基于大规模网络语料的藏文音节拼写错误统计与分析

刘汇丹，洪锦玲，诺明花，吴健

（中国科学院软件研究所，北京 100190）

摘要：针对从互联网获取的一份包含 19 万藏文网页，总计 427 万句、9328 万音节字的藏文文本语料，本文按照预定的规则对其中的藏文音节拼写错误情况进行了统计与分析。数据显示，在语料中出现的共计 20743 个藏文音节中，含有拼写错误的音节共有 9807 个，占藏文音节总数的 47.2786%，错误音节在语料中共出现 28472 次，仅占 0.0320%，说明这份语料的文本质量是相当高的。文中还详细统计了各种不同表现形式的错误音节所占比重，并分析了导致拼写错误的四个主要原因：一是输入了多余的元音符号；二是音节点或句尾空格缺失；三是同一字丁/字符存在多种表达形式；四是错误地使用了相似字符。

关键词：藏文拼写检查；拼写检查;语料；统计;藏文信息处理;中文信息处理

中图法分类号：TP391　　**文献标识码**：A

Statistics and Analysis on Spell Errors of Tibetan Syllables Based on a Large Scale Web Text Corpus

LIU Huidan, HONG Jinling, NUO Minghua, WU Jian

(Institute of Software, Chinese Academy of Sciences, Beijing, 100190)

Abstract: In this paper, a large scale corpus is built, which has about 190 thousand documents, 4.27 million sentences or 93 million syllables Tibetan text in total. Some predefined rules are used to check whether there is spell errors in each syllable. The statistical data show that there are 9807 misspelt ones out of the 20743 Tibetan syllables occurred in the corpus, which shares 47.2786%. But their occurrence is 28472, which shares only 0.0320%. It shows that the corpus has a very high quality. The different types of spell errors are also analyzed. It shows that there are mainly four causes leading to those spell errors: extra vowel sign(s); absence of syllable delimiter or sentence delimiter; characters which can be written in different forms; similar characters.

Key words: Tibetan spell check; spell check; corpus; statistics; Tibetan information processing; Chinese information processing

1、 引言

文本校对是自然语言处理的主要应用领域之一，近些年来，已有学者在藏文文本校对或拼写检查方面做了一些研究，这些研究大多针对实现藏文文本校对工具，以及为实现校对工具而构建的藏文音节规

收稿日期：2014-04-23　　**定稿日期**：2014-07-23

基金项目：核高基重大科技专项经费资助（2012ZX01039-004），国家自然科学基金资助（61202219，61202220，61303165），中国科学院信息化专项经费资助（XXH12504-1-10），新闻出版重大科技工程经费资助（0610-1041BJNF 2328/23，0610-1041BJNF2328/26）

作者简介：刘汇丹(1982－)，男，博士，工程师，主要研究方向是操作系统中文信息处理、多语言信息处理；洪锦玲(1981－)，女，硕士，工程师，主要研究方向为多语言信息处理；诺明花(1981-)，女，博士，助理研究员，主要研究方向为多语言信息处理。

则等相关知识库等方面。针对真实文本语料库的藏文拼写错误情况的统计分析工作鲜有报道。本文将通过对大规模网络藏文文本语料库中拼写错误情况进行统计分析，一方面考察真实文本中藏文拼写错误的严重程度，为藏文文本校对的研究提供依据；另一方面考察网络语料的质量，确定将网络藏文文本作为构建高质量藏文文本语料库的可靠性。

本文接下来的部分首先介绍相关领域研究现状，其次介绍大规模网络藏文文本获取的方法及利用这种方法获取的语料情况，然后对这份语料中藏文音节的拼写错误情况进行统计与分析，最后对全文进行总结。

2、 研究现状

有关藏文文本校对方面的研究可追溯到上个世纪。1998 年，扎西次仁归纳总结了藏文的拼写规则和虚词使用法则，创建了一个藏文音节库和一个藏文词表，根据藏文的拼写规则、虚词使用法则、音节库和词表，设计并开发了一个基于 DOS 的藏文拼写检查系统，并分析了拼写检查中由实词虚词兼类、词语组合型切分歧义等导致的难点问题[1]。

2002 年王维兰等人首次将藏文自动校对应用于藏文文字识别的后期处理，对识别后所形成文本中的单字进行校正[2]。

2005 年，才让卓玛提出了藏文计算机校对的两种方法：其一是将文本转换成语音，采用人听的方式发现拼写错误；其二是构建词语搭配关系表、语法规则库等知识库，利用这些知识库进行校对[3]。在后续的研究中，才让卓玛对藏文语序错误、标点使用错误、词语搭配错误等情况进行了举例分析[4]。

