



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011



International Forum for information Technology in Tamil (INFITT)  
& University of Pennsylvania (UPENN)  
jointly conduct

10<sup>வது</sup> தமிழ் இணைய மாநாடு 2011  
10<sup>th</sup> TAMIL INTERNET CONFERENCE 2011  
கூன் 17 - 19 June 17 - 19



# CONFERENCE PAPERS

**Tamil Internet Conference 2011**

*University of Pennsylvania, Philadelphia, USA*

*June 17-19, 2011*

## **Committees for Tamil Internet 2011 Conference**

### ***Conference Program Committee CPC:***

- Dr. K. Kalyanasundaram, Switzerland – Chair
- Prof. N. Deivasundaram, India
- Prof. A.G. Ramakrishnan, India
- Mr. Mani Manivannan, India
- Dr. Badri Seshadri, India
- Dr. Jean-Luc Chevillard, France
- Prof. C.R. Selvakumar, California
- Dr. Seetha Lakshmi, Singapore
- Mr. Siva Pillai, United Kingdom

### ***Local Organizing Committee LOC***

- Dr. Vasu Renganathan, University of Pennsylvania – Chair
- Prof. Harold F. Schiffman, University of Pennsylvania – CoChair
- Dr. Sankaran Radhakrishnan, University of Texas, Austin
- Dr. Arasu Chellaiah, Maryland
- Mr. Ramachandran Sivakumar, Blue Bell, Pennsylvania
- Dr. Muthumani Karuppiyah, University of Pennsylvania
- Dr. Sornam Sankar, Maryland
- Mr. Somalay Somasundaram, Avondale, Pennsylvania
- Mr. Kumar Kumarappan, California
- Ms. Jody Chavez, University of Pennsylvania

### ***International Organizing Committee IOC***

- Mr. V.M.S. Kaviarasan (US) (chair)
- Mr. Ilantamizhan (Malaysia)
- Mr. Anto Peter (India)
- Dr. M. Ponnavaikko (India)
- Dr. Appasamy Murugaiyan (France)
- Dr. S. Mohan (India)

### ***Exhibition Committee***

- Mr. Maniam, Singapore (Chair)
- Mr. Anto Peter, India
- Mr. Anandan, India

# Contents

|           |  |          |           |
|-----------|--|----------|-----------|
| <b>A.</b> | <b>Tamil Literature through Computer (கணினியினூடே செம்மொழி)</b>  | <b>/</b> | <b>01</b> |
| A1        | கை கணினியூடே செம்மொழி<br>கோ. தேவராஜன், Chennai, India  | /        | 03        |
| A2        | பாரதியின் பாடல்களுக்கு மின்னணு வழி வாசிப்புக்கருவி உருவாக்கம்<br>- ஒரு கணினி மொழியியல் அணுகுமுறை<br>டாக்டர் இரா. வேல்முருகன்<br>(தேசியக் கல்விக் கழகம் நன்யாங் தொழில் நுட்பப் பல்கலைக் கழகம், Singapore)   | /        | 09        |
| A3        | பல்லாடக வழி அற இலக்கியங்களைக் கற்றல், கற்பித்தல்<br>(Teaching and learning in ethical literature through multimedia)<br>முனைவர் வா.மு.சே. முத்துராமலிங்க ஆண்டவர்,<br>(பச்சையப்பன் கல்லூரி, Chennai, India) | /        | 14        |
| A4        | Kuralagam - Concept Relation based Search Engine for Thirukkural<br>Elanchezhiyan.K, T V Geetha, Ranjani Parthasarathi & Madhan Karky<br>(Anna University, Guindy, Chennai, India)                         | /        | 19        |
| A5        | Tamil Literature Output in National Bibliography of Indian Languages:<br>A bibliometric analysis P. Clara Jeyaseeli<br>(Madurai Kamaraj University, Madurai, India)  | /        | 24        |
| A6        | கணினி வழி தமிழ்ச் சங்க இலக்கிய ஆய்வு<br>வெ. பாலசரஸ்வதி<br>(அவினாசிலிங்கம் நிகர் நிலை பல்கலைக்கழகம், Coimbatore, India)   | /        | 30        |
| A7        | வெண்பா நிரல் - வெண்பாவிட்கான பொது இலக்கணங்களைச் சரிபார்க்கும் நிரல்<br>மு. சித்தநாதபூபதி<br>(எவர்செண்டாய் எஞ்சினியரிங், ஷார்ஜா)  | /        | 33        |
| <b>B.</b> | <b>Tamil Learning (கணினி/இணையம் வழி தமிழ் மொழி கற்றல் மற்றும் கற்பித்தல்)</b>  | <b>/</b> | <b>37</b> |
| B1        | தமிழ்மொழியும் உச்சரிப்பும்- கற்றல் கற்பித்தலில் கணினியின் பங்கு<br>டாக்டர் ஆ. ரா சிவகுமாரன்<br>(தேசியக் கல்விக்கழகம் -நன்யாங் தொழில்நுட்பப் பல்கலைக்கழகம், Singapore)                                      | /        | 39        |
| B2        | இணையம் மற்றும் கணினி மூலம் தமிழ் கற்றல் மற்றும் கற்பித்தல்<br>திருமதி. ரஜனி ரஜத்<br>(பாரதியார் பல்கலைக்கழகம், Coimbatore, India)   | /        | 44        |
| B3        | இணையவழித் தமிழ்ப்பாடங்கள்<br>முனைவர் மு.இளங்கோவன், பாரதிதாசன் அரசு மகளிர்கல்லூரி,<br>புதுச்சேரி-605 003 இந்தியா  | /        | 48        |



|     |   |   |     |
|-----|---|---|-----|
| B4  | தமிழில் தகவல் தொழில்நுட்பத்தைக் கற்பித்தல்: வாய்ப்புகளும் சிக்கல்களும்<br>வே. இளஞ்செழியன் & சி.ம.இளந்தமிழ், மலேசியா   | / | 54  |
| B5  | Teaching Tamil and Managing Tamil Schools Using Open Source Computing<br>Saravanan Mariappan (Nexus NGN Sd, Kuala Lumpur, Malaysia)   | / | 56  |
| B6  | An Innovative Software for Learning to Write Tamil Lesson Plan<br>Dr. S.K. Panneer Selvam (Bharathidasan University, Trichi, India)   | / | 65  |
| B7  | Attitudes and Motivation in Teaching through ICT among<br>Malaysian Tamil Teachers - An Overview<br>Dr. Paramasivam Muthusamy (University Putra Malaysia, Malaysia)   | / | 70  |
| B8  | தகவல் பரிமாற்றுத் திறமைகள் மூலம் தமிழ் மொழி, கலாசாரம் கற்பிக்கும்<br>வழிவகைகளைக் கட்டியெழுப்பல் - இங்கிலாந்து அரசாங்கத்தின் தேசிய கொள்கை<br>அபிவிருத்தி - அடைவையும், குறிக்கோள் சார்ந்த ஊக்கத்தையும் அதிகரித்தல்<br>சிவா பிள்ளை (Univ. of London (Goldsmith College), London, UK) | / | 74  |
| B9  | Facebook and Tamil Language in Singapore's Teacher Education<br>Seetha Lakshmi (National Inst. of Education, Singapore)   | / | 76  |
| B10 | Virtual Environment As A Collaborative Platform<br>To Enhance Pupils Information literacy skills<br>Sivagouri Kaliamoorthy (Singapore)  | / | 92  |
| B11 | இணையம் மற்றும் கணினி வழி தமிழ் கற்றல் கற்பித்தல்<br>நல்லாமுர் முனைவர் கோ. பெரியண்ணன்<br>(தமிழகக் கல்வி ஆராய்ச்சி வளர்ச்சி நிறுவனம் (இயக்குநர்), Chennai, India)   | / | 97  |
| B12 | இணையம் வழித் தேர்வுகளில்லாக் கல்வி<br>(Education Without Examination, through E-Learning)<br>முனைவர் ப.அர.நக்கீரன்<br>(இயக்குநர், தமிழ் இணையக் கல்விக்கழகம், சென்னை, தமிழ்நாடு)   | / | 100 |
| C.  | <b>Artificial Intelligence (செயற்கைத் திறனாய்வு)</b>  | / | 105 |
| C1  | Tamil Video Retrieval Based on Categorization in Cloud<br>V.Akila and Dr.T.Mala (Anna Univ. - Guindy, Chennai, India)   | / | 107 |
| C2  | Animated Story Visualizer for Tamil Text<br>M. Janani and Dr.Mala.T (Anna Univ. - Guindy, Chennai, India)   | / | 118 |
| C3  | Popularity Based Scoring Model for Tamil Word Games<br>Elanchezhyan.K, Karthikeyan.S, T V Geetha,<br>Ranjani Parthasarathi and Madhan Karky (Anna Univ. - Guindy, Chennai, India)   | / | 124 |
| C4  | Multilingual Cross - Domain Classification of Tamil Web Documents<br>based on Neural Network with Dimension Reduction<br>M.Balaji Prasath, Dr. D.Manjula<br>(Anna Univ. - Guindy, Chennai, India)   | / | 129 |

|           |   |   |            |
|-----------|---|---|------------|
| C5        | On Emotion Detection from Tamil Text<br>Giruba Beulah S E, and Madhan Karky V<br>(Anna Univ. - Guindy, Chennai, India)  | / | 133        |
| C6        | Tamil Online handwriting recognition using fractal features<br>Rituraj Kunwar and A G Ramakrishnan<br>(Indian Inst. of Science, Bangalore, India)   | / | 142        |
| C7        | Neuroscience inspired segmentation of handwritten words<br>Prof. A G Ramakrishnan and Suresh Sundaram<br>(Indian Inst. of Science Bangalore, India)   | / | 148        |
| C8        | Improving Tamil-English Cross-Language Information Retrieval<br>by Transliteration Generation and Mining<br>A Kumaran, K Saravanan & Ragavendra Udupa<br>(Multilingual Systems Research Microsoft Research India Bangalore, India.) | / | 154        |
| <b>D.</b> | <b>Computational Linguistics (கணினி மொழியியல்)</b>  | / | <b>163</b> |
| D1        | A Package for Learning Negations in Tamil<br>Dr. G. Singaravelu (Bharathiyar Univ. Coimbatore, India)   | / | 165        |
| D2        | Morphology based Factored Statistical Machine Translation<br>(F-SMT) system from English to Tamil<br>Anand Kumar M, Dhanalakshmi V, Soman K P, Rajendran<br>(Amrita Vidya Peetam, Coimbatore, India)                                | / | 171        |
| D3        | Tamil Shallow Parser using Machine Learning Approach<br>Dhanalakshmi V, Anand Kumar M, Soman K P and Rajendran S<br>(Amrita Vidya Peetam, Coimbatore, India)  | / | 175        |
| D4        | கணினிவழித் தமிழ்மொழியாய்வில் பொருள் மயக்கம்<br>இல. சுந்தரம் (SRM University, Chennai, India)  | / | 180        |
| D5        | கணினியில் ரோமன் வரிவடிவ ஒலிபெயர்ப்பு<br>முனைவர் இராதா செல்லப்பன்<br>(பாரதிதாசன் பல்கலைக்கழகம், Trichi, India)   | / | 186        |
| <b>E.</b> | <b>Electronic Tamil Dictionary (மின் அகராதி)</b>  | / | <b>195</b> |
| E1        | Agaraadhi: A Novel Online Dictionary Framework<br>Elanchezhian.K, Karthikeyan.S, T V Geetha, Ranjani Parthasarathi & Madhan Karky<br>(Anna Univ- Chennai, India)  | / | 197        |
| E2        | நவீன தமிழ் அகராதி<br>முனைவர் க. தமிழ்ச்செல்வன் (வி.ஐ.டி. பல்கலைக்கழகம், Vellore, India)   | / | 201        |
| E3        | மொழிபெயர்ப்புக் கலையில் அகராதியின் பயன்பாடு<br>இளங்குமரன் த/பெ சிவநாதன்<br>(சுல்தான் இட்ரிஸ் கல்வியியல் பல்கலைக்கழகம், Malaysia)  | / | 205        |

|           |  |   |            |
|-----------|--|---|------------|
| <b>F.</b> | <b>Blog (வலைப் பூக்கள்)</b>  | / | <b>211</b> |
| F1        | Impact of SOA and Web 2.0 in Tamil Blogs and Social Networks<br>Ferdin Joe J (Einstein College of Engineering, Tirunelveli, India)   | / | 213        |
| F2        | Tamil Classical Literature in the age of blogging and social network<br>Palaniappan Vairam Sarathy (Software Developer Virginia, US)   | / | 216        |
| <b>G.</b> | <b>Wikipedia / விக்கிபீடியா - தமிழ் நிரலிகள்</b>   | / | <b>223</b> |
| G1        | Enriching Tamil and English Wikipedias<br>Dr. N.Murugaiyan, Central Institute of Classical Tamil, Chennai, India)  | / | 225        |
| G2        | கூட்டாசிரியப் படைப்பு: தமிழ் விக்கிப்பீடியா<br>செ. இரா. செல்வக்குமார் (Univ, of Waterloo, Waterloo, Canada)  | / | 230        |
| G3        | விக்கிபீடியா - தமிழ் நிரலிகள்<br>ச. சந்திரகலா (அவினாசிசிங்கம் நிகர் நிலை பல்கலைக்கழகம், Coimbatore, India)   | / | 236        |
| <b>H.</b> | <b>E-Commerce (மின் வணிகம்)</b>  | / | <b>239</b> |
| H1        | E-Governance Activities in Tamil Nadu<br>Dr. E. Iniya Nehru, (National Informatics Center NIC, Chennai, India)   | / | 241        |
| H2        | New Media and Tamil - using softwares, tools and Technology<br>S. Gunasegaran<br>(Temasek Polytechnic, Singapore)  | / | 248        |
| H3        | தமிழில் கணினிப் பாவனை<br>சிவா அனூராஜ், ஜாஃப்னா   | / | 254        |
| H4        | Electronic Commerce<br>Dr.B.Neelavathy<br>(Avinashilingam Deemed University for women, Coimbatore, India)  | / | 257        |
| <b>I.</b> | <b>Natural Language Processing (இயற்கை மொழி பகுப்பாய்வு)</b>   | / | <b>265</b> |
| I1        | An Efficient Tamil Text Compaction System<br>N.M..Revathi, G. P. Shanthi, Elanchezhian.K, T V Geetha,<br>Ranjani Parthasarathi & Madhan Karky<br>(Anna Univ. Guindy, Chennai, India) | / | 267        |
| I2        | Tamil Summary Generation for a Cricket Match<br>J. Jai Hari Raju, P. Indhu Reka, K.K Nandavi, Dr. Madhan Karky<br>(Anna Univ. Guindy Chennai, India)                                 | / | 271        |
| I3        | Lyric Mining: Word, Rhyme & Concept Co-occurrence Analysis<br>Karthika Ranganathan, T.V Geetha, Ranjani Parthasarathi & Madhan Karky<br>(Anna Univ. Guindy, Chennai, India)          | / | 276        |

|           |   |   |            |
|-----------|---|---|------------|
| I4        | Template based Multilingual Summary Generation<br>Subalalitha C.N, E.Umamaheswari, T V Geetha,<br>Ranjani Parthasarathi & Madhan Karky<br>(Anna Univ. Guindy, Chennai, India)                                     | / | 282        |
| I5        | Special Indices for LaaLaLaa Lyric Analysis & Generation Framework<br>Suriyah M, Madhan Karky, T V Geetha, & Ranjani Parthasarathi<br>(Anna Univ. Guindy, Chennai, India)   | / | 287        |
| I6        | Tamil Document Summarization Using Latent Dirichlet Allocation<br>N. Shreeya Sowmya, Dr. T. Mala (Anna Univ. Guindy, Chennai, India)  | / | 293        |
| <b>J.</b> | <b>Language Ideology, Inscriptions, Spoken Tamil and Technology</b><br>(மொழிக் கொள்கை, கல்வெட்டுத் தமிழ், பேச்சுத் தமிழ் – தொழில் நூட்பத்தின் பங்கு)  | / | <b>299</b> |
| J1        | Mapping Language Change in Tamil:<br>Corpus analysis and Computer Database Making<br>Dr. Appasamy Murugaiyan (EPHE- UMR, Paris, France)   | / | 301        |
| J2        | Language Ideology and Technology<br>Prof. E. Annamalai (Univ. Chicago, Chicago, USA)  | / | 307        |
| J3        | Tamil: A Family of Languages<br>Prof. Harold Schiffman (Univ. Pennsylvania, Philadelphia USA)   | / | 312        |
| J4        | Tamil Literature from Sangam to Modern Period:<br>A Continuum with colorful changes:<br>What does a search of the Tamil Electronic data reveal us?<br>Dr. Vasu Renganathan (Univ. Pennsylvania, Philadelphia USA) | / | 319        |
| J5        | Open Source Tamil Computing<br>S. Gopinath, E Iniya Nehru<br>(National Informatics Center, Chennai, India)  | / | 325        |



# Messages



# भारतीय प्रौद्योगिकी संस्थान कानपुर Indian Institute of Technology Kanpur

प्रो. मु. आनंदकृष्णन्  
अध्यक्ष  
Prof. M. Anandakrishnan  
Chairman



पत्रालय — भा.प्रौ.सं.कानपुर — 208016 (भारत)  
Post Office - I.I.T. Kanpur 208 016 (India)  
Fax - +91-512-259 0260

## MESSAGE

Contrary to several negative predictions about the viability and longevity of the INFITT since its inception, it is gratifying that the organization has continued to serve the cause of Tamil Computing and Tamil Internet developments in unique and distinct manner. This has been possible on account of the dedicated involvement and participation of a large number of highly competent professionals in the spirit of community responsibility. In recent years, the involvement and contribution of a substantial number of new generations of professionals offers a renewed hope towards innovative vistas of development in Tamil internet applications. This has also enabled seamless adoption of new and emerging areas of information technology.

The pioneers who devoted a great deal of time and effort in the early stages of Tamil Computing, when the technology was in a nascent stage, deserve our wholehearted gratitude. Despite the usual organizational constraints and internal differences the INFITT has continued to show positive growth and has come to be recognized both by several National Governments as well as International Organizations as representing the voice of Tamil Computing. It is my fervent hope that this Tenth Tamil Internet Conference being held at the University of Pennsylvania will provide new directions for the future of INFITT in the context of emerging challenges.

It may be recalled that in the initial stages, the INFITT tended to rely on close collaborations with the Governments of predominantly Tamil speaking countries. This has advantages as well as constraints. The future may call for a new strategy that would provide for an arms length relationship between the INFITT and the concerned Governments, mainly to insulate the organization from the vagaries of shifting policy stances of Governments whenever there is a changes in political structure. Ideally, the INFITT should be seen as a truly professional and technical organizations engaged in monitoring the development trends in Tamil computing and offer advice to the Governments in enabling the practical applications of these developments.

I offer my sincere compliments and gratitude to the University of Pennsylvania for having taken the initiative to continue with the tasks ahead by organizing this Tenth Tamil Internet Conference.

(M.Anandakrishnan)

Chennai  
30 April 2011

---

### Address for Communication :

Science City Building, Planetarium Campus, Gandhi Mandapam Road,  
Chennai - 600 025, Phone & Fax : 044-24422415 (O)  
24916291 (R) 94440 51133 (M) E-mail : ananda1928@gmail.com





On behalf of the South Asia Studies Department here at the University of Pennsylvania, I would like to extend our warm welcome to the esteemed delegates of this conference. Our department is not only the oldest South Asia Studies Department in the US, with Tamil being part of its curriculum from the 1940s, it has also produced in the past some of the most important works on the subject of Tamil language and literature in the United States. We are honored to have members of INFITT and wish them great luck in years to come.

**Dr. Daud Ali**

Associate Professor and Chair

South Asia Studies



On behalf of the South Asia Center, I welcome the members of INFITT to the University of Pennsylvania for the 10<sup>th</sup> Annual Tamil Internet Conference. We are delighted and honored to be able to host this important conference here at Penn. I want to take this opportunity to thank Dr. Vasu Renganathan and Dr. Harold Schiffman for organizing the conference and for their significant contributions over the years to the field of Tamil language study. The South Asia Center together with the South Asia Studies Department has long been committed to supporting excellence and innovation in South Asian Language scholarship and pedagogy. Dr. Renganathan and Dr. Schiffman have provided critical leadership in these efforts, at Penn as well as nationally and internationally. We are pleased now to have the opportunity to lend our support to this conference and to the important work of INFITT. We hope that your visit to the University of Pennsylvania will be extremely productive and that you will also have time to enjoy our wonderful city of Philadelphia.

A handwritten signature in black ink, appearing to read "Kathleen D. Hall". The signature is fluid and cursive, with the first name "Kathleen" and last name "Hall" being the most prominent parts.

**Kathleen D. Hall**

Associate Professor of Education & Anthropology  
Director, South Asia Center  
University of Pennsylvania

**Professor Harold F. Schiffman**

*Honorary Conference Chair, Tenth Tamil Internet Conference  
Emeritus Professor of Dravidian Linguistics and Culture  
Department of South Asia Studies  
University of Pennsylvania*



I would like to extend a hearty welcome to the members of INFITT who are gathering in Philadelphia for the 10th annual Tamil Internet Conference (TI2011). My colleagues in the Department of South Asia Studies join me in welcoming you here. Our department is the oldest department in the US devoted to the study of South Asian languages, and Tamil is one of the languages that has been taught here since its inception in the 1948's. We can also claim to have been in the forefront of innovation in the use of information technology for the teaching and learning of Tamil and other South Asian languages, as can be seen from our Tamil Resources website at <http://ccat.sas.upenn.edu/plc/tamilweb/tamil.html> and we have also participated in the conferences sponsored by INFITT since its inception. We look forward to a successful conference and even more progress in the development of IT for Tamil in the years to come!



## Dr. Vasu Renganathan

South Asia Studies, Uni. of Pennsylvania,

Philadelphia, USA

Chair, Local Organizing Committee

அன்பு நண்பர்களுக்கு

வணக்கம். எங்களது அழைப்பை ஏற்றுக் கடல் கடந்து மா நிலங்கள் பல கடந்து இம்மாநாட்டைச் சிறப்பிக்க வருகை தந்திருக்கும் உங்கள் அனைவரையும் பென்சில்வேனியாப் பல்கலைக் கழகத்தின் தெற்காசிய மையத்தின் சார்பாக வரவேற்பதில் பேருவகை கொள்கிறேன். உத்தம நிறுவனத்தின் இப்பத்தாவது மாநாடு பல உத்தம நண்பர்களின் பேருழைப்பில் சிறப்புற நடக்கவிருக்கிறது என்பதை எண்ணும் போது நம் தமிழ் நண்பர்களின் செம்புலப் பெயனீரென ஒற்றுமையும் அயரா உழைப்புமே நம் மனதைக் கவர்வதாக இருக்கின்றன. அவனும் இவனும் அவனிவன் ஆமே என்ற திருமூலரின் கருத்துப் படி நாம் அனைவரும் அவரிவராக ஒன்றிணைந்து நம் கருத்துக் களஞ்சியங்களை அது இதுவென நம்மிடையே பகிர்ந்து கொள்ளும் வாய்ப்பை ஏற்படுத்தித் தந்த இந்த உத்தம நிறுவனத்திற்கு நம் நன்றியைத் தெரிவிப்போமாக! இந்நிறுவனத்தின் தமிழ்ப்பணி மேன்மேலும் வளர நாம் ஒன்று கூடி வாழ்த்துவோம். இணையம் வழி இணைந்த தமிழ் உள்ளங்களின் தமிழ்ப்பணி மேன்மேலும் வளர உத்தமத்தினர் அனைவரையும் வாழ்த்துவோம் உளமார!

இவ்வுத்தம நிறுவனத்தின் தமிழ்ப்பணி சந்திராதித்தவரை சாவாமுவா இளமையோடு என்றென்றும் பூத்துக்குலுங்கவே விழைவோம்!

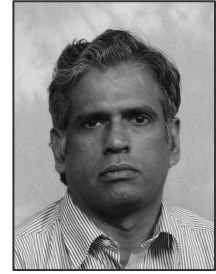
அன்புடன்

வாசு அரங்கநாதன்

**Dr. K. Kalyanasundaram**

Chair, Conf. Program Committee (CPC)

Tamil Internet 2011 Conference



## Message

Growth in wide usage of Tamil in personal computers started about 25 years ago, almost with the growth of Internet and discussions in the virtual world through Email and mailing lists. The required tools (fonts and text editors) were developed largely by few software professionals and free lancers. Creativity of Tamils led to a wide variety of fonts being used for information exchange through the Net . Soon there was the realization that networking worldwide and development of standards are essential to nurture a healthy growth of Tamil content on the net.

INFITT (international forum for information technology) was one of the initiatives launched to serve this need, with active participation of key software developers of all important Tamil speaking regions. Tamil Internet Conferences of INFITT continues to be the only annual forum that brings together all those interested in the development of softwares for computing, Tamil information and communication Technology ICT in general. I am happy that I could be part of the INFITT Management, in the team of organizers for several earlier conferences and also contribute to the present TIC 2011 to be held at the University of Pennsylvania.

A notable and very important development for Tamil Computing and INFITT is ever increasing number of academic researchers participating in the Tamil Internet Conferences and presenting their research work. Major research efforts are now in Universities. Following co-hosting of TiC2002 and TIC 2009 by the Tamil Departments of the University of California at Berkeley and University of Koeln, Germany, another Institution well known for its lead in computer-aided teaching of Tamil, University of Pennsylvania, is hosting this conference. Possibly for the first time, TIC2012 will set records as one where majority of paper presenters are from academic institutions across the world. Teachers and young students taking active interest in Tamil Computing is a very good sign for Tamil Computing. So I take this opportunity to thank sincerely the hosts INFITT and University of Pennsylvania for continuing this conference series.

**Dr. K. Kalyanasundaram**

Lausanne, Switzerland

Former Vice-Chair (2004-2006)

& Chair (2007-2009) INFITT



International Forum for information Technology in Tamil (INFITT)  
& University of Pennsylvania (UPENN)  
jointly conduct

10<sup>வது</sup> தமிழ் இணைய மாநாடு 2011  
10<sup>th</sup> TAMIL INTERNET CONFERENCE 2011  
சூன் 17 - 19 June 17 - 19



வா.மு.சே. கவிஅரசன்  
தலைவர் உத்தமம்  
தலைவர் பன்னாட்டுக் குழு  
தமிழ் இணையம் 2011  
கொலம்பசு, அமெரிக்கா.



Va.Mu.Se. Kaviarasan  
Chair, INFITT  
Chair IOC  
TI 2011  
Columbus, OH, USA

மே 30, 2011  
(தியாகிகள் திருநாள்)

## தமிழ் இணையம் 2011 வாழ்த்து மடல்

அண்டங்களை விழுங்கும் ஆழிப் பேரலைகளை ஒத்த நிகழ்வுகளால் தள்ளாடினாலும், அழிவின் விளிம்பை எட்டிவிட்டோம் என்று கண்ணை மூடித் திறக்கும் போது பிறக்கும் அமைதியின் ஆனந்தம் எண்ணிலடங்காது. அந்த ஆனந்தத்தின் எல்லையை தமிழ் இணையம் 2011 எமக்கு வழங்கியது என்றால், அது மிகையல்ல. பல்வேறு கருத்து வேறுபாடுகளுக்கு இடையிலும், மீட்கப்பட்ட உத்தமத்தின் தமிழ் இணையம் தங்கு தடையில்லாமல் மூன்றாவது ஆண்டாகத் தொடர உடனிருந்து ஒத்துழைத்த பேருள்ளங்களுக்கு எனது மனமார்ந்த நன்றி. காரிருள் கதிரவனை தற்பொழுது மறைத்தாலும், கதிரவன் மீண்டே தீரும்; இது உலக நியதி! உத்தமத்தைச் சார்ந்த காரிருள் மறைய உடன் இருந்து பணியாற்றும் எனது சக தோழர்கள், உறுப்பினர்கள், செயற்குழு உறுப்பினர்கள், இணைய மாநாட்டுக் குழு நண்பர்கள், பணிக்குழு நண்பர்கள் மற்றும் மின்மஞ்சரி ஆசிரியர் குழுவினருக்கு எனது மனமார்ந்த நன்றி.

நேற்று, இன்று போல் நாளையும் தொடரட்டும் நமது தமிழ் இணையம். நமக்குள் துளிர்க்கும் வேறுபாடுகளை மறந்து, நண்பர்களாக 2012ல் மீண்டும் ஓர் தமிழ் இணையத்தில், தமிழ், தமிழ்க் கணினி மென்மேலும் உயர்வடைய தமிழ் உலகத்தின் மற்றொரு அமைதிப் பூங்காவில் சந்திப்போம் என உறுதியேற்போம்.

### TI 2011 – Chair's Message

INFITT had its own share of organizational Tsunami, and I am very pleased to come united to celebrate the peaceful TI2011 that came after the Tsunami. In spite of varied difference of opinions, I am glad to see the revived INFITT continuing its journey to conduct the yearly conferences, popularly known to the Tamil IT world as TI. I sincerely thank all of you who have come together and offered a helping hand to make this a success. Dark clouds hiding the sun is a natural phenomenon. The Sun will raise again. I sincerely thank all friends of INFITT, Members, Executive committee members, Tamil Internet Conference committee members, Working group members and Minmanjari editorial team members who devote their personal time and money to make this happen.

Let us continue our journey like yesterday and today for yet another Tamil Internet Conference in 2012. Let us keep our differences aside and vow to meet again as friends in another conference in another peaceful part of the Tamil world for the betterment of Tamil and Tamil IT.

*International Forum for Information Technology in Tamil  
Registered as a Non-Profit Organization in U.S.A*

**www.infitt.org**



**D. S. Maniam**

*i-DNS .Net International Pte Ltd*

Singapore

Executive Director

INFITT 2010- 2011

பென்சில்வேனியா பல்கலைக்கழகத்தில் நடக்கும் பத்தாவது ஆண்டு மாநாட்டில் கலந்துகொள்வதில் நான் உண்மையிலே மிகப்பெரிய அளவுக்கு மகிழ்ச்சி அடைகிறேன். தமிழ் மென்பொருள் உருவாக்குநர்கள், இணைய வல்லுநர்கள், ஆர்வலர்கள் போன்றோர் தமிழ்க் கணிமையில் தாங்கள் பெற்றிருக்கும் அறிவையும் ஆழத்தையும் வெளிக்காட்டும் நிகழ்ச்சியாகவே வருடாந்தர உத்தம் மாநாடுகள் இருந்துவந்திருக்கின்றன. பொதுவான பிரச்சினைகளை ஆராய்ச்சி செய்து, அதன் முடிவுகளை பகிர்ந்துகொள்ளும் மேடையாகவே தமிழ் இணைய மாநாடுகள் இருந்துவந்துள்ளன.

அமெரிக்காவின் பிலடெல்பியாவிலுள்ள பென்சில்வேனியா பல்கலைக்கழகத்தில் ஜூன் 17-19 இல் நடைபெறவுள்ள மாநாடும் தமிழ்க்கணிமைக்கு ஆக்கபூர்வமான பங்கினை வகிக்கும் என்று உறுதியாக நம்புகிறேன்.

உள்ளூர் அமைப்புக் குழுவிலும் மாநாட்டு நிகழ்வுக்குழுவிலும் தலைமையாளர்களாக பொறுப்பேற்று இந்த நிகழ்வை முறையாக நடத்த முன்வந்திருக்கும் பேராசிரியர்கள் ஹரல்ட் எஃப் ஷிப்மன், முனைவர் வாசு ரங்கநாதன், முனைவர் கே. கல்யாணசுந்தரம் போன்றோருக்கு எனது நன்றிகள் உரித்தாகட்டும். அதுபோலவே பன்னாட்டு அமைப்புக் குழுவின் தலைவர் திரு வா.மு.சே.கவியரசனுக்கும் எனது நன்றிகள்

இந்த வருடாந்தர நிகழ்வு மாபெரும் வெற்றியடைய தங்கள் பணிகளை திறம்பட ஆற்றி அமர்ந்திருக்கிற அனைத்து தன்னார்வலர்களுக்கும் உதவு செய்ய நீண்ட கரங்களுக்கும் உத்தமத்தின் செயற்குழு உறுப்பினர்கள் சார்பில் எனது மனமார்ந்த நன்றிகள். பத்தாவது இணைய மாநாட்டினை நடத்தியிருக்கும் முனைவர் வாசு ரங்கநாதனுக்கும் பென்சில்வேனியா பல்கலைக்கழகத்தின் பிற நண்பர்களுக்கும் எமது சிறப்பான பாராட்டுகளும் மனமார்ந்த நன்றிகளும் உரித்தாகட்டும்.

...

I'm truly delighted to see the 10th annual conference to be held in University of Pennsylvania .The annual INFITT conference has always functioned as an occasion

for Tamil software professionals, Internet experts and enthusiasts to renew their expertise and knowledge on Tamil Computing. It has served as a platform to discuss common issues and share research and developments.

I am sure the meeting to be held at the University of Pennsylvania, Philadelphia, USA during June 17-19, 2011 would offer a constructive forum to engage issues in Tamil computing.

I would like to thank Prof. Harold F. Schiffman, Dr. Vasu Renganathan and Dr. K. Kalyanasundaram who have kindly agreed to assist in the organization of the conference as Chairs of the Local Organizing Committee (LOC) and Conference Program Committee (CPC). Va.Mu.Se. Kaviarasan, Chair of International Organizing Committee (IOC).

