International Conference and seminar on the use of Tamil in Information Technology (TamilNet 99)

We cordially invite you to the Inaugural and Valedictory functions of the International Conference and Seminar on the use of Tamil in Information Technology (TamilNet 99) to be held on 7-2-1999 and 8-8-2-1999 at IMAGE Auditorium, MRC Nagar, Raja Annamalaipuram, Chennai - 600 028.

PROGRAMME FOR INAUGURAL FUNCTION

Date: 7-2-1999 Time: 9.30 a.m.
Venue : IMAGE Auditorium, M.R.C. Nagar, Raja Annamalaipuram, Chennai - 28.

Invocation

Welcome Address :
Thiru MURASOLI MARAN, M.P..
Chairman of the Reception Committee of TamilNet 99

Presidential Address :
Justice Selvi M. FATHIMA BEEVI
Her Excellency the Governor of Tamil Nadu

Inaugural Address :
Prof. K.ANBAZHAGAN
Hon'ble Minister for Education, Govt. of Tamil Nadu

Special Address :
Dato' Seri S. SAMY VELLU
Hon'ble Minister of Works, Malaysia

Thiru S. THONDAMAN
Hon'ble Minister of Livestock Development & Estate Infrastructure, Sri Lanka

Thiru R. V. JANAKIRAMAN
Hon'ble Chief Minister of Pondicherry

Kalaignar M. KARUNANIDHI
Hon'ble Chief Minister of Tamil Nadu

National Anthem

11 - 11.30 a.m. Tea Break
12 noon Technical sessions (For Delegates only)

PROGRAMME FOR VALEDICTORY FUNCTION

Date: 8-2-1999 Time: 5.00 p.m.

Venue : IMAGE Auditorium,M.R.C. Nagar, Raja Annamalaipuram, Chennai - 28.

Invocation
Presentation of the recommendations of the Conference & Seminar :

Dr. M.ANANDHAKRISHNAN
Chairman of the Sub -committee on Tamil Computing of the State Task Force on Information Technology

Felicitation :

Dato Seri S.SAMY VELLU
Hon'ble Minister of Works, Malaysia

Thiru S. THONDAMAN
Hon'ble Minister of Livestock Development & Estate Infrastructure, Sri Lanka

Thiru R.V. JANAKIRAMAN
Hon'ble Chief Minister of Pondicherry

Dr. SAM PITRODA
Chairman and CEO, WorldTel

Veledictory Address :
Kalaignar M.KARUNANIDHI
Hon'ble Cheif Minister of Tamil Nadu

Vote of Thanks :
Dr. THAMIZHKUDIMAGAN
Hon'ble Minister of Tamil Official Language / Tamil Culture & Hindu Religious and Charitable Endowment Department.

National Anthem

Papers presented at the Tamilnet 99 Conference

1.Towards a Total internet solutions for Tamil by Naa. Govindasamy

2.Glyph Based Font Encoding Encoding Scheme, TSCII by Dr.K. Kalyanasundaram and M.Nedumaran

3. »¾ÕÃÓ ÈÀÔ²«¶Õ¡Þ Ù´×´Ôà¡Þ£ »´Â§ ×»ÔÃÕ§ ã¥½£ - S.Srinivasan

4. iDNS, a DNS System with Multi lingual support by Janes Seng and Tan Tin Wee

5. Tamil Font Encoding Standards by P. Chellappan

6.Computing with Tamil Use of the Vernacular in Design User Interfaces - by Kalyanakrishnan

7. »¾ÕÃÓ ×¼àõ´º¡´Õ§ ä»Õ¿Æ äÞ»§ by C.S. Senthilnathan

8.Towards defining the design goal while placing Tamil in Unicode by Chandrabose

9. »¾ÕÃÓ ´ºÕ¤×½ÔåÕ ÂÕÙ¶¤½ÁÙ´ É¹ ÇÙ¾é´ÄÕ¨ Ñ¯ Ð¤½Ô²é- å. ×½Ô¨ÆÙÂ¡Ø´Ô

10. Multiple Encoding Systems for different Computer Applications
by M. Ganesan

11. Tamil Standard coding Information Interchange by N. Anbarasan

12. A Tamil Speech Synthesis System by -A.G. Ramakrishnan & V. Karthigeyan

13. Tamil in Cyberspace by Ramalingam Shanmugalingam

14. ´ºÕ×¾ÔÃÕ ¶Õ- - å.×½Ô¨ÆÙÂ¡Ø´Ô

Governor Fatima Bevi Talk

I have great pleasure in Participating in the Inaugural Session of the International Conference and Seminar on the use of Tamil in Information Technology, hosted by the state level task force of the Government of Tamil Nadu on Information Technology.

I welcome the Ministers and high dignitaries from abroad and the International experts and delegate to this conference, which will set on course the process of making tamil an integral part of information superhighway.

Information technology can be used as a vehicle for all-round development in creation of new jobs and in making assess to information amazingly easy. Today's Internet links about 170 countries in the world, links people who speak many different languages on a 24 hour basis. Internet has come a long way since the today it first went online in 1969, initially conceived in huge mainframes and later survived in the era of personal computers. After the invention of world wide web, internet has become user-friendly providing access to information without any barriers between the users and no occupies the centre stage in the modern world. The world wide web consisting of documents of a staggering volume and the easy access to the entire world of data, through the internet, are among the finest contributions, after telephone, radio and television, by science to global communication and human interaction. Earlier in the 80s, the internet was considered as a people's network meant for variety of academic purposes. In the 90s, the usage of network was opened to individuals facilitating thousands to join the Internet everyday.

Ture to the scientific tradition of the state, Tamil Nadu has taken progressive strides in information Technology. The state Government has established a separate department for information technology and is very enthusiastic in making optimum use of the immense human capital of the state and its long tradition of academic excellence to attain serious growth in information technology. I congratulate the Government of Tamil Nadu on setting the pace for this growth in information Industry. The TamilNet 1999 is another laudable endeavour in this line. The Government is willing to take advantage of the competitiveness of the private sector in information industry and is happy to acknowledge the contribution being made by individuals at private firms to the development of software, especially in Tamil Language and the creation of websites carrying information on Tamil Language and literature. But all these products lack in uniform standards and global organization. In addition to the millions in tamil nadu, there are lakhs of tamil people living in different parts of the world. Their love for their mother tongue and zeal for its preservation and promotion are extraordinary. Since internet has become a daily after in individual life the world over, it i absolutely necessary to have a tamilnet with uniform codes which will be able to harmonise promotional efforts from various quarters in future. The Government of Tamil Nadu can perform this hard task and remain the central guide fro tamil network usage in future. I wish to place on record my deep appreciation of the number of individuals who have made serious and notable efforts to introduce tamil language on an international scale, by producing several tamil softwares and creating websites. With their

sincere and enthusiastic cooperation, i am sure that tamilnet will soon attains a place of prominence in the information superhighway. Already the use of Tamil
in Internet is far greater than any other indian languages. With the implementation of uniform coding standards, tamilnet will be a huge success in both emerging as a prominent network and in involving a broad community of tamil internauts in working together and accomplishing the dissemination of information on tamil and Tamil Nadu. The most significant achievement will be the ability of tamilnet to bring the large rural population of tamil nadu - the non - english speaking and non - academics - in the fold of internet. This is the progressive step which will place tamil on the information superhighway thereby rendering access to international scholars in tamil and is conducive o intensive development in tamil language. The credit for this is solely due to the hon'ble chief minister of tamil nadu who has given the needed fillip for modernising both teaching and learning of tamil language across the world. I congratulate thiru. Maran for his perseverance and determination to modernise many dimensions in the state administration for the larger benefit of the people.

I am delighted to extend my warm greetings and best wishes to all the participants in this conference for success in future endeavours. I wish te "TamilNET '99" a grand success. Jai Hind.

Towards a Total Internet Solution for the Tamil Language through Singapore Research

by Naa Govindasamy
Lecturer, National Institute of Education (NIE), Nanyang Technological University (NTU)469, Bukit Timah Road, Singapore 259756

(This Paper was presented at the SAARC Conference on Extending the Use of Multilingual & Multimedia Information Technology at Pune, India, on September 14, 1998. The present paper has added new developments which took place after September 1998. The author wish to thank Dr Tan Tin Wee, Associate Director for the Centre for Internet Research (CIR), National University of Singapore and Mr Leong Kok Yong, Research Officer of CIR, for giving valuable advice in the preparation of this paper.)

Abstract

The internet revolution has enabled the widespread dissemination of information throughout the world. Most of the content is in Romanized characters. Research is going on in some countries to enable non-Roman scripts accessible on the Internet. This paper will discuss and demonstrate how, through a successful research collaboration in Singapore, Tamil language content is now freely accessible, searchable, conveniently emailable and easily composed and edited on the Internet through all three popular platforms Unix, PC and Mac. The important part of this paper is latest development, which took place in Singapore after September 1998.

Multilingual TextEditor For Unix, Windows and Mac.
Text is saved in Unicode and Tamilnet code. This is the first Unicode Tamil Editor to be seen on the Internet for free downloading.
Multilingual multiscript URL is another important development in Singapore Research.
My presentation will focus on these two important development.

The Need for A Multilingual Internet in Singapore

Singapore is a multilingual and a multiracial country. English, Chinese, Malay and Tamil are the official languages. Most of the government and public documents are published in these four official languages. However, until recently, it was not possible for the Chinese and Tamil languages to be disseminated through the World Wide Web on the Internet.

In 1994 Dr Tan Tin Wee, my research collaborator initiated work in this area while he was the head of Technet Unit, the first Internet service provider for Research and Educational Institutions in Singapore. Technet Unit was directly under the supervision of the Computer Centre of the National University of Singapore. (Technet Unit has since been commercialised to become Pacific Internet, one of the three ISPs for Singapore. The others are Singnet and Cyberway.) By mid-1994, Technet Unit initiated the Singapore INFOMAP project which provided

a one-stop WWW home page for Singapore. He wanted the four official languages to be represented in the INFOMAP.

Since English and Malay are using the Roman script, displaying these two languages on the WEB was not a problem. By the end of 1994, Technet had successfully implemented an Experimental Chinese WEB server in Singapore. So the problem of displaying the Chinese script on the Web was solved. However, displaying Tamil script on the Web, and communicating through Tamil on Internet was a problem. There was no Tamil Information System on the Internet which provides a display system in Tamil and English simultaneously on the Text Mode using a Tamil-English single font file. There were a few servers, which were providing Tamil script using GIF image files.

Tamil Eelam Page (http://www.eelam.com) was and is still very active in this direction. Tamil Nadu Home Page, and Tamil Electronic Library .
(http://www.geocities.com/Athens/5180/index.html) are other popular Tamil Web Sites on Internet at that time. Tamil Electronic Library was using (and is still using) a mono 7bit font (Mylai) for the Tamil display on the Web.However, Mylai font cannot support native emailing at that time.

So there was a need to develop a Tamil Internet System which should go beyond Web display. In May 1995, I met Dr Tan at the Technet Unit, National University of Singapore, soon to become the Internet Research and Development Unit (IRDU) (now upgraded to Centre for Internet Research). We identified the potential solutions and agreed for a possible research collaboration between NUS and my institution, NIE, NTU, the two instititutions of higher learning in Singapore at that time.

At that time, on my own, I was in the process of developing True Type fonts and a Tamil software for Windows Applications (KanianÀ)À Tamil Software). As the service provider arm of the Technet was sold to a private consortium and renamed Pacific Internet in September 1995, the nascentTamil Internet Research was inherited by the newly formed Internet Research and Development Unit (IRDU). This Research Unit was funded by the National Science and Technology Board (NSTB) (http://www.nstb.gov.sg) of Singapore. Mr Leong Kok Yong, just graduated from the Nanyang Technological University, joined IRDU, and became one of the key member for the TamilWEB project.

Objective

During the Technet period, when Dr Tan and I, conceptulised the TamilWEB Project. We had a very clear objective.That is: to develop a bilingual font system for The Total Internet Communication in the Tamil language. That means, the system that we intended to develop: should provide display of Tamil & English Text simultaneously on the Internet Applications. (Web Browsers & Email sofware packages.)

Tamil and English text should be easily communicated and retrieved in Plain Text.
should work across Platforms. (PC, Mac & Unix) should be searchable in Tamil
should let the user type Tamil in the Web browser's Forms, and the typed word should be seen in Tamil.
Should allow the user to read Tamil & English in terminal emulation mode (telnet).

Prototype Testing, Preview and Official launch

The prototype of our system was tested during the launch of PoemWEB. PoemWeb is an electronic selection of representative poems from the four official languages, from the book , Journeys: Words, Home and Nation - Anthology of Singapore Poetry (1984-1995) which was published by The Centre for the Arts, National University of Singapore. This book was launched by H.E. Mr Ong Teng Cheong, President, Republic of Singapore on Friday 27 Oct 1995.

The preview of the first phase of TamilWEB project was shown to the public and to the Press at the Internet for Everyone 1995 at Suntec City Exibition Hall during 13 December 1995. TamilWeb was officially launched by the Honourable Member of Parliament, Dr Ong Chit Chung (Chairman, GPC for Education and MP for Bukit Batok) on 2 February 1996.

Since then the Tamil language teachers in Singapore and the Internet Users from locally and abroad, are using the Singapore System to communicate in Tamil & English over Internet and have created a significant volume of bilingual Web pages in Tamil and English. One of the most important Web site for teaching and learning of Tamil language was created by University of Pennsylvania's Penn Language Center (PLC).

Using our system, the Center created a bilingual Website for Learning and Teaching Tamil in 1996. The project is funded partially by a grant from the Consortium for Language Teaching and Learning, with the joint participation of Tamil-teaching faculty at the Universities of Chicago, Cornell, and Pennsylvania. The ruling party of Tamil Nadu Dravida Munneetra Kazhagam (DMK) Website is another important site, using our system. (http://www.thedmk.org)

Purpose of this paper

This paper will try to explain and demonstrate, how Tamil, one of the Indian languages, has achieved the Total Internet Solution, through Singapore Research. Most of my presentation will be done through Internet. When Dr Tan and I conceptualised the TamilWEB project in 1994, there were only two major graphical Web browsers on the Internet, Mosaic and Netscape2. Eudora was the one of the most preferred graphical Email program. PINE was the most preferred, Text-mode Unix Emailer, and LYNX was the Textual Browser on the Unix platform. Font Tagging was not available at that time. Content on the Web was almost exclusively in Romanized characters. English language was dominating the Web. Content in non-Roman languages was limited. With these background, we delivered these Internet Tools for our Tamil users.

Tamil Internet tools from Singapore

These Internet Tools originating from our research and software development are free for downloading:

TamilNet.ttf (PC propotional font)
TamilFix.ttf (PC fix width font)
Tamilnet.hqx (Mac propotional font)
Tamilfix.hqx (Mac fix width font)
Tamilnet18.bdf (Unix font)
Tamilfix.bdf (Unix Font)
Tamil Keyboard Manager (for PC)
Tamil Keyboard Manager (for Mac)
Xkeymap (Tamil input system for UNIX)
Mirage (CGI Application Software for rendering Multilingual encoding text into GIF images for display on web browser
Applet input sysytem for Tamil word search

Font encoding

The key tool for the project was the creation and design of a Bilingual Tamil-English single font system. We have designed a bilingual font set for the display of both Tamil and English simultaneously. This was done by making use of the upper extended ASCII character range for the Tamil characters, while retaining the basic English alphabet and punctuation intact in the lower ASCII range. This will allow most of the Web world in English (or other Romanized languages) to be traversed; at the same time, Tamil codes will be recognized and displayed correctly when they occur, without having to change font set. Figure below shows the character map for the Tamil-English font set.

One important point to note is that the upper ASCII portion does not have enough code space to include all the possible Tamil character glyphs (>>200). As such, we made use of the kerning feature built into the Postscript and the True-Type font technology to combine two Tamil characters into a new character glyph not found in the above ASCII table.

With the combination of two simpler character glyphs to give a more complex glyph, we can then include the entire Tamil character set within one single font, together with the English alphabet. To allow users to input these Tamil characters, a corresponding keyboard layout mapping has been devised by mapping the keys on a normal English (QWERTY) keyboard to the extended ASCII range where these Tamil characters reside. A toggle key enables the user to switch between the two modes. Tamilnet propotional font was developed to display Tamil & English on the Web browser. However, the variable proportional font cannot be viewed in the Web browser's Forms.

For this purpose, a fixed width font, Tamilfix was developed. This font is very similar to the courier font. The Tamilfix font is simply doing the work of the Courier font. Only with this fixed width font, Tamil can be typed in the Web browser Forms. In the Web browser, the form filling feature is a very important component for interactivity. If the user wish to communicate in Tamil to webmaster or the author of the webpage, he or she has to type Tamil into the Form.

Keying in the Tamil characters in the Forms

When the user is keying in Tamil script in the Form, the Tamil Characters should appear on the Forms as they are typed for immediate visual feedback. Only then can meaningful communication and interactivity take place. We achieved this through the Tamilfix font and the keyboard input system. Another important factor in any database creation is the Search function. If a Search function is not possible in a particular system, creating a database, is out of the question. When we launched the TamilWEB on 2.2.1996, we demonstrated the Search function.

In the search form, Tamil words were keyed in for searching against a database of Tamil text. For search and retrieval, the submitted string in extended ASCII for Tamil (and in English as well for bilingual searches) is parsed by the httpd server and submitted as a search string to any indexing engine that has multilingual capability. In the case of Tamil, we used a simple WAIS-SF indexer and demonstrated the utility.

Hits were returned in the same encoding, and displayed in the same way as described above, with bilingual capability. In fact, this powerful search function is taking place across the various platforms. In the Singapore Government Web site (http://www.gov.sg), searching for Tamil keywords by typing Tamil script is possible. The I Agent search engine will deliver the results in the form of webpages, using Singapore Tamil font encoding.

