3. SCRIPTS OF PUNJABI LANGUAGE: COMPARATIVE STUDY

3.1 Introduction

Today's world culture is an increasingly global culture, surmounting regional and country frontiers and facilitated by the abundant contributions by the Internet technology. There is now a vast amount of information accessible via the Internet where a lot of regional and cultural information is put onto the World Wide Web in different languages and scripts. Notably, there are more than six thousand living languages in the world and adding to the diversity is the fact that some languages are written in different scripts in different regions of the world. The multitude of foreign languages and mutually incomprehensible scripts of the same language pose a barrier to information as we cannot all learn every language or script being used. Therefore, if we can get around the language barrier or at least the script barrier, we would be able to access much more of the world's culture and can thus explore its plentiful richness.

Asia has 60.8\% of the world population and an impressive count of 2,322 living languages contributing (33.6\%) towards the world languages. South Asia is one of the richest regions in terms of linguistic diversity and its languages have a lot in common. For example, most of the major Indian languages use scripts that are derived from the ancient Brahmi script, have more or less the same arrangement of alphabet, are highly phonetic in nature and are very well organized. India has 22 scheduled languages\(^8\) written in ten different scripts. Indian languages can be broadly classified into five groups according to their origin and similarity. These are Indo-Aryan family (Hindi, Bangla, Assami, Punjabi, Marathi, Oriya and Gujarati); Dravidian family (Tamil, Telugu, Kannada and Malayalam); Austro-Asian family; Tibetan-Burmese family; and Andamanese.

The present research work focuses on Punjabi language. Punjabi is considered to be an ancient language. The name \textit{"Punjabi"} or \textit{"Panjabi"} comes from the region it is spoken in \textit{"the Punjab"}. The word Punjab is an amalgamation of two Persian words, \textit{"Panj"} meaning five and \textit{"Aab"} meaning water. Hence, literally, ‘Punjab’ means the land of five rivers. The Punjabi language is from the Indo-Aryan group of languages, which is the sub-group of Indo-Iranian and Indo-European group of languages.

\(^8\) http://en.wikipedia.org/wiki/Official_languages_of_India
Punjabi is spoken in both Eastern and Western Punjab, Jammu and Kashmir, and the mountainous areas of Pakistan and India. Apart from this, Punjabi is also spoken by immigrants who have migrated to USA, Canada, U.K., Australia, Singapore and other parts of the world. Hence, almost a 100 million people worldwide speak different dialects of this language as their first language. It is a fact that Punjabi also belongs to those few languages of the world that are written in two or more mutually incomprehensible scripts [134] as shown in Table 6. To record Punjabi language, Punjabi speakers use two different scripts, Perso-Arabic (referred as Shahmukhi) and Gurmukhi. Shahmukhi is used by the people of Western Punjab in Pakistan, whereas Gurmukhi is used by most people of Eastern Punjab.

<table>
<thead>
<tr>
<th>Language</th>
<th>Script</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUNJABI</td>
<td>Gurmukhi</td>
<td>India</td>
</tr>
<tr>
<td></td>
<td>Shahmukhi</td>
<td>Pakistan</td>
</tr>
<tr>
<td>KAZAKH</td>
<td>Arabic script</td>
<td>Kazakhstan, Afghanistan, China, Germany, Iran, Mongolia, Russia (Asia),</td>
</tr>
<tr>
<td></td>
<td>Cyrillic script</td>
<td>Tajikistan, Turkey (Asia), Turkmenistan, Ukraine, Uzbekistan.</td>
</tr>
<tr>
<td></td>
<td>Latin script</td>
<td></td>
</tr>
<tr>
<td>KURDISH,</td>
<td>Arabic script</td>
<td>Turkey (Asia), Afghanistan, Armenia, Azerbaijan, Georgia, Iran, Iraq,</td>
</tr>
<tr>
<td>Northern</td>
<td>Cyrillic script</td>
<td>Lebanon, Syria</td>
</tr>
<tr>
<td></td>
<td>Latin script</td>
<td></td>
</tr>
<tr>
<td>MALAY</td>
<td>Arabic script</td>
<td>Malaysia (Peninsular)</td>
</tr>
<tr>
<td></td>
<td>Latin script</td>
<td>Indonesia (Sumatra) and Singapore</td>
</tr>
<tr>
<td>MALAY,</td>
<td>Arabic script</td>
<td>Thailand</td>
</tr>
<tr>
<td>Pattani</td>
<td>Thai script</td>
<td></td>
</tr>
</tbody>
</table>

3.2 Languages of the World

Lewis, [134] presents that there are 6,909 living languages that originated under five world areas: Africa, America, Asia, Europe, and the Pacific. A living language is defined as the one that has at least one speaker for whom it is their first language; extinct languages and languages that are used only as a second language are excluded.

---

9 Source: http://www.ethnologue.org
from these counts. Table 7 summarizes the distribution of languages and their populations by these areas. Both Asia and Africa together contributes 64.10% of the world’s living languages. The Mean value gives the average number of speakers per language, while the Median reflects the middle value in the distribution of language populations (i.e., half of the languages have more speakers than that number and half have that number or fewer).

<table>
<thead>
<tr>
<th>Area</th>
<th>Living languages Count</th>
<th>Percent</th>
<th>Number of speakers Count</th>
<th>Percent</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>2,110</td>
<td>30.5</td>
<td>344,291</td>
<td>12.2</td>
<td>25,200</td>
<td></td>
</tr>
<tr>
<td>Americas</td>
<td>993</td>
<td>14.4</td>
<td>50,852</td>
<td>0.8</td>
<td>2,300</td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>2,322</td>
<td>33.6</td>
<td>1,560,194</td>
<td>60.8</td>
<td>11,100</td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>234</td>
<td>3.4</td>
<td>6,638,295</td>
<td>26.1</td>
<td>201,500</td>
<td></td>
</tr>
<tr>
<td>Pacific</td>
<td>1,250</td>
<td>18.1</td>
<td>5,144</td>
<td>0.1</td>
<td>980</td>
<td></td>
</tr>
</tbody>
</table>

| Totals  | 6,909                  | 100.0   | 862,572                   | 100.0   | 7,560 |        |

A striking fact is that there is a huge disparity between the mean size of languages and their median size. It turns out that 389 (or nearly 6%) of the world’s languages have at least one million speakers and account for 94% of the world’s population. By contrast, the remaining 94% of languages are spoken by only 6% of the world’s people. Therefore, in global view of Information technology, it is conceivable that if we can bridge the inter-communication gap between the 389 (nearly 6%) world languages, 94% of world population will benefit immensely. [134]

3.3 Historical Script Families and Derived Scripts [135]

The number of scripts currently used to write thousands of world languages is less that 30 and to simplify things further, it is rarely realized that there are, in fact, only three basic types of scripts (Clews, 1997):

a) Ideographic script, mostly used in China, Japan and Korea.

b) Brahmi script, used in scripts of Devanagari, languages of South Asia.

c) Phoenician script, now used in scripts of Europe, the Middle East and North Africa. Greek, Hebrew, Ethiopian and Arabic are derived from Phonecian
script. Afterwards, Latin, Cyrillic, Armenian and Georgian scripts are further derived from Greek.

Figure 9 Historical Script Families and Derived Scripts

These three basic scripts (see Figure 9)\(^\text{10}\) have a major influence on the scripts we use today. For example, the input Shahmukhi script that is used to record Punjabi language is a local variant of Urdu script. Further tracking of the historical scripts’ families shows that the script of Urdu and Persian languages are derived from Arabic script and Arabic script itself has originated from the ancient Phoenician script. Gurmukhi script has originated from brahmi script. Therefore, Shahmukhi and Gurmukhi scripts of Punjabi language have different roots because Shahmukhi script is derived from Urdu script and Gurmukhi script has its origin from the ancient Brahmi script. Hence, we can say that the first hurdle that transliteration system needs to tackle is to handle these Punjabi scripts with different origins, different direction of writing, different set of alphabet, and different kinds of writing system conventions.

3.4 Past and Present of the Scripts of Punjabi Language

Like many other languages, Punjabi language starts with poetry. According to Nasir Rana [136], and Sant Singh Sekhon [137] Punjabi poetry is very old and some of its

\(^{10}\) Scripts shown in CAPS are the historical source script for other scripts shown. These scripts are used in over 99% of the world's official languages Sogdian (a fourth historical script) shares some aspects of neighboring scripts.
earliest poets were Charpat Nath (840 A.D.-940 A.D.), Gorakh Nath (940-1031), Pooran Bhagat (970-1070) and Shah Shams Sabzvari (1165-1276). But most Punjabi scholars accept Sheikh (or Baba) Farid (1173 A.D.-1266 A.D.) as the first poet of Punjabi. His poetry has been preserved in the form of Shaloks. The two major traditions of medieval Punjabi literature are Sufi and qissa poetry. What is called qissa in Punjabi is about the same as Masnavi of Persian.

