doi:10.3772/j.issn.2095-915x.2023.02.011

基于敦煌古藏文语料库的字词属性统计研究

开放科学（资源服务）标识码（OSID）

三智多杰[1] 祁坤钰[1] 久仙加[2]

1. 西北民族大学中国民族信息技术研究院 兰州 730030；
2. 西北民族大学中国寓言文学部 兰州 730030

摘要：[目的 / 意义] 古藏文字符统计研究能够对机器翻译，以及从海量文本中快速定位核心内容，对情报收集工作有着重要意义。目前，藏文字符统计研究主要依据现代藏文语料库，忽视了古藏文语料库的字符统计研究。[方法 / 过程] 本文以敦煌藏文文献为主，构建了古藏文文献标注语料库。在此基础上，应用 python 语言设计出古藏文频率统计软件，对古藏文和现代藏文的元音、辅音、藏文音节频次等方面进行对比分析。[结果 / 结论] 归纳出古藏文字符的分布特征，以期为古藏文标注语料库构建和藏文文字特征研究提供参考。

关键词：敦煌古藏文文献；古藏文语料库；字符统计

中图分类号：G35

A Statistical Study of Word Attributes Based on the Dunhuang Ancient Tibetan Corpus

SANZHI Duojie[1] QI Kunyu[1] JIU Xianjia[2]

1. China Institute of Information Technology for Nationalities, Northwest Minzu University, Lanzhou 730030, China;
2. Department of Chinese Literature, Northwest Minzu University, Lanzhou 730030, China

Abstract: [Purpose/significance] The research on the statistics of ancient Tibetan characters is of great significance for machine translation, pinpointing the core content from massive texts, and intelligence collecting. At present, the research on Tibetan character statistics is mainly based on the modern Tibetan corpus, neglecting the character statistics research on the ancient Tibetan corpus. [Method/Process] Based on Dunhuang Tibetan literature, this paper constructs the annotated corpus of ancient Tibetan literature. On this basis, the software of ancient Tibetan frequency statistics is designed by python language, and the vowels, consonants and Tibetan number frequencies of ancient Tibetan and modern Tibetan are compared and analyzed. [Results/

基金项目 国家自然科学基金项目 敦煌古藏文文献中唐代汉藏文化交流研究（Z21100）；中央高校基本科研业务费专项资金 藏语句法树库构建及句法分析模型研究（31920190113）；甘肃省优秀研究生“创新之星”项目 大数据背景下敦煌藏文文献语料库字频统计研究（2022CXZX-186）。

作者简介 三智多杰（1996-），硕士研究生，研究方向为自然语言处理、敦煌古藏文数字化研究；祁坤钰（1968-），博士，教授，研究方向为自然语言处理、藏文信息处理、知识图谱，E-mail：37795386@qq.com；久仙加（1978-），博士，副教授，研究方向为敦煌学古藏文文献研究、古今藏文文学研究。

引用格式 三智多杰 , 祁坤钰 , 久仙加 . 基于敦煌古藏文语料库的字词属性统计研究 [J]. 情报工程 , 2023, 9(2): 117-127.

doi:10.3772/j.issn.2095-915x.2023.02.011

Conclusion] In order to provide reference for the construction of ancient Tibetan annotated corpus, and the study of Tibetan characters, the distribution characteristics of ancient Tibetan characters are summarized as the main content.
Keywords: Dunhuang Ancient Tibetan literature; Corpus of ancient Tibetan; Statistical characters; comparison between ancient and modern Tibetan

引言

吐蕃时期是藏语言文字形成、改革和发展的重要历史分期，敦煌藏文文献是古藏文文献的核心部分。目前，流失在海外已编目的敦煌藏文文献共计 4 967 号①，国内收藏的敦煌藏文文献 10 880 件，其中甘肃省收藏 10 340 件，占国内收藏总数的 95% 以上 [1]。这些文献不仅对宗教、历史以及医学理算等学科有着重要研究价值，尤其为语言学家研究藏语语音、词汇、句法以及汉藏语言关系等提供了重要的文献史料。