However, in some cases, users are unable to use our fonts and encoding system for unknown reasons. In this situation, we have invented another solution. Our research team has produced a CGI Application software for rendering multilingual encoding text directly into images for display on any web browser as embedded images. It is called Mirage.

When this application is added on to the server, the server is capable of rendering Unicode Tamiltext into images for the client browsers, without any helper application or any font installation. The significance of this system is that, in the client browser, the user should be able to view multilingual information, originally coded using Unicode. For that matter, any encoding can also be transformed into images using the MIRAGE system, eg. Unicode, ISCII, Kanian-Tamilnet etc. simply by modifying the code table mappings to character glyphs. Now, I will be demonstrating another of the important feature of our system.

Viewing Tamil on PC Terminal Emulation

When we developed the Tamilfix font, we knew that it will make Tamil readable in the PC Terminal emulation (eg Telnet). A Shell access user can read Tamil text in the WWW textual browser LYNX. He can also read Tamil in the Terminal Email software PINE. This is a very important development for the Tamil language. In many developing countries, the number of SHELL access users typically outnumbers the TCP account users. Most users access the internet through a character-based terminal emulation rather than a graphical user interface. As such, our system benefits a lot of SHELL access Internet users. This was made possible with our Tamilfix font.

Keyboard Input Systems

Based on the Phonetic system, a phonetic keyboard for the Hindi language was developed by Mohan Thambi in India in 1983. This was subsequently adopted as the main keyboard for the Indian languages by the Department of Electronics (DOE) of India. However, this keyboard is based on Devanagiri script. Since Tamil is from the Dravidian language group rather than the Indo-Aryan, of which numerous other Indian languages belong (e.g., Hindi, Marathi, Punjabi, Bengali) (Grimes, 1992), the DOE keyboard is not particularly suitable for the Tamil language.

To overcome the keyboard problem for the Tamil language, the author began an investigation into the frequency of occurrence of the Tamil vowels and consonants used within the language. Based on this research, a Tamil phonetic keyboard layout was introduced for Tamil computing (Govindasamy, 1989), named the Singapore Tamil Keyboard (Govindasamy, 1994a). Later in September 1994, the name was changed to Kanian Keyboard at the First Computer-Tamil Conference at the Anna University, Madras, India (Govindasamy, 1994b).

The keyboard consists of the 12 basic Tamil vowels placed on the left-hand side of the keyboard and the 18 consonants. The 28 basic vowels and consonants are placed in the lower case of the keyboard, while the 2 least frequently occurring Tamil consonants are placed at the upper case with the 5 Sanskrit sound consonants. For modern Tamil, a vowel will not appear in the middle or at the end of a Tamil word; it will appear only at the beginning of a word. These basic rules were taken into account when this keyboard layout was designed.

The advantage of this keyboard layout is that 99.5 percent of the time, Tamil characters can be typed without pressing the shift key at all. Moreover, the most frequently used vowels and the consonants are placed at the home keys (the middle row of the keyboard). This allows the user to type 68 percent of the Tamil words by using only home keys. Because of its simplicity and the incorporation of the Tamil grammar, this keyboard layout is very popular in Singapore and Malaysia and has been incorporated in numerous Tamil front-end processing software and word processors, including a commercial version available from the author (Govindasamy, Kanian Bilingual Wordproceesor for PC, Mac & Internet).

The project team has devised a Tamil type writer version for PC & Mac, using KeyMan Keyboard Manager. From our server we are using KeyMan keyboard Manager. For Unix Xkeymap was used to develop the Keyboard Manager for Tamil. These three input devises are downloadable from our Website.

Toward Java keyboard input systems for total cross platform compatibility using Unicode.

Java Input Method Engine (JIME) for Java from CIR, bundled native input methods and character display support in a set of applets. At present the users can input Chinese, Japanese or Korean text in HTML form irregardless of the locale of the user platform. The Tamilweb project team has developed the Multilingual JIMEWORD. Text editor. JIMEWord - JIMEWord is a multilingual WYSIWYG text editor.

It supports Chinese, Japanese, Korean, French, German, Thai, Cyrillic, Greek and Tamil. It's implemented in Java 1.1 and run on all JDK 1.1 compliant platform. It runs on Unix, Windows95/NT and Mac. Able to edit, load and save text in Unicode UTF-8 and UTF-7 encoding as well as other native encoding like GB2312, BIG5, JIS, KSC, TIS and Tamilnet-Kanian (Singapore Tamil encoding). (Now this Java Input Engine is used in Yahoo Chinese Search - http://www.gime.com) As for Tamil, this is the first Text editor, which runs on all platforms and saves the text in the Unicode encoding. (NT5 which is supporting Tamil and Devanagiri is yet to come on the market.)

Future Direction and Conclusion

Multilingual multiscript URL

Today, the Internet has reached the four corners of the world to a diverse community with different languages and cultures. The World Wide Web has progressed to address the localization needs of its audience with Web pages in different languages a reality today. However, the Internet Domain Name System (DNS) which started out to be strictly based on a subset of the Latin 1 alphabet, is still mainly English.

This restriction also applies to other aspects of the Internet which makes use of domain names as well, e.g. telnet, ftp,email, etc. Now CIR is creating an experimental internationalized DNS as proof of concept that it is viable. It is also creating tools and applications that will enable users to key in URLs in multilingual characters (e.g. Chinese, Japanese, Korean, Tamil etc) It is also designing a URL forwarding system for multilingual-character URLs.

When we achieve multilingual directory and filenames, we will have fully delivered a Tamil script URL in addition to Tamil content on the internet. For the proof of concept visit http://www.idns.apng.org . Now TamilWeb Project is slowly moving to Unicode text archive.

The whole Thirukural and part of Purananuru are in two coding in our server. One in Singapore "kanian/tamilnet" coding. The next coding is the Unicode. In future most of the digitalized Tamil text in our will be in these two coding. With Multilingual Domain Name System, in future, the Domain name and the URL can be typed in the Tamil language.

References

Leong Kok Yong , Tan Tin Wee, Naa Govindasamy & Lee Teck Chee (June, 1996), Multiple Language Support over the World Wide Web, INET96 Conference Paper, Montreal.

Leong Kok Yong , Tan Tin Wee & Lee Teck Chee (March, 1997), Making your Web server render Unicode text for your client users, 10th International Unicode Conference, Mainz, Germany.

Bos, Bert (1996). Internationalization/Localization W3C: Non-Western Character Sets, Languages, and Writing Systems. http://www.w3.org/pub/WWW/International/.

Grimes, Barbara F., Editor. (1992). Ethnologue, Languages of the World. 12th Edition. Consulting Editors: Richard S. Pittman and Joseph E. Grimes. Summer Institute of Linguistics, Inc. Dallas, Texas.

Nicol, Gavin T. (1996). The Multilingual World Wide Web.http://www.ebt.com: 8080/docs/multilingual-www.html.

Govindasamy, N. (1989). New Keyboard for Tamil Computer, by Naa Govindasamy, 7th International Conference of Tamil Studies Seminar Proceedings, Maritius, December 1989.

Govindasamy, N. (1994a). Computer and Tamil Teaching. 2nd International Conference of Tamil Language Teaching, Kuala Lumpur, June 1994.

Govindasamy, N. (1994b). Kanian Keyboard. Tamil and Computer Conference Proceedings, Anna University, Madras, India, August 1994.

A Glyph-Based Font Encoding Scheme
- TSCII as a candidate for Tamil Computing

K.Kalyanasundaram and Muthu Nedumaran
Singapore Internet Working Group for Tools and Standards for Tamil Computing

Introduction

Dravidian Languages such as Tamil use non-Roman characters. Historically, transliterated form of writing Tamil text using Roman characters was the common practice, particularly amongst western scholars. Dedicated softwares (text editors and word-processors) capable of rendering tamil scripts using built-in Tamil fonts made their debut in eighties and soon became popular particularly in Malaysia and Singapore region. These commercial softwares were rather expensive and so their usage was restricted largely to publishing houses.

Free, self-standing font faces became available in the Internet in the early nineties and this resulted in a major explosion in the number of people who can handle materials directly in their personal computers. Though exact numbers are not yet available, it is likely that today at least a quarter of a million can handle tamil materials in tamil script on computers. This number is likely to double as another year passes by.

Last decade also saw phenomenal growth and establishment of Internet as one of the major modes of information interchange. Enormous amount of materials are being produced using computers using different operating systems and with different softwares. Information is also being exchanged through different exchange protocols (SMTP, NNTP, POP, HTTP, ...). Facile flow of information through different computers and protocols require in a mandatory way some standardized font encoding scheme.

A font encoding scheme is an explicitly written convention for handling of language script(s) by the computer systems. An encoding scheme can be glyph-based where one uses a series of graphic characters (glyphs) stored at specific locations of a font to generate the script. All font (font faces) then use this same standardized format.

An encoding scheme can also be character-based where one simply defines the basic elements required to define the entire alphabet and leave the details of rendering of the script to softwares. Unfortunately there has not been any single encoding scheme universally accepted for Tamil and used by everyone. This has resulted in mushrooming of hundreds of Tamil font faces (7- and 8-bit monolingaul and bilingual) in the Internet and near impossible situation of exchange of information between individuals. This paper proposes a glyph-based encoding TSCII as a possible candidate for fonts used in Tamil computing.

TSCII Initiative

Two major developments were responsible for the present initiative. Almost three years ago, world-wide discussions started through two Email discussion lists "webmasters@tamil.net" and "tamilnet@tamilnews.org.sg". The participants with different background (software developers, typographers, linguists , academic scholars, users) are all interested to see an encoding standard defined soon. Earlier conference in this series TamilNet'97 held in Singapore (organized by the National University of Singapore) recognized the urgent need to define a standard and welcomed initiatives in this regard. The two mailing lists were soon merged into one. The present proposal TSCII is the outcome of discussions of an Internet Working Group for Development of Tools and Standards for Computing.

Existing Tamil Standards

During this decade nearly all major representatives of hardware and software industry joined together and launched a global encoding initiative called UNICODE. Unicode is an ambitious character-based encoding scheme to handle all world languages with specific segments allocated for each language including. While Unicode is nearly implemented for English and European Languages, it is not yet ready for Indic languages.

Before Unicode, in the eighties, Government of India proposed an Indian Standard called ISCII to handle all Indian Languages under a single character-based encoding scheme. CDAC, Pune developed softwares based on this ISCII standard and these are used in state and federal govt. establishments within India. For Indic languages, Unicode has adopted the character-based encoding schemes of ISCII. In spite of this parentage of Unicode with ISCII (as far as Indic languages are concerned), implementation of Unicode is quite different from that of ISCII. Neither the softwares nor text files/databases of one scheme applicable directly to the other.

Due to the difficulties in implementing ISCII in widely used information exchange protocols of Internet, CDAC recently proposed a "secondary layer standard" ISFOC. ISFOC is a glyph-based encoding standard with no direct inter-convertibility to the parent ISCII. The complexity of file transfers between these glyph- and character-based encoding standards takes away all benefits of cross-language advantages of this standard. So the situation is far from satisfactory. The complexity of these two schemes does not allow usage of any of the thousands of shrink-wrap softwares that are available for English language. ISCII scheme can be implemented only through dedicated hardware and/or softwares. Fortunately language is comparatively a simple language adequately handled by a glyph-based encoding scheme.

Biggest challenge today

The goal today is not to introduce one more font encoding to the arena. The biggest challenge facing us today is unification of hundreds of font encodings that are in use for today. Any standard font encoding scheme proposed today must allow facile migration of legacy

documents, texts and fonts. Through its elaborate design goals, TSCII initiative attempts to address all these pressing requirements. The initiative recognizes that soon (possibly in a decade) Unicode will firmly establish itself as the world-standard for multi-lingual computing needs. Hence the initiative proposes a glyph-based encoding scheme as an "interim" option till Unicode is firmly established in all computer OS and softwares for the market.

Design Goals of TSCII initiative

The following are some of the key Design Goals of the proposed TSCII initiative:

the encoding SCHEME MUST BE IMPLEMENTABLE ON ALL COMMONLY USED COMPUTER PLATFORMS (Unix, Windows, Mac and others). New generationsof more powerful computers and associated softwares are being released every couple of years. The encoding standard should be such that it can be used in all computers released during the last decade (backward compatibility)
the encoding SCHEME MUST BE OPEN. There will no need to get permission from anyone to implement the encoding standard in hardwares and softwares. No copyright restrictions of any kind. Nearly all of the International Standards for Information Interchange are OPEN standards.
the encoding SCHEME WILL BE AN 8-BIT BILINGUAL (ROMAN/TAMIL). The widely used lower ASCII set (roman characters and punctuation marks) take its standard location (slots 0-127). Tamil glyphs occupy the upper ASCII berth (slots 128-255). This is based on the recognition that information exchange via widely used methods such as Email and Web is best assured if key information (tags) on the nature/content of the file are indicated using the usual lower ASCII set.

the encoding SCHEME PROPOSED WILL BE GLYPH-BASED ONE WITH A UNIQUE COLLECTION OF GLYPHS to generate the entire alphabet. The encoding scheme should be such that there are no ambiguities in the interpretation of the resulting text (by search and sort engines for example), no redundancy nor repetition of old style of writing some alphabets (lai, Lai, Naa, naa and Raa). The encoding will allow text input as per the current practice of Tamils worldwide, without enforcing any language reforms. In view of the near absence of usage of numerals, their inclusion can be viewed as an exception. Two main reasons for this exception are:
the need for the encoding scheme to handle very ancient manuscripts (Etext archives) where these numerals were in use;
to allow join the main stream where many of the classical languages have their own numerals. Unicode does recognize this key fact in providing slots for these in many language segments.
The scheme will not include any idiosyncratic, personal, novel, rarely exchanged or private-use characters.

the encoding SCEME WILL BE UNIVERSAL IN SCOPE. While keeping grantha characters and numerals as part of the glyph choices, the encoding scheme is designed such that the basic glyphs are collected in widely used Latin-1 (8859-x) standards. This will ensure that the "pure

Tamil" message gets through even in the poorest/bad local implementation scenarios. Supplementary characters such as granthas, numerals and rarely used alphabets (such as nju, ngu, njU, ngU) are to be placed in rows 8 and 9. By providing a couple of encoding slot positions as vacant, the scheme will allow software developers to use them as "escape routes" to bring in additional (special characters such as old-style Lai/lai/Naa/naa/Raa) characters if necessary.

The encoding SCHEME HAS TO PROVIDE COMPATIBILITY to Unicode (and hence ISCII) scheme. Proposed inclusion of grantha characters and numerals is relevant in this context. Compatibility to Unicode will ensure easy migration of documents to and fro during and after the transition period.

Efforts will be made to provide appropriate softwares that will ALLOW SMOOTH MIGRATION FROM HUNDREDS OF CURRENT FONT ENCODING SCHEMES. For conversion of legacy documents "Converters" will be provided that allow inter-conversion of files between widely used encoding schemes and proposed TSCII. The converters will also allow any user to quickly generate converter plug-ins for any custom encoding. For text/data input, several keyboard editors will be provided allowing input as per widely used keyboard layouts and existing font encodings. This way anyone can continue to do the input in nearly the same manner as before but all with the same encode-conformant fonts. Transition to TSCII thus will be smooth and rapid.

Details of TSCII encoding

This presents in the form of a compact table glyph choices and their slot allocations (code positions) of the proposed 8-bit bilingual encoding. < tscii.gif >

It can be noted that the encoding scheme includes in addition to characters of tamil alphabet the following: single and double curly quotes and copyright sign at their code-positions of widely used Latin-1 scheme. Having these curly quotes allow ready usage of shrink-wrap software (for word-processing, graphics etc.) that are available for use in English and European languages. Copyright sign (at its ANSI slot #169) is increasingly used in many of the Internet-based documents particularly in Web pages. Presence of this will avoid un-necessary switching to other Roman font faces. Two slots (254, 255) have been left vacant as "private use area" for software developers.

TSCII encoding and desktop publishing

Undoubtedly, the major use of the tamil fonts (Tamil computing) is in the word-processing application. Desktop publishing is becoming the widely used mode for printing even in professional publishing houses. During our 2-year long discussions, it was repeatedly pointed out that the glyph choices should be such that high quality printing required by professional

publishing houses must be met adequately. Many of the tamil alphabets are complex forms graphically.

It was pointed out that excessive use of kerning (as is the case with 7-bit fonts) renders delivery of high quality glyphs rather difficult. Since the number of slots available in upper-ASCII segment (#128-255) is much less than the required ones to allocate one slot for each of 240+ tamil alphabets, choices have to be made on which of the Tamil alphabets are to be included in the native form, which are to generated using modifiers (several keystrokes in sequence).

Choice of glyphs will determine the quality of the output and the slot allocations will determine the trouble-free performance across different computer platforms. It may also be pointed out the quality of any font face (outline definition of the glyphs) will largely determine the quality of the output. Most of the freely distributed Tamil fonts in the Internet perform poorly in this context. Whatever the encoding scheme the Tamilnadu Government adopts as the standard, the world of Tamil Computing will benefit enormously if the Government distributes free at least a couple of "high quality" font faces through the Internet.

Two factors guide the selection of glyphs

One is the FREQUENCY OF OCCURRENCE OF THE ALPHABETS STRUCTURAL COMPLEXITY OF THE TAMIL ALPHABET so that they can be generated nicely in on-screen display and in print with or without the add of kerning and other basic font handling techniques already available for over a decade in all computer platforms. It is not a good idea to go for an encoding scheme where 80% of the chosen glyphs occur for less than 30% of the actual text. Assuming that the quality of the glyphs in the font face are of exceptionally good quality, more the number of alphabets in native form, better will be the quality of the Tamil text. A good balance has to be made between frequency of occurrence and the structural complexity of some of the alphabets.

