An Electronic Grammar of the Tibetan Syllables

Don Stilwell, Leonardo Gribaudo, Lee H. MacDonald, Marvin Moser, Chris J. Fynn, Pierre Robillard, Xavier Franc, Jim Parker, Uwe Ushakov, Alejandro Chaoul-Reich, Beat Steiner, Bernie Monette, Paul Hackett, Patrick Lottier, Sam Sirlin, Robert Taylor and Robert Chilton of the ACIP, and the rest of the Tibetan OCR team

Presented to the Joint Pacific Neighborhood Consortium (PNC) and Electronic Buddhist Text Initiative (EBTI) Conference,
University of California at Berkeley, Berkeley, CA — Jan. 11-17, 2000

Abstract

The Asian Classics Input Project (ACIP) has created a public database of 110 MB of Tibetan texts, giving glyph-to-glyph transcriptions of 1,031 classic books. This is a rich source of information about the allowable patterns in Tibetan syllables. In fact, these "one thousand books" contain 22.8 million syllables, a sizeable chunk of a kind of Tibetan used in scriptures, commentaries, liturgies, etc.

We created an expert system that enabled us to uncover and count the glyph patterns in the Tibetan dbu can (u-chen) script. The numeric data collected could not easily be used by human beings to make judgements on where syllable boundaries must be, but an expert system can rapidly and easily make these judgements using this kind of statistical data.

These data can be used by a machine to run down lines of glyphs without syllable indicators, and determine where the syllable boundaries most likely are and what are the "alternates," if any. Associated with each pattern is a probability that this pattern could be a syllable (p = count per 100 million syllables, for example).

But, if the glyph pattern is never seen in several large samples of Tibetan, then the probability that a glyph combination is a syllable is assigned to zero.

Tibetan syllables are generally composed of one of 259 common initial forms, with one of 5 common vowels (or 7 rare vowels), and possibly one of 25 potential common final forms. Both the initial and final forms are either one or two glyph-units wide. The vowel is attached to the last character in the initial form, either by being intrinsically there (A), or just above or below the last glyph in the initial form (as in all other vowels).

We now have tables of the frequencies of occurrence of each pattern and these can be downloaded from the Internet in Excel format (http://wwws.ghg.net/dstilwell/newdata.zip ~3-4 MB). It is of immediate use to Linguists in the form it is in now, or to CLIPS programmers in a CLIPS-formatted input file, and to all others as a series of DBF files, comma- or tab- separated lists, etc. (I'll help with each of the latter on request.)

Interestingly, the use of 600 exception syllables with 259 X 5 X 25 cells of initial, vowel, and final forms, is enough to explain virtually all of the syllables found in the 110 MB database.

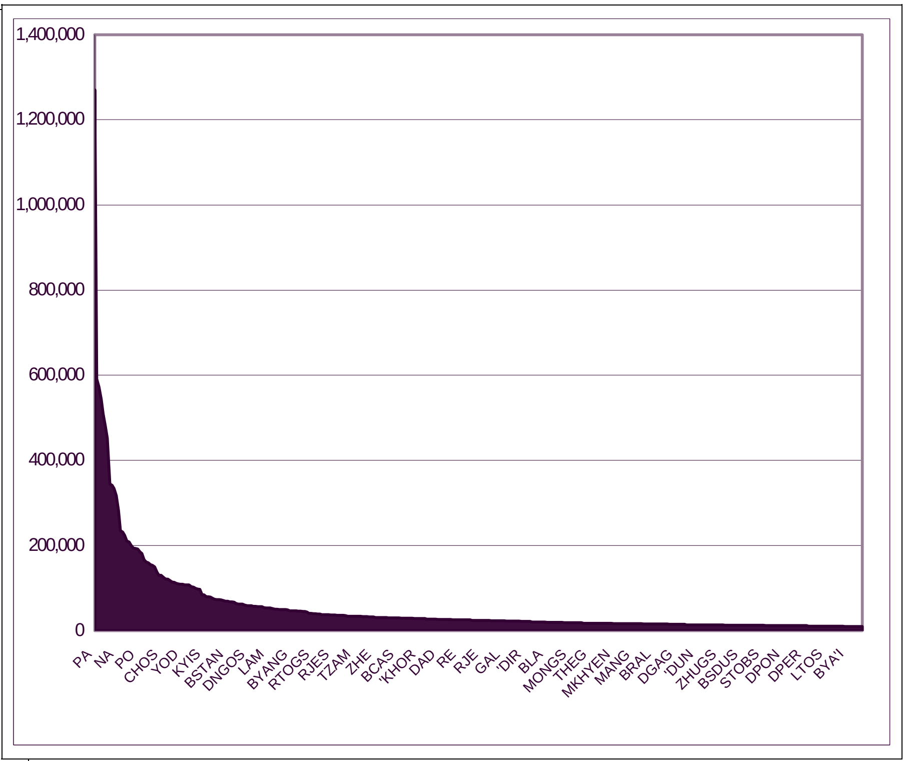


Fig. 1 - The number of times the most frequent 360 syllables were found in 1,031 Tibetan texts of the Asian Classics Input Project (ACIP) database.

The ancient Tibetan grammarians knew that there were intricate patterns in the Tibetan Language, but they did not have the modern digital computers needed to index and quantify Tibetan's deep structure.

For the first time, tools such as the NASA CLIPS expert system programming language and Excel can be applied to analyze this data.

Introduction

Tibetan existed purely as a spoken language until about the 7th Century AD, when monks and scholars created a script for it. Their sole purpose in developing a written form of Tibetan was to be able to translate the great works of Buddhism from Sanskrit into Tibetan. Since that time, much has been written about the structure of Tibetan syllables. The ancient Tibetan grammarians knew that there were intricate patterns in Tibetan script, but they did not have the modern digital computers needed to index and quantify Tibetan's deep structure.

Tibet, sitting on the high ground between the China and the Indian Subcontinent, has strategic value to the People's Republic and, hence, its army invaded Tibet in the 1950's. Since that time, more than a million Tibetans have been killed and thousands of monasteries, along with their libraries and cultural treasures, have been destroyed. Fortunately, many Tibetan texts and cultural treasures found their way out of the country on the backs of fleeing Tibetans, but many more have been lost forever.

To save what remains of this important cultural heritage, many transcription projects have arisen. Of all of the transcription programs, only the Asian Classics Input Project (ACIP) [see http://www.asianclassics.org] makes its texts freely available to scholars on the Internet. Hence, it is this database that forms the basis of the electronic syllable grammar described in this paper.

The ACIP texts used in this study include 1,031 classic texts that were typed into a standard computer using Roman characters to signify the precise Tibetan glyphs found in the texts.

Lists of these texts can be found at:
(1.) Sungbum Texts:
http://www.asianclassics.org/download/SungEng.html,
(2.) Tengyur Texts:
http://www.asianclassics.org/download/TengEng.html, and
(3.) Kangyur Texts:
http://www.asianclassics.org/download/KangEng.html.

We are grateful to the Monks of the Sera Mey Monastery, and other Gelug-pa monasteries in India, who typed these texts into computers from the original Tibetan documents. The ACIP database used in this study was 110 MB of ascii data and included 22.8 million syllables. It is the largest electronic collection of Tibetan texts available to scholars by far.

The Tibetan OCR Project is a group of about 25 individuals across the globe who are interested in providing a Free (Open Source) Optical Character Recognition (OCR) program to the Tibetan community to speed the input of these important texts into electronic format.

Although Don Stilwell is the chief author of this electronic grammar, he is indebted to many members of the Tibetan OCR group for their many contributions to this work. Dr. Lee H. MacDonald of Arizona provided software for performing syllable counts on ACIP texts. Marvin Moser, the author of Tibetan for Windows (http://members.aol.com/tib4win/), and Pierre Robillard, the author of the LTibetan font (http://www.interlog.com/%7Epierrer/) have made their software freely available on the Internet. This allowed us to easily represent the Tibetan characters in our work. And to Marvin Moser for doing some of the Tibetan typing, which the first author never learned. Leonardo Gribaudo of Florence, Italy provided syllable counts for the Rangjung Yeshe Tibetan Electronic Dictionary (http://home. earthlink.net/~rangjung/rypub/ry-dic.htm) which were used in this study. Uwe Ushakov provided the excellent ftp site at: ftp://storm.ptc.spbu.ru/pub/human-languages/tibetan/tibocr/ . Naturally, we are especially pleased with the support offered by Dr. Robert Taylor and Robert Chilton of the ACIP. Many others helped in significant ways.

Figure 1 on the previous page shows a plot of the number of times the most frequent 360 syllables were found in 1,031 Tibetan texts of the Asian Classics Input Project (ACIP) database. As we can show, by integrating the area under the curve, the first 360 most common syllables in Tibetan make up 19.8 million or 87% of the 22.8 million syllables in the database. By the time you go through the top 6,760 most common Tibetan syllables in the database, you have exhausted virtually all of the 22.8 million syllables in this 110 MB database.

The data produced by the expert system includes a 4-D mathematical model of essentially all of the syllables in a 110 MB text Corpus.

It may sounds interesting to some that the expert system codes information in 4th-Dimensional space, but we must remind everyone that this fourth dimension is only a mathematical one. For there is nothing mystical about the idea of attaching a probability to every syllable that is composed of an initial form (P1), a vowel (V), and a final form (P2).

Reading the Tibetan Syllables

We have discovered that virtually all of the syllables in database are accounted for by the top 6,760 most frequent Tibetan syllables plus a small list of about 600 syllable exceptions. Strangely, all of these syllables (except the 600 exceptions) can be built from 3 component parts: an initial part (P1) of one or two Tibetan base glyphs, a vowel (V), and a final part (P2) of one or two Tibetan base glyphs.

The initial and final forms (P1 and P2) are composed of one or two consonant glyphs. And the vowel is indicated either above or below, or both, of the final consonant in the initial form. The final form can contain vowels too, but there are a relatively small number of these final forms. And we treat these P2 vowels as just part of the final form in Tibetan syllables. By use of our expert system, we were able to determine that there are about 259 important one or two glyph initial forms and about 25 one or two glyph final forms. The number of vowels is even smaller and will be discussed after P1 and P2.

The 259 most important initial forms are shown on the next two pages. They account for the first one or two base glyphs in most Tibetan syllables. They are actually listed according to their frequency of occurence. PA (using ACIP notation) occurs the most frequently and CV occurs the least. This list is not the only one that could be made, but these represent observed frequencies in the 110 MB sample of ACIP texts. Not surprisingly, these same 259 initial forms (P1), 5 major vowels (V), and 25 final forms (P2), supplemented by the 600 exceptions (E) syllables, also account for ALL of the syllables in the Rangjung Yeshe Tibetan English Dictionary.

The final forms (P2) are only 25 in number and these too are supplemented by the list of 600 exception syllables. Like the initial forms, the final forms are either one or two base glyphs in length. There are often vowels associated with these 25 different forms, but they behave as if they are a fixed part of some final forms and need not be treated separately.

Not surprisingly, these same 259 initial forms (P1), 5 major vowels (V), and 25 final forms (P2), supplemented by the 600 exception syllables (E), also account for ALL of the syllables in the Rangjung Yeshe Tibetan English Dictionary.

ascii	tibetan	freq	P1	V	P2	C1	V1	C2	V2	C3	V3	C4	V4
PA	པ་	1,269,992	P	A		P							
PA'I	པའི་	591,208	P	A	'I	P				'	I		
DANG	དང་	572,456	D	A	NG	D				NG			
LA	ལ་	543,953	L	A		L							
BA	བ་	509,249	B	A		B							
PAR	པར་	479,586	P	A	R	P				R			
DE	དེ་	451,630	D	E		D	E						
MA	མ་	344,099	M	A		M							
NI	ནི་	341,693	N	I		N	I						
YIN	ཡིན་	333,446	Y	I	N	Y	I			N			
NA	ན་	317,321	N	A		N							
DU	དུ་	281,775	D	U		D	U						
KYI	ཀྱི་	234,872	KY	I		KY	I						
PHYIR	ཕྱིར་	231,444	PHY	I	R	PHY	I			R			
MI	མི་	223,685	M	I		M	I						
MED	མེད་	210,522	M	E	D	M	E			D			
ZHES	ཞེས་	208,355	ZH	E	S	ZH	E			S			
NYID	ཉིད་	200,347	NY	I	D	NY	I			D			
BYA	བྱ་	194,350	BY	A		BY							
YANG	ཡང་	193,393	Y	A	NG	Y				NG			
PO	པོ་	192,351	P	O		P	O						
BA'I	བའི་	185,192	B	A	'I	B				'	I		
LAS	ལས་	181,102	L	A	S	L				S			
BAR	བར་	165,982	B	A	R	B				R			
DAG	དག	160,585	D	A	G	D				G			

Fig. 2 - The 25 most common syllables in the 110 MB database. The first part of the syllable is shown as P1, the vowel as V, and P2 as the final form. The remaining breakdown is into glyphs/parts.

Fig. 3 - The next two pages show the 259 initial forms (P1) that account for the vast majority of Tibetan syllables.