According to Harbhajan Singh, the state of medieval Punjab was multicultural. In such a context there must have been a tendency to use more than one language at all levels. Persian was the court language. People used their native tongue in their homes, and for public communication they would have made use of a tongue resembling Lahori, Hindi or Urdu language. This multi-linguistic state of affairs was reinforced by the employment of more than one script. The Muslim Sufi poetry primarily comprising qissas poetry or versified tales, vars or heroic odes, and Jangnamas or war poetry employed Persian script. Therefore, we can conclude that more that 60% of the Punjabi literature of medieval period was written in Persian/Urdu based script now called Shahmukhi script. On the other hand, under the gurmat tradition set by Guru Nanak Dev, the entire writing was in Gurmukhi script which is standardised by the second Guru, Guru Angad Dev in the 16th century. Adi Guru Granth, the Sikh scripture, is the prime architecture of the gurmat tradition. It was compiled and edited by the fifth guru, Guru Arjan Dev. The Adi Guru Granth was compiled in Gurmukhi script, but its content was also received from non-Gurmukhi sources. Neither was the script used by Sheikh Farid Gurmukhi, nor even that employed by Kabir, Namdeva and Ravidas etc., whose poetry, after transliteration into Gurmukhi script, were included in this scripture. [136-138]

During the British rule, Christian missionaries used Punjabi language as a vehicle for promotion of their religious literature. Initially, William Carey wrote grammars for Bengali, Punjabi, Sanskrit, Marathi, Telugu and other languages of the country. He published the first Punjabi translation of Bible in 1811. During 1815-1826, he published five books of Moses from Bible, two history books and scriptures of the prophets in Punjabi [139]. In 1834, a missionary group set up its centre in Ludhiana and later on, they set up their branches in the cities of Jalandhar, Ambala, Hoshiarpur, Ferozpur and Rawalpindi [140]. After establishing the first Punjabi printing press with Gurmukhi font in 1835, the Christian mission at Ludhiana published their missionary literature in Gurmukhi as well as Shahmukhi or Persian.
style of writing. In 1873, the Hindu Parkash from Amritsar was the first newspaper using Gurmukhi style of writing. Later on, several Punjabi newspapers and magazines like Khalsa Akhbar, Singh Sabha Gazette, Amrit Patrika, Kavi, and Maojji etc. continued to be published. [136]

Post-Independence (since 1947) corpus of Punjabi language has been contributed to by prominent scholars and poets of both eastern (India) and western Punjab (Pakistan) in both Shahmukhi and Gurmukhi scripts. Punjabi in Gurmukhi script is the state language of east Punjab and it has got the status of 2nd language of adjoining Haryana state of India. Currently, there are many daily and periodical newspapers published in Gurmukhi Punjabi such as Ajit, Ajit Weekly, Ajj di Aawaz, Akali Patrika, Chardi Kala, Daily Aashiana, Desh Sewak, Desh Videsh Times, Doaba Headlines, Jag Bani, khuli soch, Nawan Zamana, Punjabi Tribune, Rozana Spokesman, Sanjh Savera, Sher-e-punjab, Quami Ekta etc. Similarly, in Pakistan various Shahmukhi newspapers like Sajjan, Ravel, Bhulekhā, Lōkāī, and monthly or periodical magazines such as Lahirān, Punjabi Zuban, Likhār, Punjabi Adab, Pakhērū, Sānjh, Pañcam etc. playing a crucial role in the development of Punjabi language in both Shahmukhi and Gurmukhi scripts. In particular, the Punjabi magazine Sānjh is published in two separate volumes (with identical content) from Lahore in Shahmukhi script and from Ludhiana in Gurmukhi script.

3.5 Gurmukhi Script

The Gurmukhi script, standardized by Guru Angad Dev in the 16th century, was designed to write the Punjabi language [137,138], [141]. It was modeled on the Landa alphabet. The literal meaning of "Gurmukhi" is from the mouth of the Guru. As shown in Table 8 the Gurmukhi script has syllabic alphabet in which all consonants have an inherent vowel. The Gurmukhi script has forty one letters, comprising thirty eight consonants and three basic vowel sign bearers (Matra Vahak). The first three letters are unique because they form the basis for vowels and are not consonants. The six consonants in the last row are created by placing a dot at the foot (pair) of the consonant (Naveen Toli). There are five nasal consonants (ਙ[ɲə], ਞ[ɲə], ਣ[ɳ], ਨ[n], ਮ[m]) and two additional nasalization signs, bindi ◌ਂ[ɲ] and tippi ◌ੰ[ɲ] in Gurmukhi script. In addition to this, there are nine dependent vowel
signs (or diacritics) (◌ੁ, ◌ੂ, ◌ੋ, ◌ਆ, ◌ਿ◌, ◌ੀ, ◌ੇ, ◌ੈ, ◌ੌ) used to create ten independent vowels (ਉ, ਊ, ਓ, ਅ, ਆ, ਇ, ਈ, ਏ, ਐ, ਔ) with three bearer characters: Ura ਇ, Aira ਅ and Iri ਍. With the exception of Aira ਅ independent vowels are never used without additional vowel signs. The diacritics which can appear above, below, before (not in Unicode standard) or after the consonant they belong to, are used to change the inherent vowel and when they appear at the beginning of a syllable, vowels are written as independent vowels. Some Punjabi words require consonants to be written in a conjunct form in which the second consonant is written under the first as a subscript. There are three commonly used subjoined consonants as shown here Haha ਹ (usage ਨ + ◌੍ + ਹ = ਨʰ), Rara ਰ (usage ਪ + ◌੍ + ਰ = ਪʰ) and Vava ਵ (usage ਸ + ◌੍ + ਵ = ਸ⁰).

<table>
<thead>
<tr>
<th>Gurmukhi Alphabet</th>
<th>Group(Toli) Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ਁ</td>
<td>Matra Vahak</td>
</tr>
<tr>
<td>਋</td>
<td>Mul Varag</td>
</tr>
<tr>
<td>ਕ</td>
<td>Kavarg Toli</td>
</tr>
<tr>
<td>ਘ</td>
<td>Chavarg Toli</td>
</tr>
<tr>
<td>ਢ</td>
<td>Tavarg Toli</td>
</tr>
<tr>
<td>ਣ</td>
<td>Tavarg Toli</td>
</tr>
<tr>
<td>ਤ</td>
<td>Pavarg Toli</td>
</tr>
<tr>
<td>ਢ</td>
<td>Antim Toli</td>
</tr>
<tr>
<td>ਣ</td>
<td>Naveen Toli</td>
</tr>
</tbody>
</table>

### 3.6 Shahmukhi Script

The meaning of "Shahmukhi" is literally "from the King's mouth". Shahmukhi is a local variant of the Urdu script used to record the Punjabi language. It is based on right to left Nastalique style of the Persian and Arabic script. As shown in Figure 10
the basic Shahmukhi alphabet has thirty nine letters, including five long vowel signs Alif \[\text{ا} \], Alif madda \[\text{ا̣} \], Vao \[\text{و} \], Choti-ye \[\text{ی} \] and Badi-ye \[\text{ے} \]. In total, Shahmukhi script has thirty six simple consonants including two long vowels Vao \[\text{و} \] and Choti-ye \[\text{ی} \] as a special case and eleven frequently used aspirated consonants \[133\].

<table>
<thead>
<tr>
<th>Sr.</th>
<th>Name</th>
<th>Char</th>
<th>Sr.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Noon Ghunna</td>
<td>[\text{n} ]</td>
<td>8</td>
<td>Sallallahu Alayhe Wassallam</td>
</tr>
<tr>
<td>2</td>
<td>Hamza</td>
<td>[\text{i} ]</td>
<td>9</td>
<td>Rahmatullah Alayhe</td>
</tr>
<tr>
<td>3</td>
<td>Shad</td>
<td>[\text{ّ} ]</td>
<td>10</td>
<td>Radi Allahou Anhu</td>
</tr>
<tr>
<td>4</td>
<td>Khari Zabar</td>
<td>[\text{ّ} ]</td>
<td>11</td>
<td>Alayhe Assallam</td>
</tr>
<tr>
<td>5</td>
<td>do Zabar</td>
<td>[\text{ّ} ]</td>
<td>12</td>
<td>Maddah</td>
</tr>
<tr>
<td>6</td>
<td>do Zer</td>
<td>[\text{ّ} ]</td>
<td>13</td>
<td>Khari Zer</td>
</tr>
<tr>
<td>7</td>
<td>Takhallus</td>
<td>[\text{ّ} ]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Another feature of Shahmukhi characters is that they change their shapes depending upon neighboring context. But generally they acquire one of these four shapes, namely isolated, initial, medial and final. Urdu characters can be divided into two groups, non-joiners and joiners. The non-joiners can acquire only isolated and final shape and do not join with the next character (see Table 10). On the contrary, joiners
can acquire all the four shapes and get merged with the next following character (see Table 11). A group of joiners and/or non-joiner joined together form a ligature. A word in Urdu is a collection of one or more ligatures. The isolated form of joiners and non-joiners is shown in Figure 11 & Figure 12.

Table 10 Non-Joiner Character with Consistent Shape

<table>
<thead>
<tr>
<th>Character Position</th>
<th>Shahmukhi Non-Joiner ṭ[v]</th>
<th>Highlighted</th>
<th>Gurmukhi Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Position</td>
<td>ṭ[v]</td>
<td></td>
<td>ḍhañ</td>
</tr>
<tr>
<td>Middle Position</td>
<td>ṭ[v]</td>
<td></td>
<td>ṭ[v]</td>
</tr>
<tr>
<td>Last Position</td>
<td>ṭ[v]</td>
<td></td>
<td>ṭ[v]</td>
</tr>
</tbody>
</table>

Figure 11 Non-Joiners in Shahmukhi Script

Table 11 Joiner Character with different Shapes in a Shahmukhi Word

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>First Position</td>
<td>जोड़े</td>
<td>याद</td>
<td>जोड़े</td>
<td>जा</td>
</tr>
<tr>
<td>Middle Position</td>
<td>भेरे ही की से मिले</td>
<td>भेरे ही</td>
<td>भेरे ही की से मिले</td>
<td>भेरे ही</td>
</tr>
<tr>
<td>Last Position</td>
<td>उठे जोड़े</td>
<td>उठे जोड़े</td>
<td>उठे जोड़े</td>
<td>उठे जोड़े</td>
</tr>
</tbody>
</table>
3.7 Script Mappings

3.7.1 Mapping of Simple Consonants

Unlike Gurmukhi script, Shahmukhi script has many character forms to represent similar sound. For example, the single Gurmukhi consonant ਜ਼[z] is mapped to ذ[z], ز[z], ژ[ʒ], ض[z], ظ[z], five similar sound consonants in Shahmukhi. Similarly, Gurmukhi ਤ[ṫ] and ਸ[s] have two similar sound consonants in Shahmukhi as shown in Table 12 below. In the case of non-aspirated consonants, Shahmukhi has two character forms mapped into single Gurmukhi consonant. We can also observe that Shahmukhi consonant Noon ﻦ[n] has multiple mappings to four different Gurmukhi characters ਨ[n], ਨ[ɲ], ਣ[ɳ] or ਨ[η]. Shahmukhi RNoon ﻧ[ɳ] is rarely used for Gurmukhi character ਣ[ɳ].