2009 年，刘文香采用统计方法对藏语音节的搭配规则等做了研究，创建了音节及音节搭配规则等多种知识库，以这些知识库为基础进行音节模式匹配查错，设计了音节规则模型与音节库模式匹配方法相结合的音节级查错进行藏文文本校对的原理、关键技术及可行的实现方法[5]。随后，刘文香又提出了一种将分词词表模式匹配、二元词词邻接矩阵和词间音势约束模型三种方法相结合的藏文词校对模型[6]。在上述研究的基础上，刘文香在 Windows 8 操作系统支平台上设计并实现了基于音节的现代藏文文本校对的试验系统[7]。

2009 年，多杰卓玛对藏文文本中的错误情况进行了分析，将藏文文本的错误形式归纳为音节错误、缺字和加字的错误、输入错误、人名错误、地名错误、江河名错误、知识性错误等类别，并提出了利用以字丁为单位的 N 元文法模型判断藏文音节是否错误的方法[8]。

2011 年，关白深入分析了现代藏文自动校对的研究现状[9]，分析了藏文音节字中的错误类型，并针对藏文音节字的特点，通过音节字预处理、字表匹配、混淆集匹配、二元接续关系、最小编辑距离法等方法对现代藏文音节字的自动校对进行了详细论述[10-11]。

2013 年，安见才让提出了一种对藏字的字母组合进行分段处理，根据各分段部分的构字规则进行藏字校对的方法，该方法不使用藏字字典和大规模语料库。实验表明，在一段约 130 个字符的文本中，系统成功检测出了其中的 6 处错误[12]。

2013 年，珠杰等人构建了现代藏文音节规则库，并分析了其在拼写检查、词语排序、信息抽取和文本挖掘等方面的应用[13]。在对三句藏文语料的测试中发现该模型还需要增加对藏文数字、符号、特殊音

节、梵音转写音节的特殊处理。

2013 年，洪锦玲等人综合藏文分词、音节拼写、格助词规则等多种藏文特性，提出了一种藏文词语拼写检查的方法，并提出了根据错误词语与词库词语的编辑距离给出纠错建议的方法。他们将该方法在开源办公套件 LibreOffice 中进行了实现[14]。

2014 年，陈小莹等人设计实现了一个包括藏文文本规范化处理模块、音节切分模块、黏着语的分离与还原模块和音节校对模块四个模块的藏文音节拼写自动校对系统[15]。

上述研究大多针对实现藏文文本校对工具及藏文音节规则等相关知识库的构建方面，只有多杰卓玛、关白等对藏文拼写错误情况进行了归纳，但也仅限于对个别情况的举例说明。针对真实文本语料库的藏文拼写错误情况的统计分析工作还未见有报道。本文将通过对大规模网络藏文文本语料库中拼写错误情况进行统计分析，一方面考察真实文本中藏文拼写错误的严重程度，为藏文文本校对的研究提供依据；另一方面考察网络语料的质量，确定将网络藏文文本作为构建高质量藏文文本语料库的可靠性。

3、 语料获取与处理

本节介绍大规模藏文网络文本的获取、音节切分方法和音节拼写错误的判别依据等方面的内容。

3.1 语料来源

表 1 八个新闻广播类藏文网站的基本信息

序	网站域名	网站名称	编码字符集
1	tb.chinatibetnews.com	中国西藏新闻网	扩充集
2	tb.tibet.cn	中国西藏网	基本集
3	ti.gzznews.com	康巴传媒网	基本集
4	ti.tibet3.com	中国藏族网通	基本集
5	tibet.people.com.cn	人民网藏文版	同元编码
6	www.qhtb.cn	青海藏语广播网	基本集
7	www.tibetcnr.com	中国藏语广播网	基本集
8	xizang.news.cn	新华网西藏频道藏文版	扩充集

根据我们之前对互联网藏文文本资源分布情况的考察，我们选择了八个新闻广播类的藏文网站作为文本语料的来源，这八个网站的基本信息如表 1 所示。八个网站中，中国西藏新闻网和新华网西藏频道藏文版使用国家标准藏文编码字符集扩充集，人民网藏文版使用同元编码，这三个网站的藏文文本需要做编码转换。其它五个网站均使用国际标准 Unicode 藏文基本集（小字符集）方案。在进行后续处理之前，我们将获取的语料统一转换为国家标准藏文编码字符集基本集形式（关于藏文编码转换技术请参考文献[16-17]）。编码转换过程使用了与“藏码通”相同的编码对照表和转换算法[17]。“藏码通”软件在民族出版社、社科院民族所、西藏大学、西藏编译局等单位使用近十年，并根据用户反馈情况对编码对照表进行了反复修改，因此，转换正确率是可以保证的。同时，我们对语料来源所属的网站频道进行了限制，并通过网页文种识别限定只取藏文网页，并只抽取其中的标题、正文等关键信息。以上可以最大限度地避免语料因编码转换导致的问题。

