MULTILINGUAL CONVERSATION ASCII TO UNICODE IN INDIC SCRIPT

Dr. Rajwinder Singh[1] and Charanjiv Singh Saroa[2]

[1]Department of Punjabi, Punjabi University, Patiala, India
[2]Department of Computer Engineering, Punjabi University Patiala, India

ABSTRACT

In this paper we discuss the various ASCII based scripts that are made for Indian languages and the problems associated with these types of scripts. Then we will discuss the solution we suggest to overcome these problems in the form of "Multilingual ASCII to Unicode Converter". We also explain the need of regional languages for the development of a person. This paper also contains information of UNICODE and various other issues related to regional languages.

KEYWORDS

Keywords: NLP, Punjabi, Mother Tongue, Gurmukhi, Font Conversion, UNICODE, ASCII, Keyboard Layout.

1. INTRODUCTION

According UNESCO reports About half of the 6,000 or so languages spoken in the world are under threat. Over the past three centuries, languages have died out and disappeared at a dramatic and steadily increasing pace, especially in the Americas and Australia. Today at least 3,000 tongues are endangered, seriously endangered or dying in many parts of the world. [1] A language disappears when its speakers disappear or when they shift to speaking another language. [2] It is also proved from various researches that the primary education of the child should be in the mother tongue of the child instead of in any other language.

In this new world of technology, most of the information is available on internet in e-form. But in regional languages, due to various technical issues like ASCII based fonts, keyboard layouts, lack of awareness of UNICODE, non availability of spell checkers, it is not easy. In regional languages, most of the available fonts are ASCII based instead of UNICODE. We need an intelligent code converter that can change ASCII to UNICODE based scripts.

2. REGIONAL LANGUAGE

A regional language is a language spoken in an area of a state or country, whether it is a small area, a state, a county, or some wider area. Regional languages, as defined by the European Charter for Regional or Minority Languages are traditionally used by part of the population in a state, but which are not official state language dialects, migrant languages or artificially created languages. [3]

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 83– 94, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70308

Regional language is mainly spoken in small parts. It changes with the change in culture, religion and economy of the region. In a country, there may be hundreds of regional languages and each language may have further variations. A language is not always limited within the boundaries of a country. One language may be part of more than one country. The eighth schedule of the constitution of India lists 22 scheduled languages. [4] The 22 is for scheduled languages as per the Indian Constitution. It is hard to use computer with all the languages. We need to train computer in each particular language. Computational linguistics is the study of computer system for understanding and generating natural language. [5] Linguistics is the scientific study of language. [6] V.Rajaraman writes in 1998 the government took proactive steps to promote Information Technology by giving incentives such as tax breaks and reduced import duties. [7] Communication infrastructure also improved. The cost of computers came down. All these resulted in a rapid growth of the software services industry with annual growth rate exceeding 30%. We will identify the significant events during each of the above referred periods and explain their impact on the development of IT in India.

2.1 IMPORTANCE OF REGIONAL LANGUAGES

We learn culture, religion and respect from our mother tongue. Regional languages contain lots of sources of understanding community and culture. Regional language/mother tongue gives us:

a) The connections to our roots.
b) Knowledge of our culture.
c) Sense of belonging.
d) Better linguistic skills.
f) Sharper children.
g) A better society.

2.2 NEED OF EDUCATION IN REGIONAL LANGUAGES

Primary education of the child should be in the mother tongue of the child instead of in any other language. Some of the statements are listed below also point toward this.

The following statement from the book titled “The Use of Vernaculars in Education” published by the United Nation‘s Educational Scientific and Cultural Organization (UNESCO) in 1953 is an eye opener. The book presents the essence of international research and wisdom on the issue:

It is axiomatic that the best medium for teaching a child is his mother tongue. Psychologically, it is the system of meaningful signs that in his mind works automatically for expression and understanding. Sociologically, it is means of identification among the members of community to which he belongs. [8]

Children learn best when they are taught in their mother tongue, particularly in the earliest years. Experience in many countries shows that bilingual education, which combines instruction in the mother tongue with teaching in the dominant national language, can open educational and other opportunities. In the Philippines students proficient in the two languages of the bilingual education policy (Tagalog and English) outperformed students who did not speak Tagalog at home. [9]

It is axiomatic that the best medium for teaching a child is his mother tongue. Psychologically, it is the system of meaningful signs that in his mind works automatically for expression and understanding sociologically. it is means of identification among the members of the community to which he belongs. [10] Educationally, he learns more quickly through it than through an unfamiliar linguistic medium.