<table>
<thead>
<tr>
<th>Sr.</th>
<th>Shahmukhi</th>
<th>Gurmukhi</th>
<th>Sr.</th>
<th>Shahmukhi</th>
<th>Gurmukhi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ﺑ[b]</td>
<td>ﺑ[b]</td>
<td>19</td>
<td>ﺝ[z]</td>
<td>ﺝ[z]</td>
</tr>
<tr>
<td>3</td>
<td>ﺣ[t]</td>
<td>ﺣ[t]</td>
<td>21</td>
<td>ﺫ[y]</td>
<td>ﺫ[y]</td>
</tr>
<tr>
<td>4</td>
<td>ﺤ[s]</td>
<td>ﺥ[s]</td>
<td>22</td>
<td>ﺧ[f]</td>
<td>ﺧ[f]</td>
</tr>
<tr>
<td>5</td>
<td>ﺦ[ʂ]</td>
<td>ﺦ[ʂ]</td>
<td>23</td>
<td>ﺩ[q]</td>
<td>ﺩ[k]</td>
</tr>
<tr>
<td>6</td>
<td>ﺨ[f]</td>
<td>ﺨ[f]</td>
<td>24</td>
<td>ﻥ[k]</td>
<td>ﻥ[k]</td>
</tr>
<tr>
<td>7</td>
<td>ﺩ[h]</td>
<td>ﺩ[h]</td>
<td>25</td>
<td>ﻨ[g]</td>
<td>ﻨ[g]</td>
</tr>
<tr>
<td>8</td>
<td>ﺪ[x]</td>
<td>ﺪ[x]</td>
<td>26</td>
<td>ﺪ[l]</td>
<td>ﺪ[l]</td>
</tr>
<tr>
<td>9</td>
<td>ﺪ[d]</td>
<td>ﺪ[d]</td>
<td>27</td>
<td>ﺪ[m]</td>
<td>ﺪ[m]</td>
</tr>
<tr>
<td>10</td>
<td>ﺤ[z]</td>
<td>ﺤ[z]</td>
<td>28</td>
<td>ﺦ[n]</td>
<td>ﺦ[n], ﻪ[ɲ], ﺪ[ɳ]</td>
</tr>
<tr>
<td>12</td>
<td>ﺮ[z]</td>
<td>ﺮ[z]</td>
<td>30</td>
<td>ﻨ[v]</td>
<td>ﻨ[v]</td>
</tr>
<tr>
<td>13</td>
<td>ﺯ[z]</td>
<td>ﺯ[z]</td>
<td>31</td>
<td>ﻢ[h]</td>
<td>ﻢ[h]</td>
</tr>
<tr>
<td>14</td>
<td>ﺲ[s]</td>
<td>ﺲ[s]</td>
<td>32</td>
<td>ﻬ[j]</td>
<td>ﻬ[j]</td>
</tr>
<tr>
<td>15</td>
<td>ﺵ[f]</td>
<td>ﺵ[f]</td>
<td>33</td>
<td>ﻢ[h]</td>
<td>ﻢ[h]</td>
</tr>
<tr>
<td>16</td>
<td>ﺲ[s]</td>
<td>ﺲ[s]</td>
<td>34</td>
<td>ﺮ[t]</td>
<td>ﺮ[t]</td>
</tr>
<tr>
<td>17</td>
<td>ﺤ[z]</td>
<td>ﺤ[z]</td>
<td>35</td>
<td>ﺵ[d]</td>
<td>ﺵ[d]</td>
</tr>
<tr>
<td>18</td>
<td>ﺤ[t]</td>
<td>ﺤ[t]</td>
<td>36</td>
<td>ﺦ[t]</td>
<td>ﺦ[t]</td>
</tr>
</tbody>
</table>
3.7.2 Mapping of Aspirated Consonants (AC)

In Shahmukhi script, the aspirated consonants are represented by the combination of a simple consonant and *Heh-Daochashmee* ḥ[h]. Table 13 shows eleven frequently used aspirated consonants in Shahmukhi corresponding to which Gurmukhi script has unique single character except the ṭ[rʰ] which is a character combination of ṭ[ṛ] + ḍ[ɽ].

<table>
<thead>
<tr>
<th>Sr.</th>
<th>AC</th>
<th>Gurmukhi</th>
<th>Sr.</th>
<th>AC</th>
<th>Gurmukhi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ḥ[b]</td>
<td>ḡ[b]</td>
<td>7</td>
<td>ḥ[d]</td>
<td>ḡ[d]</td>
</tr>
<tr>
<td>2</td>
<td>ḥ[p]</td>
<td>ḡ[p]</td>
<td>8</td>
<td>ḥ[t]</td>
<td>ḡ[t]</td>
</tr>
<tr>
<td>3</td>
<td>ḥ[t]</td>
<td>ḡ[t]</td>
<td>9</td>
<td>ḥ[k]</td>
<td>ḡ[k]</td>
</tr>
<tr>
<td>4</td>
<td>ḥ[d]</td>
<td>ḡ[d]</td>
<td>10</td>
<td>ḥ[g]</td>
<td>ḡ[g]</td>
</tr>
<tr>
<td>5</td>
<td>ḥ[dʒ]</td>
<td>ḡ[dʒ]</td>
<td>11</td>
<td>ḥ[ɽ]</td>
<td>ḡ[ɽ]</td>
</tr>
<tr>
<td>6</td>
<td>ḥ[tʃ]</td>
<td>ḡ[tʃ]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.7.3 Mapping of Vowels

The long and short vowels of Shahmukhi script have multiple mappings into Gurmukhi script as shown in Table 14 and Table 15 respectively. It is interesting to observe that Shahmukhi long vowel characters Vav [v] and Ye [j] have vowel-vowel multiple mappings as well as one vowel-consonant mapping.

<table>
<thead>
<tr>
<th>Sr.</th>
<th>Vowel</th>
<th>Mapping</th>
<th>Sr.</th>
<th>Vowel</th>
<th>Mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[number]</td>
<td>→ [number]</td>
<td>4</td>
<td>[o]</td>
<td>→ [number]</td>
</tr>
<tr>
<td></td>
<td>→ [number]</td>
<td></td>
<td></td>
<td>→ [number]</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>[o]</td>
<td>→ [number]</td>
<td>5</td>
<td>[e]</td>
<td>→ [number]</td>
</tr>
<tr>
<td></td>
<td>→ [number]</td>
<td></td>
<td></td>
<td>→ [number]</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>[i]</td>
<td>→ [number]</td>
<td></td>
<td>→ [number]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>→ [number]</td>
<td></td>
<td></td>
<td>→ [number]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>→ [number]</td>
<td></td>
<td></td>
<td>→ [number]</td>
<td></td>
</tr>
</tbody>
</table>
Table 15 Shahmukhi Short Vowels Mapping

<table>
<thead>
<tr>
<th>Sr.</th>
<th>Name</th>
<th>Shahmukhi Short Vowel</th>
<th>Gurmukhi Short Vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zer</td>
<td>0650 [ɨ]</td>
<td>ɿ [ɨ]</td>
</tr>
<tr>
<td>2</td>
<td>Pesh</td>
<td>064F [ʊ]</td>
<td>ɿ [ʊ]</td>
</tr>
<tr>
<td>3</td>
<td>Zabar</td>
<td>064E [ə]</td>
<td>-</td>
</tr>
</tbody>
</table>

3.7.4 Mapping other Diacritical Marks or Symbols

Shahmukhi has its own set of numerals that behave exactly as Gurmukhi numerals do with one-to-one mapping (see Table 17). Table 16 shows the mapping of other symbols and diacritical marks of Shahmukhi.