3.2 语料获取方法

在本文中，我们采用基于正则表达式的方法从藏文网页中抽取文章主题相关的信息。我们通过分析

各个网站的页面布局结构来抽取网页模板，根据之前相关的研究，分析藏文网页的板式结构，可以发现文章标题、作者、发布时间、文章正文等信息块与其他信息块之间的分隔标志，甚至可以利用 HTML 源文件中的一些注释信息进行抽取[18]。可以据此构造模板提取藏文篇章文本，举例如下：

- 中国西藏新闻网的页面模板为：

.*<!—enpproperty【文章 ID、作者、标题、副标题、所属栏目】/enpproperty--><!--enpcontent--><!--enpcontent-->【文章正文】<!--/enpcontent--><!--/enpcontent-->.*

- 中国西藏网的页面模板为：

.*<meta name="description" content="【标题】" />.*<meta name="catalogs" content="【类别编号】">.*<meta name="contentid" content="【文章 ID】">.*<meta name="publishdate" content="【发布时间】">.*<meta name="author" content="【作者】">.*<meta name="source" content="【来源】">.*<!--content-->【文章正文】<!--/content-->.*

3.3 音节切分方法

对藏文文本进行音节切分主要依据如下切分规则：

- 音节点作为音节分隔标记，切分之后附着在左边（前边）音节的结尾；
- 藏文数字和阿拉伯数字视为音节分隔标记，切分之后分别视同藏文音节参与数据统计；
- 藏文标点符号、英文标点符号和汉语标点符号视为音节分隔标记，切分之后分别视同藏文音节参与数据统计；
- 连续的英文字母视为音节分隔标记，切分之后视同藏文音节参与数据统计；
- 连续的汉字视为音节分隔标记，切分之后视同藏文音节参与数据统计。

表 2　网络语料中的藏文高频音节表

序	音节	频次	比例	序	音节	频次	比例
1	།*	4063168	4.3559%	11	ཐྱི་	852942	0.9144%
2	དུ་	2353795	2.5234%	12	གི་	742822	0.7963%
3	་	2178344	2.3353%	13	ཡོད་	737211	0.7903%
4	ཀྱི་	1354111	1.4517%	14	ཚོགས་	722484	0.7745%
5	པའི་	1294409	1.3877%	15	མི་	705641	0.7565%
6	ལ་	1008904	1.0816%	16	ཐུབ་	671467	0.7198%
7	ལས་	994194	1.0658%	17	ཏུ་	662486	0.7102%
8	བ་	973837	1.0440%	18	དེ་	565602	0.6063%
9	ནས་	955554	1.0244%	19	གའི་	551627	0.5914%
10	ཞིང་	906639	0.9720%	20	ཚེན་	539562	0.5784%

根据以上规则对获取到的网络藏文文本进行切分之后，可以统计各个藏文音节出现的频次。在上述语料中，出现频率最高的部分藏文音节如表 2 所示。

3.4 语料规模

使用上述方法获取网络藏文文本语料，并进行音节切分，统计数据显示，共计 19 万藏文网页，语料总计 427 万句、9328 万音节字（含藏文数字、汉字、英文字母、各种标点符号等）。详细的统计数字见表 3。

---

* 理论上讲，藏文单垂符是标点符号，但为反映语料真实情况，所有的标点符号、数字、汉字串、英文单词等视同“藏文音节”均参与数据统计。

表 3　获取的网络藏文文本语料的规模

	网站	网页数	段落数	句数	音节数
1	tb.chinatibetnews.com	74632	387519	1442173	38650978
2	tb.tibet.cn	13348	58540	328928	5880906
3	ti.gzznews.com	8084	61377	278138	4764235
4	ti.tibet3.com	26631	161238	736112	12701407
5	tibet.people.com.cn	29797	134541	447754	11425653
6	www.qhtb.cn	20616	125327	574326	10915022
7	www.tibetcnr.com	9559	73715	277685	4625679
8	xizang.news.cn	7707	49007	186715	4316357
合计		190374	1051264	4271831	93280237

4、 拼写错误的统计与分析

4.1　藏文音节拼写错误的判别依据

在藏文音节拼写检查的研究中，大家常用的方法是根据藏文文法中基字、前加字、上加字、下加字、元音、后加字和再后加字之间的约束关系构造藏文音节规则库来判断音节的合法性，然而，由于梵音转写和外来词音译的存在，采用这种方法构建的规则库总是不能完全覆盖真实文本中所有的情况。因此，在本文中，我们根据传统藏文文法构造一些规则来判别音节是否存在拼写错误，这些规则主要包括：