20 世纪 90 年代末开始，中国社会科学院、中国藏学、西藏大学、西北民族大学、青海师范大学等高校和科研机构，先后进行了建设藏文语料库以及开发藏文字词频统计软件。其中多拉和扎西加合著的《藏文规范音节频率词典》是目前基于语料库统计的第一个藏语音节频率词典，它从藏文字母、字丁、音节等各层面对藏文频率做出了全面的统计 [2]。此外，卢亚军的《基于大型藏文语料库的藏文字符、部件、音节、词汇频度与通用度统计及其应用研究》[3] 和才让加的《藏语语料库加工方法研究》[4] 也都是基于大规模语料对藏文字词进行研究的文书。

然而在分析现有藏语语料研究成果时，发现绝大部分语料库研究都基于近现代藏文或以藏语口语为主，很难发现以古藏文语料库为基础的相关研究文献。周季文在《论藏文的元音符号“·i”》一文中虽然用频率统计法对古藏文中的“·i”进行了计量研究，但因数据统计主要依靠人工完成，使得统计范围相对较小。2021 年三排才让的《敦煌古藏文文献的字词属性统计研究》是目前唯一一篇以古藏文语料库为基础的藏语语料库研究文献 [5]。

1 古藏文语料库与统计软件

1.1 古藏文语料库建设

敦煌古藏文文献种类繁多，藏书分布地广以及文献自身的字符特殊性等原因，对古藏文文献语料收集与整理带来了一定的困难。文中所应用的古藏文文献语料，是以法国国家图书馆和西北民族大学等合作编纂的《法国国家图书馆藏敦煌藏文文献》中的 416 篇人工录入为基础，收录了 221 个 OTDO（Old Tibetan Documents Online）②古藏文在线文献，以及 32 个分布于藏区各地的摩崖、石碑和铭文。到目前为止，语料库共收录了 680 个文本，总字数

①其中法国3375号，英国1370号，俄国214号，日本7号，美国1号。以上所述的号数与件数是不同的概念，一个号数中可能包含几件（页）。
② Old Tibetan Documents Online 网址：https://otdo.aa-ken.jp/

达近百万左右，实现了较为完整的古藏文语料样本。另外，为了凸显吐蕃时期本民族社会文书的特点，收录语料时尽量录入了古藏文文献中的社会文献，使各项主题文献达到了相对平衡。在收集文献的过程中，为保障统计结果的可信性，也最大程度上保留了原文结构。

1.2 古藏文字符统计软件的设计

敦藏文字符统计是语料库语言学的基础研究方法，敦煌古藏文字频统计不仅能对藏文信息处理提供可靠的数据支撑，对藏学以及敦煌学等学科研究也具有重要的参考价值，因此统计敦煌古藏文文献语料库首先需要设计实现古藏文字符统计软件。

现代藏文中，藏文的书写顺序或音节的组织结构是以基字为中心，由上加字、下加字、元音、前加字、后加字和再后加字七种构件以二维图形的方式构成[6]。敦煌藏文文献属于古藏文，与现代藏文具有许多相同特征。本文应用 Unicode 国际编码中的藏文基本集，结合敦煌古藏文文字特性提出了计算机统计藏文字符的方法，以音节点为切分点，应用 Python 语言设计了一种字频统计方法。

