

文章编号:1001-9081(2009)07-2029-03

藏文字形结构计量统计分析

艾金勇,于洪志,李永宏

(西北民族大学 中国民族信息技术研究院,兰州 730030)

(ajy0529@126.com; lyhweiwei@126.com)

摘要:通过对藏文词典的统计,计算出出现代藏字在藏文中的使用频度,并对藏字分别进行了部件和字丁层面上的分析,得出藏字构字方面的特征。同时依据藏字声母和韵母的结构方式的统计结果,揭示了藏字在声韵母方面的特性,为藏文的研究和信息化处理提供了一些基础数据。

关键词:现代藏字;部件;频率统计

中图分类号: TP391 **文献标志码:** A

Statistical analysis on Tibetan shaped structure

AI Jing-yong, YU Hong-zhi, LI Yong-hong

(China Minorities Information Technology Institute, Northwest University for Nationalities, Lanzhou Gansu 730030, China)

Abstract: Based on the statistics about the Tibetan dictionary, this paper calculated the use frequency of modern Tibetan words, and analyzed the Tibetan character at the level of part and whole respectively and got the characteristics of the word configuration. Meanwhile a statistic analysis was made on the structure of the initial consonant and the final. A number of characteristics of the Tibetan initial and the final were revealed, which provides some fundamental data for the future Tibetan studies and information processing.

Key words: modern Tibetan; small character parts; frequency statistics

0 引言

藏语是在梵文天成体的基础上发展而成的一种拼音文字,有 1300 年的历史。藏文一共有 30 个辅音字母,4 个元音符号,加一个零位/a/,共 5 个元音。藏文正字法中关于藏字构成有一套严格而完整的组合排列规则,根据字母在音节中的结构位置,将字母分为“基字”、“上加字”、“下加字”、“前加字”、“后加字”和“再后加字”7 个部件,除基字和元音外,其他都可为空,30 个辅音字母都可以做基字,元音不能做基字,30 个辅音字母中有 5 个做前加字(“ཀ་ཁ་ག་ང་ཅ་”),3 个上加字(“མ་,ཙ་,ཛ་”),4 个下加字(“ཨ་,ཉ་,ལ་,ཤ་”),10 个后加字(“པ་ཕ་ཆ་ཇ་མ་ཉ་ཏ་ཐ་ཇ་མ་ཉ་ཏ་ཐ་”)再后加字是后加字中的“ད་”(今不用)和“ས་”,当再后加字出现时,后加字和再后加字的组合在现代藏文中只有四种形式(“ཀ་ཁ་ག་ང་ཅ་པ་ཕ་ཆ་ཇ་མ་ཉ་ཏ་ཐ་”)。

关于藏字结构的研究,文献[1]曾对基于口语的材料进行统计,得出藏字的基本属性特征。但是由于研究材料主要基于口语,有很大的局限性,因而并未能完全表征出藏字的特征。本文语料基于常用的藏文字典和词典,以现代藏字为主要研究对象,对其频度、声韵母分布、部件频度以及相关属性进行统计分析。同时也考虑了藏文材料中存在的一部分不符合现代藏文文法的藏字,即为梵音藏文、借词和古藏文。

1 藏字属性统计研究

为了获得尽可能多的藏字,我们收集了包括《藏汉大字典》、《安多口语字典》、《拉萨口语字典》、《格西曲扎藏文辞典》、《新编藏文字典》、《藏文同音字典》在内的 6 部藏文字

(词)典共计 13 万余条词汇,包含藏字 7 062 个,其中现代藏字 6 363 个,借词、梵音藏文和古藏文 699 个(见表 1)。

表 1 藏字分布表

类别	数量	比例/%
独立成词的现代藏字	4 780	67.69
非独立成词的现代藏字	1 583	22.42
独立成词的非现代藏字	122	1.73
非独立成词的非现代藏字	577	8.17
总计	7 062	100.00

注:非现代藏字主要指藏文中出现的不符合藏文正字法的藏字,其中包括借词、梵音藏文和古藏文等。

总体上看,所有藏字中大约 90% 的为符合藏正字法的现代藏字,其他非规范藏字仅有 10% 左右。

现代藏文中大部分字都可以单独成词,但也有接近 1/3 的现代藏字只能在双音节以上的词条中出现。在实际的藏文语料库中,现代藏字的使用频率高达 99%,其中 95% 的现代藏字是具有独立意义的。