1	P	པ་	21	GY	གྱ་	41	BC	བཅ་	61	SBY	སྦྱ་	81	SN	སྣ་	101	MY	མྱ་	121	BRTZ	བརྩ་
2	D	ད་	22	RN	རྣ་	42	GR	གྲ་	62	SD	སྡ་	82	RTZ	རྩ་	102	'CH	འཆ་	122	MNY	མཉ་
3	B	བ་	23	TS	ཚ་	43	KHY	ཁྱ་	63	GC	གཅ་	83	SP	སྤ་	103	DBY	དབྱ་	123	'TS	འཚ་
4	N	ན་	24	SH	ཤ་	44	MTS	མཚ་	64	SNY	སྙ་	84	MCH	མཆ་	104	BT	བཏ་	124	RKY	རྐྱ་
5	M	མ་	25	RGY	རྒྱ་	45	GN	གན་	65	DNG	དང་	85	BZ	བཟ་	105	KHR	ཁྲ་	125	GL	གླ་
6	L	ལ་	26	LT	ལྟ་	46	'J	འཇ་	66	'	འ་	86	DM	དམ་	106	BK	བཀ་	126	SPR	སྤྲ
7	Y	ཡ་	27	GS	གས་	47	LD	ལྡ་	67	GZ	གཟ་	87	BRT	བརྟ་	107	BSNG	བསྔ་	127	BSK	བསྐ་
8	S	ས་	28	BZH	བཞ་	48	DP	དཔ་	68	LH	ལྷ་	88	BRJ	བརྗ་	108	GD	གད་	128	'KHR	འཁྲ་
9	BY	བྱ་	29	ST	སྟ་	49	SGR	སྒྲ	69	SNG	སྔ་	89	TZ	ཙ་	109	BSGR	བསྒྲ	129	BSL	བསླ་
10	R	ར་	30	RT	རྟ་	50	BST	བསྟ་	70	SL	སླ་	90	SMR	སྨྲ	110	LNG	ལྔ་	130	BSG	བསྒ་
11	G	ག་	31	GZH	གཞ་	51	K	ཀ་	71	SPY	སྤྱ་	91	SM	སྨ་	111	PHR	ཕྲ་	131	GA-Y	གཡ་
12	KY	ཀྱ་	32	NG	ང་	52	PH	ཕ་	72	'DZ	འཛ་	92	'PH	འཕ་	112	BSD	བསྡ་	132	'PHR	འཕྲ་
13	ZH	ཞ་	33	BD	བད་	53	SK	སྐ་	73	BS	བས་	93	BR	བྲ་	113	BSKY	བསྐྱ་	133	ZL	ཟླ་
14	CH	ཆ་	34	DG	དག་	54	'GR	འགྲ་	74	BSH	བཤ་	94	MDZ	མཛ་	114	'TH	འཐ་	134	BTZ	བཙ་
15	PHY	ཕྱ་	35	SKY	སྐྱ་	55	DB	དབ་	75	MNG	མང་	95	'DR	འདྲ་	115	'B	འབ་	135	DPY	དཔྱ་
16	C	ཅ་	36	GNY	གཉ་	56	BL	བླ་	76	'G	འག་	96	MD	མད་	116	DK	དཀ་	136	RG	རྒ་
17	'D	འད་	37	KH	ཁ་	57	'BR	འབྲ་	77	GT	གཏ་	97	BRGY	བརྒྱ་	117	RD	རྡ་	137	RNY	རྙ་
18	NY	ཉ་	38	Z	ཟ་	58	'BY	འབྱ་	78	J	ཇ་	98	SG	སྒ་	118	GSH	གཤ་	138	MG	མག་
19	TH	ཐ་	39	'GY	འགྱ་	59	RJ	རྗ་	79	SR	སྲ་	99	MKH	མཁ་	119	GTZ	གཙ་	139	RM	རྨ་
20	T	ཏ་	40	MTH	མཐ་	60	DR	དྲ་	80	RDZ	རྫ་	100	'KH	འཁ་	120	MKHY	མཁྱ་	140	DGR	དགྲ་

141	RL	རླ་	161	BRL	བརླ་	181	'PHY	འཕྱ་	201	SHL	ཤླ་	221	BKL	བཀླ་	241	BCV	བཅྭ་
142	BLT	བལྟ་	162	SKR	སྐཅ	182	DH	དྷ་	202	S'	སཱ་	222	M'	མཱ་	242	d	ཅ་
143	BGY	བགྱ་	163	DKY	དཀྱ་	183	H'	ཧཱ་	203	N'	ནཱ་	223	MKHR	མཁྲ་	243	Ksh	ཀྵ་
144	BSNY	བསྙ་	164	BGR	བགྲ་	184	BSN	བསྣ་	204	DZNY'	ཛྙཱ་	224	DH'	དྷཱ་	244	DZ'	ཛཱ་
145	BZL	བཟླ་	165	DMY	དམྱ་	185	MGR	མགྲ་	205	D'	དཱ་	225	BH'	བྷཱ་	245	d'	ཆ་
146	KL	ཀླ་	166	SBR	སྦྲ་	186	W	ཝ་	206	P'	པཱ་	226	BSTZ	བསྩ་	246	DPR	དཔྲ་
147	A	ཨ་	167	SH'	ཤཱ་	187	GRV	གྲྭ་	207	t'	ཊཱ་	227	LP	ལྤ་	247	RY	རྱ་
148	SGY	སྒྱ་	168	BG	བག་	188	TR	ཏྲ་	208	R'	རཱ་	228	HR	ཧྲ་	248	KV	ཀྭ་
149	BSR	བསྲ་	169	MN	མན་	189	DKR	དཀྲ་	209	T'	ཏཱ་	229	TSV	ཚྭ་	249	LB	ལྦ་
150	RK	རྐ་	170	BRDZ	བརྫ་	190	DGY	དགྱ་	210	PR	པྲ་	230	TZTS	ཙྪ་	250	BRNG	བརྔ་
151	STZ	སྩ་	171	PADM	པདྨ་	191	BRKY	བརྐྱ་	211	A'	ཨཱ་	231	MR	མྲ་	251	BHY	བྷྱ་
152	BKR	བཀྲ་	172	LJ	ལྗ་	192	SMY	སྨྱ་	212	BRN	བརྣ་	232	GANGG'	གངྒཱ་	252	KHV	ཁྭ་
153	LC	ལྕ་	173	SH'AKY	ཤཱཀྱ་	193	BRK	བརྐ་	213	B'	བཱ་	233	DBR	དབྲ་	253	BLD	བལྡ་
154	SB	སྦ་	174	LK	ལྐ་	194	SHR'	ཤྲཱ་	214	G'	གཱ་	234	sh	ཥ་	254	PY	པྱ་
155	BRD	བརྡ་	175	RNG	རྔ་	195	BRG	བརྒ་	215	GH	གྷ་	235	SHR	ཤྲ་	255	LG	ལྒ་
156	BSGY	བསྒྱ་	176	'KHY	འཁྱ་	196	BSKR	བསྐཅ	216	SV	སྭ་	236	RB	རྦ་	256	DRV	དྲྭ་
157	BRNY	བརྙ་	177	DV	དྭ་	197	n	ཎ་	217	sh'	ཥཱ་	237	L'	ལཱ་	257	DZR	ཛྲ་
158	H	ཧ་	178	BH	བྷ་	198	RTZV	རྩྭ་	218	MGY	མགྱ་	238	SHV	ཤྭ་	258	BDZ	བཛ་
159	MJ	མཇ་	179	RV	རྭ་	199	KR	ཀྲ་	219	t	ཊ་	239	BKY	བཀྱ་	259	CV	ཅྭ་
160	TZ'	ཙཱ་	180	DZ	ཛ་	200	K'	ཀཱ་	220	ZHV	ཞྭ་	240	Y'	ཡཱ་			

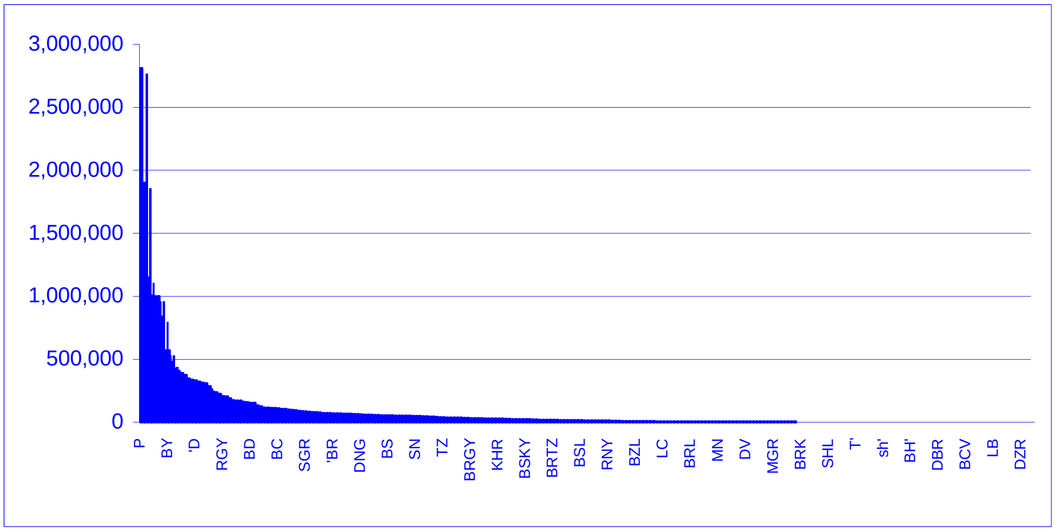
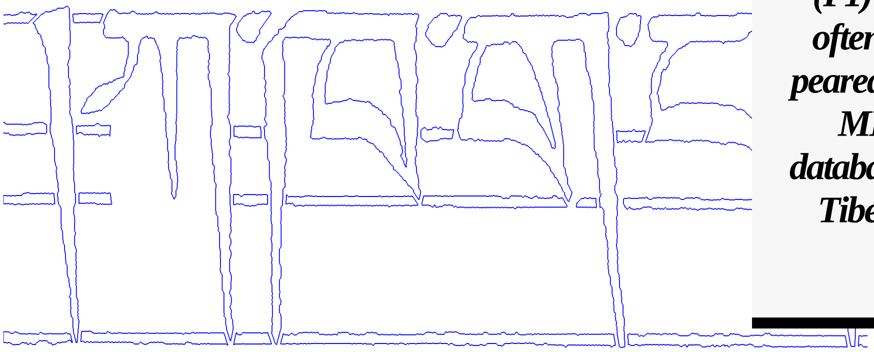


Figure 4 - This shows the number of times each of the 259 initial forms (P1) were found in the 110 MB database. With the most common intial form, PA being assigned to the lowest left-hand point on the x-axis. It is also assigned to 2.8 million on the y-axis, because that is the number of times PA was found in the ACIP database.



On the right —>

This table shows the first 32 of the 259 initial forms (P1) and how often they appeared in the 110 MB ACIP database of 1,031 Tibetan texts.

Figure 5-About 2.8 million (or 12%) of the 22.8 million syllables have PA as their initial form (P1). About 1.9 million (or another 8.3%) of all the 22.8 million syllables have DA as their initial form. Followed by BA at 1.1 million occurences, and so on.

name	Tibetan	counts
P	པ་	2,807,650
D	ད་	1,898,723
B	བ་	1,144,596
N	ན་	1,003,937
M	མ་	1,000,797
L	ལ་	999,083
Y	ཡ་	835,953
S	ས་	570,123
BY	བྱ་	569,873
R	ར་	477,139
G	ག་	424,300
KY	ཀྱ་	406,814
ZH	ཞ་	389,302
CH	ཆ་	373,161
PHY	ཕྱ	344,826
C	ཅ་	334,290
'D	འད་	330,850
NY	ཉ་	319,294
TH	ཐ་	311,994
T	ཏ་	307,753
GY	གྱ་	282,789
RN	རྣ་	240,798
TS	ཚ་	234,375
SH	ཤ་	222,292
RGY	རྒྱ་	204,311
LT	ལྟ་	201,232
GS	གས་	185,635
BZH	བཞ་	171,106
ST	སྟ་	169,469
RT	རྟ་	167,960
GZH	གཞ་	161,793
NG	ང་	159,220

	1	2	3	4	5	6	7	8	9
	xnil	xS	xN	xNG	xR	xD	x'I	xG	xGS
freq	8,012,418	2,347,871	2,022,994	1,983,920	1,611,375	1,530,514	1,060,747	1,052,529	652,336
		ས་	ན་	ང་	ར་	ད་	འི་	ག་	གས་
	11	12	13	14	15	16	17	18	19
	xL	xMS	xB	xNGS	x'O	x'	xBS	x'ANG	x'AM
freq	554,562	381,686	369,408	268,476	130,138	115,668	88,607	17,037	14,290
	ལ་	མས་	བ་	ངས་	འོ་	འ་	བས་	འང་	འམ་
	21	22	23	24	25				
	x'UR	xt	x'U'I	x'I'O	x'US				
freq	778	484	480	183	117				
	འུར་	ཏ་	འུའི་	འིའོ་	འུས་				

Figure 6 - The table above shows the 25 final forms (P2) that are found in the vast majority of Tibetan syllables.

In about 35% of all syllables there is no final form (P2 = xnil), all syllables must at least have an initial form (P1).

The 25 final forms (P2)

The 25 most frequent final forms (P2) are found in the majority of Tibetan syllables. In about 8 million of our 22.8 million syllable database, no final form was present, but the remaining 14.8 million syllables had either a one or two glyph final form. The two dozen or so P2 forms contrast sharply with the relatively high count, 259, of the P1 forms. This indicates, of course, that the range of initial forms, P1, is about 10 times greater than the range of final forms in Tibetan syllables. In our notation, we place an ‘x’ in front of the ACIP transliteration of the final forms, P2, like a variable, to indicate the position of the P1 initial form.