On behalf of fellow Executive Members of INFITT we would like to thank all supporters and volunteers who rendered their services towards the success of this annual event. I want to offer our congratulations and best wishes to Dr Vasu Renganathan and his University of Pennsylvania team for organising the 10th Tamil Internet Conference.

*D.S. Maniam*

**S.Maniam**

Singapore





# **Tamil Literature through Computer**

(கணினியினூடே செம்மொழி)



# கை கணினியினூடே செம்மொழி

## Mobile Tamil Keypads

G. Devarajan

E-mail: devarajan @gmail.com

### 1) சுருக்க முன்னுரை:

கணினிகள் சுருங்கி கைகளில் விளையாடும் நவீன தொழில்நுட்ப காலத்தில், நவீன தொழில் நுட்பங்களுக்கு ஏற்ப மொழியின் பயன்பாடு மற்றும் பயன்பாட்டிற்கான மாறுதல்களை எதிர்கொள்வது கட்டாயமாக உள்ளது.

1995-களுக்கு பிறகு ஒவ்வொரு இரண்டு ஆண்டிற்கும் கணினிகளை இயக்கும் மென்பொருள்கள் மாறிய வண்ணமாகவே உள்ளன. இவ்வகையான மாற்றத்தில், கணினிகளில் நமது செம்மொழியை பயன்படுத்துவதே மிகப் பெரிய சிக்கலாகி இருந்து வருகிறது, அப்படி இருந்தும் தமிழ் ஆர்வலர்கள் பலரின் சுய முயற்சிகளினால், ஆங்கிலமற்ற, உலக மற்றும் இந்திய மொழிகளை விட தமிழ் மொழி, கணினி மற்றும் இணைய பயன்பாட்டிலே சிறந்து விளங்குகிறது!

இருப்பினும், இப்படி மின்னல் வேகத்தில் கணினிகள் உருமாறி, இன்று கைபேசிகளாக நம் மடியிலே தவழ்ந்து கொண்டிருக்கும் இன்றைய கால கட்டத்திற்கு ஏற்ப நமது மொழியை பயன்படுத்தும் வரைமுறைகளை உற்று நோக்குவது மிகவும் அவசியமாக உள்ளது.

அவ்வகையான ஒரு உற்று நோக்கலின் பதிப்பு இந்த "கை கணினியினூடே செம்மொழி" கட்டுரை

### 2) முன்னுரை

2000 முதல் 2010 வரை, கிட்டத்தட்ட பத்து வருடங்களாக, பல்வேறு கை பேசி தயாரிக்கும் நிறுவனங்கள் வித விதமான வகைகளில் கை பேசிகளை தயாரித்து கொண்டு வந்துள்ளன இவற்றில் உலக அளவில் மிக பிரசத்தி பெற்றவை Nokia, Sony, Ericsson, Apple மற்றும் Samsung நிறுவனங்கள்.

2004 ஆண்டிற்கு பிறகு, பெரிய மனம் படைத்த சில தன்னார்வலர்கள் தமிழ் சார்ந்த கை பேசிகலுக்கான ஆய்வுகளை மேற்கொண்டனர், இருப்பினும் 2010 வரை இவ்வகையான கை பேசிகளில், தமிழை பயன்படுத்துவது மிகவும் அரிதாகவே இருந்தது, 2010-ற்கு பிறகு தொழில்நுட்ப வளர்ச்சியின் காரணமாக செம்மொழியை பரவலாக கை பேசிகளில் தமிழ் எழுத்துக்களை பார்க்க முடிந்தது.

### 3) 12 Keys/பொத்தான்களின் இயலாமை

இருந்தும் மிகப் பொதுவான மற்றும் போராட்டமான பிரச்சனையாக இருந்து வருவது, அதிக அளவில் பயன்படுத்தப்படும் கை பேசிகளில் இருக்கும் வெறும் 12 பொத்தான்களே!. ஆங்கிலம் மற்றும் ஆங்கில அடிப்படையில் சார்ந்த மொழிகளை தவிர மற்ற எல்லா உலக மொழிகளுக்கும் இந்த 12 பொத்தான்கள் மிகப் பெரிய தடையாகவே உள்ளது. இது சம்மந்தமான பல்வேறு வகையான பரிந்துரைகள் பரிந்துரைக்கப்பட்டும் மொழிகளை கை பேசிகளில் புகுத்துவதில்/பயன்படுத்துவதில் சிரமமாகவே உள்ளது.

ஆனாலும், உலக புகழ் பெற்ற Nokia மற்றும் Motorola கை பேசி தயாரிப்பு நிறுவனங்கள் தனக்கே ஆன, மற்ற இந்திய மொழிகளை சார்ந்த அடிப்படையில் தமிழ் எழுத்துக்கள் அச்சிடப்பட்ட தமிழ் விசை பலகையுடனும், முற்றிலுமான தமிழ் Menu-வுடன் கைபேசிகளை, அணைத்து விற்பனை நிலையங்களில்

வெறும் ரூபாய் 1500 முதல் விற்கப்படுவது மிகவும் மகிழ்ச்சியான மற்றும் பாராட்டக்கூடிய நிலைமையாக இருக்கிறது.

#### 4) பயனரின் இயலாமை

கை பேசியை பயன்படுத்துவோர் இடையே தமிழ் மொழியை பயன்படுத்த வேண்டும் என்ற ஆர்வம் மிக குறைவாகவே உள்ளது, இவற்றிற்கு பல காரணங்கள் இருப்பினும் மிக முக்கியமாக இருப்பது.

Lack of Nationalized standardization of Regional Mobile keypads by Linguistic organizations.

Availability of Mobile Phones with Tamil Characters Printed Keypad

No Strict Orders from State Governments to implement Tamil character printed keypads as mandatory before sale within state.

Negligence of Telecom service providers in promoting historical and classical languages.

இந்த சூழ்நிலையை மாற்ற மொழி சார்ந்த அரசு துறைகள் **Should issue a G.O. to Sell Mobile Phones within Tamilnadu with Tamil characters printed keypad as mandatory**) மற்றும் தனியார் துறைகள் முக்கியமாக (கை பேசி தயாரிப்பாளர்கள் மற்றும் தொலைபேசி சேவை நிறுவனங்கள் **பயனர்களை தமிழ் பயன்படுத்த ஊக்குவிக்கும் விதமான சலுகைகளை அறிமுகப்படுத்துவது**)

#### 5) இன்றைய கை சாதனங்கள் வகைகள்

- Hand Held Bus Ticket Printers, & TNEB Bill Generators
- Hand Held PDA & POS Terminals
- (Tamil yet to get space in these kind of devices)
- Mobile Phones (Number pad based Mobiles, Button based QWERTY keypads & Touch based Qwerty Keypads)
- Book Readers (Mostly Touch based QWERTY Keypads)



#### 6) கைபேசிகளின் இயக்க மென் பொருள்களின் வகைகள் மற்றும் பயன்படுத்துவோர் எண்ணிக்கை (OS & Device market Based on admob Web Requests 2010)

| OS Type                       | OS Owner              | Tamil Fully Rendered | Qwerty Keyboard | Number Dial Keyboard | % of Market World | % of Market India |
|-------------------------------|-----------------------|----------------------|-----------------|----------------------|-------------------|-------------------|
| iOS(iPhone, iPad, iPod Touch) | Apple Inc             | YES                  | Touch           |                      | 40%               | 3%                |
| Android                       | Google Inc            | NO                   | Touch/Keys      | Touch/Buttons        | 25%               |                   |
| BlackBerry                    | RIM                   | NO                   | Touch/Keys      |                      | 6%                |                   |
| Windows Mobile 7              | Microsoft             | NO                   | Touch/Keys      |                      | 4%                |                   |
| Symbian                       | Symbian               | YES/Partly           | Touch/Keys      | Buttons              | 20%               | 93%               |
| Unknown                       | Chinese/Korean Models | NO                   | NO              | Touch/Buttons        | 5%                | 4%                |

Note : Source metrics.admob.com. % report is based on admob requests, May 2010.

## India Mobile Devices Share


|  |      |
|--|------|
| Nokia (Symbian)                          | 42 % |
| LG & Samsung(Java OS)                    | 21%  |
| Micromax, Gfive, Lava & Karbonn(Java OS) | 14%  |
| Others Models(iPhone, Android, Windows)  | 23%  |

Note : Source based on Survey article year 2010(Approximate)

## 7) விசை பலகையின் வகைகள் & வசதிகள் (Number Dial pad, QWERTY Pad, Virtual Key Pad)

### 12-ல் வசதிகள்

| Keypad Type     | Number/ Alphabet Keys | System/Function Keys |
|-----------------|-----------------------|----------------------|
| Number Dial Pad | 9                     | 2~3                  |
| Qwerty Keypad   | 26                    | 6~8                  |
| Touch Keypad    | 26                    | 6                    |

|                           |                         |                                       |   |
|---------------------------|-------------------------|---------------------------------------|---|
| ஃ 1                       | அ, ஆ, இ ஈ, உ, ஊ 2       | எ, ஏ, ஐ,ஒ,ஓ,ஒள 3                      |  <p>Nokia 1661-2 Tamil Keypad,<br/>Rs1500</p> |
| கஙசஞ 4                    | டணதந 5                  | பமய 6                                 |   |
| ரலவ 7                     | ழளறன 8                  | ஐஷஸஹசஷஸ் 9                            |   |
| T9Options & InsertSymbols | 0<br>Space & Line Break | # T9 On/OFF & Switch<br>Tamil/English |   |

2010-ம் ஆண்டிற்கு பிறகு கை பேசிகளில் தமிழ் பயன்படுத்துவது ஓரளவில் அதிகரித்து வருகிறது, இந்த மாறுதலுக்கு மிக முக்கியமான காரணமாக இருப்பது தொடு வகை கை பேசிகள்.

இவற்றில் மிக முக்கியமாக குறிப்பிடத்தக்கது Apple நிறுவனத்தின் iOS-ல் அடிப்படையில் தயாரிக்கப் படுகின்ற திறமை வாய்ந்த கை பேசிகள்/கணினிகள். முக்கியமாக தமிழர்கள் அதிகம் வசிக்கும் நாடுகளான சிங்கப்பூர், மலேசியா, இந்தியா மற்றும் அமெரிக்கா. இதில் அதிக அளவில் தமிழ் மென் பொருள்களை பதிவிறக்கம் செய்யும் நாடுகள், வரிசையில் 1) சிங்கப்பூர், 2) அமெரிக்கா, 3) இந்தியா 4) மலேசியா.

### 26-ல் வசதிகள்

மூன்று வகையான விசை பலகையில் தொடு விசை பலகையை தவிர மற்ற இரண்டிலும் உயிர் மற்றும் மெய் எழுத்தக்களை சேர்க்க இரண்டு அல்லது மூன்று முறைக்கு மேல் ஒரு பொத்தானை அழுத்தினால்

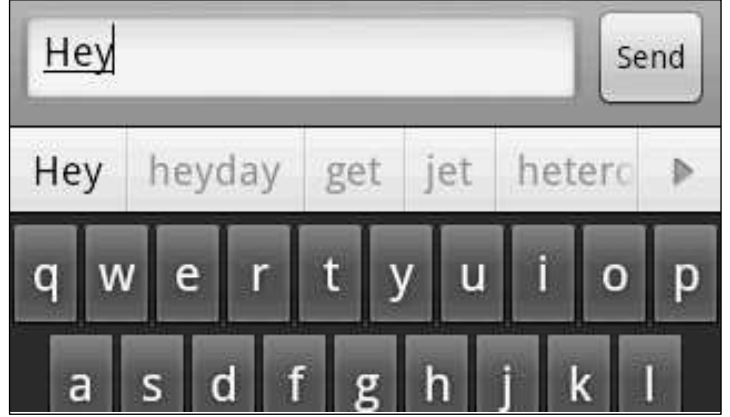
மட்டுமே எழுத்துக்களை பதிவு செய்ய முடியும், ஆனால் வேகமாக பரவி வரும் தொடு வகை மற்றும் Text Animation தொழில் நுட்பங்கள் எந்த ஒரு மொழிக்கும் சாதகமாக பயன் படுத்தி மொழியை மிக எளிதாக பயன்படுத்திக்கொள்ளலாம். அப்படிப்பட்ட Text Animation மற்றும் Text Display Graphics-ல் மிக முக்கியமான இரண்டு விசைப்பலகை தொழில்நுட்பங்கள் கீழ்வருமாறு.

#### க) Continuous Tap gives optional keys (iOS)

ஒரு பொத்தானை தொடர்ச்சியாக தொடுவதினால் அந்த எழுத்தின் Unicode Dependent Sign சார்ந்த மற்ற எழுத்துக்கள் text display animation உதவியுடன் பயன்படுத்த எளிதாக்கப்பட்டுள்ளன!



iOS Keypad



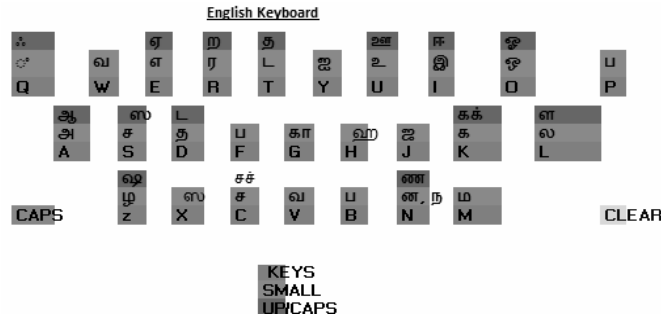
Android Keypad

#### Dictionary based Word formation (Android)

இந்த தொழில்நுட்பம் தொடர் வார்த்தைகள் மற்றும் Dictionary அடிப்படையிலான விசை பலகை, இந்த தொழில்நுட்பம் Google Transliterate அடிப்படையில் கை பேசி இயக்க மென்பொருள்களின் பயன்பாட்டில் கொண்டு வர இயலும்.

இந்த இரண்டு சக்தி வாய்ந்த தொழில்நுட்பங்கள் இனிவரும் தமிழ் விசைபலகைகளை ஒரு புதிய கோணத்தில் உற்று நோக்க வைக்கிறது.

#### 8) Present English/Tamil Phonetic Keyboard :



#### புதிய கண்ணோட்டத்தில் தமிழ் தொடு விசை பலகை:

ஒரு பொத்தானை தொட்டவுடன் Unicode Dependent Sign-ஐ பயன்படுத்தி ஒரு எழுத்தை சார்ந்த மற்ற எழுத்துக்களை மிக எளிதாக இயக்கி பதிவு செய்யலாம்.

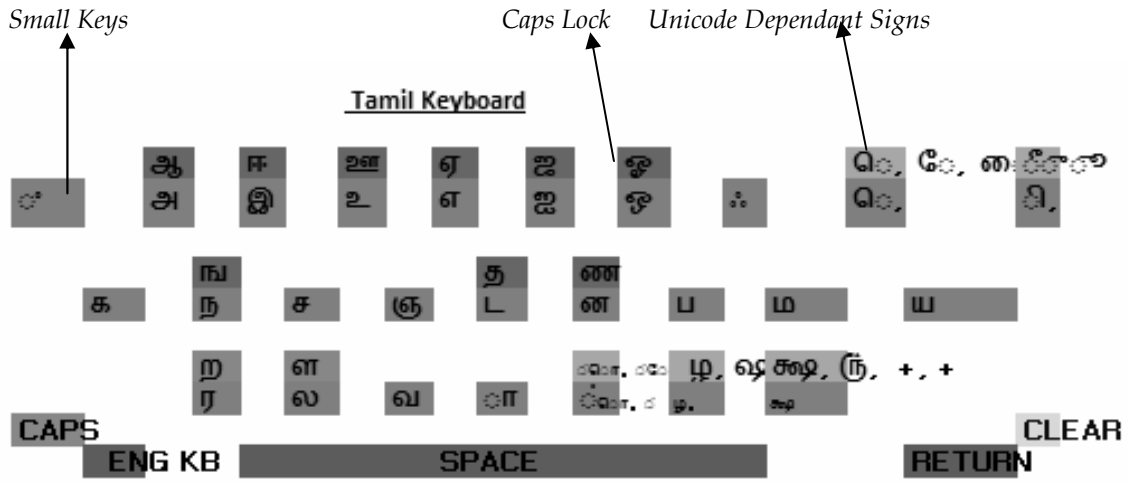
இந்த கண்ணோட்டத்தில் மிக முக்கியமான மாற்றம் இந்த கட்டுரையின் மூலம் முன் வைக்கப்படுவது, **Re-Assign the Keys away from traditional English QWERTY type writer concept.** இந்த பரிந்துரைக்கு முக்கியமான காரணம்.

இந்த முறையின் மூலம் எழுத்துக்களின் தொடர்ச்சியையும், எழுத்துகளின் நிலைகளை, அந்த தொடர்ச்சியின் மூலம் எளிதில் நினைவு படுத்த உதவும் என்பதும், அதுவும் தன்னிச்சை வாய்ந்த தமிழ் மொழியை ஆங்கில எழுத்துக்களின் வரிசை முறையில் இருந்து பிரித்து மிக எளிதான முறையில் வடிவம் செய்ய பரிந்துரைக்கப்படுகிறது.

பரிந்துரைக்கப்படும் தொடு தமிழ் விசைப்பலகை கீழ் வருமாறு :-

## 9) Touch Tamil Keypad – A Glance

(Away from QWERTY & Typewriter key position thoughts)



Hints :-

- **Caps characters** can be formed logically typing by repeating the characters specially, vowels.  
Ex : அஅ = ஆ, referred to Unicode 6.0 Tamil Chart or with caps lock.
- **Dependent Signs** may be called when a key is pressed for more than a second, which gives options to select Dependent signs.

## 10) பரிந்துரை

- 1) **Setup Workgroup** consisting experts from device manufacturers and volunteers.
- 2) Define a **Common and Open Standard** which can be unique keypad concept on all major platforms, specially the platforms which positioned as top 4 Device Manufacturers & OS Developers



- 3) A Keypad concept which can be **Continuous in Tamil Characters Positions** (Vowels, Consonant & Unicode Dependant sign Order) and easy to remember with positions, an unique model for Tamil)
- 4) Tamil keypad which can use **Dynamic Display Systems & Word dictionary**.

## 11) Download my Tamil apps for iPhone/iPad

|                       |   |  |
|-----------------------|---|--|
| Aathichoodi           | : | <a href="http://itunes.apple.com/in/app/aathichodi/id360404480?mt=8">http://itunes.apple.com/in/app/aathichodi/id360404480?mt=8</a><br>(Free)                  |
| Thamizhil Thirukkural | : | <a href="http://itunes.apple.com/in/app/tamizhil-thirukkural/id339147635?mt=8">http://itunes.apple.com/in/app/tamizhil-thirukkural/id339147635?mt=8</a> (Free) |
| Tamizh Quotes         | : | <a href="http://itunes.apple.com/in/app/tamizh-quotes/id407218410?mt=8">http://itunes.apple.com/in/app/tamizh-quotes/id407218410?mt=8</a> (Free)               |
| Tamizh Game           | : | <a href="http://itunes.apple.com/in/app/tamizh-game/id414426616?mt=8">http://itunes.apple.com/in/app/tamizh-game/id414426616?mt=8</a> (Paid)                   |
| iTunes Store keyword  | : | Tamil, Devarajan   |

# பாரதியின் பாடல்களுக்கு மின்னணு வழி வாசிப்புக்கருவி உருவாக்கம் – ஒரு கணினி மொழியியல் அணுகுமுறை

**டாக்டர் இரா. வேல்முருகன்**

ஆசிய மொழிகள் மற்றும் பண்பாட்டுத் துறை, தேசியக் கல்விக் கழகம்  
நன்யாங் தொழில் நுட்பப் பல்கலைக் கழகம், சிங்கப்பூர் 637 616

இன்றைய காலகட்டத்தில் கணினி தொழில்நுட்பத்தின் பயன்பாடு அளவிட முடியாத அளவிற்கு உயர்ந்து சென்றுகொண்டிருக்கிறது. இது மனித வாழ்வின் அனைத்துச் செயல்பாடுகளிலும் தமது பங்களிப்பைச் செய்து அதன் விளைவாய், இதன் பயன்பாட்டாளர்களுக்கு எண்ணிலடங்கா வசதி வாய்ப்புகளை அள்ளித் தந்த வண்ணம் உள்ளது. இது தன் பணியை மிக வேகமாகவும், துல்லியமாகவும், எந்தவிதமான பிழையுமின்றியும் செய்ய வல்லது. படித்தவர் முதல் பாமரர் வரை அனைவரும் இதன் பயனை அனுபவித்து வருகின்றனர்.

மொழி அறிவு எவ்வாறு அனைவருக்கும் அவசியமோ அதே போன்று கணினி தொழில்நுட்பத்துடன் தொடர்புடைய அறிவு ஒவ்வொருவருக்கும் மிகவும் அவசியமாகின்றது. கணினியறிவும், மொழியறிவும் கல்வியாளர்களுக்கு இரு கண்களாகப் போற்றப்படுகின்றன.

## மொழியும் கணினியும்

மொழியும் கணினியும் ஒன்றோடென்று நெருங்கிய தொடர்புடையவை. ஒன்று மற்றொன்றைச் சார்ந்துள்ளது. மொழி, கணினியில் பயன்படுத்தப்படுகின்றது. கணினியின் இயக்கம் மொழியைச் சார்ந்துள்ளது. இயற்கை மொழி கணினியில் பயன்படுத்தப்பட்டாலும் கணினிக்குரிய செயற்கை மொழியின் மூலம் கட்டளைகள் இடப்படுகின்றன. அதே சமயத்தில் இயற்கை மொழியின் உதவியால் கணினி இயங்குகின்றது. இவையல்லாது, இயற்கை மொழியைப் புரிந்து கொள்வதற்காக அவற்றை ஆராய்ந்து கணினியில் உள்ளீடு செய்து பேசவும், வாசிக்கவும், புரிந்து கொள்ளவும், எழுதவும் முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. இது உலகின் பல்வேறு மொழிகளில் நிகழ்வது போலத் தமிழ் மொழியிலும் நிகழ்ந்து கொண்டிருக்கிறது.

ஒருவர் எவ்வாறு மொழியின் ஆழ அகலங்களைப் புரிந்து கொண்டு, அம்மொழியின் பல்வேறு நுணுக்கங்களை உணர்ந்து செயல்புரிகின்றாரோ அதே போன்று கணினியும் செயல்புரிவதற்கான ஆய்வுகள் ஆங்காங்கே நிகழ்ந்த வண்ணம் உள்ளன.

## எழுத்துச்சோதனையும் இலக்கணச்சோதனையும்

தற்போது கணினியில் இயற்கை மொழிகளுக்குக் குறிப்பாக ஆங்கில மொழிக்கு எழுத்துச்சோதனையும் (Spell check), இலக்கணச் சோதனையும் (Grammar check) நடைபெறக் காண்கின்றோம். இவற்றின் உதவியால் ஒரு சொல்லைக் கணினியில் தட்டச்சு செய்யும் வேளையில், அதில் இடம் பெற்றுள்ள எழுத்துகள் சரியா அல்லது தவறா என அறிந்து சரியான எழுத்துகளைத் தெரிவு செய்து சரியான சொற்களைத் தட்டச்சு செய்ய வசதியளிக்கின்றது. இது போல, ஒரு வாக்கியத்தைத் தட்டச்சு செய்யும்போது அவ்வாக்கியம் சரியா அல்லது தவறா என இனம் கண்டு, தவறு நேரும் போது அந்தத் தவறை எவ்வாறு சரி செய்யலாம் என்பதற்கான அறிவுரைகளையும் வழங்கி வருகிறது.

## மொழி கற்றலும் கணினியும்

கணினியில் மொழி பயன்படுத்தப்பட்டாலும், கணினி ஒரு மொழியைக் கற்பிக்கும் சாதனமாக விளங்குகின்றது. இதன் உதவியால் மொழியின் நால்வகைத் திறன்களையும் மொழி கற்பவர் கற்றுணர முடியும். மொழி கற்றல் / கற்பித்தலுக்கு கணினியின் பயன்பாடு மிக இன்றியமையாத ஒன்றாக மாறி வருகிறது.

ஒரு மொழியை ஆசிரியரிடம் இருந்து கற்கும் பொழுது சில நன்மைகளும், சில குறைபாடுகளும் இருப்பது போலக் கணினி வழி மொழியைக் கற்கும் பொழுது சில நன்மைகளும் சில குறைபாடுகளும் இருக்கத்தான் செய்கிறது. ஒட்டுமொத்தமாகப் பார்க்கும்போது கணினிப் பயன்பாட்டால் மொழி கற்றல் / கற்பித்தல் செயல்பாடுகளில் பல நன்மைகள் இருப்பதை உணர முடியும்.

## இலக்கிய நுகர்வும் கணினியும்

கணினி என்னும் அரிய சாதனம், நாளொரு மேனியும் பொழுதொரு வண்ணமுமாய் தனது பங்களிப்பை எல்லாத்துறைகளுக்கும் செவ்வனே செய்து வருவது போல இலக்கியப் பயன்பாட்டிற்கும், அதன் சுவையை அனுபவிப்பதற்கும் அதிகப் பங்களிப்பைச் செய்து போற்றுதலுக்குரியதாக விளங்கி வருகிறது. இதன் மூலமாகக் கணினித்துறையும் இலக்கியத்துறையும் பல்வேறு பயன்களைப் பெற்றுத் திகழ்கின்றன.

பொதுவாக ஓர் இலக்கியம், அவ்விலக்கியம் உருவான மொழியைப் பேசும் மனிதர்களால் பெரிதும் கவரப்பட்டால், அவ்விலக்கியம் மற்ற அனைவராலும் நுகர்ந்து அனுபவிக்கும் வண்ணம் பிற மொழிகளுக்கு மொழிபெயர்ப்பு செய்யப்படுவது இயற்கை. பொதுவாகப் பாரம்பரியமான மொழி பெயர்ப்பின் மூலம் ஓர் இலக்கியம் மொழிபெயர்க்கப்படும் பொழுது, மூலமொழி இலக்கியத்தின் அனைத்துக் கூறுகளையும் மொழி பெயர்க்கப்படும் மொழியின் மொழிபெயர்ப்பில் கொண்டு வருவது சாத்தியமில்லை. எனவே மொழிபெயர்க்கப்பட்ட இலக்கியத்தை வாசிக்கும் வாசகர் மூலமொழி இலக்கியத்தின் சில கூறுகளை அறிய இயலாமல் போகலாம். மேலும், மொழிபெயர்க்கப்பட்ட இலக்கியத் தை நுகரும் பொழுது மொழிபெயர்ப்பாளரின் புரிதல் தன்மைக்கு ஏற்ப மொழி பெயர்க்கப்படுவதால் அவரின் புரிதலை மட்டுமே வாசகர்கள் அறிய முடியும். இதன் மூலம், அம்மூலமொழி இலக்கியத்தின் பல் வேறு புரிதல்களை அறிந்து கொள்ளக் கூடிய வாய்ப்புகளை மொழிபெயர்க்கப்பட்ட இலக்கியத்தை நுகரும் வாசகர் பெற முடியாமல் போய்விடும்.

இதே போன்று இலக்கியத்தை வாசிக்கும் அனைவரும் அவ்விலக்கியத்தின் அனைத்து விதமான புரிதல்களையும் புரிந்து கொள்ளமுடியாது. மேலும் ஒரு குறிப்பிட்ட இலக்கியம் சார்ந்த தொடர்பு இலக்கியங்கள் குறிப்பாக அவ்விலக்கியத்தின் பல்வேறு மொழிபெயர்ப்புகள், ஆய்வு முடிவுகள், திறனாய்வுக் கட்டுரைகள் போன்றவற்றை வாசிப்போர் எளிதாக, அவ்விலக்கியத்தை வாசிக்கும் பொழுதே பெறுவது என்பது இயலாத காரியம். மேலும், பாரம்பரியமிக்க இலக்கியப் பனுவல்கள், வாசிப்போரின் அனைத்து விதத் தேவைகளையும் ஒரே நேரத்தில் பூர்த்தி செய்ய முடியாது. இத்தகு சூழலில் தான் கணினியின் பங்கு மிகவும் அவசியமான ஒன்றாகிறது. கணினியின் உதவியால் உருவாக்கப்படும் இலக்கியம், அவ்விலக்கியத்தை முழுமையாக அனுபவிக்க விரும்பும் வாசகரின் அனைத்து விதத் தேவைகளையும் பூர்த்தி செய்யும் ஆற்றலைப் பெற்றுத் திகழும். அவ்வகை இலக்கியமே வாசிப்போருக்கு எல்லா வகைக் கூறுகளையும் வழங்க உதவுகின்றது.

ஆக, ஓர் இலக்கியத்தைப் பாரம்பரிய முறையில் நுகர்வது ஒரு வகை; அவ்விலக்கியம் மூலமொழி இலக்கியமாக இருந்தாலும் சரி, மொழிபெயர்ப்பு இலக்கியமாக இருந்தாலும் சரி அவற்றைக் கணினி வழி

நுகரும்போது இலக்கியத்தின் வீச்சு, வாசகரைச் சென்றடையும் போக்கு ஆகியவற்றில் பல மாற்றங்களை ஏற்படுத்தி இலக்கிய நுகர்வுக்கு ஒரு புதுப் பரிமாணத்தை வழங்குகின்றது.

இலக்கிய நுகர்வுக்குக் கணினியைப் பயன்படுத்தும் பொழுது, வாசகரின் அனைத்துத் தேவைகளையும், குறுகிய நேரத்தில், ஒரே வாசிப்பில் அளிப்பதற்கான வாய்ப்பை உருவாக்குவதால், இலக்கிய வாசிப்போரின் எண்ணிக்கை நாளடைவில் பெருகி வளரும் என்பது திண்ணம். இன்றைய அறிவியல் உலகில் இளம் வாசிப்பாளர்கள், கணினியின் பால் மிகுந்த பற்றுடையவர்களாகத் திகழ்கின்றனர். அதே சமயத்தில் பாரம்பரிய முறையில் இலக்கியம் வாசிப்போரின் எண்ணிக்கையும் குறைந்து கொண்டே வருகின்றது. இத்தகு சூழலில் இலக்கியத்தைக் கணினி என்னும் அரியாசனத்தில் அமரச் செய்து இலக்கியத்திற்குச் சிறப்பு செய்வது சாலப் பொருந்தும். அதுவே இலக்கியம் வாசிப்போர் மிகுந்த நாட்டத்தோடு இலக்கியத்தை நுகர வழிவகை செய்கின்றது.

### தமிழும் கணினியும்

இயற்றமிழ், இசைத்தமிழ், நாடகத்தமிழ் என முத்தமிழைப் பெற்ற நம் தமிழன்னையின் கைகளில் இன்று கணினித் தமிழ் என்ற நான்காவது தமிழும் புதிய வீச்சுடன் வலம் வந்துகொண்டிருக்கிறது. கணினி சார்ந்த தமிழாய்வுகள் பல்வேறு நிறுவனங்களில் மேற்கொள்ளப்பட்டு வருகின்றன.

செம்மொழியான தமிழ் மொழி இலக்கியப் பாரம்பரியத்தைப் பெற்றிருப்பதோடு, செவ்வியல் இலக்கியங்களையும், நவீன இலக்கியங்களையும் பெற்றுத் திகழ்கின்றது. இவ்வகை இலக்கியங்களைத் தமிழ் கூறும் நல்லுலகம் மட்டுல்லாது, பிறமொழி பேசுபவர்களால் கூட நுகர்ந்து அனுபவித்து மகிழ முடிகிறது. அந்த வகையில் கணினி வழியாக இலக்கியம் நுகரும் பாங்கைச் செழுமை பெறச் செய்வதன் மூலம் தமிழ் வளர்ச்சிக்கு அதிகமான பங்கினை ஆற்ற முடியும் என்னும் சீரிய நோக்கத்தை மனத்தில் கொண்டு, தமிழில் தேசியக்கவியாய், மகாகவியாய் வலம் வரும் பாரதியாரின் பாடல்களை உலக வாசிப்பாளர் மத்தியில் கொண்டு செல்ல வேண்டும் என்ற உயர்ந்த நோக்கில் அவரது ஒரு பாடலைக் கணினி வழி மின்னணு மயமாக்கம் செய்து பாடலில் உள்ள அனைத்துவகைக் கூறுகளையும் ஒரு சேர அனைவரும் நுகரச்செய்ய முயற்சிக்கின்றது இந்த ஆய்வுக்கட்டுரை.

பாரதியின் அச்சமில்லை அச்சமில்லை அச்சமென்பதில்லையே என்னும் பாடல் வரிகள் மாதிரி மின்னணுவழி வாசிப்புக் கருவி உருவாக்கத்திற்குப் பயன்படுத்தப் படுகிறது.

பாடல் நுகர்விற்குக் கீழ்க்காணும் கூறுகள் இணைக்கப்பட்டுள்ளன.

