ICS 03.080.01
CCS A 10

团　　体　　标　　准

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

Translations of Chinese key terms and expressions—Corpus construction—Part 3: Sampling inspection

2022-04-01 发布　　2022-05-01 实施

中国翻译协会　　发 布

目　　次

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3部分。T/TAC 7已经发布了以下部分：

——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、天津工商大学、传神语联网网络科技股份有限公司、中国翻译协会、中国质量标准出版传媒有限公司、对外经济贸易大学、北京悦尔信息技术有限公司。

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

中国特色话语翻译　高端语料库建设
第3部分：抽样检验

1　范围

本文件规定了利用抽样统计理论对中国特色话语翻译高端语料库（以下简称“语料库”）质量合格情况进行检验评估的基本思想、抽样方案的关键参数、抽样检验流程、检验风险处理等。

本文件适用于在无法对语料库所有内容质量进行全面检验时对语料库建设进行质量控制、验收评价、等级评定、资源管理等工作，其他语料库建设管理工作可参照使用。

本文件不包括语料库的具体检验项目、质量要求、检验技术、检验工具等。

注1：本文件所给出的对语料库进行抽样检验的程序规则，不涉及具体检验的特性、要求等，抽样检验结果以“合格/不合格”“满足/不满足”“正确/不正确”“接收/不接受”等形式给出，而不是“好/坏”或“一级/二级”等。

注2：对于语料质量的具体要求（如译文是否正确、标注是否规范等）及判别标准等，是由每个具体的语料库建设或检验任务来确定的，不在本文件范围内。

注3：无法对语料库所有内容质量进行全面检验，可能是由于语料库规模太大、无法通过软件工具自动进行，也可能是由于时间紧或成本高而无法全面展开等。

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/T 2828.1—2012　计数抽样检验程序　第1部分：按接收质量限（AQL）检索的逐批检验抽样计划

GB/T 2828.2—2008　计数抽样检验程序　第2部分：按极限质量（LQ）检索的孤立批检验抽样方案

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

3　术语和定义

T/TAC 7.1—2021界定的以及下列术语和定义适用于本文件。

3.1

检验　inspection

为确定语料库或语料批是否达到质量要求而进行的检查、测试活动

3.2

初次检验　original inspection

按照本文件的规定对批（3.3）进行的第一次检验（3.1）

注：对于已判定为不符合质量要求的批，经修改后再次提交检验，不属于初次检验。

[来源：GB/T 2828.1—2012，3.1.2，有修改]

3.12

接收质量限　acceptance quality limit

当一个连续系列批(3.3)被提交验收抽样时，可容忍的最差过程平均(3.11)质量水平(3.10)

[来源：GB/T 2828.1—2012，3.1.26]

4　符号和缩略语

下列符号和缩略语适用于本文件。

Ac：接收数(Accept)

AQL：接收质量限(Acceptance Quality Limit)

N：批量

n：样本量

Re：拒收数(Reject)

5　基本思想

本文件第6章、第7章、第8章所给出的抽样检验方法主要是从过程管理、语料库建设和使用双方风险控制的角度，从待检验的语料库中分批(批次不宜过少)、依次、随机抽取部分语料记录(即样本)进行检查和测试，根据语料库合格率(或不合格率)要求、语料总体规模、抽样过程、样品中不合格品的数量等信息，分析判断语料库整体质量是否达到指定合格率(或不合格率)的一种方法。检验过程中，应根据检验结果的好坏，动态调整检验判断规则。

若对语料库不分批或分成很少的批次进行检验，宜按照GB/T 2828.2—2008进行抽样检验。

注：本文件不是计算语料库实际合格率(或不合格率)到底是多少，而是判断其质量是否满足某个合格率(或不合格率)要求。其思想是，只要满足合格率要求即可判为质量合格，从而减少样本量和计算复杂度。