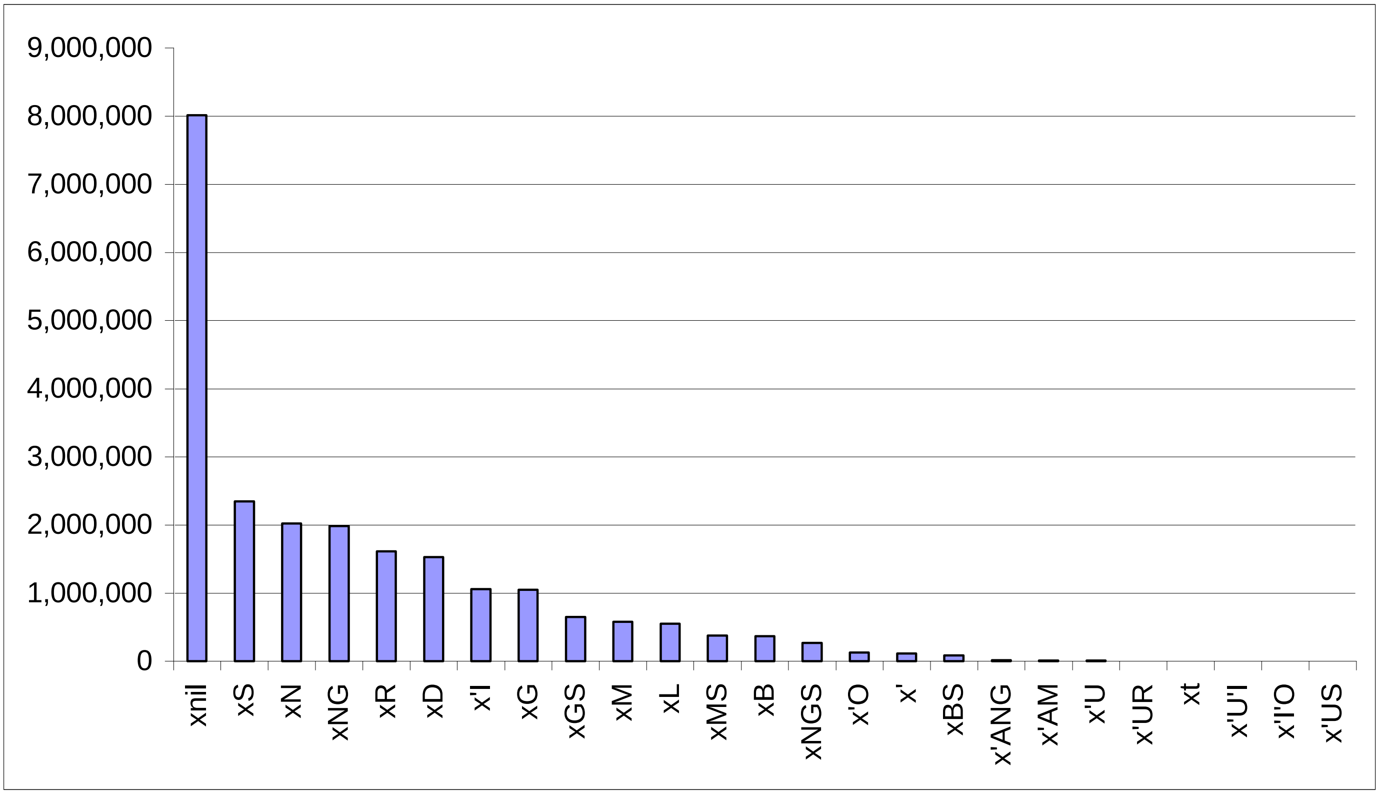
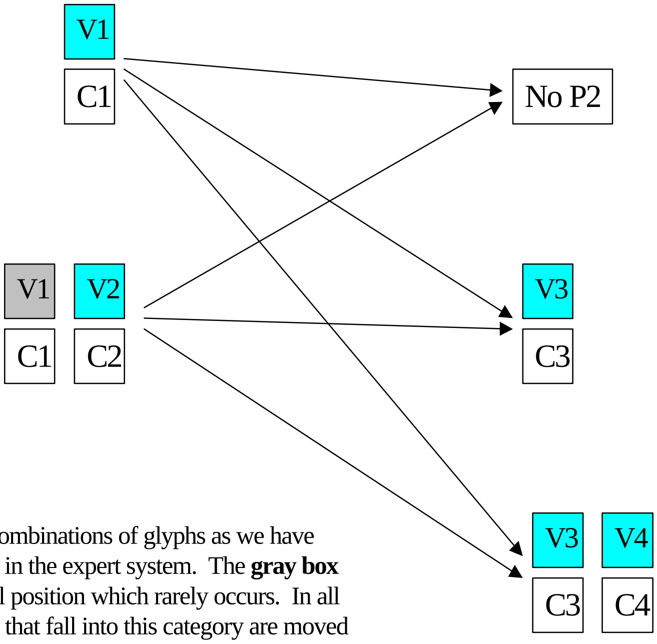


Figure 7 - A plot of the number of times each final form appeared in the 22.8 million syllable database. The x-axis represents the top-25 most frequent final forms (P2) with xnil (= syllables without final forms) at the highest point (about 8 million occurences). As we go across the top-25 P2s, it can be seen that number drops very rapidly for xS (SA) as a final form . This is a graph of the data that we showed to you numerically as "freq" in the table in Figure 6.



These are the combinations of glyphs as we have identified them in the expert system. The **gray box** show the vowel position which rarely occurs. In all cases, syllables that fall into this category are moved to the exception list. The **blue boxes** show those vowels glyphs which may or may be found in the average Tibetan syllable. The white boxes are consonants that may or may not occur in the average Tibetan syllable, except C1, which as a minimum, must always occur!.

Figure 8 - This shows the number of patterns that the gylphs can take in a standard Tibetan syllable. The way we have coded the expert system, the most common 6,760 Tibetan syllables — those syllables which are not on the exceptions list — always fit within one of these glyph patterns. The same is true for the 600 exceptions.

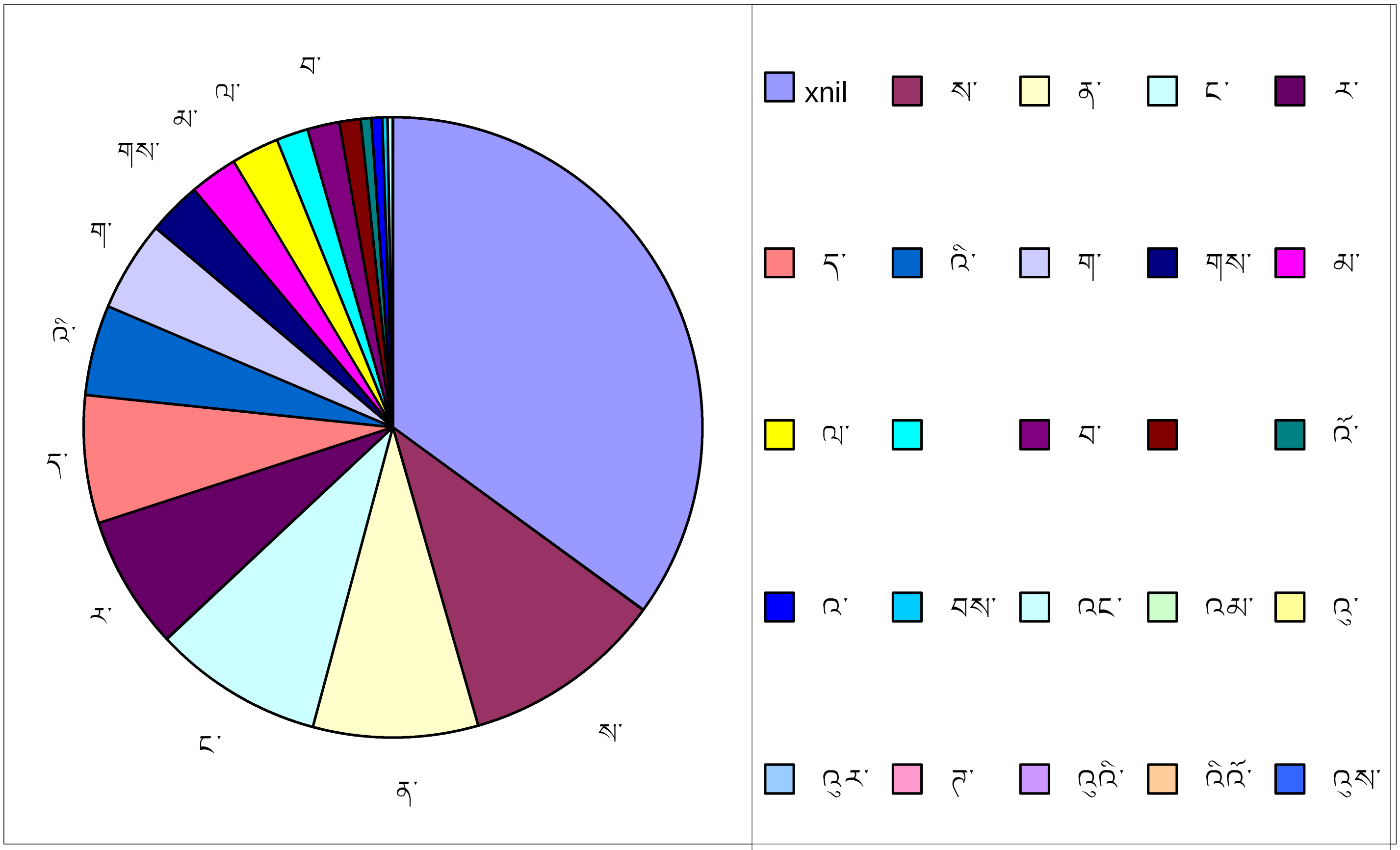


Figure 9 - The pie chart shows the percentage that each of the top-25 final forms (P2) has in the complete ACIP database. The largest purple region represents 35% of syllables in which there is no final form.

SA, is the most common final form followed by NA, NGA, and RA, in ACIP notation. Note that this is a graph for only 22,183,249 or 97.3% of the Tibetan syllables that follow the most common patterns.

1	xnil	xnil	8,012,418	35.1 %
2	xS	ས་	2,347,871	10.3 %
3	xN	ན་	2,022,994	8.9 %
4	xNG	ང་	1,983,920	8.7 %
5	xR	ར་	1,611,375	7.1 %
6	xD	ད་	1,530,514	6.7 %
7	x'I	འི་	1,060,747	4.7 %
8	xG	ག་	1,052,529	4.6 %
9	xGS	གས་	652,336	2.9 %
10	xM	མ་	580,444	2.5 %
11	xL	ལ་	554,562	2.4 %
12	xMS	མས་	381,686	1.7 %
13	xB	བ་	369,408	1.6 %
14	xNGS	ངས་	268,476	1.2 %
15	x'O	འོ་	130,138	0.6 %
16	x'	འ་	115,668	0.5 %
17	xBS	བས་	88,607	0.4 %
18	x'ANG	འང་	17,037	0.1 %
19	x'AM	འམ་	14,290	0.1 %
20	x'U	འུ་	6,003	0.0 %
21	x'UR	འུར་	778	0.0 %
22	xt	ཏ་	484	0.0 %
23	x'U'I	འུའི་	480	0.0 %
24	x'I'O	འིའོ་	183	0.0 %
25	x'US	འུས་	117	0.0 %
		TOTAL	22,803,065	

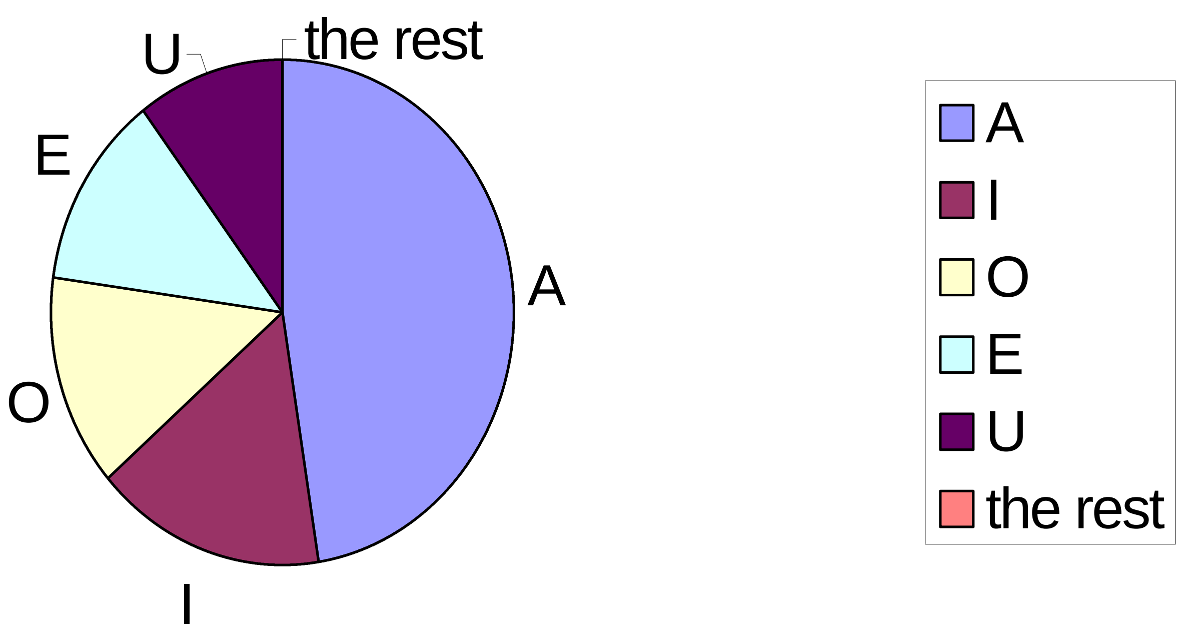


Fig. B - The table above shows the 5 common vowels (A, I, O, E, U in ACIP notation) and a tiny,microscopic red line near A to indicate the rest of the vowels (Am, EE, Em, i, Im, Om, OO, and Um), which in total acount for only 8,662 out of the 22,183,249 syllables that follow the 259 initial form X 5 vowel X 25 final form pattern.

Valid Tibetan Inital Forms in a Systematic Sort Order

x xY xR xL Gx Dx DxY DxR Bx BxY BxR BxL Mx MxY MxR 'x 'xY 'xR Rx RxY Lx Sx SxY SxR BRx BRxY BLx BSx BSxY BSxR

K KY KR KL DK DKY DKR BK BKY BKR BKL RK RKY LK SK SKY SKR BRK BRKY BSK BSKY BSKR

The Vowels (V)

The Main Vowel (V) of each syllable adheres to the last glyph in the initial form (P1).