软件进行字符统计的步序如下：

➢ 对藏文元音、辅音，前加字和后加字等藏文构件列表进行赋值定义；

➢ 创建藏文字符和频度的｛“藏文字符”：“统计数值”｝统计字典；

➢ 从文件中以藏文音节点为单位读取一个藏文字；

➢ 判定读取的音节是否藏文音节，若不是就将放置另一个文件中；

➢ 对读取的藏文音节按前导字符切分音节；

➢ 处理藏文字丁；

➢ 判断该字符是否在字典中，若不在则把该字符存入字典，并将“统计数值”设为 1，若存在则其数值 +1；

➢ 转到 3）步直至文本结束。

软件流程图如图 1 所示。

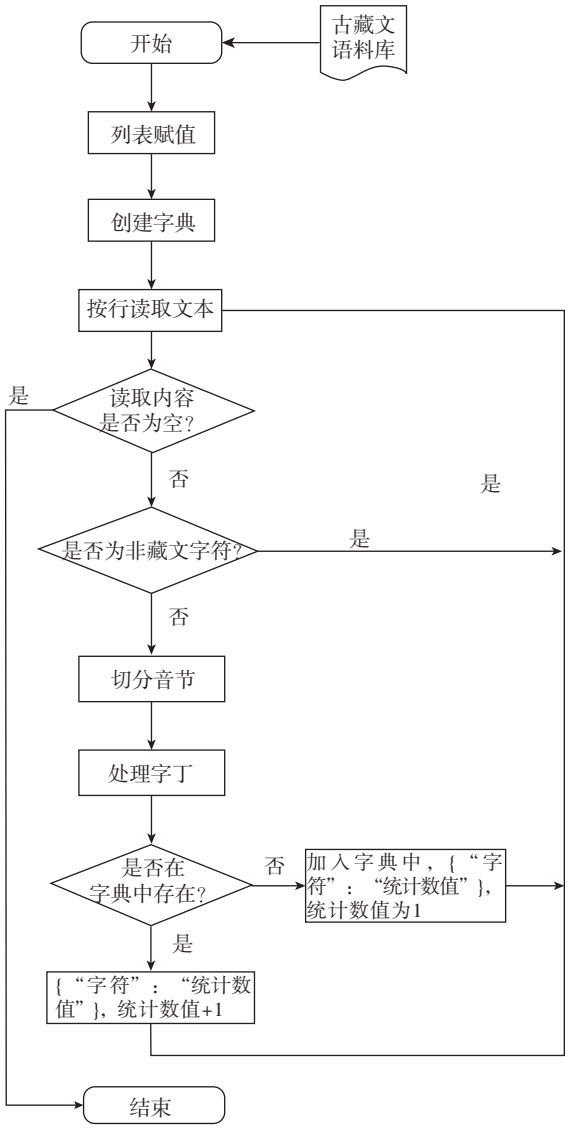

图 1 字符统计流程图

doi:10.3772/j.issn.2095-915x.2023.02.011

1.3 古藏文字符频度统计软件的实现

传统藏文语法理论中，藏文公认为是参照梵文创制的。公元 7 世纪初智者吞米桑布扎从梵文 16 个元音中选取 4 个与藏语发音相近的字母；从 34 个辅音中选取了 24 个与藏语发音相似的字母，在此基础上新创 6 个符合藏语发音的特殊字母，最后创制了 30 个藏文辅音字母和 4 个元音符号。对于古今藏文文献中的藏文标点符号以及反写元音“i”等特殊字符缺乏统一的观点。为了探究古藏文中的特殊字符，以及对古藏文字符分布特征进行进一步了解，本文应用以上设计软件对敦煌藏文文献为主的古藏文语料库进行字符统计的结果如表 1 所示：

表 1 古藏文字符统计表

序号	字符	频度	序号	字符	频度	序号	字符	频度	序号	字符	频度
1	space	340505	18	pa	30890	35	ja	4229	52	ai	13
2	sa	98569	19	• i	26812	36	ha	2924	53	4	11
3	ga	79105	20	ta	18063	37	dza	2218	54	5	9
4	da	72764	21	ka	16020	38	§ <ya>	826	55	rra	8
5	ba	66470	22	cha	13448	39	wa	714	56	au	8
6	ra	62779	23	.<ja>	12964	40	.<jd>	607	57	6	6
7	o	59618	24	zha	11525	41	.<jb>	593	58	0	6
8	ya	58140	25	pha	10341	42	(cv)mq	548	59	8	5
9	na	57796	26	ca	9141	43	a	396	60	9	5
10	nga	56119	27	tsa	8358	44	aa	333	61	7	4
11	i	50006	28	sha	7572	45	space	219	62	nna	3
12	u	49066	29	nya	7243	46	(cv)hq	101	63	dda	2
13	ma	47951	30	kha	7080	47	§ <sd>	90	64	(c)x(/-)	1
14	,	45773	31	tha	6019	48	o+e	27	65	××	1
15	e	43334	32	tsha	5660	49	1	19	66	ssha	1
16	la	43333	33	(c)f(v)	5614	50	2	19	67	tta	1
17	va	38885	34	dza	4841	51	3	18	68	.....	1