Fortunately several of the Tamil alphabets are written as a composite of two or three basic components (referred to as "modifiers"). e.g. aakara, ekara, Ekara, okara, Okara, aikara and aukara varisai alphabets. Along with the basic consonants (mei), it suffices to have a select collection of modifiers (aakara, ekara, Ekara, aikara, aukara) in the encoding scheme to generate all these compound (uyirmei) characters. There is no need to have these as unique glyphs in the encoding scheme. Similar logic can be applied to grantha series as well. It suffices to include the special ukara and uukara modifiers and can use the Tamil modifier glyphs for the rest. After an in-depth analysis of various options, it was decided to invoke modifiers for the "ikara" and "iikara" varisai alphabets and rest of the series are generated directly.

There have been several analyses of the frequency of occurrence of Tamil alphabets and they have been used earlier in determination of the keyboard layout. With the choice of glyphs discussed above , we have NEARLY 87.07% OF THE TAMIL CHARACTERS ACCOMMODATED IN NATIVE FORM in the encoding scheme: meis with puLLis 28.85%;

basic meis (akaram eRRiya meis) 23.50%; ukara varisai 11.88%; entire uyirs: 7.00 %; aakara varisai (with stand alone "aa" modifier) 6.39 %; aikara varisai (with stand alone "ai" modifier) 4.41%; eekara varisai (with stand alone "ee" modifier) 1.88%; ekara varisai (with stand alone "e" modifier) 1.44%; ti and tii 1.06%; uukara varisai 0.62% and aukara varisai (with e, au modifiers) 0.04%. It means that, nearly 87% of the Tamil characters are rendered as native ones without any kerning. Their quality will be dependent purely on the quality of the font face design.

Even in the ca. 13% generated via kerning (used mainly in the ikara and iikara varisai), majority of them can be generated in quite satisfactory way using kerning procedures. Kerning is a routine font handling technique now available in all of the common computer platforms/OS. As a right-end modifier, the ikara and iikara varisai uyirmeis can be rendered fairly precise on all platforms.

So it is likely that, using the proposed glyph encoding scheme, over 98% of the Tamil characters can be rendered easily on screen and in print without any loss of quality. Techniques such as pair-wise kerning can handle even the residuals adequately. Professional publishing houses with more stringent requirements on the glyph display invariably use more sophisticated printing equipment and high-end computer systems. Advanced font handling techniques such as glyph substitution (GSUB) through (or without) Open True Type fonts are already implemented at the OS level. Hence it should not be problem for these cases to use dedicated software where single form of these alphabets are stored elsewhere and brought in wherever they are needed.

TSCII encoding and information exchange

A second major area of application of the font encoding is information exchange through Email and WWW. We will discuss each of these one by one. With the emergence of 16-bit Unicode as the encoding for multi-lingualism, nearly all of the widely used computer operating systems now can handle correctly information at this 16-bit level. Computers released in this decade (target coverage of proosed TSCII) all can handle 8-bit encoded messages.

EMAIL: Nearly all of Email softwares (including those that are used in shell account access such as PINE) are 8-bit compliant. A routinely used communication protocol is MIME (also known as Quoted-printable or base 64 encoding). MIME was designed to allow email exchanges in all of the world languages without worrying much into the details of the font encoding used. MIME simply sends the code positions of the characters (A0, B2, EA etc) and the user/client-software recodes the information as the local choice of font face and associated font encoding.

Using MIME it is possible to exchange information across different computer platforms very reliably. During the testing of TSCII encoding, we have successfully used most of the commonly used Email softwares on all three computer platforms. We have already a handful of email discussion lists where TSCII-based tamil exchanges are taking place routinely and the participants use many different softwares running on Mac or Unix or Windows-based computers.

WEB: Current versions of both of the dominant Web browsers Internet Explorer and Netscape are Unicode-intelligent. Using the "user-defined" case for the font-encoding, we have shown successfully that it is possible to present formatted tamil texts (TSCII based) in the form of Web pages. Users can read the Tamil materials locally using their preferred web-browser and font face of his personal chocie. We have also shown successful demonstrations of information exchange of formatted tamil text materials via "portable document format (PDF)" on all three Mac/windows/Unix platforms. PDF format is increasingly becoming the preferred mode of distribution of formatted materials (e.g. catalogs and annual reports of business establishments).

TSCII encoding and database applications

Another major area of concern is on Database Applications. Any encoding scheme should allow facile search and sorting of stored/saved tamil information. The information searched could be a tamil text viewed on a word-processing application or large database in a business or governmental organization. Database handling can be considered in three stages; storing, sorting and searching. The database could be directly as plain 8-bit text as per the TSCII encoding. Sorting of the 8-bit tscii data can be done through an intermediate layer where glyphs are substituted by de-coupled characters and use any of the standard sorting algorithms. A demo software "varisai characters

With the very likelihood of Unicode taking the place as the international encoding standard, alternate possibility would be in Unicode. An on-fly convertor associated with the application can convert tscii data -> unicode for saving in the database and also render data back to the application via another unicode -> tscii conversion. The intermediate layering can be transparent only to the application developers. Except for an encoding scheme that lists the entire 240+ alphabets in the required sorting sequence, usage of an intermediate layer in any glyph-based scheme in inevitable.

Unicode has already released some standardised sorting options as RFCs and there are already software developers working on developing softwares based on this option. Double-byte sorting has also been proposed as an option. Clearly there will be more than one way one can do the search and sorting.

Concluding Remarks

The proposed glyph-based encoding is the outcome of nearly three years of discussions in a public forum accompanied by extensive field testing by a group of intemet-linked volunteers group. It has been shown to be a very viable "interim" option for Tamil Computing. It has the support of a broad spectrum of Internet Tamil community. In a short span of two months since the present encoding scheme was adopted as the final form by the Internet Working Group, many Tamil commercial software developers have produced several TSCII-encoding based

tools and softwares and agreed to distribute them FREE: Tamil font faces and keyboard editors for use in Windows, Mac platforms, text converters to go between TSCII and popular Tamil font faces and vice versa, Email software that allow exchanges directly in Tamil, On-fly converters of web-pages, etc.

A dedicated Web site for TSCII.

http://www.tamil.net/tscii has also been formed to provide all the necessary technical assistance for quick implementation of the standard and to serve as "the site" where anyone can download above type of TSCII-based tools. Hence we strongly believe that the proposed standard is a very viable one, guaranteed to deliver the goods it promises. We sincerely hope that the Tamilnadu Government will give a fair hearing to this proposal and possibly adopt it for Tamil Computing as soon as possible.

---

iDNS, a DNS SYSTEM WITH MULTILINGUAL SUPPORT

Mr Ching-Hong Seng & Dr Tin-Wee Tan Singapore

Abstract

The Internet is based on IP technology. Packets are routed using IP address which is in numerical form, for example 137.132.19.110. While this is useful for routing purposes, it is difficult to remember. Thus, we have a domain name system or DNS which translate domain names, eg www.nus.edu.sg to IP addresses and vice versa. Navertheless, the current DNS only allows a very limited set of characters. While this is not a problem for most English speaking countries, this is not so for many countries in the Asia Pacific region where its own national language is more used. Thus, with the internationalisation of the Internet, and the proliferation of non-English content, there is a growing need for a DNS system which is able to handle multilingual and multiscript features. In addition, with the development of Unicode which encompass most languages in the world, including Tamil, it is high time a DNS system is designed for non-English speaking countries.

The iDNS aims to address this issue by

1.Providing a mechanism for the DNS system to adopt multilingual character set thus allowing non-English language domain names.
2.Retains compatibility with the current DNS standards.
3.Transparently transform different language encoding to a universally adopted format Unicode
4.Minimum changes on the client/server easy installation for the system administrator allowing end-users to immediately use multilingual domain names

References

RFC1034, P. Mockapetris, Nov. 1987 Domain Names - Concepts and Facilities
RFC1035, P. Mockapetris, Nov. 1987 Doman Names - Implementation and Specification
draft-duerst-dns-i18n-02.txt,
M. Duerst, July 1988 Interationalization of Domain Names draft-skwan-utf8-dns-01.txt,
Stuart Kwan/James Gilroy, Mar 1998 Using the UTF-8 Character Set in the Domain Names System

---

TAMIL FONT ENCODING STANDARDS - A CRITICAL STUDY

P.Chellappan,
Partner, Pallaniyappa Brothers, Chennai,
chellappan@vsnl.com

Synopsis

The use of tamil in the word processing / DTP environment has been well established for nearly a decade now. The use of Tamil was made possible by the development of a number of fonts by many enthusiastic tamils and software houses. Although every one was aware of the non-standard encoding schemes used by these developers, no serious effort was made to standardise them as it was not a major concern because interchange of text between users was very restricted and in the few cases were it was required, the problem was overcome by translating programs. But with the advent of Internet Computing in the 90s people started communicating with each other and suddenly everyone started feeling the urgency for standardisation. Fortunately or unfortunately we are in the process of standardising at a time when the world is moving towards a global standard (UNICODE) for information interchange. This is a transition stage and one has not only to think of the present but also the future while evolving a standard.

Another major factor to be considered during this process is the fact that in the near future the use of Tamil in computers will not only be restricted to mere word processing or DTP but also database operations. Hence if due and proper consideration is not given to these major issues the future of Tamil Computing will be at stake. It is with this serious concern in mind that this paper attempts to critically analyse the existing and proposed standards. This paper does not deal with the Tamil Keyboard Layout standards as it is a separate issue all together.

Introduction

Before proceeding to analyse the available and proposed standards, one has to decide on the criteria or design goals these standards have to meet. Any standard must necessarily satisfy the following condition<%0>s:

1. It must be capable of accommodating all the Tamil and grantha characters and their variants.

2. It should enable unique character to glyph mapping as far as possible because the quality of the printed text is as important as text interchange. Any attempt to create character representation by combining glyphs through keming will greatly affect the quality of the printed output because:

a) Not all word processors support kerning. b) In those word processors that support it, aesthetics of the Tamil character will be lost. c) Softwares with only basic justification capabilities pull the combining glyphs apart. This is not a desirable feature in the print media. And even in softwares where you can prevent addition of intercharacter spacing during justification, the look of the output will suffer. d) It inhibits the creativity of the font designer from freely designing character shapes. Examples of these problems are given in Annexure-A.

3. It should facilitate easy sorting and searching as it is essential for database applications.

4. It should be usable on all computing platforms available today and of course without doubt on future platforms too. The text based Operating Systems have given way to the graphics based OS and are almost obsolete except may be in POS terminals and hence need not be considered. Having stated the broad design guidelines, we now have to look at the existing encoding schemes in terms of its advantages and disadvantages. This is a very daunting task as there are number of different schemes in the market today.

However for the purpose of this study these different schemes are grouped under the following heads and each of the heads analysed separately :

1.Bilingual Vs. Monolingual
2.128 Characters Vs. 256 Characters 3.Single Byte Vs. Double Byte

Bilingual Vs. Monolingual :

In Bilingual fonts Character Sets of two different languages co-exist and they share the available 256 slots in the font. The concept of Bilingual fonts came into existence in Europe where almost all the European languages shared a common script and varied only in a few accented characters. It is the similarity of the script which prompted the emergence of the Bilingual font system and not the need for simultaneous usage of two languages. A Bilingual font system meant that font developers could sell the same font for all languages. At the same time and more importantly no major compromises were made to accommodate two languages in the same font. There was a perfect one to one character to glyph mapping. Of course an added advantage was that several European languages could be used simultaneously without change of font. The use of a Bilingual font system for Tamil with the lower 128 character set being used by the Roman Script means that only 128 slots are available for the Tamil Character set and even out of these slots a few are unavailable because of some OS restrictions. In the Multilingual font system as envisaged by the ISCII standard too only 128 slots are available. Although the Tamil language has 256+ characters which are unique, the design of these individual characters fortunately allow them to be constructed by adjacent placement of two or more component or glyphs. By using this advantage we were able to accomodate the entire Tamil character set in a 256 character system and at the same time did not compromise

on the individual characteristic of each character. Kerning was completely avoided in this system and hence the print quality of the text and the display quality too did not suffer in any way. If we have to live with only 128 slots then only two options are available. The first is to use sophisticated Glyph Substitution techniques to ensure that each tamil character had a unique glyph and the second is to compromise on the quality of the printed text.

The first altemative is just evolving and is currently not available on low end software packages. Hence only those users who can afford these expensive packages can benefit form it. The second alternative is really not an alternative and is only a poor man's choice to say the least. Having said that, the pertinent question now is Why Compromise at all? Why should English be used along with Tamil in the first place? Is it to satisfy the needs of a few who would like to use English along with Tamil without changing the font? In the few cases where arguments are forwarded for the necessity of co-existence of English and Tamil can we not find a work around? Is it really impossible to use Tamil on the Intemet without providing for the English script in the Tamil font? Is it because of convenience that we choose to have a bilingual font? Is it that this is the only solution which will ensure a trouble free use in database applications? Should the quality of the Printed Output suffer because of this unnecessary compromise? By using monolingual fonts have not English-Tamil Dictionaries been printed by using available off the shelf DTP packages? Have not ordinary users printed great looking letters and articles containing Tamil and English text by using monolingual fonts even on low end word processors? Should the Publishers and Printers, who produce millions of books and magazines, go in for expensive software just to satisfy the needs of a few who need English along with Tamil in the same font? The above questions which repeatedly come to our mind cannot be simply ignored or wished away. The argument that by using a bilingual font like TSCII vl.6 one can print a text that has nearly 87.07% of its characters looking correct is simply not acceptable. Does it mean that the remaining 13% of the characters should be overlooked because of the compromise? Is it acceptable to us if 87% of an image in a photograph is in good focus and the remaining not? Well, compromises can be made and should be made if there is no other alternative and if the future of Tamil is at stake. But till then let us not talk or even think about compromises.

One final thought on this subject. If the Tamil scholars agree to simplify the tamil script and form words just by placing consonants and vowels like the Roman Script then of course it will be possible to even have a multilingual font without any compromises instead of a bilingual one !!

128 Characters Vs. 256 Characters :

Most of the above arguments in favour of monolingual fonts apply here too. Suffice it to say that only a 256 character based monolingual font will be capable of meeting the needs of the Tamil Text Processing enviromnent. As far as the database applications go both the monolingual and bilingual font systems based on a 128 character set will pose

varying degree of problems in areas for sorting and searching which is so vital to these applications. But several attempts have been made and it is definitely possible to arrive at a solution once we decide on a standard character encoding scheme. This may not be a big stumbling block.

Single Byte Vs. Double Byte :

The double byte system has been widely used by the Chinese, Japanese and Koreans. This has helped them to accommodate the several thousand characters in their script. The far eastern edition of Microsoft Windows edition supports several double byte character sets. Why not try to develop a double byte character set system for Tamil also ?

On the face of it, it might be a ridiculous suggestion. But in reality it is not. We are only now in the process of standardisation. We should have an open mind and discuss the pros and cons of this system and if the pros outweigh the cons we should definitely adopt the double byte system. The single byte based Tamil fonts have been in use for over a decade and we have had no problem except that of standardisation. In fact we are all very comfortable with this system and all readily available software packages are able to handle Tamil easily. Then Why a double byte system? Let me now answer that question.

Mind you the thoughts that I am now making have not been researched but only serve to kindle us to think with an open mind before making a final decision on Standardisation. Till a few years back the use of computers in India was mainly limited to business houses whose requirement of Tamil, if at all any, was limited only to the word processing environment. But the present policy of the Government of India and several State Governments has brought the computer within reach of the common man. In this changed scenario we can now think of extending the use of Tamil to database applications also. As we are aware sorting and searching processes are very vital to these application. The present single byte encoding schemes can be used provided we have a separate sorting algorithm for Tamil.

he computing world is now moving towards UNICODE which has established a global standard for text information interchange. The
UNICODE standard only aims at having a single comprehensive coding scheme for all the different world languages and is not concerned about transliteration of these languages as envisaged by the ISCII standard for Indic languages. The ISCII standard specifies only a 128 character set for each of the Indic languages and all the disadvantages of the 128 character set as discussed earlier applies to the ISCII standard. The ISCII standard gives unique location for each of the vowel, consonant, and modifier characters and leaves the screen rendering process to the software / Hardware. This has been adopted to make transliteration between the Indian languages possible. But transliteration is not perfect because of the inherent differences in the languages. Even presuming without admitting that the available transliteration is sufficient, the use of

transliteration is very insignificant and the standardisation process should not consider this as a major criteria. In cases were transliteration is required separate transliteration software can be developed. The simple solution will be to disregard the ISCII standard and evolve a separate standard for Tamil. But the problem does not end here.

Unfortunately UNICODE has adopted the ISCII standard and allotted only 128 locations for Tamil (U+0B80 to U+0BFF). It will not be out of place to mention that several languages have been allotted more than 256 locations in the UNICODE structure. Examples are Ethiopian (U+1200 to U+137F) and Canadian syllabics (U+ 1400 to U+167F) in addition to of course the CJK languages. This being the case why can't we get the required number of locations in the UNICODE structure to accommodate each of the 256+ Tamil characters individually and not rely on the advantage of creating these characters by adjacent placement of two or more glyphs. By doing this we have a simple one to one mapping between the glyphs and each of the 256+ tamil characters. This will greatly simplify the sorting and searching process in database applications also. The next question that comes to our mind is what about software? More and more applications are now becoming UNICODE compliant and the day is not far off when almost all software will become UNICODE compliant. Hence the question of Software compatibility is only a temporary phase. Let us try to utilise the big opportunity given by the UNICODE structure to our advantage. Let us not be myopic and still think that only a maximum of 128 or 256 character sets are available to us and that we have to cut our feet to match our shoes.