由于抽样的随机性，对语料库质量的判断会存在一定误判或漏判风险。通过合理确定抽样参数等措施可以有效降低这些风险。

6　抽样方案的关键参数

6.1　批和批量

总体而言，批和批量的确定主要和语料库建设人员专业水平、语料库建设流程、质量要求高低、语料库错误影响程度、管理工作量大小、检验时间要求、成本要求等因素有关。

总体而言，在语料规模相同的情况下，增加批次，检验工作的误判和漏判风险会下降，检验所需总样本量会增加，检验总成本、管理成本也会增加。

6.2　检验水平

检验水平反映了对抽样检验工作质量管理要求的严格程度。本文件中，检验水平从低到高依次为：水平Ⅰ、水平Ⅱ和水平Ⅲ。

如希望漏判风险和误判风险较低、提高鉴别能力、提高质量管理水平，则需要较高的检验水平，需要的样本量也较多；如对漏判风险、误判风险、鉴别能力等要求不高，则可降低检验水平，以减少样本量。

6.3　AQL

本文件中，AQL是语料库所有方所允许的最大不合格率，以不合格记录所占百分比来表示。

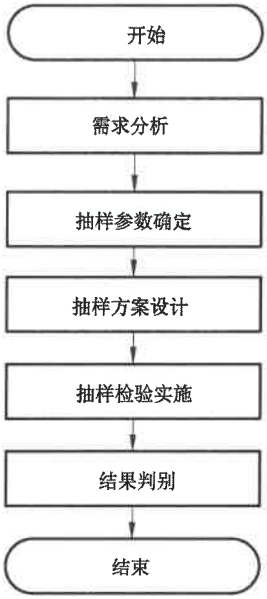

图 1　抽样检验流程

7.2　需求分析

结合具体语料库内容、检验要求、建设要求、应用需求、管理成本等信息，从以下方面综合分析语料库抽样检验工作具体要求：

a）　质量水平要求；
b）　语料库规模；
c）　语料库结构；
d）　语料库建设流程；
e）　单个样本检验工作难度；
f）　检验工作误判和漏判可能带来的风险和危害；
g）　经费、时间及其他检验工作所依赖的资源限制。

7.3　抽样参数确定

7.3.1　批和批量的确定

根据 T/TAC 7.1—2021，语料库通常由多个表组成，每个表存储不同类型的数据，不同表的结构可能不同，每个表内存储多条语料数据。当语料库规模较大或持续建设时，可按照下列步骤（见图 2）将语料库分解为多个待检验批之后再进行分批检验。

求和确定的参数，按照 GB/T 2828.1—2012 中 10.3 的要求，检索、确定抽样方案。示例见附录 A。

7.4.2 AQL 值不可直接检索到

当已确定的 AQL 值不能直接在 GB/T 2828.1—2012 的表 2 和表 3 中检索到时，可按照 7.4.1 的要求，从本文件的表 2 中按与原 AQL 相近的 AQL 进行检索，也可根据图 3 和本文件的表 2 计算求得抽样方案。

图 3 利用 AQL 计算抽样方案流程

表 2 抽样方案计算汇总表

方案类型	样本量系数	累计样本量系数	AQL—正常检验											
			12.5/n	50/n	80/n	125/n	200/n	315/n	—	500/n	—	800/n	—	1 250/n
			Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re
一次	1	1	0 1	1 2	2 3	3 4	5 6	7 8	8 9	10 11	12 13	14 15	18 19	21 22
二次	0.63	0.63	*	0 2	0 3	1 3	2 5	3 6	4 7	5 9	6 10	7 11	9 14	11 16
	0.63	1.26		1 2	3 4	4 5	6 7	9 10	10 11	12 13	15 16	18 19	23 24	26 27
			20/n	80/n	125/n	200/n	315/n	—	500/n	—	800/n	—	1 250/n	—
			接收质量(AQL)——加严检验											
注：* 表示采用一次抽样方案。														