从表 1 古藏文字符统计表中可以看出，古藏文字符分布总体上与现代藏文保持相对一致性，都是由藏文元辅音以及数字和标点符号等四大字符类型组成，单个字符和频度也与现代藏文较为相似。但对表中字符进行分门别类与现代藏文进行对比，我们也能够发现古藏文字符分布的独特之处，挖掘藏文演变长期历程以及古今藏文字符的细微差异。

2 古今藏文字符对比分析

语言是表达思维的工具，而文字又是从属与语言的语言符号，这是亚里士多德在内的国内外诸多学者的共识[7]。藏文在长久的历史演

变中，依照藏语口语变化以及书写方便的需要进行了三次大规模厘定，因此在古今藏文字符计对比中也能够看出这一变化情况。

2.1 古藏文元辅音字符分析

2.1.1 古藏文辅音字母

元音和辅音是构成藏文的主要要素。藏文文法根本著《三十颂》中明确指出“藏文可分为元音和辅音两种类型。‘ai’等四个组成元音符，‘kaa’等三十个组成辅音字母。”从表2的统计结果来看，古藏文中总出现了37个辅音字符，这与传统藏文文法中公认的30个藏文辅音多出7个字符。为了探究这一问题，我们对统计数据按频次进行降序排列，并计算出每个字符的占比频率。显然在37个字符排列中，序号30号以后的字符基本上都超出了30个藏文辅音的传统范围，唯独字母“a”排列在32号的位置。

表2 古藏文辅音统计表

序号	字符	频次	频率/千	序号	字符	频次	频率/千	序号	字符	频次	频率/千
1	sa	98569	115.273	14	ka	16020	18.735	27	ja	4229	4.946
2	ga	79105	92.510	15	cha	13448	15.727	28	ha	2924	3.420
3	da	72764	85.095	16	zha	11525	13.478	29	dza	2218	2.594
4	ba	66470	77.734	17	pha	10341	12.093	30	wa	714	0.835
5	ra	62779	73.418	18	ca	9141	10.690	31	(cv)mp	548	0.641
6	ya	58140	67.992	19	tsa	8358	9.774	32	a	396	0.463
7	na	57796	67.590	20	sha	7572	8.855	33	aa	333	0.389
8	nga	56119	65.629	21	nya	7243	8.470	34	nna	3	0.004
9	ma	47951	56.077	22	kha	7080	8.280	35	dda	2	0.002
10	la	43333	50.676	23	tha	6019	7.039	36	ssha	1	0.001
11	va	38885	45.474	24	tsha	5660	6.619	37	tta	1	0.001
12	pa	30890	36.125	25	(c)f(v)	5614	6.565	—	—	—	—
13	ta	18063	21.124	26	za	4841	5.661	—	—	—	—

在这些字符中，可以看出排列倒数的“tta”、“ssha”、“dda”、“nna”四个字符是藏文梵音转写字，与古藏文字符基本没有太大关系，在统计数据中频率最高的也不超过0.0004%几乎可以忽律不计。排列倒数第5、第7的“(cv)mp”和“aa”较为特殊，这两个字符在形态上也同样跟梵音转写字中的字符较为吻合，但古藏文中“(cv)mp”和“aa”也同样当做藏文字符应用在藏文文献中，只不过古藏文中的p”是藏文“(cv)m字母“ma”的缩写体，“aa”是表示长元音的特殊符。

而多出的7个古藏文字符中，排列靠前的“(c)f(v)”其在古藏文中的原形是“vfa”，是字符“va”和“(c)f(v)”的组合形式，在古藏文中“vfa”只是“va”的另一种写法，并不能成为单个藏文字符。因此，可以看出古藏文辅音字母与现代藏文或与传统藏文文法理论中的三十个藏文辅音字母基本吻合，虽有“vfa”和

doi:10.3772/j.issn.2095-915x.2023.02.011

“(cv)mp”等个别变体字母的存在，但总体上原形占比高于变体，继而回顾藏文千年的演化历程，出现极个别变体字母也实属正常。

2.1.2 古藏文元音字母

对于传统藏文辅音符，跟文中上一段所述的一样公认是 4 个字符，但从统计结果表 3 中可以看出，古藏文文献共出现了 7 个与现代藏文元音符相似的符号。

表 3 古藏文辅音统计表

序号	元音	频次	频率%
1	0	59618	29.510
2	i	50006	24.753
3	u	49066	24.287
4	e	43334	21.450
5	• i	26812	13.272
6	o+e	27	——
7	ai	13	——