- 包含多个藏文元音符号的音节视为拼写错误。但由འཇུག་རྟགས་（紧缩标志）构成的合法紧缩音节例外。由相邻两个 ི 或 ུ 重叠构成的长元音视为一个元音。
- 包含多个紧缩标志的音节视为拼写错误；
- 紧缩标志出现在第四字丁或更靠后位置的音节视为拼写错误；
- 包含五个或更多个字丁的音节视为拼写错误；
- 包含在国家标准藏文基本集、扩充集 A 和扩充集 B 以外字丁的音节视为拼写错误。
- 前加字、上加字、基字、下加字、后加字和再后加字之间搭配不符合藏文文法约束关系的视为拼写错误。

为确保上述规则包容梵音转写和外来词音译形成的音节，达到对真实语料形成完全覆盖的目的，我们的检测规则中充分考虑了梵音转写和外来词音译的情况。由于约束关系检测方法不能保证百分之百的正确率，我们对被该规则判断为存在拼写错误的情况进行了人工确认。

为了考察各种不同表现形式的拼写错误的情况，在设计上述规则时未考虑规则的互斥性。由于在真实文本中，一个藏文音节存在多种拼写错误形式是可能的，因此，这样的设计符合实际情况。除上述规则之外，我们还对对包含紧缩标志འཇུག་རྟགས་的音节做出标记。

4.2　对拼写错误的统计与分析

本文所用语料中，共有 20743 个藏文音节，总出现频次 89059463 次，占语料总量的 95.4752%。藏文数字共出现 130808 次，占语料总量的 0.1402%，两项合计占比 95.6154%，语料中另外 4.3846%是其它文种的字符串，其各自出现频次和比例如表 4 所示。

表 4　语料中各种不同成分的频次和比例

序	类别	频次	比例
1	藏文音节（含藏文标点符号）	89059463	95.4752%
2	其他标点符号	2835236	3.0395%
3	阿拉伯数字	770924	0.8265%
4	拉丁字母和单词	467992	0.5017%
5	藏文数字	130808	0.1402%
6	汉字（串）	14039	0.0151%
7	URL	1775	0.0019%
合计		93280237	100.0000%

根据前述规则，对语料中出现的所有藏文音节进行拼写检查，获得的统计数据如表 5 所示。可以看出，在这些包含拼写错误的音节中，大部分具有两个或者两个以上的表现形式，这主要是由于拼写错误判别规则之间并不是严格互斥所导致。从表 5 中可以看出，在本文所用语料中，紧缩标志位置错误也同时意味着紧缩标志太多和元音太多，而紧缩标志太多，大部分情况下也意味着元音太多。

表 5　藏文音节拼写错误情况总表

序号	正确	元音太多	紧缩标志	紧缩标志太多	紧缩标志位置错误	字丁太多	非法字丁	约束关系错误	数量	比例	频次	比例
1	Yes	No	Yes	No	No	No	No	No	1312	6.3250%	4355740	4.8908%
2	Yes	No	No	No	No	No	No	No	9624	46.3964%	84675251	95.0772%
3	No	Yes	Yes	Yes	Yes	Yes	No	Yes	4	0.0193%	5	0.0000%
4	No	Yes	Yes	Yes	Yes	No	No	Yes	12	0.0579%	371	0.0004%
5	No	Yes	Yes	Yes	No	No	No	Yes	38	0.1832%	378	0.0004%
6	No	Yes	Yes	No	No	Yes	Yes	Yes	2	0.0096%	5	0.0000%
7	No	Yes	Yes	No	No	Yes	No	Yes	182	0.8774%	367	0.0004%
8	No	Yes	Yes	No	No	No	Yes	Yes	69	0.3326%	207	0.0002%
9	No	Yes	Yes	No	No	No	Yes	No	10	0.0482%	22	0.0000%
10	No	Yes	Yes	No	No	No	No	Yes	468	2.2562%	1272	0.0014%
11	No	Yes	Yes	No	No	No	No	No	3	0.0145%	4	0.0000%
12	No	Yes	No	No	No	Yes	Yes	Yes	19	0.0916%	24	0.0000%
13	No	Yes	No	No	No	Yes	No	Yes	1222	5.8911%	2178	0.0024%
14	No	Yes	No	No	No	No	Yes	Yes	445	2.1453%	1265	0.0014%
15	No	Yes	No	No	No	No	No	Yes	2626	12.6597%	6745	0.0076%
16	No	Yes	No	No	No	No	No	No	19	0.0916%	26	0.0000%
17	No	No	Yes	Yes	No	No	No	Yes	6	0.0289%	16	0.0000%
18	No	No	Yes	Yes	No	No	No	No	5	0.0241%	35	0.0000%
19	No	No	Yes	No	No	Yes	No	Yes	33	0.1591%	49	0.0001%
20	No	No	Yes	No	No	No	Yes	Yes	13	0.0627%	27	0.0000%
21	No	No	Yes	No	No	No	Yes	No	15	0.0723%	134	0.0002%
22	No	No	Yes	No	No	No	No	Yes	175	0.8437%	437	0.0005%
23	No	No	No	No	No	Yes	Yes	Yes	5	0.0241%	6	0.0000%
24	No	No	No	No	No	Yes	No	Yes	1098	5.2934%	1817	0.0020%
25	No	No	No	No	No	No	Yes	Yes	496	2.3912%	2239	0.0025%
26	No	No	No	No	No	No	Yes	No	286	1.3788%	3490	0.0039%
27	No	No	No	No	No	No	No	Yes	2556	12.3222%	7353	0.0083%
合计									20743	100%	89059463	100%

