中 文 信 息 学 报
第 19 卷 第 1 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 19 No. 1

文章编号:1003 - 0077(2005)01 - 0071 - 05

现代藏字全集的属性统计研究①

高定国,龚育昌

(中国科学技术大学 计算机科学技术系,安徽 合肥 230027)

摘要:藏文基本属性的研究是藏文信息处理技术的基础,现代藏字的研究是藏文信息处理的重点。藏字全集是有限集,为了更好地研究现代藏字,本文以现代藏字为研究对象,按照现代藏文文法的规律,对全部现代藏字用计算机辅助统计了藏字全集的个数、藏字的字长、藏字的结构方式、位置特征、字符频度以及所有现代藏字中的整基字丁,并且简要地分析了这些数据。这些数据可以较全面地反映现代藏字的本质特征,可为藏文研究和藏字信息处理提供基础数据。

关键词:计算机应用;中文信息处理;藏字全集;藏字结构;藏字频度

中图分类号:TP391 **文献标识码**:A

A Statistically Study on the Qualities of All Modern Tibetan Character Set

GAO Ding-guo, GONG Yu-chang

(Department of Computer Science & Technology, University of Science & Technology of China, Hefei, Anhui 230027, China)

Abstract: A study of the basic qualities of the Tibetan language forms the basis for the Tibetan information processing. Study of modern Tibetan character is an important aspect in developing Tibetan information processing. All modern Tibetan characters set is finite, and useful for better researching modern Tibetan character, This thesis is concerned with the modern Tibetan character and how to, according to Tibetan grammar rules and using computer, do the following: calculate the total number of character, length of character, structural mode, quality of position, letter frequency, and entire character. Moreover, this thesis will also examine in a summary manner the above figures. This thesis will use modern Tibetan language analysis to better understand the nature of the language, thus offering a basic understanding for the study of the Tibetan language and Tibetan information processing.

Key words: computer application; Chinese information processing; Tibetan character set; structural mode; letter frequency

1 引言

藏语语音的特点是单音节性,每一组单音节藏文字符串代表藏语里的一个音节,每个音节可能代表藏语里的一个词,也可能代表一个词素。为此,我们把每个音节藏文字符组合称为藏字[1]。藏字由三十个辅音字母和四个元音符号(简称为元音)拼写组合而成,元音不能独立书写,只能加在辅音字母的上部或下部(图 1 中 5 的位置)。辅音字母中有些特殊的辅音字母,以一个辅音(基字)为基础,加在前、后、上、下,也可兼而有之。这些辅音按所处的位置分别命名为前加字(图 1 中 1 的位置)、上加字(图 1 中 2 的位

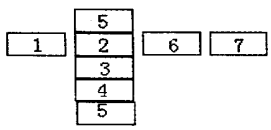

图 1 一个藏文音节的构成

① **收稿日期**:2004 - 04 - 16
作者简介:高定国(1972 —),男,硕士研究生,讲师,主要研究藏文信息处理.

置)、基字(图 1 中 3 的位置)、下加字(图 1 中 4 的位置)、后加字(图 1 中 6 的位置)、再后加字(图 1 中 7 的位置),其实它们都是构成藏字的一个字符,统称为藏字的构件[2](以下简称为构件)。其中,藏文的纵向叠加只是在基字的上下,而前加字、后加字、再后加字均为无叠加的单一辅音字母[2]。在现代藏文文法中,对藏文字符构成藏字有很严格的约束,一个藏字可有一到七个字符构成,藏文的音节最多由七字符构成(如图 1 所示,每个方块表示一个字符),并且每个位置上的字符有严格的限制。符合现代藏文文法的藏字称为现代藏字。所有的现代藏字的集合称为现代藏字全集(简称为藏字全集)。