So the order of each syllable is P1, V, P2 (if P2 is even present).

The main vowel, V, of each P1 X P2 combination adheres to the last glyph of the initial form (P1).

So the order of each syllable is P1, V, and P2 (if P2 is even present).

The absence of an explicit vowel indicator (I, O, E, U, Am, EE, Em, i, Im, Om, OO, and Um) means the implied (or inherent) vowel A is the main vowel of the syllable.

A	**10,863,087**	**47.6 %**
I	**3,624,685**	**15.9 %**
O	**3,141,609**	**13.8 %**
E	**2,836,071**	**12.4 %**
U	**2,328,925**	**10.2 %**
the rest	**8,662**	**0.0 %**

BG- B G SG NG SN BC C

MJ 'J RJ LJ BRJ

MNY RNY SNY BRNY BSNY

NY GNY

x xY xR xL Gx Dx DxY DxR Bx BxY BxR BxL Mx MxY MxR 'x 'xY 'xR Rx RxY Lx Sx SxY SxR BRx BRxY BLx BSx BSxY BSxR

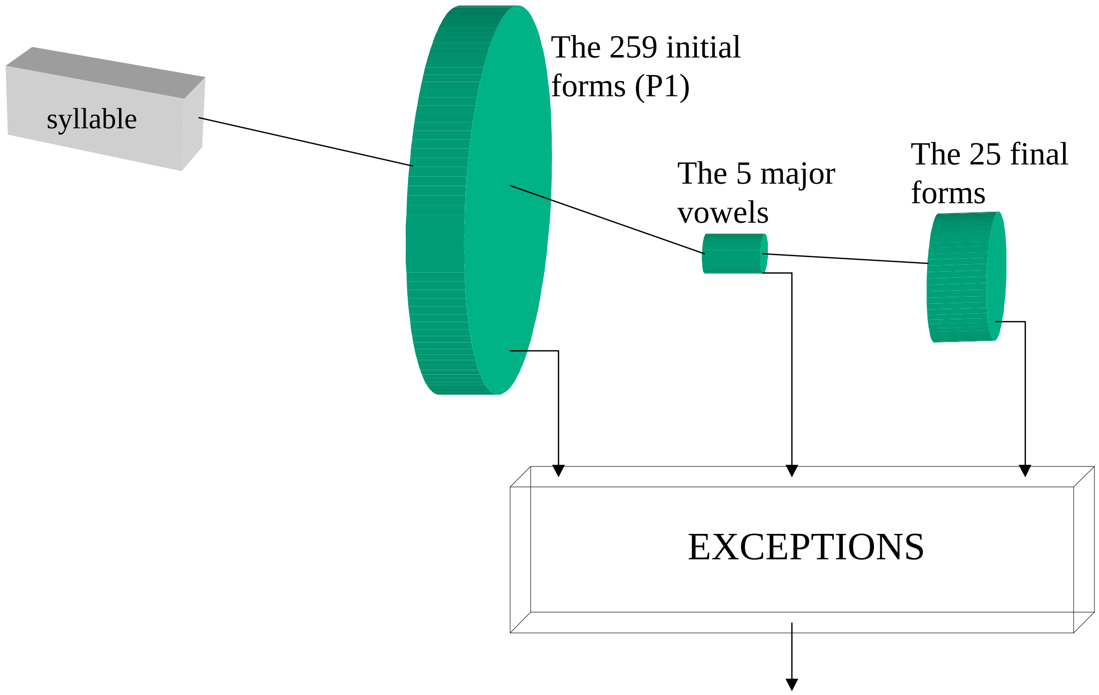

syllable
The 259 initial forms (P1)
The 5 major vowels
The 25 final forms
EXCEPTIONS
Residue = a tiny number of syllables that future users can choose to add to the exceptions list.

Figure 12 - Shows the structure of the model of Tibetan script that is encoded in our expert system. The use of an exceptions list allows the model to absorb new glyph patterns by assimilation — that is, to learn under a human's guidance. But the number of exceptions that we have so far is small — perhaps 600. If a syllable does not meet the most common patterns (259 X 5 X 25), they are compared against the exception list and if the whole syllable doesn't match any of those, it flagged, listed and counted.

Figure 11 shows counts for most of the punctuation marks.

ACIP	tibetan	freq
,	།	2,560,906
*	༄༅	31,196
;		23,264
`	༈	9,817
...	▪▪▪	2,099
#	༄༅༅	474
&		358
:	་	34
:*	་༄༅	22
*:	༄༅་	10
	total	**2,628,180**

The Expert System

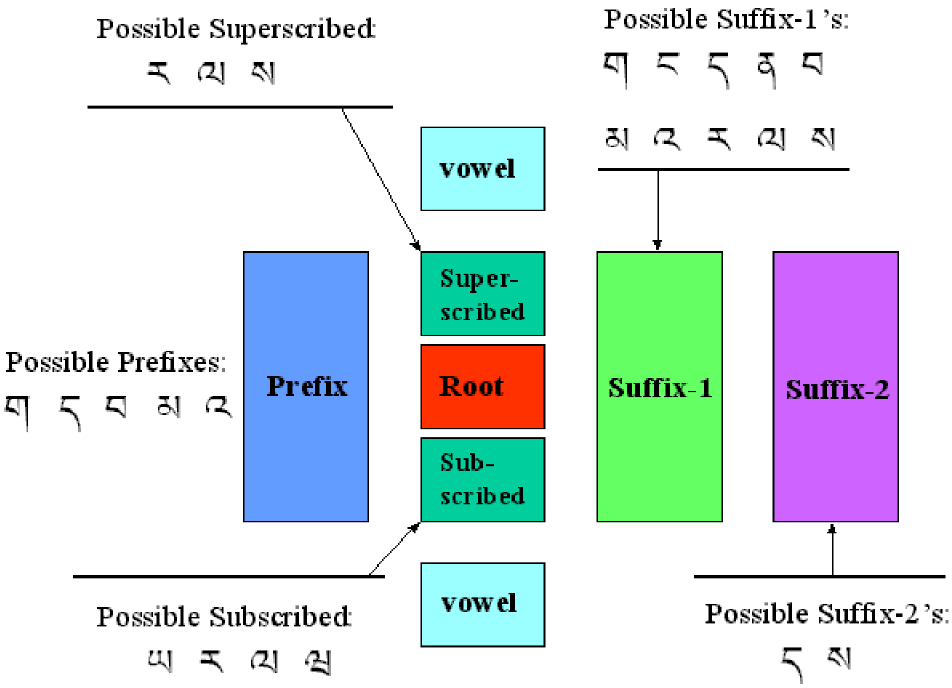
The current expert system matches more than 5 vowels, but we plan to add the rare vowels (Am, EE, Em, i, Im, Om, OO, and Um) to our exception list instead of matching them in our regular Tibetan syllable format (P1, V, P2).

We hope to soon do a detailed analysis on the Tibetan glyphs themselves. What frequency each is used in the 110 MB database, etc. , is temporarily unknown. The glyph analysis, which is extremely germane to the question of OCR, determines the frequency of occurrence of the various glyphs used in Tibetan.

The Expert System ideally should be structured as shown in Figure 12 above. It is a little different from our current system, but not much.

Our expert system analyzes the patterns in Tibetan dbu can (u-chen) script by concentrating on the frequency of each pattern in the 110 MB Corpus. It is much harder for us to imagine the number of patterns which are theoretically possible, but which do not occur. The value of this analysis lies as much in what it excludes from the realm of possibility as in what it includes. For, in practice, knowing what glyph combinations are <u>not</u> found in Tibetan, along with knowing what glyph patterns actually <u>are</u> found, is real knowledge that can be used in the creation of an intelligent Tibetan OCR system.

Figure 13 shows the ancient grammatical understanding of regular Tibetan syllables in a simple form. Syllables are made up of glyphs that have certain forms and not others. When we break these syllables up we see that only certain glyphs can be in certain positions. Of course, this only applies to syllables that follow this regular form. Many syllables do not! The use of an expert system allows matching and counting of more sophisticated patterns, including all exceptions that one can identify.



Our expert system analyzes the patterns in Tibetan dbu can (u-chen) script by concentrating on the frequency of each pattern in a 110 MB text Corpus.

It may sound interesting to some that the expert system codes information in 4th-Dimensional space, but we must remind everyone that this fourth dimension is only a mathematical one. For there is nothing mystical about the idea of attaching a probability to every syllable that is composed of an initial form (P1), a vowel (V), and a final form (P2).

The Expert System

There are about 259 common Initial Forms in Tibetan and they are shown in Figure 3 on pages 4 and 5.

We list these Initial Forms (P1s) in a special order — that is, from the most common ,PA, to the least common, CVA.

Following the list of the most common Initial Forms (P1s) there is a much shorter list of Final Forms (P2s).

The 25 common Final Forms (P2s) are shown in Figure 6 on page 7.

There are only 5 common vowels (which are denoted A, I, O, E and U as shown on page 10.

The general format for the expert system is shown in figure 12 on page 11. We do the following now, but the next version of the expert system will have a bit more power.

In this version of the expert system, the initial form (P1) of a group of syllables is detected by the first stage of the expert system. Then it will be passed on to the next stage for vowel matching. If the major vowel is detected as anything other than A, I, O, E, and U, then the current version of the expert system tallies them and breaks them down into P1, V, and P2. In the current version of the expert system, V was allowed to include 13 vowel planes rather than the 5 later settled on. In the current version of the expert system, the rare vowels (Am, EE, Em, i, Im, Om, OO, and Um) have their own vowel planes. On inspection of them, we decided we would include these in the exceptions list for the next version of the system.

In both the current and the future version of the expert system, the number of times each combination was triggered is continually summed as the data is analyzed. The current and future system also breaks down syllables into ordered glyphs (C1, V1, C2, V2, C3, V3, C4, V4) for the glyph analysis and data mining we will do next.

By filling a series of cells in a 259 X 5 X 25 three-dimenisional cube with counts of how many times each syllable occurred, we store a number related to the Bayesian probability that a combination of glyphs exists. This is a fourth dimension in our data.

By adding an exception list we allow learning.

For each cell or item in the exception list, a combination of glyphs (C1, V1, C2, V2, C3, V3, C4, V4) is identified and is a fifth dimension of information in our syllable grammar, which we will analyze next.

In the new system, if no common Tibetan Initial Form (P1) is found, then the exception list (E) will be consulted, if the syllable is not found in E, then this data is passed back to the CLIPS programmer as a list of non-matches with frequencies.

But the really interesting information about the Tibetan syllables comes when the structure of the syllables can be dissected by computer pattern matching. Simultaneously, we break down the tsyllables into their component glyphs.

The expert system provides the knowledge base during pattern matching. Furthermore, such a system can provide extremely powerful data handling, such as making global counts of key linguistic features of Tibetan script. The data produced by the expert system includes a 4-D mathematical model of essentially all of the syllables in a 110 MB text Corpus.

Our Results

What you can see from the previous graphs is remarkable. But, by far, the most interesting result comes from 5 separate 3-Dimensional graphs, called vowel planes. These vowel planes encode information about the frequencies of each combination of P1 X P2 for each major vowel (A, I, O, E, U).

By using the complete 110 MB ACIP version 4 data set, we were able to produce a very accurate electronic grammar of Tibetan syllables. It is a grammar in that it tells us, based on one 110 MB corpus, whether a combination of glyphs could be a syllable or not. In the texts we have used, however, we may be missing a greater density of mantras, and other syllables which have no lexical meaning, that might be in Tantric sources not in the public ACIP database. Christopher Fynn pointed out that there may be a much richer soil for Sankritized Tibetan in Tantric materials. We hope to negotiate a contract on ACIP's terms that would allow us to analyze the Tanric material next. If we can not negotiate a contract on ACIPs terms, then we can only ask for the initiations necessary based on recognition of the need to do this archaeology of the Tibetan syllables.

ascii	tibetan	Echars	freq	C1	V1	C2	V2	C3	V3	C4	V4
PAndI	པཎྜི	5	2,920	P		nd	I				
X	X	1	1,572	X							
PAn		3	1,357	P		n					
x	x	1	1,342	x							
BANDE		5	670	B		ND	E				
BADZRA	བཛྲ	6	595	B		DZR					
t'IK	ཊཱིཀ	4	535	t'	I	K					
DHARMA	དྷརྨ	6	491	DH		RM					
SARBA	སརྦ	5	444	S		RB					
MAndAL	མཎྜལ	6	425	M		nd		L			
WARMA	ཝརྨ	5	345	W		RM					
MANYDZU	མཉྫུ	7	325	M		NYDZ	U				
TH'A	ཐཱ	4	320	TH'							
RAKshI	རཀྵི	6	309	R		Ksh	I				
A'A:	ཨཱཿ	4	297	A':							
AUTPA	ཨུཏྤ	5	294	A	U	TP					
SENGGE	སེངྒེ	6	293	S	E	NGG	E				
BUDDHA	བུདྡྷ	6	287	B	U	DDH					
TZANDRA	ཙནྡྲ	7	286	TZ		NDR					
t'IKA	ཊཱིཀ	5	283	t'	I	K					
PRADZNY'A	པྲཛྙཱ	9	263	PR		DZNY'					
MAH'A	མཧཱ	5	258	M		H'					
RATNA		5	257	R		TN					

Figure 14 shows some of the most common syllables in the exception list. An exception list allows the expert system to simply look up syllables that don’t fit into our selection of the set of P1, P2, and V forms.

It would be possible to expand the 259 P1s, 25 P2s, and 5 major Vs to include all of the combinations found, but it is only reasonable to limit the time and memory used to represent the Tibetan language. We account for syllables that do not fit our most common P1s, P2s, and Vs by matching them against syllables in the exception list directly and immediately.