Table 16 Mapping of other Diacritical Marks or Symbols

<table>
<thead>
<tr>
<th>Sr.</th>
<th>Shahmukhi</th>
<th>Gurmukhi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Noon ghunna</td>
<td>ɲ [n]</td>
</tr>
<tr>
<td>2</td>
<td>Hamza</td>
<td>ɲ [n]</td>
</tr>
<tr>
<td>3</td>
<td>Sukun</td>
<td>ɲ [n]</td>
</tr>
<tr>
<td>4</td>
<td>Shad</td>
<td>ɲ [n]</td>
</tr>
<tr>
<td>5</td>
<td>Khari Zabar</td>
<td>ɲ [n]</td>
</tr>
<tr>
<td>6</td>
<td>do Zabar</td>
<td>ɲ [n]</td>
</tr>
<tr>
<td>7</td>
<td>do Zer</td>
<td>ɲ [n]</td>
</tr>
</tbody>
</table>

Table 17 Mapping of Digits

<table>
<thead>
<tr>
<th>Digits</th>
<th>Shahmukhi</th>
<th>Gurmukhi</th>
<th>Digits</th>
<th>Shahmukhi</th>
<th>Gurmukhi</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>•</td>
<td>०</td>
<td>5</td>
<td>५</td>
<td>Ґ</td>
</tr>
<tr>
<td>1</td>
<td>ɿ</td>
<td>१</td>
<td>6</td>
<td>६</td>
<td>ɿ</td>
</tr>
<tr>
<td>2</td>
<td>ɿ</td>
<td>२</td>
<td>7</td>
<td>ɿ</td>
<td>ɿ</td>
</tr>
<tr>
<td>3</td>
<td>ɿ</td>
<td>३</td>
<td>8</td>
<td>ɿ</td>
<td>ɿ</td>
</tr>
<tr>
<td>4</td>
<td>ɿ</td>
<td>४</td>
<td>9</td>
<td>ɿ</td>
<td>ɿ</td>
</tr>
</tbody>
</table>

3.8 Transliteration Challenges

The major challenges of transliteration between Shahmukhi-to-Gurmukhi script are:

3.8.1 Dealing with Scripts of Different Origin

The first challenge that the Shahmukhi-to-Gurmukhi transliteration system has to deal with is to handle scripts with different origins, different direction of writing, different set of alphabet, and different kind of writing system conventions. Notably,
the transliteration between languages that share similar alphabet and sound is usually not difficult, because the majority of letters remain the same which makes the transliteration task comparatively very easy as in the case of language pairs among Devanagari, Gujarati, Bengali, Gurmukhi, Oriya, Telugu, and Kannada etc., having the same script of origin called Brahmi script. Similarly, Chinese, Japanese or Korean language pairs have the same origin from Ideographic script. However, the task is significantly more difficult when the language-script pairs are considerably different, for example, English-Arabic, English-Chinese, and English-Japanese and this research work.

3.8.2 Recognition of Shahmukhi Text without Short Vowels

Most Semitic languages in both ancient and contemporary times are usually written without short vowels and other diacritic marks, often leading to potential ambiguity [34]. Arabic orthography does not provide full vocalization of the text, and the reader is expected to infer short vowels from the context of the sentence. Inferring the full form of a word is useful when developing Arabic language processing tools [35]. Take, for example, the Arabic word, written in Latin transliteration as ( nj(#) /ktaab/). One possible interpretation is the noun “book” (pronounced /kitaab/) and another is the plural of the noun “secretary”, (pronounced /kuttaab/) as shown in Table 18.

Table 18 Problem with Arabic Language Processing

<table>
<thead>
<tr>
<th>Word without Diacritic</th>
<th>Short Vowel</th>
<th>With Diacritic</th>
<th>Latin Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>كتَب</td>
<td>◐ Zer</td>
<td>كَتَب</td>
<td>Book</td>
</tr>
<tr>
<td>كتَب</td>
<td>◐ Pesh</td>
<td>كُتَّاب</td>
<td>Secretary</td>
</tr>
</tbody>
</table>

Similarly, in the written Shahmukhi script, it is not mandatory to put short vowels, called Aerab, below or above the Shahmukhi character to clear its sound leading to potential ambiguous transliteration to Gurmukhi as shown in Table 19.

In our findings, Shahmukhi corpus has just 1.66% coverage of short vowels ◐ [ʊ] (0.81415%), ◐ [i] (0.7295%), and ◐ (0.1234%) whereas the equivalent ਿ ◐ [ɪ] (4.5462%) and ◐ [ʊ] (1.5844%) in Gurmukhi corpus has 6.13% usage. The detailed analysis of Shahmukhi corpus and its comparison with Gurmukhi corpus is explained
in Section 3.9. Hence, it is a big challenge in the process of machine transliteration process to recognize the right word from the written (without diacritic) text because in a situation like this, correct meaning of the word needs to be corroborated from its neighboring words or, in worst cases, through deeper levels of n-gram analysis.

### Table 19 Ambiguous Shahmukhi Words without Short Vowels

<table>
<thead>
<tr>
<th>Sr.</th>
<th>Shahmukhi Word (without diacritics)</th>
<th>Possible Gurmukhi Transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ਗੱਲ /gall/, ਿਗੱਲ /gill/, ਗੁੱਲ /gull/, ਗੁਲ /gul/</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ਤਕ /tak/, ਤੱਕ /takk/, ਤੁਕ /tuk/</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ਮੱਖਿ /makkī, ਮੁਖਿ /mukhī</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ਹਨ /han/, ਹੁਣ /huṇ</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ਜਥੇ /jithē, ਜਤੇ /jathē</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ਵੰਦਾ /dissdā/, ਦੱਸਦਾ /dassdā</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ਅੱਕ /akk/, ਇੱਕ /ikk</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>ਟਟ /jatt', ਟੁਟ /jutt'</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>ਇਸ /is/</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>ਅਤੇ /atē', ਉੱਤੇ /uttē'</td>
<td></td>
</tr>
</tbody>
</table>

### 3.8.3 Filling the Missing Script Maps

There are many characters in the Shahmukhi script, corresponding to those there are no characters in Gurmukhi, e.g. Hamza اء [I], Do-Zabar ً[ən], Do-Zer ُ[ɪn], Aen ع[7] etc. Handling such missing script mapping makes the system more complicated. Table 20 shows typical Shahmukhi words having correct spellings with ambiguous character Aen ع[7].

While we transliterate first two words into Gurmukhi the ع[7] is transformed into two different Gurmukhi characters ع → {਋ or ਋ਠ} and interestingly, in next words it has been dropped for correct output. Clearly, a simple rule-based system is not sufficient enough to handle this situation.
Table 20 Shahmukhi Words with Correct Spellings but Ambiguous Character

<table>
<thead>
<tr>
<th>Sr</th>
<th>Shahmukhi</th>
<th>Character Ambiguity</th>
<th>Gurmukhi</th>
<th>Roman</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>عورت</td>
<td>ع→ع</td>
<td>مفت</td>
<td>aurat</td>
</tr>
<tr>
<td>2</td>
<td>ودوده</td>
<td>ع→نگ</td>
<td>تفتیغ</td>
<td>vādā</td>
</tr>
<tr>
<td>3</td>
<td>نفخ</td>
<td>ع→nil</td>
<td>ندی</td>
<td>rafī</td>
</tr>
<tr>
<td>4</td>
<td>الواواع</td>
<td>ع→nil</td>
<td>افستیغ</td>
<td>alvidā</td>
</tr>
</tbody>
</table>

3.8.4 Multiple Mappings

It is observed that there are multiple possible mappings in Gurmukhi script corresponding to a single character in the Shahmukhi script as shown in Table 21. Furthermore, the shown characters of Shahmukhi have vowel-vowel, vowel-consonant and consonant-consonant mapping.

Table 21 Multiple Mapping into Gurmukhi Script

<table>
<thead>
<tr>
<th>Sr</th>
<th>Name</th>
<th>Shahmukhi Character</th>
<th>Multiple Gurmukhi Mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Vav</td>
<td>[v]</td>
<td>द [v], दे [o], देय [ə], दु [u], द [o]</td>
</tr>
<tr>
<td>2</td>
<td>Yeh</td>
<td>[j]</td>
<td>झ [j], झि [ɪ], झे [e], झॆ [æ], झै [i], झै [i]</td>
</tr>
<tr>
<td>3</td>
<td>Noon</td>
<td>[n]</td>
<td>ठ [n], ठ [n], ठ [n], ठ [n]</td>
</tr>
</tbody>
</table>

Some of the Shahmukhi words expressing multiple mappings due to presence of one or more ambiguous characters Vav [v], Yeh [j] and or Noon [n] are collectively shown in Table 22.

Table 22 Shahmukhi Words with Multiple Gurmukhi Mappings

<table>
<thead>
<tr>
<th>Sr</th>
<th>Shahmukhi Word</th>
<th>Ambiguous Character</th>
<th>Possible Gurmukhi Transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>عورت</td>
<td>[v]</td>
<td>बेट /वोट/</td>
</tr>
<tr>
<td>2</td>
<td>ودوده</td>
<td>[j]</td>
<td>खुब /खुह/, खेट /खोह/</td>
</tr>
<tr>
<td>3</td>
<td>نفخ</td>
<td>[n]</td>
<td>विण /पिण/, वौण /पिण/</td>
</tr>
<tr>
<td>4</td>
<td>الواواع</td>
<td>[v] and [n]</td>
<td>चेत /cēn/, चीत /cīn/, चेत /cain/</td>
</tr>
<tr>
<td>5</td>
<td>الواواع</td>
<td>[n]</td>
<td>संच /जान्दा/, सच्च /जान्दा/</td>
</tr>
<tr>
<td>6</td>
<td>الواواع</td>
<td>[v] and [n]</td>
<td>मुच्छ /सुचना/, मिच्छ /sōcanā/</td>
</tr>
<tr>
<td>7</td>
<td>الواواع</td>
<td>[j] and [n]</td>
<td>बेट /dēn/, चीठ /dīn/</td>
</tr>
</tbody>
</table>
In particular, the first Shahmukhi word has two adjoining Vav [v] characters showing additional inflexible form of ambiguity. Similarly, another form of complex ambiguity is commonly observed with the presence of more that one ambiguous characters such as [v] with ؇[n] and ؇[j] with ؇[n]. Therefore, the system must be intelligent enough to resolve this character level ambiguity.