表 6 列出了各种不同类型的拼写错误音节的数量及其在语料中的比例。在所有的藏文音节中，拼写

正确的藏文音节共有 10936 个，占 52.7214%，共出现 89030991 次，占 99.9680%。其中，含有前述四个紧缩标志的音节共有 1312 个，占 6.3250%，出现总次数为 4355740，占 4.8908%。含有拼写错误的藏文音节共有 9807 个，占 47.2786%，在语料中共出现 28472 次，占 0.0320%。错误形式最多的是约束关系类错误，共有 9469 个音节，占比 45.6491%，在语料中出现频次累计 24761 次，占比 0.0278%。其次是元音太多类错误，共有 5119 个音节，占比 24.6782%，在语料中出现频次累计 12869 次，占比 0.0144%。再次是字丁太多类错误，共有 2565 个音节，占比 12.3656%，在语料中出现频次累计 4451 次，占比 0.0050%。包含非法字丁的音节共有 1360 个，占比 6.5564%，在语料中共出现 7419 次，占比 0.0083%。紧缩标志太多的音节共有 65 个，占比 0.3134%，出现频次为 805，占比 0.0009%。紧缩标志位置错误的音节共有 16 个，占比 0.0771%，出现频次为 376，占比 0.0004%。

表 6　藏文拼写错误类型及其在语料中的比例

序	错误类别	音节数	比例	频次	比例
1	约束关系错误	9469	45.6491%	24761	0.0278%
2	元音太多	5119	24.6782%	12869	0.0144%
3	字丁太多	2565	12.3656%	4451	0.0050%
4	非法字丁	1360	6.5564%	7419	0.0083%
5	紧缩标志太多	65	0.3134%	805	0.0009%
6	紧缩标志位置错误	16	0.0771%	376	0.0004%
错误总数		9807	47.2786%	28472	0.0320%
正确总数		10936	52.7214%	89030991	99.9680%
合计		20743	100%	89059463	100%

表 7 和图 1 显示了不同错误形式在所有出错音节中的比例。

表 7　藏文拼写错误类型及其比重

序号	错误类别	音节数	比例	频次	比例
1	约束关系错误	9469	96.5535%	24761	86.9661%
2	元音太多	5119	52.1974%	12869	45.1988%
3	字丁太多	2565	26.1548%	4451	15.6329%
4	非法字丁	1360	13.8676%	7419	26.0572%
5	紧缩标志太多	65	0.6628%	805	2.8273%
6	紧缩标志位置错误	16	0.1631%	376	1.3206%
错误总数		9807	100%	28472	100%

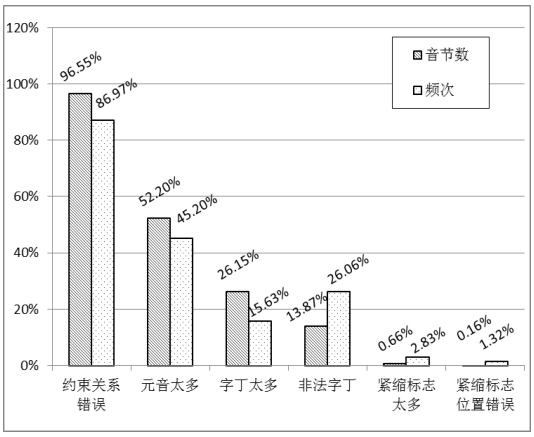

图 1　各种拼写错误表现形式所占比重

在所有的出错音节中，表现为约束关系错误的音节数量和频次占比分别达到了 96.5535%和 86.9661%，占据了出错音节的绝大部分。部分典型的拼写错误音节如表 8 所示。这些错误中，一部分是因为输入了多余的元音导致，其它大部分都是因音节点或句尾空格缺失导致。