Conclusion :

The above discussion clearly leads to the conclusion that for the present we should standardise on a Single Byte Monolingual font system which facilitates easy sorting and searching. The existing Monolingual font systems should be analysed from this point of view and a suitable one could be adopted as the standard. A possible starting point could be the Chennai code Version 1.4 which has been proposed by Dr.V.Krishnamoorthy.

These fonts should also without doubt ensure a high quality of printed text which is of utmost importance not only to the Print media but also to the cause of Tamil in general. But the future of computing lies in UNICODE and we should try to exploit the wonderful opportunities thrown open by it. We should seriously think about the Double byte encoding scheme as it will greatly help in database applications and we should certainly not try to live within the confines of the 256 character limitation.

Annexure - Problems associated with Kerning

1. Kerned characters loose their aesthetics. As can you clearly see below a single kokki glyph will not suffice. The gap between the character and the glyph will be inconsistent and in fact for the letters `ma' and `sha' it will almost coincide with the letter itself while for

the letters `ka' and `sa' it will be far apart.

2. During justification the glyphs of the kerned charactrer will be pulled apart. As you can see below the Tamil letter `sa' and the `kokki' are pulled apart in the first row. The letter `ma' looks ok in the first row but in the second row it is pulled apart.

3. Because of kerning, characters such as those shown below in the first row cannot be designed as you have a choice of only one `kokki'. With kerning you can only have characters as shown in the second row.

A PROPOSAL FOR TAMIL STANDARD CODE FOR INFORMATION INTERCHANGE

N. Anbarasan Applesoft,
Bangalore - 560 010
Tel: 3386167, Telefax: 3357167
e-mail: aplesoft@vsnl.com, anbu.arasan@axcess.net.in

Introduction

Tamil is the only living classical language of the world. It creates interest among the researchers world wide. Various universities, Institutions and individuals around the world doing linguistic analysis, are involved in Tamil literature archive and analysis for various requirernents. The efforts of these universities and institutions are funded by various Governments, semi-governments and non- government institutions, which involve crores of rupees. For the lack of a standard - which could meet their requirements, Tamil data portability is lost. The data of various projects are being archived in their own coding schemes, which is accessible only by those, who created such data bases. There is an invisible difficulty in exchanging the information and processing such information, by others - other than those who had created these data bases. Computers have become an inevitable part of lifestyle of mankind. The Tamilnadu Government headed by Dr. Mu. Karunanidhi is having aspirations of making Tamilnadu a computer literate state.

Computers are being introduced at school levels, it is the foundation to prepare trained manpower to meet the future requirements. Amidst Computerisation at various levels in government offices, departments, undertakings, educational institutions etc. a vacuum felt in data portability of Tamil digital texts, be it a simple letter or a volume of literature. It s noteworthy mentioning that, wayback during 1995 I came forward to offer my software christened Surabhi Professional ver 2.00 to the Government of Tamilnadu, free of cost to be distributed to the needy users by a centralised agency. However, a response is yet awaited

Need for standard code

Tamil being a well structured and having well defined grammar even at word level, a need arises for everyone to convert the available data to a character level coding, which helps in various processing requirements. It can be mentioned that, the advanced technical development such as spellchecker translation, OCR etc., could happen only on the basis of Character coding not on the basis of glyph coding. character coding is a must for information processing and interchange.

Need for separate standard for Tamil

Tamil language is well structured with rich grammar. Tamil language is having different

character order than any other Indian Languages, which can in no way be achieved along with any other language. In principle the ISCII is designed based on Devanagari script (please note that ISCII is not phonetic or character coding scheme), which is having contrasting script structure than Tamil. In unicode, all Indian languages are coded as separate standard, which will help to maintain the traditional identity of the languages. Tamil language is wrongly coded in the Unicode standard. As far as ISCII is concerned, the revision of ISCII is ruled out in the near future. The way in which Tamil characters are coded in ISCII is a bottleneck in all the projects. With these backgrounds it becomes necessary to have a separate standard for Tamil. Unicode will emerge as an international standard when the stand alone operating system like Windows 2000 comes supporting it or when the Windows NT becomes the standard desktop operating system. Microsoft has already developed Windows NT lab version with Tamil Unicode. If we could Come out with a standard character code for Tamil, the Tamilnadu Government can appraise Microsoft in using the standard and push to Unicode.

Need for Multiple coding Schemes

Any of the present OS can't handle Tamil character codes. These 0Ss are supporting fonts of different formats. However, there are certain application softwares which do not allow selection of fonts. One such application could be terminal emulation softwares. To meet the various requirements of the users, the proposal for Tamil standard consists of three coding systems, namely:

1. Character based coding
2. Glyph based, monolingual font coding
3. Glyph based, bilingual font coding The ISCII is also having different coding system to meet different requirements. In any character based coding system, a font is a must for rendering of proper text.

ISCII compatibility

The Government of India is spending crores of rupees on projects of National importance and on Indian Language enabling technologies. To make use of the benefits of such projects and for portability of data it is a must for any software developed for Tamil to support ISCII atleast in the form of export or import facility. Irrespective of the problems or shortcomings in the ISCII, it is a National standard for Indian Languages.

1. Character based coding

We are witnessmg a new horizon blooming for Tamil computing. There is a lot of projects initiated by the Govt of India, which is of national importance. These projects involve lot of processing on data, which is in Indian Languages.

It is evident from the National and International standards, the coding has to be based on characters and not on glyphs. Following table shows Tamil character coding scheme.

Data identification

In the present day computers, English is playing a vital role as every aspect of computer is designed using English. It becomes more than necessity to co-exist with English. This bilingual character based coding system is flexible in data identifsication. This is an important feature for data processing, communication, rendering etc.

Information processing

Character coding provides Language analysis and processing facilities, which is the basic requirement for any further development of Tamil specific applications.

Phonetic coding and archiving Tamil literature

As the coding is, in its basic phonemic level, it becomes handy to archive old and new literature alike. The display rendering system could take care of the script pattern to be used for the data. It means that, by using the phonetic coding system, the literature can be well preserved.

Less space requirement

As the language is being coded in its basic form, the archiving of data will require less space.

Phonetic spelling

The spelling of a word is the basic constituent alphabets (characters phonetic spelling is essential for grammar based processing such as spell checkers, dictionaries and OCRs. These phonetic spelling is also important in search operations.

Portability

As the data is in its pure alphabetic code, the data can be shared amongst the various researchers, developers for their various requirements with converting the data.

Word boundary

In applications, the minimum feature supported is search. It entirely depends on isolating words, this coding meets such requirement.

Conjuncts

When two or more consonants occur together, these consonants join together to form a syllable, and take a different shape, which is very different from its constituent consonants. As these syllables are derived and can be converted to its constituent characters, are not coded.

Abbreviations or signs

To increase the writing speed of Tamil script, certain abbreviations, were used. Theses abbreviations are difficult to decipher into character codes, without a lookup table. These abbreviations are coded as symbols. This is not exhaustive, and could be finalised to meet the requirement.

Convertibility to 7-bit systems

There are installations of 7-bit UNIX systems, which are being used for linguistic analysis and processing. This coding system is designed in such a way that, with simple `bit' operations or addition/subtraction operation, the data can be converted from 8-bit to 7-bit and vice versa.

Usability on 7-bit systems

On conversion from 8-bit coding, the data is mapped on to the English alphabets, which is a must for processing.

Numerals

Tamil is having its own distinct numerals

2. based, monolingual font coding s 1. Character based coding , which is in Indian Languages.

……..
…
…

TOWARDS DEFINING THE DESIGN GOALS WHILE PLACING TAMIL IN THE UNICODE

Dr. P. Chandra Bose
Reader in English
Presidency College (Autonomous), Chennai 600 005.

Abstract

This paper tries to seek and define certain objectives based on the experience of working with various Tamil Desk Top Publishing packages and Word processors while thinking seriously about the present attempts of various forums in finding slots in the UNICODE for Tamil. The approach is basically as that of a user and at the end a researcher in Tamil computing. The objectives are termed Design Goals and an attempt is made to make a study of the existing codes at present in view of these Design Goals. Based on the study an earnest appeal is made to the forums concerned to ask for more slots for Tamil in UNICODE.

The study finds and arrives at the following Design Goals:

1.We must have unique Tamil fonts, which could be used in all widely used computer systems throughout the world.

2.The font system should accommodate all the possible variants of Tamil fonts and the few grantha characters. Each one should have a glyph on its own without any kerning.

3. Any attempt in evolving a unique font system should not sacrifice the aesthetic quality of the printed Tamil text.

4.The font system to be devised should not end up with the ultimate goal of promoting DTP possibilities alone. It should provide easy way for data processing by yielding to sorting and searching operations

5.The font system should give adequate place to the creative imagination of the font designer.

The study at the end justifies the finding that a minimum of 256 slots are to be asked for Tamil and any encoding scheme must utilise them to have unique glyph based characters. Tamil has subjected itself to various and varying factors of influence and has changed at various points of time. No doubt it has changed succumbing to the demands of its developing society and its people. Now it has to face a major force viz. the spread of information technology without which no walk of life could continue to grow.From the time that computers have started setting in the land of Tamils, throughout the world attempts have been made to use Tamil in computers. While English continues to be the language in the operating systems and software packages in the computers even today,

Tamil enthusiasts have not attempted to alter it and introduce Tamil. Most of their efforts ended in finding the language in the screen and to use it to have just texts in Tamil. Even capable Tamil technocrats in this field did not think of the laymen in Tamilnadu who have no other language to communicate to the world. They must have been and are carried away by the faith that most of the Tamil population are educated up to a level of understanding certain core words in English which are found necessary to operate a package in the systems. Unfortunately it is not the case. The use of the systems has penetrated to a great extent in the society that the uneducated while attempting to use the packages found it difficult to comprehend and grasp the use of it. This factor is to be borne in mind when one takes up the task of fixing Tamil in UNICODE.

The language should be placed in such a way that it must give way for all computing jobs involving data analysis including in the web and not just for the jobs related to Desk Top Publishing alone. This should be the primary Design Goal while placing Tamil in the UNICODE. So far, lot of work has been carried out in different parts of the world in developing Tamil word processors. The developers are many in the field and hence it has resulted in many non standard-encoding schemes. The enthusiasm and happiness that one achieved in creating and setting Tamil fonts on the monitor and just a text on the paper overshadowed the handicap in such non-standardization. But of late the proliferation of the Internet usage and the World Wide Web has made the Tamil using community to go in for standardization. And incidentally the intense striving for a standardization of Tamil has got up at a time when the computing world itself is moving towards a UNICODE, which will provide a universal standard for exchange of data and products. Hence our responsibility in recommending and endorsing a standard code for Tamil has become of vital and utmost importance. To achieve this end we must have our Design Goals defined and designed properly targeting all our needs not only at present and also in future. Here again we are placed in an embarrassing situation. There have already been attempts towards standardization. Each group and individuals have propelled their word processing packages based on different coding systems to the public. Facility for transmission between these packages is provided only within a few.

The developers of these coding systems continue to work with the same zeal and interest. With the result now each one would like to have his or the group's coding system get standardized. This tendency must be avoided and future needs alone be the criteria in setting the standardization. Among all the attempts made, the coding given in TSCII scheme is of much significance. The Tamil using community must stand grateful to the group that has evolved it. An important point to recall is that it was evolved after hectic work for two years by leading Tamil technocrats on the globe. Among most of the information interchange packages (like Murasu Anjal) also follow the same coding system. This system deprives us of lot of flexibility and benefits, which are available in a monolingual fonts system which is prevalent in most of the DTP packages in Tamilnadu.

First of all if an encoding system for Tamil is to be universal in scope it can be achieved in the given context in monolingual font system alone. In a bilingual font system kerning has to be done and it makes the execution slower. Secondly if one takes into account of the number of users of bilingual fonts and compare them with the number of users who use only monolingual, the second category outnumbers the first. Moreover a driver that switches to the other font can be easily made available. Hence one need not sacrifice the benefits of a monolingual font system for the sake of the existing bilingual systems. As a third criterion we can think of the future needs where we require a package like MS Office in Tamil. We may need to search or sort out large databases. This can easily be done at a good speed only in a monolingual font system. Though it could also be tackled in a bilingual system it will take lot of time. Anyhow as many desire phonetic encoding could also coexist as a standard just for information exchange. Most of us think of inputting the text and getting it in print without much quality. While embedding the glyphs through kerning the beauty of the printed text is lost. The printed text does have an aesthetic appeal, especially to the young learners and at any cost it should be preserved. This can be maintained only in the case of a monolingual font system with reference to Tamil. While designing a font, a font designer should have adequate place for his creative imagination. This can be made possible only if we have place for all the characters in Tamil through monolingual font system. Each possible character should be represented as a separate glyph. Then only it will be easy for a graphic based system to work fast. We are sure that we will not reverse ourselves back to text based system. Hence to accommodate all the Tamil characters along with the needed few grantha characters we have to follow the monolingual font system. A change, though for more power and facilities, is looked at with askance by those who are reluctant to change the track. Hence nowadays voice of objection is heard from the quarters of the technocrats who were responsible for the early well-knit attempts in Tamil computing. They even recommend to have some adjustment and sacrifice with the language use. What we do today will be passed to our future generations. Let our successors do not blame us in future for not resorting to a change for the best. First of all why should we give up certain benefits and why should we sacrifice for a few.

Hence the above arguments place before us the following as our Design Goals before us:-

1.We must have unique Tamil fonts, which could be used in all widely usedcomputer systems throughout the world.

2.The font system should accommodate all the possible variants of Tamil fonts and the few grantha characters. Each one should have a glyph on its own without any kerning.

3.Any attempt in evolving a unique font system should not sacrifice the aesthetic quality of the printed Tamil text.

4.The font system to be devised should not end up with the ultimate goal of promoting DTP possibilities alone. It should provide easy way for data processing by yielding to sorting and searching operations.

5.The font system should give adequate place to the creative imagination of the font designer.

These Design Goals naturally lead us to have at least a minimum of256 locations in the UNICODE system. But unfortunately the Unicode Consortium
based on ISCII has assigned only 128 locations to Tamil. This was done without taking into consideration the complexities of character formation in Tamil. ISCII's goal was to provide a computing environment where users without knowing the Indian languages can learn them through transliteration. In the process of achieving this a lot was sacrificed and end goal was also achieved only to the extent of transliterating proper nouns. Moreover there are lot of difficulties in the ISCII standards which are obviously felt in tasks like sorting, indexing and in identifying the characters. With the above Design Goals and the encoding schemes available, the Chennai Code Version 1.3 satisfies most of the needs enumerated above. But our needs grow and timely revisions and formation of new codes are necessary. Let us get the required locations from the Unicode consortium and work on. We are safe in the sense that we have time to press forward our request for more locations. Now it should be our endeavour to seek for a minimum of 256 locations and if possible more. We are not greedy in asking for more; in future we may include a few more characters in Tamil due to the spread of knowledge in new vistas or we may modify some of the existing characters. In the light of some of the world languages that have got more than 256 locations let us also ask for more. More locations will have more flexibility in providing a suitable system in this fast changing and growing world. The situation that we meet today is only a transitory one and we can get things done if we construct or reconstruct our theories first. Let us all work together keeping the future needs and generations in mind. In our world nothing is permanent except the change. Let us change for the best.

MULTIPLE ENCODING SYSTEMS FOR DIFFERENT COMPUTER APPLICATIONS

Dr. M. Ganesan
Central Institute of Indian Languages, Manasagangotri, Mysore - 570 006
Ph: 0821-515558 (Off) Ph: 0821-410740(Res) e-mail: ciil@giasbgO1.vsnl.net.in

---

Introduction

Any language needs a standard encoding system for easy transportability of data across platforms (DOS, Windows, Unix, Mac, etc.) and application softwares. ASCII is one such standard for all the languages using Roman Script. There are many basic differences between these languages and Tamil. The writing system of the languages like English is basically a spelling based one whereas Tamil, for that matter any Indian language, is a syllabic language. Due to this difference and the complexity of Tamil scripts, there are problems in constituting a single encoding system. Again due to the lack of a standard code, many codes came to exist and created hindrances in data transportability and information interchange in Tamil.

Standard Encoding System:

Whether Glyph Based or Character Based

In information technology Characters are abstract information elements in the domain of coding for data interchange whereas glyphs are abstract presentation elements in the domain of presentation processing..... Further - A character conveys distinctions in the meaning or sound. A character has no intrinsic appearance. -A glyph conveys distinction in form or appearance. A glyph has no intrinsic meaning (ISO/IEC Working Draft, 1996. page 3). It means that characters represent sound and meaning of a language whereas glyphs represent the writing of the language. Even historically speech preceeds the writing system of a language. Writing.... is a late cultural invention and it comes later than articulacy in the history of individuals and writing is a device to record and a tool to represent speech (Srivatsava and Gupta 1990:12-13). It is clear that the basic encoding system for information interchange must be a character based rather than a glyph based one.

Glyph Based Encoding System

As far as Tamil is concerned, character based encoding system alone cannot meet the requirements of Tamil computing. The reasons are

1) not having one-to-one correspondence between character and glyph,
2) more number of graphic characters due to the syllabic nature of the language,
3) not having uniformity in consonant-vowel clustering and more importantly

4) limitation of technology at system level for handling the complexity of writing systems of Tamil like anguages. For different applications like document preparation, database creation, e-mail, web page creation, browsing, and searching, etc. on popular platforms like Windows, at present, there is a need for glyph based encoding system.

Disadvantage of Glyph Based Encoding System

Data on a glyph based encoding cannot be usable directly for a number of applications, listed under category III later in this paper. For any linguistic study on Tamil, characters have to be identified and recognized by the system, which is not easy with glyph based encoding system. Characters are represented with varying number of glyphs.