1.1 声韵母组成结构

在七世纪吐弥桑布扎创制藏文时,撰写了《文法根本三十颂》和《字性法纲要》两本有关藏文文法的经典著作,九世纪大规模进行藏文厘定时,对这两本书也进行了修订,形成了藏语古音韵及文法的基本理论框架,根据这个框架,确定出古藏语的声母一共有 220 个,其中单辅音声母 30 个,二合辅音 115 个,三合辅音 69 个和四合辅音 6 个。古藏语韵母一共有 98 个,其中有 5 个单元音韵母,8 个复元音韵母,50 个单尾韵和 35 个复尾韵。藏字声母由基字、前加字、上加字及下加字组

收稿日期:2008-11-24。 基金项目:国家自然科学基金资助项目(60773052),国家民委 2007 年重点科研项目(民委发[2007]10)。

作者简介:艾金勇(1983-),男,湖北襄樊人,硕士研究生,主要研究方向:中文信息处理; 于洪志(1947-),女,山东龙口人,教授,博士生导师,主要研究方向:多语言信息处理; 李永宏(1979-),男,山西临汾人,博士研究生,主要研究方向:计算语言学。

成,韵母是由元音、后加字及再后加字组成,具体组合形式见表 2、表 3^[1]。

表 2 声母结构方式统计

组合类型	部件	个数	字数	频度 / %
基字	1	30	1576	24.77
前加字 + 基字	2	47	1651	25.95
上加字 + 基字	2	34	962	15.12
基字 + 下加字	2	34	806	12.67
前加字 + 基字 + 上加字	3	20	354	5.56
前加字 + 基字 + 下加字	3	32	567	8.91
上加字 + 基字 + 下加字	3	14	324	5.09
基字 + 下加字 + 下加字	3	3	5	0.08
前加字 + 上加字 + 基字 + 下加字	4	6	118	1.85
总计		220	6363	100.00

表 3 韵母结构方式统计

组合类型	个数	字数	频度 / %
元音(包含复合元音)	13	1167	18.34
元音 + 后加字	50	4072	63.99
元音 + 后加字 + 再后加字	35	1124	17.66
总计	98	6363	100.00

声母构造方式的复杂程度同所构成的藏字数量大致成反比关系,结构越简单其构字能力越强,有接近 1/4 的藏字用单部件做声母,单部件和双部件声母的藏字总数接近所有现代藏字的 4/5,而四部件声母的藏字只有 118 个不到现代藏字总数的 2%。在所有的组合中,基字与前加字作声母出现的频度远大于其他组合,基字与上加字和基字与下加字做声母的情况在数量上相差不大。整体上来看,藏字声母结构更偏重于横向叠加,韵母结构相对简单,以“元音 + 后加字”为主要构成方式。

2 字丁统计


藏文大字符集编码方案是以上下叠加的垂直预组合(字丁)作为一个整体进行编码。例如“”就是一个带有上加字(𑀓)和下加字(𑀔)和元音(𑀅)的字丁,藏字字丁在现代藏字的使用频度如表 4 所示。

表 4 藏文字丁频度表

序号	字丁	频率	累积	序号	字丁	频率	累积
1	𑀓	10.50	10.50	10	𑀔	2.74	58.90
2	𑀔	9.76	20.26	11	𑀅	1.49	60.38
3	𑀅	9.21	29.47	12	𑀆	1.01	61.40
4	𑀆	5.40	34.86	20	𑀇	0.25	63.72
5	𑀇	5.33	40.19	40	𑀈	0.20	68.06
6	𑀈	5.05	45.24	51	𑀉	0.18	70.13
7	𑀉	4.76	50.00	118	𑀊	0.12	80.01
8	𑀊	3.23	53.24	219	𑀋	0.08	90.03
9	𑀋	2.92	56.16	491	𑀌	0.01	100.00

从表 4 得知:

1) 藏文字丁的频度分布极不均匀。统计发现在至少出现 1 次的 491 个字丁中,序号在前的 7 个字丁就覆盖了所有词条字丁总数的一半,即占总字数不到 2% 的字丁却占据了 1/2 的字丁数目。如序号 1 的字丁 𑀓 的频度最高,它占据字丁总数的 10.50%,而序号为 219 的 𑀋 的频度仅为 0.08%,其

累计频度为 90.03%,它后面的 272 个字丁的频度累计值尚不及第一个字丁的频度值。

2) 频度累计值增长不均匀。在前的较少一部分字丁,频度的累计值增长较快,而后面很大一部分字丁,累计值增长十分缓慢。如从序号 1 到 10,频度累计值从 10.50% 增加到 58.90%,净增值为 48.40%;序号从 10 到 20,频度累计值从 58.90% 增加到 63.72%,净增值为 4.82%;序号从 20 到 40,频度累计值从 63.72% 增加到 68.06%,净增 4.34%;从累计频度 80.01% 的序号为 118 的字丁 𑀊 起,到累计频度为 90.03% 的 219 号字丁 𑀋 之间经过了 101 个字丁;而累计频度值从 90.03% 增加到 100.00%,之间经过了 272 个字丁。