从后往前看，最后两个音符明显与前五个音符不同。这两个音符由两个常见藏文音符组合而成，而其中字符“ai”又跟梵音转写字中的长元音符“e”及其相似，使得这两个音符显得更加复杂。可实际上，这两个“元音符”是古藏文中虚词“vi”在特殊条件下的一种缩写形式，与藏文梵音转写字中的“ai”大有不同。才项南杰在《敦煌藏文文献断代研究—以古藏文“poe”和“pi”等俗体字的演变分析》中表示，“这两个音符是古藏文中对虚词“vi”的一种缩写方式[8]”不能看做是藏文元音，同时他也表明这一缩写体不是分布于每个古藏文文献中，而在某一时期的文献当中，这也就符合统计表中此类符号相对较少的结果。

此外，排列第 5 的“·i”在古藏文中也呈现出明显的特点。王尧在他的《吐蕃文献学概述》一文中提到：“元音符号的反书：这是一个十分醒目的标记，几乎成为吐蕃文献共同特征，不用赘举。”“把它作为吐蕃文献的标志之一却是无庸置疑的。”[9] 对此周季文先生也表示完全符合事实[10]。

2.2 古今藏文元辅音频率差值分析

2.2.1 辅音字母频率对比分析

由下表 6 古今藏文辅音字母对比表中可以看出，虽然古今藏文字母频率分布不是完全相等，但三十个辅音字母频率平均差不超过 0.5%，其中大于 1% 的有“ra”、“na”、“ga”、“ya”、“nga”和“da”等字母。虽然这六个字母差值最大的原因还无法判定，但在现有的古藏文特征研究中大致还能推断个别字母出现差值的原因。如：“ra”在古藏文占比中减去现代藏文占比后得出负的 2.14%，这表明比起古藏文现代藏文中字母“ra”的频率明显较高。与此相反的是“ya”在古藏文占比中减去现代藏文占比后得出 1.35，表明古藏文中字母“ya”出现的频率高于现代藏文。而这两个字母的差值变化也正好和古藏文中下加字“ya”通常代替下加字“ra”的现象。

除此之外字母“da”的差值得 1.05% 也符合古藏文中再后加字以现形式存在的特点。但因为此对比表未能按藏文每个部位的字频进行细化对比，因此统计结果不仅受构词特征的不同影响，也有可能受古今藏文句法表达等其它因素的制约，故不能仅靠词表来断定古今藏文字词构成的不同差异。

表 6　古今藏文辅音字母对比表

序号	字母	（古）频率/%	（现）频率/%	古-现=差值	序号	字母	（古）频率/%	（现）频率/%	古-现=差值
1	ka	1.887	2.164	-0.28	16	ma	5.712	5.788	-0.08
2	kha	0.834	1.501	-0.67	17	tsa	0.984	0.668	0.32
3	ga	9.316	10.856	-1.54	18	tsha	0.667	1.143	-0.48
4	nga	6.609	5.301	1.31	19	dza	0.261	0.469	-0.21
5	ca	1.076	0.915	0.16	20	wa	0.084	0.155	-0.07
6	cha	1.584	1.433	0.15	21	zha	1.357	0.968	0.39
7	ja	0.498	0.576	-0.08	22	za	0.57	0.767	-0.20
8	nya	0.853	0.952	-0.10	23	va	4.579	4.639	-0.06
9	ta	2.127	1.605	0.52	24	ya	6.847	5.493	1.35
10	tha	0.709	1.103	-0.39	25	ra	7.393	9.532	-2.14
11	da	8.569	7.516	1.05	26	la	5.103	4.877	0.23
12	na	6.806	5.241	1.56	27	sha	0.892	0.900	-0.01
13	pa	3.638	4.570	-0.93	28	sa	11.608	11.217	0.39
14	pha	1.218	1.020	0.20	29	ha	0.344	0.417	-0.07
15	ba	7.828	8.086	-0.26	30	a	0.047	0.130	-0.08
序号	字母	（古）频率/%	（现）频率/%	古-现=差值	序号	字母	（古）频率/%	（现）频率/%	古-现=差值