2.3 CHALLENGES TO USE REGIONAL SCRIPTS IN E-FORM OF INFORMATION

It is very challenging to provide information in e-form using regional languages. Some of the challenges are:

2.3.1 Fonts & Keyboard Layouts:

There are 100s of fonts for every language. In most of the regional languages mainly each font has its own keyboard layout. That result in changing the content of the matter with the change of font and most of the information become useless. Like if Correct sentence is " I am Going " in font Arial , with change in font it becomes something like "r kj pjras". This will never happens in English because all the fonts are created with same keyboard layout and same coding system. But this type of problem is very common in regional languages. A problem with ASCII based fonts for Regional scripts is that there is no standardization of mapping of script characters with keyboard keys. We presently work on 5 Indic Scripts and some of the ASCII based tables of various fonts of these scripts are:

Table 1. For Gurmukhi script of various ASCII Based Fonts

	Remington Style							**Phonetic Style**							
Decimal Code	Joy	Asees	BJanmeja5A	Prime Ja	GurmukhiLys 010	TERAFONT-Maharaja	Gul-P5Bold	AnmolLipi	LMP_Amrik	Akhar	Amritboli	DRChatrikWeb	GurmukhiIIGS	Chatrik	Sukhmani
65	A	A	A	A	A	A	**A**	A	A	A	A	**A**	A	**A**	A
66	B	B	B	B	B	B	**B**	B	B	B	B	**B**	B	**B**	B
67	C	C	C	C	C	C	**C**	C	C	C	C	**C**	C	**C**	C
68	D	D	D	D	D	D	**D**	D	D	D	D	**D**	D	**D**	D
69	E	E	E	E	E	E	**E**	E	E	E	E	**E**	E	**E**	E
70	F	F	F	F	F	F	**F**	F	F	F	F	**F**	F	**F**	F
71	G	G	G	G	G	G	**G**	G	G	G	G	**G**	G	**G**	G
72	H	H	H	H	H	H	**H**	H	H	H	H	**H**	H	**H**	H
73	I	I	I	I	I	I	**I**	I	I	I	I	**I**	I	**I**	I
74	J	J	J	J	J	J	**J**	J	J	J	J	**J**	J	**J**	J
75	K	K	K	K	K	K	**K**	K	K	K	K	**K**	K	**K**	K
76	L	L	L	L	L	L	**L**	L	L	L	L	**L**	L	**L**	L
77	M	M	M	M	M	M	**M**	M	M	M	M	**M**	M	**M**	M
78	N	N	N	N	N	N	**N**	N	N	N	N	**N**	N	**N**	N
79	O	O	O	O	O	O	**O**	O	O	O	O	**O**	O	**O**	O
80	P	P	P	P	P	P	**P**	P	P	P	P	**P**	P	**P**	P
81	Q	Q	Q	Q	Q	Q	**Q**	Q	Q	Q	Q	**Q**	Q	**Q**	Q
82	R	R	R	R	R	R	**R**	R	R	R	R	**R**	R	**R**	R
83	S	S	S	S	S	S	**S**	S	S	S	S	**S**	S	**S**	S
84	T	T	T	T	T	T	**T**	T	T	T	T	**T**	T	**T**	T
85	U	U	U	U	U	U	**U**	U	U	U	U	**U**	U	**U**	U
86	V	V	V	V	V	V	**V**	V	V	V	V	**V**	V	**V**	V
87	W	W	W	W	W	W	**W**	W	W	W	W	**W**	W	**W**	W
88	X	X	X	X	X	X	**X**	X	X	X	X	**X**	X	**X**	X
89	Y	Y	Y	Y	Y	Y	**Y**	Y	Y	Y	Y	**Y**	Y	**Y**	Y
90	Z	Z	Z	Z	Z	Z	**Z**	Z	Z	Z	Z	**Z**	Z	**Z**	Z
91	[	[	[	[	[	[	**[**	[	[	[	[	**[**	[	**[**	[
92	\	\	\	\	\	\	****	\	\	\	\	****	\	****	\