3.8.5 Word Boundary Issues in Shahmukhi Text

Shahmukhi text is written in cursive Urdu script. According to Durrani and Sarmad [142] the concept of space as a word boundary marker is not present in Urdu script (though with the increasing usage of computer it is now being used to generate correct shaping and also to separate words); therefore, the word boundary identification for Shahmukhi text is not simple due to irregular use of space in this cursive script. There are following main reasons for space insertion and omission problem in Shahmukhi:

− The space within a word is also used to generate correct shaping while writing Shahmukhi words. Therefore, space is introduced more as a tool to control the correct letter shaping and not to consistently separate words
− Many Shahmukhi words which are written as combination of two words are written as single word in Gurmukhi script.
− Urdu Zabata Takhti (UZT) 1.01 has the concept of two types of spaces. The first type of space is normal space and the second type of space is given the name, Hard Space (HS). The function of hard space is to represent space in the character sequence that represents a single word. In Unicode character set, this Hard Space is represented as Zero Width Non Joiner (ZWNJ). It is observed that in the written text normal space is used instead of hard space resulting in generation of two or more tokens of the same Shahmukhi word in Gurmukhi script. [143]
− While writing in Urdu/Arabic script, a common user finds that it is unnecessary to insert space between the two Urdu words because the correct shape is produced automatically when the first word ends with a non-joiner Urdu character [144]. The same case is observed in Shahmukhi text where many times the user omits word boundary space between the consecutive words when the first word ends with a non-joiner character.
Due to Nastaliq style, typing mistakes of space are commonly observed in Urdu (http://bbc.co.uk/Urdu) and Shahmukhi (http://www.wichaar.com) text. Thus, Shahmukhi word segmentation has to deal with both space omission and space insertion problems. Therefore, to develop a system with high transliteration accuracy it must handle the challenging task of Shahmukhi word segmentation. The examples of cases involving space insertion problem due to correct shaping in Shahmukhi text are shown in Table 23 and a few cases of wrong word interpretation due to extra space are shown in Table 24. The Shahmukhi words with space omission problem are shown Table 25.

<table>
<thead>
<tr>
<th>Sr</th>
<th>Correct Shape (with space)</th>
<th>Incorrect Shape (without space)</th>
<th>Transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>کیان دار</td>
<td>کیانلا دار</td>
<td>عربی الفاظ</td>
</tr>
<tr>
<td>2</td>
<td>ایمان دار</td>
<td>ایمانلا دار</td>
<td>عربی الفاظ</td>
</tr>
<tr>
<td>3</td>
<td>نازن پر</td>
<td>نازنپر</td>
<td>عربی الفاظ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sr</th>
<th>Words with Space</th>
<th>Wrong</th>
<th>Desired</th>
<th>Transliteration</th>
<th>Romanized</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>بیز دول</td>
<td>بیزدل</td>
<td>بیزدل</td>
<td>buzdil</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>بکر عیر</td>
<td>بکریع</td>
<td>بکریع</td>
<td>bakrīd</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>نیو پن</td>
<td>نیوپن</td>
<td>نیوپن</td>
<td>bhōlepān</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>سوے پن</td>
<td>سوےپن</td>
<td>سوےپن</td>
<td>bētukām</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word with Space Omission</th>
<th>Merged words</th>
<th>Romanized</th>
</tr>
</thead>
<tbody>
<tr>
<td>ایمانلا دار</td>
<td>عربی الفاظ</td>
<td></td>
</tr>
<tr>
<td>نازنپر</td>
<td>عربی الفاظ</td>
<td></td>
</tr>
<tr>
<td>بیزدل</td>
<td>عربی الفاظ</td>
<td></td>
</tr>
<tr>
<td>بکریع</td>
<td>عربی الفاظ</td>
<td></td>
</tr>
<tr>
<td>نیوپن</td>
<td>عربی الفاظ</td>
<td></td>
</tr>
<tr>
<td>سوےپن</td>
<td>عربی الفاظ</td>
<td></td>
</tr>
<tr>
<td>bētukām</td>
<td>عربی الفاظ</td>
<td></td>
</tr>
</tbody>
</table>
3.9 **Comparison on the basis of Corpus Analysis**

The rich source of written form in both scripts of Punjabi language is found particularly in eastern and western Punjab of India and Pakistan. Despite the influence of local dialect both scripts have common vocabulary. To estimate such facts and figures a comparative study of both the scripts has been carried out at word and character level. We have performed quantitative analysis of both Gurmukhi and Shahmukhi corpus. The size of the Gurmukhi corpus is 7.8 million having 1,59,272 unique words. On the other hand the size of Shahmukhi corpus is 8.5 million having 1,79,537 unique words. The text of Gurmukhi corpus comprises daily and regional newspapers’ reports, periodicals, magazines, short stories and Punjabi literature books etc. The Shahmukhi soft data has been obtained from various sources of Pakistan. The text was found in the form of InPage software files. This soft data was converted to Unicode format before performing the statistical analysis.

3.9.1 **Shahmukhi and Gurmukhi Character Analysis**

The character level statistical analysis of both Shahmukhi and Gurmukhi corpora is undertaken. After calculating the frequency of occurrence of each character in both corpora we ranked Gurmukhi and Shahmukhi characters as shown in Table 26. The rank vs. frequency distribution of Gurmukhi and Shahmukhi characters in the respective corpus is plotted and shown in Figure 13. It is observed that the first character of Gurmukhi spans 8.89% of the corpus whereas the first character of Shahmukhi spans much more i.e. 13.57% of the corpus. This is because ◌ਾ is used for single vowel [ɘ] sound in Gurmukhi. Where as, in Shahmukhi script the Alif character is used with other long and short vowels to produce different vowel sounds like ۆ [v, o], ۆ + mada [a], ۆ [æ, ဗ], ۆ [i] and ۆ [i]. The first 10 characters of Gurmukhi corpus contributes to 50% of the corpus whereas first 7 Shahmukhi characters contribute the same corpus coverage. Further, 95% of Shahmukhi corpus is covered by first 28 characters whereas the same coverage in Gurmukhi corpus is made by first 36 characters. The reason is that the top three characters of Shahmukhi cover 30% of its corpus and have multiple mappings into Gurmukhi character set and additionally, the 2nd and 3rd positional characters Vao ۆ[v] and Choti-ye ۆ[j] act as both vowel as well as consonant in Shahmukhi text. Another difference is that the
Shahmukhi character noon ں has high frequency of occurrence at 6.57% while in Gurmukhi corresponding mapping characters are ੢ (3.94%), ੍ (2.24%) and ੲ (1.27%) which have nearly same combined frequency of occurrence. It is also observed that there are 36 Shahmukhi characters having less that 1% occurrence in the corpus where as in Gurmukhi there are 25 such characters.