藏文基本属性的研究是藏文信息处理技术的基础,现代藏字的研究是藏文信息处理的重点。藏字全集的统计对揭示现代藏字全集本身的属性、计算机软件的设计、藏文的研究以及支持藏文今后的发展等方面有重要的意义。因此,本文以现代藏字为研究对象,按照现代藏文文法[4,5]的规律把所有的现代藏字录入到计算机,用计算机辅助统计了藏字全集中元素的个数、藏字的字长、藏字的结构方式及对应的字数、位置特征对应的字数、构件频度以及整基字丁结构对应的字数等数据。这些数据也是我们所开展的藏文处理研究所需要的。

2 藏字全集中藏字的字长统计

一个藏字的字长是指构成该藏字的构件的多少。一个藏字可由一到七个构件构成,因此藏字的构造上是不等长的。藏字的构件数的研究对计算机藏文内码、输入编码设计和藏字识别等研究有重要的理论和实践意义。藏字全集中藏字字长的统计结果如表 1 所示。

表 1 藏字全集中藏字字长的统计结果

字符长度(字长)	对应的藏字个数	占所有藏字的百分比(%)
一个字符	30	0.155
二个字符	466	2.505
三个字符	3018	15.557
四个字符	7062	36.402
五个字符	6474	33.371
六个字符	2162	11.144
七个字符	168	0.866
藏字全集	19380	100

现代藏文的藏字全集是有限集,共有 19380 个藏字;藏字全集中总共出现了 84782 构件·次。从表 1 中可以看出:藏字全集中字长为四、五个构件的藏字占全藏字的近 70 %,是构成藏字最主要的方式。藏字全集中藏字的平均长度为 4.3723 字符。

3 藏字全集中藏字的结构方式统计

现代藏字可由前加字、上加字、基字、下加字、元音、后加字和再后加字构成。其中基字是构成藏字必不可少的构件,其它位置上构件的有无因字而不同。藏字的结构可细分为 48 种。表 2 是藏字全集中藏字的结构方式及所对应的藏字的统计结果。

在表 2 数据的基础上做个简单的计算,可以看出,无纵向叠加的藏字(不含上加字、下加字、元音)只有 1326 个,只占藏字全集的 6.84 %;而 93.16 %的藏字都是含有纵向叠加的字符。再按位置特征做个简单的计算,可以统计出藏字全集中字符的位置特征对应的字数,如表 3。

