An Analysis of Difficulties in Punjabi Language Automation due to Non-standardization of Fonts

Dharam Veer Sharma

Department of Computer Science, Punjabi University, Patiala

dveer72@hotmail.com

Abstract: *Non-standardization of fonts and some inherent characteristics of Indian scripts make it difficult to use them in automation of systems. This is true for all the Indic scripts because of no direct support of the character sets of Indic scripts in ASCII, which had been the prevalent standard till recent past. Though Unicode solves most of the problems but conversion of existing data from ASCII based font to Unicode is a major challenge in itself. This paper covers properties of Punjabi language, difficulties encountered in font conversion of Punjabi language.*

1. Introduction

Language automation is dependent on uniform support for different fonts of the language. System and applications software are generally developed on the basis of Roman script. Even ASCII code is meant for Roman script and there is no direct support for other language scripts in ASCII code. This makes representation of character set of other languages very difficult. Moreover, ASCII code set proves inadequate for some languages like Oriental where character set may be consist of even more than 5000 characters. Same is the case with Gurmukhi script, used for writing Punjabi language along with Shahmukhi script. For providing support of Gurmukhi script using ASCII codes, various fonts were developed, without following any standards. This resulted in different fonts with different key mapping. Some fonts are phonetic while some are based on Remington. Due to different key mapping

it becomes difficult to convert from one font to another, whenever some data composed using one font is transferred to some other location for use (editing or printing), where the font used is not available, then there is a need either to send the font, used to edit the document, along with the document or to convert it to some font which is available on the target machine.

Unicode was developed to overcome this problem and it has been successful in achieving the goal to some extent. But still Unicode has not been widely adopted for automation purposes and the huge amount of already existing data with different fonts add to the complexity of conversion to Unicode. This paper concentrates on difficulties problems of non-standardization of fonts, problems encountered in font conversion and some rules for conversion from one font to another. The rest of the paper has been organized as follows: section 2 covers overview of some of the Indian languages, section 3 covers a brief description and characteristics of Punjabi language, section 4 covers the difficulties faced in font conversion and paper end with enlisting the references used.

2. A Brief Overview of Indian Languages

There are 15 officially recognized Indian scripts [1]. These scripts are broadly divided into two categories namely Brahmi scripts and Perso-Arabic scripts. The Brahmi scripts consist of Devanagari, Gurumukhi, Gujarati, Oriya, Bengali, Assamese, Telugu, Kannada, Malayalam, and Tamil. And the Perso-Arabic scripts include Urdu, Sindhi and Kashmiri. Devanagari script is used by Hindi, Marathi and Sanskrit languages. The characteristics of the languages within the family are quite peculiar. They have the common phonetic structure, making the common character set. Within the same family again north Indian scripts like, Hindi, Marathi, Gurumukhi, Gujarati, Oriya, Bengali, Assamese have common features while Southern scripts like Tamil, Telugu, Kannada and Malayalam have common features.

All these scripts mentioned above are written in a nonlinear fashion. Unlike English, the width of the characters is different even on a same script. The division between consonant and vowel is applied for all Indian scripts. The vowels getting attached to the consonant are not in one (or horizontal)

directions; they can be placed either on the top or the bottom of consonant. This makes the use of the scripts on computers more complicated to represent them.

3. Brief Description and characteristics of Punjabi Language:

Punjabi is classified as a member of the Indo-Aryan subgroup of the Indo-European family of languages. The Punjabi language is a descendent of the Sauraseni Prakrit, a language of medieval northern India. It is believed to have developed as a distinct language from the Shauraseni Apabhramsha language around the 11th century. Other early influences on Punjabi include Indo-Aryan and pre-Indo-Aryan languages.

India is a country of 122 languages; among these 22 are official languages declared by Government of India. Punjabi language is world's 12th most widely spoken language [2]. Punjabi Language is used in both parts of Punjab, in India and also in Pakistan. Punjabi is syllabic in nature. It consists of 41 consonants called *vianjans*, 9 vowel symbols called *laga* or *matras* and 2 symbols for nasal sounds (. , °) .Modern Punjabi is a very tonal language, making use of various tones to differentiate words that would otherwise be identical. Three primary tones can be identified: high-rising-falling, mid-rising-falling, and low-rising. Following are characteristics of Punjabi language [3] [4].