For example,

1) single glyph for two characters (phonemes) h = k + u
2) two glyphs for single characterC,= i
3) three glyphs for two characters L = k + o, etc That is, number of bytes per character is not uniform.

So it is difficult to locate the character boundary. Sorting, which is a minimum requirement for word indexing, dictionary making, etc. is not possible with glyph encoding.

Character Based Encoding System

As mentioned earlier, character encoding system is the basis for information interchange. Standard codes, ASCII, Unicode, ISCII, etc. are character based encoding system. The advantages are

1) sorting is simple,
2) easy data transportability across platforms and applications softwares and 3) Straight away usable for NLP related works.

Short-comings of ISCII for Tamil Computing ISCII code is a character based encoding system. It is a standard code common to all Indian Languages. But it cannot be taken as a standard for Tamil, for the following reasons:

i) The inherent `a' in a consonant does not have a code. This is because of the considerations that L + . = e; this derivation is unnatural. Because of the lack of code for `a' matra, it is impossible to separate the vowel `a' from a consonant. In Tamil, for example, `a' is a morpheme denoting relative participle(RP) marker(peyarecca vikuti), etc
eg:TzR=Tz++Averb+Past tense+RP marker

ii) Vowels and their corresponding matras are having two different codes, though they are phonetically same. eg:7=+ B| When segmenting `a: Tu' from `ni:r,'one will get "| which is first of

all meaningless and while searching for B|, "| will not be counted automatically, as they have different codes. Similar problems are there for all other vowels.

iii) ISCII follows the alphabetic order of Hindi, and therefore a number of characters are (^, \, Z, [,]) dislocated from their regular alphabetic order of Tamil. So, sorting is not possible without an additional programming.

iv) Indian language numerals are not in the regular slots of Arabic numerals and therefore any computation, when it is needed, is not possible.

v) Each of the five plossive sounds (L, N, T, P, R) has three more consonants as its varg-as<%-2>pirated, voiced and voiced appirated- in all Indian languages except Tamil. In the case of Tamil, ISCII represents all the four consonants with single letter and thereby a consonant is enco<%0>ded with four codes. Thus, unambigious representation of character is not possible with ISCII. For the reasons mentioned above, ISCII cannot be adopted as a standard for Tamil. Similar is the case with Unicode, as it almost follows the pattern of ISCII. So, there is a need for a separate character based encoding as a standard for Tamil.

Applications of Computers in Tamil Studies

It is mentioned in the proposal for TSCIIpage 5) that it is likely that over 90% of Tamil computing is in the form of simple wordprocessing of plain text. It is worth to mention here that Tamil texts have to be first inputted mostly through a word processor to have the data in machine readable form. A number of research Organisations and institutions have been working for the last one decade on various aspects of Tamil language. To mention a few,

1) the Central Institute of Indian Languages, Mysore has built a machine readable corpora to a size of above three million words for many Indian languages including Tamil; software tools for the processing of corpora have been developed and currently working on Lexical Resources development for 5 Indian languages including Tamil.

2)IIT, Kanpur and University of Hyderabad are collaborately working on a Machine Aided Translation between Tamil and Hindi.

3) C-DAC Pune is working on Spell Checker for Tamil.

4) Tamil University, Thanjavur has prepared a word index for Sangam literature. In most of these works Roman characters are used to represent Tamil, due to nonavailability of a standard character based encoding system. If an appropriate character encoding is not made available to the researchers, they are, further encouraged indirectly to use Roman characters. The greatest advantage of computer itself is that data are stored in electronic media and therefore any manupulations on the data can easily be done. Here the type of applications of computer on Tamil are broadly grouped under three categories.

Category I:
Document preparation, Information exchange through Internet and e-mail, searching, CAI and CALL packages, Multimedia Titles, etc.

Category II:
Calligraphic works - Desk Top Publishing in publishing houses, Newspapers, Magazines, Advertisement agencies, etc. Category III:Sorting, Indexing, Information processing and retrieval, Corpus building including literary archives, NLP works which includes morphological analysis, syntactic parsers, semantic analysis, etc, Machine Aided Translation, (Electronic)Dictionary compilation, Language Standardisation, Style analysis,Text-to-speech(Speech Synthesis) and Speech-to-text (Speech recognition), etc.The above classification is made on the basis of the requirement of different encoding systems for Tamil. They are

1. = A Bilingual Glyph based encoding for category I
2. A Monolingual Glyph based encoding for category II
3. A Bilingual Character based encoding for categoy III O. S. like windows has the feature of storing the text in 1) the format of the packages used 2) print-files and 3) ASCII format. It shows that data created can be stored in different formats depending upon the future use of the data. In case of Tamil, the same principle can be extended to the coding level, by considering the innate complexity of Tamil writing system and other facts mentioned earlier in this paper.

Three Standards for Encoding of Tamil

1. Bilingual Glyph based encoding system:

At present a number encoding systems are in use. I studied the characteristic features of four such bilingual encoding systems, produced by 1) ING-TSC(TSCII), 2) C-DAC, Pune, 3) Govindaswamy, Singapore and 4) Krishnamoorthy, Chennai.
The chart given below compares the features of them. No., Features, TSCII(1), C-DAC(2), Govindaswamy(3), Krishnamoorthy(4) 1. Slots between 128 and 160 are left free, No, No, Only 3 charactersin 150,, 151,, 157, No 2., Whether all characters are represented or not, Yes, Yes, Yes, Yes 3., Whether character representation is unambiguos, expect, No, No, Yes 4. Tamil numerals included or not, Yes, No, No, No 5., Special symbols included or not, No, Yes, No, only The encodings is for bilingual use (i.e using Tamil and English together) for the applications listed under category I. So the features 4 and 5 are not seriously considered. All the four satisfy the feature given in 2. The features (1) and (3) are more crucial. If the slats between 128 and 160 are defined with Tamil characters, on terminals there is a chance for not getting them displayed. So, it is advisable that codes should be above 161. Generally the encoding system of Govindaswamy is better and only 3 characters, that too rarly used ones are defined in those locations. The other vital features to be considered is that the encoding is unambigious in generating graphic characters. Encoding system of Krishnamurthy is unambigious, if we take

the glyphs defined only in Upper ASCII. To suit the requirements by keeping the constraints in view, a bilingual encoding system has been defined and proposed herewith. (Table 2.4.1). It satisfies all the three features which are necessary for the applications mentioned under category I.

2. A monolingual glyph based encoding

The purpose of this encoding is to meet the requirements of DTP works. If the encoding takes care of all the graphics characters, numerals, and special symbols with no or minimum of kerning, that will be a better suitable one for the purposes mentioned under category II. Though different monolingual encodings meet this requirements, the one proposed by Anbarasan M/s Applesoft has 3 additional features.

1. The fonts defined below 128, match with typewriter keyboard layout; 2) All the glyphs defined below 128 can make all graphics characters, and therefore, usable in the 7-bits systems like pagers, and 3) the alternant forms which was there earlier in use; Nai,'lai' and 'Lai' are also available in the system. Because of these merits, the monolingual encoding systems of Anbarasan can be taken as a standard.

3. Bilingual Character based encoding system

To my knowledge only Anbarasan, M/s AppleSoft has defined the bilingual character based encoding systems for Tamil, other than ISCII and Unicode. The character codes are defined such a way that 1) Tamil Characters occupy safely, above 193 as various standard packages often use some slots which are normally below 192, 2) Through bit manipulation Tamil characters will occupy only the slots of Roman Characters but not the symbols, 3) Tamil Numerals are assigned the codes of Arabic numerals, so that any computing can, even, be done with Tamil numerals. 4) Any addition of special symbols, need to be included can be done without disturbing the system. Further this encoding overcomes the shortings of ISCII and therefore it can be taken as a standard for character based encoding system.

Conclusion

This paper establishes the need for three standards of encodings, till technology develops to handle languages like Tamil only with a character encodings. Provision have to be made for conversion from one encoding to another.

References:

1. ISO/IEC Working Draft TRXXXXX, Information TEchnology- An Operational model for characters and glyphs 24 June 1996.
2. Srivastava. R.N. and R. S.Gupta, Diamentions of Applied Linguistics, Mysore: Central Institute of Indian Languages, 1990

A Tamil Speech Synthesis System

Ramakrishnan, A. G., Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012.
And V. Karthigeyan, Ncore Systems, Bangalore

Abstract

This paper describes the development of a limited vocabulary, preliminary speech synthesis system for Tamil, incorporating some rules representing the knowledge sources. A speech synthesis system has many applications, such as natural language interface for computers, multimedia education packages in Tamil, automatic telephone based enquiry systems in TamilNadu Government organizations, audio on-line help in Tamil based IT software, Virtual teacher on CD ROM/ Internet, computer based Tamil teaching, automatic document reading machines in Tamil for the blind, speech output in Government Online projects (such as that proposed with World Tel), and intelligent toys/multimedia aids for children. Speech synthesis involves synthesizing intelligible and natural-like speech from text in coded form (ISCII or unicode). This involves two major phases (i) the text analysis phase and (ii) the speech synthesis phase. The text analysis phase parses the input text into a sequence of basic units of speech. The speech synthesis phase involves the concatenation of the parameters of these units and synthesis, after the application of bath segmental and suprasegmental rules for n naturalness. Since Tamil is phonetic in nature, characters are chosen as the appropriate basic units. Acoustic parameters such as linear prediction coefficients, formant frequencies, bandwidths, pitch and gain are presto red for the basic speech sound units corresponding to the orthographic characters of Tamil. The parameters are concatenated based on the imp~ text for synthesizing speech. In order for the synthesized speech to appear natural, we need to adequately model the various features of natural speech, such as courticulation, intonation and duration. These knowledge sources are stored in the form of rules. Courticulation rules specify the pattern of joining the basic units. Intonation rules specify the overall pitch contour for the utterance. Duration roles modify the duration of the basic units based on the linguistic context in which they occur.

Introduction

Speech synthesis involves synthesizing,

~. ~.

Intelligible and natural-like speech from unrestricted text of ~a language. Progress in speech synthesis has been made possible by advances in linguistic theory, acoustic-phonetic characterization of sound patterns, mathematical modeling of speech production, ~structured programming and computer hardware design. Potential Applications of a speech synthesis ~system in Tamil include

1. Natural language interface for computers

2. Self-learning multimedia education packages in Tamil

3. Automatic telephone-based enquiry systems in all Thailand Government organizations.

4. Audio on-line help in (futuristic, but positively feasible) Tamil based IT software

5. Virtual teacher on CD ROM/ Internet (for distance learning/ Open Universities)

6. Computer based Tamil teaching (speaking books, multimedia dictionaries)

7. Text to speech in Tamil and voice mail

8. Automatic document reading machines in Tamil for the blind

9. Learning IT in Tamil

10. Speech Output in Government Online projects (such as that proposed with World Tel)

11. Intelligent teaching toys/ multimedia aids for children.

Developing a complete speech synthesis system requires proper insight into the human speech production mechanism and knowledge of the perceptual significance of various acoustic parameters of the speech model. The major issues in the design of a speech synthesis system are the following:

? Choice of an appropriate speech synthesis model. ? Collection of data required for segmental synthesis. ? Acquisition, representation and incorporation of various ~ knowledge sources that account for the naturalness of speech.

~The common approaches followed for speech synthesis are:

81 '? Concatenation method.
? Synthesis by rule method.

~ The concatenation method involves collecting and storing the basic units of speech `[O' Shaugnessy 1984, Lukaszewicz et al 1987]. The basic units in the text to be synthesized are ~identified and the stored basic units are concatenated to get the speech output. In synthesis by rule ~method, a set of rules operate on the given text to produce a parameter sequence which is then d~synthesized. The rules are for the generation of basic speech units (acoustic -phonetic knowledge) ~and for the naturalness of the synthesized speech (Allen 1976).

Materials and Methods

Characters are more compact than phonemes, but their sounds vary with the place of occurrence in a word, especially for non-phonetic languages. Whereas for phonetic languages, each character has a unique script and pronunciation. Since Tamil is phonetic in nature, character could be an appropriate choice as basic unit, which we have chosen. A character in Tamil can be any one of (i) consonant-vowel sequence (CV) (ii) consonant alone (C) (iii) vowel alone (V). Cluster characters such as CCV and CCCV etc can be generated by combining appropriate Cs and a CV.
The other advantages in choosing the character as the basic unit are

'? The number of characters are not so large (around 250 in Tamil). '? Consonant to vowel transitions are preserved in the characters itself. '? characters have a more natural pronunciation than the phonemes, diphones, etc.

Model of speech production

Speech sounds are produced as a result of acoustic excitation of the human vocal tract. When the vocal tract is excited by a series of nearly periodic pulses generated by the vocal cords, the sounds produced are termed as voiced sounds. All the vowels come under this category. When the excitation I is produced by air passing turbulently through constrictions on the vocal tract, the sounds produced - are called unvoiced. The consonants like \pa\ & \sa\ come under this category. The speech production model has been developed with an assumption that the variations in the vocal tract shape with time can be approximated with sufficient accuracy by a succession of stationary shapes. These set of stationary shapes of the vocal tract are modeled as discrete time varying digital filters. With this basic idea of speech production model, the various acoustic parameters can be understood well. The acoustic parameters and their relation to the human speech production model are described below.

? Pitch : It is the fundamental frequency, ${F_0}$ of the vocal cords. The typical values are

50-200 Hz for men, 150-300 Hz for women and 200-400 Hz for children.

? Formants: These are the resonant frequencies of the vocal tract. Their values depend on what is uttered. Only the first three formant frequencies are perceptually significant and they lie below 3000 Hz.

? Bandwidths: Bandwidths determine the concentration and distribution of energy near the formant frequencies in the spectra of the signal.

? LPC : These are the linear predictive coefficients, i.e, filter coefficients of the all-pole model of the vocal tract.

? Gain : Energy of vibration of the vocal cord.

Construction of the basic unit database

The basic unit database has to be developed with great care as the quality of the synthetic speech 'predominantly depends on the quality of the basic units. Certain guidelines need to be followed in developing the database. They are as follows.

? The basic speech unit has to be embedded in some carrier word which needs to be meaningless in ~order to avoid any prosodic bias.

'? The carrier word should be so chosen that the basic unit of interest is easily separable *om it. The motivation behind extracting the basic units from some carrier word rather than uttering it in isolation is that, we can use the duration of the basic unit extracted from the carrier word as the default length for synthesis. The carrier words recorded are digitized at a sampling rate of 10 KHz. Waveforms of the basic speech units are carefully separated from the carrier words using an interactive audio

•editor package, Goldwave.

|Pitch extraction

'We have used the simplified inverse filtering technique (SIFT)} for pitch extraction (Markel 1972). ! The speech waveform s (t) is first filtered by a low pass filter with a cut-off at 0.8 KHz. After sampling

•The filter output at 2 kHz, the first five terms of the short-term autocorrelation sequence are calculated ~for an appropriate input length (corresponding to 32 msec) of data. Then a set of linear equations are solved for inverse filter coefficients a_i. Knowing a_i, the inverse filter output y_n can be calculated. The output autocorrelation sequence r n is then calculated as the autocorrelation sequence of y_n. Atter r n is obtained, the largest peak within specified limits is Found. If the peak is above 'a threshold (0.4), then that frame is classified as voiced and the Reciprocal of the location of the peak is the pitch, F_O. If the peak is less than the threshold, the corresponding frame is classified as unvoiced, with F_O =0. The advantage of the SIFT algorithm over other algorithms is that, the voiced ~and the unvoiced regions are satisfactorily classified with the same threshold for all the frames.

Linear prediction coefficients and Gain computation

~We have used the Levison-Durbints recursion algorithm to extract the linear prediction coefficients •and gain for each frame of the data. Linear prediction synthesis There are two parametric models of ~speech synthesis. They are

? The linear prediction synthesis. "7 The formant synthesis

The linear prediction (LP) method (Makhoul, 1975) is based on the speech production model, where the vocal tract is modeled as an all-pole digital filter of order 'p' and gain, G. Here every 'speech sample is computed as the linear combination of a certain no of previous speech samples and a sample of the excitation signal. The advantages of the LP model are the following: It can represent fairly well most of the speech sounds except nasals and voiced fricatives. '? Efficient methods for automatic extraction of linear prediction coefficients are available. The disadvantages are the ~following: ? It cannot model zeros in the speech spectrum and hence it does not model speech sounds.like nasals satisfactorily. '? It is very rigid against spectral level modification. Because of these disadvantages, adjustments of peak frequencies (formants) in the spectrum are not possible. The second disadvantage is very severe making it impossible to incorporate the knowledge due to coarticulation 'which modify the spectral peaks contextually. The formant synthesis requires the estimation of bandwidths from the spectrogram of the signal. The spectrogram required for the accurate ~measurement of bandwidth should be of very good resolution. Since we do not have the necessary ~equipment to get a very good resolution spectrogram, we could not use the formant synthesis.

Excitation source

The quality of the synthetic speech generated by these parametric models depends on the excitation~ signal used as input. The excitation signal should be such that its spectrum is white. This is because,' it is easier to shape the white spectrum according to a model, compared to a colored spectrum. Based~ on the type of the sound to be synthesized, the excitation signal should be either a white noise. Sequence (for unvoiced sounds) or periodic pulses a pitch period apart (for voiced sounds). We have' used the Fant's excitation model for excitation signal generation (Childers, 19). In this model, the' energy is distributed over one pitch period. Hence the synthesized speech obtained using this type; of excitation is significantly better compared to the speech generated using impulse excitation. The~ parameters of the model are T1 =50% of pitch period and T2 = 12% of pitch period.