1. அச்சமில்லை அச்சமில்லை அச்சமென்பதில்லையே என்னும் பாடலை ஒரு சட்டகத்தில் (frame) மின் தட்டச்சு செய்தல்.
2. அவ்வாறு உருவாக்கப்பட்ட பாடலைத் தமிழைத் தாய்மொழியாகக் கொண்டவரும், சொற்களைச் சரியான முறையில் உச்சரிப்பவருமான ஒருவரைக் கொண்டு பிழையில்லாமல் வாசிக்கச் செய்து அவரது வாசிப்பைப் பதிவு செய்தல். அவ்வாறு வாசிக்கும் பொழுது வாசிப்புக்கு ஏற்ப ஒவ்வொரு வரியும் திரையில் தோன்றுதல்.
3. முகப்புத்திரையில் தமிழ் விளக்கம், ஆங்கில விளக்கம், ஆங்கில மொழிபெயர்ப்பு என மூன்று பொத்தான்களை உருவாக்குதல். அப்பொத்தான்களை வாசிப்போர் அழுத்தும் வேளையில் தமிழ் விளக்கமும், ஆங்கில விளக்கமும், ஆங்கில மொழிபெயர்ப்பும் தோன்றும். மேற்கண்ட மூன்றில் வாசிப்போர் எதைத் தெரிவு செய்து அழுத்தினாலும் அவரது விருப்பத்திற்கு ஏற்ப அக் கூறுகள் திரையில் தோன்றும். அதோடு இவை, வாசிப்போரின் வேண்டுதலுக்கு ஏற்பவோ அவர்களின் பிற தேவையின் அடிப்படையிலோ தோன்றும். அவ்வாறு தோன்றும் பொழுது, அவை எழுத்து வடிவில்

மட்டுமல்லாது, ஒலி வடிவிலும் வந்தமையும். தமிழ் விளக்கம் அப்படியே வாசிக்கப்படும். ஆங்கில விளக்கம் அல்லது ஆங்கில மொழிபெயர்ப்பும் அப்படியே வாசிக்கப்படும்.

4. இவை தவிர ஆங்காங்கே சில இடங்களில் பச்சை நிறக் குறியீடுகள் இருக்கும். அவற்றை அழுத்தும் போது, அவை பற்றிய சிறப்பு விளக்கங்கள் தோன்றும். அதாவது குறிப்பிட்ட ஒரு சொல்லை அழுத்த அச்சொல் தொடர்புடைய அனைத்துச் செய்திகளும் தோன்றும். அதாவது,

**இச்சகத்து** ஸோரெலாம் எதிர்த்து

நின்ற போதிலும்

என்னும் வரியில் **இச்சகத்து** என்ற சொல்லை அழுத்தும் பொழுது, கீழ்க்காணும்

தரவுகள் தோன்றும்.

- 1) சொல்லின் உச்சரிப்பு
- 2) சொல்லின் இலக்கணக் கூறு
- 3) சொல்லின் பொருள்
- 4) சொல்லினைப் பிரித்தெழுதல்
- 5) சொல்லினை வேறு ஒரு சூழலில் பயன்படுத்திப் புரியச் செய்தல்

மேற்கூறிய விளக்கங்கள் தமிழிலும் ஆங்கிலத்திலும் தோன்றும். இதன் மூலம், இப்பாடலைத் தமிழ் அறிந்தவர் மட்டுமல்லாது கல்வியாளர்கள் அனைவரும் வாசித்து அனுபவிக்க முடியும்.

5. பண்பாடு தொடர்பான சொற்கள் இருப்பின் அச்சொல் தொடர்பான பண்பாட்டு விளக்கங்கள் இடம் பெறும். தெரிந்தெடுக்கப்பட்ட பாடலில் பண்பாட்டுச் சொற்கள் ஏதும் இடம் பெறவில்லை என்பதால் பழமையான சொற்களுக்குச் சிறப்புக் கவனம் செலுத்தி அச்சொல்லின் விளக்கம் அளிக்கப்படும். சான்றாக,

**துச்சமாக எண்ணி நம்மைத் தூறு செய்த போதிலும்**

என்னும் வரியில் தூறு என்ற சொல்லுக்குக் 'கெடுதல்' என்ற பொருளைக் கூறி விளக்கம் கொடுத்தல்.

6. ஒரு சொல் பல பொருள் தோன்றும் மொழிக்கூறுகளுக்கு விளக்கமளித்தல். இதில் அச்சொல் அல்லது தொடருக்குரிய பல பொருட்களை விளக்கி விட்டு இப்பாடலில் இடம் பெற்றுள்ள சூழல் பொருளை விளக்குதல். சான்றாக,

**துச்சமாக எண்ணி நம்மைத் தூறு செய்த போதிலும்**

என்னும் வரியில் **எண்ணி** என்ற சொல்லுக்கு இரு பொருள்கள் உள்ளன. முதல் பொருள் ஒன்று, இரண்டு என எண்ணுதல்; மற்றொன்று சிந்தித்தல். இவ்விரு பொருள்களையும் விளக்கி, மேற்கண்ட பாடலில் 'சிந்தித்தல்' என்னும் பொருளிலேயே 'எண்ணி' என்னும் சொல் பயன்படுத்தப்பட்டுள்ளன என விளக்குதல்.

7. மேற்கண்ட பாடலில் பயன்படுத்தப்பட்டுள்ள இலக்கிய உத்திகளான எதுகை, மோனை, உவமை, உருவகம், அணி போன்ற கூறுகளை வரிசைப்படுத்தி வாசிப்போருக்கு உதவுதல்.
8. அச்சமில்லை அச்சமில்லை என்ற பாடல் தொடர்பான ஆய்வுகள், திறனாய்வுகள், விளக்கங்கள் போன்றவை ஆங்கிலத்திலும் தமிழிலும் வழங்கி அவை உள்விளக்கங்களாக (Hypertext) கொடுக்கப்படுதல். அவை வாசிப்போரின் வேண்டுதலின் பேரிலேயே திரையில் தோன்றும்.

9. மேற்கண்ட பாடல்களில் இடம் பெற்றுள்ள அனைத்துச் சொற்களுக்கான விளக்கங்கள் பட்டியலாக இடம் பெற்றிருக்கும். அவற்றைப் பார்த்து வாசிப்போர் பொருள் புரிந்து கொள்ள முடியும்.
10. இறுதியாக, வாசிப்போரின் கருத்துகளைப் பதிவு செய்யும் முகத்தான், ஒரு பகுதி ஒதுக்கப்படும். அப்பகுதியில் வாசிப்போர் தங்களது கருத்துகளைப் பதிவு செய்யலாம்.

### முடிவுரை

இம் மின்னணு சாதனம் பாரதியின் பாடலில் உள்ள அனைத்து வகையான கூறுகளையும் ஒரே வாசிப்பில் எளிதாக, ஒரே இடத்தில் இருந்து கொண்டு அறிந்துகொள்ள வாய்ப்பளிக்கின்றது. இதன் மூலம் தமிழைத் தாய்மொழியாகக் கொண்டவர்கள் மட்டுமல்லாது பிற மொழி பேசும் மக்கள் அனைவரும் இலக்கியத்தைச் சுவைத்து மகிழ முடியும். இதே முறையைப் பின்பற்றிப் பாரதியின் பாடல்கள் அனைத்தையும் மின்னணுமயமாக்கம் செய்தால், நவீன யுகத்தில் கணினியின்பால் ஈர்ப்புடையோர் அனைவரும் பாரதியின் இலக்கிய இன்பத்தைச் சுவைத்து அனுபவிக்க முடியும் என்பது திண்ணம்.

# பல்லாடக வழி அற இலக்கியங்களைக் கற்றல், கற்பித்தல்

## (Teaching and learning in ethical literature through multimedia)

முனைவர் வா.மு.சே. முத்துராமலிங்க ஆண்டவர்,  
தமிழ் இணைப் பேராசிரியர், பச்சையப்பன் கல்லூரி, சென்னை  
sethuandu@yahoo.co.in | sethuandavar@yahoo.co.in

தமிழ் மொழி, செம்மொழியாக அறிவிக்கப்பட்ட பிறகு, பயன்பாட்டு அடிப்படையில் பல ஆய்வுகள் நிகழ்ந்து வருகின்றன. கணினித் தமிழ், நவீன இலக்கியங்களுக்கு எந்த அளவிற்குப் பயன்பட்டதோ, அந்த அளவிற்குத் தொன்மையான இலக்கிய ஆய்விற்கும் பயன்படுகிறது. இந்தக் கண்ணோட்டத்தில் கணினி வழியாக ஆற்ற வேண்டிய பற்பல பணிகள், நம் முன் நிற்கின்றன.

கணினியில் மொழி, இலக்கியச் செயல்பாடுகளை இரண்டு வகையாகப் பிரிக்கலாம்.

1. கணினிக்குச் செயற்கை மொழிக்கு மாற்றாக, இயற்கை மொழியினை ஊட்டுவதற்குச் செய்யப்படும் ஆராய்ச்சி முயற்சிகள் தொடர்பானவை.
2. கணினி வழியாக, இலக்கண, இலக்கிய, மொழிசார் பணிகளைப் பல்லாடக அடிப்படையில் எவ்வாறு மேற்கொள்வது என்பதைத் திட்டமிடுதல்.

பழந்தமிழ் இலக்கியங்களைப் பல்லாடக வழி அறிமுகப்படுத்துவதற்கு இது வரை ஆராய்ச்சியாளர்கள் பலர், தன்னார்வ அடிப்படையில் செயல்புரிந்துள்ளனர். செம்மொழித் தமிழாய்வு நிறுவனமும் தொல்காப்பியத்தையும் சங்க இலக்கியத்தையும் குரல் வழி அறியும் ஒலிப் பேழைகளை உருவாக்கியுள்ளது. காணொலி அடிப்படையில் அசைவூட்டத்தின் அடிப்படையிலும் சில முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன.

தொலைக்காட்சி வழியாகவும் இணைய வழியாகவும் நேரடியாக மொழி, இலக்கியம், இலக்கணம் ஆகியவற்றைக் கற்கும் - கற்பிக்கும் பணிகளும் நிகழ்ந்து வருகின்றன. தமிழ்ச் சங்கத்தினரும் தமிழ் ஆர்வலர்களும் சில முயற்சிகளில் ஈடுபட்டுள்ளனர். அவர்களின் தன்னார்வ முயற்சிகளைத் தரப்படுத்துவதற்கும் மேம்படுத்துவதற்கும் நாம் என்ன செய்ய வேண்டும் என்றும் இக்கட்டுரை ஆராய்கிறது.

அறிவுசார் முயற்சிகளையும் தொழில்நுட்பம் சார் முயற்சிகளையும் ஒருங்கிணைக்க, நாம் என்ன செய்யலாம்? அற நெறி இலக்கியங்களை வயதுக்கு ஏற்றபடி எவ்வாறு அறிமுகப்படுத்தலாம்? எடுத்துக்காட்டாக, ஆத்திசூடி, கொன்றை வேந்தன் ஆகியவற்றைத் தொடக்கப் பள்ளிகளுக்கும் நன்னெறி, மூதுரை போன்வற்றை உயர்நிலைப் பள்ளிகளுக்கும் திருக்குறள் போன்ற நீதி இலக்கியங்களை மேல்நிலைப் பள்ளிகளுக்கும் ஏனைய பதினெண் கீழ்க்கணக்கில் உள்ள அற நெறி இலக்கியங்களை தொடக்கக் கல்வி, உயர் கல்வி, இளங்கலை, முதுகலைப் பட்டப் படிப்புகளுக்கும் பல்லாடக வழி .

இன்றைய கல்விசார் சூழலில் அற நெறி என்பது, முற்றிலும் புறக்கணிப்பட்டு, வணிகம் சார்ந்த முறை முன்வைக்கப்படுகிறது. இதற்கு மாற்றாக, பண்பாட்டுடன், அற உணர்வுடன் கூடிய சமூகக் கல்வியை நாம் கற்பிக்க முயல்வோம்.

பல்லாடக வழியே பழந்தமிழ்ப் பணுவல்களைக் கற்கும் பொழுதும் கற்பிக்கும் பொழுதும் ஏற்படுகின்ற சிக்கல்கள் பல. சங்க இலக்கியங்களைப் பல்லாடக வழி கற்பிக்கும் பொழுது அதிகப் புரிதலை உருவாக்கலாம். காதல், வீரம், கொடை, புகழ், வள்ளல் தன்மை போன்ற பாடல்களின் கருத்துகளை அறிவுறுத்தலாம். படக் காட்சி அடிப்படையில் பட வரைகலை விளக்கத்துடன் எளிமையான சொற்களைக் காட்சிவழி அறிமுகப்படுத்தலாம்.

மாணவர்களுக்குப் புரியும் வகையில், பழந்தமிழ்ச் சொற்களுக்கு இணையான புதிய சொற்கள் அறிமுகம், வினா விடை, அருஞ்சொற்பொருள்கள், முரண் சொற்கள், சொல்வளம் போன்ற அடிப்படைகளில் பல்லாடக வழி சங்க இலக்கிய கற்றல் பணியினைத் தொடங்கலாம்.

இன்றைய நவீன மொழியிலாளர்கள், தாய்மொழி கற்றலுக்கான புதிய உத்திகளின் அடிப்படையில் திட்டங்கள் அமைய வேண்டும். சங்கம் மருவிய கால அறநெறி இலக்கியங்களை மாணவர்களுக்குப் பல்லாடக வழி கற்பிக்க, தனித்த நடைமுறைகளைத் திட்டமிட வேண்டும். முதலில் பல்லாடக வழி அறநெறி இலக்கியங்களைக் கற்பிப்பதற்குப் பாடத் திட்டங்களைத் உருவாக்க வேண்டும்.

தமிழகத்தில் தமிழைத் தாய்மொழியாகக் கொண்ட மாணவர்களுக்கு ஏற்ற வகையில் பாடத் திட்டங்கள் அமைய வேண்டும். இப்பாடத் திட்டங்கள், மாணவர்கள் அறநெறி இலக்கியங்களில் ஆர்வம் கொண்டு, தன்னார்வ அடிப்படையில் கற்றுக்கொள்ளுமாறு அமைய வேண்டும். இவ்வகையான திட்டமிடலுக்குப் பல்லாடகத்தையும் துணைக் கருவிகளாகக் கொண்டு இத்திட்டத்தினைச் செம்மையாகச் செயல்படுத்த முடியும்.

தொடக்க முயற்சியாக, திருக்குறள் என்ற அற இலக்கியத்தைத் தொடக்கப் பள்ளி முதல் அறிமுகம் செய்யும் அலகுகளை ஆராயலாம். அறநெறி இலக்கியங்கள் தமிழ்நாடு அரசின் பாடத் திட்டத்தில் இடம் பெற்றிருக்கின்றன. ஆனால் இப்பாடத் திட்டங்களை அலசி ஆராய்ந்து, தேவையானவற்றை ஏற்று, காலத்திற்கேற்ற மாறுதல்களுடன் புதிய முறையில் உருவாக்க வேண்டும். தொடக்கப் பள்ளி, உயர்நிலைப் பள்ளிகளில் திருக்குறளும் நாலடியாரும் தொடர்ந்து அறிமுகப்படுத்தப்பட்டு வருகின்றன.

தொடக்கப் பள்ளிகளில் திருக்குறள், ஐந்தாம் வகுப்பு வரை 10 குறள்கள் இடம் பெற்றுள்ளன. 6 முதல் 8ஆம் வகுப்பு வரை 20 திருக்குறள்கள் இடம் பெற்றுள்ளன. 9 முதல் 12ஆம் வகுப்பு வரை 50 திருக்குறள்கள் இடம் பெற்றுள்ளன. இதே போலவே பழமொழி நானூறு என்ற அறநெறி இலக்கியமும் நாலடியாரும் பள்ளிகளின் வகுப்பு நிலைக்கேற்ப, பாடல்களின் எண்ணிக்கை அமைந்திருக்கும். ஆனால் பல்லாடக வழி, பாடங்களைக் கற்கும்பொழுதும் கற்பிக்கும் பொழுதும் இவை முழுமையாக மாற்றப்பட வேண்டும்.

8ஆம் வகுப்புக்குப் பிறகு தான், மாணவர்களுக்கு நேரடியான செயல் கல்வியைத் தொடங்க வேண்டும். இதுவரை அமைந்துள்ள கல்வி முறை, மனப்பாடத் திறனை வளர்ப்பதாக அமைந்து, மாணவரின் ஆக்க சிந்தனைக்கு எதிராக அமைந்துள்ளது. தொடக்கப் பள்ளிகளில் திருக்குறளை நேரடியாகக் கற்பிக்காமல், கதை வடிவில் கற்பிக்க வேண்டும். கதை வழி கற்பிப்பதற்குப் பல்லாடக வழி காட்சி உருக்களை அசை உருக்களை உருவாக்கி, கதை வழி மாணவர்களுக்கு அறச் சிந்தனையைப் புரிய வைத்து, திருக்குறளை இறுதியாகக் கூறவேண்டும்.

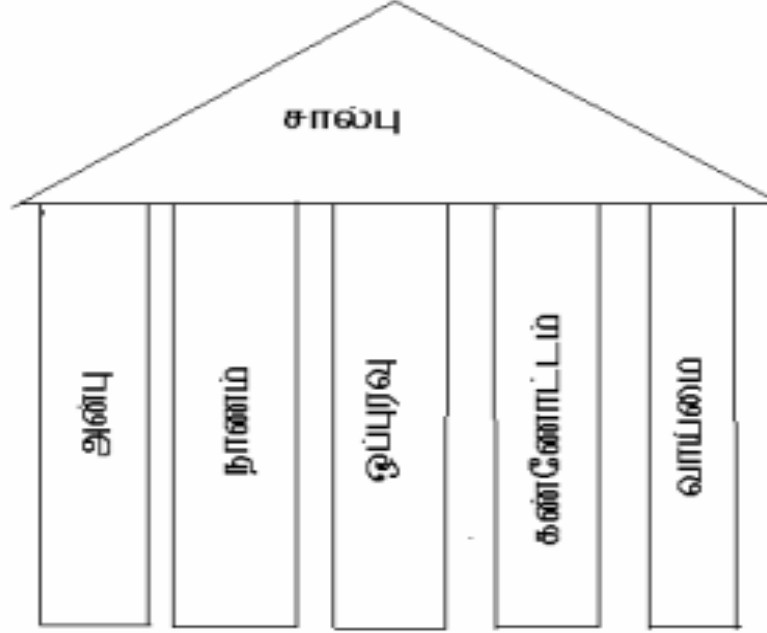
ஒரு தட்டிலே காய், இன்னொரு தட்டிலே கனி, காய்-கெட்ட சொற்கள் கனி-நல்ல சொற்கள். மாணவரிடம் விளக்கும் பொழுதே மாணவரே கனி-நல்ல சொல், காய்-கெட்ட சொற்கள் எனக் கூறும்படி செய்து, திருக்குறளை விளக்கலாம்.

இனிய உளவாக இன்னாத கூறல்  
கனியிருப்பக் காய்கவர்ந்தற்று<sup>1</sup> (குறள் - 100)



திருக்குறளைப் படவிளக்க அடிப்படையிலும் அசை உருக்களை உருவாக்கி மாணவர்களுக்குக் கற்பிக்கலாம். உதாரணமாக, சான்றான்மையின் அதிகாரத்தில் சால்பு தூண்களாக அன்பு, நாணம், ஒப்புரவு, கண்ணோட்டம், வாய்மை ஆகியவற்றைக் குறிப்பிடுகிறார்.

அன்பு நாண் ஒப்புரவு கண்ணோட்டம் வாய்மையொடு  
ஐந்துசால்பு ஊன்றிய தூண்<sup>2</sup> (குறள் - 983)



தெரிந்ததைக் கொண்டு, தெரியாததைக் கற்பதுதான் கற்றல் (known to unknown). சால்பு என்னும் சொல்லுக்குரிய பொருளை மாணவர்களுக்குப் புரிய வைக்க, அதை மாளிகையுடன் ஒப்பிட்டு, அம்மாளிகை உறுதியாக இருக்கப் பயன்படும் தூண்களைப் போல, மனிதன் சான்றோனாக வாழத் தேவையான பண்புகளாக அன்பு, நாணம், ஒப்புரவு, கண்ணோட்டம், வாய்மை ஆகியவற்றைக் குறிப்பிட்டு விளக்கலாம். இவற்றைப் பல்லாடக வழி காட்சிப்படுத்தி, வரைகலை அடிப்படையில், மாணவர்களுக்கு அறநெறியை விளங்க வைப்பது எளிது. இவ்வாறு பல்லாடக வழி காட்சிப்படுத்துவதற்கு உரிய திருக்குறள்களைக் கண்டறிய வேண்டும்.

அன்பு, பண்பு, ஆசை, நன்மை, தீமை போன்ற பண்புசார் சொற்களை மாணவர்களுக்கு விளங்க வைப்பது அரிது. அதற்கு மாற்றாக, திருக்குறளில் உள்ள பெயர்ச் சொற்கள், வினைச் சொற்கள் அறிமுகம் முதலில் அமைய வேண்டும். பல்லாடக வழி கற்பித்தலுக்குப் பெயர்ச் சொற்களே பெரிதும் பயன்படும். திருக்குறளில் இடம் பெற்றுள்ள விலங்கு, பறவை தொடர்பான பெயர்ச் சொற்களை முதலில் பட்டியலிடலாம். விலங்கையோ, பறவையையோ முதலில் அறிமுகப்படுத்த வேண்டும். அதற்குப் பிறகு, திருக்குறளில் எந்தச் சூழலில் பயன்படுத்துகிறார் என்பதை விளக்கி, குறள்வழி பெறப்படும் அற உணர்வினைப் பிறகு விளக்க வேண்டும்.

உதாரணத்திற்கு, (புலி - பசு) பசுத்தோல் போர்த்திய புலி என்ற இல்பொருள் அணியினை விளக்குவதற்குப் பல்லாடக வழி கற்பித்தல், பெரிதும் உதவும்.

வலியின் நிலைமையான் வல்லுருவம் பெற்றம்  
புலியிந்தோல் போர்த்துமேய்ந் தற்று<sup>3</sup> (குறள் - 273)

பெயர்ச் சொற்களில் பறவைகள் தொடர்பாக, (கொக்கு - மீன்) கொக்கு, மீனை எவ்வாறு கொத்துகிறதோ, அது போல், வாய்ப்பு வரும்போது அதைப் பயன்படுத்த வேண்டும். கொக்கு, மீனைக் கொத்துவது போல், வரைபடம் வரைந்து, அதைக் காட்சி வழியில் அறிமுகப்படுத்தினால், மாணவர்களுக்கு இக்குறளின் பொருள் எளிமையாகப் புரியும்.

கொக்கொக்கக் கூம்பும் பருவத்து மற்றதன்  
குத்தொக்க சீர்த்த இட்த்து<sup>4</sup> (குறள் - 490)

பல்லாடக வழி மொழிக் கற்றல் தொடர்பான ஆராய்ச்சிகள் நிறைய நிகழ்ந்துள்ளன. ஆனால் அந்த ஆராய்ச்சியின் அடிப்படையில் தற்காலத் தகுநிலைக்கு ஏற்ப, பல்லாடக வழி இலக்கியம் கற்பித்தலுக்கான நிலை உருவாக்கப்பட வேண்டும். இவ்வடிப்படையிலேயே பல்லாடக வழி சங்க இலக்கியமோ, அறநெறி இலக்கியமோ கற்பிக்கப்படவேண்டும்.

இதுவரை திருக்குறள் தொடர்பான பல்லாடக வழி கற்பித்தலுக்கு உரிய சில அணுகுமுறைகள் சுட்டப்பெற்றன.

முனைவர் எல். இராமமூர்த்தி அவர்கள், பல்லாடகக் கருத்தாடல், பின்வரும் ஐந்து நிலைகளை உள்ளடக்கியதாக அமையும் எனக் குறிப்பிடுகிறார்.<sup>5</sup>

- கருத்துப் பரிமாற்றம் (interaction)
- உடனடி விளக்கம் (immediate feedback)
- தவறு பற்றிய விளக்கம் (error analysis)
- தானே திருத்துதல் (self correction)
- பாராட்டுகள் (reinforcement)

மேலும் பல்லாடகக் கருத்தாடல் வழி ஆசிரியர் விளக்கம் (tutorial )

அடிப்படையில் திருக்குறளைக் கற்பிப்பதற்குத் திட்டங்களை வரையறுக்க முனையலாம். மேலும் ச. இராசேந்திரன் அவர்கள், பல்நோக்கு ஊடகம் பற்றிக் குறிப்பிடும்பொழுது, பின்வருமாறு குறிப்பிடுகிறார்.<sup>6</sup>

ஒருங்கிணைந்த கணினிவழி கற்றல், இரு முக்கியமான தொழில்நுட்ப முன்னேற்றங்களை அடிப்படையாகக் கொண்டு அமைந்தது.

1. பன்னோக்கு ஊடகக் கணினிகள் (Multimedia computers)
2. இணையம் (Internet)

பன்னோக்குத் தொழில்நுட்பம், ஒரு இயந்திரத்திலேயே பனுவல், வரைபடம், ஒலி, உயிரியக்கம், கட்டில் காட்சி போன்ற பல ஊடகங்களுடன் தொடர்புகொள்ள வசதி செய்தது. பன்னோக்கு ஊடகம், உயர் ஊடகத்தை உள்ளடக்கிய காரணத்தால் மிகச் சக்தி வாய்ந்ததாய் அமைந்தது. பன்னோக்கு ஊடகத் திறன்கள் / வழிமுறைகள் எல்லாம் ஒன்று சேர்க்கப்பட்டு, குறிப்பாணைச் சுண்டிக் கற்பவர்கள் தங்கள் வழியில் மிதக்க / பவனி வர வகை செய்யப்பட்டது.

இக்கருத்து, குறிப்பிடத்தகுந்தது.

ஒரு இலக்கியத்தின் தகவல்களை மட்டுமே தருவது பல்லாடக வழி கற்பித்தலாகாது. இலக்கியத்தின் தகவல்களின் அடிப்படையில் மாணவர்களின் சிந்தனைத் திறனை மேம்படுத்த, தூண்டுகோலாக அமைவதே சரியான கற்பித்தல் முறையாகும். இதற்கெல்லாம் மொழி பற்றிய அறிவும் சூழல் அறிவும் இலக்கியத்தைப் பயன்பாட்டில் பொருத்தும் செயல்பாட்டு அறிவும் அவசியமானவை. மொழித் திறனைப்

பல்லாடக வழி மேம்படுத்தி, அதன்வழி இலக்கியத் திறன்களை மேம்படுத்த வேண்டும். அப்பொழுதுதான் பல்லாடக வழி அறநெறி இலக்கியங்கள் கற்றல், கற்பித்தல் சிறப்பாக அமையும்.

**துணை நின்ற நூல்கள்:**

1. திருக்குறள் 100
2. திருக்குறள் 983
3. திருக்குறள் 273
4. திருக்குறள் 490
5. மொழியும் அதிகாரமும், எல்.இராமமூர்த்தி, 1999
6. தமிழியல் ஆய்வு - இன்றைய போக்குகள், இணையவழி அயல்நாடுகளில் தமிழ் கற்றல், கற்பித்தல், 2002

# Kuralagam - Concept Relation based Search Engine for Thirukkural

*Elanchezhiyan. K, T V Geetha, Ranjani Parthasarathi & Madhan Karky*

*Tamil Computing Lab (TaCoLa)*

*Department of Computer Science and Engineering*

*College of Engineering Guindy, Chennai – 600025*

*E-mail: chezhiyank@gmail.com, madhankarky@gmail.com*

## **Abstract**

Thirukkural is one of the most popular literatures in Tamil Language. Thirukkural is being quoted in speeches, news articles, blogs, micro-blogs and has a very strong reach in the Internet. Various interpretations of Thirukkural have been proposed by eminent Tamil scholars. This paper aims at presenting the world's first conceptual search framework for Thirukkural. The Framework uses CoReX [1]; a concept relation based indexing and presents a ranking model based on concept strength and popularity of Thirukkural, obtained by a Thirukkural statistic crawler. The search Framework is evaluated using Average Precision and Mean Average Precision (MAP) was found to be 0.83 compared to 0.52 with traditional keyword based search.

## **1. Introduction**

Thirukkural is the one of the outstanding accomplishment of Tamil literature. It had been translated in many languages. Thirukkural has totally 133 chapters. These are classified in to Aram (Virtue), Porul (Wealth) and Kamam or Inbam (Love). Each chapter has ten Thirukkural; Thirukkural in the form of couplets illustrates various aspects of life. Most of the present day Thirukkural search engines are keyword-based and Bilingual Keyword-based. Thirukkural search engines are available for Tamil and English language. Those keyword-based search engines fail to satisfy the user requirements. For example, in keyword-based search user won't get the result for the common word "பணம்" (Money). For the reason that actual keyword "பணம்" was not used in the Thirukkural.

The Kuralagam is a Conceptual and bilingual Thirukkural search engine. It is designed to clear up the complication in the traditional Thirukkural Search engine using CoReX Frame work. CoReX is designed such that the documents retrieved through it are semantically relevant to the query. The data structure used by the CoReX helps in storing concepts of terms rather than storing just words, thereby retrieving Thirukkural that are semantically relevant to the query. The main purpose of such indexing techniques is cross lingual information retrieval by an intermediate representation called the Universal Networking Language [2] (UNL). The Universal Networking Language is an electronic language for computers to express and exchange information. Kuralagam search system fetches Thirukkural based on keywords, concepts and expanded query words using the Concept Based Query Expansion technique using CoReX Framework.

In this paper we propose Kuralagam, a Concept Relation Based Thirukkural Search. Kuralagam aims at understanding the Thirukkural and its meaning by indexing them based on concepts and their relations rather than indexing the keywords and their frequencies. The Kuralagam was implemented and tested with 1330 Thirukkurals with 4 explanations (Kalaingar Karunanidi, Mu Va, Soloman Poppaiya, and G.U.Pope). The search results were compared against traditional keyword based search for precision and relevance.

## **2. Background**

CoReX [1] is a concept based semantic indexing technique used to index Universal Networking Language (UNL) expressions. CoReX retains the semantics captured by the UNL expressions. Since UNL expressions are stored as graphs, CoReX uses graph properties to index the UNL graphs. CoReX considers the out degree of each node and frequency of the same for indexing which helps in capturing the relations between the concepts in UNL expressions thereby retaining the semantics of the same. CoReX is simple and efficient and helps in retrieving documents which are semantically relevant to the query. The Thirukkural popularity score is computed by giving a Thirukkural sequentially to the web and finding its frequency distribution across the popular blogs, news articles, social nets etc.

## **3. Thirukkural Search Framework**

Thirukkural search framework presented in figure1 can be divided into two major divisions, online and offline, in terms of the time of processing. This section describes the various components of the Thirukkural search in detail.

### **3.1 Offline Processing**

The offline process comprises indexing Thirukkurals and their interpretations and crawling the web for usage of each Thirukkural.

#### **3.1.1 Web Crawler**

A Thirukkural statistics crawler crawls the news and blog documents on the web to find the usage of each individual Thirukkural. The usage on internet is recorded for measuring the popularity score for each Thirukkural, which is explained in detail later.

#### **3.1.2 En-conversion**

Here a Thirukkural and its meaning are passed to a rule based system to identify the various concepts in the Thirukkural and the rules are used to identify one of the 44 UNL relations [2]. Enconversion [4] uses the Morphological Analyser [3] to recognize various morphological suffixes of a word and uses this information along with syntax and semantics to identify the relationship between concepts. UNL graphs are generated for every sentence constituent. The UNL graph is then sent to CoReX indexer along with information such as Thirukkural Number, positional index and original keyword, its frequency in the document etc.

#### **3.1.3 Indexer**

The Kuralagam Indexer is designed based on CoReX Techniques. The Indexer stores and manages the UNL graphs in two different indices. Concept only index (C index), and Concept-Relation-Concept

index (CRC index) are the two indices maintained by the indexer. The UNL graphs are stored in the indices by their concept for efficient retrieval.

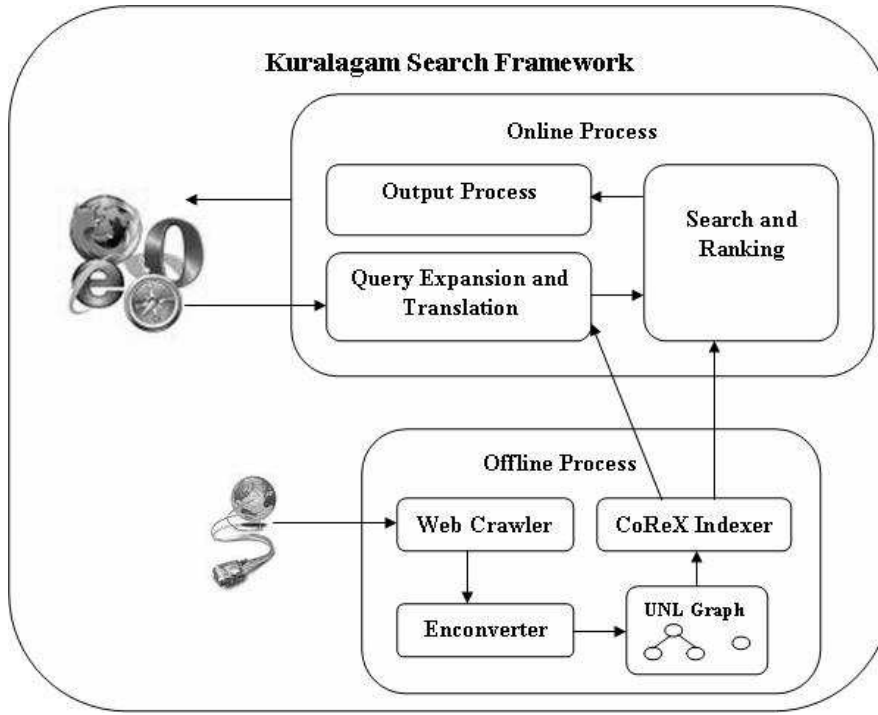


Fig 1: Kuralagam Search Framework

### 3.2 Online Processing

A user's query is processed, converted to UNL graph(s), expanded and sent to a search and ranking module, where the Thirukkural(s) that match the concept relation similarity are ranked and sent for output processing. The output processing module displays the retrieved Thirukkural(s) with its meaning and sends them to the user.