Additionally, when we use an exception list to reduce the memory and computational times required to match Tibetan syllables, it then becomes our policy to relegate to the exceptions list any correct Tibetan syllable that does not fit into the 259 most common P1s, the 25 most common P2s, and the 5 major vowels (A, E, I, O, U). This means that we relegate any syllable that contains the minor vowels (Am, EE, Em, i, Im, Om, OO, and Um) to the exception list, because they do not occur frequently enough for us to separately map their P1 and P2 vowel planes. You will see this shortly in subsequent Figures. Similarly, P1s and P2s that are rare or which do not form many combinations are also arbitrarily assigned to.the exception list.

Although one could say that we are motivated in part by speed and memory limitations, the use of an exceptions list is actually a speed increasing method if the number of times each exception syllable occurs is fairly low. We expect that the use of such a system would eventually result in the learning of many exceptions over time.

			xnil	xS	xN	xNG	xR	xD	x'I	xG	xGS	xM	xL	xMS	xB	xNGS	x'O	x'	xBS	x'ANG
name	tibetan	r		ས་	ན་	ང་	ར་	ད་	འི་	ག་	གས་	མ་	ལ་	སྨ་	བ་		འོ་	འ་	བས་	འང་
P	པ་	1	1,465,278	160,255	411	188	496,620	1,532	622,714	75	394	23	50			7	48,349	106		4,348
D	ད་	2	791,058	67,319	115,204	576,546	32,530	29,180	69,686	166,176	4,404	38,785	2,766	30	432	184	2,307	21	9	1,910
B	བ་	3	628,916	56,301	6,036	3,119	182,504	7,669	203,587	7,473	139	16,230	1,475	160	1,875	55	22,100	64	970	1,413
N	ན་	4	737,173	186,780	2,974	16,636	11,715	5,311	2,632	3,815	1,482	17,807	84	7,789	2,709	395	3,671	6	4	2,802
M	མ་	5	611,759	18,375	33,632	33,112	17,255	216,163	30,808	15,402	21	620	2,450			19,060	1,286	5		259
L	ལ་	6	572,686	209,596	37,414	21,814	1,240	338	2,789	37,201	41,075	53,887		42	324	8,742	3,004		66	4,640
Y	ཡ་	7	45,575	12,068	368,769	194,876	1,672	133,781	97	3,416	106	2,975	36,969	99	2,061	33,426	36			24
S	ས་	8	232,320	2,203	3,983	16,847	15,501	1,380	5,737	695	113,372	28,183	6,585	107,662	99	34,720	317		10	284
BY	བྱ་	9	206,296	66,757	16,103	83,771	4,162	141,315	10,197	687	225	9	284	5,686	10	219	33,851	12	12	42
R	ར་	10	111,787	10,297	12,234	141,661	863	732	4,625	35,406	46,993	17,852	29,406	2,364	58,721	2,084	90		1,902	64
G	ག་	11	172,980	71,375	1,041	131,934	7,520	757	3,686	226	1,826	2,995	23,586	4,220	392	1,053	315	3		136
KY	ཀྱ་	12	236,806	84,807	94	83,981	165	32	96	60	4	4	180		7	4	464	3		80
ZH	ཞ་	13	48,847	212,558	9,600	34,341	1,201	65	639	54,015	13,714	1,637	5,997	9	1,306	8	305		4,856	
CH	ཆ་	14	48,738	138,002	62,322	12,933	9,184	28,697	4,037	5,685	24,260	684	338	359	36,903	265	537		128	31
PHY	ཕྱ་	15	29,069	4,439	25,020	2,423	231,687	3,960	1,344	15,227	31,438	36	46		6	57	49			
C	ཅ་	16	39,856	31,024	107,633	34,697	4,140	58,215	7,396	50,786	5	92	181		4	8	72			124
'D	འད་	17	120,491	53,987	16,285	505	23,650	74,733	7,523	8,497	6,486	1,136	5,750	873	1,338	100	606	3,063	5,709	94
NY	ཉ་	18	32,192	17,206	39,420	2,796	6,644	200,467	151	457		1,124	2,380	16,375		42	31			
TH	ཐ་	19	46,810	18,041	8,622	3,680	12,326	1,969	99	39,727	14,763	1,141	57,222	57,903	35,946	128		17	13,537	
T	ཏ་	20	276,408	397	14,882	7,475	1,032	13	651	5,313	19	625	368		21		304	7		167
GY	གྱ་	21	156,544	68,054	1,095	137	56,155	268	25	6	18	5	200		3		227			32
RN	རྣ་	22	2,894	69	940	11	54		15	392	26	118,437	6,718	111,232	10					
TS	ཚ་	23	42,455	2,433	8,621	7,457	9,187	31,527	910	23,401	45,467	5,689	50,759	442	773	4,666	101		382	79
SH	ཤ་	24	15,213	137,229	13,461	28,736	7,564	1,457	1,112	11,001	3,428	602	1,902	38	48	165	143	99	42	
RGY	རྒྱ་	25	68,073	47,370	18,091	1,215	3,704	22,446	3,874	365	558	31	36,986		898	150	463			16
LT	ལྟ་	26	73,851	11,593	70	15,550	98,282	144	850	488	90	5			24	3	184		16	67
GS	གས་	27	3,200	645	2,403	14,554	8,954	2,523	37	1,259	69	67,060	34,163		272	50,488	4	4		

Figure 15 - This is a plot of the P1 X P2 plane collapsed across all of the vowels, both major and minor vowels. This table shows the upper right hand region (near the axis) of the the major verb planes (planes A, I, O, E, U) that we plot in Figures 16 through 20.

The P1 X P2 Plane

Figures 16 through 20 show the combinations of P1 X P2 and the number of times they occurred in the 110 MB ACIP database for each of the major vowels (A, I, O, E, U in ACIP notation).

One way to look at this P1 X P2 plane is to examine a table of points near the axis. These are where most of the activity is taking place in creating Tibetan syllables. Notice the figure above, Figure 15, for a look at the top-26 initial forms (P1) along with the top-18 final forms (P2).

The P1 X P2 Vowel Planes

Figures 16 thru 20 show the combinations of P1 X P2 and the number of times they occurred in the 110 MB ACIP database for each of the major vowels (A, I, O, E, U in ACIP notation).

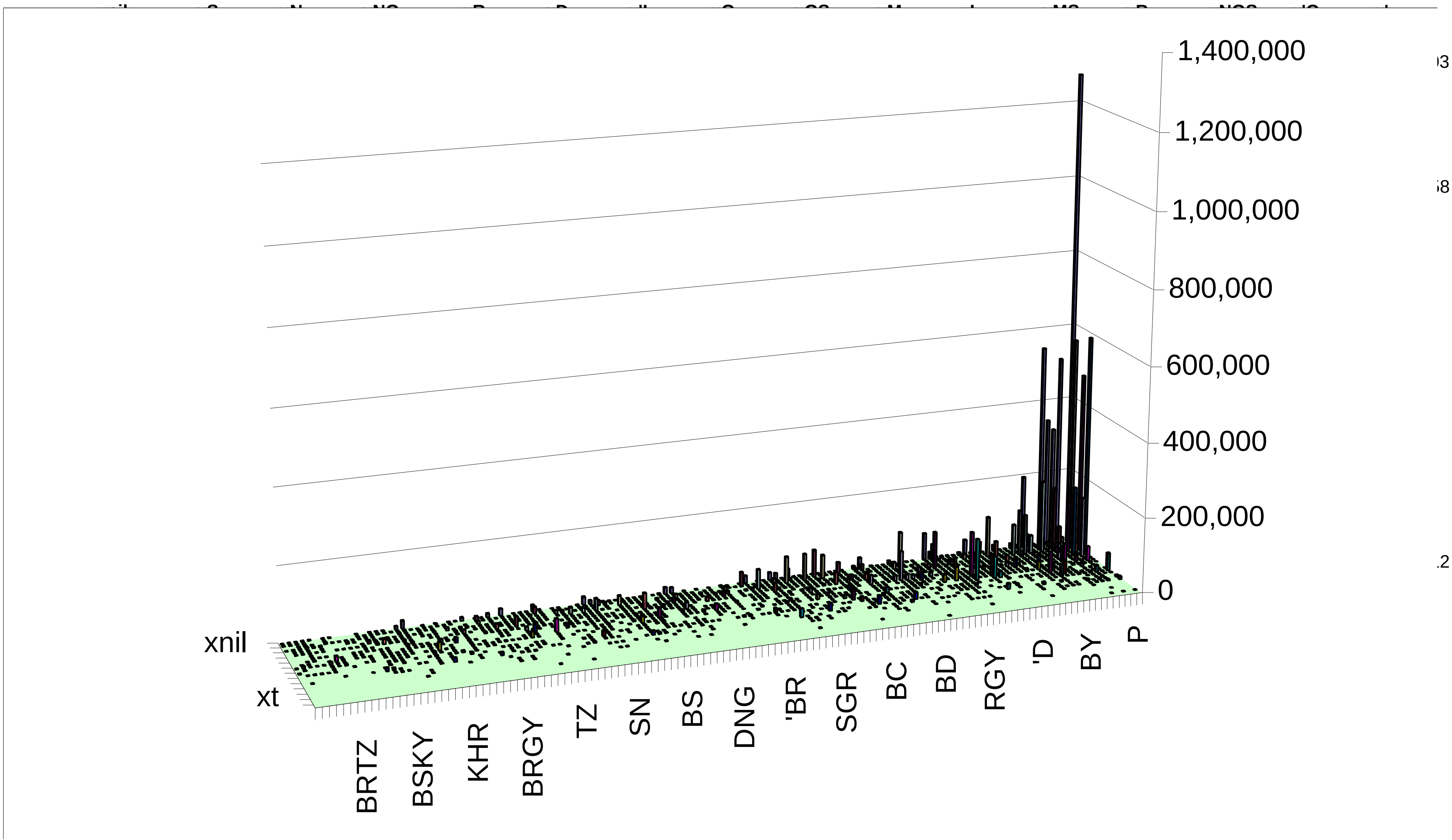
Each of the plots of P1 X P2 shows the number of times the P1 and P2 combination occurred for each of the major vowels (A,I, O, E, U).

From scanning the vowels planes (V), we see that there are clearly many combinations in each graph in figures 16 to 20 which do not occur in the 110 MB database. The 259 most common initial forms (P1), the 5 major vowels (V), and the 25 most common final forms (P2) do not contain anywhere near the number of possibilities that could be formed from the basic Tibetan glyphs. We can give it just a little thought to realize that there are far more combinations that do not exist, than those that do. It is far more likely that we would reject any random combination of allowable glyphs as a potential syllable, rather than accept it. With such tight constraints on what is possible in Tibetan, it is clear that sliding a 1-, 2-, 3-, or 4- glyph window along recognized glyphs will reject far more combinations as syllables than it accepts. With the addition of frequency of occurence information, this sliding window technique can provide a “betting man” with an excellent indication of where syllable boundaries are and where they can not occur. This is just one use of this data in Tibetan OCR. We believe we can do the latter matching with Fuzzy CLIPS.

Additionally, because the syllables are automatically broken out into individual glyphs and vowel indicators (C1, V1, C2, V2, C3, V3, C4, V4), a frequency distribution of glyphs can also be constructed from this data. We plan to complete that in the next month or two. Figure 8 way back on page 8 showed how these are encoded in the expert system’s 259 X 5 X 25 matching and counting universe.

In general, Tibetan syllables are 1-4 glyphs wide. They are formed by combining a single or dual glyph intial form (P1 = C1, V1, C2, V2) with either a one or two glyph final form (P2 = C3, V3, C4, V4).

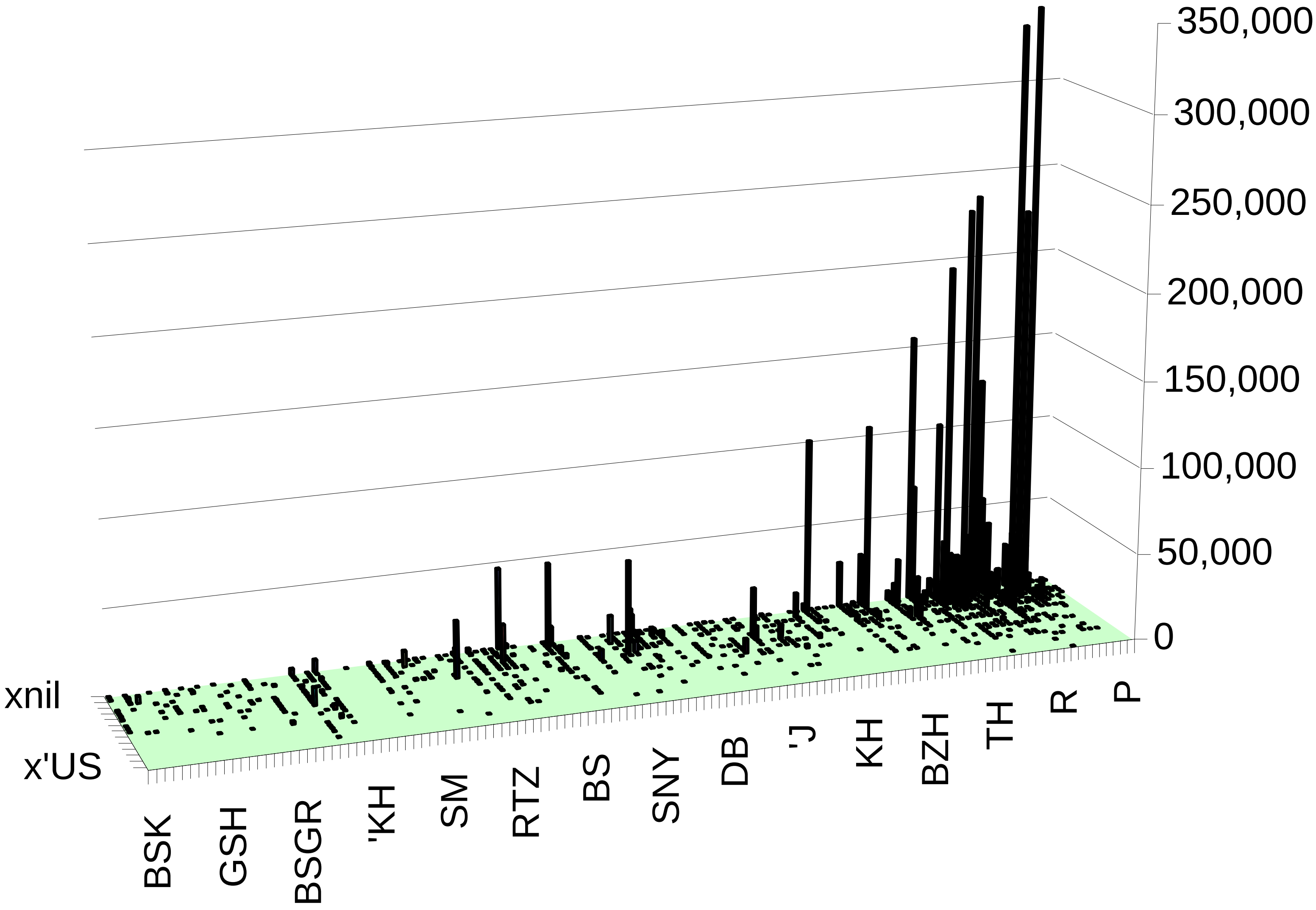
In general, Tibetan syllables are 1-4 glyphs wide. They are formed by combining a single or dual glyph intial form (P1 = C1, V1, C2, V2) with either a one or two glyph final form (P2 = C3, V3, C4, V4).