Table 26 Rank order of Gurmukhi and Shahmukhi Characters

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gurmukhi</th>
<th>%</th>
<th>Cumulative</th>
<th>Shahmukhi</th>
<th>%</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>੢</td>
<td>8.8975</td>
<td>8.8975</td>
<td>ں</td>
<td>13.57977</td>
<td>13.5798</td>
</tr>
<tr>
<td>2</td>
<td>ਠ</td>
<td>5.7737</td>
<td>14.6712</td>
<td>ੲ</td>
<td>10.71357</td>
<td>24.2933</td>
</tr>
<tr>
<td>3</td>
<td>ਛ</td>
<td>5.2240</td>
<td>19.8952</td>
<td>ੱ</td>
<td>7.43898</td>
<td>31.7323</td>
</tr>
<tr>
<td>4</td>
<td>ࢆ</td>
<td>4.7471</td>
<td>24.6423</td>
<td>ੲ</td>
<td>6.57107</td>
<td>38.3034</td>
</tr>
<tr>
<td>5</td>
<td>ਵ</td>
<td>4.6514</td>
<td>29.2937</td>
<td>ੱ</td>
<td>5.08077</td>
<td>43.3842</td>
</tr>
<tr>
<td>6</td>
<td>ਫ</td>
<td>4.5462</td>
<td>33.8399</td>
<td>ੲ</td>
<td>4.89593</td>
<td>48.2801</td>
</tr>
<tr>
<td>7</td>
<td>ਲ਼</td>
<td>4.1342</td>
<td>37.9741</td>
<td>ੲ</td>
<td>4.76675</td>
<td>53.0468</td>
</tr>
<tr>
<td>8</td>
<td>ਟ</td>
<td>4.1341</td>
<td>42.1082</td>
<td>ੲ</td>
<td>4.70883</td>
<td>57.7557</td>
</tr>
<tr>
<td>9</td>
<td>ਜ</td>
<td>4.1216</td>
<td>46.2298</td>
<td>ੲ</td>
<td>4.32901</td>
<td>62.0847</td>
</tr>
<tr>
<td>10</td>
<td>ਰ</td>
<td>3.9412</td>
<td>50.1710</td>
<td>ੲ</td>
<td>4.09279</td>
<td>66.1775</td>
</tr>
<tr>
<td>11</td>
<td>ਲ</td>
<td>3.3821</td>
<td>53.5530</td>
<td>ੲ</td>
<td>3.50742</td>
<td>69.6849</td>
</tr>
<tr>
<td>12</td>
<td>ਦੋ</td>
<td>3.3529</td>
<td>56.9060</td>
<td>ੲ</td>
<td>3.40208</td>
<td>73.0870</td>
</tr>
<tr>
<td>13</td>
<td>ਤੁ</td>
<td>3.1309</td>
<td>60.0369</td>
<td>ੲ</td>
<td>2.94626</td>
<td>76.0332</td>
</tr>
<tr>
<td>14</td>
<td>ਸੇ</td>
<td>3.6210</td>
<td>62.6579</td>
<td>ੲ</td>
<td>2.72562</td>
<td>78.7588</td>
</tr>
<tr>
<td>15</td>
<td>ਹੋ</td>
<td>2.3337</td>
<td>64.9916</td>
<td>ੲ</td>
<td>2.53606</td>
<td>81.2949</td>
</tr>
<tr>
<td>16</td>
<td>ਦੁ</td>
<td>2.2434</td>
<td>67.2349</td>
<td>ੲ</td>
<td>2.10253</td>
<td>83.3974</td>
</tr>
<tr>
<td>17</td>
<td>ਦਸ</td>
<td>2.1418</td>
<td>69.3767</td>
<td>ੲ</td>
<td>1.95542</td>
<td>85.3529</td>
</tr>
<tr>
<td>18</td>
<td>ਪ੍ਰ</td>
<td>2.1193</td>
<td>71.4960</td>
<td>ੲ</td>
<td>1.81686</td>
<td>87.1697</td>
</tr>
<tr>
<td>19</td>
<td>ਤੇ</td>
<td>1.9981</td>
<td>73.4941</td>
<td>ੲ</td>
<td>1.44817</td>
<td>88.6179</td>
</tr>
<tr>
<td>20</td>
<td>ਕੇ</td>
<td>1.9814</td>
<td>75.4755</td>
<td>ੲ</td>
<td>1.39257</td>
<td>90.0104</td>
</tr>
<tr>
<td>21</td>
<td>ਘਾ</td>
<td>1.9558</td>
<td>77.4313</td>
<td>ੲ</td>
<td>0.81415</td>
<td>90.8246</td>
</tr>
<tr>
<td>22</td>
<td>ਖੌ</td>
<td>1.8517</td>
<td>79.2831</td>
<td>ੲ</td>
<td>0.72950</td>
<td>91.5541</td>
</tr>
<tr>
<td>23</td>
<td>ਜਲ</td>
<td>1.5844</td>
<td>80.8675</td>
<td>ੲ</td>
<td>0.69393</td>
<td>92.2480</td>
</tr>
<tr>
<td>24</td>
<td>ਨਾ</td>
<td>1.5617</td>
<td>82.4292</td>
<td>ੲ</td>
<td>0.68380</td>
<td>92.9318</td>
</tr>
<tr>
<td>25</td>
<td>ਘਹ</td>
<td>1.5197</td>
<td>83.9488</td>
<td>ੲ</td>
<td>0.66233</td>
<td>93.5942</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>1.3705</td>
<td>85.3193</td>
<td>ڑ</td>
<td>0.65794</td>
<td>94.2521</td>
</tr>
<tr>
<td>27</td>
<td>ئ</td>
<td>1.2735</td>
<td>86.5927</td>
<td>ڑ</td>
<td>0.63326</td>
<td>94.8854</td>
</tr>
<tr>
<td>28</td>
<td>ے</td>
<td>1.2325</td>
<td>87.8253</td>
<td>ا</td>
<td>0.57755</td>
<td>95.4629</td>
</tr>
<tr>
<td>29</td>
<td>ئ</td>
<td>1.2267</td>
<td>89.0520</td>
<td>ع</td>
<td>0.54146</td>
<td>96.0044</td>
</tr>
<tr>
<td>30</td>
<td>ئ</td>
<td>1.1215</td>
<td>90.1735</td>
<td>ف</td>
<td>0.53095</td>
<td>96.5353</td>
</tr>
<tr>
<td>31</td>
<td>ئ</td>
<td>1.0524</td>
<td>91.2259</td>
<td>ح</td>
<td>0.50380</td>
<td>97.0391</td>
</tr>
<tr>
<td>32</td>
<td>ئ</td>
<td>1.0366</td>
<td>92.2625</td>
<td>ق</td>
<td>0.47949</td>
<td>97.5186</td>
</tr>
<tr>
<td>33</td>
<td>ئ</td>
<td>0.9525</td>
<td>93.2151</td>
<td>ز</td>
<td>0.43060</td>
<td>97.9492</td>
</tr>
<tr>
<td>34</td>
<td>ئ</td>
<td>0.8905</td>
<td>94.1056</td>
<td>س</td>
<td>0.38820</td>
<td>98.3374</td>
</tr>
<tr>
<td>35</td>
<td>ئ</td>
<td>0.8070</td>
<td>94.9126</td>
<td>ض</td>
<td>0.36593</td>
<td>98.7033</td>
</tr>
<tr>
<td>36</td>
<td>ئ</td>
<td>0.8060</td>
<td>95.7194</td>
<td>ض</td>
<td>0.29034</td>
<td>98.9937</td>
</tr>
<tr>
<td>37</td>
<td>ئ</td>
<td>0.6342</td>
<td>96.3537</td>
<td>ط</td>
<td>0.18004</td>
<td>99.1737</td>
</tr>
<tr>
<td>38</td>
<td>ئ</td>
<td>0.5158</td>
<td>96.8695</td>
<td>ظ</td>
<td>0.13409</td>
<td>99.3078</td>
</tr>
<tr>
<td>39</td>
<td>ئ</td>
<td>0.4775</td>
<td>97.3470</td>
<td>ض</td>
<td>0.12678</td>
<td>99.4346</td>
</tr>
<tr>
<td>40</td>
<td>ئ</td>
<td>0.4178</td>
<td>97.7648</td>
<td>ظ</td>
<td>0.12340</td>
<td>99.5580</td>
</tr>
<tr>
<td>41</td>
<td>ئ</td>
<td>0.3611</td>
<td>98.1259</td>
<td>ظ</td>
<td>0.11304</td>
<td>99.6710</td>
</tr>
<tr>
<td>42</td>
<td>ئ</td>
<td>0.3330</td>
<td>98.4589</td>
<td>ظ</td>
<td>0.09044</td>
<td>99.7615</td>
</tr>
<tr>
<td>43</td>
<td>ئ</td>
<td>0.2592</td>
<td>98.7181</td>
<td>ظ</td>
<td>0.07294</td>
<td>99.8344</td>
</tr>
<tr>
<td>44</td>
<td>ئ</td>
<td>0.2578</td>
<td>98.9759</td>
<td>ظ</td>
<td>0.06689</td>
<td>99.9013</td>
</tr>
<tr>
<td>45</td>
<td>ئ</td>
<td>0.1833</td>
<td>99.1593</td>
<td>ظ</td>
<td>0.02440</td>
<td>99.9257</td>
</tr>
<tr>
<td>46</td>
<td>ئ</td>
<td>0.1779</td>
<td>99.3372</td>
<td>ظ</td>
<td>0.02029</td>
<td>99.9460</td>
</tr>
<tr>
<td>47</td>
<td>ئ</td>
<td>0.1620</td>
<td>99.4992</td>
<td>ظ</td>
<td>0.01997</td>
<td>99.9659</td>
</tr>
<tr>
<td>48</td>
<td>ئ</td>
<td>0.1420</td>
<td>99.6412</td>
<td>ظ</td>
<td>0.01264</td>
<td>99.9786</td>
</tr>
<tr>
<td>49</td>
<td>ئ</td>
<td>0.1143</td>
<td>99.7556</td>
<td>ظ</td>
<td>0.00842</td>
<td>99.9870</td>
</tr>
<tr>
<td>50</td>
<td>ئ</td>
<td>0.0832</td>
<td>99.8388</td>
<td>ظ</td>
<td>0.00581</td>
<td>99.9928</td>
</tr>
<tr>
<td>51</td>
<td>ئ</td>
<td>0.0777</td>
<td>99.9165</td>
<td>ظ</td>
<td>0.00280</td>
<td>99.9956</td>
</tr>
<tr>
<td>52</td>
<td>ئ</td>
<td>0.0480</td>
<td>99.9645</td>
<td>ظ</td>
<td>0.00216</td>
<td>99.9978</td>
</tr>
<tr>
<td>53</td>
<td>ئ</td>
<td>0.0254</td>
<td>99.9899</td>
<td>ظ</td>
<td>0.00208</td>
<td>99.9999</td>
</tr>
<tr>
<td>54</td>
<td>ئ</td>
<td>0.0042</td>
<td>99.9941</td>
<td>ظ</td>
<td>0.00006</td>
<td>99.9999</td>
</tr>
<tr>
<td>55</td>
<td>ئ</td>
<td>0.0035</td>
<td>99.9976</td>
<td>ظ</td>
<td>0.00005</td>
<td>100.0000</td>
</tr>
<tr>
<td>56</td>
<td>ئ</td>
<td>0.0017</td>
<td>99.9993</td>
<td>ظ</td>
<td>0.00004</td>
<td>100.0000</td>
</tr>
<tr>
<td>57</td>
<td>ئ</td>
<td>0.0007</td>
<td>100.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The character distribution of two corpora is divided into six different groups as consonants, long vowels, short vowels (diacritics), nasalized (Nasal) characters, Adak/Shad symbol and others character symbols for further comparison. The group wise distribution of Gurmukhi and Shahmukhi corpora is plotted as pie chart and shown in Figure 14 and Figure 15 respectively.

It is observed that the distribution of Gurmukhi and Shahmukhi consonants is nearly the same at 48.59% and 48.42% respectively. Similarly, the usage of nasalized characters in both Gurmukhi and Shahmukhi corpus is also the same as 10.81% and 10.90% respectively. But the distribution of long and short vowels is different. In Gurmukhi corpus, the distribution of long vowels is 29.79% and short vowels is 6.13%. Where as in Shahmukhi the long vowels cover 37.20% of the corpus and short vowels coverage is comparatively very low at 1.66%.

![Figure 13 Character Rank with Frequency Distribution of Gurmukhi and Shahmukhi Corpus](image)

This is a significant finding that supports the fact that like Arabic and Urdu, Shahmukhi script writing also omits short vowel signs in the text. The usage of Adak ◇ in Gurmukhi corpus is relatively high 1.99% where as, the Shahmukhi equivalent Shad ◦ is just found in 0.09% in the corpus. In Gurmukhi, other symbols include
Virama or halant\textsuperscript{11} 昶 and Nukta 昶 (dot at the foot) are found to be 2.65%. The other symbols in Shahmukhi like Hamza ِ [i], Do-Zabar ٰ [ən], Do-Zer ٍ [ın] etc. are 1.7% which adds another complexity because they do not have direct mapping in the Gurmukhi script.