表 8　约束关系错误音节典型实例

序	音节	Unicode 编码序列	频次
1	ཀཱཨ་	0F40 0F62 0FA8 0F0B	2070
2	ཤྲཱ་	0F64 0F71 0F40 0FB1 0F0B	1010
3	མཚོའུའི་	0F58 0F5A 0F7A 0F60 0F74 0F60 0F72 0F0B	348
4	ཡུངྷལ་	0F68 0F74 0F4F 0FA4 0F63 0F0B	141
5	མེཏྲེ་	0F66 0F7A 0F44 0F92 0F7A 0F0B	129
6	ཤྲཱུའི་	0F64 0F71 0F40 0FB1 0F60 0F72 0F0B	122
7	ཀཱཌྜ་	0F64 0FB0 0F40 0FB1 0F0B	119
8	ཀྱིཏྲེ་	0F64 0F74 0F51 0FA1 0FB7 0F7A 0F0B	108
9	འདུགའེ་	0F60 0F51 0F74 0F42 0F51 0F7A 0F0B	106
10	སྒྲིཪ་	0F62 0F92 0FB1 **0F74 0F74** 0F0B	105

在所有的出错音节中，表现为元音太多的音节数量和频次占比分别达到了 52.1974%和 45.1988%，占据了出错音节总数的过半。部分典型的拼写错误音节如表 9 所示。这些错误中，一部分是因为输入了多余的元音导致（表中第 8 个实例），其它大部分都是因音节点或句尾空格缺失导致。

表 9　元音太多的错误音节典型实例

序	音节	Unicode 编码序列	频次
1	ཤྲཱ་	0F64 0F71 0F40 0FB1 0F0B	1010
2	མཚོའུའི་	0F58 0F5A 0F7A 0F60 0F74 0F60 0F72 0F0B	348
3	ཡུངྷལ་	0F68 0F74 0F4F 0FA4 0F63 0F0B	141
4	མེཏྲེ་	0F66 0F7A 0F44 0F92 0F7A 0F0B	129
5	ཤྲཱུའི་	0F64 0F71 0F40 0FB1 0F60 0F72 0F0B	122
6	ཀྱིཏྲེ་	0F64 0F74 0F51 0FA1 0FB7 0F7A 0F0B	108
7	འདུགའེ་	0F60 0F51 0F74 0F42 0F51 0F7A 0F0B	106
8	སྒྲིཪ་	0F62 0F92 0FB1 **0F74 0F74** 0F0B	105
9	ཐུགཤེས་	0F55 0FB1 0F74 0F42 0F64 0F7A 0F66 0F0B	86
10	པཱཌེ་	0F54 0F74 0F60 0F74 0F60 0F72 0F0B	80

表现为字丁太多的音节数量和频次占比分别达到了 26.1548%和 15.6329%。部分典型的拼写错误音节如表 10 所示。这些错误中，几乎全部是因音节点和句尾空格缺失导致。

表 10　字丁太多的错误音节典型实例

序	音节	Unicode 编码序列	频次
1	འདུགཚང་	0F60 0F51 0F74 0F42 0F41 0F7C 0F44 0F0B	38
2	འདུགཚོན་	0F60 0F51 0F74 0F42 0F60 0F7C 0F53 0F0B	36
3	འདུགངས་	0F60 0F51 0F74 0F42 0F44 0F66 0F0B	27
4	འདུགའདི་	0F60 0F51 0F74 0F42 0F60 0F51 0F72 0F0B	25
5	འདུགཡིན་	0F60 0F51 0F74 0F42 0F61 0F72 0F53 0F0B	24
6	གཉིགཐེངས་	0F42 0F45 0F72 0F42 0F50 0F7A 0F44 0F66 0F0B	21
7	ཀུངདབང་	0F40 0FB2 0F74 0F44 0F51 0F56 0F44 0F0B	15
8	རིའཛོགས་	0F62 0F7A 0F60 0F72 0F5A 0F42 0F66 0F0B	15
9	གས་རའཚོད་	0F42 0F66 0F62 0F60 0F42 0F7C 0F51 0F0B	14
10	འཕྲིན་ཡུངས་	0F60 0F55 0FB2 0F72 0F53 0F63 0F9F 0F62 0F0B	14

含有非法字丁的音节数量和频次占比分别达到了 13.8676%和 26.0572%。部分典型的拼写错误音节如表 11 所示。这些错误中，大部分是因同一个字丁在 Unicode 编码框架中可以有多种表示形式或者部分相似字符所导致。例如字丁“ཐྭ”是 Unicode 中的一个字符 U+0F57，但也可以视为两个字符 U+0F56 和 U+0FB7 叠加而成。表中第 1 例、第 2 例和第 8 例均属此类。U+0F71 和 U+0FB0 的相似性、U+0F62 和 U+0F6A 的相似性也导致了拼写错误的发生，在表 11 中第 3~6 例均属于此类。表中第 7 例是因错误地将字符 U+0F97 录入为 U+0FAB 所致。还有一部分是因为存在多余的元音符号导致错误，如第 9 例和第 10 例。含有非法字丁的拼写错误中，大部分属于字符使用不规范或无规范可依所致，严格地讲，其中一些情况不能视为拼写错误。