#### 3.2.1 Query Translation and Expansion

A user query is first sent to Query Translation module. Query Translation module converts the user query to UNL graph. The Concepts in UNL graph are sent to the Query Expansion module. Query Expansion uses CRC (Concept Relation Concept) CoReX indices to fetch similarity thesaurus and co-occurrence list to populate the Multi list Data Structure.

#### 3.2.2 Search and Ranking

The functionality of searching and ranking is to fetch the Thirukkural number and its details. Thirukkural(s) for a given query are fetched using the two types of concept relation indices namely CRC and C. The query concept is expanded using related CRC indices pointing to the query concept. This helps in retrieving many Thirukkural(s) conceptually related to the query. This kind of conceptual retrieval is highly impossible with key word Thirukkural search engines. The ranking is done by giving priority to the indices in the order CRC>C. The ranking is also based on the usage score and frequency occurrence of the query concept. Hence the search results are based not only on the query

term but also on the concepts related to the query term. The search results and performance analysis is discussed in the next section.

## 4. Kuralagam Search Results & Analysis

Kuralagam is a conceptual search framework for Thirukkural. Kuralagam, unlike traditional keyword based searches, identifies the concepts in a Thirukkuaral and their relationship with each other. All 1330 Thirukkural and their meaning were crawled, enconverted and indexed for search.

### 4.1 Tabbed Layout

In this paper, we propose a Tabbed Layout for displaying the results to the user. The Tabbed layout shown in figure 2, displays the results in 3 tabbed boxes to a class of results based on the concepts and relationship between concepts. Figure 2 displays the results for the query நட்பின் சிறப்பு (*Natpin sirappu*). The first cell displays the results of the concepts that contain the actual keywords which are sorted by the relation they have between them. Second tab identifies results that contain concepts of actual keywords, relation between them and displays the results corresponding to நட்பு பெருமை (*Natpu Perumai*). The third cell is based on expansions of the query. Here results corresponding to நட்பு துன்பம் (*Natpu thunbam*), நட்பு கொள் (*natpu koL*), நல்ல நட்பு (*Natpu nalla*) etc are displayed. The snapshot presented in figure 2 is from our engine implemented from the CoReX framework.

### 4.2 Performance Evaluation

The accuracy of the Thirukkural search engine was measured using the Precision. Precision can be computed using the formula given below [5]. We compute the precision for the first 5, 10 and 20 Thirukkural. The average precision and mean average precision for a set of queries will indicate the performance of the system.



Fig 2: Tab Layout

The comparisons between concept based search and keyword based search were measured using Average Precision methodology and the result is shown in figure 3.

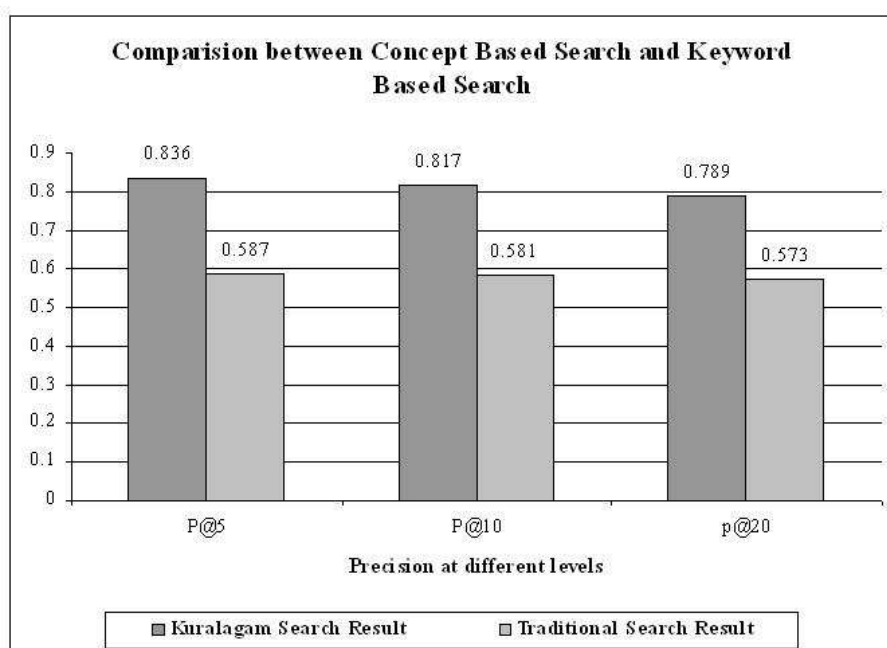


Fig 3: Average Precision Comparison

## 5. Conclusion and Future Work

This paper describes Kuralagam, a framework for concept relation based Thirukkural search in Tamil as well as in English using CoReX Techniques. Kuralagam unlike traditional keyword based Thirukkural searches retrieves Thirukkural that are conceptually relevant to the Query. When compared to traditional search techniques, our conceptual search methodology has higher precision. In future enhancement can be made to increase the precision and recall score of the conceptual Thirukkural search.

## Reference

1. Subalalitha, T V Geetha, Ranjani Parthasarathy and Madhan Karky Vairamuthu. CoReX: A Concept Based Semantic Indexing Technique. In SWM-08. 2008. India.
2. Foundation, U., the Universal Networking Language (UNL) Specifications Version 3 3ed. December 2004: UNL Computer Society, 2004. 8(5).Center UNDL Foundation
3. Anandan, R. Parthasarathi, and Geetha, Morphological Analyser for Tamil. ICON 2002, 2002.
4. T.Dhanabalan, K.Saravanan, and T.V.Geetha. 2002. Tamil to UNL Enconverter, ICUKL, Goa, India.
5. Andrew, T. and S. Falk. User performance versus precision measures for simple search tasks. In 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval 2006. Seattle, Washington, USA.



# Tamil Literature Output in National Bibliography of Indian Languages: A bibliometric analysis

**P. Clara Jeyaseeli**

*Ph.D. Research Scholar*

*Dept. of Library and Information Science,*

*Madurai Kamaraj University, Madurai.*

*e-mail: loyolaclara@gmail.com*

## Abstract

Tamil literature is one of the most classical and ancient South Indian Languages. The present study analyses the growth of Tamil literature based on the NBIL (*National Bibliography of Indian Languages*) database. Bibliometric studies such as literature growth study, authorship pattern, language distribution and document type and subject dispersion are reported. Literature growth study emphasis tremendous growth during 1946 to 1953. The authorship pattern analysis results that Kotaiyammai, Vai, Mu had contributed the most and is 4.36% among the 1147 publications. Tamil is the most predominating language reported from language dispersion study and moreover, it is found that 77.50% Tamil documents are microfilmed. Documents on philosophy and religion are given first preference followed by history, biography and travel.

**Keywords:** Tamil literature; Ancient Tamil Literature – growth analysis; Bibliometric analysis, NBIL-bibliometric analysis, Digital South Asia Library – bibliometric analysis

## Introduction

Tamil is one of the most classical and ancient South Indian Languages. In ancient times, the assembly or academy of most learned men of Tamil land was called “Sangam” and the literature produced from these assemblies is known as the “Sangam literature”<sup>1</sup>. *The National Bibliography of Indian Literature (NBIL)* is a selective bibliography with a compilation of works "of literary merit, and important and significant books on Philosophy, Religion, History and the other aspects of the Humanities". The Bibliography covers the period from 1901-1953<sup>2</sup>.

## Objectives

Egghe defined bibliometrics as the quantitative study of any literature as they are reflected in bibliographies. Thus bibliometrics may be generalized as a study of the quantitative analysis of the characteristics, behavior and productivity of all aspects of written communications. The objective of this study is to apply bibliometric techniques on the Tamil literature available from Digital South Asia Library's NBIL, since DSA's NBIL is a freely available bibliography in the internet. This bibliography is widely used by many Tamil scholars and researchers and therefore a study on this database is a valuable one which can exploit the characteristics of ancient digital Tamil literature provided by NBIL.

---

<sup>1</sup> [http://www.culturopedia.com/Literature/tamil\\_literature.html](http://www.culturopedia.com/Literature/tamil_literature.html)

<sup>2</sup> <http://dsal.uchicago.edu/cgi-bin/nbil.py>

## Significance of the study

The literature from Digital South Asia Library's NBIL is used as the bibliographic source database for this study since it is freely available and covers nearly 56,000 titles with imprints prior to 1954 in 22 Indian languages. This is the apt bibliographic database for analysis of ancient Tamil literature from 1901 to 1953.

## Research Methodology

The NBIL is searched for "tamil" (not case sensitive) in the subject search strategy and retrieved 1218 hits. The retrieved data is processed using MS-Word and MS-Excel.

## Literature Growth Study

One of the important aspects in bibliometric study is the prediction of the pattern of growth of literature<sup>3</sup>. Figure 1 depicts actual growth of Tamil literature.

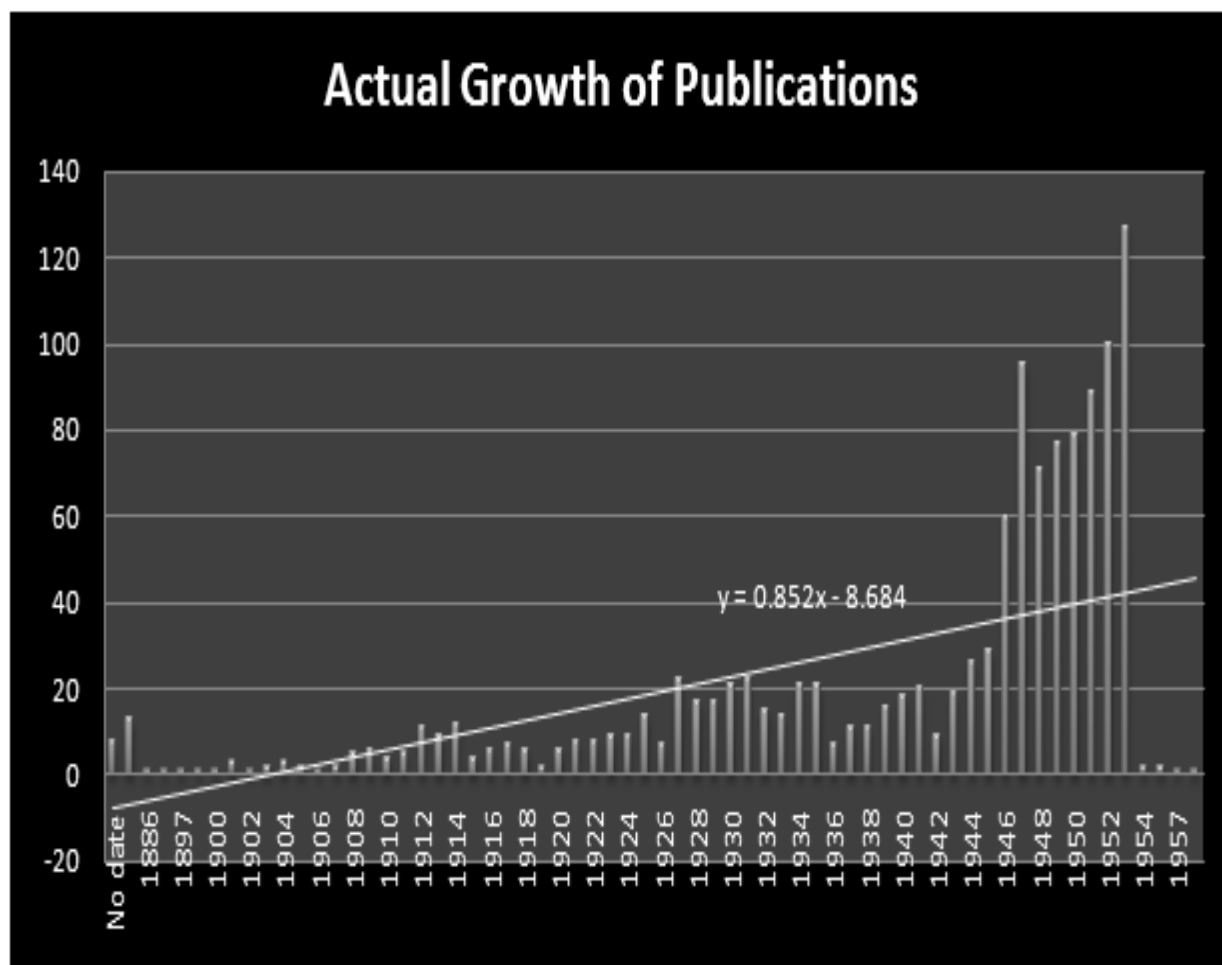


Figure 1. Actual Growth of Tamil Literature (1886 to 1958)

<sup>3</sup> Jeyaseeli, P. Clara. Growth Pattern Analysis no Ascidians Literature: A Scientometric Study (1998 to 2008). *Journal of Library, Information and communication Technology*, 2(1-4):51-59 (2010).

| Table 1. Growth of Tamil Literature (Top 8 years) |        |                     |            |
|---|--------|---------------------|------------|
| Rank  | Year   | No. of publications | Percentage |
| 1   | 1953   | 127                 | 10.43      |
| 2   | 1952   | 100                 | 8.21       |
| 3   | 1947   | 95                  | 7.80       |
| 4   | 1951   | 89                  | 7.31       |
| 5   | 1950   | 79                  | 6.49       |
| 6   | 1949   | 77                  | 6.32       |
| 7   | 1948   | 71                  | 5.83       |
| 8   | 1946   | 60                  | 4.93       |
| 9 to 64   | Others | 520                 | 42.69      |
| Total   |        | 1218                | 100        |

From Table 1, it is clear that 698 publications were authored during the top 8 years and the remaining 520 publications were authored from 1886 to 1958. The top 8 years productivity was 57.31 percentages when compared with the remaining 42.69 percentage. The years 1946 to 1953 were the most productive years and among them 1953 stands first with 127 publications. The linear trend calculates to  $0.852x - 8.684$ .

### Authorship Pattern

The kind of authors, nature and degree of collaboration among the authors are dealt in authorship pattern studies<sup>4</sup>. Table 2 depicts the list of authors who had produced more than 10 publications.

| Table 2. Tamil Literature Output - Authorship Pattern (Productivity $\geq 10$ ) |  |                     |            |
|---|--|---------------------|------------|
| Rank  | Author Name  | No. of publications | Percentage |
| 1   | Kotaiyammai, Vai. Mu., 1901-1960.                        | 50                  | 4.36       |
| 2   | Sambanda Mudaliar, Pammal, 1873-1964.                    | 35                  | 3.05       |
| 3   | Kaliyanasundaranar, Thiruvarur Viruddhachala, 1883-1953. | 31                  | 2.70       |
| 4   | Jagannathan, Krishnarayapuram Vasudeva, 1906-            | 28                  | 2.44       |
| 5   | Venkatanatha, 1268-1369, Alias Vedantadesika.            | 26                  | 2.27       |
| 6   | Varadarajan, M., 1912-1974.                              | 19                  | 1.66       |
| 7   | Maraimalaiyathikal, 1876-1950.                           | 15                  | 1.31       |
| 8   | Appadurai, K.  | 14                  | 1.22       |
| 9   | Caminata Carma, Ve., 1895-1978.                          | 13                  | 1.13       |
| 10  | Kantaiya Pillai, Na. Ci.                                 | 13                  | 1.13       |
| 11  | Cuppiramaniya Pillai, Ka., 1888-1945.                    | 12                  | 1.05       |
| 12  | Iracamanikkanar, Ma., 1907-1967.                         | 11                  | 0.96       |
| 13  | Suddhananda Bharati, 1897-                               | 10                  | 0.87       |
| 14  | Others (07-01)   | 870                 | 75.85      |
| Total   |  | 1147                | 100.00     |

From Table 2, it is inferred that 13 authors had contributed 10 to 50 publications. Out of 1218 records, 71 records don't have statement of responsibility. Kotaiyammai, Vai, Mu had contributed the most

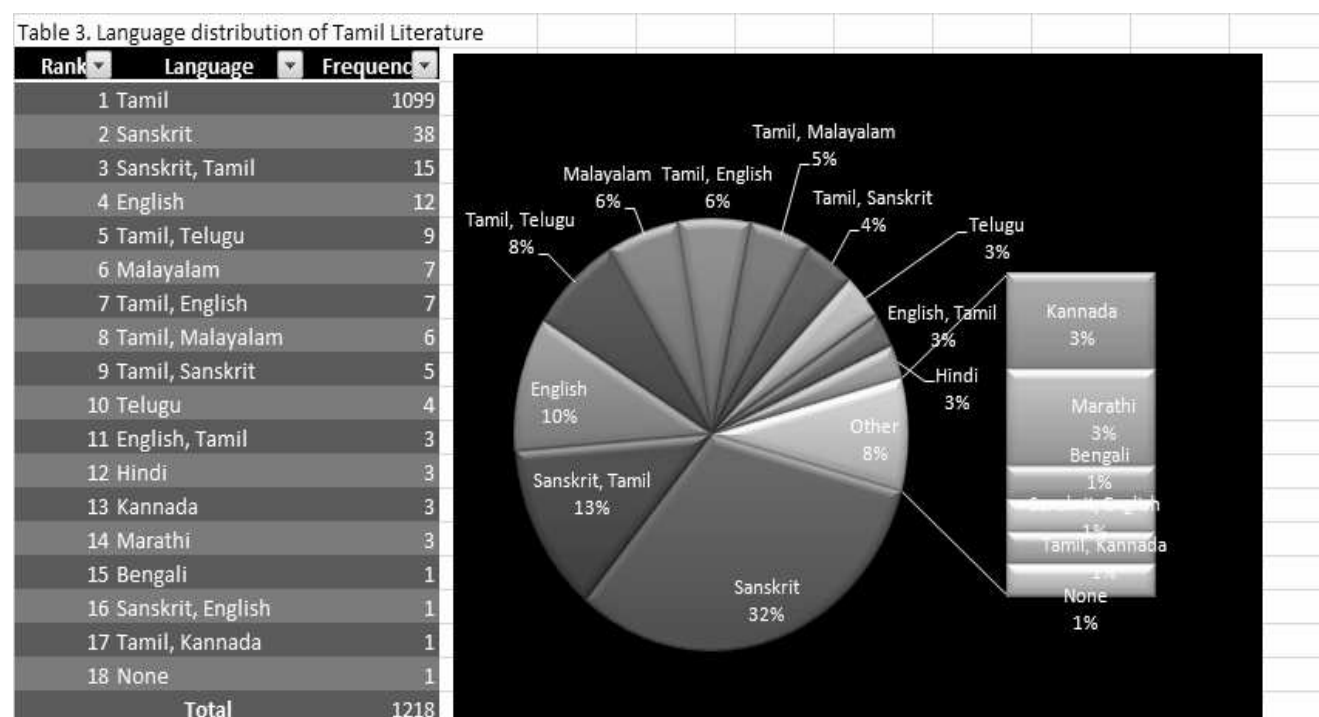
<sup>4</sup> Mahapatra, Gayatri, *Bibliometric Studies: in the internet era*, 2<sup>nd</sup> ed. (New Delhi: Indiana Publishing House, 2009), 78.

and is 4.36% among the 1147 publications. The difference between the second ranking author and up to thirteenth ranking author is not quite high. The 277 publications were authored by 13 authors calculating to 24.15%. The others had contributed one to seven publications and reached 75.85 % which is high.

### Language dispersion

The language of the document is one of the most important factors in bibliometric studies, since the references cited by the authors depend upon the language of the documents. If the authors don't know a language, then the citations for these documents are not made. Of course there are some percentage of publications published in regional languages may contain ideas on the subject and referred by the authors. **Table 3** explores the usage of languages in the ancient Tamil literature.

Tamil ranks first to a maximum of 90.23% and the rest of the languages occupy 9.77%. The Pie diagram shows rest 119 publications' language distribution. Sanskrit follows the second rank followed by bilingual language Sanskrit and Tamil as third. English language ranks fourth position.



### Document Type

Microforms may be of any form namely films or paper containing micro reproductions (reduced to about 25 times of size) of documents for storage, retrieval, transmission, and printing. In Digital South Asia Library, the NBIL microfilms the Indian publications under the Microfilming of Indian Publications Project (MIPP)<sup>5</sup>. The Government of India has approved a project, originally proposed by the National Library in Calcutta for the Preservation of Early Twentieth-Century South Asian

<sup>5</sup> <http://dsal.uchicago.edu/bibliographic/nbil/aboutmipp.html>

Books. Microfilming of Indian Publications Project (MIPP) is preserving and making accessible all 55,992 books listed in The National Bibliography of Indian Literature: 1901-1953 (NBIL) together with the pre-1954 titles in the NBIL supplement. In this analysis, out of 1218 documents, 944 documents are microfilmed which accounts to 77.50%. This is an appreciable mode of digital preservation for reference.

### Subject Dispersion Study

Subject dispersion study is one of the useful bibliometric study to know about the concentration of subject areas of documents. It is also useful for the funding agencies to disburse the grant based on the strong subject areas and also to enhance research in the needy and weaker areas. Table 4 tabulates the subject dispersion of Tamil literature from NBIL. The retrieved 1218 documents are categorized under 615 subject headings.

| Rank | Subject  | Frequency   | %             |
|------|--|-------------|---------------|
| 1    | Philosophy and religion.   | 69          | 5.61          |
| 2    | History, biography and travel.   | 50          | 4.07          |
| 3    | Literature - General works, histories of literature, literary criticism, general anthologies, etc. | 37          | 3.01          |
| 4    | Literature - Poetry.   | 31          | 2.52          |
| 5    | Tamil literature History and criticism.  | 31          | 2.52          |
| 6    | Linguistics.   | 21          | 1.71          |
| 7    | Arts.  | 20          | 1.63          |
| 8    | Literature - Fiction.  | 17          | 1.38          |
| 9    | Tamil poetry To 1500 History and criticism.  | 16          | 1.30          |
| 10   | Statesmen India Biography.   | 15          | 1.22          |
| 11   | Gandhi, Mahatma, 1869-1948.  | 13          | 1.06          |
| 12   | India Politics and government 1919-1947.   | 12          | 0.98          |
| 13   | Social Sciences.   | 12          | 0.98          |
| 14   | Hindu hymns, Tamil.  | 11          | 0.90          |
| 15   | Tiruvalluvar. Tirukkural.  | 11          | 0.90          |
| 16   | Others (<11)   | 863         | 70.22         |
|      | <b>Total</b>   | <b>1229</b> | <b>100.00</b> |

Since it is too lengthy to display all the 615 subject areas, the top 15 ranking subject areas are displayed in Table 4. Here the number of subject areas is greater than the number of documents analyzed. This is because; one document may belong to more than one subject area. Therefore, while consolidating the subject areas, it is found that 5.61% of documents fall under philosophy and religion subject keyword followed by History, biography and travel (4.07%).

### Conclusion

The growth of the Tamil literature analyzed highlights that productivity reached its peak during 1946 to 1953. There exist ups and down during the growth. The authorship pattern study shows that Kotaiyammai, Vai, Mu had contributed the most and is 4.36% among the 1147 publications. Collaborative authorship pattern did not exist and confirms only single author contribution. But there

exist translators and editors, but authorship pattern determines solo authorship dominance during this period of study.

The language dispersion analysis resulted that Tamil is the most predominant language and nearly 77.50% of documents are microfilmed. Philosophy and religion subject areas were given first preference followed by History, biography and travel.

Since Tamil is one of the most ancient and classical literature, NBIL concentrates from 1901 to 1954, the national policy may be framed to digitize almost all the Tamil literature documents which in turn results in reducing unemployment and provides a wealthy literature output for the future generation. Uniformity in transliteration and subject term specification may be given due priority to increase precision in NBIL.

## Reference

- [http://www.culturopedia.com/Literature/tamil\\_literature.html](http://www.culturopedia.com/Literature/tamil_literature.html)
- <http://dsal.uchicago.edu/cgi-bin/nbil.py>
- Jeyaseeli, P. Clara. Growth Pattern Analysis no Ascidians Literature: A Scientometric Study (1998 to 2008). *Journal of Library, Information and communicationTechnology*, 2(1-4):51-59 (2010).
- Mahapatra, Gayatri, *Bibliometric Studies: in the internet era*, 2<sup>nd</sup> ed. (New Delhi: Indiana Publishing House, 2009), 78.
- <http://dsal.uchicago.edu/bibliographic/nbil/aboutmipp.html>

# கணினி வழி தமிழ்ச் சங்க இலக்கிய ஆய்வு

வெ. பாலசரஸ்வதி

தமிழ்த்துறை முனைவர் பட்ட ஆய்வாளர்,

அவினாசிலிங்கம் நிகர் நிலை பல்கலைக்கழகம், கோவை, தமிழ்நாடு, இந்தியா.

Email: balasaraswathyramachandran63@gmail.com

சங்க இலக்கியங்கள் கருத்துக்காருவனங்கள். நம் முன்னோர் நமக்கு சேர்த்துக் கொடுத்திருக்கும் செல்வம். இதை நாம் அறிந்து நமக்குப் பின்னால் வரும் பரப்பரைக்கு கொடுத்துவிட்டுச் செல்லவேண்டும். இது பொருட் செல்வம் அல்ல துய்த்துவிட்டு அனுபவித்து முடித்துவிட, அனுபவித்துக் கொண்டே இருக்கும் அறிவுச் செல்வங்கள், இவை அனைத்தையும் நாம் படித்துமுடிப்பது என்பது "கல்வி கரையில் கர்பவர் நாள் சில" என்பது போல வாழ்நாளில் படித்து முடிக்க முடியாத ஒன்று. எனவே உலக மக்கள் அனைவரும் சேர்ந்து அறிந்து கொள்வோமானால் முழுமையும் அறியப்படும்.

பழையன கழிதலும் புதியன புகுதலும்

வழுவல கால வகையினான்'

என்பது நம் தமிழ் இலக்கணம் கூறுவது காலத்திற்கு ஏற்ப நம் தமிழ் இலக்கியங்களை கொண்டு செல்வதாகும் தமிழ் இலக்கிய வரலாறு முழுவதுமாக சங்க இலக்கியம் முதலாக தற்கால இலக்கியம் வரை கணினிமயமாக்கப் படவேண்டும்.

சங்க இலக்கிய தரவுகளை அலசி ஆய்தல்; தமிழ் இலக்கியவரலாறு எட்டுத்தொகை பத்துப்பாட்டு பதினெண்கிழக்கணக்கு நூல்கள் கப்பியங்கள் அனைத்தும் மென் பொருக்கப்பட்டு கணினியில் கிடைத்திட பாடல்கள் அப்படியே தமிழில் அமைந்திருக்கவேண்டும். இதற்கான ஆங்கில ஒலிப்பெயர்ப்பு கொடுக்கப்படலாம் சான்றாக தொல்கப்பியம் பொருலதிகாரத்தில் இடம் பெறும்.

"முதலெனப்படுவது நிலம் பொழுது இரண்டின்

இயல்பென மொழிப இயல்புணர்ந்தோரே"

muthal enapaduvathu nilam pozhthu irandin

iyalbena mozhipa iyalpunarthaore

என்று ஒலிபெயர்ப்பில் தரும் பொழுது அனைவரும் உச்சரிப்பை அறிந்துகொள்ள முடியும். தமிழர் நிலத்தையும் காலத்தையும் முதற்பொருளாக மதித்துப் போற்றினர் (According to tholkappiam the first thing is land and time) என்று வழங்கலாம். பொருள் அடிப்படையில் இது போதுமானதாகும் இதில் இயல்பென மொழிப இயல்புணர்ந்தோரே' என்பதுதான் தொல்காப்பியர் கொடுக்கிற அடிக்குறிப்பு அதாவது- reference தொல்காப்பியத்திற்கு முன்பே தமிழ் நூல்கள் தமிழ் புலவர்கள் தமிழ் அறிஞர் இருந்திருக்கிறார்கள் என்பதை அனைவரும் அறியமுடியும். தமிழ் இலக்கியங்களின் காலம் மேலும் பின்னோக்கி இருப்பதை உலகிற்கு உணர்த்தமுடியும். தொல்காப்பியம் நமக்கு கிடைத்த முதல் இலக்கண நூல் கால அடிப்படையில் முதலில் நிற்பது. "தொல்காப்பியர் கிறித்து பிறப்பதற்கு ஐநூறு ஆண்டுகளுக்கு முன்னர் வாழ்ந்தவர் என்பது பொருந்துவதாக தமிழ் இலக்கிய வரலாற்று அறிஞர்கள் கருதுகின்றனர். தொல்காப்பியர் குறிப்பிடும், என்மனார் புலவர், மொழிப, யாப்பறிபுலவர், நூல் நவில் புலவர் என்று கூறும் அடிக்குறிப்புச்சொற்கள் வழி அவருக்கு முன்னால் இலக்கிய இலக்கணங்கள் இருந்திருக்க வேண்டுமென முடிவு செய்ய முடிகிறது. இவ்வாறு காலத்தைப் பற்றி அறியப்படும் பொழுது சங்க இலக்கிய ஆய்வில் அதிக மாணவர்கள் ஆர்வம் கட்டுவர். படக்காட்சிகள் ஒவ்வொரு பாடல்களுக்கும் முடிந்தவரை வழங்குதல் சிறப்பானதாகும்.

சான்றாக நிலப்பாகுபாடு

"மலைச் சார்பு குறிஞ்சி நிலம் கடற்சார்பு நெய்தனிலம்  
காடுசார் நிலமுல்லை நாடுசார் நிலமருதம்  
நீரின்றி வேனிற் நெறு நிலம் பாலை" (திவாகரம்)

என்பதை படங்களாக கொடுக்கலாம். (மலை, கடல், காடு, வயல், வறட்சி-காட்சிகள்)

தமிழ் இலக்கியங்களில் உள்ள அறிவியல் மற்றும் பிற துறை கருத்துக்கள் இடம் வகையில் பிரித்து செய்திகள் வழங்கலாம். அவ்வாறு வழங்கும் பொழுது சங்க

இலக்கியத்தில் காணப்படும் அறிவியல் என்பதைவிட , பழந்தமிழ் தாவரவியல் , பழந்தமிழ் விலங்கியல் ; பழந்தமிழ் இயற்பியல், உளவியல் சுற்றுப்புறவியல், புவியியல் எனத் தனித்தனியான நுட்பமான முறையில் பிரிக்கப்பட்டு அந்த அந்த செய்திகளைக் கருவாகக் கொண்ட பாடல்கள் அந்த தலைப்பில் இடம்பெறச் செய்திடல் வேண்டும்.

அறிவியல் சார்ந்த தரவுகளில் சங்க இலக்கிய பாடல்களை சான்றுகளாக தரலாம். சான்றாக தாவரவியலில் மிக மென்மையான பூக்கள்பற்றிய தரவுகளில் மோப்பக்குழையும் அணிச்சம் இன்ங்கக் இவ்வாறு கொடுக்கும் பொழுது அறிவியல் சார்ந்த ஆய்வு மாணவர்களுக்கு சங்க இலக்கியத்தை அறிந்து கொள்ள ஆர்வம் ஏற்படும்.

'கையும் காலும் தூக்கத் தூக்கும்  
'ஆடிப்பாவை போல'(குறுந்தொகை)

இக் குறுந்தொகைப் பாடல் நமது பிம்பத்தை காட்டக்கூடிய கண்ணாடி இரண்டாயிரம் ஆண்டுகளுக்கு முன்பு தமிழ் இலக்கியத்தில் இருந்த செய்தி இயற்பியல் துறையில் இருந்திட வேண்டும். உயிரியல் செய்திகள் மிக அதிகமாக சங்க இலக்கியத்தில் காணப்படுகின்றன. இவ்வாறு நாம் அனைத்துத் துறைகளிலும் சங்க இலக்கிய பாடல்களை இணைத்து கொடுத்தல் சிறப்பாக இருக்கும்.

தேடு பொறிகள் அமைத்தல்;

சங்க இலக்கியத்தில் தேடு பொறிகள் அமைக்கின்ற பொழுது இலக்கியவரலாறு அடிப்படையில் கால வரிசையில் நூல்கள் பாடல்கள் பாடல் ஆசிரியர்கள் போன்றவற்றை வழங்குவதோடு இன்னும் நுட்பமாக நூற்பொருளை பிரித்து வழங்குதல் நன்று; சன்றாக புறநானூறு நூல் பற்றிய செய்திகளை அறியவேண்டுமானால் , நானூறு பாடல்களைப் படித்து செய்திகள் சேக்ரிப்பது விரைவான தவல் சேகரிப்புக்கு தடையாகும், காலதாமதமாகும் செய்திகள் அடிப்படையில் தலைப்புகள் பிரிக்கப்பட்டு தரவுகள் வழங்குதல் சிறப்பாக இருக்கும். சான்றாக தமிழ் இலக்கியம் --எட்டுத்தொகை- புறநானூறு

வீரம் பற்றிய பாடல்கள்

ஆட்சி

கொடை

கல்வி

ஒழுக்கம்

என்று தனித்தனியாக பிரிக்கப்பட்டு பாடல்கள் பொருள் விளக்கத்துடன் வழங்கப்படும் பொழுது ஆய்வு எளிமையானதாகவும் தூண்டுவதாகவும் அமையும். இன்றைய நவீன அறிவியலான தகவல் தொடர்பு கூறுகள் மேலாண்மை என அனைத்தும் வழங்கப்படவேண்டும்.



