Shahmukhi Corpus Report

I. Character Analysis:

The character analysis has been performed on the Shahmukhi corpus at sentence level and below is the list of all characters in the corpus with percentage of their occurrence.

Shahmukhi Char	Unicode Decimal	Percentage	SN
ا	1575	13.21	1
ی	1740	10.16	2
و	1608	7.59	3
ن	1606	6.53	4
ے	1746	5.53	5
ہ	1729	4.85	6
ر	1585	4.75	7
د	1583	4.55	8
ں	1722	4.42	9
ک	1705	4.12	10
ل	1604	3.32	11
ت	1578	3.24	12
ھ	1726	3.04	13
س	1587	2.97	14
م	1605	2.44	15
ب	1576	1.99	16
پ	1662	1.98	17
ج	1580	1.91	18
گ	1711	1.59	19
ِ	1616	1.53	20
چ	1670	1.46	21
ُ	1615	1.4	22
ء	1569	1.1	23
ڑ	1681	0.78	24
ٹ	1657	0.71	25
آ	1570	0.7	26
ڈ	1672	0.66	27
ش	1588	0.53	28
ف	1601	0.35	29
ح	1581	0.34	30
ع	1593	0.34	31
ق	1602	0.33	32
ز	1586	0.31	33
خ	1582	0.3	34

ص	1589	0.22	35
◌َ	1614	0.17	36
ط	1591	0.13	37
غ	1594	0.09	38
ض	1590	0.08	39
◌ّ	1617	0.08	40
ظ	1592	0.07	41
ذ	1584	0.05	42
ث	1579	0.05	43
◌ِ	1648	0.01	44
الله	65010	0.01	45
ـَ	1556	0.01	46
ـْ	1554	0.01	47
◌ً	1611	0	48
ء	1552	0	49
ئ	1574	0	50
ؤ	1654	0	51
ژ	1688	0	52
ـٴ	1553	0	53
ـٵ	1555	0	54
ة	1731	0	55
ي	1610	0	56
ـٔ	1550	0	57
ؤ	1572	0	58
ـئـ	1537	0	59
ٔ	1620	0	60
ـَ	1619	0	61
ٕ	1622	0	62
◌ٍ	1613	0	63
◌ّ	1612	0	64
◌ْ	1618	0	65
	65279	0	66

II. Unigram or Word Analysis:

The word analysis has been performed on the Shahmukhi corpus at sentence level and below is the list of top 100 words.

Shahmukhi Word	Percentage	SN
تے	2.54	1
اے	2.05	2
دے	1.75	3
دی	1.61	4
دا	1.41	5
نوں	1.29	6
کے	1.26	7
سی	1.13	8
وچ	0.95	9
نہیں	0.94	10
وی	0.94	11
اوہ	0.81	12
نال	0.79	13
ای	0.72	14
میں	0.7	15
ایہہ	0.66	16
توں	0.66	17
نیں	0.65	18
تاں	0.58	19
نے	0.57	20
نہ	0.54	21
وچ	0.52	22
اوس	0.47	23
پر	0.47	24
ہے	0.45	25
اِک	0.44	26
کر	0.42	27
کوءی	0.42	28
کہ	0.41	29
ایس	0.41	30
ہو	0.4	31

32	0.38	اوہناں
33	0.34	گل
34	0.34	سن
35	0.34	گیا
36	0.33	لءی
37	0.24	آپنے
38	0.24	گءی
39	0.24	مینوں
40	0.23	اُتے
41	0.22	پءی
42	0.22	اوہدے
43	0.21	جے
44	0.2	میرے
45	0.2	جو
46	0.2	گھر
47	0.2	کیہ
48	0.19	اوہنوں
49	0.19	ہووے
50	0.19	آ
51	0.19	ہویا
52	0.18	ہور
53	0.18	اک
54	0.18	جی
55	0.17	گءے
56	0.17	کُجھ
57	0.17	جاندا
58	0.17	جا
59	0.17	پھیر
60	0.16	دِیاں
61	0.16	آپ
62	0.16	بے
63	0.16	لگ
64	0.16	ہوندا
65	0.16	آپنی
66	0.15	ہر

67	0.15	کم
68	0.15	ہن
69	0.15	ہُن
70	0.15	آکھیا
71	0.15	آں
72	0.15	ہوءی
73	0.15	جیویں
74	0.15	پیا
75	0.14	ول
76	0.14	ایہناں
77	0.14	کرن
78	0.14	کِسے
79	0.14	دیاں
80	0.14	جدوں
81	0.14	والے
82	0.14	بولی
83	0.14	رہیا
84	0.14	اسیں
85	0.13	کیتا
86	0.13	اندر
87	0.13	ویلے
88	0.13	دو
89	0.13	لے
90	0.13	ہتہ
91	0.12	ہوءے
92	0.12	ماں
93	0.12	کول
94	0.12	اوہنے
95	0.12	بندے
96	0.11	پتہ
97	0.11	اج
98	0.11	سارے
99	0.11	اوہدی
100	0.11	میری

III. Bi-Gram Analysis:

The bi-gram analysis has been performed on the Shahmukhi corpus at sentence level and below is the list of top 100 bi-gram words.

