维吾尔文字母频率统计及其应用

艾尼瓦尔 · 麦麦提　吐尔根 · 依布拉音
（新疆大学信息科学与工程学院，乌鲁木齐 830000）；
E-mail：enwer@xju.edu.cn

摘要：本文在超过 2000 万词汇的大量科学统计的基础上，提供了维文尔字母频率表。这对今后语言文字研究和制定科学的维吾尔文字输入键盘布局方案、维吾尔文字压缩算法的设计、语文类课程内容的补充等众多方面有着非常重要的参考价值。

关键词：维吾尔文、字母频率统计、键盘布局

引言

维吾尔文是由 32 个字母组成的拼音文字。众所周知，在实际使用过程中这 32 个字母出现的频率是不同的。虽然大家都知道这个客观事实，但从小学生的维吾尔文字母表到大学语文课程都没有提到有关维文字母使用频率的问题。而使用频率恰恰在维吾尔文科学的键盘输入布局方案的设计、维吾尔文压缩算法设计、维吾尔文智能输入法的设计和维吾尔文研究领域里有着非常重要的作用。

虽然语文研究工作者对维吾尔文字母频率有一个大概的理解，绝大多数人都不知道经过大量统计基础量化的指标。据我们所知，目前为止还没有一个科学的统计结果。这也是维吾尔文语文研究工作中的一个空白。

鉴于以上原因，我们对这几年收集的维吾尔文语料库进行了一次字母频率统计。统计语料库包括最具权威性的维吾尔语详解词典、汉维大词典以及具有代表性的各类文学作品、杂志、报纸等内容（所涉及的范围有科技、文化、教育、文学、历史、语言、经济、法律、工业、农业和牧业等），超过 2000 万词汇。统计结果见表 1。

1.频率统计表的应用

1.1 科学的维吾尔文字母键盘布局

目前正在使用的计算机维吾尔文键盘布局已经使用了 15 年。虽然它也是基于一定的频率统计，但是当时没有根据统计语言学的方法对诸多的领域语料库进行科学的统计。根据以上统计结果，很容易看出，维吾尔文字母键盘布局可以更有效。目前维吾尔文字母键盘布局中维吾尔文每个元音字母前的艾木扎（“ئ”）和字母“پ”都都属于高频率的符号和字母。因此可以把这些符号字母安排在键盘中心位置。当然在这里不是说非要修改已经使用了十几年的键盘布局，只是说明可以提出更有效的键盘布局。当然已经用了 10 几年有标准不容易改变，但是制定更有效、更科学的键盘布局，对专业打字业和专业排版业的文字输入工作是一份贡献。

1.2 数字键维吾尔文字母输入方案（小键盘、手机、电话机输入维吾尔文字母）

随着手机的普及，短信、彩信等业务的出现，而家加 e（固定电话短信）的投入使用，对维吾尔文字母的输入提出了一个新问题。使用 0-9 个数字键，手机或固定电话输入维吾尔文字母怎么办？通过 10 个键输入 32 个字母，有些字母可能通过两次甚至三次按键可以实现。那么这个顺序怎么确定？手机数字键上的 ABC、DEF 排列可以使用吗？第一，通过英文 26 个字母输入维吾尔文 32 字母时，需要借助于上档（SHIFT+）键，而手机键盘里没有 SHIFT 键。那么多余的 6 个维吾尔文字母没有办法输入，也就是说使用目前在手机键盘上的“ABC，DEF 字母布局不能输入维吾尔文字母。如果说通过烦琐的转换键可以输入维吾尔文字母时，其输入效率也会非

常低。根据以上频率统计结果，如果使用目前手机上的 ABC、DEF 布局，输入“ل”，“ر”，“ئا”等频率最高的字母需要按三次实现，甚至“س”需要四次按键才可以输入，而频率非常低的“ھ”，“ۇ”等字母安排在第一个位置上。

根据以上频率统计表和目前手机键盘上 ABC、DEF 键盘布局的需要的按键次数，不难计算出，输入效率（平均按键次数 / 每个维吾尔文字母）大于 3。也就是说要输入有 100 个维吾尔文字母的一句话，需要按下手机键 300 次以上。这种输入用户会很难接受，手机输入维吾尔文字母普及工作会受到很大影响。

根据以上统计结果，我们提出最有效的数字键盘维吾尔文字母输入布局。它有两个最大特点。一是有效性，二是容易记忆性。

（1）有效性

在这个键盘布局中，通过 1 键输入的维吾尔文频率为 58.37%，2 键输入的为 25.56%，3 键输入的为 12.16%，4 键输入的为 3.89%。

平均按键次数（输入效率）等于 1.6153。如果没有此统计表，这种键盘布局最坏时平均按键次数达到 3.3837。计算结果表明，在科学的键盘布局的键盘上，通过 10 个键输入 34 个字母（目前维吾尔文键盘布局中多两个字符 和 ），平均不是 3.4，而只是 1.6。这说明科学的键盘布局将大大提高文字输入速度。也就是说要输入以上举例的有 100 个维吾尔文字母的一句话，只需要按下手机键 161 次。通过 9 个数字键（0 作为空格和其它符号处理）输入 34 个不同的字符，同时可以实现自动选形，这样效率会更高的。