## மொழி நடை

மொழி நடை நமது சங்க இலக்கியம் செய்யுள் வடிவில் அமைந்திருக்கிறது. உரைநடைகள் இடை இடையே இருந்தாலும் (சிலப்பாதிகாரம்) அவையும் எதுகை மோனை பெற்று சிறந்த நடையில் இருப்பதைக் காண்கிறோம். உரையாசிரியர்கள் மொழிநடை சில இடங்களில் கடினமானதாக இருபதாம் நூற்றாண்டு உரையாசிரியர்களின் உரையை பயன்படுத்தலாம். சிறிய சொற்றொடர்களை பயன்படுத்துவது சிறந்ததாகும்.

இலக்கியங்களில் கால நிர்ணயம்;

நம் தொன்மை நூல்களின் காலத்தை பல குறிப்புகளைக் கொண்டு ஓரளவுக்கு முடிவு செய்ய இயலும் "பிற நாட்டு அறிஞர்கள் குறிப்புக்கள், அவர்களைப் பற்றிய செய்திகள், சில வரலாற்று உண்மைகள் சமுதாய பழக்கவழக்கங்களின் குறிப்புகள் சொல்லாட்சிகள், போன்றவை கால ஆரய்ச்சிக்கு துணை நிற்பனவாகும்." (சங்கத் தமிழ் டாக்டர் ச. அகத்தியலிங்கம்) மேலும் "ஒரு நூலில் காணப்படும் கருத்துக்களை கூறுவதுடன் அதன் ஆசிரியரையோ அல்லது நூலையோ குறிப்பிட்டு கூறும் போது நிச்சயமாக எடுத்தாளப்பட்ட நூல் காலத்தால் முந்தியது என்றும் எடுத்தாள்கின்ற நூல் பிந்தியது எனவும் முடிவு செய்யலாம். (எ.கா.) திருக்குறளில் காணப்படும் கருத்தினை சிலப்பதிகாரம் 'தெய்வம் தொழாஅள் கொழுநன் தொழ்வானை' எனவும் 'தெய்வம் தொழாள் கொழுநன் தொழுதெழுவாள்

பெய்யெனப் பெய்யும் பெரு மழை என்ற அப்

பொய்யில் புலவன் பொருளுரை தேறாய்'

என்று கூறுவதில் இருந்து திருக்குறள், சிலம்பு மணிமேகலை ஆகிய காப்பியங்களுக்கு முந்தியது என அறிய முடிகிறது.

## முடிவுரை

சங்க இலக்கியம் முழுவதும் பொருள் அடிப்படையில் எளிய மொழி நடையில் கொடுக்கப்படுகின்ற நிலையிலும் நுட்பமான தேடு பொறிகளிலும் அமைத்திடல் வேண்டும்.

# வெண்பா நிரல்

வெண்பாவிடிகா஢ பொது இலக்கணங்களைச் சரிபார்க்கும் நிரல்

## மு. சித்தநாதபூபதி

ஂவர்செண்டாய் எஞ்சினியரிங்

### சுருக்கம்

தமிழ்ச் செய்யுட்களில் வெண்பாவுக்கு சிறப்பான இடம் உண்டு. அசை, சீர், தளை என யாவற்றிலும் கட்டுக்கோப்பான யாப்பைப் பெற்றிருக்கும் வெண்பா, பா வகைகளில் கடினமானதாகக் கருதப் படுகிறது. 'வெண்பா நிரல்' என் வழங்கப் படும் இம்மென்பொருளால், நாம் இயற்றும் செய்யுட்களில் வெண்பாவின் இலக்கணங்கள் பயின்று வருகின்றனவா என்று சரிபார்த்துக் கொள்ள முடியும். அதைத் தவிர எதுகை, மோனை, ஒற்றளபெடை உள்ளிட்ட இலக்கணக் கூறுகள் அமைந்திருக்கின்றனவா என்றும் சரிபார்த்துப் பிழைகளைத் திருத்திக் கொள்ளலாம்.

### முன்னுரை

கணிணியைப் பயன்படுத்தி தமிழின் தனிச்சிறப்பாகக் கருதப்படும் யாப்பிலக்கணத்தையும் சரிபார்க்க இயலும் என்பது நம்மொழியின் என்றும் குன்றா இளமைக்கு ஒரு எடுத்துக்காட்டு , எழுத்து , அசை, சீர், அடி, தொடை என எல்லாக் கூறுகளிலும் திட்டவட்டமான கட்டமைப்பைப் பெற்றுள்ள பாவகைகளில் ஒன்றாகிய வெண்பாவைச் சரிபார்க்க இந்நிரல் உதவும் வண்ணம் வடிவமைக்கப் பட்டுள்ளது.

### எழுத்து

தமிழ்மொழியின் 247 எழுத்துக்களும் பயன்படுத்தப் பட்டுள்ளன . வடமொழி எழுத்துக்களான ஹ,ஸ,ஷ,ஐ,சஷ,ஸ்ரீ ஆகியவை தவிர்க்கப் பட்டுள்ளன. இவற்றிற்கு இணையான ஒலிப்புக் கொண்ட தமிழ் எழுத்துக்களைப் உள்ளிடவும் (அ,ச,ச,ச,ட்ச,சிறி). மெய்யெழுத்துக்கள் மொழி முதலாக வாரா. அவ்வெழுத்துக்களை மொழிமுதலாக உள்ளிட்டால் தவறு சுட்டிக் காட்டப்படும். எனினும் ட , ர, ல, ழ, ற , ன இவற்றை மொழி முதலாக உள்ளிட வகை செய்யப் பட்டுள்ளது. ஒரே சொல்லைச் சீர் பிரித்தாலன்றி அவற்றை மொழிமுதலாகப் பயன் படுத்த வேண்டாம்.

### அசையும் சீரும்

இந்நிரலில் அமைக்கப் பட்டுள்ள கட்டங்களில் ஒவ்வொரு சீரையும் அதற்கான கட்டங்களில் உள்ளிட வேண்டும். முதலாவதாக இந்நிரல் அச்சீர்களை தனி எழுத்துக்களாகப் பிரித்து , குறில் , நெடில் , ஒற்று என வகைப்படுத்திக் கொள்ளும். கிரந்த எழுத்துக்கள் அல்லாத தமிழ் ஒருங்குறி எழுத்துருக்களைத் தவிர்த்துப் பிற எழுத்துக்களை உள்ளிட்டால் ஏற்றுக்கொள்ளப் படா.

எ.கா: நீங்கள் கீழ்க்கண்டவாறு உள்ளிட்டால்

|     |      |               |     |
|-----|------|---------------|-----|
| அகர | முதல | எழுத்தெல்லாம் | ஆதி |
|-----|------|---------------|-----|

|       |         |      |
|-------|---------|------|
| பகவன் | முதற்றே | உலகு |
|-------|---------|------|

கீழ்க்கண்டவாறு சரிபார்க்கப் படும்.....

| சீர்கள் | எழுத்<br>து<br>வகை | முதல<br>சை | 2-ஆம்<br>அசை | 3-ஆம்<br>அசை | சீர்   | தளை            | தளை                 |     |
|---------|--------------------|------------|--------------|--------------|--------|----------------|---------------------|-----|
| அகர     | குக்கு             | நிரை       | நேர்         |              | புளிமா | மாமுன்<br>நிரை | இயற்சீர்<br>வெண்டளை | சரி |
| முதல    | குக்கு             | நிரை       | நேர்         |              | புளிமா | மாமுன்<br>நிரை | இயற்சீர்<br>வெண்டளை | சரி |

**தளை:** வெண்பாவுக்கு உரிய இயற்சீர் வெண்டளை , வெண்சீர் வெண்டளைகள் தவிர ஏனைய தளைகள் வரின் சிவப்பு நிற எழுத்துக்களில் எச்சரிக்கை தோன்றும்.  
எடுத்துக்காட்டாக முதற்சீர் காய்ச்சீராக இருக்க வரும் சீரின் முதலசை நேரசையாக இல்லாவிடின் கீழ்க்கண்டவாறு எச்சரிக்கை தோன்றும்.

| எழுத்தெ<br>ல்லாம் | குக்குகு<br>நெஓ | நிரை | நேர் | நேர் | புளிமாங்<br>காய் | காய் முன்<br>நிரை | கலித்தளை                | தளை<br>தட்டுகிறது |
|-------------------|-----------------|------|------|------|------------------|-------------------|-------------------------|-------------------|
| பகவன்             | குக்குகு        | நிரை | நேர் |      | புளிமா           | மாமுன்<br>நிரை    | இயற்சீர்<br>வெண்ட<br>ளை | சரி               |

இவ்வாறாக ஒவ்வொரு சீருக்கும் இடையே பயின்று வரும் தளைகளையும், தளை தட்டும் இடங்களையும் கண்டு சரிசெய்ய வேண்டிய இடங்களைச் சரி செய்யலாம் . ஈற்றடியின் ஈற்றுச்சீர் (மூன்றாவது சீர்) , நாள் , மலர் , காசு , பிறப்பு என்பவற்றுள் அடங்கும் ஏதேனும் ஒரு ஓரசைச் சொல்லாகவோ ஈரசைச் சொல்லாகவோ இருத்தல் வேண்டும் . இந்நிரல் அவற்றையும் இனங்காணும்.

**எதுகை :** நேரிசை வெண்பாக்களில் ஒரு விகற்ப எதுகையோ , இரு விகற்ப எதுகையோ அடிகளில் அமையப் பெறுதல் அவசியம். அடி எதுகை , சீர் எதுகை போன்றவற்றையும் சிறப்பாகச் சரிபார்க்கலாம். எடுத்துக்காட்டாக அன்பும் அறனும் உடைத்தாயின் இவ்வாழ்க்கை பண்பும் பயனும் அது என்ற குறளில் அமைந்த இன எதுகையைக் கூட இந்நிரல் சரியாகக் கண்டுபிடிக்கும்.

**மோனை :** மோனை வெண்பாவில் கட்டாயமில்லை எனினும் இந்நிரலில் கீழ் மோனையைக் கண்டுபிடித்தல் இடம் பெறுகிறது. பாடலில் கீழ் பயின்று வரும் மோனையைக் கண்டுபிடிப்பதற்கான தர்க்கம் எதுகையை விட வெகு எளிதானதே. இவ்வாறே அளபடைகளையும் கண்டு பிடிக்கலாம்.

**தனிச்சொல் :** தனிச்சொல்லுக்கு முன் ‘-’ என்ற குறி காட்டப்பட்டுள்ளது. எனினும் தனிச்சொல்லுக்கு முந்தைய சீர் இலக்கணப்படியும் முற்றுப் பெற்றதா இல்லையா என்பதை பயனரே உறுதி செய்து கொள்ள வேண்டும். முந்தைய சீர் முற்றுப்பெற்று விட்டதா என்பதைச் சரிபார்த்தல் எதிர்காலத்திட்டத்தில் இணைக்கப் பட்டுள்ளது.

**நிரல் சுட்டிக்காட்டும் பிழைகள் :**

ஒருங்குறி அல்லாத தமிழ் எழுத்துருக்கள் வரின் :- # NA

ஒருங்குறியாயினும் கிரந்த எழுத்துக்கள் வரின் :- # NA

ஒற்று முதலெழுத்தாக வரின் :-

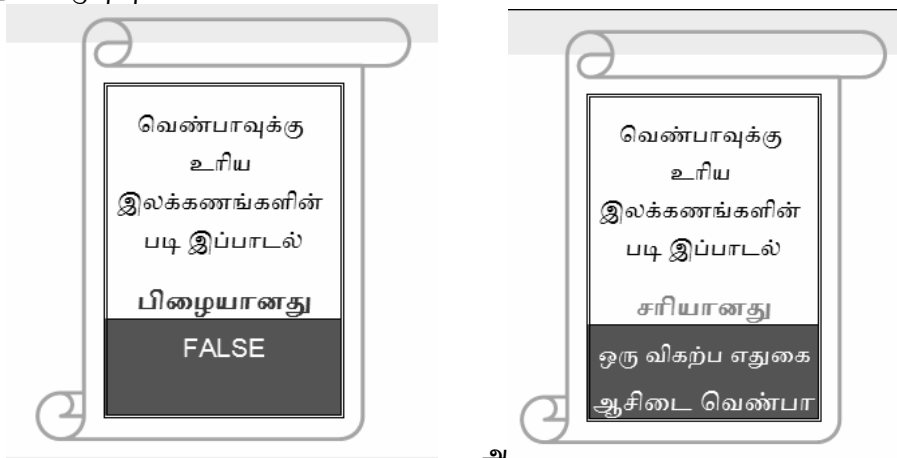
| சீர்க்<br>ள் | எழுத்து<br>வகை | முதல<br>சை | 2-ஆம்<br>அசை | 3-ஆம்<br>அசை | சீர்       | தளை            | தளை                     |      |
|--------------|----------------|------------|--------------|--------------|------------|----------------|-------------------------|------|
| க்ரி<br>யா   | ஒகுநெ          | okn        |              |              | #N/<br>A   | #N/A           | #N/A                    | #N/A |
| கிரி<br>யா   | குகுநெ         | நிரை       | நேர்         |              | புளி<br>மா | மாமுன்<br>நிரை | இயற்சீர்<br>வெண்ட<br>ளை | சரி  |

வெண்பாவிற்சூரிய தளைகள் வராவிடின் :- தளை தட்டுகிறது

எதுகைகள் வராவிடின் :-

| அடிச்<br>சீர்கள் |   | முதலெழு<br>த்து |    | 2-ஆம் எழுத்து                             | 2-ஆம்<br>எழுத்துக்கள்   | முதலெழு<br>த்து             |   |
|------------------|---|-----------------|----|---|-------------------------|-----------------------------|---|
| அரியபெ<br>ரிய    | ர | அ               | ரி | ர   | ரண                      | கு                          | அடி எதுகை<br>அமையப்<br>பெறவில்லை<br><br>அடி எதுகை<br>அமையவில்லை |
| கண்டு            | ண | க               | ண் | ண   | எதுகை<br>அமையவில்<br>லை | கு                          |   |
| கண்டு            | ண | க               | ண் | ண   | ணர                      |                             |   |
| தெரியின்         | ெ | தெ              | ரி | ர   | எதுகை<br>அமையவில்<br>லை | கு                          |   |
| கரிய             | ர | க               | ரி | ர   | ரண                      | கு                          | அடி எதுகை<br>அமையப்<br>பெறவில்லை                                |
| எண்டு            | ண | எ               | ண் | ண   | எதுகை<br>அமையவில்<br>லை | கு                          | அடி எதுகை<br>அமையவில்லை   |
|                  |   |                 |    | 2 மற்றும் 3-ஆம்<br>அடிகளுக்கு<br>இடையே :- | எதுகை                   | பாடல்<br>முழுவதி<br>லும் :- | FALSE   |

ஒட்டு மொத்தமாக முடிவு



அ

**எதிர்கால விரிவாக்கம்:** இந்நிரல் தொகுப்பில் குறள் வெண்பா , நேரிசைச் சிந்தியல் வெண்பா , இன்னிசைச் சிந்தியல் வெண்பா , நேரிசை அளவியல் வெண்பா , இன்னிசை அளவியல் வெண்பா ஆகியவற்றைச் சரிபார்க்கவும் , சீரெதுகை , அடியெதுகை , இன எதுகை ஆசிரியப்பாவின் வகைகள் ஆகியவற்றைச் சரிபார்க்கவும் வகைசெய்யப் பட்டுள்ளது. எதிர்கால மேம்படுத்தலில் பிற பாவகைகளைச் சரிபார்க்க ஒ நிரல்கள் உருவாக்கப் படும்.

**முடிவுரை :** தன்னேரிலாத தமிழின் அரிய செல்வங்களான யாப்பின் அனைத்துக் கூறுகளையும் எளிதில் நிரல்படுத்தும் நாள் வெகு தொலைவில் இல்லை.

### **இணைப்பு**

நேரிசை வெண்பாவிற்கான சரிபார்த்தல்

**துணைநூற் பட்டியல்/ உதவிய வலைத்தளங்கள்**

1. தொல்காப்பியம் கேசிகன் உரை
2. தமிழ் இலக்கணம் -நுஃமான் எம்.ஏ வாசகர் சங்கம் , கொழும்பு
3. யாப்பருங்கலக் காரிகை
4. [www.tamilvu.org](http://www.tamilvu.org)
5. [www.venbavadikkalamvanga.com](http://www.venbavadikkalamvanga.com)

**பயன்பட்ட மென்பொருட்கள்**

1. மைக்ரோசாஃப்ட் விண்டோஸ்- எக்செல்



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# தமிழ்மொழியும் உச்சரிப்பும் - கற்றல் கற்பித்தலில் கணினியின் பங்கு

டாக்டர் ஆ ரா சிவகுமாரன்

இணைப் பேராசிரியர், தலைவர் - தமிழ்மொழி பண்பாட்டுப் பிரிவு  
ஆசியான் மொழிகள் மற்றும் பண்பாட்டுத்துறை தேசியக் கல்விக்கழகம் :  
நன்யாங் தொழில்நுட்பப் பல்கலைக்கழகம், சிங்கப்பூர் 637616.

உலகெலாம் தமிழோசை பரவும் வகை செய்வோம் என்று கூறிச்சென்றனர் நம் கவிஞர்கள். தமிழ்மொழி இரட்டைவழக்குச்சூழல் (Diglossic situation) கொண்டது. பேச்சுத்தமிழ், எழுத்துத்தமிழ் என இரண்டு வழக்குகளும் தமிழர்களின் மொழிச் செயல்பாடுகளுக்குப் பயன்படுகின்றன.

‘தமிழ்மொழியினைத் தெளிவாக உச்சரிக்கும்போது அது மேலும் அழகு பெறுகிறது. உச்சரிப்பு சரியில்லை என்றால் மொழியின் பொருளும் மாறுகிறது. ‘ஒரு மொழியின் பொருள்தரு ஒலிகளை - ஒலியன்களை அத்தாய்மொழியாளர் நன்குணர்ந்திருப்பர். இஃது அன்னாரின் உளவறிவு (Psychological image). ஆனால் அவ்வொலியன்கள் இடம், சூழல்களுக்கு ஏற்ப மாறிவருவதனை அவர் யாரும் சாதாரண நிலையில் அறிந்திருப்பதில்லை. காரணம் மனிதமுனை ஒலிகளை, அவை மொழியில் பொருள் அல்லது இலக்கணக் கூறுகளை வெளிப்படுத்தும் பணிபினைக்கொண்டே தேர்வே செய்கிறது. அலகுகளாகக் கொள்கிறது. மற்ற ஒலிகளைப் பிறர் கூற்றாலோ ஒலியியல் அறிவாலோதான் அறிகின்றது’ என்று ஒலியியல் பேராசிரியர் க. முருகையன் கூறுவார்.

சிங்கப்பூர்வாழ் தமிழ் மாணவர்கள், இப்போது அன்றாட உரையாடல்களுக்கு ஆங்கிலத்தையும் எழுத்துத் தமிழையும் மிகுதியாகப் பயன்படுத்தும் வழக்கம் அதிகரித்துவருகிறது. சிங்கப்பூரில் சுமார் 58 விழுக்காட்டுக் குடும்பங்களில் தமிழ்மொழி வீட்டில் பேசப்படாததால் தமிழ்க்குழந்தைகள் தமிழை அறியாமல் பள்ளிக்கு வருகின்றனர். மேலும் தமிழ்மொழியின் சரியான உச்சரிப்பைக் கைவரப் பெறுவதில் சிரமத்தையும் எதிர் நோக்குகின்றனர். தமிழ்மொழியின் இலக்கணத்தை ஓரளவு புரிந்து கொண்டாலும்கூட அதனைச் சரிவர உச்சரிப்பதில் பிழை புரிகின்றனர். ஆசிரியர் எவ்வளவுதான் உச்சரித்துக்காட்டினாலும் அதனைக் கேட்டு மீண்டும் உச்சரிப்பதில் சிக்கலை எதிர்நோக்குகின்றனர்.

‘மாற்றொலிகள் வேறுபடுத்தி ஒலிக்கப்படாவிட்டாலும் பொருள்கொள்வதில் குழப்பம் விளைவதில்லை. ஆனால் அவ்வகை உச்சரிப்பு தமிழ்மொழியின் இயல்பு வழக்கினின்றும் வேறுபட்டு அயற்றன்மை உடையதுபோல் ஆகிவிடும். இயல்பான தமிழ்ப்பேச்சாக அமையாது’ என்று ஒலியியல் பேராசிரியர் க. முருகையன் கூறுவார்.

இப்பிரச்சினையை எவ்வாறு களைவது? இதில் உள்ள சிக்கல்களைத் தீர்ப்பதில் கணினியின் பயன்பாடு என்ன என்பது பற்றியே இக் கட்டுரை விவரிக்கிறது. (இக்கட்டுரையில் கூறப்பட்டிருக்கும் செய்திகளைக் கணினியின் துணையோடு படிக்கும்போதே இக்கட்டுரையில் கூறப்பட்டிருக்கும் செய்திகள் நன்கு விளங்கும்)

மொழிக் கல்வியில் / பயிற்சியில் பயன்படுத்தப்படுகிற கேட்டல் - பேசுதல்முறை ( audio-lingual method) மாணவர்களுக்குத் தமிழ் உச்சரிப்பை முறையாகக் கற்றுக்கொடுக்கப் பயன்படும். அதற்குக் கணிணித் தொழில்நுட்பம் மிகவும் உதவும் என்பதே இக்கட்டுரையின் கருதுகோளாகும்.

தமிழ் உச்சரிப்புப் பயிற்சியில் மிகவும் கவனம் செலுத்தவேண்டிய பிரச்சினைகளாகக் கீழ்க்கண்டவை அமைகின்றன.

1. ஒலியன்களிடையே (Phonemic) பிரச்சினை : ல,ள,ழ - ந,ன,ண - ர,ற ஆகியவற்றை உச்சரிப்பதில் காணப்படும் பிரச்சினை.
2. மாற்றொலிகளில் (Allophonic) பிரச்சினை : உயிர் ஒலியன்களில் முற்றுகரம், குற்றுகரம் உச்சரிப்பு, வல்லின ஒலியன்களின் மாற்றொலிகளின் உச்சரிப்பு ஆகியவற்றில் காணப்படும் பிரச்சினை.
3. மேற்கூற்றுஒலிகளில் (Suprasegmental / Prosodic feature) பிரச்சினை : மேலும் தொடர்களை வாசிக்கும்போது, ஏற்றம், இறக்கம் ஆகியவற்றைப் பொறுத்துப் பொருளே மாறுபடும். ஆங்கிலத்தில் ஒரு சொல்லில் அமைந்துள்ள அசைகளின் அழுத்தம் (Syllabic stress) பொருள் மாறுபாட்டைத் தரும் ("permit" - இதில் இரண்டு அசைகள் உள்ளன. முதல் அசைக்கு அழுத்தம் அளித்தால், அச்சொல் பெயராக அமையும். இரண்டாவது அசைக்கு அழுத்தம் அளித்தால், அச்சொல் வினையாக அமையும்) . தமிழில் இப்பிரச்சினை கிடையாது. ஆனால் தொடர்களின் ஏற்றம், இறக்கம் (Intonation) பொருள் வேறுபாட்டைத் தருகிறது.

ல, ள, ழ - ந,ன,ண - ர,ற ஒலியன்களில் மாற்றொலிகள் பிரச்சினைகள் இல்லை. இருப்பினும் அவற்றிற்கிடையே உள்ள சில ஒலி ஒற்றுமைப் பிரச்சினைகளால் மாணவர்களுக்குக் குழப்பம் ஏற்படுகிறது. வல்லின ஒலியன்களில் ஒவ்வொரு ஒலியனுக்கும் ஒன்றுக்கு மேற்பட்ட மாற்றொலிகள் உள்ளன. இது மாணவர்களுக்கு ஒரு பிரச்சினையாக அமைகிறது.

மேற்கூறிய பிரச்சினைகளை முறையாகத் தீர்ப்பதற்குத் தமிழ் ஒலியியல், ஒலியனியல் விதிகள் மிகவும் உதவும். ஆனால் அவ்விதிகளை மாணவர்களுக்கு நேரடியாகக் கற்றுக்கொடுப்பதால் உச்சரிப்புப் பிரச்சினையைத் தீர்த்துவிட முடியாது. அவ்விதிகளின் அடிப்படையில் முறையான உச்சரிப்புப் பயிற்சிகளை உருவாக்கி மாணவர்களுக்குப் பயிற்சி அளிக்கவேண்டும். இதற்கு இன்றைய கணினித் தொழில்நுட்பம் மிகவும் பயன்படும்.

தமிழ்மொழியின் ஒலியன்களின் மாற்றொலிகளைச் சரியாக உச்சரிக்கும்பொழுதே மொழியின் தூய்மை பாதுகாக்கப்படுகின்றது. பொதுவாக மாணவர்கள் செய்யும் பிழைகளைப் பொறுத்துக் கீழ்க்கண்டவாறு நாம் அவர்களுக்குப் பயிற்சி அளிக்கலாம் .

1. தமிழ் உயிர்ஒலிகளைப் பொறுத்தமட்டில், குறில் உகரத்தில் மட்டும் இரண்டுவகையான உச்சரிப்புகள் உள்ளதை நாம் அறிவோம். அவை முற்றியலுகரம், குற்றியலுகரம் ஆகும். அவற்றைச் சரியாக உச்சரிக்க வேண்டியது அவசியம். முற்றுகரத்திற்கு உதடு (இதழ்) குவியும். குற்றுகரத்திற்கு உதடு குவியாது. தனிக்குற்றெழுத்தை அடுத்து வராத ஒரு சொல்லின் இறுதியில் வல்லினத்தோடு இணைந்து வரும் உகரம் குற்றுகரமாகும். எடுத்துக்காட்டுகள் - நாடு, பத்து, அங்கு, பழகு, வயிறு, அஃது. பிற இடங்களில் வரும் எல்லா உகரமும் முற்றுகரமாகும். தனிக்குற்றெழுத்தை அடுத்துச் சொல்லின் இறுதியில் வல்லினத்தோடு இணைந்து வரும் உகரமும் முற்றுகரமே. கொசு, பசு என்று சொல்லுகின்றபொழுது உதடு குவிந்தே வருவதைக் காணலாம் .
2. தமிழிலுள்ள 18 மெய்யெழுத்துகளில் (ஒலியன்களில்) வல்லின மெய்களான ஆறுக்கும் மாற்றொலிகள் உள்ளன.



3. க /k/ என்னும் வல்லின ஒலியனுக்கு மூன்று மாற்றொலிகள் உள்ளன.

ஒலிப்பில்லா கடையண்ண வல்லின மாற்றொலி [k] - சொல் முதல் மற்றும் சொல் இடையில் இரட்டித்து வரும்போது வரும். எடுத்துக்காட்டு (காக்கை, அக்காள்)

ஒலிப்புள்ள கடையண்ண வல்லினமாற்றொலி [g]- சொல் இடையில் மெல்லின ஒலியன்களுக்கு அடுத்து வரும். எடுத்துக்காட்டு (தங்கம், பங்கு, இங்கே)

அடுத்து கடையண்ண உரசொலி [x] - சொல் இடையில் இரண்டு உயிர்களுக்கு இடையில் வரும் ககரம் வேறு விதமாக ஒலிக்கும். எடுத்துக்காட்டு (பகல், நகம்)

4. மற்றொரு வல்லின ஒலியான /c/ சகரத்திற்கும் மூன்று மாற்றொலிகள் உள்ளன. சொல் முதலிலும் இடையிலும் இரட்டித்தும் மெல்லின எழுத்துகளுக்கு அடுத்தும் வரும்பொழுது இந்தச் சகரத்தின் ஒலிப்பில் மாற்றம் இருப்பதை உணர/கேட்க முடியும்.

[C] சகரம், சொல் நடுவில் இரட்டித்து வரும்பொழுதும் சொல் நடுவில் வல்லினத்தை அடுத்து வரும்பொழுதும் இம்மாற்றொலியைக் கேட்க முடியும். எடுத்துக்காட்டு. பச்சை, மொச்சை, கட்சி, பட்சி.

[ j ] சகரத்தின் மற்றொரு மாற்றொலியைச் சொல் நடுவில் மெல்லினத்தை அடுத்து வருகின்றபொழுது கேட்க முடியும் எடுத்துக்காட்டு மஞ்சள், பஞ்சு, கொஞ்சு, கெஞ்சு.

[s] சகரத்தின் மூன்றாவது மாற்றொலியைச் சகரம் சொல் முதலிலும் சொல் நடுவில் இரண்டு உயிர்களுக்கு இடையிலும், ல் ஒலியனுக்கும் ஒரு உயிருக்கும் இடையிலும் வருகின்றபொழுதும் உணர முடியும். எடுத்துக்காட்டு சாப்பிடு. சட்டை, பசி, ஊசி, வல்சி.

5. வல்லின ஒலியன்களில் மற்றொரு ஒலியன் டகரம் / t / . இந்த வல்லின எழுத்திற்கும் மூன்று மாற்றொலிகள் உள்ளன.

[t] சொல் நடுவிலும், இரட்டித்து வரும்பொழுதும் மாற்றொலியாக ஒலிக்கும். எடுத்துக்காட்டு. பட்டு, தட்டு, வெட்கம், தட்பம் நுட்பம்.

[d] மற்றொரு மாற்றொலி டகரம், சொல் நடுவில் மெல்லினத்திற்குப்பின் வரும்பொழுது வரும். எடுத்துக்காட்டு தொண்டு, மண்டு, குண்டு, நண்டு.

[r] டகரத்தின் இன்னொரு மாற்றொலி இரண்டு உயிர்களுக்கு இடையில் டகரம் வரும்போது ஒலிக்கும். எடுத்துக்காட்டு. தவிடு, செவிடு, குருடு, முரடு.

6. தகரத்திற்கும் /t/ மூன்று மாற்றொலிகள் உண்டு.

[t] ஒரு சொல்லின் முதலிலும், சொல் நடுவில் இரட்டித்தும் வருகின்றபோது தகரத்தின் மாற்றொலியைக் கேட்க இயலும். எடுத்துக்காட்டு, தாய், தந்தை, தட்டு, தவிர, பத்து, அத்தை, சொத்து.

[d] தகரம், சொல்லின் நடுவில் மெல்லினத்தை அடுத்து வரும்பொழுது மாற்றொலியாக ஒலிக்கும். எடுத்துக்காட்டு பந்து, சந்தை, நொந்து.

[ð] தகரம் சொல் நடுவில் இரண்டு உயிர்களுக்கு இடையில் வருகின்ற பொழுதும் சொல் நடுவில் ய், ர், ழ் ஒலியன்களுக்கும் உயிருக்கும் இடையில் வருகின்ற பொழுதும் இவ்வொலியைக் கேட்க இயலும்.

7. 'ப' என்ற வல்லின ஒலியனுக்கும் /p/ மூன்று மாற்றொலிகள் உள்ளன.

[p] சொல் முதலில் வருகின்ற பகர ஒலி போலவே சொல் நடுவில் இரட்டித்து வரும்போதும், சொல் நடுவில் ற், ட் ஒலியன்களுக்கு முன் வருகின்ற பொழுதும் பிறக்கும் ஒலி இருக்கும். எடுத்துக்காட்டு பாடம், பாட்டு, பட்டினம், உப்பு, செப்பு, மப்பு, கற்பு, நட்பு, நுட்பம்.

[b] சொல் நடுவில் மெல்லினத்தை அடுத்து வருகின்ற பகரம் வேறொரு மாற்றொலியாக ஒலிக்கும். எடுத்துக்காட்டு அன்பு, என்பு அம்பு, வம்பு, கம்பு, தம்பு.

[β] பகரத்தின் மற்றொரு மாற்றொலி சொல்நடுவில் இரண்டு உயிர்களுக்கு இடையில் வருகின்றபொழுதும் சொல் நடுவில் ம், ர், ல், ழ் ஆகிய மெய்யொலியன்களுக்கும் உயிர் ஒலியனுக்கும் இடையில் வருகின்ற பொழுதும் ஒலிக்கும். எடுத்துக்காட்டு சபை, அவை, இயல்பு, சால்பு, சாய்ப்பு, தொடர்பு.

8. வல்லின ஒலியன்களில் இறுதியாக இருக்கின்ற றகரத்திற்கும் மூன்று மாற்றொலிகள் உள்ளன.

[ɽ] நடுவில் இரட்டித்து வரும்போதும் சொல் நடுவில் வல்லின ஒலியனுக்குமுன் வரும்போதும் றகரம் மாற்றொலியாக ஒலிக்கிறது. எடுத்துக்காட்டு பற்று, நேற்று, காற்று, போற்று, கற்பு.

[ɽ̌] நடுவில் மெல்லினத்திற்கு முன் வருகின்றபொழுதும் மாற்றொலியாக ஒலிக்கின்றது. குன்று, ஒன்று, நன்று சென்று இன்று, கன்று.

[ɽ̌̌] சொல் நடுவில் இரண்டு உயிர்களுக்கு இடையே வருகின்ற 'றகரம்' மாற்றொலியாக ஒலிப்பதையும் நாம் உணரலாம்.

மெய்யொலிகளில் ங், ஞ், ண், ந், ம், ன்- ம், ர், ல், வ், ழ், ள் ஆகிய ஒலியன்களுக்கு ஒன்றுக்கு மேற்பட்ட மாற்றொலிகள் கிடையா.

9. தமிழ்மொழிக்கே உரிய சிறப்பு முகரத்தை மாணவர்கள் பலர் தவறாக ஒலிப்பதைக் கேட்டிருக்கின்றோம். அவற்றை எவ்வாறு ஒலிக்க வேண்டும் என்பதையும் கணினி வழி அறிய இயலும். பெரும்பான்மையான மாணவர்கள் 'தமிழ்' என்பதைத் 'தமில்' என்றும் 'கழகம்' என்பதை 'கலகம்' என்றும் உச்சரிப்பதைக் கேட்டிருக்கின்றோம். இவற்றைச் சரியான முறையில் உச்சரித்துக் கேட்கக் கணினி உதவிபுரியும்.