表 11　含有非法字丁的错误音节典型实例

序	音节	Unicode 编码序列	频次
1	ཐྭ་	0F56 0FB7 0F0B	673
2	ཐྭང་	0F56 0FB7 0F44 0F0B	464
3	བཀྲན་	0F56 0F6A 0F99 0F53 0F0B	267
4	ཏྱི་	0F4F 0FB0 0F0B	200
5	ཙྱི་	0F59 0FB0 0F0B	190
6	ཁྲིད་	0F6A 0F99 0F72 0F44 0F0B	171
7	ལཕྱོངས་	0F63 0FAB 0F7C 0F44 0F66 0F0B	148
8	ཨོཾ་	0F68 0F7C 0F7E 0F0B	126
9	ཐུཿ	0F62 0F92 0FB1 0F74 0F74 0F0B	105
10	ཨ་ཅི་	0F42 0F7C 0F72 0F0B	9

含有多个紧缩标志的音节数量和频次占比分别达到了 0.6628%和 2.8273%。部分典型的拼写错误音节如表 12 所示。这部分错误基本都是因为音节点缺失导致。

表 12　含多个紧缩标志的错误音节典型实例

序	音节	Unicode 编码序列	频次
1	མཇོད་བྱའི་	0F58 0F5A 0F7A 0F60 0F74 0F60 0F72 0F0B	348
2	ཕུག་བའི་	0F54 0F74 0F60 0F74 0F60 0F72 0F0B	80
3	སྐྱེ་བའི་	0F66 0FA4 0FB2 0F7A 0F60 0F74 0F60 0F72 0F0B	64
4	རྒྱབའི་	0F45 0F74 0F60 0F74 0F60 0F72 0F0B	59
5	ཚེ་བའི་	0F63 0FA1 0F7A 0F60 0F74 0F60 0F72 0F0B	36
6	ཁེ་བའི་	0F56 0FB1 0F7A 0F60 0F74 0F60 0F72 0F0B	19
7	ཅེ་བའི་	0F56 0F7A 0F60 0F74 0F60 0F72 0F0B	18
8	ཁའི་	0F60 0F74 0F60 0F72 0F0B	14
9	རྒྱབའི་	0F67 0FB2 0F74 0F60 0F74 0F60 0F72 0F0B	14
10	ལེ་བའི་	0F63 0F7A 0F60 0F74 0F60 0F72 0F0B	13

表 13　紧缩标志位置错误的音节典型实例

序	音节	Unicode 编码序列	频次
1	མཇོད་བྱའི་	0F58 0F5A 0F7A 0F60 0F74 0F60 0F72 0F0B	348
2	མཇོད་བྱའི་	0F58 0F5A 0F7C 0F60 0F74 0F60 0F72 0F0B	5
3	མཇོད་བྱའོ་	0F58 0F5A 0F7A 0F60 0F74 0F60 0F7C 0F0B	4
4	གཅོད་བྱའི་	0F42 0F45 0F7C 0F60 0F74 0F60 0F72 0F0B	2
5	བའིའོག་	0F54 0F60 0F72 0F60 0F7C 0F42 0F0B	2
6	རྟོན་པའུའི་ཉའིའོ་	0F51 0FB2 0F7C 0F53 0F60 0F72 0F56 0F74 0F53 0F72 0F4F 0F63 0F72 0F60 0F7C 0F0B	2
7	པའིགྱུར་	0F54 0F7C 0F60 0F72 0F40 0FB2 0F74 0F60 0F74 0F0B	2
8	འོག་གྱུར་	0F60 0F7C 0F42 0F40 0FB2 0F74 0F60 0F74 0F0B	2
9	ལིའོཚའེ་	0F63 0F72 0F60 0F7C 0F5A 0F60 0F7A 0F0B	2
10	གནའིསའི	0F42 0F53 0F60 0F72 0F56 0F7C 0F60 0F72 0F0B	1

紧缩标志出现在第四个字丁或者更靠后位置的错误音节数量和频次占比分别为 0.1631%和 1.3206%。部分典型的拼写错误音节如表 13 所示。这部分错误基本都是因为音节点和句尾空格缺失导致。

综合上述拼写错误的各种情况，导致拼写错误的原因主要包括四个方面：一是输入了多余的元音符号；二是音节点、单垂符或句尾空格缺失；三是同一字丁/字符存在多种表达形式；四是使用了错误的相似字符。