93	]	]	]	]	]	]	]	]	]	]	]	]	]	]	]
94	^	^	^	^	^	^	^	^	^	^	^	^	^	^	^
95	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_
96	`	`	a	`	`	`	`	`	`	`	`	`	`	`	`
97		a	a	a	a	a	a	a	a	a	B	a	a	a	a
98	B	b	b	b	b	b	b	b	b	b	b	b	b	b	b
99	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
100	d	d	d	d	d	d	d	d	d	d	d	d	d	d	d
101	e	e	e	e	e	e	e	e	e	e	e	e	e	e	e
102	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f
103	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
104	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h
105	i	i	i	i	i	i	i	i	i	i	i	i	i	i	i
106	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j
107	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k
108	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l
109	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m
110	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
111	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
112	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p
113	q	q	q	q	q	q	q	q	q	q	q	q	q	q	q
114	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r
115	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s
116	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
117	u	u	u	u	u	u	u	u	u	u	u	u	u	u	u
118	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v
119	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w
120	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
121	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
122	z	z	z	z	z	z	z	z	z	z	z	z	z	z	z

Table 2. For Hindi(Devnagri), Gujrati, Malayalam, Tamil script of various ASCII Based Fonts

Hindi (Devnagri)					Gujrati				Malayalam				Tamil			
Decimal code	Aakriti	Chanakya	xdvng	Kundli	Gujrati Saral-1	LMG-Laxmi	Saumil_guj2	Chitra	Manorama	Aruna	Deepa	Gayathri	ELCOT-Kanchi	TM-TTValluvar	Baamini	divya
65	A	A	A	A	A	A	A	A	A	A	A	A	A	அ	A	A
66	B	B	B	B	B	B	B	B	B	B	B	B	B	ஆ	B	B
67	C	C	X	C	C	C	C	C	C	C	C	C	C	இ	C	C
68	D	D	Δ	D	D	D	D	D	D	D	D	D	D	ஈ	D	D
69	E	E	E	E	E	E	E	E	E	E	E	E	E	உ	E	E
70	F	F	Φ	F	F	F	F	F	F	F	F	F	F	ஊ	F	F
71	G	G	Γ	G	G	G	G	G	G	G	G	G	G	எ	G	G
72	H	H	H	H	H	H	H	H	H	H	H	H	H	ஏ	H	H
73	I	I	I	I	I	I	I	I	I	I	I	I	I	ஐ	I	I
74	J	J	ϑ	J	J	J	J	J	J	J	J	J	J	ஒ	J	J

