Joint UNESCO and ITU
Global Symposium on Promoting the Multilingual Internet

Multilingualism on the Internet: an Indian Perspective

Dr. Gautam Sengupta

Professor, University of Hyderabad, India

Multilingual India

- Multilingualism is the norm in India.
- It is not uncommon for an Indian to speak one or two languages at home, another on the streets and yet another at school or in business transactions.
- With a population of over one billion, more than four hundred living languages and ten major scripts, India is a veritable fortress of multilingualism.
- Yet it is more important now than it has ever been at any time before to realize that this fortress is under siege.
- Multilingualism, especially in written form involving the use of multiple scripts is in rapid decline in India.

Multilingual India (Contd)

- A quick informal personal survey will reveal the fact that most of our children graduating from urban Indian schools are virtually illiterate in the original script associated with their mother tongue.

Current State of Affairs

- This state of affairs is a result of several factors, including low literacy, inequitable distribution of wealth, the urban-rural divide and globalization without due “recognition of language diversity on an equitable basis” (Pouzin, Abstract).

Current State of Affairs (Contd)

- The decline in the use of the indigenous scripts of India has an obvious and predictable impact on multilingualism on the Internet.
- In what follows I shall briefly discuss some of these issues and move on to some other technical ones pertaining to script encoding.
- Despite having an ostensive public policy on multilingual computing, ineffective interventions by the Govt. of India in technical matters relating to script encoding has resulted in a situation where the multinationals rather than public policy-makers have become the arbiters of the fate of multilingualism on the Indian Internet.

Multilingualism on the Internet

Literacy

- Growth of multilingualism on the Internet presupposes a minimal degree of literacy in the languages under consideration.
- The level of literacy in India is far from what could be expected in a burgeoning economy with nearly 8% annual growth.
- The situation is further complicated by the fact that the socio-economic advantages of being literate in English by far outweigh those of being literate in any of the indigenous languages.
- The state of primary education is abysmal, and quality higher education through a native language is virtually impossible.

Urban-Rural Divide

- Economic development is mostly in high-tech areas and confined to urban India.
- Rural India with its agriculture-based economy is poor, with inadequate educational and healthcare facilities.
- The literate in urban India with its English-medium higher education have enough knowledge of English to be able to access and use English language content on the Internet.
- They have no urgent need to be able to access internet content in an Indian language and are often illiterate in the script associated with their mother tongue.

Urban-Rural Divide (Contd)

- Rural India, on the other hand, is preoccupied with acquiring the basic means of subsistence and can scarcely afford the luxury of web-surfing.

Inequitable Distribution of Wealth

- Due to inequitable distribution of wealth, the benefits of economic growth remain confined to urban centers and play no role whatsoever in alleviating rural poverty.
- As a result, the population that stands to benefit most from multilingual content on the web is hardly able to access the web at all.

Inequitable Distribution of Wealth (Contd)

- It is perhaps necessary to emphasize the fact that contrary to expectations, even in urban centers of higher education the infrastructure for internet access is often far from satisfactory, let alone the facilities available in rural India.
- It is quite unreasonable to expect rapid growth in multilingual internet content under such adverse circumstances.

Technical Issues

- Having briefly looked at the socio-economic factors affecting the growth of multilingual internet content in India, let us now focus on some pertinent technical issues.
- The advent of Unicode has undoubtedly provided a major impetus to the growth of multilingualism on the internet.
- But there are major problems with the way Unicode is being applied to encode the Brahmi-derived scripts of India.
- The Unicode Consortium is dominated by multinational corporations.

Technical Issues (Contd 1)

- The governments of India and Pakistan, despite being “institutional members” almost never send representatives to any of the crucial meetings at which major issues relating to script encoding are discussed and decided.
- This has encouraged some multinational corporations that are full members of the consortium to assume the role of arbiters on matters relating to encoding of Indian scripts, notwithstanding their utter lack of expertise and domain knowledge.

Technical Issues (Contd 2)

- Let me substantiate the point with a few concrete examples.
- **The Devanagari Letters Short A and Candra A**
  - The code page for Devanagari contains a character named “DEVANAGARI LETTER SHORT A” at 0904.
  - I am not aware of the existence of any such character in Devanagari or any other Indic alphabet.
  - The whereabouts of this mysterious character is known only to the members of the Unicode Technical Committee (overwhelmingly dominated by representatives of multinational corporations) that approved it.
  - But, thanks to them, we are now stuck with a letter that we shall never use.

Technical Issues (Contd 3)

- **The Devanagari Letters Short A and Candra A (Contd)**
  - On the same code page we find a character at code point 090D named “DEVANAGARI LETTER CANDRA E”.
  - It is composed of the half-moon shaped glyph for CANDRA mounted on the glyph for DEVANAGARI LETTER E, and is used in the script to represent the “a” sound in “cat”.
  - Marathi users of Devanagari use this, as well as a CANDRA mounted on A – mostly the latter – to denote this sound.
  - Yet, Unicode recognizes no “DEVANAGARI LETTER CANDRA A”.
  - Evidently, the cause of multilingual computing would be served much better if the consortium worked in consultation with linguists who had the required domain knowledge.