3.1 Morphological Characteristics

Morphologically, Punjabi is an agglutinative language. That is to say, grammatical information is encoded by way of affixation (largely suffixation), rather than via independent freestanding morphemes. Punjabi nouns inflect for number (singular, plural), gender (masculine, feminine), and declension class (absolute, oblique). The absolute form of a noun is its default or uninflected form. This form is used as the object of the verb, typically when inanimate, as well as in measure or temporal (point of time) constructions. There are seven oblique forms in Punjabi, corresponding more or less to the case forms: agentive/nominative, accusative-dative, instrumental, ablative, genitive, locative, and vocative. All cases except for the vocative are distinguished by means of postpositions. The vocative form takes no

postposition, but may be preceded by a vocative particle or term of address. Punjabi verbs inflect for tense, aspect (perfective, imperfective), mood (indicative, imperative, and subjunctive, conditional), voice (active, passive), person, number, and gender. In this way, Punjabi verbs agree with their subjects, as is the case with other Indic languages. Adjectives inflect for gender and number and thus agree with the nouns they modify. Adverbs do not inflect. With respect to morphology, Punjabi and Gujarati are nearly identical.

3.2 Syntactic Characteristics

General syntactic structure of Punjabi language is Subject, Object and Verb (SOV). Punjabi sentences are mainly simple in structure but complex and compound sentences are also found in literature. Punjabi sentence structure is flexible. Depending on the context or mood of the speaker, it might vary. Punjabi sentences are mostly analytic in structure but the feature of synthesis is still found at dialectal level.

3.3 Technical Characteristics

Each Language has its own script suited to its particular needs and there are certain rules governing its writing and usage. These rules pertain to its script, vowel-signs, correct pronunciation, numerals, spellings and its dialectical variations, punctuation marks, phonemes and their nearest equivalents in other languages, standard terminology, forms of verbs, declensions and other grammatical subtleties.

3.4 Punctuation Marks

Punctuation marks are symbols that indicate the structure and organization of written language, as well as intonation and pauses to be observed when reading aloud. By analyzing a syntactic construction and word formation, punctuation marks delimiting them to the concerned fractions explicate the underlying meaning perfectly. In the ancient Punjabi single full stop bar (.) or double full stop bar (..) had been in vogue only. However, to the present day Punjabi besides Dandi (.) many other marks are used out of which the mark of Dandi (.) has been retained from old Punjabi while all other marks have been derived from English.

4. Difficulties faced in font conversion

Due to non-standardization of keyboard layouts, font conversion is not an easy task to achieve. There are some other additional problems which make the task, of conversion from one font to another, difficult. These problems are discussed below:

4.1 Different keyboard layouts: There is no standardization of Punjabi keyboard layouts. There are more than forty keyboard layouts and more than 500 fonts commonly being used, which means that the same Punjabi word can be internally stored in forty different ways. As for example, the word gi/kph is internally stored in following fonts by using different key map [5] [6].

Table 1: Key map of word gi/kph *in different fonts*

Font Name	Key Map
Akhar	pMjwbI
Amrit-Lipi2	pMj`bI
Anandpur	pµj;bI
Asees	Gzikph
Satluj	ê³ÜÅìÆ
Sukhmani	P^JABI

The thesaurus has to deal with each of these cases separately and read the whole word. Even in the same font, a character can be typed and stored in more than one ways.

4.2 Punjabi language has phonetic nature: One of the unique features of Punjabi, in the variety of modern South Asian Languages, is the presence of pitch contours. These change the meaning of the word depending on the way it sounds. In technical terms these are called 'tones' and these are of three types: low, high and level.

Table 2: Example of one word having different tones

Low Tone			Level Tone			High Tone		
ਘੜੀ	Ghadī	Watch	ਕੜੀ	kadī	link of a chain	ਕੜ੍ਹੀ	kadhī	Turmeric curry

4.3 Complexity in typing: Punjabi typing is much more complex as compared to English typing, as 57 characters have to be typed on the standard QWERTY keyboard. One has to memorize the Punjabi characters corresponding to the English keys and search out each character and then worry whether to type it with SHIFT or without SHIFT.