75	K	K	K	**K**	K	K	K	K	K	K	K	K	K	ஓ	K	K
76	L	L	Λ	**L**	L	L	L	L	L	L	L	L	L	க	L	L
77	M	M	M	**M**	M	M	M	M	M	M	M	M	M	ங	M	M
78	N	N	N	**N**	N	N	N	N	N	N	N	N	N	ச	N	N
79	O	O	O	**O**	O	O	O	O	O	O	O	O	O	ஞ	O	O
80	P	P	Π	**P**	P	P	P	P	P	P	P	P	P	ட	P	P
81	Q	Q	Θ	**Q**	Q	Q	Q	Q	Q	Q	Q	Q	Q	ண	Q	Q
82	R	R	P	**R**	R	R	R	R	R	R	R	R	R	த	R	R
83	S	S	Σ	**S**	S	S	S	S	S	S	S	S	S	ந	S	S
84	T	T	T	**T**	T	T	T	T	T	T	T	T	T	ப	T	T
85	U	U	Y	**U**	U	U	U	U	U	U	U	U	U	ம	U	U
86	V	V	ς	**V**	V	□	V	V	V	V	V	V	V	ய	V	V
87	W	W	Ω	**W**	W	□	W	W	W	W	W	W	W	ர	W	W
88	X	X	Ξ	**X**	X	X	X	X	X	X	X	X	X	ல	X	X
89	Y	Y	Ψ	**Y**	Y	Y	Y	Y	Y	Y	Y	Y	Y	வ	Y	Y
90	Z	Z	Z	**Z**	Z	Z	Z	Z	Z	Z	Z	Z	Z	ழ	Z	Z
91	[	[	[	**[**	[	[	[	[	[	[	[	[	[	ள	[	[
92	\	\	∴	\	\	\	\	\	\	\	\	\	\	ற	\	\
93	]	]	]	**]**	]	]	]	]	]	]	]	]	]	ன	]	]
94	^	^	⊥	**^**	^	^	^	^	^	^	^	^	^	ஸ	^	^
95	–	–	_	**–**	–	–	–	–	–	–	–	–	–	ஜ	_	_
96	`	`	α	`	`	`	`	`	`	`	`	`	`	ஷ	`	`
97	a	a	α	**a**	a	a	a	a	a	a	a	a	a	ஹ	a	a
98	b	b	β	**b**	b	b	b	b	b	b	b	b	b	க்ஷ	b	b
99	c	c	χ	**c**	c	c	c	c	c	c	c	c	c	ஸ்ரீ	c	c
100	d	d	δ	**d**	d	d	d	d	d	d	d	d	d	க்	d	d
101	e	e	ε	**e**	e	e	e	e	e	e	e	e	e	ங்	e	e
102	f	f	ϕ	**f**	f	f	f	f	f	f	f	f	f	ச்	f	f
103	g	g	γ	**g**	g	g	g	g	g	g	g	g	g	ஞ்	g	g
104	h	h	η	**h**	h	h	h	h	h	h	h	h	h	ட்	h	h
105	i	i	ι	**i**	i	i	i	i	i	i	i	i	i	ண்	i	i
106	j	j	φ	**j**	j	j	j	j	j	j	j	j	j	த்	j	j
107	k	k	κ	**k**	k	k	k	k	k	k	k	k	k	ந்	k	k
108	l	l	λ	**l**	l	l	l	l	l	l	l	l	l	ப்	l	l
109	m	m	μ	**m**	m	m	m	m	m	m	m	m	m	ம்	m	m
110	n	n	ν	**n**	n	n	n	n	n	n	n	n	n	ய்	n	n
111	o	o	o	**o**	o	o	o	o	o	o	o	o	o	ர்	o	o
112	p	p	π	**p**	p	p	p	p	p	p	p	p	p	ல்	p	p
113	q	q	θ	**q**	q	q	q	q	q	q	q	q	q	வ்	q	q

114	r	r	ρ	**r**	r	r	r	r	r	r	r	r	r	ਪ੍	r	r
115	s	s	σ	**s**	s	s	s	s	s	s	s	s	s	ਜ਼	s	s
116	t	t	τ	**t**	t	t	t	t	t	t	t	t	t	ਫ਼	t	t
117	u	u	υ	**u**	u	u	u	u	u	u	u	u	u	ਕ਼	u	u
118	v	v	ϖ	**v**	v	v	v	v	v	v	v	v	v	ਖ਼	v	v
119	w	w	ω	**w**	w	w	w	w	w	w	w	w	w	ਗ਼	w	w
120	x	x	ξ	**x**	x	x	x	x	x	x	x	x	x	ਜ਼	x	x
121	y	y	ψ	**y**	y	y	y	y	y	y	y	y	y	ੜ	y	y
122	z	z	ζ	**z**	z	z	z	z	z	z	z	z	z	ਫ਼	z	z

Above tables explain the variation in various fonts with same code. Like in ASCII code 65 in Joy, Asees, Bjanmeja5A, GurmukhiLys 010 represents Bindi ("ਂ"), in Prime Ja it represents adak("ੱ") in AnmolLipi, AmritLipi, Amritboli 65 is used to represents aira("ਅ")in Akhar font it is for ura+hora("ੳ") and in Chatrik font it is for ura("ੳ"),. And when we change the font of the text information of the text is also changed and become useless as shown in following table.