3) 累计频度达到 61.40% 的字丁只需要 12 个,这些字丁就是构成常用的高频字丁了。

4) 出现次数很少的字丁,出现次数在 10 次以下的字丁有 219 个,占字丁总数的 44.60%。

根据字丁在所有藏字中的使用频率,将其划分为高频字丁、次高频字丁、中频字丁和低频字丁四个等级,图 1 列出了字丁数目和使用频率的关系。

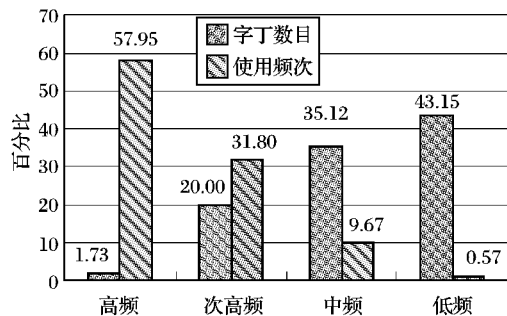


图 1 藏文字丁使用频率分布表

从图 1 中可以很直观的看出仅仅不到 3% 的字丁在构成藏字时其使用频率却高达 61.40%,而后面 44.60% 的字丁在构成藏字时仅出现了不到 10%。由此可见,藏字的差异主要集中在少量常用的字丁的不同组合方式上。

3 藏字部件统计

3.1 藏字部件数分布情况统计

部件是构字时能反复出现的、并能从字形分割出来的有固定形体的笔划组合块。一个藏字可由一到七个部件构成。藏字部件数分布情况有助于计算机藏文内码的设计、藏字字模库的建立、藏字输入编码方式以及藏文识别(见表 5)。

表 5 不同部件数藏字的频数统计

部件数	藏字数	频度 / %	累积 / %
1	30	0.47	0.47
2	486	7.64	8.11
3	1862	29.26	37.37
4	2434	38.25	75.62
5	1260	19.80	95.43
6	271	4.26	99.69
7	20	0.31	100.00

注: 元音符号 a 不计入统计之列。

从以上统计可以看出,部件分布极不均匀,两部件、三部件和四部件的藏字数目就覆盖了接近 80% 的样本数据,而多于五部件数的藏字的频度总共不到 5%,以上统计充分说明藏字结构的合理性,尽量用有限的字符构造出最佳的区别字

形的形式。以上数据经统计可得藏字的平均部件个数为3.8331个,图2也清晰地表明了不同部件数藏字的分布。

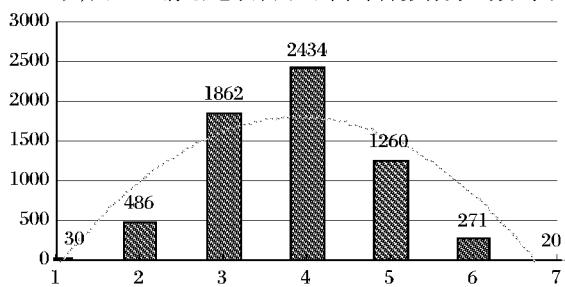


图2 藏字部件分布

3.2 部件频度统计

藏字是由字符依据正字法规定横向和纵向叠加拼写而成的,为了使得统计上有更好的可比性,以及以后对藏文基本构成单元的分析研究,我们以部件为单位对其频度进行统计,同时为了更清楚了解部件的不同变形体的频度情况,我们把部件中的变形体也单独做了统计(见表6)。

表6 藏文部件频度及累加频度

序号	部件	频率	累频	序号	部件	频率	累频
1	ཨ	11.91	11.91	20	ཨ	0.94	91.85
2	ཨ	8.49	20.40	21	ཨ	0.86	92.71
3	ཨ	8.22	28.61	22	ཨ	0.83	93.54
4	ཨ	8.14	36.75	23	ཨ	0.67	94.21
5	ཨ	6.28	43.03	24	ཨ	0.66	94.87
6	ཨ	5.53	48.56	25	ཨ	0.66	95.53
7	ཨ	5.35	53.91	26	ཨ	0.63	96.16
8	ཨ	4.80	58.71	27	ཨ	0.63	96.79
9	ཨ	4.68	63.40	28	ཨ	0.60	97.39
10	ཨ	4.08	67.48	29	ཨ	0.57	97.96
11	ཨ	4.05	71.53	30	ཨ	0.53	98.49
12	ཨ	4.00	75.53	31	ཨ	0.53	99.02
13	ཨ	3.83	79.37	32	ཨ	0.52	99.55
14	ཨ	3.26	82.63	33	ཨ	0.21	99.75
15	ཨ	2.71	85.34	34	ཨ	0.17	99.92
16	ཨ	2.11	87.45	35	ཨ	0.03	99.95
17	ཨ	1.28	88.72	36	ཨ	0.03	99.98
18	ཨ	1.22	89.95	37	ཨ	0.01	99.99
19	ཨ	0.96	90.91	38	ཨ	0.01	100.00