Knowledge sources for naturalness

Mere concatenation of the signals corresponding to the basic units of speech does not produce' intelligible and natural sounding speech. To produce speech from a given text, human beings use~ several knowledge sources such as phonetics, phonology, morphology, syntax, semantics and' pragmatics (V.R. Ramachandran, 1991). It is necessary to incorporate these knowledge sources in a~ suitable form in the system to accomplish the same task. Human speech is characterized by segmental ~ and suprasegmental features which collectively contribute to the naturalness of speech. A segment' rel`ers to some small chosen unit of speech (eg phonemes, syllables etc). Segmental features refer to I those, which decide the phonetic quality of the segment. Suprasegmental or prosodic features have i their domain extended over more than

one segment. The suprasegmental features are influenced by~ factors such as phonetic and syntactic context, semantics and emotional state of the speaker. In' continuous speech, the segmental features are subjected to changes decided by phonetic context. These ~ changes are caused by the co articulation effect and give rise to certain joining patterns between' adjacent speech units. One needs to acquire the knowledge pertaining to these features from natural specch. Their incorporation into the system makes the synthesized speech sound intelligible and natural. Some of the knowledge sources identified here are co articulation, intonation and duration. ~ Intonation in Tamil

j

With the knowledge of the general declining tendency of the pitch contour of declarative sentences ~ (A.S.Madhukumar et al,1991), we have developed a model to synthesize intonation pattern for simple I declarative sentences. The model is developed based on the observations from a set of 25 declarative sentences read out by an adult, male, native speaker of Tamil. The model assumes that? F_0 for~ any simple declarative sentence fluctuates between two abstract lines, a base line and a top line. ?

The top line is determined by the value of the first peak of the first word and the last peak of the~ last word of the sentence.

? The base line is determined by the valley following the first peak of the first word and the valley of the final word of the sentence.

? The intonation contour is characterized by the local falls and rises of F_0 based on the stress on the words of the sentence.

The F 0 values of the valleys and the peaks in each word are computed from the corresponding values taken from the base line and the top line respectively. The intonation knowledge modifies the default pitch values of the basic units. The nature of the underlying sentence (declarative, interrogative, yes/no type, etc.) is determined and based on that, the corresponding intonation is provided to the

Pitch contour, prior to synthesizing.

Incorporation of duration knowledge

Studies have been made to determine the change of duration of the basic units in different contexts (B.Yegnanarayana, 1994). From these studies, rules are formulated to modify the duration of the basic units during synthesis, depending on the context in the input text. Some of the rules are the. Following.

? Positional effect: The duration of a character at the beginning of a word is increased by 10% 736 and that at the final position, is increased by 30%.

.

? Prepausal lengthening effect: The duration of a character appearing before a pause is increased. If the pause is due to a punctuation like comma, then the final character of the word preceding the pause is increased by 35\%. If the pause is due to sentence ending, then the increase will be by 35\%. If the pause is due to any other syntactic boundary, then the duration of the character will be increased by 30%.

The base duration, D_O, of the basic unit is taken from the carrier word. Since the carrier word is spoken in isolation, the actual base duration in continuous speech is about 70% of the corresponding D_O. Therefore, the actual base duration, D for each basic unit is taken to be 70% of D_0. Now, depending upon the context of each basic unit, various rules may be applied. Each of these modifies the base duration, D for that particular unit using the expression D_{final} =D + alpha*D/ 1 00 where alpha is the percentage value specified in each rule. The basic units are synthesized for D_{final} duration. The duration knowledge does not modify any of the other parameters used for synthesis.

Conclusion

The present speech synthesis system for TtS conversion in Tamil is successful in generating al] vowels, all CV units except those that have the consonants \r/, \vl & \p/. The consonants have not been synthesized properly owing to their very short duration. Words without consonants have been synthesized with a reasonable amount of intelligibility using the present speech synthesis system. Attempts have also been made to synthesize some declarative sentences. Proper pitch contour has been provided to the declarative sentences synthesized, in order to give proper intonation.

The main contributions of this work are the following:

1. The basic units of speech that are suitable for speech synthesis in the context of TtS Conversion is collected from natural speech.

2. Some knowledge sources that account for the naturalness of speech are identified and acquired.

3. A representation scheme is proposed for the basic speech units which is flexible enough to

4.
Allow incorporation of the knowledge sources.

Demonstration of a limited vocabulary speech synthesis system for a TtS conversion system in Tamil, incorporating some of the rules that represent the knowledge sources.

Text to Speech conversion has been one of the focus areas of research for the speech scientists for over three decades. Currently available TtS conversion systems in English, that are capable of producing highly intelligible speech, are the result of this research. The foci of research have only been English, French, German and Hindi. As far as our knowledge goes, no work has been done in developing a TtS conversion system for Tamil. So, this is the first attempt in this direction.

References

1. B. Yegnanarayana, "Formant Extraction from linear-prediction phase spectra," J. Acoust. Soc. Am., Vol. 63, No. 5, pp. 1638-1640, May 1978.

Hema A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," Signal Processing, Vol. 22, pp. 259-267, 1991.

Dennis H. Klatt, "Review of text-to-speech conversion for English," J. Acoust. Soc. Am., Vol. 82, No. 3, pp. 737-793, May 1987.

4. B. Yegnanarayana, S. Rajendran, V. R. Ramachandran and A. S. Madhukumar, "Significance of knowledge sources for a text-to-speech system for Indian languages," Sadhana, Vol. 19, Part 1, pp. 147-169, Feb. 1994.

5. Eric Moulines and Francis Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication, Vol.9, pp. 453-467, 1990.

6. Francis Charpentier and Eric Moulines, "Text-to-speech algorithms based on l;Fl synthesis," Proc. ICASSP'88, pp. 667-670, 1988.

7. Bishnu S. Atal and Nancy David, "On synthesing natural-sounding speech by linear prediction," Proc. ICASSP'79, pp. 44-47.

8. S. E. G. Ohman, "Coarticulation in VCV utterances: spectrographic measurements," pp.; 151-
168, 1965.

9. A. S. Madhukumar, S. Rajendran, and B. Yegnanarayana, "Intonation component of a text-to- speech system for Hindi," Computer Speech and Language, Vol. 7, pp. 283-301, 1993.
t`

10. Andrew Varga and Frank Fallside, "A technique for using multipulse linear predictive speech synthesis in text-to-speech type systems," Proc. ICASSP'87, pp. 586-587.

11. D. G. Childers and H. T. Hu, "Speech synthesis by glottal excited linear prediction," J. Acoust. Soc. Am., Vol. 96, No. 4, pp. 2026-2036, Oct. 1994.

12. N. Yiourgalis and G. Kokkinakis, "A TtS system for the Greek language based on concatenation of formant coded segments," Speech communication, Vol. 19, pp. 21-38, 1996.

t`

13. Jonathan Allen, "Synthesis of speech *om unrestricted text," Proc. IEEE, Vol. 64, No. 4, pp. 433-442, Apr. 1976.

14. J. D. Markel and A. H. Gray, "A linear prediction vocoder simulation based upon the | autocorrelation method," IEEE Trans. ASSP, Vol. No. pp. 124-134, Apr. 1974.

15. D. H. Klatt, "Software for a cascade/ parallel formant synthesizer," J. Acous. Soc. Am., Vol.
67, No. 3, pp. 971-995, 1980.

16. B. S. Atal and S. L. Hananer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acous. Soc. Am., Vol. 67, No. 3, pp. 637-655,-1971.

17. J. D. Markel, "Digital inverse filtering - a new tool for formant trajectory estimation," IEEE~
Trans. Audio, Electroacoust., Vol. 20, No. 2, pp. 129-137, June 1972.

18. Noriko Umeda, "Linguistic rules for Text-to-Speech synthesis," Proc. IEEE, Vol. 64, No. 4,' pp. 443-451, Apr. 1976.

19. J. D. Markel, "The SIFT algorithm for fundamental *equency estimation," IEEE Trans. Audio,' Electroacoust., Vol. 20, No. 5, pp. 367-377, 1972.

20. J. P. Olive, "Mixed spectral representaion-formants and linear predictive coding, J. Acoust.- Soc. Am., Vol. 92, No. 4, pp. 1837-1841, 1992. 1

TAMIL IN CYBERSPACE

Ramalingam Shanmugalingam (Appu Archie)
36, Farrell Court, Marblehead, MA 01945 (USA)

Courage is the keyword that comes to mind as one ponders about the rapid changes that are taking place before our eyes, and ears. One needs courage to face new situations that at the outset may be perplexing and signal deep changes in our ways of life. Cyberspace to put it simply is, ``The virtual space created by Computer Systems. There are many numbers of questions that may arise out of making this world into a ``Global Village.Until such lofty ideas take any shape, life has to go on and we have to find ways to overcome fears of new challenges and opportunities that will serve to make our lives better.

Tamil, more than anything-else, is the mother tongue of more than 70 M people. The tradition and culture of the Tamils allow them to mix and assimilate easily with other ethnic groups and yet maintain their identity in tact in most cases. Absence of a free Tamil country has begun to show signs of loss of Tamil identity. Nevertheless, the enthusiasm and interest in Tamil shown by a few Tamil expatriates in the so called Developed countries is a good omen to future retention of Tamil interest via Cyberspace as well. It becomes necessary to ignore the realities of most Tamil lives in developing countries, as economic conditions will dictate a more somber attitude towards investment in more sophisticated technological advances when there is still the hang-ups about basic human needs. Thus hampering much needed technological advances. In the meantime, we should not feel increasingly isolated, dependent on others, apparently without needed resources to find a lifestyle we deserve. The days of a rigid agrarian society are a distant memory and happily so, even though a larger population is dependent on Agri based industries.

Technology proved to be a major gift in changing the rigid society in the last 50 years into a society open to its promise and given the environment, learn well enough to use technology properly. But unlike the Americans. Tamils should not replace a solid based hardware industrial complex with a more abstract cyberspaced virtual reality type of bubble economy. It is said that, that the most valuable corporation in the US and the world is Microsoft, the software company. If the world economy is dominated by `information' companies, financial shocks are bound to happen and this may lead to the inevitable

Big church some scientists predict will happen sooner than later. In short we should develop our own software that we can use to make our life better but not to spell disaster by depending on others. Computers can break through barriers in life and connect with other people. Computers are beginning to bring the whole to us. It is true that many of us are threatened by computers and all the mysterious things connected with it. The car is a complicated piece of equipment and we are not threatened by its ability to function with the least effort and knowledge of it by us. Yet its dangers on the road are not minimized by our affinity for the car. Similarly, the computer in addition to being a hobby can also cut down distance and time to

bring people closer in spirit through esprit de corps. The computer can send messages in numbers unlike the telephone, where it is a one-on-one mode of communication. Today, we can face each other though in virtual reality as photo-images, discuss and decide matters as if we are physically present together. Also unlike TV with its limited channels, time and distance of access, the INTERNET is almost an endless ocean of data in Cyberspace, where much of the world's information can be found. I am no Technophile, nor a neophyte, on computers. I am learning to be an expert in the use of a computer, and leave the web of designing and developing the more complicated technology to experts. There still are some important useful people in the USA who claim to live out their days Computer-free.I have found the computer to be a source of inspiration to learn more, communicate better, and less strain on the purse. It will be even more economical, if its use is widespread and in native languages. It is this desire that prompted me to get involved in the design and development of a computer tool for Tamil. I was disappointed in the many English translations of THIRUKKURAL

Most translations appeared to me as either paraphrases or non-couplets English according to Richard Lederer, in his book Miracle of English as a preferred second language outnumbers those who use it natively. English is first or official language of 45 countries. Thirukkural, in its native Tamil is unmatched for brevity, the soul of wit.Yet, English with its large vocabulary, burrowed or added to its dark age low Germanic tongue in the mid 5th century, through the Anglo Saxon Old English until William Duke of Normandy led Normans invade England in 1066, conquered Saxons and Danes. Duke William became King of England. Free Mix of French words into Old English gave way to Middle English.European renaissance between the 14th to 17th centuries began as educated people rediscovered the world of ancient Greece and Rome. Love affair for anything classical led to taking of Latin and Greek words that could describe new discoveries in Medicine, Art, Science, Geography, etc., These classical words entered into English through writing-often scholarly writings.It is estimated, that there are more than one million words in the English language although only about 615,000 have been listed in the Oxford English Dictionary. Poet Carl Sandberg said, The English Language hasn't got where it is by being pure Naturally. I began to wonder, why, that THIRUKKURAL that can claim to biblical usefulness, has no English version in couplet form e.g. `perukkattu vEnhdum panhital;

Humility in Prosperity
Dignity in Adversity.

English by Appu Archie *I am using and will use Yarzhan Key to pronunciation when using English Phonemes in all my writings. An Alphabet Chart in Roman Script for Tamil is given below: I had difficulty typewriting Tamil in computers before 1993. Whatever software or Fonts available for Tamil then were based on the Typewriter Keyboard. One look at the Keyboard Matrix convinced me that there must be a better way to make Tamil typewriting with computers easy and logical, a prerequisite for any Tamil software. My engineering background and love for my mother tongue prompted me to look for a Tamil answer to a word processing problem in Tamil depending on a Western or English origin computer for Tamil. Every language has its share of Literature, Music and Drama, but Tamil is perhaps the only language that has given

language a life and body, namely `ujir and mej' and all Tamil letters are either life or body and a combination of body and life give body with life letters or `ujirmej ezuttukkaL.'Tamil has 247 letters, depending on 31 Basic Letters thus: Vowels - `ujir', `aA, iI,, uU, eE, y, oO, w(av)(Of the 12 `ujir' 5 are short and 7 are long vowels.)Velar Fricative - (<$Esymbol 1Consonants - `mej'18

Total Basic LettersEach Consonant in combination with each Vowel can produce a third letter or a dependent letter called `ujirmej.' 18 Consonants and 12 Vowels produce Dependencies -

Total number of Tamil 247 The 5 pairs of short and long vowels can be represented by 5 upper and lower cases of the QWERTY Keyboard

Number of Keys required to represent 31 basic letters is (31-5)s26

Please refer to the CPD- YARZHAN TAMIL-EDITOR CHARACTER MATRIX below:

Obviously, Tamil language with its renown antiquity falls easily into modernity by being easily represented by the 26 Keys of the QWERTY Keyboard without exception. There is one shortcoming in using all 26 Keys to represent Tamil, as any transliteration will not be accurate, since letters represented by sounds not in Tamil will give rise to confusion.

நிறைவு விழா சிறப்புரை
தமிழக முதல்வர் கலைஞர் மு. கருணாநிதி

நேற்றும் இன்றும் தமிழ் அறிஞர்கள் ஒன்றாக கூடி முடிவுகளை அறிவித்திருக்கிறார்கள். இங்கு பேசிய மாறன் குறிப்பிடுகையில் இரண்டு தமிழர்கள் ஒன்றாக கூடி ஒருமுடிவை எடுக்க முடியுமா? என்ற ஐயப்பாட்டை சொல்லி கேள்விக்குறியாக விட்டுச்சென்றார். மேலும் நான் ஏமாறப்போகிறேன் என்று குறிப்பிட்டார். நான் ஒருபோதும் அறிஞர் பெருமக்களிடம் என்றுமே ஏமாறமாட்டேன் என்று திட்டவட்டமாக தெரிவித்துக்கொள்கிறேன்.

இங்கே எடுக்கப்பட்ட முடிவுகள் அதற்கு கிடைத்த ஆதரவு, அறிஞர்கள் வரவேற்பு இவற்றையெல்லாம் பார்க்கும் போது நான் முதல்அமைச்சராக கோட்டையில் நாற்காலியில் அமர்ந்திருக்கும் மகிழ்Óசியைவிட இÓங்கே அறிஞர்கள் மத்தியில் அமர்ந்திருப்பது வெறும் நாற்காலியில் அமர்ந்திருப்பதாக கருதவில்லை கோபுரத்தின் உச்சியிலேயே அமர்ந்திருப்பதாக உணர்கிறேன். தமிழர்கள் ஒன்று கூடி ஒரு நல்ல முடிவை எடுத்திருப்பது இதுவரை உள்ள ஏழு அதிசியங்களுக்கு அப்பாற்பட்டு எட்டாவது அதிசியமாக தோன்றுகிறது.

இந்த மாநாட்டினால் என்ன பயன்? இந்த கருத்தரங்கு என்ன சாதிக்கும்? என்றெல்லாம் கேள்விகள் எழுந்தன. இந்த முடிவுகளுக்கு பின்னும் பல்வேறு விமர்சனங்கள் எழும். கிணறு வெட்ட புÛதம் புறப்பட்டதுபோல என்று ஒரு பழமொழி உண்டு. ஒரு கவியரங்கில் இப்பழமொழிக்கு ஒரு புதிய விளக்கத்தை நான் கொடுத்தேன். அதாவது, கிணறு வெட்ட மண் குவியும் என்பது முதலாவது புÛதம், வெளÕவரும் அடைபட்ட காற்று இரண்டாவது புÛதம், பொங்கிவரும் நீர் மூன்றாவது புÛதம், நீர் பிரதிபளிக்கும் வானம் நான்காவது புÛதம், ஆதவனிÓன் கதிரால் நீர்தெளÕவடையும் என்பது ஐந்தாவது புÛதம். கிணறு வெட்ட இந்த ஐம்புÛதங்கள் கிளம்பின என்று சொன்னேன். இங்கே அறிஞர்கள் கிணறு வெட்டினார்கள். ஐம்புÛதங்கள் புறப்பட்டன.