5、 结束语

在本文中，我们从互联网获取了共计 19 万藏文网页，进行篇章抽取之后获得了一份总计 427 万句、9328 万音节字的藏文文本语料，按照预定的规则对其中的拼写错误情况进行了统计与分析。数据显示，在所有 20743 个藏文音节中，拼写正确的藏文音节共有 10936 个，占 52.7214%。含有拼写错误的藏文音节共有 9807 个，占 47.2786%，在语料中共出现 28472 次，占 0.0320%，这说明这份语料的文本质量是相当高的。

导致拼写错误的原因主要包括四个方面：一是输入了多余的元音符号；二是音节点和句尾空格缺失；三是同一字丁/字符存在多种表达形式；四是使用了错误的相似字符。

参考文献：

[1] 扎西次仁.一个藏文拼写检查系统的设计[C].//1998 中文信息处理国际会议论文集.1998:371～376.

[2] 王维兰，丁晓青，戴玉刚等.藏文识别后处理研究[J].术语标准化与信息技术,2002,(2):30-34.DOI:10.3969/j.issn.1007-2489.2002.02.008.

[3] 才让卓玛.藏文字自动校对系统初探[C].//第十届全国少数民族语言文字信息处理学术研讨会论文集.2005:292-294.

[4] 才让卓玛,才智杰.藏文文本自动校对系统开发研究[J].西北民族大学学报(自然科学版),2009,30(1):25-28.DOI:10.3969/j.issn.1009-2102.2009.01.007.

[5] 刘文香.藏文音节校对模型建设研究[J].西北民族大学学报（自然科学版）,2009,30(2):13-16,32.DOI:10.3969/j.issn.1009-2102.2009.02.004.

[6] 刘文香.藏文文本词校对模型研究[J].西藏大学学报（自然科学版）,2009,24(2):70-74.

[7] 刘文香.现代藏文文本校对设计方案研究[J].西藏大学学报（自然科学版）,2012,(2):66-69.

[8] 多杰卓玛.N 元模型在藏文文本局部查错中的应用研究[J].计算机工程与科学,2009,31(4):117-119,123.DOI:10.3969/j.issn.1007-130X.2009.04.035.

[9] 关白，洛藏，才科扎西等.现代藏文自动校对现状分析[J].西藏科技,2011,(8):78-80.DOI:10.3969/j.issn.1004-3403.2011.08.035.

[10] 关白.自动校对中现代藏文音节字研究[J].西藏大学学报（自然科学版）,2011,26(1):69-75.

[11] 关白,才科扎西.现代藏文音节字自动校对研究[J].计算机工程与应用,2012,48(29):151-156.DOI:10. 3778/j.issn.1002-8331.2012.29.031.

[12] 安见才让.基于分段的藏字校对算法研究[J].中文信息学报,2013,27(2):58-64.DOI:10.3969/j.issn.1003-0077.2013.02.009.

[13] 珠杰,欧珠,格桑多吉等.藏文音节规则库的建立与应用分析[J].中文信息学报,2013,27(2):103-112.

[14] 洪锦玲,刘汇丹,吴健.一种在办公套件中支持藏文拼写检查的方法[C].//第 14 届中国少数民族语言文字信息处理学术研讨会论文集,2013:116-122

[15] 陈小莹,艾金勇.藏文音节拼写自动校对系统的设计[J].语文学刊,2014,(5):31-32.

[16] 刘汇丹,芮建武,吴健等.藏文网页的编码识别与转换[C].//中文信息处理前沿进展——中国中文信息学会二十五周年学术会议.2006:573-580.

[17] 刘汇丹,诺明花,赵维纳等.藏文编码转换软件“藏码通” 的设计与实现[C].//第三届全国少数民族青年自然语言信息处理、第二届全国多语言知识库建设联合学术研讨会论文集.2010:217-221.

[18] 刘汇丹，诺明花，高墨赤等.面向新闻广播网站的藏文文本采集和语料库构建[C].//第 14 届中国少数民族语言文字信息处理学术研讨会论文集,2013:85-94

[19] 周季文.藏文拼音教材(拉萨音)[M].北京：民族出版社,1983.

[20] 胡书津.简明藏文文法[M].昆明：云南民族出版社,2000.

[21] GB16959-1997 信息技术-信息交换用藏文编码字符集——基本集[S]. 中国标准出版社, 1998.

[22] GB/T 20542-2006 信息技术-藏文编码字符集——扩充集 A [S]. 北京：中国标准出版社, 2006.

[23] GB/T 22238-2008 信息技术-藏文编码字符集——扩充集 B [S]. 北京：中国标准出版社, 2008.

[24] ISO/IEC 10646:2012 Information technology - Universal Coded Character Set (UCS) [S]. International Organization for Standardization, 2012.

[25] The Unicode Standard, Version 6.1 [S]. Mountain View, CA: The Unicode Consortium, ISBN 978-1- 936213-02-3, 2012.