表 2　藏字的结构及所对应的藏字数的统计结果

字符长度	结　构　方　式	组成的藏字个数	占藏字全集的藏字的百分比(%)
一个字符	辅音字母	30	0.155
两个字符	基字 + 元音	120	0.619
	基字 + 后加字	270	1.495
	上加字 + 基字	33	0.170
	基字 + 下加字	43	0.222
三个字符	前加字 + 基字 + 后加字	480	2.474
	前加字 + 基字 + 元音	192	0.990
	前加字 + 上加字 + 基字	20	0.103
	前加字 + 基字 + 下加字	31	0.160
	上加字 + 基字 + 元音	132	0.680
	上加字 + 基字 + 下加字	15	0.077
	特殊的两个字(基字 + 下加字 + 下加字)	2	0.010
	上加字 + 基字 + 后加字	297	1.531
	基字 + 下加字 + 元音	172	0.887
	基字 + 下加字 + 后加字	387	1.995
	基字 + 元音 + 后加字	1080	5.567
	基字 + 后加字 + 再后加字	210	1.082
四个字符	前加字 + 上加字 + 基字 + 元音	80	0.412
	前加字 + 基字 + 下加字 + 元音	124	0.639
	前加字 + 基字 + 元音 + 后加字	1728	8.907
	前加字 + 上加字 + 基字 + 下加字	6	0.031
	前加字 + 上加字 + 基字 + 后加字	180	0.928
	前加字 + 基字 + 下加字 + 后加字	279	1.438
	前加字 + 基字 + 后加字 + 再后加字	336	1.732
	上加字 + 基字 + 下加字 + 元音	68	0.351
	上加字 + 基字 + 元音 + 后加字	1188	6.124
	上加字 + 基字 + 下加字 + 后加字	153	0.789
	上加字 + 基字 + 后加字 + 再后加字	231	1.190
	基字 + 元音 + 后加字 + 再后加字	840	4.330
	基字 + 下加字 + 元音 + 后加字	1548	7.979
	基字 + 下加字 + 后加字 + 再后加字	301	1.552
五个字符	前加字 + 上加字 + 基字 + 下加字 + 元音	24	0.124
	前加字 + 上加字 + 基字 + 下加字 + 后加字	54	0.278
	前加字 + 上加字 + 基字 + 元音 + 后加字	720	3.711
	前加字 + 上加字 + 基字 + 后加字 + 再后加字	140	0.722
	前加字 + 基字 + 下加字 + 元音 + 后加字	1116	5.753
	前加字 + 基字 + 下加字 + 后加字 + 再后加字	217	1.119
	前加字 + 基字 + 元音 + 后加字 + 再后加字	1344	6.928
	上加字 + 基字 + 下加字 + 元音 + 后加字	612	3.155
	上加字 + 基字 + 下加字 + 后加字 + 再后加字	119	0.613
	上加字 + 基字 + 元音 + 后加字 + 再后加字	924	4.763
	基字 + 下加字 + 元音 + 后加字 + 再后加字	1204	6.206
六个字符	前加字 + 上加字 + 基字 + 下加字 + 元音 + 后加字	216	1.113
	前加字 + 基字 + 下加字 + 元音 + 后加字 + 再后加字	868	4.474
	前加字 + 上加字 + 基字 + 元音 + 后加字 + 再后加字	560	2.887
	前加字 + 上加字 + 基字 + 下加字 + 后加字 + 再后加字	42	0.216
	上加字 + 基字 + 下加字 + 元音 + 后加字 + 再后加字	476	2.454
七个字符	前加字 + 上加字 + 基字 + 下加字 + 元音 + 后加字 + 再后加字	168	0.866

从表3中可以看出,除基字外后加字是构成藏字最主要的部分,藏字全集的94.372%中都有后加字,这是因为后加字不仅数目多(10个后加字),而且其构字能力强的原因。含上加字、下加字、再后加字的藏字所占的比例较少,它们本身数量少,而且添加受基字或后加字的严格限制。

表3 藏字全集中字符位置特征对应的藏字数的统计结果

藏字的位置特征	藏字数	占藏字全集的藏字的百分比(%)
含有前加字的藏字	8925	46.005
含有上加字的藏字	6458	33.289
含有下加字的藏字	8245	42.500
含有元音符的藏字	15504	79.918
含有后加字的藏字	18308	94.372
含有再后加字的藏字	7980	41.134

4 藏字全集中构件的频度统计

字符在构成藏字时,三个上加字和四个下加字有的会产生变形,为了更清楚地了解各个构件的频度,我们把上加字和下加字作为与该辅音字母不同的构件进行统计,这在藏文键盘的布局、藏字识别等方面更有利。这样,总共就有41个构件(三十个辅音、四个元音、三个上加字和四个下加字),统计各个构件的频度结果如表4。

表4 藏字全集中各构件的频度统计结果

序号	构件	藏字全集中的出现次数	占藏字全集中所有构件的百分比(%)	序号	构件	藏字全集中的出现次数	占藏字全集中所有构件的百分比(%)
1	b	7380	8.705	22	l下	935	1.103
2	d	7025	8.286	23	t	935	1.103
3	s	6380	7.525	24	p	850	1.003
4	g	5765	6.800	25	kh	850	1.003
5	m	4405	5.196	26	ts	765	0.902
6	i	3876	4.572	27	ny	680	0.802
7	u	3876	4.572	28	ph	680	0.802
8	e	3876	4.572	29	j	510	0.602
9	o	3876	4.572	30	z	425	0.501
10	r下	3145	3.710	31	h	425	0.501
11	n	3045	3.592	32	dz	425	0.501
12	ng	2960	3.491	33	sh	425	0.501
13	s上	2890	3.409	34	zh	340	0.401
14	y下	2720	3.208	35	tsh	340	0.401
15	r	2620	3.090	36	c	340	0.401
16	l	2450	2.890	37	th	340	0.401
17	r上	2380	2.807	38	ch	255	0.301
18	k	1955	2.306	39	y	170	0.201
19	v	1748	2.062	40	?	85	0.100
20	w下	1530	1.805	41	w	85	0.100
21	l上	1020	1.203				