சாம்பிட்ரேடோவும் டத்தோ சாமிவேலும் பேசும் போது இந்தமாநாடு உரிய நேரத்தில் உரிய இடத்தில் நடக்கிறது என்று குறிப்பிட்டார்கள். தமிழை பொருளாதார மொழியாக ஆக்கவேண்டும் அதற்கு உலக அங்கÑகாரம் வேண்டும், உலகம் முழுவதும் தமிழ் பயன்படும் மொழியாக ஆகவேண்டும். தமிழ் இணையம் சிறப்பாக செயல்படவேண்டும் என்றார் அந்த அடிப்படையில் தான் முடிவுகள் எடுக்கப்ட்டன.

கணினி இணையம் இரண்டின் பயன்பாடும் பெருகிக் கொண்டிருக்கும் இந்தவேளையில் தமிழர்கள் அவற்றிற்கு அந்நியமாகி விடக்கூடாது. அவற்றை பயன்படுத்த பழகவேண்டும் என்று ராமசாமி சிதம்பரம் பிள்ளை அவர்கள் கூறியதன் அடிப்படையிலும் முடிவுகள் எடுக்கப்பட்டுள்ளன. சிங்கப்புÛரில் இந்த தமிழ்நெட் கருத்தரங்கிற்கு விதை போடுவது போல் தமிழ் நெட் 97ஐ நடத்திய கோவிந்தசாமி அதை தொடர்ந்து இந்த மாநாட்டிலும் பங்கேற்று இதை வரவேற்றது குறித்து மகிழ்ச்சியடைகிறேன்.

இந்த மாநாட்டில் திரு அனந்த கிருஷ்ணன் அவர்கள் திட்ட குழுவும் இணைந்து செயற்படும் திட்டங்களை உருவாக்கியுள்ளார்கள். உலகளாவிய தமிழ் சமுதாயம் உருவாக்கிய திட்டம் என்பதில் நான் பெருமிதம் கொள்கிறேன்.

தமிழ் மொழியின் தனித்தன்மை குறையாமலும் அதை மேம்படுத்தும் வகையில் முடிவுகள் எடுக்கப்பட்டுள்ளன. விசைப்பலகை பிரச்னை குறித்து ஒத்த கருத்து எடுக்கப்பட்டுள்ளது இது குறித்த சிலருக்கு அதிருப்தி இருக்கலாம். ஏற்கனவே பழகிய விசைப்பலகையில் இருந்து மாறுவது என்பது கடினம் தான். நான் விசாரித்த வகையில் பரிந்துரைக்கப்பட்ட விசைப்பலகைக்கு பயிற்சி எதுவும் பெரிதாக

தேவையில்லை இரண்டு மூன்று நாட்கள் பயிற்சி பெற்றால் போதும், The best and efficient என்று Phonetic விசைப்பலகையை அறிவித்திருக்கிறார்கள் தனியார் துறையும் பெருமளவில் அவற்றை தயார் செய்ய ஒப்புக்கொண்டுள்ளது.

எங்களுக்கு எதையும் வலுக்கட்டாயமாக திணிப்பதில் நம்பிக்கை கிடையாது. வெளÕநாட்டில் வாழுகிற தமிழர்களுக்கு ஆங்கிலம் தான் தெரியும். தமிழ் பேசத் தெரியும் ஆனால் எழுதத்தெரியாது அதனால் அங்கே romanized விசைப்பலகை உகந்தது என்பதால் அதில் நாங்கள் குறுக்கிடவில்லை.

ஆடுத்தது coding system எனப்படுகின்ற குறியிட்டு முறை. நாம் இன்று இடஒதுக்கூட்டிற்காக போராடிக்கொண்டிருக்கிறோம் இந்த குறியூட்டு முறையிலும் இடஒதுக்கூட்டு பிரச்சனை எழுந்திருக்கிறது. இப்போது 8bit முறைப்படி ஒவ்வொரு மொழிக்கும் 256 characters ஒதுக்கப்பட்டடிருக்கிறது. விசைப்பலகையில் உள்ள control keys, Number keys, போன்றவைகளுக்கு 64 இடங்கள் ஒதுக்கப்பட்டுள்ளன. எஞ்சியிருப்பது 192 இடங்கள்தான் ஆனால் நமக்கு தேவையோ 512 இடங்கள் எனவே இப்பிரச்சனைக்கு இப்போது தÞர்வேற்படாது. இன்னும் இரண்டு மூன்று ஆண்டுகளில் 16 bit system ஆக மாற்றப்பட்டால் 65000 இடங்கள் கிடைக்கும் அப்போது 512 இடங்களை எளிதாக பெற்றுக்கொள்ளலாம் என்று எனக்கு சொல்லியிருக்கிறார்கள். œஎனக்கு தெரியாது அவர்கள் சொன்னார்கள்œ.

International unicode consortium தில் உறுப்பினராக தமிழகம் தÞர்மானதிÕரூக்கிறது. சிங்கப்புÛர், மலேசியா, இலங்கை போன்ற நாடுகள் ஏற்கனவே அந்த அமைப்பில் உறுப்பினர்களாக இருக்கின்றன. எனவே பிற்காலத்தில் இதன் தொடர்பாக நமக்கு பிரச்னை ஏற்படும் போது அவர்கள் துணைபுரிவார்கள் என்பதில் எனக்கு சந்தேகமே இல்லை.

குறியூட்டு முறையில் தற்போது monolingual பரிந்துரை செய்யப்பட்டுள்ளது தமிழ் ஆங்கிலம் இரண்டிலும் வேண்டுமெனில் சிறிய மாற்றம் செய்தால் போதும் bilingual முறையை பயன்படுத்தலாம். ஆனால் இந்த முடிவுகள் கல்லால் செதுக்கப்பட்ட மாற்ற முடியாத முடிவுகள் அல்ல. உலகில் நிலையானது என்பது எதுவென்றால் œமாறுதல்œ மட்டும்தான் என்று கார்ல் மார்க்ஸ் சொன்னார்.

இப்போது கொடுக்கப்பட்டிருப்பது வெறும் டிராப்ட்தான். தமிழர்கள் இதைப்பயன்படுத்தி இதிலுள்ள நிறைகுறைகளை சுட்டிக்காட்டலாம். சிங்கப்புÛர், மலேசியா போன்ற பல்வேறு நாடுகளில் உள்ள அறிஞர்களை கொண்டு குழு அமைத்து மாற்றங்களை ஆராய்ந்து இந்த முடிவுகளைப் பற்றி மேலும் ஆய்வுகள் செய்யப்படும். இறுதி முடிவுகளுக்கு இது போல மாநாடு நடத்தவேண்டிய அவசியம் இல்லை. விடியோ கான்ஃபரன்சிங்Õ மூலமாகவே நான் இங்கிருந்தும், மலேசியாவில் டத்தோவும், இலங்கையில் தொண்டைமானும் சேர்ந்து இறுதிமுடிவெடுத்துக்கொள்ளலாம். மே மாதத்திற்குள் இப்போது எடுக்கப்பட்ட முடிவுகளைப் பற்றிய கருத்துக்கள் இறுதிமுடிவு எடுக்கப்பட்டு செயல்படுத்தப்படும்.

டத்தோ சாமிவேலு அவர்கள் பேசும் போது virtual tamil world உருவாகவேண்டும் என குறிப்பிட்டார்கள். அதன் அடிப்படையில், ஒரு விர்சுவல் பல்கலைக்கழகம் தொடங்க எண்ணியுள்ன் அதற்கு உங்கள் அனுமதியுடன் œஉலக தமிழ் இணையப் பல்கலைக் கழகம்œ என்று பெயரிட்டுள்ளேன்.

தமிழில் மென்பொருள் உருவாக்கத்திற்கான தேவை அதிகமாக இருக்கிறது. மென் பொருள் வளர்ச்சிக்காக தமிழ் மென் பொருள் நிதியம் œ(Tamil Software fund)ஒன்று உருவாக்கப்படும்.

Web browsing, voice processing, handwriting processing. Optical character recogniton இவற்றை தமிழிலேயே உருவாக்க முயற்சிகள் எடுக்கப்படும்.

சென்னையில் உள்ள அண்ணா பல்கலைக்கழகம் கோவை பாரதியார் பல்கலைக்கழகம், திருச்சி மண்டல பொறியியல் கல்லூரி ஆகிய கல்வி அமைப்புகளில் மூன்று ஆய்விருக்கைகள் ஏற்படுத்தப்படும்.

தமிழில் இன்டர்நெட் ஆய்வு மற்றும் வளர்ச்சிக்காக 5 கோடி ரூபாய் நிதி ஒதுக்கப்பட்டுள்ளது.

இந்த தகவல் தொழில் நுட்பப் புரட்சி பாமர மக்களை அடையவேண்டும் என்பதுதான் எங்கள் நோக்கம்.

Electronic Government அமைய இந்த மாநாடுதான் முதல் படி அப்போது இப்போது இருப்பது போல ஒரு transparant Government அதாவது ஒளÕவுமறைவு இல்லாத வெளÕப்படையான அரசாங்கம் அமையும். இன்னும் ஓராண்டிற்குள் தமிழகம் முழுவதும் 1000 சமுதாய இன்டர்நெட் மையங்கள் அமைக்கப்படும். அரசு துறையில் உள்ள எல்லா படிவங்களும் தமிழ், ஆங்கிலம் என இரண்டு மொழியிலும் கிடைக்கும்படி செய்யப்படும். இங்கு ராமசாமி சிதம்பரம் பிள்ளை குறிப்பிட்டதுபோல தமிழ் Tamil Global village is in the making என்று குறிப்பிட்டார். அதையே நானும் குறிப்பிட விரும்புகிறேன். அந்த(Tamil Global Village) உருவாக அனைவரும் பாடுபடவேண்டும்.

உலக தமிழ் இணையக் கருத்தரங்க மாநாடு

தமிழ் நெடுங்கணக்கில் புதியன புகுதல்

தமிழில் ஒலி??யர்?பு: Ø?Ôதிய வரிவடிவங்கள் இன்Ù?யும் யூனிகோட்(?) கண்ணோட்டத்தில் தீர்வுக்காண முயற்சியும்

செ. செந்தில் நாதன்
துணை ஆசிரியர், இந்தியா டுடே
சென்னை

?Õழிவு

தமிழ் நெடுங்கணக்குக்கான எழுத்துரு குறியீட்டு அÙ??பும் விசைப்பலகை அÙ??பும் தரநெடுத்தநெடும் இந்த தருணத்தில், தமிழ் நெடுங்கணக்கிலேயே சில புதிய வரிவடிவங்களை சேர்க்கவேண்டும் என்று இந்த ஆய்வுரை கோருகிறது. பிறமொழிகளிலிருந்து ஒலி??யர்க்கும் போது சில ஒலியன்களுக்கான துல்லிய வரிவடிவம் தமிழில் இல்லாததை சுட்டிக்காட்டும் இந்த ஆய்வு அவற்றுக்குரிய புதிய வடிவங்கள் உருவாக்கப்பட்டு அவையும் தரநெடுத்தப்பட வேண்டும் என்கிறது.

g. ch,d,dh,b,x,z, ஆகிய(ஆங்கில) ஒலியன்களுக்? இணையாக, அவற்றுக்கு தற்போது பயன்படுத்தநெடும் எழுத்துகளின் ?Ö்சியாக கூடுதல் வரிவடிவங்களை உருவாக்க முனையும் எழுத்துக்களின் முக்கியத்துவம் குறித்து தகவலியல் கோட்பாடுகளின் வழியாக சில முடிவுகளை ஆய்வு முன்வைக்கிறது.

œœமாற்றுவதும் திருத்துவதும் யாருக்கும் எதற்கும் இழிவாகவோ குற்றமாகவோ ஆகிவிடாது. Ø?ன்Ù?யÙ?வும் காலத்தோடு கலந்து ?சல்லவும் எதையும் மாற்றவும் திருத்தவும் வேண்டும். மொழி என்பது உலக? Ø?Ôட்டி? Ø?Ôராட்?த்துக்? ஒரு Ø?Ôர்க்கருவி. Ø?Ôர்க்கருவிகள் காலத்துக்கு ஏற்ப மாற்றப்படவேண்டும்.œœ

- தந்தை? ?குரியார்

ஒலி??யர்?பும் தமிழ் இலக்கண?ரபும் - ஒரு ?Öள்பார்வை

தமிழ் பல காலகட்டங்களில் சமஸ்கிருதம், ?Ôலி, ?Õராகிருதம், அரபு, லத்தீன், ஆங்கிலம், இந்தி Øமான்று பல மொழிகளோடு தொடர்புகொள்ள Ø?ரிட்?து. அவற்று?ன் சொல், கட்?Ù??பு ரீதியிலான ?கால்வினை ?காடு?மானைகளையும் Ø?ற்?கா?டிருக்கிறது. இந்த

வரலாற்று மான்னணியில், ஒலி ??யர்?பு பிரச்னையை தமிழ் எப்படி கையாண்டது என்பதை சுருக்கமாக பார்ப்போம். ஒலி??யர்?பு குறித்த தமிழ் ?ரÙ? அறியாமல் அதை மாற்றுவதோ ?Öறுவதோ இயலாதது. முறையற்றது.

வடமொழியிலிருந்து சொற்களை தமிழுக்கு கொண்டுவரும் போது அதை எப்படி எழுத என்பது குறித்து தொல்காப்பியர் காலத்திலேயே ஒரு மரபு இருந்திருக்கிறது.

œœவ?சொற் கிளவி வ? ?வழுத்?தார°இ எழுத்?தாடு புணர்ந்த சொல்லா கும்Ø?œœ

என்ற நூற்பாவின் மூலம் தொல்காப்பியர் கூறிய வரம்மான் அடிப்படையில் தற்சகும்,தற்?வம் ஆகிய விதிகள் உருவாகின. இந்த இரண்டுவிதிகளின் மூலமும் தமிழில் வ?சொல் ஒலி??யர்க்கநெடுகிற போது வடமொழிக்கான வரிவடிவம் கலக்காமல் சேர்த்துக்கொள்ளமாட்டாது. இணையான வரிவடிவங்கள் இருக்கும்போது அப்படியே மாற்றிக்கையாளலாம் என்பது தற்ச? முறை (உதாரணம் :?ங்?கும்).

இல்லாத போது தமிழில் மாக?ஒருங்கிய, விதிக்க??ட்? எழுக்களை பயன்படுத்த வேண்டும் என்பது தற்?வம் (உதாரணம்: ?க்ஷம் - ?க்கம்).

?Õற்காலத்தில் தமிழகத்தில் சமஸ்கிருதம் Ø?லாதிக்கம் வகித்த போதும் மணிப்பிரவாள ?Ù? கோலச்சியபோதும் கூட ?ஒரும்மான்Ù?யான மக்கள் இந்த விதிகளையே பின்பற்றினார்கள். இந்த ?ரØ? ?Õற? அரபு, லத்°ன், ?Õற ஐரோப்பிய மொழிகளிலிருந்து ஒலி??யர்க்க??ட்?போதும் மான்குற்ற??ட்?து. உதாரணம்:செய்யது, முகம்?து, கிறித்து, ?Õ?ர?சு, இங்கிலாந்து, ?ல்லாந்தர் Øமான்ற ??யர்ச்சொற்கள்.

?ணி??Õரவாளம் மூலமாக தமிழ் உரை?Ù?யில் ?ல கிரந்த எழுத்துக்கள் பு?ந்தன. ஆனால் அவற்றில், இன்று ஜ, ஸ,ஷ,ஹ ஆகிய கிரந்த

எழுத்துக்கள் ?ட்டும் ?ஒரும்மான்Ù?யானவர்களால் ஏற்றுக்கொள்ள?ட்டிருக்கிறது. இந்த மான்? கிரந்த எழுத்துக்களும் தமிழில் ?Öடி??தற்? காரணம்,

இடபோது இவ்?வழுத்துக்கள் வடமொழி சொற்கள் எழுதுவதற்காக அல்லா?ல் ஆங்கில ஒலி??யர்?புகளை எழுதுவதற்காகவே மாகவும்

பயன்படுத்தநெடுகின்றன என்பதே. ஹிமாச்சல பிரதேஷை இடÔச்சல??Õரதேசம் என்று எழுதுவதை சங்க?மான்றி எற்றுக்?கால்கிற இன்றைய

??Ôதுவழக்? ஜ?மானை ச?மான் என்றோ ஆஸ்திரேலியாவை ஆசுத்திரேலியா என்றோ எழுதுவதை ஏற்றுக்கொள்ளாததன் ரகசியமும் இதுவே. இந்த எழுத்துக்களின் தேவை அதிகரித்ததால் அவை தொலைத்துவிட்டன.

ஆனால், சமஸ்கிருத நெடுங்கணக்கோடு ?ஒருங்கிய ?தா?ர்?Õருந்தும் ?Õறதிராவிட மொழிகள் அவற்றை அப்படியே ஏற்றுக்?கா?? போதும், குறைந்த ?ட்சம் க, ச,?, த,? ஆகிய வல்லினங்களுக்கான ஹகரம் ஏறிய, அதிர்வுÙ?ய, அதிர்வும் ஹகரமும் ஏறிய kha,ga, gha முதலான (வர்க்க) எழுத்துக்களை தமிழ் ஏன் ஏற்றுக்கொள்ளவில்லை என்ற கேள்வி எழு?????லாம். இதற்கான விÙ?களை சொல்வது இங்?கும் நோக்க?Õல்லை என்றாலும் ??க்?த் தேவையான ?Õரச்சனையோடு இதற்?ள்ள ?தா?ர்Ù? ?ட்டும் விளக்கமுற்?டுவோம். உதாரணமாக, தமிழ் க, ka, ga,ha ஆகிய ?Ôற்?றாலிகளை (allophones)?கா?டிருக்கிறது. எனவே œகœ வர்க்க எழுத்துக்களில் ga ஒலிக்குறி?Ù? தமிழ் œகœ ஓரளவுக்? ஈடு?ச?து வந்தது. Ø?லும் ஹகரØகுறிய aspirated ஒலியன்கள் தமிழ் இயல்புக்? ?Ôறானது என்று வி???ட்?ன. அத்துடன் வடமொழி சொல் வரவு ?தா?ர்மாக ?Õலவிவந்த தற்ச?-தற்?வ விதிகளை மாற்றும் அல்லது கைவிடும் தேவை எழவில்லை. எழுந்திருந்தால் புதிய வரிவடிவங்கள் முன்Ø? வந்திருக்கும். உதாரணமாக, F என்ற ஆங்கில ஒலியினை குறியீ? ஃ? என்ற புதிய கூட்??ழுத்து உருவாக்கிக் கொள்ள??ட்?தை? Ø?Ôல்.