（2）容易记忆

键盘布局中的维吾尔文字母在保证高频字母优先的前提下尽量安排在与英语字母或该数字键同音或相似的位置，使得记忆过程比较容易。

1 ل و ز خ	2 ر ك چ ج	3 ن د ۇ ف
4 ت ۇ ې ۆ	5 ى ش ب ۋ	6 ا ئ غ ھ
7 ە ي لا ژ	8 م س گ	9 ق پ ك
	0 空格和其它符号	

表 3　数字小键盘维吾尔文字母部局图（字母从右到左优先）

（3）本频率统计表可以作为小学或中学语文或语法课程中的补充内容，以提高学生对本民族语言文字的认识和掌握。

2. 结论

结论，本文对维吾尔文字母的频率进行了全面和科学的统计，并给出了频率统计表，弥补目前没有一个科学的、权威的，根据大量统计基础上的统计结果的空白。同时在该统计表的基础上给出了几种用途，特别提出了数字小键盘输入维吾尔文字母最有效的键盘部局方案。该统计表将对语言研究者和计算机处理维吾尔文字工作有很高的使用价值。

序号	字母	国际音标	各类词典	报刊杂志	合计	频率(%)
1	ى	i	1,505,112	1,305,841	2,810,953	15.41
2	ا	a	910,327	591,312	1,501,639	8.23
3	ە	æ	600,657	506,689	1,107,346	6.07
4	ل	l	562,026	412,457	974,483	5.34
5	ر	r	542,791	420,097	962,888	5.28
6	ن	n	486,331	455,511	941,842	5.16
7	ت	t	503,421	379,719	883,140	4.84
8	ق	q	490,608	276,462	767,070	4.21
9	م	m	442,254	256,894	699,148	3.83
10	ئ	amze	382,612	281,900	664,512	3.64
11	ۇ	u	367,898	288,574	656,472	3.60
12	ك	k	378,910	234,538	613,448	3.36
13	ي	j	285,552	237,299	522,851	2.87
14	س	s	326,469	186,886	513,355	2.82
15	د	d	254,907	236,685	491,592	2.70
16	ش	ʃ	228,212	198,074	426,286	2.34
17	و	o	251,544	155,545	407,089	2.23
18	پ	p	220,424	144,254	364,678	2.00
19	ب	b	189,947	171,328	361,275	1.98
20	ې	e	206,748	132,377	339,125	1.86
21	ۋ	y	155,068	112,788	267,856	1.47
22	ز	z	146,874	104,784	251,658	1.38
23	چ	tʃ	151,769	90,812	242,581	1.33
24	غ	gh	128,244	89,522	217,766	1.19
25	ڭ	ŋ	94,813	103,021	197,834	1.08
26	گ	g	100,960	70,118	171,078	0.94
27	لا	la	92,619	76,599	169,218	0.93
28	ۋ	w	91,036	76,120	167,156	0.92
29	ھ	h	74,816	75,480	150,296	0.82
30	ۆ	ø	85,925	55,281	141,206	0.77
31	خ	x	66,139	57,994	124,133	0.68
32	ج	dʒ	51,901	42,321	94,222	0.52
33	ف	f	22,467	6,253	28,720	0.16
34	ژ	ʒ	1,650	1,097	2,747	0.02
	合计		10,401,031	7,834,632	18,235,663	100.00

参考文献：

[1]计算机操作技术 艾尼·阿布拉等 新疆科技卫生出版社(W)1995年
[2]维吾尔语正字正音水平测试大纲 麦麦提艾力等 新疆人民出版社 2003年10月
[3]数据结构 严蔚民 吴伟民 清华大学出版社 1992年6月
[4]C语言常用数值计算常用程序 苏海岛等 警官教育出版社 1996年10月

作者简介：艾尼瓦尔·麦麦提 硕士、出生于1975年5月。新疆喀什人。在新疆大学信息与工程学院工作、讲师。主要研究方向“自然语言处理”。

Statistics and Application of Uighur Alphabet Frequency

Ainiwaer.Mamai Tuergun.Yibulayin

[1](*Information Science and Engineering College of Xinjiang University, Urumqi 830000, China*);
E-mail: enwer@xju.edu.cn

Abstract: This article presented the Uighur Alphabet frequency based on the statistics of the over 20 million words. This is of great importance to make more efficient keyboard layouts, design the Uighur text compression methods, and can be used as the supplement material for the language and literature courses.

Key words: Uighur Alphabet Frequency, Keyboard layout, Data Compression