2.2.2　元音字母符频率对比分析

虽然元音不能在藏文中单独成字，但元音在藏文构字或构词中也是不可或缺的存在。表 7 中可以看出比起辅音古今藏文中的元音差值明显较大，元音平均差值甚至是辅音字母的 7 倍多。但需要注意的是藏文元音表和辅音表中的字符数量不同，辅音个数是元音的 7 倍多，而这两个差值也就能看出元辅音的两个平均差之间也不超过 1。因此此表中的差值大小只能在表中元音间比对，不能与以上辅音表中的差值直接对比得出错误的结论。

表 7　元音字母符频率对比

序号	字符	（古）频率/%	（现）频率/%	古-现=差值
1	i	33.569	26.247	7.322
2	u	21.442	23.893	-2.451
3	e	18.937	18.474	0.463
4	o	26.053	31.385	-5.332

另外，表中古藏文元音“i”是古藏文中两个正反“i”元音的频率，若将两者分开“i”元音在古藏文中的频率为 21.852%，与现代藏文中“i”元音的差值是 4.395。

3　古藏文音节统计

藏语语音的特点是单音节性，每一组元音和辅音字符串代表藏语里的一音节，每个音节可能代表藏语里的一个词，也可能代表一个词素 [11]。藏文音节都由一个音节点符号“·”或分句符“།”隔开，每个音节都由 1~4 个不同的字丁构成。有关古藏文研究的今现代著述中，也基本都会谈到古藏文音节在行文中的规律，而这一点基本上也可以包含在古藏文书写特征中。

doi:10.3772/j.issn.2095-915x.2023.02.011

如表 8 的统计数据来看，古藏文音节基本上与现代藏文遵循着同样的特点，没有太大差异，但在单个字丁组成方面，古藏文显得更加自由。1984 年民族出版社出版的《吐蕃碑刻钟铭选》中提出十大特征，这也是较为系统的提出古藏文书写特征的最早观点，而这些特征基本上都能在音节统计表中找到相对应的例子，这不仅证实了这几大特征的可信性，同时也论证了古藏文音节构件的自由性。而对书中第十个特点，因超出该段所述的音节特征范围，因此不在此加以探讨。

表 8　古藏文音节统计表

序号	音节	频次	序号	音节	频次	序号	音节	频次	序号	音节	频次
1	pa	40696	28	myed	6413	55	so	4488	82	che	3315
2	dnga	34998	29	myi	6341	56	ky • i	4448	83	ltr	3315
3	la	34137	30	rpa	6286	57	bzhin	4412	84	pvi	3258
4	na	30005	31	dga	6103	58	pha	4335	85	stsogs	3171
5	du	28463	32	zhes	5908	59	rnmsa	4272	86	g • i	3153
6	ma	24893	33	gnga	5824	60	re	4248	87	ppa	3115
7	ba	22231	34	bva	5819	61	kyis	4111	88	do	3103
8	de	20663	35	sems	5806	62	gsol	3988	89	chda	3066
9	po	19360	36	lha	5732	63	ldna	3961	90	pkva	3013
10	pra	15739	37	kynga	5728	64	cda	3935	91	phyir	2963
11	ni	14848	38	psa	5656	65	gshegs	3912	92	vog	2924
12	v	14116	39	rgyla	5591	66	zhig	3864	93	pcom	2918
13	ynga	13800	40	bdga	5529	67	pvi	3834	94	to	2870
14	nsa	13749	41	gyis	5339	68	pho	3790	95	runga	2855
15	myi	13113	42	ptsna	5284	69	stong	3660	96	yin	2845
16	n • i	12255	43	chen	5097	70	pzng	3653	97	vd • i	2819
17	te	10694	44	su	5071	71	rgya	3636	98	lta	2793
18	lsa	9732	45	plon	5069	72	Po	3635	99	va • i	2784
19	ste	9384	46	vi	4915	73	v • i	3625	100	pysa	2780
20	bra	8705	47	shes	4801	74	ptpa	3616	101	sa	2768
21	gyi	8190	48	yul	4773	75	gis	3501	102	vdi	2748
22	mo	7638	49	rje	4758	76	gnyis	3479	103	vo	2625
23	ngo	7168	50	gi	4752	77	pynga	3399	104	mnga	2594
24	bu	7161	51	lo	4752	78	zhing	3393	105	gcig	2577
25	bya	6820	52	chos	4672	79	zhnga	3381	106	gsum	2557
26	kyi	6567	53	Pv • i	4621	80	psa	3346	107	pva	2520
27	gy • i	6423	54	mchis	4584	81	nva	3343	108	gyur	2515