Tabel 3: Various Punjabi Fonts with same key impression

Font Name	**Text**
AnmolLipi	pMjwbI XUnIvristI, pitAwlw
Akhar	pMjwbI XUnIvristI, pitAwlw
Chatrik	**pMjwbI XUnIvristI, pitAwlw**
Joy	pMjwbI XUnIvristI, pitAwlw
Asees	pMjwbI XUnIvristI, pitAwlw
Bjanmeja5A	pMjwbI XUnIvristI, pitAwlw
GurmukhiLys 010	pMjwbI XUnIvristI, pitAwlw

Computer understand data in the form of 0 and 1. According to that ASCII codes are created for Roman script. But these codes are not compatible with other Indic scripts. Some of the variations are;

Table 4: Script and code Compare

Dec Code	Joy	Asees	BJanmeja5A	Prime Ja	GurmukhiLys 010	AnmolLipi	AmritLipi	Akhar	Amritboli	DRChatrikWeb
65	A	A	A	A	A	A	A	A	A	**A**
69	E	E	E	E	E	E	E	E	E	**E**
70	F	F	F	F	F	F	F	F	F	**F**
87	W	W	W	W	W	W	W	W	W	**W**

Table 5: Script and code Compare

Hindi (Devnagri)				Gujrati			Malayalam		Tamil	
	Aakriti	Chanakya	xdvng	Gujrati Saral-1	LMG-Laxmi	Saumil_guj2	Manorama	Aruna	ELCOT-Kanchi	TM-TTValluvar
65	A	A	A	A	A	A	A	A	A	அ
69	E	E	E	E	E	E	E	E	E	உ
77	M	M	M	M	M	M	M	M	M	ங
87	W	W	Ω	W	□	W	W	W	W	ற

2.3.2 Non Unicode Fonts:

Mostly the data available in regional languages are in ASCII based Indic scripts. we want to display that data on website we have to upload the required font for that data and user firstly have to download and install that font, Only then the user can view that information. There are thousands of fonts used for create data in Indian Scripts. Only in Punjabi (Gurmukhi) alone has more than 225 popular fonts which are still in use to publishing books, magazine, news paper etc. Publishers are still working with ASCII coding based fonts. They have not used Unicode based fonts to following reasons:

- People resist to change, due to Unicode typing issues. [11]
- Lack of awareness of Unicode standard.
- Little support of Unicode system in publishing software that they are using.
- Less availability of Unicode fonts to represent text in different style and designs

It is always better to display information in Unicode based fonts while displaying the information on website. The information presently available to us is mainly in ASCII based fonts. So we convert that ASCII based information into Unicode based fonts so that it can be available on internet. The information displayed in Unicode can be seen on any computer without installing font. Other advantage of Unicode based fonts is that it is searchable on search engines like Google, Yahoo, Ask, Bing etc.

2.3.3 Some special Symbols:

Some text can contain some unique type of symbols that are not available in ASCII codes and even not in UNICODE system.

2.3.4 Typing problem:

By default Mainly each computer contain English (roman) keyboard. And most of the user are not aware of UNICODE based system and fonts. Without the knowledge of UNICODE based fonts user cannot type in Unicode.

2.3.5 Spell check:

All the available spellcheckers mainly work with English language.

3. ASCII/UNICODE

ASCII abbreviated from American Standard Code for Information Interchange, is a character encoding standard (the Internet Assigned Numbers Authority (IANA) prefers the name US-ASCII). ASCII codes represent text in computers, telecommunications equipment, and other devices. Most modern character-encoding schemes are based on ASCII, although they support many additional characters ASCII coding system can code only 128 characters [0-127] in ASCII 7bit and 256 characters (0-255) in ASCII 8bit. [12] These are allocated to characters of roman script, special symbols and to alphanumeric characters. No place for other scripts in ASCII .On the other hand the Unicode coding system provide much more range of codes that help to give unique code to various scripts. The Unicode Standard, the latest version of Unicode contains more than 110,000 characters covering 100 scripts. [13]