Previous word	word	Frequency	SN
	میں	8473	1
	اوہ	7923	2
	ایہہ	7083	3
	پر	5917	4
	اوس	4744	5
	تے	4526	6
	ایس	4187	7
اے	تے	4181	8
نہیں	سی	3804	9
کر	کے	3759	10
جاندا	اے	**2941**	**11**
	اوہناں	2598	12
	اِک	2563	13
سی	تے	2512	14
ہوندا	اے	2446	15
اے	کہ	2416	16
	جے	2302	17
	مینوں	2051	18
اوہناں	نوں	2033	19
اوہناں	دے	1843	20
ہو	گیا	1842	21
	ہُن	1838	22
تے	اوہ	1793	23
	پھیر	1754	24
ہو	کے	1654	25
سی	کہ	1587	26
	کوءی	1569	27
	اوہنے	1553	28
	جدوں	1547	29
ای	نہیں	1544	30
	اسیں	1517	31

32	1511	اے	سکدا
33	1487	اے	ہوندی
34	1472	دی	اوہناں
35	1434	ایہناں	
36	1432	نوں	اوس
37	1406	اے	جاندی
38	1393	میرے	
39	1390	گل	دی
40	1375	تے	نیں
41	1355	نہیں	وی
42	1340	توں	
43	1329	اے	گیا
44	1299	کے	لے
45	1298	نال	دے
46	1203	اج	
47	1195	گےء	ہو
48	1188	اوہدے	
49	1176	کیہ	
50	1172	دے	اوس
51	1151	جیویں	
52	1147	پر	اے
53	1146	ہر	
54	1142	نہیں	
55	1113	نہ	
56	1113	سی	گیا
57	1097	کے	جا
58	1054	سی	رہیا
59	1048	اے	ہویا
60	1042	دا	اوہناں
61	1026	سی	رہی
62	1025	کے	ویکھ
63	1025	میں	تے
64	1005	اوس	تے
65	1002	نہیں	پتہ
66	1002	اوہناں	تے

67	993	اوہنوں	
68	982	وی	نوں
69	972	نیں	جاندے
70	958	پےء ی	اے
71	957	میرا	
72	953	کے	آ
73	952	پھیر	تے
74	951	جو	اے
75	944	پیا	لگ
76	943	بولی	ماں
77	937	اک	
78	936	ای	نال
79	926	نہیں	کوءی
80	919	نے	اوس
81	916	آپ	آپنے
82	911	پر	سی
83	900	اوہ	تاں
84	886	بے	
85	882	اے	گےء ی
86	880	دی	اوس
87	872	سی	گےء ی
88	863	ساڈے	
89	857	تے	سن
90	855	میری	
91	845	سی	ہویا
92	842	اوہ	کہ
93	841	گےء ے	ہو
94	838	تے	گیا
95	838	گل	ایہہ
96	837	سکدا	ہو
97	828	لےء ی	ایس
98	818	اے	گل
99	814	ایتھے	
100	802	اے	کردا

IV. Tri-Gram Analysis:

The tri-gram analysis has been performed on the Shahmukhi corpus at sentence level and below is the list of top 100 tri-gram words.

Previous to Previous Word	Previous Word	word	Frequency	SN
		میں	8473	1
		اوہ	7923	2
		ایہہ	7083	3
		پر	5917	4
		اوس	4744	5
		تے	4526	6
		ایس	4187	7
		اوہناں	2598	8
		اِک	2563	9
		جے	2302	10
		مینوں	2051	11
		ہُن	1838	12
		پھیر	1754	13
		کوءی	1569	14
		اوہنے	1553	15
		جدوں	1547	16
		اسیں	1517	17
		ایہناں	1434	18
		میرے	1393	19
		توں	1340	20
		اج	1203	21
		اوہدے	1188	22
		کیہ	1176	23
		جیویں	1151	24
		ہر	1146	25
		نہیں	1142	26
		نہ	1113	27
		اوہنوں	993	28
		میرا	957	29
		اک	937	30
		بے	886	31

32	863	ساڈے		
33	855	میری		
34	814	ایتھے		
35	802	ایہدے		
36	781	ٹجھ		
37	764	ہن		
38	752	کیوں		
39	708	سو		
40	705	اوہدی		
41	687	گل		
42	681	تُسیں		
43	669	جد		
44	669	بس		
45	657	ماں		
46	648	تاں		
47	646	آپنے		
48	644	ایسے		
49	642	جو		
50	625	ایہو		
51	617	گھر		
52	606	اس		
53	597	جیہڑا		
54	576	فیر		
55	576	اوہدا		
56	572	ہور		
57	562	کِسے		
58	557	تیرے		
59	556	ہاں		
60	549	سانوں		
61	548	اے	سکدا	ہو
62	544	پنجابی		
63	539	سِمرن		
64	534	اساں		
65	529	دو		
66	522	مُڑ		

67	511	وچ		
68	501	اِنج		
69	500	سارے		
70	497	لوک		
71	495	آکھیا	اوس	
72	489	اوتھے		
73	485	سبہ		
74	481	آپ		
75	478	علی		
76	474	ٹوں		
77	466	ڈاکٹر		
78	465	نوں	آپ	آپنے
79	462	تینوں		
80	**446**	اے	جاندا	ہو
81	435	جس		
82	422	کے	کر	ایس
83	414	پاکستان		
84	413	دے		
85	411	جس		
86	410	نالے		
87	401	جیہڑے		
88	401	آپنی		
89	398	رب		
90	394	پنجاب		
91	393	لءی	ایس	
92	389	اتے		
93	387	پتہ		
94	385	کءی		
95	382	کجھ		
96	375	نال		
97	375	بندے		
98	371	وچ		
99	370	دے	اوہناں	
100	366	شاہ		

PIN- 147 002
Phone: +91-0175-3046171, 3046172
Fax: +91-0175-3046313
email: gslehal@yahoo.com