10. தமிழ்மொழியில் சில தொடர்களை உச்சரிக்கும் விதத்தில் பொருள் வேறுபாடு அடையும் என்பதை மாணவர்கள் உணர்ந்து இருக்கவேண்டியது அவசியம்; அப்பொழுதே மாணவர்கள் எந்த இடத்தில் எந்தபொருளில் எப்படி உச்சரிக்க வேண்டும் என்பதை உணர்வர். குறிப்பாக 'அவன் வந்தான்' என்னும் தொடரை ஏற்ற இறக்கத்தோடு உச்சரிக்கின்ற விதத்தினால் பல வேறுபட்ட பொருள்கள் கிடைக்கும் என்பதை மாணவர்கள் உணர வேண்டும். இந்த உச்சரிப்பு முறையைக் கணினியின் உதவியால் அறிந்து கொள்ளலாம்.

இவ்வாறு சில ஒழுங்குகளுக்கு உட்பட்ட விதிகள் பல உள்ளன. அவற்றை நாம் எழுத்துவடிவத்தையும் ஒலி வடிவத்தையும் ஒருங்கே பெறச் செய்து கற்பிக்கலாம். பொதுவாகத் தனித் தனி ஒலியன்களாக அமைத்துக்காட்டுவதைவிடத் தொடர்களில் அமைத்துக் கற்பிக்கும்பொழுது மாணவர்களுக்கு அவ்வொலிகள் எளிதாக விளங்கும். பொதுவாக ஒலிகளைத் தனித்தனியாக அடையாளம் காட்டுவதைவிடச் சொல்லிலும் தொடரிலும் பாடலிலும் உரையாடலிலும் அமைத்துக் காட்டும்பொழுது இவ்வொலிகளை நன்கு அடையாளம் காண முடியும்.

இவ்வாறு அனைத்துக் கூறுகளையும் எடுத்துக்காட்டுகளோடு காட்டி அவற்றைக் கணினிவழிப் பல்லாடகத்தின் (Multi media) உதவியால் பலவிதங்களில் வெளிப்படுத்திக் காட்டுகின்றபொழுது

மாணவர்களுக்குப் எழுத்துத்தமிழையும் ஒலிப்பு முறைகளையும் ஒருங்கேபெற வாய்ப்புகள் கிடைக்கின்றன. (மேலும் ஏற்ற இறக்கம் காட்டி உச்சரிக்கின்றபொழுது ஏற்படும் பொருள் வேறுபாடுகளையும் அறிய முடிகிறது. அன்றியும் சரியான உச்சரிப்புகளை அறிந்துகொள்வர். மேலும் கணினியில் தயாரிக்கப்பட்ட இக்கூறுகளை மாணவர்கள் சொந்தமாகப் பலமுறை மீண்டும் மீண்டும் இயக்கிக் கேட்பதன்வழி அவர்கள் மனத்தில் அவை நன்கு பதியும். ஆர்வமும் பெருகும். மாணவர்கள் பொருள் உணர்ந்து படிக்கவும் கேட்கவும் வாய்ப்புகள் அதிகரிக்கும்; இந்த முயற்சியில் கணினி வல்லுநர்களும் மென்பொருள் தயாரிப்பாளர்களும் ஈடுபட்டு உதவுகின்றபொழுது பல்வேறு பரிமாணங்களில் கணினித் திரையில் ஆர்வமூட்டும் உச்சரிப்பு முறைகளைக் கேட்க இயலும். இவை தாய்த் தமிழகத்திலிருந்து வெளியேறி எழுத்துத்தமிழ், பேச்சுத்தமிழ் சூழல் அமையப்பெறாமல் வாழ்கின்ற வெளிநாட்டுத் தமிழ்க் குழந்தைகளுக்குப் சரியான உச்சரிப்பை அறியவும் கற்கவும் பெரிதும் உதவும்.

#### கட்டுரை ஆக்கத்திற்கு உதவிய நூல் :

- ந. தெய்வசுந்தரம் (Diglossic Situation in Tamil - A Sociolinguistic approach , Ph.D. Thesis submitted to the University of Madras, 1980, Chennai)
- முனைவர் புனல் க. முருகையன் ( பன்னிரு திருமுறை ஒலிபெயர்ப்பு, 2010, காந்தளகம், சென்னை)



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# இணையம் மற்றும் கணினி மூலம் தமிழ் கற்றல் மற்றும் கற்பித்தல்

திருமதி. ரஜனி ரஜத்

முனைவர், பட்ட ஆய்வாளர், பாரதியார் பல்கலைக்கழகம்

## முன்னுரை

மனிதனால் பேசப்படும் மொழிகளுள் வளமும் துல்லியமும் கொண்ட திருந்திய மொழிகளுள் செம் மொழியாம் தமிழ்மொழி ஒன்று. மனித உள்ளத்துக்கும் உணர்ச்சிக்கும் அறிவுக்கும் மொழி மிகமிக நெருங்கிய தொடர்பு கொண்டது தமிழ் மொழியை கற்றலும் கற்பித்தலும் எளிதானதல்ல, இணையம் மற்றும் கணினி மூலம் தமிழைக் கற்றலும் கற்பித்தலும் ஒரு சவால் என்றே கொள்ளலாம். தமிழ்மொழியைக் கற்றலின் அடிப்படைத் திறன்கள். கற்பித்தல் குறித்த நிகழ்வுகள். மொழிச்சிக்கல்கள். மொழியைப் பயன்படுத்தும் திறமை ஆகியவற்றோடு கணினிமற்றும் இணையத்தின் பயன்பாடுகள் குறித்த அறிவு, இணையம் மற்றும் கணினி வழி தமிழ் கற்பதற்கும் கற்பித்தலுக்கும் தேவைப்படுகிறது.

## ஆய்வுக்களம்

இதுகுறித்து சென்னை தரமணியில் இயங்கிவரும் “தமிழ் இணையக் கல்விக் கழகத்தை அணுகி அங்கு ஆலோசகராகப் பணிபுரியும் திரு.ஜேம்ஸ் (M.A P.hd) உடன் நிகழ்ந்த நேர்காணல் மூலம் சேகரித்த தகவல்கள் மிக முக்கியமானவை தஞ்சைத் தமிழ்ப்பல்கலைக் கழகத்தின் ஓர் அங்கமாக இயங்கி வரும் இந்தக் கல்வி நிறுவனம் உலகிலேயே இணையம் மற்றும் கணினி மூலம் தமிழ் கற்பிக்கும் ஒரே நிறுவனமாக விளங்குகிறது இவர்களின் பணிகளை மூன்று விதமாக வகைப்படுத்தலாம்.

1. வெளி மாநில / நாடுகளில் வசிக்கும் தமிழ் கற்க விரும்புவவர்களுக்கு இணையம் மூலம் தமிழ்க் கல்வி கற்பித்தல்.
2. தொன்மையான இலக்கண, இலக்கிய நூல்களைக் கணினியில் ஏற்றுதல் இதுவரை ஒரு லட்சம் சொற்களைக் கொண்ட400க்கும் மேற்பட்ட நூல்கள் கணினியில் ஏற்றப்பட்டுத் தமிழ் ஆர்வலர்களால் பயன்படுத்தப் படுகின்றன.
3. தமிழில் மென்பொருள் செய்பவரை இனங்கண்டு ஊக்குவித்தல்.

## தமிழின் பெருமை

2000 ஆண்டுகளுக்கு மேற்பட்ட தொன்மையும், இலக்கண வளமும், செம்மையான இலக்கியமும் உடையது. தமிழ் மொழி காலத்தைக் கடந்த மொழி. தமிழில் உள்ள இலக்கியங்கள் காதல், வீரம் போன்ற பல்வேறு செய்திகளையும், வாழும் முறைகளையும் கூறுகின்றன. உலக மொழிகள் அனைத்தும் எழுத்துக்கும் சொல்லுக்கும் இலக்கணம் வகுக்கும் போது தமிழ் மட்டுமே வாழ்க்கைக்கும் இலக்கணம் வகுத்த பெருமை பெற்றது.

“இலக்கணமும் இலக்கியமும் தெரியாதான்  
ஏடெழுதுதல் கேடு நல்கும்”

என்கிறார் பாரதிதாசன்.

## தமிழ் கற்றலுக்கான காரணம்

“தமிழன் என்று சொல்லடா  
தலைநிமிர்ந்து நில்லடா”

என்று பெருமை கொள்ளவும், “தமிழன்” என்ற அடையாளத்தை நிலைநாட்டிக் கொள்ளவும் இலக்கியங்களின் நயங்களையும், உணர்வுகளையும் ரசிக்கவும், தமிழ் கற்கின்றனர். இணையவழி தமிழ் கற்றலில் “இலக்கிய ரசனையின் வீச்சு” அதிகமாகக் காணப்படுகிறது.

## நோக்கம்

இணையம் மூலம் தமிழ் கற்க விரும்புவர்களின் நோக்கம் அவர்கள் வசிக்கும் நாடுகளைப் பொறுத்தே அமைகிறது. கீழை நாடுகளான, மலேசியா, சிங்கப்பூர், இலங்கை போன்ற நாடுகளில் வசிப்பவர்கள் வேலை வாய்ப்புக்காகவும், தாய்மண்ணின் கலாச்சாரத்தைப் போற்றிக் காக்கவும், கற்கின்றனர். மேலை நாடுகளான அமெரிக்க, ஐரோப்பிய நாடுகளில் வசிப்போர் தாய்மொழியை விட்டுவிடாமல் நினைவுபடுத்திக் கொள்ளவும், சொந்த ஊர்களில் உள்ள உறவுகளின் தொடர்பு விட்டுவிடாமல் இருக்கவும் கற்றுக் கொள்கின்றனர். ஒரு மொழியைக் கற்கும்போது அந்த மொழிக்குரிய கலாச்சாரமும் கற்பிக்கப்படுகிறது.

## கற்பித்தலின் நோக்கம்

அறிவியல் முன்னேற்றத்திற்கேற்ப புதியவற்றை ஏற்கக்கூடிய மொழி என்ற பெருமை பெற்றது தமிழ் மொழி மின்னணு சாதனங்கள் உதவியால் உலக உருண்டை சுருங்கி உள்ளங்கையில் அடங்கி விட்டது. “உலகமயமாதல்” மொழிக்கும் பொருந்தும் முண்டாசுக் கவிஞன் பாரதி

“திறமையான புலமை யெனில் வெளி நாட்டினர் அதை வணக்கம் செய்திட வேண்டும்”

என்று கூறுவது போல் தமிழின் பெருமையை உலகெங்கும் பரப்பவும், இலக்கண, இலக்கியங்களைப் பாதுகாக்கவும் அதன் மேண்மையினை பரப்புவதற்கும் இணையம் மற்றும் கணினி மூலம் தமிழைக் கற்பித்தல் கட்டாயமாகிறது. ஆரம்பகால கட்டங்களில் ஆங்கிலத்திற்கு அடுத்தபடியாக கணினி வழி கற்றல், கற்பித்தலில் தமிழ் இடம் பெற்று இருந்தது.. காலஓட்டத்தில் சீனம், ஃபிரென்ச் போன்ற பலநாட்டு மொழிகள் தமிழைப் பின்தள்ளிவிட்டன. தமிழ் மொழியைக் கணினியின் மூலம் உலகின் மூலை முடுக்குக் கெல்லாம் பரப்புவதற்குத் தீவிர முயற்சிகள் எடுக்கப்பட்டு வருகின்றன.

## பாடத்திட்டங்கள்

இணையம் மற்றும் கணினி மூலம் தமிழ்கற்க, கற்பிப்பதற்கான பல்வேறு நிலைகள் உள்ளன. அடிப்படை நிலையிலிருந்து பட்டப்படிப்பு வரை கற்பிப்பதற்கான பாடப்பகுதிகள் முன்னரே திட்டமிடப்பட்டு இணையம் மூலம் அறிவிக்கப்படுகின்றன. சுருக்கமான முன்னுரை ஆங்கிலத்தில் வழங்கப்படுகிறது. தேவையான இடங்களில் ஆங்கிலத்தில் மொழிபெயர்த்தும் கூறப்படுகிறது. தமிழ் கற்பதை எளிதாக்கும் வகையில் கற்பிக்கும் முறைகள் கையாளப்படுகின்றன. மொழிப்பற்று, இலக்கண இலக்கியங்களில் புலமை, எடுத்துக் கூறும் ஆற்றல், குரலில் ஏற்றத்தாழ்வு அமைத்துப் பேசும் திறன். திறமையாக எழுதும் திறன், உளநூல் வல்லுநரின் நகைச்சுவை, மொழிபெயர்ப்புத் திறன் இவையாவும் ஒருசேரப்பெற்ற ஆசிரியர் மூலம் பாடத்திட்டங்கள் தயாரிக்கப்படுகின்றன.

## கற்றல் முறை

இணையம் மற்றும் கணினி மூலம் தமிழ்க்கல்வியில் ஆசிரியர் மாணாக்கர் நேரடித் தொடர்பு இல்லை கேட்டல், கேட்டலின் வழிகற்றல், கேட்டல் பழக்கத்தை வளர்த்தல், பேசுதலைக் கேட்டறிதல், படித்தலைக் கேட்டறிதல் போன்ற வழிகளிலேயே கற்றல் அமைகிறது. கற்றலில் கேட்டல் முறையே முக்கிய இடம் பெற்று

“செல்வத்துள் செல்வம் செவிச்செல்லவம்” அச்செல்வம் செல்வத்துள் எல்லாம் தலை”

– என்ற வள்ளுவரின் வாக்கை நினைவு படுத்துகிறது.

## உச்சரிப்பு

தமிழ் மொழி மிகுதியான எழுத்துக்களைக் கொண்டது. ஒரே ஒலியைப் போன்ற பல எழுத்துக்கள் இருக்கின்றன. பேச்சொலிகள், உயிரொலிகள், பேச்சு உறுப்புகள் , இதழ்களின் குவிநிலை, விரிநிலை, நாளும் உயரம், ஒலிப்பான் ஒலிப்படம், ஒலிப்பு முறை, குறில், நெடில் போன்றவற்றை எல்லாம் அறிந்து மொழியைக் கற்க வேண்டும். ஒலிபிறக்கும் இடங்களை நன்கு உணர்ந்தால் செம்மையாகப் பேச இயலும். இணையக் கல்விக் கழகத்தின் மூலம் தயாரித்து மாணவர்க்கு அனுப்பப்படும் குறுந்தகடுகள், விளக்கப்படங்கள் தெளிவான ஒலிப்பு, திருத்தமான உச்சரிப்பு ஆகியவற்றை கற்பிக்கவும் கற்கவும் உதவுகின்றன. திரும்பத் திரும்ப எழுத்தொலிகளைப் பயிற்சி செய்தால் "செந்தமிழும் நாப்பழக்கம்" என தமிழ் கற்பவர் வசப்படும். தமிழில் ஒலியியலுக்கும் எழுத்துக்களுக்கும் வேறுபாடில்லை.

## கற்பித்தல் முறைகள்

தமிழ் கற்பித்தலில் பயிற்று முறை அறிவுடன் அதைத் தெளியப் பயன்படுத்தும் ஆற்றலும் இன்றியமையாதது. கணினி மற்றும் இணையம் வழியே கற்பிக்கும் போது ஆசிரியர்தம் பட்டறிவு, மொழிப்புலமை இவற்றுடன் கணினி மற்றும் இணையம் இவற்றின் பயன்பாடு குறித்த அறிவும் அவசியம். எனவே அதுகுறித்த சிறப்புப் பயிற்சியும் அளிக்கப்படுகிறது. இலக்கண அமைப்பு, தமிழ்மொழிக் கட்டமைப்பு, மொழிக்கூறுகள் ஆகியவற்றில் செம்மையான தெளிவைக் கொண்டு பாடத்திட்டங்களுக்கான குறுந்தகடுகள் தயாரிக்கப்படுகின்றன.

தமிழ் கற்பித்தலில் வாய்மொழிமுறை உரையாடல் முறை, தடைவிடைமுறை, வினாவிடை முறை, விதிவிளக்கமுறை காரணகாரிய முறை, போலக்கற்றல் போன்ற பலமுறைகள் உள்ளன. இவற்றில் உடையாடல் முறை தவிர பிற முறைகள் பாடத்தயாரிப்பின்போது செயல்படுத்தப்படுகின்றன.

முன்னுரை கூறும் போதும் மற்றும் தேவையான இடங்களிலும் தமிழில் மட்டுமல்லாமல் ஆங்கிலத்திலும் மொழிபெயர்த்துக் கூறுவது கட்டாயமாகிறது. சுருங்கக்கூறின், தமிழ் கற்பதற்கு தமிழ் ஆங்கிலம் மூலம் கற்பிக்கப்படுகிறது.

## அகராதிகள்

“தமிழ் இணையக் கல்விக் கழகம்” மூலம் இணையத்தில் இதுவரை 21 அகராதிகள் 9,44,000 சொற்களுடன் இடம் பெற்றுள்ளன, இதனால் அதிகமான சொற்களின் பொருளை குறுகியகாலகட்டத்தில் அறிந்து கொள்ள முடியும். தேடுகுறி (search engine) மூலமாக ஒரு சொல் உபயோகப்படுத்தப்படும் இடங்கள், ஒரு செய்தியைப் பல்வேறு நூல்கள் கூறும் இடங்கள் ஆகியவை பற்றி அறிவது இணையவழிச் கல்வி மூலம் எளிதாகிறது.

## பேச்சுவழக்கு, எழுத்துவழக்கு

தமிழ் பேச்சு வழக்கு, எழுத்து வழக்கு எனும் இருவழக்குளைக் கொண்டது ஆரம்பக் கட்டத்திலிருந்து பாடங்கள் பதிவு செய்யப்பட்ட குறுந்தகடுகள் விளக்கப்படங்கள் மூலம் உச்சரிப்புப் பயிற்சி அளிக்கப்படுகின்றது. பல்வேறு நிலைகளில் நடத்தப்படும் தேர்வுகள் மூலம் எழுத்துக்களின் உபயோகம், பொருள் உணர்ந்து சொற்களை அமைப்பது போன்றவற்றிற்கான பயிற்சி அளிக்கப்படுகிறது. உரையாடல் தமிழைக் கற்பதற்கும் அடிப்படைத் தேவைக்கான தமிழ்வார்த்தைகள் குறுந்தகடுகளில் பதிவு செய்யப்படுகின்றன.

## கற்றலின் நன்மைகள்

கற்றல் கொள்கையின் படி கற்பதற்கு வயது ஒரு தடையல்ல இணையவழிக் கல்வி பெரும்பாலும் வயதுவந்தோருக்கான கல்வியாகவேகருதப்படுகிறது. It is nothing but adult education – காலம், நேரம் கற்பவர்களின் வசதிக் கேற்ப அமைகிறது. இன்றைய சூழ்நிலை, வேலைப்பளு, நேரக்குறைவு காரணமாக உலகின் எந்தப் பகுதியாயினும், இருந்த இடத்தில் இருந்து கொண்டே, கிடைக்கும் நேரத்தில் இணையம் மூலம் தமிழ் கற்க முடிவது மிகப் பெரிய வரமாகும். ஆர்வமும் சுறுசுறுப்பும். ஊக்கமும் திறமையும் எந்த அளவிற்குக் கற்பவரிடம் உள்ளதோ, அந்த அளவிற்கு வேகமாகத் தமிழை நன்கு கற்க முடியும்.

## புதிய கோணம்

எந்தத் துறையிலும் புதிய நோக்குகள் வரவேற்கத்தக்கவை. இணைய வழிக்கற்றல், கற்பித்தல் முறையில் கற்பவருக்கே முக்கியத்துவம் தரப்படுகிறது. ஆசிரியரும் மாணவரும் உடனுறைந்து பயிலுவதே சிறந்தது ஆனால் காலச்சுழற்சியில் அறிவியல் யுகத்தோடு மனிதரும் தம்முடன். சமயம் மொழி போன்றவற்றைப் பிணைத்துக் கொள்ளவேண்டும்.

"உலகத்தோடு ஒட்ட ஒழுக்கல் பலகற்றும்  
கல்லார் அறிவிலாதார்"

என்கிறார் வள்ளுவர் இந்தக்கணினியுகத்தில் செம்மொழியான தமிழ் மொழியின் பெருமையும், இலக்கியச் செழுமையும் உலகத்தின் முலை முடுக்கெல்லாம் வலம் வரவேண்டுமானால் இணையம் மற்றும் கணினி மூலம் தமிழ்கற்றல் மற்றும் கற்பித்தல் மிக மிக அவசியமாகும்.

## முடிவுரை

மொழியின் ஆளுமையும், தனித்தன்மையும் செல்வாக்கும் குறைந்து விடாமல் அறிவியல் முன்னேற்றங்களுக்கேற்ப தமிழ் மொழியை இணையம் மற்றும் கணினி மூலம் கற்றல் மற்றும் கற்பித்தல் ஆரம்பக் கட்டத்திலேயே உள்ளது எனக் கூறலாம். ஆசிரியர் துணையின்றி சுயமுயற்சியின் அடிப்படையில் கல்விகற்க எந்த அளவுக்கு மாணவர் பயிற்சிபெற்றுள்ளார் என்பதைப் பொறுத்தே இணையம் மற்றும் கணினி வழிக் கற்றல் மற்றும் கற்பித்தலின் வெற்றி அமைகிறது.





## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# இணையவழித் தமிழ்ப்பாடங்கள்

முனைவர் மு.இளங்கோவன்

பாரதிதாசன் அரசு மகளிர்கல்லூரி,

புதுச்சேரி-605 003 இந்தியா muelangovan@gmail.com

தமிழ் இணையப் பல்கலைக்கழகம், தமிழம்.நெட் தளங்களிலும், பிற வெளிநாட்டுப் பல்கலைக் கழகங்களின் தளங்களிலும் தமிழ் எழுத்துகளை அறியவும், படிக்கவும், எழுத்துகளைக் கூட்டிச் சொற்களைப் படிக்கவும், சொல்வளம் பெருக்கவும் வசதிகள் உள்ளன. தமிழ் வழியில் தமிழ் படிக்கவும், ஆங்கில வழியில் தமிழ் படிக்கவும் வசதிகள் உள்ளன. தமிழ்க்கல்வி குறித்த பாடங்களை உலக அளவில் அமைக்கும்பொழுது பிறநாட்டுச்சூழல் உணர்ந்து வடிவமைக்க வேண்டியுள்ளது. தமிழகத்துக் குழந்தைகளுக்கு உருவாக்கும்பொழுது தமிழகத்துச் சூழலை உணர்ந்து வடிவமைக்க வேண்டும். இலங்கை, மலேசியா, சிங்கப்பூர், அமெரிக்கா, இங்கிலாந்து, கனடா உள்ளிட்ட நாடுகளில் வடிவமைக்கப்படும் பாடங்கள் அந்தந்த நாட்டுச் சூழலை உணர்ந்து வடிவமைக்க வேண்டும். ஆனால் அண்மைக்காலம் வரை தமிழகத்தைச் சார்ந்து, பாடநூல்கள் வடிவமைக்கப்பட்டுள்ளன. இணையத்தில் தமிழ்க் கல்விக்குப் பயன்படும் செய்திகள் பாடல்களாகவும், கதைகளும் பகுதிகளாகவும் யு டியூப் தளங்களில் பல உள்ளன. இணையத்தில் உள்ள தமிழ்க்கல்வி சார்ந்த செய்திகள் தொடக்க நிலை, அடிப்படை நிலைகளைக் கொண்டு மட்டும் உள்ளது. இவற்றின் தன்மைகளை இக்கட்டுரை அறிமுகம் செய்கின்றது.

மேலும் உயர்நிலை, மேல்நிலை, கல்லூரி, பல்கலைக்கழகம், ஆய்வுசார்ந்த பாடத் திட்டங்கள், பேச்சுரைகள், காட்சி விளக்கங்கள் உருவாக்கப்பட வேண்டும். குறிப்பாகத் தொல்காப்பியம், சங்க இலக்கியம், திருக்குறள், காப்பியங்கள், பக்திப் பனுவல்கள், நன்னூல், இக்கால இலக்கியம் முதலான பாடங்கள் அறிஞர்களின் பேச்சுகளாகவும் (ஒலி-ஒளி), காட்சியுரைகளாகவும் (Power Point) உருவாக்கப்பட வேண்டும். இதனை எவ்வாறு உருவாக்குவது, பராமரிப்பது, இதன் பயன்பாடு, பற்றிய செய்திகளைத் தாங்கி இக்கட்டுரை அமைகின்றது.

## பென்சில்வேனியா பல்கலைக்கழகத்தின் முயற்சி

பென்சில்வேனியாவில் பேராசிரியர் ஷிப்மேன், முனைவர் வாசு ஆகியோரின் முயற்சி யில் இணையம் வழியாகத் தமிழ்கற்றல், பயிற்றுவித்தலுக்குரிய பாடப்பகுதிகளை உருவாக்கி இணையத்தில் வைத்துள்ளனர். (<http://ccat.sas.upenn.edu/plc/tamilweb/> & <http://www.southasia.upenn.edu/tamil>). இதில் இடம்பெற்றுள்ள பாடப் பகுதிகள் பிறமொழிச்சூழலில் தமிழ் கற்போருக்கு உதவும் வகையில் உள்ளன.

நெடுங்கணக்கு அறிமுகப் பகுதியில் தமிழ் உயிர் எழுத்துகளையும், உயிர்மெய் எழுத்துகளையும் கிரந்த எழுத்துகளையும் எழுதவும் ஒலிக்கவுமான பயிற்சிகள் உள்ளன. தமிழ் எழுத்துகளும் அதனை ஒலிக்க உதவும் ஆங்கில எழுத்துகளும் இருப்பதால் ஆங்கிலம் அறிந்தார் தமிழ் கற்க இந்தப் பகுதி பயன்படும். இந்தத் தளத்தைப் பயன்படுத்த எழுத்துருக்கள் தரவிறக்கம், ஒலிப்புக்கருவி மென்பொருள் தரவிறக்கம் செய்ய வேண்டும். தமிழ் எழுத்துகளை (உயிர், மெய்) நினைவுப்படுத்திக்கொள்ள அமைக்கப்பட்டுள்ள படக்காட்சிகள் தமிழ் எழுத்துகளை அறிவோருக்கு நன்கு பயன்படும். மெய்யெழுத்தும் உயிர் எழுத்தும் இணைந்து எவ்வாறு உயிர்மெய் எழுத்து உருவாகின்றது என்ற வகையில் படக்காட்சி வழியாக நன்கு விளக்கப்பட்டுள்ளது.

எ.கா. க்+அ = க ; ச்+அ = ச; ண்+ஓ = ணோ

ஒரு எழுத்துக்கு விளக்கம் கொடுத்து அதுபோல் பிற எழுத்துகளும் எவ்வாறு மாறும் என்பதைப் பயிற்சியில் அறிய வாய்ப்பு உருவாக்கித் தரப்பட்டுள்ளது.

மரம்+ஐ= மரத்தை என்று சாரியை உருவாகும் விதமும் காட்டப்பட்டுள்ளது.

வீடு+இல்= வீட்டில் என்று மாறுவதும் காட்டப்பட்டுள்ளது.

பேச்சுத் தமிழுக்குரிய அடிப்படைக் கருவிகளும் உருவாக்கப்பட்டுள்ளது. வகுப்பறை, வீடு, பொது இடங்களில் பயன்படுத்தப்படும் உரையாடலில் இடம்பெறும் சொற்களை அறிமுகப்படுத்தும் பயிற்சிகளும் ஒலிப்பு வசதிகளுடன் உள்ளன. இவற்றில் வினா-விடை முறை காணப்படுகின்றது. தமிழ் எழுத்துருக்கள் தரவிறக்கி இதனைப் படிக்க வேண்டியுள்ளது ஆங்கிலம்-தமிழ் ஒலிப்பு வசதிகள் இருப்பதால் பிறமொழியினர் தமிழைக் கற்க இந்தத் தளம் பேருதவியாக இருக்கும்.

வினா விடை வடிவம், ஆம் இல்லை வடிவம் எனப் பல வடிவங்களில் சொற்களையும் தொடர்களையும் அறிமுகப்படுத்தி ஆங்கிலத்தின் துணையுடன் தமிழ் கற்பிக்க இந்தத் தளம் பலவகையான நுட்பங்களைக் கொண்டு விளங்குகின்றது.

விடுபட்ட சொற்களைப் பொருத்துதல், பொருத்தமான சொற்களைத் தேர்ந்தெடுத்துத் தொடர்களை உருவாக்குதல் என்ற வகையில் இடம்பெற்றுள்ள பயிற்சிகளும் நிகழ்காலம், எதிர்காலம், இறந்தகாலம் காட்டும் பயிற்சிகளும் சிறப்பாக வடிவமைக்கப் பட்டுள்ளன. தமிழகத்தில் வாய்மொழியாக வழங்கப்பட்டுவரும் நாட்டுப்புறக் கதைகளை அறிமுகப் படுத்துதல் தமிழ் மரபு அறிவிக்கும் செயலாக உள்ளது. தமிழர் பண்பாடு உணர்த்தும் கலைகள், பழக்கவழக்கங்கள் காட்சிப்படுத்தப்பட்டுள்ளமை தமிழர் மரபு அறியவிழைவார்க்குப் பேருதவியாக இருக்கும்.

ஒருங்குறி எழுத்துகளைப் பயன்படுத்தும்பொழுதும் தொழில்நுட்பம் எளிமைப் படுத்தித் தரும்பொழுதும் அனைத்துத் தரப்பு மக்களாலும் விரும்பப்படும் தளமாக இது விளங்கும்.

**தமிழ் இணையக் கல்விக்கழகத்தின் தளம்** <http://www.tamilvu.org/core/site/html/cwhomepg.htm>

தமிழ் இணையக் கல்விக்கழகத்தின் தளத்தில் தமிழ் கற்பதற்குரிய பலவகை வசதிகள் உள்ளன. தமிழ் இணையக் கல்விக்கழகத்தில் உள்ள வசதிகள் யாவும் தமிழ்ச்சூழலில் தமிழ் கற்பாருக்கு உதவும் பொருள்களாக உள்ளன. தமிழை அறிமுக நிலையிலிருந்து பட்டக்கல்வி வரை இந்தத் தளம் சிறப்பாக அறிமுகப்படுத்தியுள்ளது. மழலைக்கல்வி, பாடங்கள், பாடநூல்கள், இணையவகுப்பறை, நூலகம், அகராதி, கலைசொற்கள், சுவடிக்காட்சியகம், பண்பாட்டுக் காட்சியகம் எனும் தலைப்புகளில் உள்ள செய்திகள் யாவும் தமிழையும் தமிழ்ப் பண்பாட்டையும் அறிய விழைவார்க்கு அறிமுகப்படுத்தும் வகையில் அமைக்கப்பட்டுள்ளன.

மழலைக்கல்வி என்ற பகுதியில் பாடல்கள், கதைகள், உரையாடல், வழக்குச்சொற்கள், நிகழ்ச்சிகள், எண்கள், எழுத்துகள் என்னும் தலைப்புகளில் அமைந்து தொடர்புடைய செய்திகள் பொருத்தமுடன் காட்சிப்படுத்தப்பட்டுள்ளன.

**பாடல்கள்** என்ற பகுப்பில் கோழி, காக்கை, கிளி, பசு, முத்தம் தா, நாய் என்று சிறுவர்களுக்குக் கதைப்பாட்டு வழியாகத் தமிழ் அறிமுகம் செய்யப்படுகின்றது. இந்தப் பாடல்கள் இசையமைப்புடனும், படக்காட்சியுடனும் தரவிறக்குவதில் எந்தச் சிக்கலும் இல்லாததால் அனைவராலும் விரும்பிப்பார்க்கப்படும்.

பாடல்களும் பயிற்சிகளும் எனும் பகுதியில் பாடல்களைக் குழந்தைகள் கற்பதற்குரிய வசதிகள் உள்ளன. பயிற்சி பெறுவதற்குரிய கட்டளைகள் எளிமையாக உள்ளதால் குழந்தைகள் தாமே கற்க இயலும். பயிற்சி

பெறுவதற்குரிய பகுதியில் நிலா, கைவீசம்மா, காகம், என் பொம்மை, எங்கள்வீட்டுப்பூனை, பம்பரம் எனும் தலைப்பில் மாணவர்களுக்குப் புரியும்படியான பாடல்கள் உள்ளன.

**கதைகள்** என்னும் தலைப்பில் குப்பனும் சுப்பனும்(கோடரிக்கதை), கொக்கும் நண்டும், புத்தியின் உத்தியால் பிழைத்த குரங்கு, தாகம் தணிந்த காகம் எனும் தலைப்பில் கதைகள் உள்ளன. இந்தக் கதைகள் முன்பே தமிழகத்துக் குழந்தைகளுக்கு அறிமுகமான கதைகள் அல்லது பின்புலங்களைக் கொண்டவை என்பதால் எளிமையாகப் புரியும். இந்தக் கதைகளை எடுத்துரைக்கும் முறையில் அமைத்துள்ளதால் பிறர் உதவியின்றிக் குழந்தைகள் தாமே கதைகளைப் புரிந்துகொள்ள வாய்ப்பு உண்டு. காட்சி, ஒலி வழி அமைந்துள்ளதால் எளிமையாகப் புரிந்துகொள்வர்.