说明:表中的藏文字符用藏文的转写符号表示,有脚注"上"、"下"字样的表示该字符为上加字或下加字。

从表4可以看出,部分构件所构成的藏字的数目相同,这表明这些构件构成藏字的能力相同,但实际使用时各藏字的出现频度不同,致使这些构件的出现频度也不同。

5 藏字全集中藏文整基字丁的统计

藏文不仅有横向拼写性,同时也有纵向拼写性(如图 1),把每一个横行基本单位称为字丁[6]。由于藏字书写是非线性的二维阵列方式[7],增加了藏文信息处理的难度。所以,在藏字字模库的建立等藏文信息处理研究中把藏字纵向叠加的部分作为一个整体来处理。把纵向拼写叠加的结构作为一个整体,称为整基字丁,即藏字中的上加字、基字、下加字、元音的组合(图 1 中的 2、3、4 及 5)。现代藏字中,藏文的整基字丁是一至四个字符叠加而成的,最多不超过四层。把藏字纵向叠加的部分作为一个整体后,藏字可以看出是前加字、整基字丁、后加字和再后加字的线性的排列,这就与其它文字的处理方式相同了,可减少了藏文处理的难度。表 5 是藏字全集中整基字丁的结构方式及所对应的个数的统计结果。

表 5 藏字全集中整基字丁的结构方式及对应的个数的统计结果

字符长度	结 构 方 式	组成整基字丁的个数	占所有整基字丁的百分比(%)
1 个字符	基字	30	4.878
2 个字符	基字 + 元音	120	19.512
	上加字 + 基字	33	5.366
	基字 + 下加字	43	6.992
3 个字符	上加字 + 基字 + 元音	132	21.463
	上加字 + 基字 + 下加字	15	2.439
	基字 + 下加字 + 下加字	2	0.325
	基字 + 下加字 + 元音	172	27.967
4 个字符	上加字 + 基字 + 下加字 + 元音	68	11.057

从表 5 中可以看出,2 个构件叠加的整基字丁占所有整基字丁的 31.87 %;3 个构件的叠加构成的整基字丁占所有整基字丁的 52.194 %,占所有整基字丁的一半多,是构成藏文整基字丁最主要的方式,总共有 615 个整基字丁。

6 结束语

梵音藏文和古藏文的处理在藏文信息处理中虽相当重要,但在现代藏文的使用中出现的概率很低,所以,藏文信息处理应该以现代藏文为主要的研究对象。对现代藏字基本属性的统计是研究现代藏文信息处理的基础,也是我们所开展地藏文处理研究工作的依据之一。

参 考 文 献:

[1] 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.

[2] 于洪志. 藏文内码扩展体系[J]. 中文信息学报,1999,13(1):50 - 58.

[3] 道布. 中国少数民族文字[M]. 北京:中国藏学出版社,1991.

[4] 胡书津. 简明藏文文法[M]. 昆明:云南民族出版社,1987.

[5] 瞿霭堂. 藏族的语言和文字[M]. 北京:中国藏学出版社,1996.

[6] 王浩军,赵南元,邓钢铁. 一种现代藏文笔段提取算法[J]. 中文信息学报,2001,15(4):41 - 52.

[7] 江狄,董颖红. 藏字叠加结构线性处理统计分析[J]. 中文信息,1994,(4):44 - 46.