இந்த ஒலி??யர்?பு வரலாறு இரண்டு முக்கியமான உ?Ù?களைச் சுட்டிக்காட்டுகிறது.

முதலாவதாக, தேவை ஏற்பட்டபோது - அதன் ?Õர்?ந்தம் அழுத்தியபோது - தமிழ் அயல்மொழி ஒலியன்களுக்காக புதிய வரிவடிவங்களை ஏற்றது.

இரண்டாவதாக, இந்த ஏற்பும் ஐரோப்பிய மொழிகளிலிருந்து ஒலி??யர்க்க வேண்டிய ?Õர்?ந்தம் வந்தØ?Ôதே இரவலாக ஏற்கப்பட்டது. ?Öடிக்ககாரணமானது.

எனினும் இந்த மாற்றங்கள் போதுமானவை அல்ல. க, ?, த, ? ஆகிய மான்? முக்கிய வல்லினங்களும் சொல்லுக்? ?டுவில் ??ல்லினத்தை ?தா?ர்ந்து வரும்போது ga,da,dha,ba ஒலிகளை குறிக்கின்றன. எனவேதான். London என்கிற சொல் ஒலி??யர்க்க??ட்? போது இயல்மாகவே துல்லியமாக ல??ன் என்று அÙ?ந்தது. என்றாலும் இதே ஒலிகள் ஒலி ??யர்க்க??ட்? சொல்லுக்? முதலிலும் இறுதியிலும் ????யழுத்தாக தனித்தும் வரும்போது துல்லியமாக அந்த ஒலிகளை சுட்??பயன்படுத்த முடிவதில்லை.(உதாரணத்துக்? Goliath, CAD) இதற்கான தீர்வு, இந்த ஏழு

ஒலியன்களுக்? சொல்லுக்? முதலிலும் இÙ?யிலும் இறுதியிலும் எந்த இடத்தில் வந்தாலும் ?ரே?Ôதிரி ஒலி?Ù?த் தருகிற வரிவடிவங்களை உருவாக்குவதே ஆகும்.இதற்கான தீர்வை நோக்கிய முயற்சியாக சில ஆலோசனைகளை விவாதத்துக்? வைக்கவிரும்புகிறேன்.

?Õலவும் வரிவடிவங்களின் ?Öட்சியாக புதிய வரிவடிவங்களை வருவித்தல் உலகத்தின் எல்லா மொழிகளிலும் உள்ள எல்லா ஒலிக்குறி?புகளுக்கும் தமிழில் எழுத்துக்களை உருவாக்கித் தீர்த்துவிடுவது என்?தல்ல இந்த ஆய்வுரையின் நோக்கம். ஆங்கிலத்தில் உள்ள எல்லா

ஒலியன்களுக்கும் தமிழில் இணைகளை உருவாக்குவதும் கூ? இதன் நோக்க?ல்ல. ?ன்னாட்?ளவில் ??Ôதுவாக ஏற்கநெடுகிற, ஒலியியல் ரீதியாக தரநெடுத்தும் செய்ய??ட்? சொற்களை ஆங்கிலம் வழி ஒலி??யர்க்கும் போது சந்திக்கிற ?குரிய பிரச்னையை தவிர்க்கவும் ஒலி??யர்?மான் மூலமாக தகவல் சிதைவதற்? இடம் தரக்கூ?Ôது என்?தற்காகவும் ?ட்டுØ? இந்த மாற்று ஏற்பÔடு வேண்டும் என்று இந்த ஆய்வுரை வலியுறுத்துகிறது.

அதற்கான தீர்வை நோக்கிய ஆலோசனைகள் என்ன?

க்,ட், த்,?, ஆகிய மான்? அதிர்வற்ற வல்லினங்களின் அதிர்வுள்ள ?Ôற்?றாலிகளான g,d,dh,b ஆகியவற்றுக்? தனியாக மான்? புதிய எழுத்துக்களை உருவாக்கி சேர்த்துக் கொள்ளுதல்.

கிரந்த எழுத்துக்களின் உதவியோடு ச்-இன் ?Ôற்?றாலிகளான ஜ்-உம் ஸ்-உம் ஏற்கனவே ஏற்க??ட்டுள்ளன. ஆனால் Ch என்கிற ஒலியனுக்கான தமிழ் வடிவம் உருவாக்கப்பட வேண்டும்.

X,Z ஆகிய ஒலியன்களுக்கும் வரிவடிவம் காணப்பட வேண்டும்.

இதற்? தேவையான வரிவடிவங்களை எப்படி உருவாக்குவது என்ற கேள்விக்கான ?திலை தமிழில் ஜ, ஸ,ஷ, ஹ,ஃ?, ஆகியவை ஏற்க??ட்?தின் மான்னணியிலிருந்து வருவிக்கலாம். முதல் மான்? எழுத்துக்களும் தமிழ் வரிவடிவத்தோடு பட்டிய, தமிழ் ?Ôட்டில் வடமொழியை எழுதுவதற்காக பயன்படுத்த??ட், தமிழ் எழுத்துக்களோடு வடிவ குற்றுÙ?மாகவும் ?கா?? கிரந்த எழுத்துக்களிலிருந்து எடுக்க??ட்?ன என்?தால் அதில் அந்நியத்தன்Ù?யும் வரிவடிவ வித்தியாசமும் இல்லாமல் Ø?Ô?விட்டன. ஃ? என்பதும் ?Ù?முறையில் உள்ள இரு தமிழ் எழுத்துக்களின் கூட்டுதான்.

இந்த அடிப்படையில் சேர்க்கையில், புதிய வரிவடிவங்களை உருவாக்குவது என்?தற்? ????Õல் உள்ள வரிவடிவங்களின் ?Öட்சியாக, சில மாற்றங்களை ?ச?து புதிய வடிவங்களாக அவற்றை அறிமுகநெடுத்துவது என்று அர்த்தநெடுத்திக்?கா??Ôல் அது சரியான வழியாக இருக்கும் என்று தோன்றுகிறது. கீழ்கண்ட ஆலோசனைகளை யோசிக்கலாம்.

'\'அல்லது '/'அல்லது œ.œ என்கிற குறியீடு ஒன்றை வித்தியாசநெடுத்தும் குறியீமாக பயன்படுத்தலாம். உதாரணத்துக்?, œகœ என்ற வரிவடிவத்துடன் இந்த மூன்றில் ஏதோ ஒரு குறியீட்Ù? இணைத்து œகœ அல்லது œகœ அல்லதுœகœ Øமான்ற ஒரு வடிவத்தை உருவாக்கி அதை ga என்று உச்சரிக்க வேண்டும் என்று விதிக்கலாம்.அதே ச?யம் இந்த புதிய குறியீடுகள் ஒலி??யா?புக்? ?ட்டுØ? என்று வரம்?Õ?வேண்டும். இந்த உத்தேச வடிவ அÙ???Õல் எடுத்துக்கொள்ள??ட்? ஏழு ஒலியன்களுக்கான தமிழ் வடிவங்களை கீழே அட்?வணை இட்டிருக்கிறேன்.

இந்த உத்தேச வரிவடிவங்களின் முக்கிய அம்சங்கள்:

1.இந்த வரிவடிவங்கள் தற்போது மாற்றாக பயன்படுத்தநெடும் வரிவடிவங்களின் ?Öட்சியே. எனவே புதிய வடிவங்கள் தன் தா? வடிவத்தின் ஒலியின் ?Ôற்?றாலியே என்பதை எளிதில் ?Õரூமாக்கும். எல்லா புதிய வடிவங்களுக்கும் ?ரே வகையான புதுக் குறியீட்Ù? அளி??தன் மூலமாக அவை ?ரே இனத்தைச் சேர்ந்தவை என்பதையும் காட்? முடியும் (உதாரணமாக, குறியீடு இருந்தால்
அது அதிர்?வாலி என்பது Ø?Ôல)
2.புதிதாக தரநெடுத்த??வுள்ள எழுத்துரு குறியீட்டு முறை(Character encoding), விசைப்பலகை இட ?துக்கீடு (keyboard layout) ஆகியவையும் இந்த மாற்று எழுத்துக்களை ஏற்பது எளிது. கூடுதல் வடிவங்களான இந்த ஏழினையும் குறியீட்டு?பட்டியலில் கூடுதல்
எழுத்துக்களாக ஏற்கமுடியும். விசைப்பலகையை? பொறுத்தவரை, கூடுதல் குறியிட்டுக்காக ?ரே?யாரு விசையை ?துக்?வதன் மூலம் ?Õரச்சனையை தீர்த்துவி?லாம். ஆக, ஒரு சின்ன?சிறு குறியீட்Ù?த் சேர்??தன் மூலØ? ஒலி??யர்?பு சவாலை முறியடிக்கமுடியும் என்று தெளிவாக தெரிகிறது.

இங்? இதே பிரச்னைசம்பந்தப்பட்ட ஒரு முக்கிய முன்னுதாரணத்தை எடுத்துக்காட்டுவது ?ல்லது. Ø?லும் அது யூனிகோட் சம்பந்தப்பட்ட ஒன்று என்?தால் தற்Ø?Ôதைய விவாதத்துக்? அது முக்கியத்துவத்தையும் அளிக்கும்.

புதிய எழுத்துக்களை சேர்த்து யூனிகோட் தரும் வாய்ப்பு - தேவநாகரி முன்னுதாரணம் - ஒரு பார்வை வீச்சில் The Unicode Consortium வெளியிட்டுள்ள The Unicode Standard, Version 2.0 புதிய எழுத்துக்களை சேர்க்க விரும்புவர்களுக்கு ஆறுதலைத் தருகிறது.இந்திய மொழிகளை? பொறுத்தவரை, ISCII-II வையே அது மூல தரநெடுத்தமிக எடுத்துக் ?கா?டிருக்கிறது. இந்த தரநெடுத்தம் தமிழை? பொறுத்தவரை முழுமையானதோ முறையானதோ அல்ல என்கிற குற்றச்சாட்Ù? ?றுக்க முடியாது என்பதை ஏற்றுக்கொள்ளும் இந்த ஆய்வுரை வேறொரு கோணத்திலிருந்து அதன் அம்சங்கள் சிலவற்றை ஒலி??யர்?பு குறித்த ஆய்வுக்கு பயன்படுத்திக்கொள்வதோடு நிறுத்திக் கொள்கிறது.

எல்லா இந்திய மொழிகளுக்கான குறியீட்டுப்பட்டியல்களிலும் சில இடங்களை எதிர்காலத்தில் சேர்க்கக்கூடிய வரிவடிவங்களுக்காக என்று காலியாக ஒதுக்கி வைத்திருக்கிறது. எடுத்துக்காட்டாக, தமிழுக்?கன ஒதுக்க??ட்? OB80 - OBF2 ஆகியவற்றுக்கிÙ?யிலான 115 இடங்களில் 54 இடங்கள் காலியாக இடமாக வைத்திருக்கிறது. இந்த இடங்களில் கூடுதல் வரிவடிவங்களை முறைப்படி சேர்த்துக்கொள்ளலாம். உதாரணமாக, கன்னகும் f க்கான புதிய வரிவடிவத்தை அவ்வாறே சேர்த்துக் ?காட்டிருக்கிறது. தேவநாகரிக்கான குறியீட்டு? பட்டியலை
ஆராயும்போது, ஒரு முக்கிய மாற்றம் ?தன்?டுகிறது. Various Signs என்கிற இனத்தில் œ.œ என்கிற ஒரு கூடுதல் வரிவடிவம் சேர்க்கப்பட்டு, அதற்கான காரணமாக for extending the alphabet to new letters எனவும் குறியீ???ட்டிருக்கிறது. இதே முறை குஜராத்தி, குரியா, வங்காளம் ஆகிய மொழிகளின் குறியீடுகளிலும் மான்குற்ற??ட்டிருக்கிறது.

இந்த விரிவாக்கமுறையை தேவநாகரி எப்படி பயன்படுத்தியிருக்கிறது என்பதை அதே பட்டியலில் அது சேர்த்துக்கொண்டுள்ள சில புதிய எழுத்துக்களைக் கொண்டு விளக்கமுடியும்.

திராவிட மொழிகளிலிருந்தும் குறியீக தமிழில் இருந்தும் சில ?Õரத்யேக ஒலிகளை ஒலி??யர்த்து எழுதுவதற்காக ?Õலவும் வரிவடிவங்கள் சிலவற்றின் கீழே புள்ளி ஒன்றை வைத்து அதை புதிய வரிவடிவமாக ?Ö்ட்டித்துக்கொள்ளும் உத்தியை இந்த முறை மான்குற்றுகிறது. உதாரணம்: தேவநாகரியின் ரகர எழுத்துக்? கீழே ஒரு புள்ளிவைத்து அதை தமிழ் வல்லின œறœ வை

ஒலி??யர்??தற்கான வரிவடிவம் என்று ISCII-II வரையறுத்திருக்கிறது பட்டியலிலிருந்து

0931 DEVANAGARI LETTER LLLA(for transcrbing Tamil alveolar r) இது Ø?Ôலவே for transcribing Tamil alveolar n, for

trnascribing Tamil l என்று குறியீட்டில் மூன்று வரிவடிவங்களை சேர்த்துக் கொண் டிருக்கிறது. Ø?லும்திராவிட மொழிகளின்

உயிர்க்குறில்களான எ, ஓ ஆகியவற்றை ஒலி??யர்க்க இரு கூடுதல் வரிவடிவங்களை அது சேர்த்துக் கொண்டிருக்கிறது.

தேவநாகரிக்கான யூனிகோட் முறையில் இந்த மாற்றங்களுக்கு மான்னுள்ள அணுகுமுறை ?Ôராட்?த்தக்கது. எளிதில் குற்ற மொழிகளிலும் மான்குற்றத்தக்கது. இதே அடிப்படையில் தமிழ் யூனிகோட் குறியீட்டிலும் ஒரு கூடுதல் குறியீட்Ù? சேர்த்து புதிய வரிவடிவங்களை சேர்க்க வழிசெய்யலாம்.

முடிவுரையாக

?Ö?டும் ஒரு எடுத்துக்காட்Ù? சுட்டிக்காட்டலாம் என்று தோன்றுகிறது. ?ல்லாயிரக்கணக்கான கருத்?தழுத்துக்களையும் (Ideographs) ஹிரகானா

,க?கானா என்று இரு வேறு?ட்? ஒலி??Õயல் நெடுங்கணக்கையும் ஒருசேர மான்குற்றும் ஐ?மானிய மொழியும் கூ?, அந்நிய மொழிகளிலிருந்து

ஒலி??யர்க்க Ø?ரிடும் போது துல்லியமாக செய்யவேண்டும் என்?தற்காக க?கானா எழுத்துக்கள் சிலவற்றை ?Ö்ட்டித்து சு?Ôர் 20 புதிய

வரிவடிவங்களை உருவாக்கிக் ?கா??து. ஐப்பானிய மொழியின் அல்லது இந்தியின் அணுகுமுறை இந்த வகையில் முற்நோக்கானது.

காலத்தை ஓட்டியது. இது வளரும் மொழிகளுக்கான அÙ?யாளம்.

இந்த விதி தமிழுக்கும் ??Ôருந்தும். கும்?Õ?Ø? அதற்கு முன்னுதாரணம் இரு??தையும் முன்Ø? ?Ôர்த்தோம். இடபோது சற்று கூடுதலான வரிவடிவ

தேவைகளை ?Õர??வேண்டிய தருணத்தில் உள்ளோம். தற்போதுள்ள தமிழ் நெடுங்கணக்கில் புதிய வரவுகளாக இவற்றை இந்த ஆய்வுரை

முன்மொழிந்தாலும் விரு??த்தின் அடிப்படையில் இவற்றை பயன்படுத்துவதற்கான ஒரு வா???Ôவது அளிக்க???லாம்.

இது குறித்து விரைந்து விவாதித்து துணிந்து முடி?வடுத்து அவற்றையும் தரநெடுத்திவிட்டÔல் அது தமிழுக்? கூடுதல் ?லத்தை தரும் என்பதை

?தாழில் முறை மொழி??யர்??Ôளன் என்கிற என் சொந்த அனு?வத்திலிருந்து ?Õச்சயம் அடித்துச் சொல்ல முடியும்.

Selective Bibliography

The Unicode Standard, Version 2, The Unicode Consortium

ISCII and Tamil - A Perspective By N. Anbarasan, paper presented in Tamilnet 97 Conference

Ø?ச்சொலியியல், சி.சி??Õர?ணியன், ?Ô?Ôர் வழக்காற்றியல் ஆய்வு Ù?யம், ?Ôளையங்கோட்Ù?, 1998

??Ôது மொழியியல், மாக்?ர். ??Ôற்கோ, புÛம்மொழில், 1997

அறிவியல் தமிழ், மாக்?ர். வா.?ச. ?ழந்தைசா?Õ, ?Ôரதி ?தி??கம், 1985


உலகத் தமிழ் இணையக் கருத்தரங்க மாந


உலகத் தமிழ் இணையக் கரு


உலகத் தமிழ் இணையக் கருத்தரங்க


உலகத் தமிழ் இணையக் கருத்