注:元音符号 a 不计入统计之列。

由表6可以得出:藏字部件频度分布极不均匀,其中最高频度的部件“ཨ”占样本部件总数的11.91%,频级在前的4个部件就覆盖了统计样本的36.75%,即不到2%的部件的频度接近样本部件总频度的40%;而频级为19的部件“ཨ”的频度仅为0.96%,累计频度为90.91%,它后面的共计19个部件的频度累计值不及第一个高频部件。

4 结语

本文通过对藏字本体以及在部件层次上的统计分析,得出在藏语中不符合藏文正字法的梵音藏文和古藏文在现代藏语中使用频率较低。而对于藏字本身,不同部件数的藏字分布趋近于一条抛物线,出现频率极不均匀。在藏字声韵母结构方式上,似乎更倾向于横向叠加。

通过对藏文字形结构的统计分析,可以提取藏文文字编码的研究所需的规范部分作为编码的基本元,以及对部件变形显现形式层面上大样本的抽样统计,可以确定藏文合适的编码方式,同时藏文字形结构的统计分析还可以抽取出藏文文字特征,辅助计算机对藏文文字的识别,对于藏文信息化处理的研究起着非常重要的意义。

参考文献:

- [1] 格桑居冕. 藏文字性法与古藏语音系[J]. 民族语文, 1991(6): 12-22.
- [2] 江荻,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.
- [3] 于洪志. 计算机藏文编码概述[J]. 西北民族学院学报: 自然科学版, 1999,20(3): 15-19.
- [4] 于洪志. 藏文编码字符集部件集[J]. 西北民族学院学报: 自然科学版, 1998,1: 11-16.
- [5] 胡书津. 简明藏文文法[M]. 昆明: 云南民族出版社, 1987.
- [6] 王维兰. 现代藏文语言单位频率和频级关系的统计分析[J]. 科学技术与工程, 2004,4(5): 413-417.
- [7] 周毛仁增. 藏文字形结构分析与编译分析[J]. 西藏大学学报: 汉文版, 1999(8): 29-30.
- [8] 格桑居冕, 格桑央京. 实用藏文文法教程[M]. 成都: 四川民族出版社, 2004.
- [9] 陈玉忠, 俞士汶. 藏文信息处理技术的研究现状与展望[J]. 中国藏学, 2003(4): 97-107.
- [10] 胡坦. 藏语研究文论[M]. 北京: 中国藏学出版社, 2002.

(上接第2028页)

- [8] SIMARD M, CANCEDDA M, CAVESTRO B. Translation with non-contiguous phrases [C]// Proceedings of Human Language Technology Conference and Conference on Empirical Methods in NLP (HLT/EMNLP). Morristown, NJ: Association for Computational Linguistics, 2005: 755-762.
- [9] GUVENIR H A, CICEKLI L. Learning: Translation templates from examples[J]. Information Systems, 1998,23(6): 353-363.
- [10] CHIANG D. A hierarchical phrase-based model for statistical machine translation [C]// Proceedings of the 43rd Annual Meeting of the ACL. Morristown, NJ: Association for Computational Linguistics, 2005: 263-270.
- [11] KOEHN P. Pharaoh: A beam search decoder for phrase-based statistical machine translation models[EB/OL]. [2008-09-20]. <http://www.iccs.inf.ed.ac.uk/~pkoeHN/publications/pharaoh-amta2004.ps>.
- [12] ZENS R, NEY H, WATANABE T, et al. Reordering constraints for phrase-based statistical machine translation[C]// Proceedings of the 20th international conference on Computational Linguistics. Morristown, NJ: Association for Computational Linguistics, 2004: 205-211.
- [13] ZENS R, NEY H. A comparative study on reordering constraints in statistical machine translation[C]// Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Morristown, NJ: Association for Computational Linguistics, 2003: 144-151.
- [14] 陈晴, 姚天顺, 张俐, 等. 基于谓词驱动模板的汉日机器翻译方法[C]// 中国中文信息学会二十五周年学术会议. 北京: [s. n.], 2006: 439-446.