First version of Unicode (1.0.0) is released on October 1991 that contain total 7,161 of 24 scripts some of the scripts are Arabic, Armenian, Bengali, Bopomofo, Cyrillic, Devanagari, Gujarati, Gurmukhi, Hangul, Hebrew, Hiragana, Kannada, Katakana, Lao, Latin, Malayalam, Oriya, Tamil, Telugu, Thai, Tibetan etc. Version 7.0 is released in June 2014 that contain 113,021 characters of 123 scripts new scripts that are included are Bassa Vah, Caucasian Albanian, Duployan, Elbasan, Grantha, Khojki, Khudawadi, Linear A, Mahajani, Manichaean, Mende Kikakui, Modi, Mro, Nabataean, Old North Arabian, Old Permic, Pahawh Hmong, Palmyrene, Pau Cin Hau, Psalter Pahlavi, Siddham, Tirhuta, Warang Citi, and Dingbats. [14] The latest version was released in June 2016 that contain 128,237 characters and 135 scripts.

3.1 HOW UNICODE WORKS

As in ASCII code each roman character get its unique code so on every computer it will display as the user type it. when user write anything he/she did not worry about choosing which font to write in. User knows that other users will be able to read this article without any problems. This is not happened with regional languages ASCII based fonts. User need to provides font with the information so that other user can read it. But in Unicode each character gets its own individual code. But when user use Unicode font, users would not have to decide which font to use. ASCII uses the limited set of codes to store the character information whereas Unicode gives a uniqe code to every character which it recognises. Thats why ASCII may change its characters when the font is changed.

3.2 ADVANTAGES OF UNICODE

1) Allows for multilingual text in single document without bothering about fonts.
2) Support of Unicode is available on all modern technologies which extend life and scope of application.
3) Full internet support for Unicode system so information written using Unicode based font is easily viewed on internet.
4) Text in any language can be exchanged worldwide.

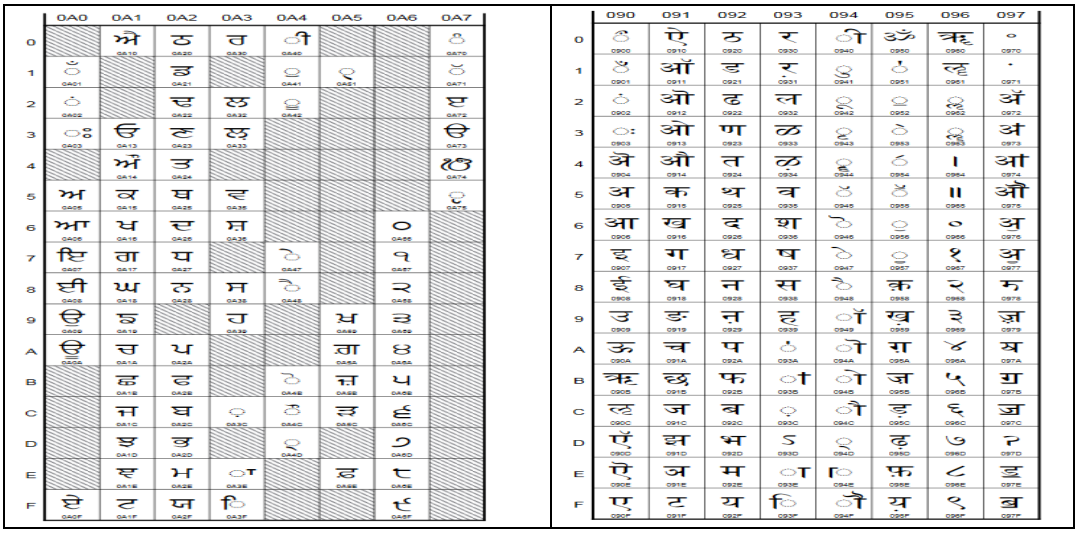

Figure 1: Unicode Character Code to Rendered Glyphs for Gurmukhi and Devnagri Script