Many syllable combinations of P1 and P2 do not result in an allowable syllable with an inherent A as the major vowel in the syllable.

The P1 X P2 Vowel Plane for A

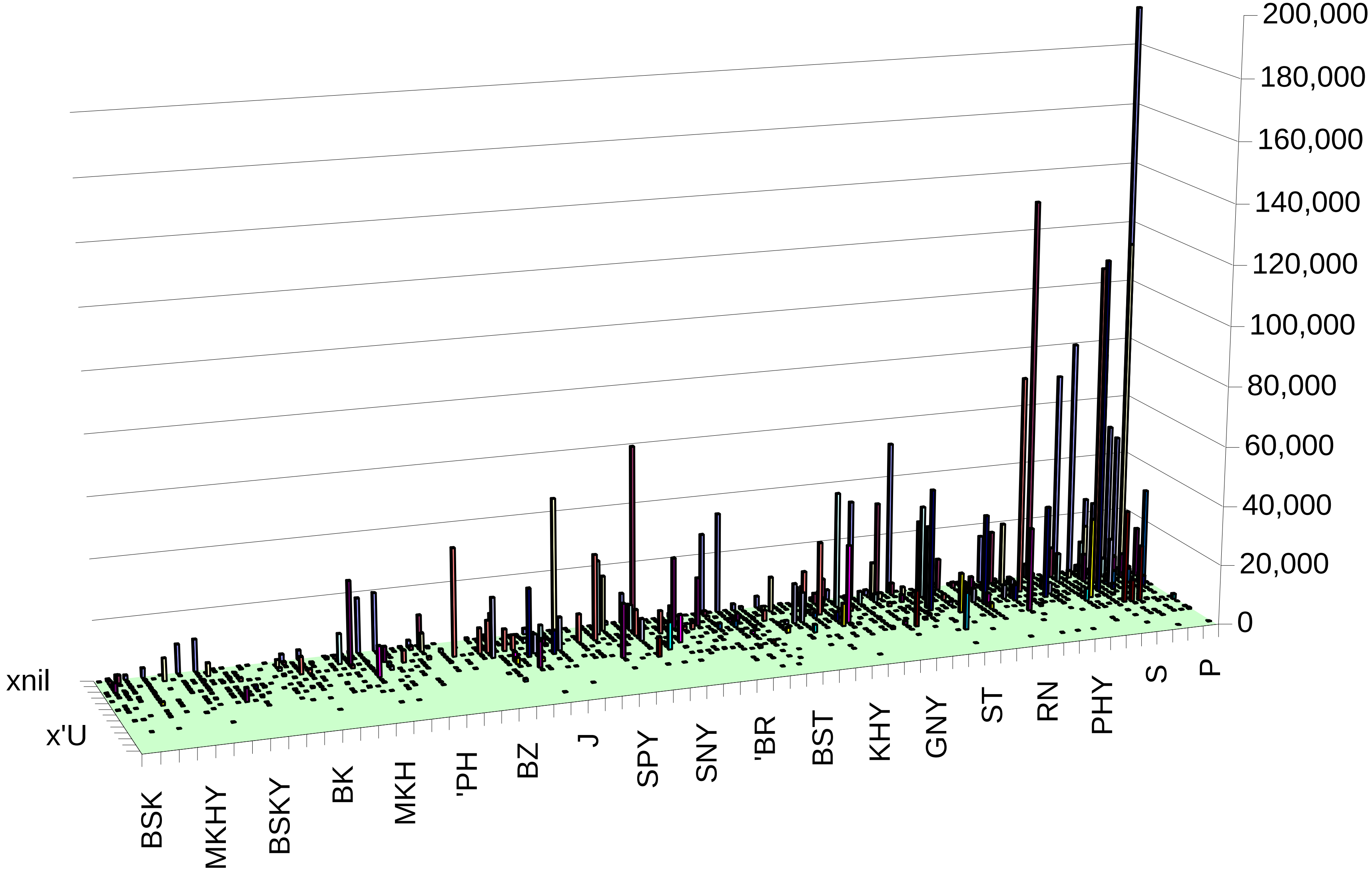
Figure 16 shows how many syllables used combinations P1 and P2 for the inherent vowel A. The vowel A is indicated by the absence of a vowel mark at the last glyph of P1.



There are a very large number of P1 and P2 combinations that do not result in allowable syllables when I is the major vowel.

The P1 X P2 Vowel Plane for I

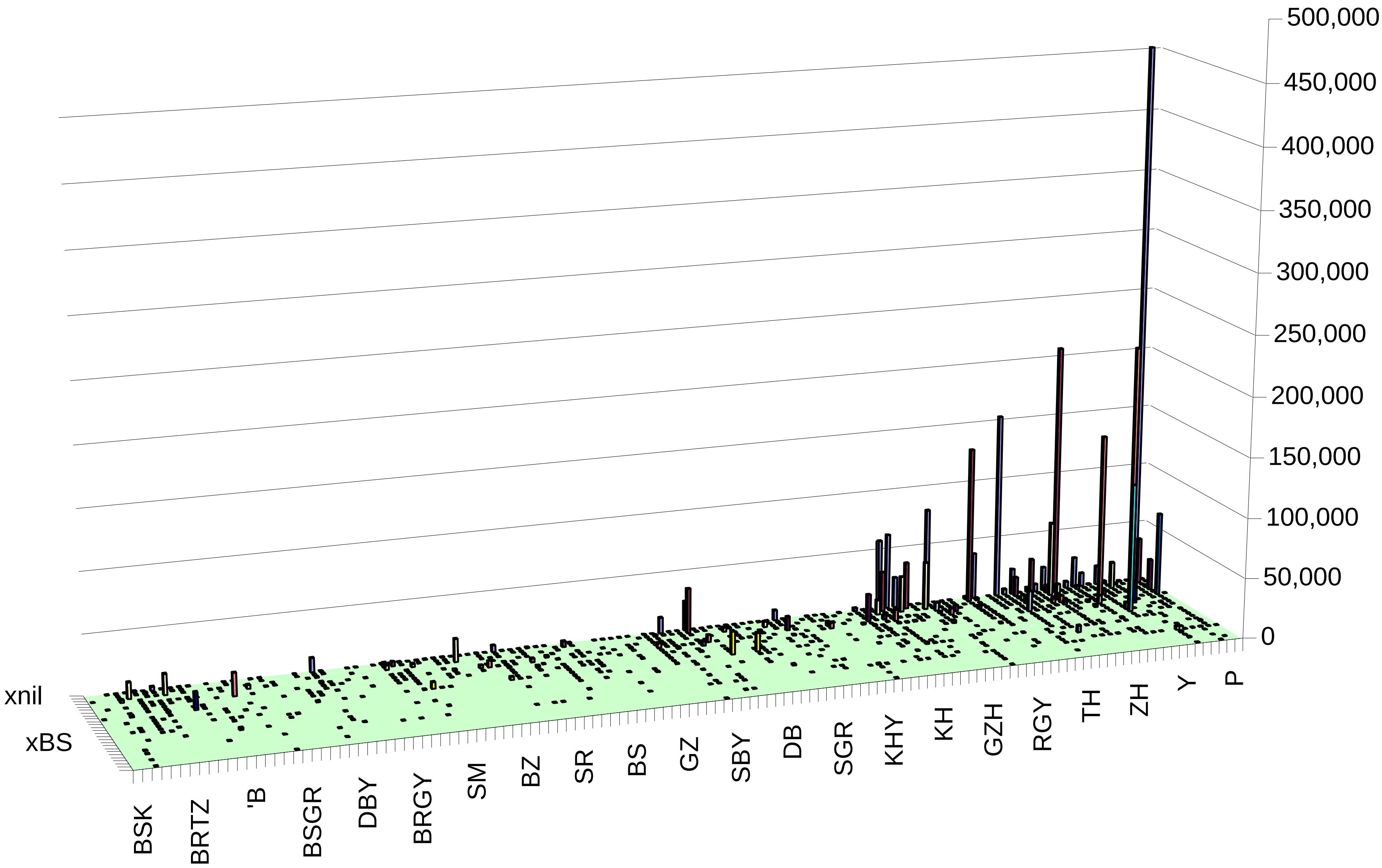
<u>Figure 17</u> shows how many syllables used combinations P1 and P2 for the vowel I. The vowel I is indicated by a vowel mark, I, at the last glyph of P1.



The P1 X P2 Vowel Plane for O

Figure 18   shows how many syllables used combinations P1 and P2 for the vowel O. The vowel I is indicated by a vowel mark, O, at the last glyph of P1.

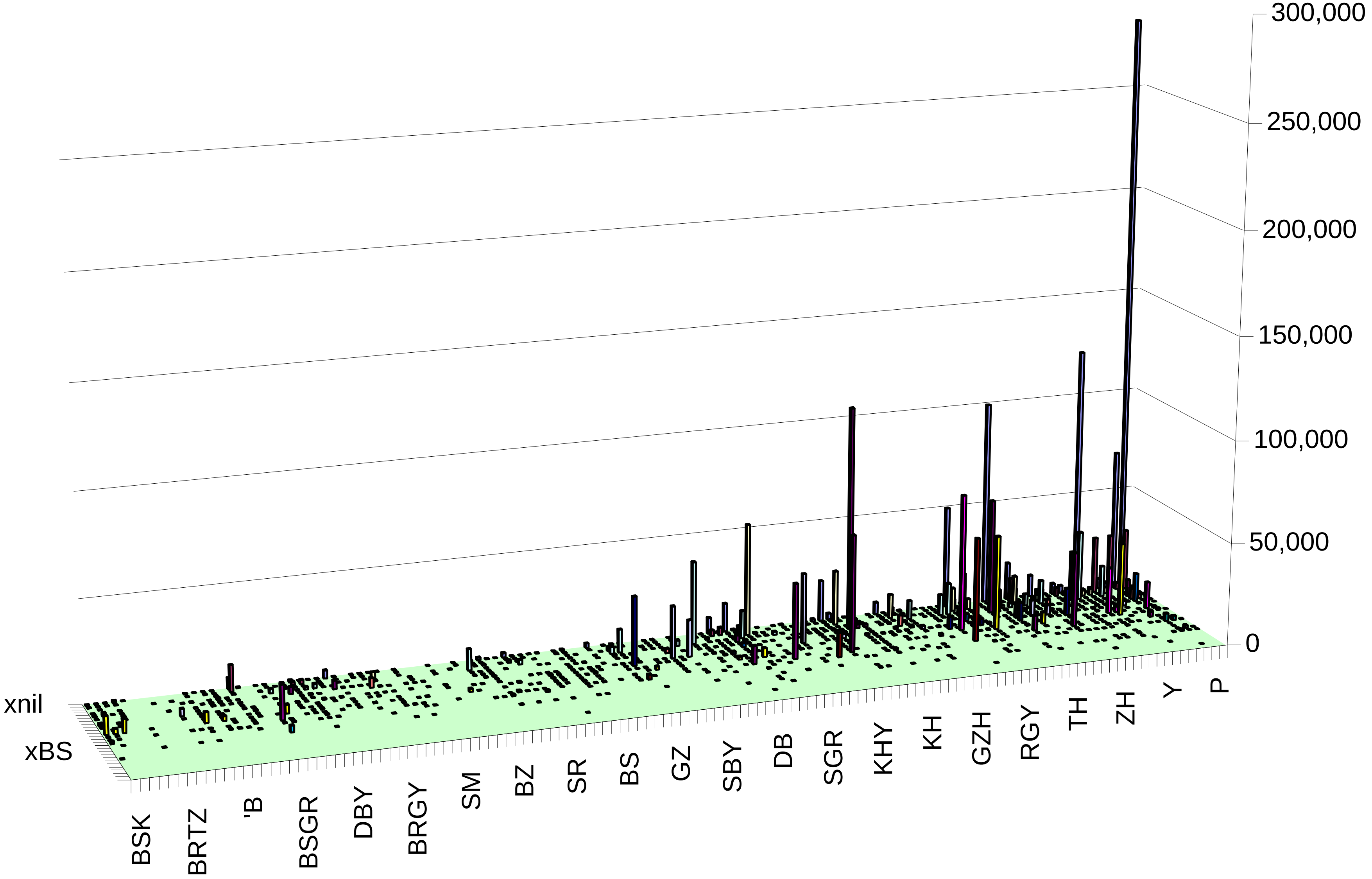
Many combinations of P1 and P2 do not result in allowable syllables with O as the major vowel .