7.5.4 转移规则及结果判别

对连续各批进行检验和判别时，应遵循图 4 所给出的转移规则。

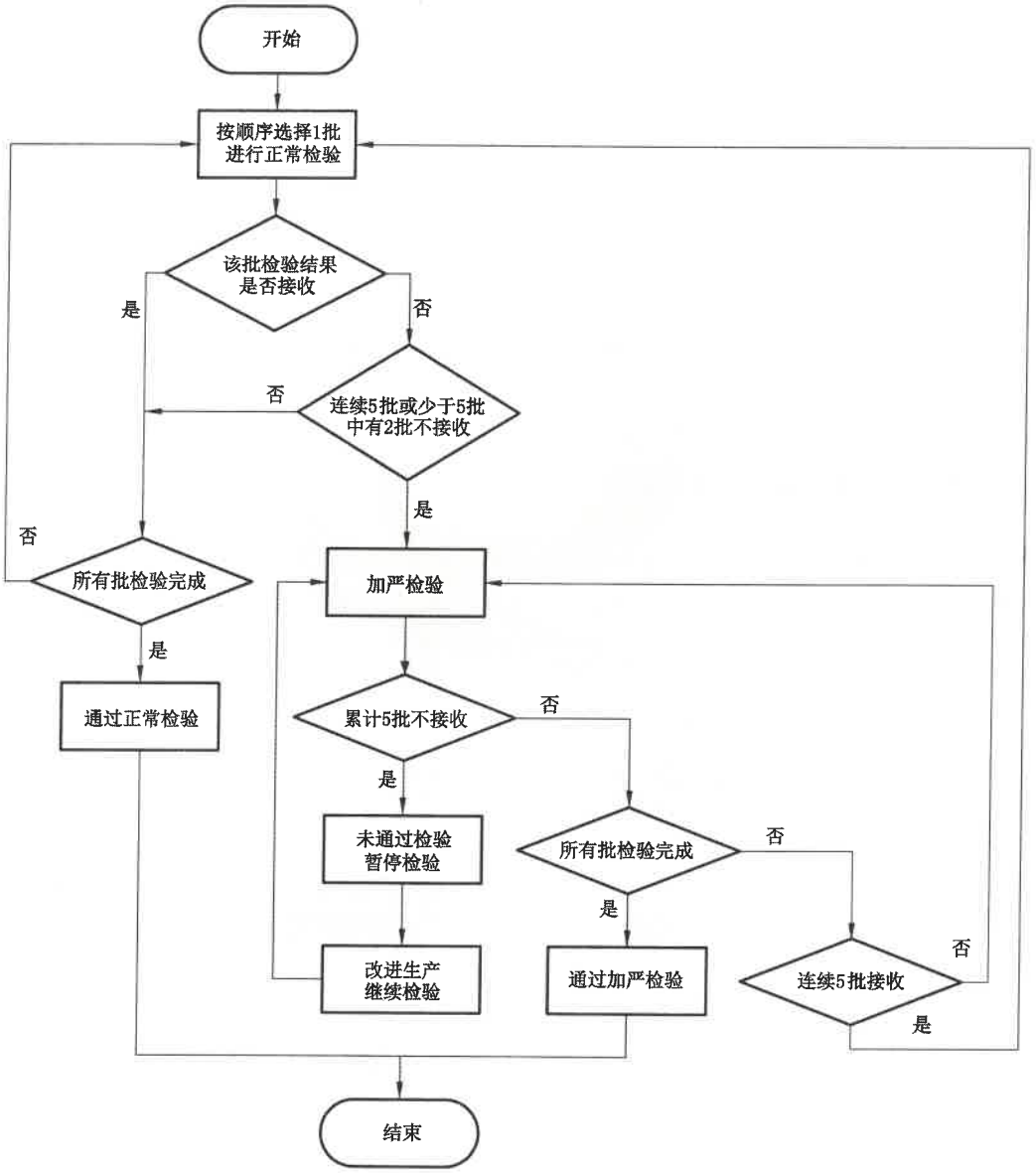

图 4 转移规则

加严检验时，应按照 GB/T 2828.1—2012 中的表 2-B、表 3-B 检索抽样方案并执行相应的检验。

对于检验中不接收的批，宜对其错误进行修改，并再次提交检验。

7.5.5 暂停检验及改进生产

如因不接收批次数量累计达到上限，应暂停检验，并对语料库建设过程进行全面检查，分析不合格记录产生原因，改进语料库建设工作。

7.6 批的再提交

对于拒收批，如果可能，应对整批进行筛选，剔除所有不合格记录，并对不合格记录加以修改或者替换（剔除的不合格记录经过修改校正后可选择归入原批），然后再次组批交验，即再次提交。再次提交

