率及对齐单位合规率的平均值介于 40% - 59%之间。	60 - 69
E	原文与译文的对应率及对齐单位合规率的平均值介于 0% - 39%之间。	0 - 59

（3）数据质量由语料清洗程度及领域一致性（语料实际领域与“领域”字段的描述一致）决定，可分为五个等级。不同等级对应不同得分，详见表 4。

表 4　语料数据质量打分标准

数据质量等级	评判标准	得分
A	语料清洗程度及领域一致性的平均值介于 90% - 100%之间。	90 - 100
B	语料清洗程度及领域一致性的平均值介于 80% - 89%之间。	80 - 89
C	语料清洗程度及领域一致性的平均值介于 60% - 79%之间。	70 - 79
D	语料清洗程度及领域一致性的平均值介于 40% - 59%之间。	60 - 69
E	语料清洗程度及领域一致性的平均值介于 0% - 39%之间。	0 - 59

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。

附录：参考文献

胡开宝. 语料库翻译学概论[M]. 上海：上海交通大学出版社，2011.

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》）.

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》.

中国国家标准 GB 13715《信息处理用现代汉语分词规范》.

中国国家标准 GB/T 19682-2005《翻译服务译文质量要求》.

BAKER P. Querying keywords: questions of difference, frequency and sense in keywords analysis[J]. *Journal of English Linguistics*, 2004, 32(4): 346-359.

SINCLAIR J. The search for units of meaning[J]. *TEXTUS*, 1996, 9(1): 75-106.

SINCLAIR J. *Corpus, concordance, collocation*[M]. Oxford: Oxford University Press, 1991.

STUBBS M. Two quantitative methods of studying phraseology in English[J]. *International Journal of Corpus Linguistics*, 2002, 7(2): 215-244.