3.3 MULTILINGUAL ASCII TO UNICODE SCRIPT CONVERTER:

(www.gurmukhifontconverter.com)

To overcome the problem of ASCII based fonts we create font converter software that can convert many fonts of Indic scripts into other scripts without changing its original meaning and content. Some of the software that are available before it are not very accurate and it convert text without formatting, if a document contains some text in other scripts it converts the whole document to target font. But the font converter that we have created converts the only required code to target code without effecting the text of any other scripts. This software is then converted into an online website (www.gurmukhifontconverter.com) So that everyone can use it. Now more than 30,000 users are using it to convert their text from one font to another. To create this converter we had done manual mapping of around 168 fonts that results around 8,000 page document. This document has helped us to create this font converter. This converter also converts ASCII based fonts to Unicode based fonts.

i. Algorithm:

ii. Input text in the converter.

iii. Identify the script.

iv. By matching words from corpus

v. By matching tries from corpus

vi. Select maximum frequency of word and tri in each language script.

vii. Source script is automatically identified from step 2.

viii. Converter automatically identifies the target Unicode system.

ix. Convert the ASCII characters one by one into Unicode system.

x. Repeat stem 2 to 5 for each word and sentence.

Our developed system that will convert multilingual (Devnagri, Gurmukhi , Malyalam, gujrati, Telgu, Roman) data which automatically identify the script and then system automatically convert it into its UNICODE based script. To identify the script we created a database of around

200 Thousands of words of each script. Further tries are created comprising 400 Thousands tries of each script. "Multilingual Conversation ASCII to Unicode in Indic Script" fulfils following requirements:

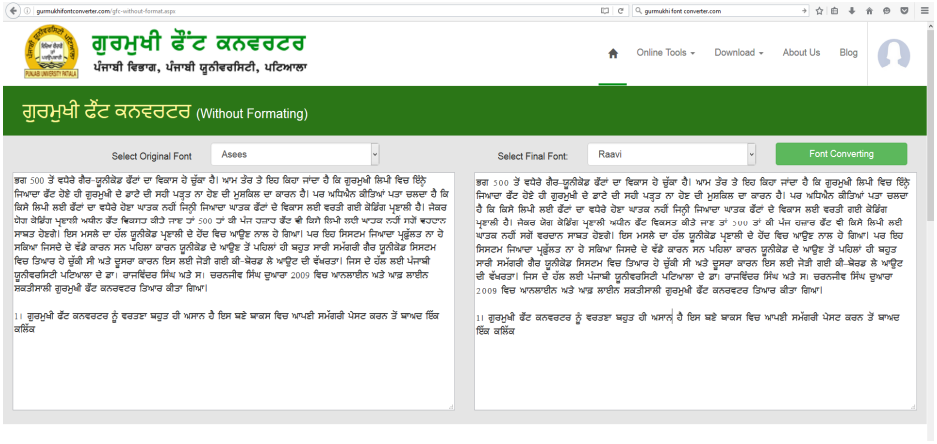

Figure 2. http:/gurmukhifontconverter.com/gfc-without-format.aspx

IDENTIFICATION

Identification is done on the basis of words. We need to identify the language from minimum words which are typed/paste by the user.

Convert To Unicode

The system efficiently converts all the ASCII based scripts in to Unicode system.

Retain Formatting

"Multilingual Conversation ASCII to Unicode in Indic Script" retains the original formatting of the text. It converts the text without changing its formatting.