Technical Issues (Contd 4)

- **Bangla Khanda-Ta and Malayalam Chillaksarams**
  - The Unicode Standard 4.1 introduced a new character named “BENGALI LETTER KHANDA TA” at code point 09CE.
  - Khanda Ta has always been recognized as a distinct letter of the Bangla alphabet.
  - Yet users of the Bangla script and Bangla-speaking linguists had to engage in a mindless conversation that did not always remain cordial, over a period of more than a year, to convince the UTC to assign a distinct code point to Khanda Ta.
  - Now the same process is being repeated for the Malayalam Chillaksarams, which have as much, if not more, reason to be distinctly encoded as the Bangla Khanda Ta.

Technical Issues (Contdd 5)

- **Bangla Khanda-Ta and Malayalam Chillaksarams**
  - Whether these characters get distinctly encoded will depend on the tenacity of the Malayalee community and the degree to which they are able and willing to politicize the matter, rather than the persuasiveness of their arguments.
  - This is not a happy state of affairs and it is certainly not helping the cause of multilingualism on the internet.

Technical Issues (Contd 6)

- **The Use of ZWJ in South-Asian Scripts**
  - The Unicode Standard is concerned with encoding characters not glyphs or display variants.
  - This is an avowed policy of the Unicode Standard.
  - Yet it made an exception in this regard in the case of Devanagari, apparently for the sake of preserving backward compatibility with ISCII.
  - Variant forms of consonantal conjuncts in Devanagari were assigned Unicode encodings, one with and the other without a ZWJ (U+200D):
  - क्ष U+0915 U+094D U+0937
  - क्‍ष U+0915 U+094D U+200D U+0937

Technical Issues (Contd 7)

The Use of ZWJ in South-Asian Scripts (Contd 1)

- It should be noted that the examples cited above are display variants of the same consonantal conjunct in Devanagari, and that their Unicode representations differ only in terms of the presence or absence of a ZWJ (U+200D).
- Using the ZWJ in this manner is in itself not a bad thing since it was originally conceived of as a control character signaling ligation.
- However, encoding display variants was a major deviation from an avowed policy explicitly stated in the standard.
- Having done so, for the sake of backward compatibility, one would assume that the UTC would refrain from further deviations from explicitly stated norms.

Technical Issues (Contd 8)

The Use of ZWJ in South-Asian Scripts (Contd 2)

- However, as if to add insult to injury, the UTC decided to use the ZWJ to encode semantic distinctions as well.
- This was done to encode the Bangla Khanda Ta and the Marathi Eyelash Ra.
- In the latter case, it amounted to using the ZWJ to encode a semantic distinction in a script – namely Devanagari – in which it had already been used to encode display variants of consonantal conjuncts.
- Consider now the plight of someone trying to design a spell-checker for Marathi.

Technical Issues (Contd 9)

The Use of ZWJ in South-Asian Scripts (Contd 3)

- If he programs his spell-checker to ignore ZWJs then it would fail to distinguish between two semantically distinct Marathi words
- दर्या U+0926 U+0930 U+094D U+092F U+093E
- and
- दर्‍या U+0926 U+0930 U+094D U+200D U+092F U+093E

Technical Issues (Contd 10)

The Use of ZWJ in South-Asian Scripts (Contd 4)

- If, on the other hand, he programs his spell-checker NOT to ignore ZWJs then every conjunct in his dictionary would have to be encoded twice to ensure that words like
- रक्षा U+0930 U+0915 U+094D U+0937 U+093E
- and
- रक्‍षा U+0930 U+0915 U+094D U+200D U+0937 U+093E
- (which are display variants of the same word) are both recognized as legitimate words.
- Neither of the two situations is acceptable.

Technical Issues (Contd 11)

The Use of ZWJ in South-Asian Scripts (Contd 5)

- Clearly, the use of ZWJ to encode semantic distinctions has to be completely prohibited.
- However, this would imply that the software vendors that had already started to produce rendering engines for Indian scripts would have to revise a lot of their code, and that would cost them a substantial amount of money.
- So, once again, instead of reversing a faulty decision, the UTC decided to legitimize it by incorporating it into the standard.
- Peter Constable of Microsoft Corporation produced a Public Review Document (PR-37) that proposed a systematic use of ZWJ to encode semantic distinctions in South Asian scripts, and the UTC accepted it despite stiff resistance from the user community.

Conclusion

- Contrary to appearances, technological hurdles are often not the real impediments to progress.
- The real hurdles are socio-economic and political.
- The real impediments to the rapid growth of multilingual internet content in Indian languages and scripts are low levels of literacy, the rural-urban divide, inequitable distribution of wealth leading to poverty, the self-appointed multinational custodians of our multilingual heritage and the failure on the part of our successive governments to keep them at bay.