附 录 A
（资料性）
抽样方案检索示例

A.1 AQL 值可直接检索到

对某语料库进行连续分批的抽样检验，根据实际情况确定的抽样参数为：批量为 5 000，采用检验水平Ⅱ，AQL 值为 0.1，使用一次抽样方案进行检验，初次检验时使用正常检验。抽样方案检索步骤如下：

a) 检索样本量字码，依据 GB/T 2828.1—2012 中的表 1，根据批量 5 000 所在的行和检验水平Ⅱ所在的列找到交叉位置的样本量字码为 L（如图 A.1 所示）；

b) 检索正常检验的抽样方案，依据 GB/T 2828.1—2012 中的表 2-A，根据样本量字码 L 所在行和 AQL 值 0.1 所在列交叉位置寻找抽样方案，查表可知此交叉位置为向上箭头（如图 A.2 所示），根据 GB/T 2828.1—2012 选择箭头所指向的上一行的方案作为抽样方案，因此确定的抽样方案为（125，0，1）；

c) 按照抽样方案（125，0，1），抽取 125 个样本进行检验，接收数为 0，拒收数为 1，即：如果未发现不合格则该批检验通过，如果发现有 1 项检验不合格则该批不能通过检验。

在连续批的检验中，如果出现连续 5 批或少于 5 批中有 2 批为不接收，则实施加严检验。加严检验的抽样方案依据 GB/T 2828.1—2012 中的表 2-B 进行检索，检索方法同正常检验。

批 量	特殊检验水平				一般检验水平		
	S-1	S-2	S-3	S-4	Ⅰ	Ⅱ	Ⅲ
2～8	A	A	A	A	A	A	B
9～15	A	A	A	A	A	B	C
16～25	A	A	B	B	B	C	D
26～50	A	B	B	C	C	D	E
51～90	B	B	C	C	C	E	F
91～150	B	B	C	D	D	F	G
151～280	B	C	D	E	E	G	H
281～500	B	C	D	E	F	H	J
501～1 200	C	C	E	F	G	J	K
1 201～3 200	C	D	E	G	H	K	L
3 201～10 000	C	D	F	G	J	L	M
10 001～35 000	C	D	F	H	K	M	N
35 001～150 000	D	E	G	J	L	N	P
150 001～500 000	D	E	G	J	M	P	Q
500 001及以上	D	E	H	K	N	Q	R

图 A.1 样本量字码检索示意图

b) 根据批量 3 000，检验水平Ⅱ，检索 GB/T 2828.1—2012 中的表 1 得到样本量字码为 K，再根据 GB/T 2828.1—2012 中的表 3-A，可知样本量字码为 K 对应的样本量为(80，80)，把(80，80)作为参照样本量；

c) 从表 A.2 中可知，样本量 (63，63)与参照样本量 (80，80)最接近，因此选定样本量(63，63)；

d) 根据表 2，样本量(63，63)对应的列与二次抽样方案对应的行交叉处的接收数和拒收数分别为(0，2)和(1，2)，因此抽样方案为(63，63 | 0，2；1，2)。

加严抽样检验方案的获得方法与正常方案类似，计算样本量时应选择表 2 最后一列值进行计算。

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3部分：抽样检验
T/TAC 7.3—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 1.25 字数 36千字
2022年3月第一版 2022年3月第一次印刷
*
书号：155066·5-4238 定价 31.00 元

T/TAC 7.3-2021

码上扫一扫 正版服务到