#### **உரையாடல்**

உரையாடல் பகுதியில் ஏழு உரையாடல் பகுதி உள்ளது. குழந்தைகளுக்கு நற்பண்புகளை ஊட்டும் செய்திகள் இதில் இடம்பெற்றுள்ளன. இவை யாவும் படக்காட்சியுடன் விளக்கப்பட்டுள்ளன. சிறவர்களுக்கு நற்பண்புகளை ஊட்டும் இந்தப் பயிற்சியின் வழியாகச் சொற்கள் நன்கு அறிமுகம் செய்யப்பட்டுள்ளன.

#### **வழக்குச்சொற்கள்**

பறவைகளின் ஒலிகள், காய்கள், வீடுகள், விலங்குகளின் ஒலிகள், பழங்கள், கிழமைகள், உறவுப் பெயர்கள், நிறங்கள், சுவைகள் இதில் இடம்பெற்றுள்ளன. ஆறுவகைப் பறவைகளின் ஒலிகள் இங்குக் காட்டப்பட்டுள்ளன. காய்களின் பெயர்கள் ஒலித்துக்காட்டப்படுவதால் சொற்களை எளிமையாகக் குழந்தைகள் அறிவார்கள்.

**நிகழ்ச்சிகள்** என்ற பகுதியில் நிகழ்காலம், இறந்தகாலம், எதிர்காலம் குறித்த காலம் அறிவிக்கும் பயிற்சிகள் உள்ளன.

**எண்கள்** என்ற தலைப்பில் ஒன்று, இரண்டு, மூன்று என்று எண்கள் அறிமுகம் செய்யப்பட்டுள்ளன. பாடம் - பாடல் - பயிற்சி என்னும் பகுப்பில் உள்ள செய்திகள் உள்ளன. இதில் உள்ள பயிற்சிகள் பகுதியில் எண்களின் ஒலியைக் கேட்டுப் பொருத்தமான படத்தைச்சுட்டும் பகுதி அமைந்துள்ளது. குறிப்பாக ஒன்று என்னும் ஒலியைக் கேட்டு, ஒரு பொம்மை உள்ள படத்தைச் சுட்டியால் சுட்ட வேண்டும். பொருத்தமானவற்றைச் சுட்டினால் சரியான விடை எனவும் பொருத்தம் இல்லை என்றால் தவறான விடை என்றும் குறிப்புகள் ஒலிக்கும்.

**பாடல்** என்ற பகுப்பில் ஒன்று முதலான எண்கள் பாடல்வடிவிலும் காட்சி வடிவிலும் விளக்கப்பட்டுள்ளன.

**எழுத்து** என்னும் பகுப்பில் பாடம் - பயிற்சி - பாடல்கள் என்ற தலைப்பில் செய்திகள் உள்ளன. உயிர் எழுத்துகள், மெய்யெழுத்துகள், ஒரெழுத்துச் சொற்கள், ஈரெழுத்துச் சொற்கள், மூன்று எழுத்துச் சொற்கள், நான்கு எழுத்துச் சொற்கள், ஐந்து எழுத்துச் சொற்கள் அறிமுகம் செய்யப்பட்டுள்ளன.

தமிழ் இணையக் கல்விக்கழகத்தில் எளிமையிலிருந்து கடுமைக்குச் செல்வது என்ற அடிப்படையில் பாடங்கள் கதையும் பாட்டுமாகத் தொடங்கி நிறைவில் எழுத்து, சொல் அறிமுகமாக வளர்ந்துள்ளது. தமிழ் இணையக் கல்விக்கழகத்தில் மழலைக்கல்வி- சான்றிதழ்க்கல்வி, மேற்சான்றிதழ்க்கல்வி, இளநிலைக் கல்வி (B.A) உள்ளிட்ட பாடப்பகுதிகளின் பாடங்களும் உள்ளிடப்பட்டுள்ளன. இவ்வாறு உருவாக்கப்பட்டுள்ள பாடங்கள் மாணவர்களின் கல்விநிலையை மனத்தில் கொண்டு உருவாக்கப் பட்டிருப்பினும் அவர்களின் உள்ள நிலையை மனத்தில் கொண்டு உருவாக்கப்படவில்லை. இணைய வகுப்பறை விரிவுரைகள் என்னும் பகுப்பில் சான்றிதழ்க்கல்விக்கான பாடங்கள் அடிப்படைநிலை,

இடைநிலை, மேல்நிலை என்று வகுக்கப்பட்டுத் தரப்பட்டுள்ளன. பேராசிரியர்கள் நன்னன், சித்தலிங்கையா ஆகியோர் இதற்குரிய பாடங்களை அறிமுகம் செய்கின்றனர். ஆங்கில வழியிலும் தமிழ்ப்பாடப்பகுதிகள் சித்தலிங்கையா அவர்களால் அறிமுகம் செய்யப்பட்டுள்ளன. சைவ சமயம் சார்ந்த பகுதிகள் அறிமுகம் செய்யப்படுவது போல் தமிழின் சங்க நூல்கள், இலக்கணங்கள், இலக்கியங்கள் தமிழகத்து அறிஞர்களால் பாடமாக நடத்தப்பட்டுத் தமிழ் இணையக் கல்விக்கழகத்தின் பாடப்பகுதிகளில் வைக்கப்பட வேண்டும். ஒரு பாடத்தைப் பல அறிஞர்கள் நடத்தி அந்தப் பகுதிகள் பயன்பாட்டுக்கு இருந்தால் தேவையானவர்களின் விரிவுரைகளை மாணவர்கள் தேர்ந்தெடுத்துப் பயில முடியும்.

**பொள்ளாச்சி நசனின் முயற்சி** <http://www.thamizham.net/>

பொள்ளாச்சி நசனின் தமிழம்.நெட் தளத்தில் தமிழ் கற்கும் வசதி அமைந்துள்ளது. இத்தளத்தில் தமிழை ஆங்கிலம் வழியாகப் பயிற்றுவிக்கும் வகையில் செய்திகள் உள்ளன. மற்ற தளங்கள் நெடுங்கணக்கு அடிப்படையில் தமிழை அறிமுகம் செய்வதிலிருந்து மாறுபட்டு எழுதுவதற்கு எளிய எழுத்துகளை முதலில் அறிமுகம் செய்து பிறகு மற்ற எழுத்துகளை நசன் அறிமுகம் செய்கின்றார். ட, ப, ம என்று இவரின் எழுத்து அறிமுகம் உள்ளது. எடுத்துக்காட்டாக ஐந்து நிலைகளில்(Level) 19 பாடங்களை (Lesson) இவர் அமைத்துள்ளார். அதனைத் தொடர்ந்து பயிற்சிகளை அமைத்து அடுத்த ஐந்து நிலைகளில் பதினாறு பாடங்களை அமைத்துள்ளார். ஐந்து பாடல் பகுதிகளை அமைத்து அதில் 247 எழுத்துகளையும் பாடி அறியும்படியும் நசன் செய்துள்ளார். அதுபோல் ஓரெழுத்துச்சொற்கள் ஈரெழுத்துச்சொற்கள், மூன்றெழுத்துச் சொற்கள் இவற்றையும் பட்டியலிட்டு அறிமுகம் செய்துள்ளார்.

**தமிழ்க்களம்** (<http://tamilkalam.in/>)

தமிழ்க்களம் தளத்தில் குறியிலக்கும் நோக்கங்களும், தமிழ் கற்றல் கற்பித்தல், வகுப்பறை என்னும் மூன்று பகுப்பில் செய்திகள் உள்ளன. தமிழ்க்களத்தில் பாடங்கள் மூன்று பெரும் பகுதிகளாக அமைந்துள்ளன.

பகுதி 1:

எழுத்துகள் அறிமுகமும் அவற்றாலான சொற்களைப் படித்தலும் எழுதலும்.

பகுதி 2:

சொற்களஞ்சியம் பெருக்கம். பகுதி 3:கேட்டல், பேசுதல், படித்தல், எழுதுதல் ஆகிய திறன்களில் உயர்நிலை எய்துதல்

பகுதி 1: பகுதி ஒன்றில் பதினான்கு பாடங்கள் தமிழ் எழுத்துகளின் வரிவடிவம் அறிந்து ஒலித்துப் பயிற்சி பெற அமைந்துள்ளன. அவ்வெழுத்துகளாலான எளிய சொற்களைப் படிக்கவும் , எழுதவும் பயிற்சிபெற விரும்புவோர் அப்பாடங்களில் தொடர்ச்சியாகப் பயிற்சி பெறவும் வழியமைக்கப்பட்டுள்ளது. வரிவடிவ எழுத்துப் பயிற்சிக்கு என எட்டுப் பாடங்கள் அமைந்துள்ளன ஒவ்வொரு பாடமும் மூன்று கூறுகளை உட்கொண்டுள்ளன. தமிழ்ப் பாடங்கள் பேராசிரியர் திரு.வி.கணபதி புலவர் இ.கோமதிநாயகம் ஆகியோரால் எழுதப் பெற்றுள்ளன. தமிழ்க்களத்தில் எழுத்துரு தரவிறக்கம், ஒலிப்புக்குரிய மென்பொருள் தரவிறக்கம் தேவைப்படுகின்றன. தமிழ்நாடு தொடக்கப்பள்ளி ஆசிரியர் கூட்டணியின் சார்பில் இத்தளம் உருவாக்கப்பட்டுள்ளது.

**பள்ளிக்கல்வி** <http://www.pallikalvi.in/>

பள்ளிக் கல்வி என்னும் தமிழக அரசின் தளத்தில் பள்ளிக்கல்விக்குரிய பாடநூல்கள் இடம்பெற்றுள்ளன (கட்டுரை உருவான சமயத்தில் சமச்சீர் கல்வி குறித்த சிக்கலால் பாடநூல்கள் இடம்பெறவில்லை. எனவே விரிவாகப் பார்வையிட இயலவில்லை).

**தமிழமுதம்** <http://www.tamilamudham.com/Jan11.html>

தமிழமுதம் என்ற இணையவழி வானொலியில் தமிழ் இலக்கியங்கள் சார்ந்த பாடல்கள் ஒலிவழியாகக் கேட்கும் வசதியைக் கொண்டுள்ளது. குறிப்பாகத் திருவெம்பாவைப் பாடல்கள், திருமுறைகள்(பித்தா பிறைசூடி) கேட்கும்வகையில் இனிய முறையில் தொகுத்து வைக்கப்பட்டுள்ளன.

**இந்தியமொழிகளின் நடுவண் நிறுவனம்** <http://www.tamil-online.info/Introduction/learning.htm>

மைசூர் இந்தியமொழிகளின் நடுவண் நிறுவனத்தின் சார்பில் தமிழ்ப்பாடங்கள் இணையத்தில் உள்ளன. இதில் இணையம் வழியாகத் தமிழ் கற்க 500 உருவா கட்டணம் கட்டிப் படிக்க வேண்டும்(அமெரிக்க டாலர் 50). மாதிரிப்பாடங்கள் சிறிது வைக்கப்பட்டுள்ளன. ஒலிப்பு வசதி உண்டு. எழுத்துகளைத் தரவிறக்கிக் கற்க வேண்டும். தளம் புதுப்பிக்கப்பட வேண்டும். தரமான முயற்சியில் இத்தளம் தமிழை அறிமுகப்படுத்துகின்றது.

**வடக்குக் கரோலினா பல்கலைக்கழகம்** <http://www.unc.edu/~echeran/paadanoor/home.html>

வடக்குக் கரோலினா பல்கலைக்கழகத்தளத்தின் தளத்தில் தமிழ் கற்பதற்குரிய பல வசதிகள் உள்ளன. முன்னுரையுடன் மூன்று பகுதிகள் உள்ளன. பன்னிரு இயல்கள் உள்ளன. 38 பாடங்கள் உள்ளன. பின்னிணைப்புகளும் உள்ளன. தமிழ் கற்பதற்குரிய அடிப்படைச்செய்திகள் எழுத்துருச் சிக்கல் இன்றி அமைக்கப்பட்டுள்ளன. ஒலிப்பு வசதி, எழுதிக்காட்டும் வசதி யாவும் கொண்டு தரமான தளமாக இந்தத் தளம் உள்ளது

**இந்தியானா பல்கலைக்கழகம்** [http://www.iu.edu/~celtie/tamil\\_archive.html](http://www.iu.edu/~celtie/tamil_archive.html)

இந்தியானா பல்கலைக்கழகத்தின் தளத்திலும் தமிழ் கற்பதற்குரிய வசதிகள் உள்ளன.

**தமிழ் டியூட்டர்** <http://www.tamiltutor.com/>

தமிழ் டியூட்டர் என்ற தளத்தில் பதிவு செய்துகொண்டால் தமிழைக் கற்கும் வசதியை இந்தத் தளம் தருகின்றது..

**குழந்தைகளுக்கான தளம்** <http://www.kidsone.in/>

குழந்தைகளுக்கான பன்மொழி கற்கும் வாய்ப்புடைய தளம் இது. இதில் இந்தி, தெலுங்கு, தமிழ் மொழி அறிமுகம் எனிய நிலையில் செய்யப்பட்டுள்ளது.

**உறுப்புலவர் தமிழ்மொழி நிலையம்** <http://www.uptlc.moe.edu.sg/>

இனிய இசைகொண்ட அறிமுகப்பாடலுடன் இந்தத்தளம் விரிகின்றது. சிங்கப்பூர் கல்வி அமைச்சின் சார்பிலான தளம். சிங்கப்பூரில் தமிழ் கற்பிக்கும் சில கானொளிப் பகுதிகளைக் கொண்டுள்ளது. சிங்கப்பூரில் மாணவர்களுக்குப் பயிற்றுவிக்கப்படும் பாடத்திட்டங்கள் சிலவும் காட்சி விளக்கவுரைகளும் இந்தத் தளத்தில் இடம்பெற்றுள்ளன (<http://www.uptlc.moe.edu.sg/index.php/ntlrc/primary>).

**எஸ்.ஆர்.எம்.பல்கலைக்கழகத்தின்முயற்சி** [http://www.srmuniv.ac.in/tamil\\_perayam.php#](http://www.srmuniv.ac.in/tamil_perayam.php#)

எஸ்.ஆர்.எம்.பல்கலைக்கழகத்தின் வழியாக இணையவழிக் கல்வி, கணினித்தமிழ்க் கல்வி அளிக்கும் முயற்சிகள் நடந்துவருகின்றன. தமிழ் முதுகலை,இளம் முனைவர் பட்டத்திற்குரிய பாடப்பகுதிகள் உருவாக்கப்பட்டு வருகின்றன. எதிர்காலத்தில் எஸ்.ஆர்.எம் பல்கலைக்கழகத்தின் தளம் தமிழ் உயர்கல்விக்கான தேவைகளை நிறைவுசெய்யும் என நம்பலாம்.

இணையத்தில் இடம்பெற்றுள்ள தமிழ்ப்பாடங்கள் அவரவர்களின் வாய்ப்புகள், தேவைகளுக்கு ஏற்ப உருவாக்கப்பட்டுள்ளன. உயர்கல்விக்குரிய பாடங்கள் இனிதான் உருவாக்கப்பட வேண்டும் அகவை

முதிர்ந்த நிலையில் உள்ள தமிழ்ப்பேரறிஞர்களின் வாய்மொழி வடிவில் தொல்காப்பியம், திருக்குறள், சிலப்பதிகாரம் உள்ளிட்ட நூல்கள் பாடமாக நடத்தப்பெற்று இணையவெளியில் பாதுகாக்கப்பட வேண்டும். அதுபோல் பிறநாட்டுத் தமிழறிஞர்களின் வாய்மொழியிலும் தமிழ்ப்பாடங்கள் நடத்தப்பெற்றுத் தொகுக்கப்பெற வேண்டும். தமிழ் சார்ந்த பாடங்கள் உருவாக்கும் முயற்சி உலக அளவில் நடந்தாலும் இவற்றை எல்லாம் ஒரு குடையில் பார்க்கவும், ஆராயவும், பாடத்திடங்களுக்கு இடையே ஓர்மை காணப்படவும் அறிஞர்கள் சிந்திக்கவேண்டும்.

#### **இணையவழித் தமிழ்க் கல்விக்குரிய தளங்கள்**

- <http://www.pallikalvi.in/Schools/Samacheerkalvi.htm>
- <http://tamilkalam.in/>
- <http://www.tamil-online.info/Introduction/design.htm>
- <http://www.plc.sas.upenn.edu/tamilweb/>
- <http://www.uptlc.moe.edu.sg/>
- <http://www.tamilvu.org/>
- <http://ccat.sas.upenn.edu/~haroldfs/tamilweb/webmail.html>
- <http://www.maharashtraweb.com/learning/learningTamil.htm>
- <http://www.tamilamudham.com/tamil-resources.html>
- <http://www.tamil-online.info/Introduction/learning.htm>
- <http://www.talktamil.4t.com/>
- <http://www.tamiltoons.com/view/14/tamil-alphabet-/>
- <http://www.thamizham.net/>
- <http://ethirneechal.blogspot.com/2010/06/learn-tamil-online.html>
- <http://www.thetamilanguage.com/>
- <http://www.unc.edu/~echeran/paadanoool/home.html> <http://www.learnTamil.com/>
- <http://www.tamilo.com/learn-tamil-education-57.html>
- <http://www.languageshome.com/>
- [http://www.google.com/search?q=learn+tamil&hl=en&prmd=vnb&source=univ&tbs=vid:1&tbo=u&ei=4AfqS5utMIStgODn7WiDg&sa=X&oi=video\\_result\\_group&ct=title&resnum=4&ved=0CDkQqwQwAw](http://www.google.com/search?q=learn+tamil&hl=en&prmd=vnb&source=univ&tbs=vid:1&tbo=u&ei=4AfqS5utMIStgODn7WiDg&sa=X&oi=video_result_group&ct=title&resnum=4&ved=0CDkQqwQwAw)
- <http://www.saivam.org.uk/saivamTamil.htm>
- <http://www.ukindia.com/zip/ztm1.htm>
- <http://www.tamilcube.com/tamil.aspx>
- <http://www.mylanguageexchange.com/Learn/tamil.asp>
- <http://kids.noolagam.com/>
- <http://www.tamilunltd.com/>
- [http://languagelab.bh.indiana.edu/tamil\\_archive.html#basic](http://languagelab.bh.indiana.edu/tamil_archive.html#basic)
- [http://www.srmuniv.ac.in/tamil\\_perayam.php](http://www.srmuniv.ac.in/tamil_perayam.php)





## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011



## தமிழில் தகவல் தொழில்நுட்பத்தைக் கற்பித்தல்: வாய்ப்புகளும் சிக்கல்களும்

வே. இளஞ்செழியன் & சி.ம.இளந்தமிழ்  
மலேசியா (tamiliam@gmail.com)

தகவல் தொழில்நுட்பத்தைத் முழுக்க முழுக்கத் தமிழிலேயே கற்பிக்கும், கற்கும் சாத்தியம் இன்று ஏற்பட்டுள்ளது. அதே வேளையில், அதற்குச் செயல்வடிவம் கொடுக்கும் போது பல சிக்கல்களும் எழுகின்றன. சுமார் ஆறு மாதங்களுக்கு முன், மலேசியாவிலுள்ள ஒரு குழு, தமிழ் தகவல் தொழில்நுட்பப் பாட திட்டமொன்றை மேம்படுத்தத் தொடங்கியது. ஆரம்பப் பள்ளி மாணவர்களுக்கான பாட திட்டமிது. அதனை மேம்படுத்தும் போது அக்குழு எதிர்கொண்ட சிக்கல்களையும், பாட திட்ட நடைமுறையில் திட்டம் கண்டுள்ள வெற்றிகளையும் இக்கட்டுரை அலசும்.

தகவல் தொழில்நுட்பம் நமது உட்பத்தித்திறனை அதிகரிக்கச்செய்கிறது; மாணவர்கள் கற்கும் விதத்தையும், தகவலைப் பெற்று, அதனைப் பயன்படுத்தும் விதத்தையும் மாற்றியமைத்து வருகிறது. இருப்பினும், கணினியைத் தமிழிலேயே இயக்க முடியும் என்று நம்மில் பலர் அறிந்திருக்கவில்லை. கணினி கற்பதற்கும் தகவல் தொழில் நுட்பத்தைப் பயன்படுத்துவதற்கும் ஆங்கிலம் தேவை என்ற தவறான எண்ணம் நிலவுகிறது.

இந்நிலையை மாற்ற வேண்டுமெனில் குழந்தைகளுக்குக் கணினி கற்பிற்கும் முறையில் மாற்றம் ஏற்பட வேண்டும். தொடக்கம் முதலே அவர்கள் கணினியைத் தமிழில் கற்பார்களேயாயின் -- தமிழை இடைமுகப்பு மொழியாகப் பயன்படுத்துவார்களாயின் -- சிக்கலின்றி தகவல் தொழில்நுட்பத்தை அவர்கள் பயன்படுத்தலாம். ஆங்கிலம் தமிழர்களின் தகவல் தொழில்நுட்ப வளர்ச்சிக்கு ஒரு தடைக்கல்லாக இருக்காது.

இதன் சாத்தியத்தைச் சோதித்துப் பார்க்க மலேசியாவிலுள்ள ஒரு குழு எண்ணம் கொண்டது. இந்நாட்டில் மொத்தம் 523 தமிழ்ப்பள்ளிகள் இயங்கி வருகின்றன. ஏறக்குறைய 110,000 மாணவர்கள் இப்பள்ளிகளில் பயில்கின்றனர். ஆங்கிலம், மலாய் (மலேசியாவின் தேசிய மொழி) தவிர இதர பாடங்கள் அனைத்தும் தமிழிலேயே கற்பிக்கப்படுகின்றன. ஆக இம்மாணவர்களுக்கும் தகவல் தொழில்நுட்பத்தைக் கற்பிக்க தமிழே சிறந்த மொழியென இக்குழு கணித்தது. ஆகவே, இங்குள்ள இரண்டு தோட்டப்புற தமிழ்ப்பள்ளிகளில் கணினிக்கூடங்களை அமைத்து பள்ளியிலுள்ள அனைத்து மாணவர்களுக்கும் வாரத்திற்கு ஒரு மணி நேரம் தகவல் தொழில்நுட்ப பாடத்தைக் கற்பிக்க முடிவு செய்தனர். அதற்கு லினக்ஸின் விநியோகத்தில் ஒன்றான உபுண்டுவை (குறிப்பாக எல்.டி.எஸ்.பி. சேவையைக் கொண்ட எடுபுண்டுவை) தெரிவுசெய்தனர். மிக மலிவான விலையில் கணினிக்கூடங்களை அமைப்பதற்கான வசதியை இந்த இயங்குதளம் கொண்டிருந்தது. இக்கணினிக் கூடங்கள் பள்ளி ஆசிரியர்களிடமும், மாணவர்களிடமும் நல்ல வரவேற்பைப் பெற்றது. மேலும் பல கூடங்களை அமைப்பதற்கு அழைப்புகள் வந்தன. இதுவரை ஏழு கணிக்கூடங்கள் அமைக்கப்பட்டுள்ளன. 2,500 மாணவர்கள் அதன்வழி பயன்பெறுகின்றனர். தற்போது இன்னும் ஐந்து பள்ளிகளில் இக்குழு கணினிக்கூடங்களை அமைத்து வருகின்றனர்.

கணினிக் கூடங்கள் அமைப்பதோடு நின்று விடாது, கற்பிப்பதற்கான பாட திட்டத்தையும் இக்குழு தற்போது தயாரித்து வருகின்றது. நான்கு படிநிலைகளிலான பாடதிட்டங்களை இக்குழு தயாரிக்க எண்ணம் கொண்டுள்ளது. முதல் படிநிலை தயாரிக்கப்பட்டுவிட்டது. இரண்டாவது படிநிலை தயாரிக்கும் பணிகள் நடந்து வருகின்றன. தமிழை இடைமொழியாகப் பயன்படுத்தி தகவல் தொழில்நுட்பக் கற்றல் கற்பித்தலுக்கு இப்பாடதிட்டம் வழிவகுக்கின்றது.

ஏறத்தாழ ஓராண்டு காலமாக இத்திட்டம் நடைமுறைப் படுத்தியதில் நாம் இதனை அறிகின்றோம்: தமிழை இடைமொழியாகப் பயன்படுத்துவதன் வழி கடை நிலை மாணவன் உட்பட அனைத்து மாணவர்களும் கணினி அறிவையும் தகவல் தொழில்நுட்ப பயன்பாட்டு அறிவையும் இலகுவாகப் பெறுகின்றனர்.

இருப்பினும், திட்ட நடைமுறையாக்கத்தில் பல சிக்கல்கள் இருக்கவே செய்கின்றன. அவற்றில் முக்கியமானவை--

ஒன்று: உபுண்டு இயங்குதளமும் மாணவர்கள் அதிகம் பயன்படுத்தும் மென்பொருள் களும் இன்னும் முழுமையாகத் தமிழாக்கப்படாமலுள்ளன. தமிழாக்கத்திலும் சிக்கல்கள் உள்ளன. எடுத்துக்காட்டிற்கு, தட்டச்சுப் பழகுவதற்கு ஆங்கிலத்தில் பல மென்பொருள்கள் உள்ளன. இம்மென்பொருள் தமிழாக்கப்பட்டிருந்தாலும் முழுமை பெற்றிருக்கவில்லை; வழுவுடையதாக இருக்கின்றது.

இரண்டு: தமிழ்க்கணிமத்தை முன்னெடுத்துச் செல்ல வேண்டுமெனில் அதற்குத் தேவையான கலைச்சொற்களை தொடர்ந்து உருவாக்கிக் கொண்டே இருக்க வேண்டும் என்று பலர் அறிந்திருக்கின்றனர். இதன் விளைவாக தனி நபர்கள் முதற்கொண்டு அமைப்புகளும் பல்கலைக்கழகங்களும் கலைச்சொல் உருவாக்கப் பணியில் ஈடுபட்டு வந்திருக்கின்றனர், வருகின்றனர்[4]. இவ்வாறு பலரும் கலைச்சொல் திரட்டுகளை வெளியிட்டிருந்தாலும், அது தொடர் முயற்சியாக நடப்பது இல்லை. தவிர்த்து, அத்திரட்டுகளிடையிலும் ஒரு ஒற்றுமை இல்லாமையை நம்மால் காண முடிகின்றது. எடுத்துக்காட்டிற்கு, கணினியை இயக்க பயன்படுத்தப்படும் 'mouse' என்ற கருவியைப் பார்ப்போம். அதனை 'சுட்டெலி' என்றும், 'எலியன்' என்றும், 'சுட்டி' என்றும், 'மவுஸ்' என்றும் பலவாறாக பலரும் அழைக்கின்றனர். எது சரி? எது தவறு?

மூன்று: மாணவர்களுக்குத் தேவையான வலைத்தளங்கள் தமிழில் போதுமான அளவிற்கு இல்லை. மிகவும் பயனுள்ள தளங்களில் ஒன்றான விகிபீடியா போன்ற தளங்கள் கூட குறைவான தமிழ் கட்டுரைகளை கொண்டுள்ளன. இந்நிலை மாற வேண்டும். மாணவர்களை மையமாகக் கொண்ட இன்னும் அதிகமான தளங்களை நாம் ஏற்படுத்தல் அவசியம்.

தகவல் தொழில்நுட்பத்தை தமிழில் கற்பிக்கும் கற்கும் முயற்சி இன்று மெல்ல, மெல்ல தவழத் தொடங்கியுள்ளது. அது எழுந்து ஓட வேண்டுமெனில், நாம் அனைவரும் ஒன்றிணைந்து செயல்படல் வேண்டும்.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Teaching Tamil and Managing Tamil Schools Using Open Source Computing

*Saravanan Mariappan*

*Nexus NGN Sdn. Bhd., Malaysia.*

*Email: nexusngn@gmail.com*

## **Abstract**

Teaching and managing Tamil schools in Malaysia are going through a enormous development and synchronization with setting up and commissioning of sustainable Computer Learning, Teaching and ICT Skill Development Laboratories successfully. To-date there are over 25 computer labs commissioned over a period of two years using open source computing advancement in Malaysian Tamil schools. With the use of open source computing, cost effective solutions for ICT labs were now made available to schools in Malaysia, providing infrastructures needed for teaching and managing educational systems. The key to this achievement were laid upon the innovation of our research and development team. After a painstaking 3 years of hard work, dedication and a lot of financial difficulties, we were able to implement a reliable and cost effective solutions using open source computing. This pioneering work brings integration to Student Management, Classroom Management, Teacher Management and School Management. The implemented school computer lab infrastructure consist of 41 thin clients connected to a server which delivers the required computing speed enabling the students to access wide spectrum of knowledge freely giving equal opportunity in education. Furthermore, a school management application were proposed using open source school ERP (Educational Resource Planning). Managing the educational system were simplified to upgrade the level of school's teaching and management to be comparable with private educational institutions. This open sourced ERP proven to be the cost effective and affordable in term of development implementations and maintenance. This paper will address in detail, how the server based open source computing along with the integrated open source school ERP for schools in Malaysia implemented and how it is gearing up students with sound computing knowledge.

*Keywords:* Thin Client, Open Source, Server Based Computing, Free-ware, ERP For Schools.

## **1. Introduction**

This paper addresses the key area of institutional concern for the education sector, that of delivering effective and efficient school and class room management system in a flexible, secure and accessible way in Malaysian Tamil schools. The system will adopt server based open source computing technology linked with centralized server to implement school and classroom management.

The proposed system will have secure integration with other key educational systems (student records, module registration, examination scheduling, conducting trial exams and distribution of

teaching materials), which will be delivered via network services and a centralized server technology meeting the following requirements:

1. System is required to be online and can be accessed from anywhere and anytime.
  - ⤴ The system should have a user-friendly interface which is easy to use.
  - ⤴ Provide security functions to avoid any unauthorized access.
  - ⤴ Able to have user friendly database search engine.
  - ⤴ Able to update the particulars of individual or organization involved.

Built using the latest open source technology '*ruby on rails*' which works on a web based platform, this school management system automates school's diverse operations, with the objective of :-

2. Systematic User Management
3. Integrated Student Management
4. Incorporated Exam Management
5. Control over Attendance Management
6. Allow for Timetable Management
7. Uploading school news management
8. Other miscellaneous settings

Apart from that, NexusEdu ERP also brings teaching and educational management to a whole new level where all the information (data) is managed by full suite of integrated ERP application as shown in Figure 1.

### Logical Model of Proposed System

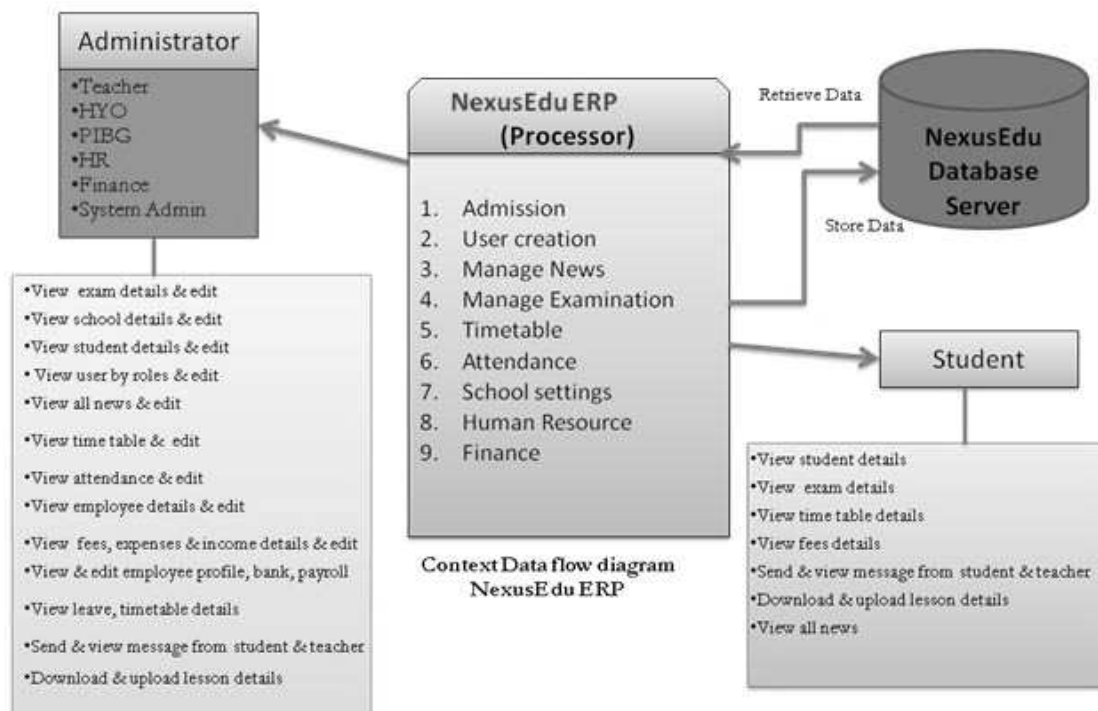


Figure 1 Data Flow Diagram for NexusEdu ERP

## 2. Requirement for System Integration

The integration of School Management System and teaching via Open Source Computing platform is achievable through:

- a) *School Management System* - application that is designed to automate a school's diverse operations from classes scheduling, examination schedule to school events and calendar in order to create a powerful online community, with parents, teachers and students on the common interactive platform.
- b) *Open Source File/Application Server* - that integrates data storage functionality as well as structured database modules.
- c) *LAN (Local Area Network)* that physically connects disk-less Thin Clients to the LTSP Server via DHCP (Dynamic Host Configuration Protocol) in the PXE (Pre-boot execution Environment).
- d) *WAN (Wide Area Network)* that acts as a super highway to access valuable information and Data Centre.
- e) *Thin Client* that is made up of a fully functional computer desktop set minus the hard disk as data is stored on the LTSP server.

