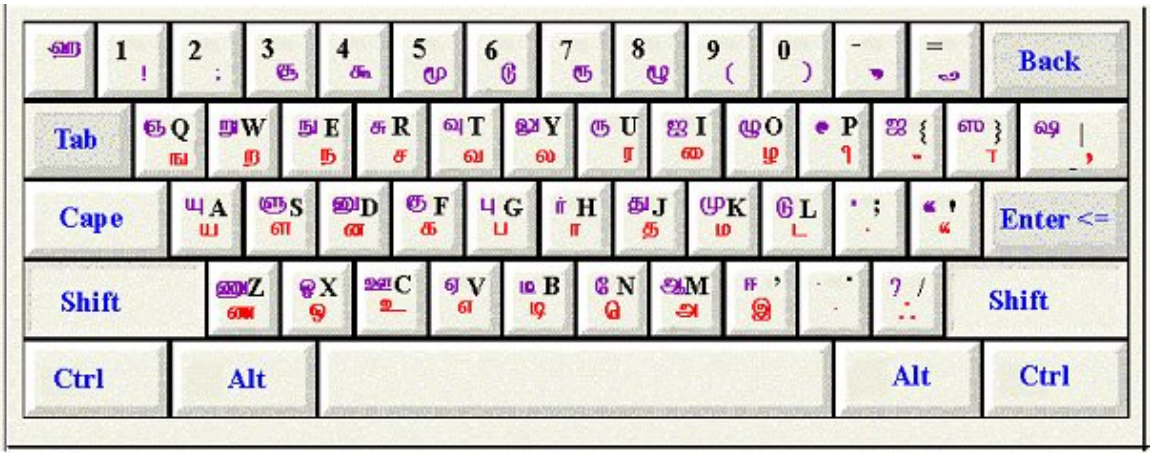
L2/00-405

Visual Order in Indic Languages

(and other Indic issues blocking adoption of Unicode)

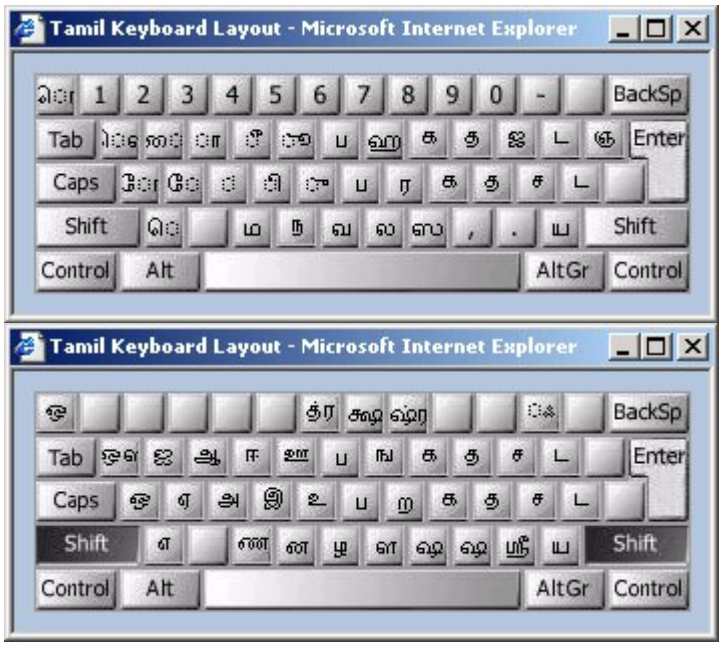
Note: I am centering the bulk of my discussion here on Tamil, but the same issue does apply to other Indic languages, including Hindi.

A wide variety of Tamil font faces and specialized word-processors are currently being used for Tamil word-processing. By having different font-encoding schemes (non-charitably named "font hacks" -- referring to the use of fonts that map standard latin letters on an English keyboard to specific glyphs), and different input/output tools, many practical difficulties of Tamil composing exists. Huge number of Tamil websites are being created everyday using Tami, (over 15 weekly Tamil newspapers being published in Toronto alone -- Tamil population ~250,000), standardization becomes crucial. Let me present the typical keyboard layouts used for the Tamil font face:



This is a typical one used in Canada (by the majority of the aforementioned 250,000 people!). Another example, showing a keyboard more commonly used in Sri Lanka:

There are some obvious similarities, but the most glaring of them relates to the fact that they are both based on **VISUAL** ordering of letters. Contrast this with the Microsoft Tamil keyboard layouts:



In Unicode, when I am typing a word such as கோவில் (KOVIL, which means Temple), I would type க ோ வ ி ல ் ("iabfnd" on the keyboard), whereas with these visual layouts, I would type the equivalent keys (using the first keyboard) of "Nfhtpy;". This is not a good example to show actual advantage to these keyboards; that is a fact that comes more into play when there are cases where both Unicode and ISCII are based on particular combinations of code points, where the language itself is not taught in these terms.