The P1 X P2 Vowel Plane for E

Figure 19 **shows how many syllables used combinations P1 and P2 for the vowel E. The vowel E is indicated by a vowel mark, E, at the last glyph of P1.**

Many syllable combinations of P1 and P2 do not result in an allowable syllable with an E as the major vowel .



Many syllable combinations of P1 and P2 do not result in an allowable syllable with an U as the major vowel .

The P1 X P2 Vowel Plane for U

__Figure 20__ shows how many syllables used combinations P1 and P2 for the vowel U. The vowel U is indicated by a vowel mark, U, at the last glyph of P1.

ascii	tibetan	freq	P1	V	P2	C1	V1	C2	V2	C3	V3	C4	V4
SAm	སཾ་	293	S	Am		S	Am						
PAm	པཾ་	236	P	Am		P	Am						
LAm	ལཾ་	203	L	Am		L	Am						
BAm	བཾ་	115	B	Am		B	Am						
TAm	ཏཾ་	115	T	Am		T	Am						
BHAm	བྷཾ་	73	BH	Am		BH	Am						
RAm	རཾ་	72	R	Am		R	Am						
WAm	ཝཾ་	44	W	Am		W	Am						
MAm	མཾ་	43	M	Am		M	Am						
NAm	ནཾ་	41	N	Am		N	Am						
RNAmS	རྣཾས་	40	RN	Am	S	RN	Am			S			
KAm	ཀཾ་	37	K	Am		K	Am						
HAm	ཧཾ་	34	H	Am		H	Am						
YAm	ཡཾ་	27	Y	Am		Y	Am						
BSAm	བསཾ་	25	BS	Am		B		S	Am				
DAm	དཾ་	25	D	Am		D	Am						
THAmS	ཐཾས་	24	TH	Am	S	TH	Am			S			
DZAm	ཛཾ་	23	DZ	Am		DZ	Am						
KHAm	ཁཾ་	22	KH	Am		KH	Am						
AAm	ཨཾ་	20	A	Am		A	Am						

Figure 21. As we can see in the attached table, there are vowels that occur fairly infrequently. These vowels are:

OO	2,500
Am	1,712
Om	1,686
Um	1,314
EE	1,021
i	349
Im	41
Em	39
Total	8,662

and they occur so infrequently that they are better classed as exceptions. By doing this, we reduce the amount of memory used by the expert system and speed the symbol matching process. In essence, we are just leaving them out of the vowel planes because most of their P1 X P2 combinations have “nils” or zeros in their cells.

This will allow us in our next version of the expert system to provide real probablities and other knowledge to standard C language programs or many other forms of CLIPS interfaces to Linux and MS Windows based machines.

Hence, a reasonable speed system can be provided that can spell-check syllables, identify whether a presented character patten is a syllable, and to perform the logic needed to sweep down a line of recognized characters and apply 1, 2, 3, and 4 glyph windows and determine the most likely syllable breaks.

ascii	tibetan	freq	P1	V	P2	C1	V1	C2	V2	C3	V3	C4	V4
BEE	བཻ	479	B	EE		B	EE						
MEE	མཻ	116	M	EE		M	EE						
KYEE	གྱཻ	94	KY	EE		KY	EE						
TEE	ཏཻ་	87	T	EE		T	EE						
TZEE	ཙཻ་	46	TZ	EE		TZ	EE						
NEE	ནཻ་	45	N	EE		N	EE						
BHEE	བྷཻ	32	BH	EE		BH	EE						
DZEE	ཛཻ་	32	DZ	EE		DZ	EE						
KEE	གཻ་	21	K	EE		K	EE						
MKHYEEN	མཁྱཻན་	17	MKHY	EE	N	M		KHY	EE	N			
DEE	དཻ་	8	D	EE		D	EE						
TREE	ཏྲཻ་	8	TR	EE		TR	EE						
YEE	ཡཻ	7	Y	EE		Y	EE						
AEE	ཨཻ	6	A	EE		A	EE						
YEES	ཡཻས་	5	Y	EE	S	Y	EE			S			
PEE	པཻ་	5	P	EE		P	EE						
BHEER	བྷཻར་	4	BH	EE	R	BH	EE			R			
BEEN	བཻན་	3	B	EE	N	B	EE			N			
REE	རཻ་	3	R	EE		R	EE						
B'EE	བཱཻ	3	B'	EE		B'	EE						
	total	1021											

Figure 22 This is the tally of all syllables with the vowel EE that were recognized and decomposed into individual glyphs.

This figure and the others immediately surrounding it show all the rare vowel syllables that we plan to code as exceptions instead of having the expert system match on them. By removing these vowel planes, and coding them as exceptions, we create a system that is simpler, that has the capacity to learn, and which will provide its knowledge at a faster rate.

Figures 21 thru 25 show the rare vowels (OO, Am, Om, Um, EE, i, Im, and Em). These vowels occur so rarely that matching on these is better done in the exception list. This saves the expert system from creating 8 extra P1 X P2 vowel planes which would mostly be filled with nils or zeros. To do otherwise creates an inefficient expert system model because it causes a tripling of the memory required for the system and creates a concomittant slow down in expert system speed.

The Rare Vowels:

OO	*2,500*
Am	*1,712*
Om	*1,686*
Um	*1,314*
EE	*1,021*
i	*349*
Im	*41*
Em	*39*
Total	*8,662*

ascii	tibetan	freq	P1	V	P2	C1	V1	C2	V2	C3	V3	C4	V4
SEmS	སེཾས་	39	S	Em	S	S	Em			S			

ascii	tibetan	freq	P1	V	P2	C1	V1	C2	V2	C3	V3	C4	V4
LIm	ལིཾ	10	L	Im		L	Im						
KshIm	ཀྵིཾ	10	Ksh	Im		Ksh	Im						
SIm	སིཾ	7	S	Im		S	Im						
BIm	བིཾ	5	B	Im		B	Im						
RIm	རིཾ	3	R	Im		R	Im						
DZRIm	ཛྲིཾ	3	DZR	Im		DZR	Im						
DZIm	ཛིཾ	3	DZ	Im		DZ	Im						
	total	41											

ascii		freq	P1	V	P2	C1	V1	C2	V2	C3	V3	C4	V4
AOm	ཨོཾ	1,676	A	Om		A	Om						
BCOm	བཅོཾ	5	BC	Om		B		C	Om				
SGOm	སྒོཾ	5	SG	Om		SG	Om						
	total	1,686											

Figure 23. More rare vowels.

The Rangjung Yeshe Dictionary

While we continue to parade lists of rare vowels before your eyes, let's look at a different approach to knowledge about Tibetan syllables — the lexicographical dimension. Dictionaries seek to be fairly complete wordlists and provide definitions to allow non-natives to understand the meaning of sequences of syllables. One would expect a dicitonary to express the breadth or the range of possible syllables in Tibetan. Rare words that might not find their way into 110 MB of Tibetan texts would certainly find their way into a good Dictionary, as by all accounts the Rangjung Yeshe Dictionary is.

One of our team members, Leonardo Gribaudo, of Florence, Italy, provided a frequency count of the syllables in the Rangjung Yeshe Dicitonary and, following conversion from Wylie to ACIP notation, we applied the expert system to analyze the data in the same way that we looked at the 110 MB of classic texts in the ACIP database.

Testing the expert system on the Rangjung Yeshe Dictionary tells us nothing of the relative frequency of the syllables in natural Tibetan wrting, as does the ACIP database, but it does test the scope

ascii	tibetan	freq	P1	V	P2	C1	V1	C2	V2	C3	V3	C4	V4
KOO	ཀཽ	1,120	K	OO		K	OO						
GOO	གཽ	582	G	OO		G	OO						
SHLOO	ཤླཽ	401	SHL	OO		SHL	OO						
MOOD	མཽད	302	M	OO	D	M	OO			D			
AOO	ཨཽ	46	A	OO		A	OO						
SOO	སཽ	19	S	OO		S	OO						
MOO	མཽ	11	M	OO		M	OO						
KOO'	ཀཽའ	4	K	OO	'	K	OO			'			
KOO'U	ཀཽའུ	4	K	OO	'U	K	OO			'	U		
SOOR	སཽར	4	S	OO	R	S	OO			R			
MOO'U	མཽའུ	4	M	OO	'U	M	OO			'	U		
GLOO	གླཽ	3	GL	OO		GL	OO						
		2,500											

Figure 24. More rare vowels.

The Rangjung Yeshe Dictionary (continued)

of our model of Tibetan script and ensures that we are capturing the greater variety inherent in the Tibetan syllables in an effectively larger body of work that 110 MB. The fitness of our choices of P1, V, and P2 are in question and a good dictionary can test our inclusiveness.

Figures 26 and 27 on pages 27-28: But forgive me for nestling some of the prettier pictures toward the end. In the previous vowel plane graphs, we only showed the top 130 most common P1s plotted against the 25 most common P2s. We did that so you could see some of the patterns in the top 130 most common P1s, where most of the action is occurring. But here we decided to show you the whole range of P1s. If nothing else, to get you a bull's eye view of the patterns. Just so you can see that, except for the big size differential in the two samples, the Rangjung Yeshe Dictionary and 110 MB sample of ACIP texts a reasonably similar pattern. The latter is fairly remarkable given that a dictionary is a fairly complete wordlist and that our large sample of ACIP texts is not. They are surprisingly similar when seen with the eyes of an eagle.

Of course, there are some differences. The areas near the bottom left hand

ascii		freq	P1	V	P2	C1	V1	C2	V2	C3	V3	C4	V4
HUm	ཧཱུྃ	1,160	H	Um		H	Um						
GSUm	གསཱུྃ	89	GS	Um		G		S	Um				
HUm	ཧཱུཾ	23	H	Um		H	Um						
SUm	སཱུཾ	18	S	Um		S	Um						
BUm	བཱུཾ	8	B	Um		B	Um						
YUm	ཡཱུཾ	5	Y	Um		Y	Um						
BHUm	བྷཱུྃ	5	BH	Um		BH	Um						
KUm	ཀཱུཾ	3	K	Um		K	Um						
GUm	གཱུཾ	3	G	Um		G	Um						
		1,314											

Figure 25. More rare vowels from the ACIP version 4 text database.

corner of the Rangjung Yeshe distribution in Figure 26 show some blank bands in some Initial Forms (P1s). These include, CV, BDZ, BHY, RY, d', DZ', Ksh, d, and BCV, amongst others. The Rangjung Yeshe had only about a dozen rare vowel entries (spread among Am, EE, and OO) and only about 50 exception syllables. Overall, the much larger ACIP database had more variability than the Rangjung Yeshe Dictionary and included forms that weren't found in the Rangjung Yeshe Dicitionary. Given the size of the ACIP database and the fact that a greater proportion of ACIP syllables are probably mantras, rare names, Sanskritized forms, and other things which may not have a strictly lexical meaning, this is not all that surprising. Such forms would not readily find their way into a dictionary.

It is also ture that forms were found in the Rangjung Yeshe Dictionary that were not in the ACIP database. But again, this is to be expected in that not every available word or syllable would likely be used in sample of text no matter how large.

http://www.liv.ac.uk/~ms2928/wordsmith/index.htm
Support for WordSmith Tools at author's site.

Guide to our code and data

The source code and data used to do the analysis are found at: http://www.ghg.net/dstilwell/newdata.zip (~3-4 MB) or at: ftp://storm.ptc.spbu.ru/pub/human-languages/tibetan/tibocr/newdata.zip.

We began our analysis of the ACIP version 4 data by pre-processing it with the UNIX tools 'sed,' 'diff,' a variety of custom shell scripts, and a good text editor in X-Windows. Occasionally, it was necessary to use a spreadsheet to help in writing the shell scripts. But by a combination of the above tools one is able prepare the ACIP data for processing by the WordSmith Tools (obtainable from Oxford University Press — http://www1.oup.co.uk/elt/catalogu/multimed/4589846/4589846.html) and Excel in the Microsoft Windows 98 environment. The need for pre-processing doesn't come from the change in UNIX to MS Windows environments, but rather from the reliance on common ASCII punctuation symbols by the ACIP notation in order to transcribe the richer set of Tibetan glyphs. So, for example, a comma might need to be separated from other ASCII characters by a space on both sides and then protected from being ignored by WordSmith by being converted into a tag: zcommaz, or a period into zperiodz, or the colon into zcolonz. WordSmith also wants to treat upper-case and lower-case letters the same — if not in the input, then at least in the output. So they too must be protected as zaz, zbz, zcz, etc. And numbers such as 1 are protected as zonez and 2 as ztwoz. Then English comments and other extraneous notes must be removed. ACIP page number notations were removed. Through out all of the preprocessing, the UNIX diff command was used along with a text editor to view the changes made and to decide if the intention was always carried out properly. To do this safely required human scanning of all of the results rather than sampling in some of the more complex transformations wreaked by ‘sed’ or “replace” in a text editor.

Following the pre-processing, WordSmith Tools were used to count all of the tokens (syllables and tagged data). These counts were then imported directly into Excel. This part was easy compared to the much more time consuming and difficult task of trying to prepare the raw ACIP data for processing.