![Figure 14 Character Groups Distribution in Gurmukhi Corpus](image)

![Figure 15 Character Groups Distribution in Shahmukhi Corpus](image)

3.9.2 Entropy and Redundancy of Gurmukhi and Shahmukhi Corpora

Two concepts play a principal role in the mathematical theory of communication: the “entropy” or uncertainty and the “redundancy” of a set of symbols or a language. Generally, it is desirable to maximize the relative entropy of a set of symbols and therefore minimize the redundancy associated with that set.

\textsuperscript{11} Halant is used to form Conjunct forms or to signify muting of inherent vowel sound of a full consonant
The statistical result of Shahmukhi characters obtained from the corpus analysis can now be used for the calculation of Entropy (H) and redundancy of Shahmukhi script. The entropy is a statistical parameter which measures how much information is produced on average for each letter of a text in a language Shannon [145, 146]. It is also important because if text is encoded into the binary digits 0 and 1 by using the most efficient way, the entropy would measure the average number of binary digits required per letter of the original language.

According to Yannakoudakis and Angelidakis [147] the entropy (H) of a discrete set of probabilities \((p_1, p_2, \ldots, p_n)\) is defined as:

\[
H = -k \sum_{i=1}^{i=n} p_i \log_2 p_i 
\]

(1)

where \(k\) is a positive constant referring to the units selected and

where \(n\) is the number of symbols in the encoding alphabet and \(p_i\) is the probability of \(i\)th symbol

The \(\log\) here in the equation, can in principal be computed on any base but we use \(\log\) base 2 to express entropy (E) in bits per character.

According to equation 1 we have calculated the entropy (H) of Gurmukhi as well as Shahmukhi alphabet as shown in Table 27. These values are called the actual entropy and are represented as \(H_{\text{actual}}\). Therefore, the \(H_{\text{actual}}\) of Gurmukhi and Shahmukhi alphabet are 5.19 and 4.54 bits/character respectively.

\(H_{\text{maximum}}\) is reached when the probabilities of occurrence of the symbols of a particular sequence are equal, i.e. when \(p_1 = p_2 = \ldots = p_n = 1/n\) and therefore

\[
H_{\text{maximum}} = -n(1/n)\log_2 (1/n) = \log_2 n 
\]

(2)

According to equation 2 the \(H_{\text{maximum}}\) for Gurmukhi and Shahmukhi alphabet is

\(\log_2 57 = 5.832890014\) and

\(\log_2 56 = 5.807354922\) respectively.

The relative entropy or the relative uncertainty is given by the formula:

\[
H_{\text{relative}} = \frac{H_{\text{actual}}}{H_{\text{maximum}}} 
\]

(3)

Hence, The relative entropy of Gurmukhi and Shahmukhi alphabet is

\(H_{\text{relative}} = 0.874\) and

\(H_{\text{relative}} = 0.782\) respectively.
The redundancy measures the amount of constraint imposed on a text in the language due to its syntactical rules where for every syntactic rule there is a constraint that must introduce some redundancy. For example, symbol Hamza [i] will be inserted if two vowels are appearing together in Shahmukhi. The statistical structure of the language itself introduces a certain amount of redundancy. In Shahmukhi script, for example, there is a strong tendency for Alif [i] to be followed by اء [æ], [t] to be followed by اء [æ] and in Gurmukhi script the symbol ṣ has a strong tendency towards symbol े [e] and is strongly associated with ṭ [æ] symbol.

The redundancy can be defined as the difference between $H_{\text{maximum}}$ and $H_{\text{actual}}$ expressed as a fraction of $H_{\text{maximum}}$ as:

$$\text{Redundancy} = \frac{H_{\text{maximum}} - H_{\text{actual}}}{H_{\text{maximum}}}$$

or

$$\text{Redundancy} = 1 - H_{\text{relative}} \quad (4)$$

According to equation 4, we have calculated the redundancy of both Gurmukhi and Shahmukhi as:

$$1 - H_{\text{relative}} = 0.126 \text{ and } 0.2176 \text{ respectively.}$$

Therefore, for Gurmukhi the relative uncertainty is 87.4% and the redundancy is 12.6%. Whereas, for Shahmukhi the relative uncertainty is 78.2% and the redundancy is 21.8%.

Yannakoudakis and Angelidakis [147] studied the entropy (or uncertainty) and redundancy of for English alphabet in a dictionary of 93,681 words and the results are:

$$\text{Entropy (}H_{\text{actual}}\text{)} = 4.2 \text{ bits/character}$$

$$H_{\text{relative}} = 0.893$$

$$\text{Redundancy} = 0.107$$

As shown in Figure 16, we can conclude that the Shahmukhi script has higher level of redundancy than Gurmukhi script and that data compression is more expensive in Shahmukhi as compared to Gurmukhi whereas Gurmukhi script has higher level of redundancy than English and data compression is more expensive in Gurmukhi as compared to English. Hence, among Shahmukhi, Gurmukhi and English, Shahmukhi has highest level of redundancy. [147]
Figure 16 Comparison of Entropy and Redundancy of Gurmukhi, Shahmukhi and English

Table 27 Information Content of Letters in the Shahmukhi Corpus

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gurmukhi</th>
<th>Shahmukhi</th>
<th>Probability P(g)</th>
<th>Probability P(s)</th>
<th>Entropy -P(g)log2P(g)</th>
<th>Entropy -P(s)log2P(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.19</td>
<td>0.874</td>
<td>0.126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4.54</td>
<td>0.782</td>
<td>0.2176</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4.2</td>
<td>0.893</td>
<td>0.107</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.088975</td>
<td>0.1357977</td>
<td>0.391161072</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.057737</td>
<td>0.1071357</td>
<td>0.345243593</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.05224</td>
<td>0.0743898</td>
<td>0.278868865</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.047471</td>
<td>0.0657107</td>
<td>0.258093748</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.046514</td>
<td>0.0508077</td>
<td>0.2184126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.045462</td>
<td>0.0489593</td>
<td>0.213084252</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.041342</td>
<td>0.0476675</td>
<td>0.203080853</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.041341</td>
<td>0.0470883</td>
<td>0.207588184</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.041216</td>
<td>0.0432901</td>
<td>0.19609632</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.039412</td>
<td>0.0409279</td>
<td>0.188709197</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.033821</td>
<td>0.0350742</td>
<td>0.169529251</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.033529</td>
<td>0.0340208</td>
<td>0.165934381</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.031309</td>
<td>0.0294626</td>
<td>0.14981648</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>भ</td>
<td>0.021418</td>
<td>0.0195542</td>
<td>0.118763504</td>
<td>0.110997024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>न</td>
<td>0.021193</td>
<td>0.0181686</td>
<td>0.117838767</td>
<td>0.105058275</td>
<td></td>
<td></td>
</tr>
<tr>
<td>घ</td>
<td>0.019981</td>
<td>0.0144817</td>
<td>0.112797289</td>
<td>0.08847776</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ङ</td>
<td>0.019814</td>
<td>0.0139257</td>
<td>0.112094456</td>
<td>0.085867347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>भ</td>
<td>0.019558</td>
<td>0.0081415</td>
<td>0.111013112</td>
<td>0.056505997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ॠ</td>
<td>0.018517</td>
<td>0.007295</td>
<td>0.106565443</td>
<td>0.051786303</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ऌ</td>
<td>0.015844</td>
<td>0.0069393</td>
<td>0.094745846</td>
<td>0.04976168</td>
<td></td>
<td></td>
</tr>
<tr>
<td>य</td>
<td>0.015617</td>
<td>0.006838</td>
<td>0.093713539</td>
<td>0.049180331</td>
<td></td>
<td></td>
</tr>
<tr>
<td>र</td>
<td>0.015197</td>
<td>0.0066233</td>
<td>0.091790938</td>
<td>0.047940996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>द</td>
<td>0.013705</td>
<td>0.0065794</td>
<td>0.084822354</td>
<td>0.047686361</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ढ</td>
<td>0.012735</td>
<td>0.0063326</td>
<td>0.080167554</td>
<td>0.046246891</td>
<td></td>
<td></td>
</tr>
<tr>
<td>।</td>
<td>0.012325</td>
<td>0.0057755</td>
<td>0.07816846</td>
<td>0.042945685</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ऻ</td>
<td>0.012267</td>
<td>0.0054146</td>
<td>0.077884087</td>
<td>0.040766142</td>
<td></td>
<td></td>
</tr>
<tr>
<td>अ</td>
<td>0.011215</td>
<td>0.0053095</td>
<td>0.072655554</td>
<td>0.040124997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MediaPlayer.res default font</td>
<td>0.010524</td>
<td>0.005038</td>
<td>0.069144501</td>
<td>0.038454717</td>
<td></td>
<td></td>
</tr>
<tr>
<td>रि</td>
<td>0.010366</td>
<td>0.0047949</td>
<td>0.06833264</td>
<td>0.036941269</td>
<td></td>
<td></td>
</tr>
<tr>
<td>णि</td>
<td>0.009525</td>
<td>0.004306</td>
<td>0.063951471</td>
<td>0.033842731</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ष</td>
<td>0.008905</td>
<td>0.003882</td>
<td>0.060653457</td>
<td>0.031090877</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ट</td>
<td>0.00807</td>
<td>0.0036593</td>
<td>0.05611245</td>
<td>0.029619167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ढ</td>
<td>0.008068</td>
<td>0.0029034</td>
<td>0.056101429</td>
<td>0.024469974</td>
<td></td>
<td></td>
</tr>
<tr>
<td>धि</td>
<td>0.006342</td>
<td>0.0018004</td>
<td>0.046301968</td>
<td>0.016415087</td>
<td></td>
<td></td>
</tr>
<tr>
<td>य</td>
<td>0.005158</td>
<td>0.0013409</td>
<td>0.0391955</td>
<td>0.012795649</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ल</td>
<td>0.004775</td>
<td>0.0012678</td>
<td>0.036816604</td>
<td>0.012200619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ञ</td>
<td>0.004178</td>
<td>0.001234</td>
<td>0.033018616</td>
<td>0.011923453</td>
<td></td>
<td></td>
</tr>
<tr>
<td>र</td>
<td>0.003611</td>
<td>0.0011304</td>
<td>0.029297436</td>
<td>0.01106543</td>
<td></td>
<td></td>
</tr>
<tr>
<td>रि</td>
<td>0.00333</td>
<td>0.0009044</td>
<td>0.027406773</td>
<td>0.009144164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>सि</td>
<td>0.002592</td>
<td>0.0007294</td>
<td>0.022269735</td>
<td>0.007601079</td>
<td></td>
<td></td>
</tr>
<tr>
<td>झि</td>
<td>0.002578</td>
<td>0.0006689</td>
<td>0.022169594</td>
<td>0.007054167</td>
<td></td>
<td></td>
</tr>
<tr>
<td>झ</td>
<td>0.001833</td>
<td>0.000244</td>
<td>0.016664862</td>
<td>0.002928203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ञ</td>
<td>0.001779</td>
<td>0.0002029</td>
<td>0.016250663</td>
<td>0.002488963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ञ</td>
<td>0.00162</td>
<td>0.0001997</td>
<td>0.015017061</td>
<td>0.002454289</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ञ</td>
<td>0.00142</td>
<td>0.0001264</td>
<td>0.013433049</td>
<td>0.001636844</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The word level statistical analysis of both Shahmukhi and Gurmukhi corpora is performed. After calculating the frequency of occurrence of each word in their respective corpora we ranked them according to their frequency of occurrence and selected top 20 words of Gurmukhi and Shahmukhi as shown in Table 28.