In fact, the language is taught in terms of 247 characters, divided into Uyir (vowels), Mei (consonants), and Uyirmei (syllables). Although each Uyrimei can be split into a vowel and consonant, they are treated as separate letter in terms of learning and using the language. In fact, Kani Thamizh Sangham (the Tamil Computing Society, a professional organization I belong to) is working to make recommendations to the government of India regarding a Tamil syllabic block, similar to the Hangul syllable block and that of the Canadian Aboriginal Syllabics already present in Unicode.

It is quite simply fact that Windows 2000 (the first OS to support Unicode Tamil in font, keyboard, etc.) has not garnered nearly as much interest as any of the other encoding attempts. What they want, more than anything: they want an encoding not tied to ISCII. ISCII is not very widely used even in Tamilnadu, and it is almost completely unused outside of India, largely due to the perception that their language is being "Devanagrized" by it (a term used by several translators). Many feel the same way about Unicode, as it is largely based on ISCII.

Now, I am not specifically advocating that any of these specific schemes need to be adopted, as I think it even if it is not a step backwards, it would be a step sideways. But in this current world Unicode is very seldom used (to date I have been in contact with 193 localizers/translators in Tamilnadu, Sri Lanka, Malaysia, Singapore, and Canada, and only one supports Unicode -- that one only does so because I wrote a parser that would take visual layouts and converted them to Unicode!). Currently, the following (entirely separate!) projects are underway or developed:

- TSCII, a recently developed 8-bit encoding standard for Tamil that is designed to support Tamil and English, is available at http://www.tamil.net/tscii/ and was adopted by STC (Standards for Tamil Computing).
- TAM, a monolingual 8-bit encoding scheme.

- TAB, a bilingual 8-bit encoding was proposed by the Tamilnadu government, and could be thought of as a competing standard with TSCII.
- The Anjal encoding, widely used originally in Singapore and Malaysia, is gaining greater acceptance, in part due to the fact that Murasu's Anjal2000 and similar programs provide tools for converting between TSCII, TAB, Anjal, Unicode, and other common "font hack" encodings.

Although Unicode currently has the strength of being able to consider itself the "volume platform" in regards to an encoding standard, the fact that it is not being widely accepted by many languages/scripts such as Tamil. This is largely due to the fact that most of these scripts are using entirely different means that are either based on:

- the "font hack" principles used by keyboard layouts similar to those above, or
- encoding standards such as TSCII or TAB, aimed at providing support for Tamil along the same lines as ISO-8859* or Microsoft's 125x code pages.
- The syllabic approach being developed (currently used in schools quite a bit, and fostered by font hacks to help people learn with the aid of computers when available)

Mapping between these standards and Unicode or ISCII is very problematic (a problem shared with other Dravidian scripts), and since they see their needs met in their current solution, their desire to move to Unicode is thus seriously impaired (if not simply killed outright). Unfortunately, many of the people who have been dismissing ISCII for over a decade find Unicode easy to dismiss since it is based on the same principles as ISCII.

Separately, ISCII and Unicode are also under fire from many linguists who feel it does not properly encode Indic lanuages (as the paper by S.P. Mudur highlights, that paper will also be made available when this one is).

Conclusion

Nothing to vote on at this point, really. Sorry. I just want the concepts out there, and an action item to work on the changes that would need to be made to the Tamil block, in order to garner the support of the people who would most want to use it. They clearly see Unicode as a possible future, but currently it does not suit their needs so it is thought of as "the road not taken."

The same types of issues exist for many other Indic languages, and similar actions should likely be undertaken to get people involved with explaining how to have their languages/scripts best represented by Unicode.

November 7, 2000 -- Michael Kaplan