Once in Excel, the creation of format for input of the data into CLIPS is also fairly straight forward. The result is the very long table of “deffacts” (or CLIPS input facts or input data) in ACIPv4.clp, the source code for the ACIP data analysis:

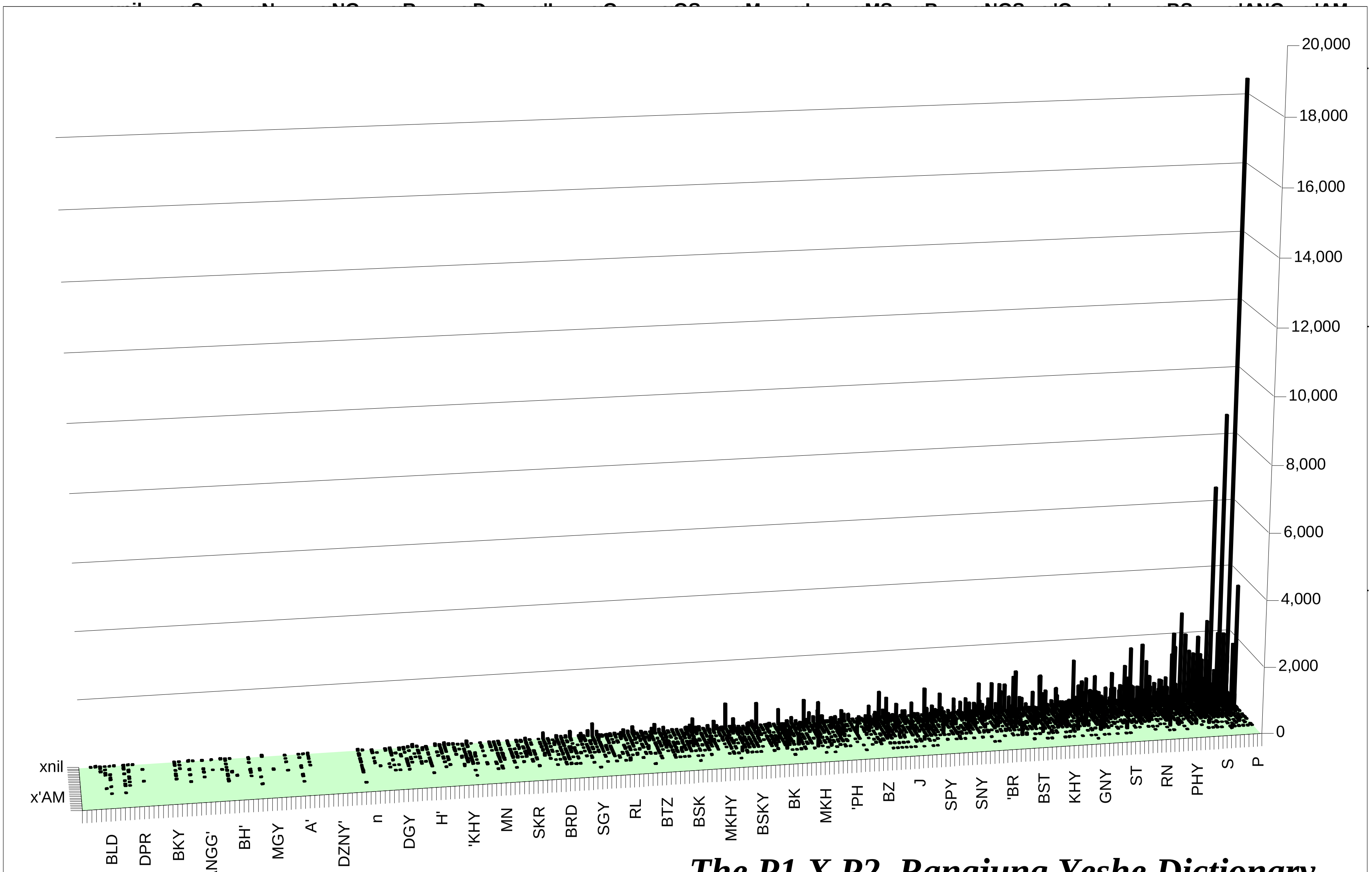
```
(deffacts acipdata "The 110 MB of ACIP
texts"

(acip (ascii PA )(freq 1269992))
(acip (ascii PA'I )(freq 591208))
(acip (ascii DANG )(freq 572456))
(acip (ascii LA )(freq 543953))
(acip (ascii BA )(freq 509249))
(acip (ascii PAR )(freq 479586))
(acip (ascii DE )(freq 451630))
(acip (ascii MA )(freq 344099))
(acip (ascii NI )(freq 341693))
(acip (ascii YIN )(freq 333446))
(acip (ascii NA )(freq 317321))
(acip (ascii DU )(freq 281775))
(acip (ascii KYI )(freq 234872))
(acip (ascii PHYIR )(freq 231444))
```

and so on.

Finally, the CLIPS analysis program is run. The CLIPS program, ACIPv4.clp in newdata.zip, breaks down each of the Tibetan syllables into parts P1, V, and P2 by matching. Why do we use the CLIPS expert system shell to do this analysis? Simple. Expert systems or production systems work on matching symbolic patterns and they do it very well. To determine what patterns there are in Tibetan and to count the number of times that the patterns occur in Tibetan, a tool that can make symbolic matches and then operate on them, such as counting and separating or sorting, is really what you need. We guess someone who knew a lot more about expert systems than we did could have written such a program in PERL or C or C++, but we can’t imagine being able to do so. The ability to match symbols and count them and to separate matches from non-matches is precisely what you need to discover what the actual combinations are. Hence, as those who are accustomed to “data mining” will readily know, our CLIPS analysis program is a very good one.

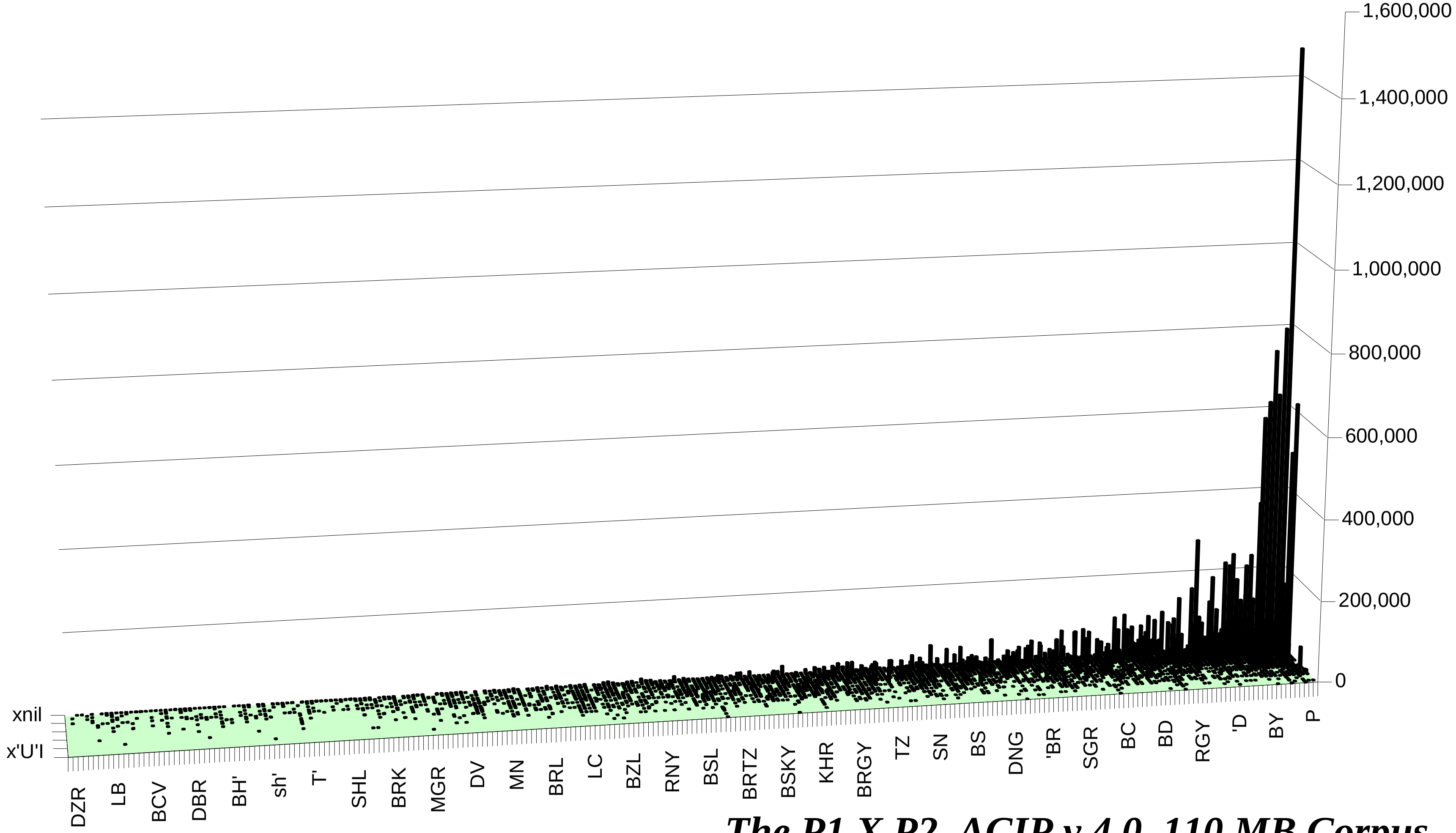
We also stored information about the breakdown of the syllables into base and vowel glyphs so that we could use it later in analysing the combinations on a glyph by glyph basis. The analysis is different. It will use a window of up to 4 base glyphs and one (or occasionally 2) vowel stacks. We will then use this method of analysing the syllables to achieve rapid lookups of the probabilistic data as the window slides along text. We think we can use FuzzyCLIPS to move across the syllables in the whole ACIP version 4 database when it is broken down into base and vowel glyphs. We think we can easily determine the separations between syllables for each text and we will compare it against the original. We will also determine how often alternate syllable divisions present themselves and how often they affect our ability to make the divisions between syllables. We believe that it will be possible to name a relatively small number of potential syllable divisions and alternates for each text. This can highlight “problem” areas for human inspection. But unfortunately, we’ll have to wait until later to prove that, but those who know Tibetan well can imagine the process themselves and slide along a sequence of glyphs in a book. How many places could you have put the tseg’s (the dots between syllables)?



The P1 X P2 Rangjung Yeshe Dictionary

<u>Figure 26</u>   This is the distribution, collapsed accross all vowels, sut shown the stunning length of P1 = PA =1 to P1= CVA =259.

In previous vowel planes, we only showed the top 130 P1s and 25 P2s. In this one we show the whole range from P1=PA=1 to P1=CVA, 259

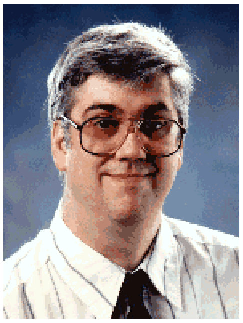


The P1 X P2 ACIP v 4.0, 110 MB Corpus Collapsed across all vowels and shown full length.

<u>Figure 27</u>   This is the distribution, collapsed accross all vowels, sut shown the stunning length of P1 = PA =1 to P1= CVA =259. See figure 15 on page 15 for a numerical representation of the upper right hand corner.

In previous vowel planes, we only showed the top 130 P1s and 25 P2s. In this one we show the whole range from P1=PA=1 to P1=CVA, 259

Don Stilwell
http://www.ghg.net/dstilwell/
http://www.serve.com/dstilwell/index.html and index1.html



I'd like to dedicate any work that I have done to the Tibetan people, who have suffered far more than I ever have. But I'd also especially like to dedicate it to my Tibetan teachers, Geshe Tenzin Wangyal Rinpoche (http://www.ligmincha.org/), His Holiness Lungtok Tenpa'i Nyima Ripoche, Alejandro Chaoul-Reich, and Geshe Chongtul Rinpoche (http://www.ccatech.com/bon/). Additionally to my Grandparents, Charles and Mabel Hassel who were the kindest people I ever met and an example I could never hope to fully emulate. But who could forget to mention our debt to the ancient Tibetan grammarians who would have loved to be able to do this kind of mathematical analysis, but who did not have the electronic computers and the vast Tibetan textbases that make it possible. And we should not forget our debt to the monks of Sera Mey and other Gelug-pa Monasteries who spent so many thousands of hours typing this data into computers. And to the creators of the Rangjung Yeshe Dictionary of the Tibetan Language (http://home.earthlink.net/~rangjung/rypub/ry-dic.htm) and to all their teachers and all those they hold dear. May all attain happiness.

Of course, we can thank everyone best by making all of this data and code available under the GNU General Public License so that everyone can have this information in usable formats, including all programs, data, files, reports, etc. (see: http://www.gnu.org/copyleft/lgpl.html)

How to Sponsor a Tibetan Bon Monk:
http://www.yungdrung-bon.org/english/SPONSORSHIP.HTML

**You will find the complete data set and source code for the work in this paper at:
http://www.ghg.net/dstilwell/newdata.zip (Approx 4 MB)**

**This paper is available, in its latest form, at:
http://www.ghg.net/dstilwell/paper2.PDF (Approx 1.5 MB)**

**You can see the hopelessly outdated websites at:
http://www.serve.com/dstilwell/index1.html and index.html**

**You can see our archive of Tibetan Calligraphy, information, etc., at:
ftp://storm.ptc.spbu.ru/pub/human-languages/tibetan/tibocr/**

**Learn about everything in Tibetan Computing at Chris Fynn's site:
http://www.users.dircon.co.uk/~cfynn/**

**See Rangjung Yeshe Dictionary site at:
http://home.earthlink.net/~rangjung/rypub/ry-dic.htm**

**See Marvin Moser's *Tibetan For Windows* and Pierre Robillard's *Tibetan on the Macintosh* site at:
http://members.aol.com/tib4win/ and
http://www.interlog.com/%7Epierrer/**

**See Leonardo Gribaudo's BOD_X software at:
http://space.tin.it/scienza/vfassio/index.htm**

**Asian Classics Input Project
http://www.asianclassics.org/**