现代藏文中，一个藏文音节最少可由一个部件构成，最多可包含七个构件，这七种构件每个位置上都有严格的字符限制。单个字符组成音节时，必须是辅音字母构成的基字，而元音符只能在基字的基础上充当上下构件，不能单独使用。然而，在古藏文中却不尽相同。如下表 8 所示，在语料库音节统计表中还出现了“d”、“bsduste”、“gyg·i”和“gsumq”、“dpaa”等现代藏文中极为罕见的音节。

这些音节若出现在当今某个藏文著述中，可能上会便认成藏文梵音转写字，甚至在单个语境里会理解成为错别字，但在古藏文中，这只是两个音节或一个音节的特殊缩写形式。由下图 9 可知，这类两音节缩写体最大的特点就是，一个音节中包含了两个基字，也就是说这一音节原本是两个音节 [12]。

表 9　古藏文缩写体统计表

序号	紧缩字	频次	序号	紧缩字	频次
1	dpngo	254	16	rikso	28
2	psduste	119	17	lgso	27
3	stsldo	69	18	pgyiste	26
4	gyg·i	66	19	myedo	26
5	dgumo	55	20	pdkis	26
6	pdgi	50	21	Pdk·i	26
7	pzhgste	49	22	vduste	26
8	pogste	47	23	mchisn	25
9	mchis	44	24	phyogsu	24
10	pzngo	44	25	phroks	24
11	ppste	37	26	psogs	24
12	drngste	32	27	yongsu	24
13	phste	30	28	chungu	24
14	stsldo	30	29	gnngo	23
15	ptgste	29	30	gygi	23

再回首现代藏文，现代藏文中我们用虚词“vi”“vo”“vnga”时，也常会把它与上一个音节间的音节点去掉，缩写成同一个音节，而这就跟以上古藏文中的缩写形式如出一辙。因此“ pvi”“byvo”“nvnga”等与“va”元音相关的音节缩写，是唯一种沿用至今的缩写方式。但再仔细对比，“dbngo”、“bsduste”、“gygi”等古藏文缩写与“ pvi”“byvo”“nvnga”还有个区别之处。如图 2 可见，古藏文中“dbngo”等缩写体在组成同一个音节时，被缩写的两个音节中有一个共同的音节构件，这也是此类缩写体形成的主要基础。而“ pvi”“byvo”“nvnga”等现代藏文缩写体没有共同利用同一个音节构件，因此与其说这些缩写体是古藏文缩写方法的沿用，不如更接近与“vi”等虚词与前一个音节之间免去音节点的说法。而解释这一质疑，则需要追究第三次藏文厘定前的单辅音形态，若对第三次藏文厘定有所了解，这一表层现象也就不攻自破。

如上所诉，《吐蕃碑刻钟铭选》中对古藏文特征的第七个观点就指出，古藏文在单独构成音节时另加元音“va”为后加字。也就是说在古藏文中，“ pvi”“byvo”“nvnga”等缩写体在拆解后，前面的音节“pa”“bya”“na”后面具有后加字“va”。如此以来在古藏文中“pa”“bya”“na”等单元音音节就是以“ pvi”“byvo”“nvnga”的形式存在，而这也就完全对应了语料库统计表 10 中出现的“pa”“bya”“na”等字符，证实了现代藏文中“ pvi”“byvo”“nvnga”等缩写体与古藏文中“双音节”缩写体之间的同类关系，使表 11 中的缩体分解有了一定的实证依据。

doi:10.3772/j.issn.2095-915x.2023.02.011

（a） （b） （c）

图 2 古藏文“双音节”缩写体拆解图

表 10 古藏文中带有后加字“va”的音节统计表

序号	辅音	频次	序号	辅音	频次
1	pva	1407	16	gva	21
2	nva	786	17	duv	20
3	bva	616	18	steva	18
4	byva	94	19	tev	17
5	lva	74	20	skyv	17
6	phva	68	21	sv	14
7	dev	62	22	rgyva	14
8	mva	56	23	dbva	14
9	dva	44	24	ngov	14
10	pov	39	25	ngva	11
11	zva	27	26	lhva	11
12	ltva	27	27	C • iv	11
13	kva	23	28	j • iv	11
14	buv	22	29	rev	9
15	sova	21	30	civ	8