It is interesting to note that 15 words are same out of top 20 most frequent words. The most frequent Gurmukhi word is ਦੇ [दे] which is 2nd most frequently occurring word in Shahmukhi corpus. The top most word in Shahmukhi is ਤੇ [ਤੇ] which is on 6th position in Gurmukhi. In Gurmukhi word ਤੇ [ਤੇ] is very frequently used and the same word is also of common usage in Shahmukhi. Another, Shahmukhi word ਈ [ਈ] is very frequent in use but it is not frequently used in Gurmukhi writing because in Shahmukhi writing ਈ [ਈ] is used in place of ਈ [ਈ]. For example, consider a sentence in Gurmukhi as:

ਪੰਜਾਬੀ ਮੇਰੀ ਮਾਂ ਬੋਲੀ ਹੈ

The same sentence in Shahmukhi can be expressed as:

pañjābī mērī māṃ bōlī ē
Another interesting point is that in Gurmukhi writing ਅਤੇ [ਅਤੇ], ਦਾਰੂ [ਕਾਰੂ] and ਮਿਸ਼ਨ [ਸ਼ਾਹ] are commonly used but in Shahmukhi text they are not so common, instead the usage of ਦੇ [ਦੇ] and ਹਨ [ਹਨ] is common in place of ਅਤੇ [ਅਤੇ] and ਦਾਰੂ [ਕਾਰੂ] respectively and unlike Sikh names (e.g. ਸ਼ਮਸ਼ੇਰ ਸ਼ਿਸ਼ਨ) in Gurmukhi corpus having terminating word ਮਿਸ਼ਨ [ਸ਼ਾਹ], Muslim names (e.g. ਚਾਰਈਨਾਂ) have different terminating words.

Table 28 Top 20 Rank of Gurmukhi and Shahmukhi Words

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gurmukhi</th>
<th>Percentage</th>
<th>Shahmukhi</th>
<th>Matched</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ਦੇ</td>
<td>2.3228</td>
<td>ਦੇ</td>
<td>[ਦੇ,6]</td>
<td>2.7475766</td>
</tr>
<tr>
<td>2</td>
<td>ਹੈ</td>
<td>2.1070</td>
<td>ਹੈ</td>
<td>[ਹੈ,1]</td>
<td>2.19781251</td>
</tr>
<tr>
<td>3</td>
<td>ਦੀ</td>
<td>1.9403</td>
<td>ਦੀ</td>
<td>[ਦੀ]</td>
<td>1.89280506</td>
</tr>
<tr>
<td>4</td>
<td>ਨਾਂ</td>
<td>1.7634</td>
<td>ਨਾਂ</td>
<td>[ਨਾਂ,3]</td>
<td>1.74061775</td>
</tr>
<tr>
<td>5</td>
<td>ਨਾਂ</td>
<td>1.4828</td>
<td>ਨਾਂ</td>
<td>[ਨਾਂ,5]</td>
<td>1.52065101</td>
</tr>
<tr>
<td>6</td>
<td>ਨਾਂ</td>
<td>1.3535</td>
<td>ਹਿੰਦੀ</td>
<td>[ਹਿੰਦੀ,4]</td>
<td>1.40521212</td>
</tr>
<tr>
<td>7</td>
<td>ਨਾਂ</td>
<td>1.2535</td>
<td>ਵਿੱਚ</td>
<td>[ਵਿੱਚ,7]</td>
<td>1.23940783</td>
</tr>
<tr>
<td>8</td>
<td>ਅਤੇ</td>
<td>1.0487</td>
<td>ਕੇ</td>
<td>[ਕੇ,17]</td>
<td>1.03915735</td>
</tr>
<tr>
<td>9</td>
<td>ਦੀ</td>
<td>1.0389</td>
<td>ਓ</td>
<td>[ਓ,10]</td>
<td>0.98657701</td>
</tr>
<tr>
<td>10</td>
<td>ਦੀ</td>
<td>0.9062</td>
<td>ਵਿੱਚ</td>
<td>[ਵਿੱਚ,12]</td>
<td>0.88345669</td>
</tr>
<tr>
<td>11</td>
<td>ਦੀ</td>
<td>0.8871</td>
<td>ਟਾਨ</td>
<td>[ਟਾਨ,15]</td>
<td>0.81584092</td>
</tr>
<tr>
<td>12</td>
<td>ਦੀ</td>
<td>0.8635</td>
<td>ਰਾਜ</td>
<td>[ਰਾਜ,14]</td>
<td>0.76271503</td>
</tr>
<tr>
<td>13</td>
<td>ਦੀ</td>
<td>0.8178</td>
<td>ਨਿੰਦਾ</td>
<td>[ਨਿੰਦਾ,20]</td>
<td>0.7238826</td>
</tr>
<tr>
<td>14</td>
<td>ਦੀ</td>
<td>0.8048</td>
<td>ਪਾਁਂਹਰ</td>
<td>[ਪਾਁਂਹਰ]</td>
<td>0.69085486</td>
</tr>
<tr>
<td>15</td>
<td>ਦੀ</td>
<td>0.7792</td>
<td>ਨੀ</td>
<td>[ਨੀ,9]</td>
<td>0.66463561</td>
</tr>
<tr>
<td>16</td>
<td>ਦੀ</td>
<td>0.7647</td>
<td>ਨਿੰਦਾ</td>
<td>[ਨਿੰਦਾ]</td>
<td>0.65451016</td>
</tr>
<tr>
<td>17</td>
<td>ਦੀ</td>
<td>0.7398</td>
<td>ਕੇ</td>
<td>[ਕੇ,11]</td>
<td>0.60673068</td>
</tr>
<tr>
<td>18</td>
<td>ਦੀ</td>
<td>0.7369</td>
<td>ਨਾਂ</td>
<td>[ਨਾਂ,2]</td>
<td>0.58856378</td>
</tr>
<tr>
<td>19</td>
<td>ਮਿਸ਼ਨ</td>
<td>0.7086</td>
<td>ਫਿਲਚ</td>
<td>[ਫਿਲਚ]</td>
<td>0.55552513</td>
</tr>
<tr>
<td>20</td>
<td>ਨਾਂ</td>
<td>0.6572</td>
<td>ਨਾਂ</td>
<td>[ਨਾਂ]</td>
<td>0.51275382</td>
</tr>
</tbody>
</table>
3.10 Conclusion

There are more than six thousand living languages in the world and adding to the diversity is the fact that some languages are written in different scripts in different regions of the world. The multitude of foreign languages and mutually incomprehensible scripts of the same language pose a barrier to information as we cannot all learn every language or script in use worldwide. Therefore, if we can get around the language barrier or at least the script barrier, we can access much more of the world's culture and can explore its abundant richness. Language analysis shows that, in global view of Information technology, it is conceivable that if we can bridge the intercommunication gap between the 389 (nearly 6%) world languages, 94% of world population will benefit immensely.

To record Punjabi language, Punjabi speakers use two different scripts, Perso-Arabic (referred as Shahmukhi) and Gurmukhi. Shahmukhi is used by the people of Western Punjab in Pakistan, whereas Gurmukhi is used by most of the people of Eastern Punjab. This research work also contributes towards IT at the global level by bridging the intercommunication gap between the two scripts of Punjabi Language. We have presented a comparative study of the two scripts of Punjabi language that are mutually incomprehensible to each other. They have different alphabet, different writing system rules and direction. Grapheme or character level script mapping has shown missing and multiple character mappings to generate character level ambiguity.

The corpus-based analysis of Shahmukhi script has shown the fact that Shahmukhi text is affected with missing diacritics and word segmentation problems. The character and word level comparison of both the scripts suggested that despite of few differences at both character and word level both scripts are representing the same language.

Thus, we can apply direct orthographical approach between the two scripts and for handling of missing short vowels and word segmentation problems we might used target script resources because lexical resources of Gurmukhi script has the missing sound and word segmentation problem is not present in Gurmukhi script.