4.4 Absence of well defined word boundaries: Unlike English, there is no well defined word boundary for Punjabi words written in different Punjabi fonts. As for example, in *Asees* font the following punctuation marks are encoded as Punjabi characters and thus are part of the word (‘ “ + / : ; ? [ ] \ { }). But there are many other fonts such as *Akhar, Satluj* etc. which do not encode the above punctuation marks as Punjabi characters. So the extraction of word boundary is font dependent in case of Punjabi. In English and in many other languages, special characters and delimiters separate one word from another word. But in various different fonts of Punjabi, with the help of these special characters and delimiters some letters to be put in word. This is clear from the following table:

Table 3: Keymap of words consist of delimiters and special characters in different fonts

Word	Font Name	KeyMap
ਉਦਾਸ	Asees	Tüdk;
ਬੇਚੈਨ	Asees	p/u?B
ਪੰਜਾਬੀ	Anandpur Sahib	pµj;bI
ਪੰਜਾਬੀ	Asees	Gzikph
ਪੰਜਾਬੀ	Satluj	ê³ÜÅìÆ
ਪੰਜਾਬੀ	Sukhmani	P^JABI

4.5 Non-linear writing style: Punjabi is not written in linear fashion. The structure of the Gurmukhi script, the script for Punjabi, is non-linear i.e. besides 41 consonants of the language; there are other symbols such as Laga, Lagakhar etc, which are used to represent the phonetic structure of the word. These symbols inherently decorate the consonant. For example, the word 'COMPUTE' as written in English, the character 'O' is called the colleague of 'C', 'M' is called the colleague of 'O' and so on, but in Punjabi it will be written as 'e/gT{N where character e is said to be wearing a cap, g is holding a stick and T is wearing shoes.

4.6 Lack of spelling standardization: There is no standardization of Punjabi spellings. A word may be spelled in more than one way and all the forms may be acceptable. This problem mainly exists because of presence of too many dialects in Punjab. Punjabi language has many different dialects, spoken in different sub-regions of greater Punjab. Different dialects of Punjabi are Majhi, Malwi, Doabi, Pothohari, jhangvi, Multani etc. Residents of one dialect pronounce one word in different manner from residents of another dialect. The problem arises when they write the words as they actually pronounce it. Following are some words which are pronounced in different manner only with one sound change but written in different manner depending upon pronunciation. These words are sometimes called homonyms.

- nkgDk, nkgBk
- fpgohs, ftgohs
- r{Vk, r{VkQ
- jBbk, jBQk
- nByk, nBÿk

4.7 Zero width of some characters: In some of the Punjabi fonts, the Punjabi characters such as *bindi, lava, onkar, dulainkar* etc. have zero width and so if by mistake a user makes multiple entries of such characters only a single

entry is visible. Following table shows some words written by single entry in some fonts but some font require more than one entry to spell same word.

Table 4: Same words make use of different key entry in different fonts

Word	Font Name	Ascii Code
ਵਿ»	Nanak	96;124;121
ਵਿ»	Merapunjab	126
ਉ	e-Panjabi30	85
ਉ	EKTA-DUNIYA	84;91

References

[1] Meenu Bhagat, “Spelling Error Pattern Analysis of Punjabi Typed Text”, M. Tech. thesis Report, Thapar University, Patiala.

[2] Nirma Garg, “Font Independent Spell Checker for Punjabi Using Gurumukhi Script”, M. Tech. thesis report, Department of Computer Science, Punjabi University, Patiala, 2010.

[3] Rupinderdeep Kaur, R.K.Sharma, Suman Preet and Parteek Bhatia, “Punjabi WordNet Relations and Categorization of Synsets”, M. Tech. thesis report, Thapar University, Patiala

[4] Dharam Veer Sharma and Aarti, “Punjabi Language Characteristics and Role of Thesaurus in Natural Language Processing”, International Journal of Computer Science and Information Technologies, vol. 2, pp. 1434-1437, 2011.

[5] G S Lehal, “Design and Implementation of Punjabi Spell Checker”, International Journal of Systemics, Cybernetics and Informatics, pp. 70-75, Jan 2007.

[6] <http://www.learnpunjabi.org/intro1.asp> accessed on 15 May, 2011.