表 11 古今藏文“双音节”缩写体对比表

主题	A类				B类	
古藏文中的原形	dbngo	bsduste	gyg·i	pvi	byvo	nvnga
古藏文的拆分体	dbnga + ngo	bsdus + ste	ayga + g·i	pva + vi	byva + vo	nva + nga
现代藏文中的原形	dbnga ngo	bsdus ste	ayga g·i	pvi	byvo	nvnga
现代藏文的拆分体	现代藏文中A类没有缩写体			pa + vi	bya + vo	na + nga
	缩体消失				缩体保留	

值得注意的是，古藏文音节统计中还出现的“lstsogs”也是两个音节的缩写体，不同的是这一频度较高的缩写体音节，只是将词缀“l”与“lstsogs”之间的音节点省略，与上

述古藏文中常见的两音节缩写体形成鲜明的对比。除此之外，对于古藏文中同样应用较广的“gsump”和“dpaa”等单音节缩写体，在第三次藏文厘定后，基本从现代藏文中丢弃。而这些单音节的缩写，即借用了梵文文法中的书写方法，又与藏文梵音转写字区分差别。列如：Pt_0983、Pt_0986、Pt_1047、Pt_1051、Pt_1283、ITJ_0407、ITJ_0731、ITJ_0733、Or_8212_0187、insc_Zhva_W、等文献中，只是藏文元音“ma”的缩写，与藏文梵音转写字有着根本区别。

4 小结

语言是文化的载体，文字是打破时空界限的语言变体。敦煌古藏文文献是古人淡墨执笔书写的千年文书，应用语料库语言学方法对敦煌古藏文字词进行研究，不仅是探知千年文化的基础，也是现代藏文信息化的重要研究领域。本文所得出的结论基本立足于统计数据的结果以及领域学者的研究成果，虽有较强的实证依据，但因读取敦煌藏文原始文献的专业性要求，以及语料内容欠缺，所提出的观点还需有关专家学者推敲验证。

参 考 文 献

[1] 党燕妮 , 郭向东 , 陈军敦 . 煌少数民族文献举要 [J]. 图书与情报 , 2014, 2: 31-38.
[2] 多拉 , 扎西加 . 藏文规范音节频率词典 [M]. 北京 : 中国社会科学出版社 , 2015: 1.
[3] 卢亚军 . 基于大型藏文语料库的藏文字符、部件、音节、词汇频度与通用度统计及其应用研究 [J]. 西北民族大学学报（自然科学版）, 2003.06.15.
[4] 才让加 . 藏语语料库加工方法研究 [J]. 计算机工程与应用 , 2011.2.14.
[5] 三排才让 . 敦煌古藏文文献的字词属性统计研究 [D]. 拉萨 : 西藏大学硕士论文 , 2021: 6.
[6] 边巴旺堆 , 卓嘎 , 陈延利 , 等 . 藏文构件元素识别算法研究 [J]. 中文信息学报 , 2015.5.
[7] 王文斌 , 柳鑫淼 . 汉语会意字构造与意合表征方式的相承关系 [J]. 当代修辞学 , 2020, (1): 18-28.
[8] 才项南杰 . 敦煌藏文文献断代研究—以古藏文“poe”和“pi”等俗体字的演变分析 [A]. 第四届全国藏文古籍文献整理与研究高层论坛论文集（下）[C]. 2021 年 10.
[9] 王尧 . 吐蕃金石录 [M]. 北京 : 文物出版社 , 1982.
[10] 周季文 . 论藏文的元音符号“i”[J]. 东方语言学 , 2009, (2): 67-83.
[11] 江狄 , 董颖红 . 藏文信息处理属性统计研究 [J]. 中文信息学报 , 1995, 9(2): 37-44.
[12] 梁金宝 . 藏语历史文献词汇统计研究 [D]. 北京 : 中国社会科学研究院研究生院博士学位论文 , 2013.6