4. RESULTS

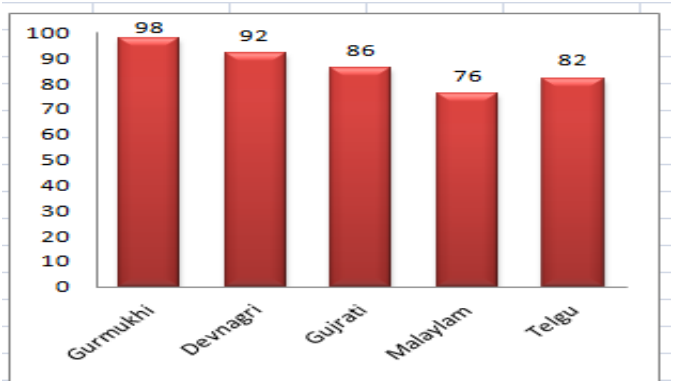

Figure 3. Accuracy of various scripts

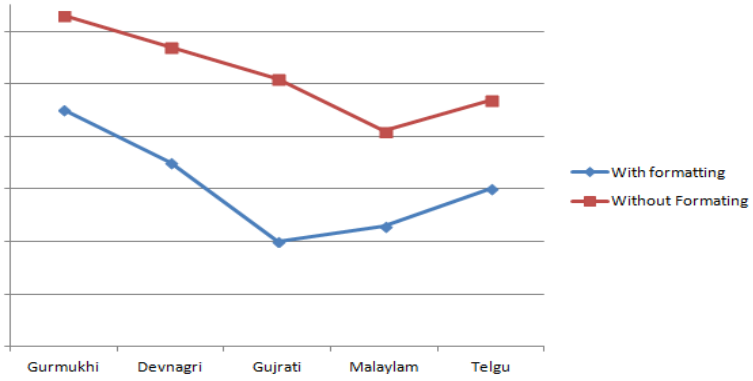

Figure 4. Speed with formatting and without formatting

REFERENCES

[1] UNESCO Bangkok, (2008) Improving the Quality of Mother Tongue-based Literacy and Learning: Case Studies from Asia, Africa and South America, pp7.

[2] Marcia Langton and Zane Marhea, (2003) Traditional Lifestyles and Biodiversity use Regional Report: Australia, Asia and The middle east, PP21

[3] Strasbourg, (1992) EUROPEAN CHARTER FOR REGIONAL OR MINORITY LANGUAGES, ETS 148–Charter for Minority Languages, 5.XI, pp14-15.

[4] Government of India, (2007), The Constitution of India, Govt. of India Ministry of law and justice (As modified up to the 1st December 2007), pp358-360.

[5] Ralph Grishman, (1986) Computational linguistics an introduction, Cambridge University Press, pp24.

[6] John Lyons, (1981) Language and Linguistics an Introduction, Cambridge University Press, pp33

[7] V. Rajaraman, (2012) History of computing in India, 1995-2010, Supercomputer education and research centre Indian institute of science, Bangalore 560012, pp14.

[8] Unesco Education Position Paper, (2003) Education in a multilingual World, Published by the United Nations Education, Scientific and Cultural Organization, pp13-14

[9] Human Development Report, (2004) Cultural Liberty in Today‟s Diverse World, Carfax Publishing, Taylor and Francis Ltd. Customer Services Department, pp77

[10] UNESCO. 1953. The Use of Vernacular Languages in Education. Monographs on Fundamental Education, No. 8. Paris, pp54

[11] Gurpreet Singh LEHAL1 Tejinder Singh SAINI, (2012) An Omni-font Gurmukhi to Shahmukhi Transliteration System, Proceedings of COLING, Mumbai PP314

[12] https://en.wikipedia.org/wiki/ASCII

[13] Ms.M.Kavitha1 , Ms.S.Kawsalya, (2013) Secured Crypto-Stegano Communication through Unicode and Triple DES, International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 2, PP396

[14] The Unicode Consortium, (2015), The Unicode Standard Version 8.0 – Core Specification, Published in Mountain View, CA, pp913.

AUTHORS

Dr. Rajwinder Singh is Assistant Professor in Department of Punjabi, Editor and Coordinator Punjabipedia (www. punjabipedia.org) world famous project in Punjabi University Patiala, where he has been since 2009. He also currently working various projects related to technical development for Punjabi language, literature and culture. He has completed his Ph.D. (Linguistics) at Punjabi University, Patiala (2008) and his undergraduate also complete this University. His research interests Computational Linguistics, NLP, Grammar, Punjabi Linguistics and area of programming languages.

Er. Charanjiv Singh Saroa is Assistant Professor in Computer department of Engraining, Punjabi University Patiala. He is also working co-coordinator Punjabipedia world famous project in Punjabi University Patiala. His area of interest is NLP.