Figure 2 shows the integrated network architecture of the ERP.

## 3. School Management System

The school management system integrates the following management functions on to a software to improve the efficiency of school management.

### - User management

Manages the authentication and authorization for different users. For example, students can't access certain management system for security and privacy issues. This management facility provides security, integrity and privacy to the data managed under the ERP system.

### - Student management

Students' information are centralized under the database for easy administration purposes. Student data can be extracted from this database for other management purposes.

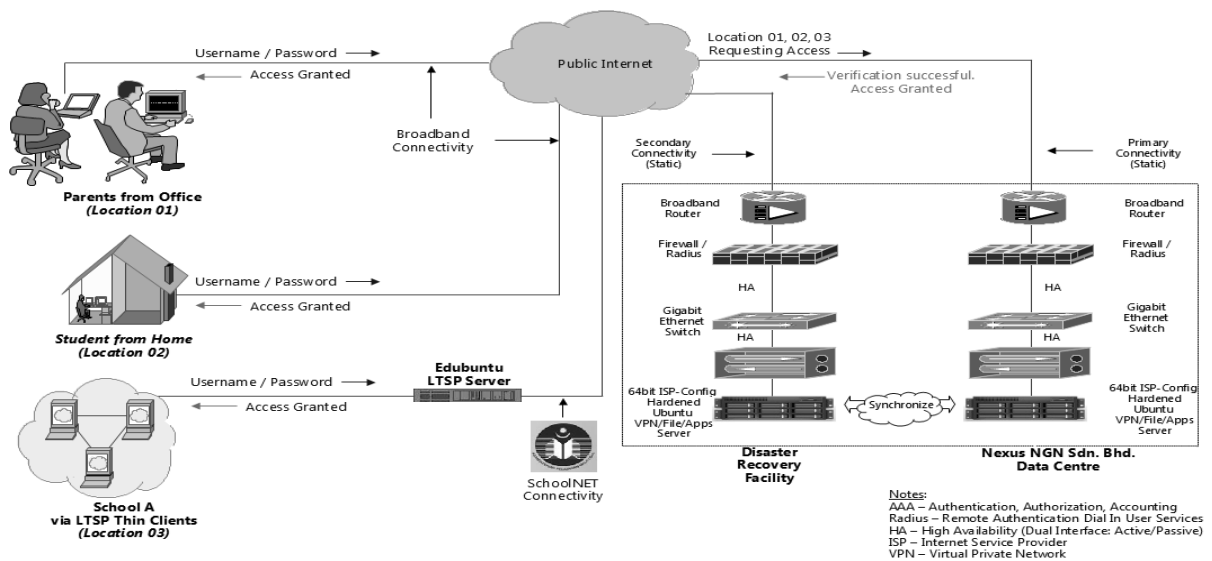


Figure 2 – Integrated network architecture of the ERP

#### - Exam management

Administrators can schedule examinations, set grading systems, generate examination reports while students and parents will be able to view examination schedules and reports. This would assist to monitor a student's overall progress and performances. Other than that, this also eliminates/lessens the need of written progress books and manual update works.

#### - Attendance management

This system benefits teachers and administrators to record and generate daily, weekly and monthly attendance reports. Students can view their records and parents would be able to monitor their children attendance.

#### - Timetable management

Provides the flexibility to create timetable of subjects, classes and view them. We can also change weekday and weekend settings.

#### - News management

Students, employers, and administrators will be able to communicate with each other, with the integrated news management system. News regarding holidays, examinations and special events can be spread to all the parties involved within seconds with this system.

#### - Human Resource Management

Human resource of the school (example: teachers, administration staffs) can be managed efficiently with this system. Employee details, payslip, and attendance could be managed and released using this system.

#### - Finance Management

Fees, assets, donations and payslips can be issued and monitored using this system. This would simplify financial dealings and accounts matters of the school.

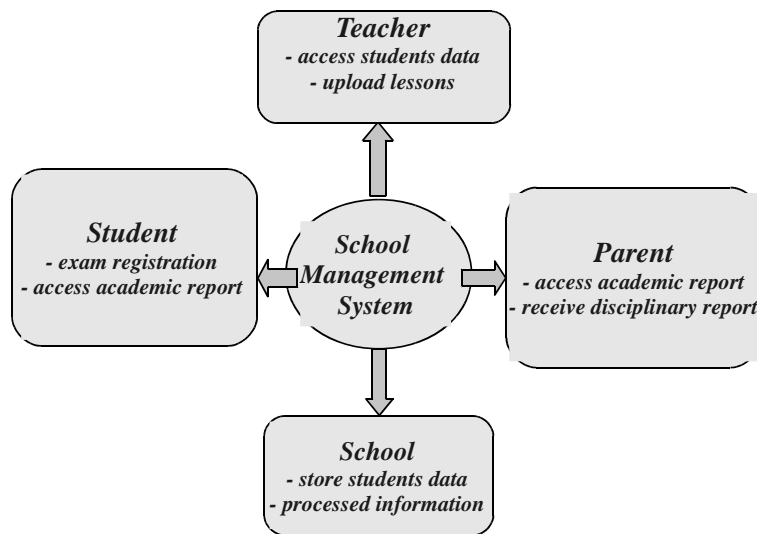


Figure 3 – Integration of School Management System

### 3.1 TEACHING MADE SIMPLE

This is an extra feature included in the school management system where the online learning is implemented. Teachers would be able to upload the teaching materials and post a message to students. Students would be able to download, view or print the lessons exercises prior to the actual lesson hour. This would enable the students to access their educational material anytime and anywhere. Teaching would be made simple and effective through the usage of this online learning tool where interactive materials such as animations, videos and audios could be uploaded for the students to view and learn on their own.

This system also benefits the teachers where the whole syllabus could be loaded into the system and released phase by phase to the students according to the lesson plan. This reduces the teachers work load and makes them productive and effective in teaching and guiding the students. Information sharing between schools is made possible through the existence of centralized server. Schools can share resources to create and use standardized materials and examinations through this system.

## 4. Advantages and System Security of the School ERP

This system benefits all parties in various of ways. The benefits for the school management are as follows:

- 1) Easy performance monitoring of individual teaching modules.
- 2) Automated and quick report generation along with process turnaround time.
- 3) Centralized data repository for trouble-free data access.
- 4) Authenticated profile dependent access to data.
- 5) User friendly interface requiring minimal learning and IT skills.
- 6) Design for simplified scalability.
- 7) Elimination of people dependent processes.
- 8) Minimal data redundancy.



#### *4.1 Advantages to parents are:*

- Frequent interaction with teachers.
- Reliable update on child's attendance, progress report and fee payment.
- Prior information about school events and holidays.
- Regular and prompt availability of school updates such as article's discussions forums, image gallery and messaging system.

#### *4.2 Security*

This ERP system integrates the information security elements, confidentiality, integrity and availability (CIA) for the security and privacy. Confidentiality is where the system prevents disclosure of information to unauthorized individuals or systems. The open source ERP system also provides integrity where data cannot be modified by everyone and undetectably. Apart from that, the system also promises availability where the data is available to authorized users anywhere and anytime. Authorized users can access and view the data through the web based platform which serves as a user interface to the user.

### **5. ADVANTAGES OF OPEN SOURCE COMPUTING SYSTEM**

The thin client system in Tamil school computer labs are capable of providing affordable server-based open source computing solutions. The main advantages of using this system in schools are to increase reliability and consistency of technology. Through a password-accessible account, students, teachers, and administrators can store and can access saved documents and personal settings. As all files and programs are stored centrally, users can access their work from any computers on the network.

Eg. when a teacher or student “logs in”, the server provides them with their “desktop configuration”. Users can even access their “desktop” from home or other remote locations. The other advantages of open source computing thin client system can be listed as below:

- Less Administration – Central management of users, patches, software, data, and backups.
- Higher Security – Elimination of viruses, Trojans or other vulnerabilities on the user desktops.
- Hardware Independence – Support of virtually all client devices and computer hardware, with low system requirements ( eg - Pentium 3 with 512MB RAM).
- Easy Access – Teacher and students can access their documents and applications from any computers in the local area network.
- Reduction in TCO – Total Cost of Ownership reduction by up to 50%

Benefits for school in using open source computing thin client system:

- Lowers cost of technology over time
- Secure data and equipment
- Less downtime and greater efficiency
- Reduces administrator staffing costs

- Lessen the risk of data theft
- Disaster recovery: Data is more Secure
- Reduced time for technical Support
- Lower power consumption: save electricity
- Zero licensing management
- Minimum Maintenance
- Highly trained individuals are not required

## 5. Advantages of Thin Client Technology

By using thin client technology rather than standalone computers, it is possible to deliver a wide range of computer based educational and examination materials while restricting other resources that are usually accessible to the students if conventional computer system is to be used. With conventional computer based technology, it is difficult to prevent access to the Internet, chat services, mobile devices such as USB drive, documents previously stored by other students etc., which could allow simple cutting and pasting of answers into the assessment or exam sheets by students without thin client technology in place. It is simple for an administrator to disable USB port on thin client terminals for the duration of the assessment or examination time, thus further limiting the ability for student's accessing disallowed information to assist them in the assessment or examination.

Another major attraction of the thin client technology for assessment purpose is that it is very resilient, given the fact that they have no software or moving parts. Therefore there are unlikely to be an issue when the assessment are not been delivered due to faulty desktop devices. This causes unnecessary pressure on the affected student and the additional works involved to the invigilator.

The issue of ensuring that computers have the appropriate software available also affects computers which are located in teaching spaces. Traditionally such computers are left switched off when not in use which means that any automated software updates tend to fail or, worse, try to start when a teacher turns the computers on for a class. This can lead to anti-virus software not being updated, operating system vulnerability not being patched etc. the start up time of a computers system also causes difficulties, when a lecturer arrives in a class room, there will be about 5~10 minutes start up time for the conventional computers and to get the necessary software up and running; if any updated needed to be done this could delay the start of the class. Using thin client technology there is no need for the software updates and no need to worry about viruses. The user will always get the appropriate version of all the software via central server. The new upload of teaching material will be ready for teaching immediately as the student or teacher starts the class.

### 5.1 Thin Client System

- ⤴ Thin client is a general term for a device that relies on a server to operate.
- ⤴ Thin client has display device, keyboard with mouse and basic processing power in order to interact with the server.
- ⤴ An ideal thin client device contains no hard drives and CD or DVD-ROM

Plate 1 – Shows an image of ideal thin client.

Plate 2 shows image of computer labs before and after with students using refurbish machines operating as thin client.



*Plate 2: SJK (T) Bukit Raja, Klang before and after setting up computer laboratory*

## **6. Conclusion**

The awareness of benefits and advantages of thin client and server based computing technology have resulted in the growth of Tamil schools implementing this technology in Malaysia, with the supports from governmental and non-governmental bodies. With the use of thin client technology and school management system, the teaching system will now look forward into a new age of centrally manageable teaching technology, with equal access to information will be given to all students regardless of their background and geographical location. Through implementation of this system, Tamil schools in Malaysia will soon become community information hub where resources can be maintained and shared for the uplifting of the Malaysians. The students benefited from this technology will become independent learners and one day become knowledge based skilled leaders.

With the open source applications and thin client technology it was possible to decrease the cost of installation and the cost of maintaining the computer lab. With servers installed at each and every schools, the setting up of centralized school management system is possible. Currently we have successfully implemented server base teaching with thin clients for about 25 over Tamil schools in Malaysia. Currently we are working and developing further improvement into open source by performing R & D into the implementation of Open Source base ERP for schools to manage the 25 schools systematically using centralized server. Most importantly design system are required to be scalable, sustainable, maintenance free and most importantly able to eliminate the digital gap between poor and rich students and build the digital bridge between urban and rural students.

## **7. Acknowledgement**

The authors wish to extend special thanks to: all the individuals and institutions involved directly and indirectly by providing financial and material support for this research work, especially Selangor State Government, National Land Finance Co-Operative Society (NLFCs), Malaysian Community, Education Foundation (MCEF) and Hindu Youth Organisation, Klang.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# **An Innovative Soft Ware for Learning to Write Tamil Lesson Plan**

***Dr. S. K. Panneer Selvam***

*Assistant Professor, Department of Education,  
Bharathidasan University, Tiruchirappalli, TN  
skpskpbd@gmail.com*

## **Introduction**

Quality Education is indispensable for ameliorating the secondary education to stand on the platform of globalization. It can be acquired by revamping the quality of Secondary Teacher Education through the degree course of Bachelor of Education. Teachers are responsible facilitators for ensuring the quality education. Teacher Education can assist to enrich the quality of teachers by implementing innovative methods in teaching learning process. Most of the students are unable to read and write Tamil without mistakes in Tamilnadu. Even if the Tamil is the mother-tongue of the learners, they are unable to write correct Tamil. Parents and learners are attracted by learning English and neglect learning Tamil language. It paves way to the learners for committing more mistakes. Errors of learning Tamil can be rectified through effective teaching learning process. Innovative effective method has to be investigated for easy understanding of the language. Hence innovative teaching-learning software was prepared for error free successful learning to write Tamil lesson plan.

## **Need of the Study**

Effective teaching is based on the planning of lesson plan which is depending upon unit plan. Writing and preparing a lesson plan for macro-classroom is essential for a perfect teacher-trainee or in service teacher. Most of the teacher-trainees of selecting their optional-I as Tamil at Bachelor Education level were unable to write lesson plan in Tamil from some Educational colleges situated around the Trichy district. Lesson plan is the skeleton of the teaching learning process. If the student-teacher is unable to write an appropriate lesson plan for particular unit of the lesson for /poem/prose/grammar, his teaching can not be more effective. Acquiring practice in lesson plan writing may eliminate the problems of the teacher-trainees in teaching learning process. Learning to write Tamil lesson plan among the teacher-trainees was identified by administering a diagnostic-test. Teacher-trainees of Tamil scored very less mark in writing lesson plan. Hence the researcher endeavoured to prepare innovative Software for the teacher-trainees to eliminate the problems in conventional methods of learning to write Tamil lesson plan.

## **Statement of the Problem**

Teacher-trainees of optional I as Tamil in Bachelor of Education have problems in writing lesson plan in Tamil by conventional method.

## **Objectives of the study**

1. To measure the learning hurdles in writing Tamil lesson plan.
2. To find out the significant difference in achievement mean score between pre-test of control group and post-test of control group in learning to write Tamil lesson plan.
3. To find out the significant difference in achievement mean score between pre-test of experimental group and post -test of experimental group in learning to write Tamil lesson plan.
4. To measure the effectiveness of innovative software in learning to write Tamil lesson plan.

## **Hypotheses of the study**

1. Teacher-trainees of optional I as Tamil in Bachelor of Education have problems in writing lesson plan in Tamil by conventional method.
2. There is no significant difference in achievement mean score between pre-test of control group and post-test of control group in learning to write Tamil lesson plan.
3. There is no significant difference in achievement mean score between pre-test of experimental group and post-test of experimental group in learning to write Tamil lesson plan.
4. Innovative software is more effective than conventional methods in learning to write Tamil lesson plan.

## **Methodology**

Experimental method was followed in the study.

### **Sample**

One hundred Teacher trainees of B.Ed were selected from Indira Ganesan college of Education, Trichy as sample for the study. Fifty Teacher-trainees were considered as Controlled group and another Fifty Teacher-trainees were considered as Experimental group.

### **Tool**

Researcher's self-made criterion reference test was used as a tool for the study.

### **Reliability of the tool**

The co-efficient correlation was found 0.78 in the tool through split-half method.

### **Validity of the tool**

Face validity and Content validity was established for the test through expert suggestions. Hence reliability and validity were properly established for the study. 't' test was used as a statistical technique for the study.

### **Procedure of the study**

1. Problems identification by administering diagnostic test.
2. Preparation of innovative software and validation.
3. Pre-test-treatment-post-test.
4. Finding the effectiveness of the software.
5. Implementation of the study.

## Data collection

The researcher administered a diagnostic test to identify the problems of the Teacher-trainees in learning to write Tamil lesson plan with permission of Principal of the college. Pre-test –Treatment-Post-test was used for the control group in conventional method. Pre- test- using the innovative software and post -test was administered for the research.

## Hypothesis testing

### Alternative Hypothesis-1

**Teacher-trainees of optional I as Tamil in Bachelor of Education have problems in writing lesson plan in Tamil by conventional method.**

Teacher-trainees scored 24% of marks in writing Tamil lesson plan. Seventy six percentages of teacher-trainees committed the mistakes in writing lesson plan. It substantiates that the problems existing in writing lesson plan among the teacher-trainees in Tamil by conventional method. Hence the Teacher-trainees of optional I as Tamil in Bachelor of Education have problems in writing lesson plan in Tamil by conventional method.

### Null-Hypothesis-2

**There is no significant difference in achievement mean score between pre-test of control group and post-test of control group in learning to write Tamil lesson plan.**

**Table-1**

| Tests                    | N  | Mean | S.D. | do | t- value | Result        |
|--------------------------|----|------|------|----|----------|---------------|
| Pre-test control group   | 50 | 8.62 | 2.32 | 98 | 0.214    | Insignificant |
| Post- test control group | 50 | 8.68 | 2.42 |    |          |               |

**Achievement means scores between pre-test of control group and post-test of control group.**

The calculated t value is (0.214) less than table value (1.96). Hence null hypothesis is accepted at 0.05 levels. Hence there is no significant difference between the pre-test of control group and post-test of control group in achievement mean scores of the teacher-trainees in writing Tamil lesson plan through conventional methods.

### Null Hypothesis- 3

**There is no significant difference in achievement mean score between pre-test of experimental group and post-test of experimental group in learning to write Tamil lesson plan.**



Table-2

| Tests                         | N  | Mean  | S.D. | df | t- value | Level of significance  |
|-------------------------------|----|-------|------|----|----------|------------------------|
| Pre-test Experimental group   | 50 | 14.39 | 3.68 | 98 | 9.32     | P>0.05<br>significance |
| Post- test Experimental group | 50 | 18.12 | 3.02 |    |          |                        |

**Achievement means scores between pre-test of experimental group and post-test of Experimental group.**

The calculated 't' value is (9.32) greater than table value (1.96). Hence null hypothesis is rejected at 0.05 levels. Hence there is significant difference in achievement mean score between the pre-test of Experimental group and post-test Experimental group in achievement mean scores of the teacher-trainees of B.Ed College in learning to write Tamil lesson plan.

#### **Hypothesis- 4**

**Innovative software is more effective than conventional methods in learning to write Tamil lesson.**

The above two tables prove and confirm the **innovative software** is more effective than traditional approaches in learning to write Tamil lesson plan. Mean scores in pre-test of Experimental group by conventional method is (14.39) less than the mean score of post-test of Experimental group by using **innovative software** in learning to write Tamil lesson plan (18.12). It substantiates that **innovative software** is more effective than conventional methods in learning to write Tamil lesson plan.

#### **Findings**

1. Teacher-trainees of optional I as Tamil in Bachelor of Education have problems in writing lesson plan in Tamil by conventional method.
2. There is no significant difference in achievement mean score between pre-test of control group and post-test of control group in learning to write Tamil lesson plan.
3. There is significant difference in achievement mean score between pre-test of experimental group and post-test of experimental group in learning to write Tamil lesson plan.
4. Innovative software is more effective than conventional methods in learning to write Tamil lesson.

#### **Recommendation of the study**

1. Preparing more software's for learning of Tamil may simplify the learners of Tamil in mother tongue as well as Tamil as second language learners.
2. It may be implemented in Diploma in Teacher Education also.
3. It may be implemented in School Education also.
4. It may be implemented in Collegiate Education also.

## Conclusion

Like this study can eliminate the problems of the learners in all levels of education.

## References

1. **Kwok-Keung LAU(1992)** On The Objectives of Teacher, *Education Journal Vol. 20 No,* pp. 43-48 by The Chinese University of Hong Kong Faculty of Education,
2. **Wahab, Norshahriah and Zaman, Halimah Badioze (2007)** Multimedia courseware Package for learning English based on learning styles (mel-e). *in: ICEEI2007* Bandung, Indonesia.
3. **Britten, J.S. and Cassady, J.C. (2005)** The Technology Integration Assessment Instrument: Understanding Planned Use of Technology by Classroom Teachers: Computers in the Schools Vol. **22**, No. 3/4, pp. 49-61.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# **Attitudes and motivation in teaching through ICT Among Malaysian Tamil Teachers: an overview**

*Paramasivam Muthusamy, Ph.D.,*

*University Putra Malaysia, Malaysia*

*E-mail: param@fbmk.upm.edu.my*

## **Abstract**

It is an accepted fact the progressive motivation and positive attitudes play significant role in attaining learning achievements. Computer based learning too is not an exception to this observation as communication and Information Technology (ICT) have made deep inroads to teaching and learning among teachers and students. The Ministry of Education in Malaysia have up-graded the language laboratories in schools and have installed sufficient computers for both teachers and students to use in the teaching and learning process. It is needless to say that the role of ICT is very important in helping learners in comprehending as ICT provides avenues through unlimited collection of text, sound, pictures, videos, animation and hypermedia (Bruner, 1986). This can support the findings of Fisher (1996), who argues that computational environment is needed to support 'new frameworks' to education. The aim of this study is to explore the attitudes and motivational levels of in-service teachers who are serving in Tamil schools. A questionnaire adapted from (Wong, 2002) will be used to identify teachers' attitudes and their motivational levels in teaching. Besides that an interview will also be carried out to further question teachers' on their attitudes and motivation. The data from the questionnaire and interview will be used to analyze the Tamil teachers' knowledge, attitude and motivation in using ICT in classrooms. This information will equip the researcher if ICT is being explored to the fullest by the teachers in Tamil schools since facilities had been provided by the Education Ministry and ICT is taught as a subject in the Tamil schools.

## **Introduction**

During the past couple of decades Information communication technology (ICT) and its tremendous growth have made remarkable and significant inroads into almost all the disciplines. One of the instruments for the fast developments of ICT is the growth of computers. As a result, one should know very well the power and the potentiality of this medium. Subsequently, one cannot afford to be a computer illiterate in this era of globalization. Not only that, apart from the knowledge in his/her discipline the success of the person depends mainly on his/her extent of knowledge and the potentiality to use and exploit the computer technology in his discipline. The field of education is not an exception to this rule. In fact, one can assert that the field of education can contribute remarkably by exploiting the potentialities of computers and its allied areas such as, multimedia, internet, software development, need based computer assisted language learning/teaching (CALL,CALT) etc.

## **Schemata for the Current Study**

The present study on Attitudes and Motivation in Teaching through ICT among Malaysian Tamil Schools has been viewed from two angles. The first perspective is to view from the student's point of view and the second is to view from the teacher's point of view. Seeing through these two angles is very important because there is a significant gap between the teachers and the taught as far as ICT is concerned. In other words, the teacher's attitude and motivation get reflected on the student's and the student's attitude and motivation get reflected on the teacher's. This paradigm shift gets reflected on the achievement levels of the learners as well as the teachers.

## **Ict and Students in Malaysia**

Though the knowledge level in the field of ICT among the student population in Malaysia depends on several social, economic, linguistic and educational factors, it is an established fact that the students born after 1980 in general are considered as having digital mind and also known as N-Gen-Net Generation (Tapscott,1998). These groups of students are highly motivated and influenced with internet, computer application and have changed their learning attitudes and achievement levels (Adone et al. 2007). Subsequently, these students are more at ease in the use of computers and also have the expertise to exploit the potentialities of computers and its related areas of education and in other areas of acquiring knowledge. As a result, they are fully aware of the use of computers, multimedia packages, internet etc. They are also aware how these can help them remarkably in the form of its collection of texts, sound and pictures, video graphics and hypermedia in order to increase their knowledge and learning process. Subsequently, they use them extensively whenever necessary and their motivational and attitudinal levels are very high as far as the use of ICT is concerned.

## **ICT and Teachers in Malaysia**

In this situation the development of the basic positive attitude towards the acceptance and the use of the computer are necessary among the teachers. Attitude here is referred to the tendency to behave positively or negatively towards an object, situation, concept or a person (Aiken 1976).

As opposed to the highly motivated and with positive attitude of the general student population in the use of computers and other ICT applications, the teacher's population can be grouped into three categories on the basis of their age, education level in computer application and the number of years of exposure to computers and its applications in general and education in particular. On the basis of the three factors mentioned above the teachers in Malaysian educational context are concerned there are four dimensions which act on them in the formation of attitudes towards ICT. They are, age, background of the teacher, belief and the teacher's extent of exposure to ICT. Age can be divided into two categories namely, those who are above 50 and those who are below 50; background or the opportunity to acquire the computer related knowledge; the extent of exposure to the computer application coupled with the opportunity to use the ICT applications and overall belief system among the teachers. These are the four dimensions which contribute towards the formation of the teacher's motivational factors in ICT. On the basis of the four dimensions listed above the motivational factors for attaining confidence, development of computer based supporting skills, building positive environment around them and the varying degrees of anxiety which affect the formation of motivation will be determined.

## Objectives

The main objectives of this study are,

- To identify the attitudes of the teachers towards teaching through ICT
- To identify the motivational factors which contribute for teaching through ICT
- To identify the relationship between the attitude and motivation in achieving the goal of teaching through ICT

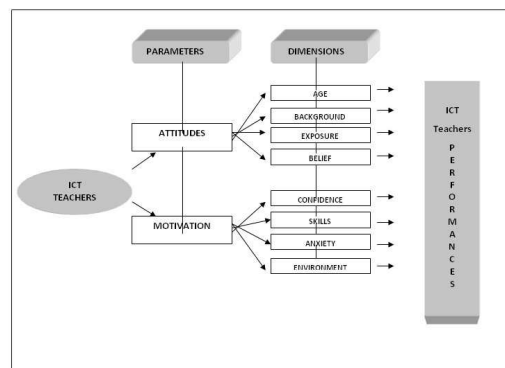
## Hypotheses

- Exposure to ICT contribute positively for the development of positive attitude towards teaching through ICT
- Higher the age group lower the motivation to use ICT in teaching
- Lower age coupled with higher exposure to ICT has a positive role to play on the development of efficient teaching strategies

## Methodology

This study which is mainly focusing on the quantitative approach to the study of attitude and motivation will be undertaken in the 10 Tamil schools situated in Klang Valley, Peninsular Malaysia. Five teachers (including ICT teachers) from each Tamil schools will be taken as the subjects of the study (n=50). These teachers are directly involved in teaching ICT in schools. A questionnaire adapted from Wong (2002), will be used to identify teachers' attitudes and their motivational levels in teaching ICT in the selected schools. The questions in the questionnaire are classified under two categories namely, the attitudinal questions and the questions related to motivational aspects. Each category mentioned above has 4 dimensions. For attitudes, the selected teacher's background, age, belief and exposure to ICT are included. In order to obtain information regarding motivational aspects of the teachers the ways through which they gain confidence, attain teaching skills and manner through which the teachers try to avoid the mounting pressure on them which result into anxiety. The data will be analyzed using SPSS software. Descriptive and inferential analysis such as frequency, percentage, mean and standard deviation will be used to describe the general data of the study. Besides this, analysis such as independent T-test, ANOVA and Pearson Correlation will be employed to discover any relationship and differences between the dependent and the independent variables of the study.

## Framework



The study will look at ICT teachers' performances toward teaching ICT from two parameters and eight dimensions. The two dimensions are teachers' attitudes and motivation towards imparting ICT knowledge to their students. To measure teachers' attitudes several information based on their age, background, exposure and belief in ICT will be gathered through a specially designed questionnaire. The other dimension is motivation. Teachers' confidence, skills, environment and anxiety in using computers will be analysed. These information will reflect ICT teachers performance in their classrooms.

## Conclusion

The achievements of using ICT in all the Tamil schools nationwide in Malaysia does not depend only on the ICT laboratory which are well equipped and other facilities provided by the government but also on teachers involvement. Teachers' attitudes and motivation towards ICT teaching plays an important role in promoting and imparting ICT knowledge to students. Therefore, this study looks into how teachers' attitudes and motivation helps in promoting ICT usage in all the Tamil schools.

## References

- Adone,D., D.,Dron, J.,Pemberton, L. & Bagne,C.(2007). E-Leaning environments for digitally minded students. Interactive leaning Research, V 18(1)pg 41-53
- Aiken,.(1976). Update and other affective variables in learning Mathematics. Review of Educational Research, 46,293-311
- Bruner,J.(1986). Toward a Theory of Instruction. Cambridge, MA: Haward University Press.
- Fisher,G.(1996). Making learning a part of life beyond the 'gift Wrapping" approach to technology. Retrieved act 26, 1999 from the world wide web: <http://www.CS.Colorado.edu/~13d/presentations /gf-wlf/>
- Paramasivam.M.(2002). Malaysia Tamil School and ICT Usage. TI 2002 Conference Proceeding p192-197.
- Tapscott,D.(1988). growing up digital: how the web changes work, education, and the ways people learn. change Magazine, pg 11-20
- Wong (2002). Development and Validation of an information Technology (IT) Based instrument to Measure teachers IT preparedness. PhD Thesis.





## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011



# தகவல் பரிமாற்றுத் திறமைகள் மூலம் தமிழ் மொழி , கலாசாரம் கற்பிக்கும் வழிவகைகளைக் கட்டியெழுப்பல்:

இங்கிலாந்து அரசாங்கத்தின் தேசிய கொள்கை அபிவிருத்தி  
-அடைவையும், குறிக்கோள் சார்ந்த ஊக்கத்தையும் அதிகரித்தல்  
**Raising achievement and aspiration**

சிவா பிள்ளை

லண்டன்

## தேசிய பாடத்திட்டம்

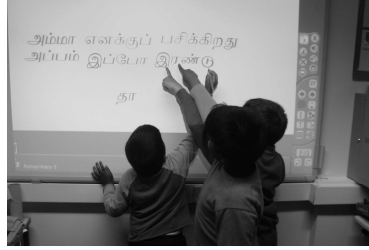
"மாணவர்கள் நமது மனிதாபிமானத்தையும் வேறுபாடு ஒற்றுமைகளையும் கற்று உணர்ந்து மதித்துக் கொள்ளத் தெரிந்து கொள்வதால் தங்களது சொந்த பந்தங்களுக்கு இடையே நிறைவான சிறப்பான புரிதலோடு இயங்க முடிகிறது என்பதே வாழ்க்கையைப் படித்து உணர்ந்து கொண்டதன் முதன்மையான பகுதியாகும்". " (National Curriculum, 1999, p (.136

இதன் அடிப்படையில் இங்கிலாந்தில் தமிழ் மொழிக்கென ஒரு பாடத்திட்டம் ஏனைய ஐரோப்பிய மொழிகளுக்கு இருப்பது போல் உருவாக்கப்பட்டு -2006ல் இருந்து இங்கிலாந்து அரசாங்க முன் பள்ளிகளிலும் வார முடிவில் ,பள்ளி முடிந்த பின் மாலை நேரத்தில் நடக்கும் தமிழ் பள்ளிகளிலும் பின்பற்றப்பட்டு தமிழ் மொழி கற்பிக்கப்படுகிறது .இதில் கற்பிக்கும் ஆசிரியர் பலர் தாய்நாட்டில் பயிற்றப்பட்ட ஆசிரியர்கள் ஆவார்கள் ,சிலர் தமிழ் கற்பிக்க வேண்டும் என்று ஆர்வம் கொண்ட சமூகத் தொண்டர்கள் ஆவார்கள் .

இந்நாட்டிற்கு ஏற்றவாறு பயிற்றப்பட்ட ஆசிரியர்கள் எண்ணிக்கை மிகக் குறைவாகவே உள்ளது , இதனால் இந்நாட்டு மாணவர்களிடம் அணுகும் வழிவகைகள் ஆசிரியர்களிடம் காணப்படுவது குறைவாகவே உள்ளது .இதற்கென பல பயிற்சிப் பட்டறைகள் இந்நாட்டு பல்கலைக்கழக ஆர்வலர்களால் நடத்தப்பட்டிருந்தும் அதில் பங்கெடுத்து அனுபவம் பெறுபவர்கள் எண்ணிக்கை மிகக் குறைவாகவே உள்ளது.

கற்பிக்கும் வழி முறைகளில் பல மாற்றங்கள் உண்டாகி வருகிறது .இவ்வளவு காலமும் ஒரு கட்ட அமைப்பு தமிழ் மொழிக்கென இருந்ததில்லை .சமீபத்தில் வெளியிடப்பட்ட பாடத்திட்ட கட்டமைப்பு இந்நாட்டில் மொழிகற்பிக்கும் வழி முறைகளுக்கு ஏற்ப அமைந்திருப்பதுடன் தேவையான வளங்களையும் அணுகுமுறைகளையும் அது கொண்டிருக்கிறது .அத்தோடு கேம்பிறிஜ் பல்கலைக்கழகம் ஏற்படுத்தி இருக்கும் தராதரம் அளிக்கும் அசெற் மொழிகள் (ASSET Languages) (assetlanguages.org.uk) திட்டம் பல சமூக மொழிகளினால் வரவேற்கப்பட்டுள்ளது . இந்நாட்டில் -1000க்கு மேற்பட்ட மக்களால் பேசப்படும் 26மொழிகள் தேர்ந்தெடுக்கப்பட்டு அவற்றில் 12மொழிகளுக்கு) தமிழ் உட்பட (எல்லாவற்றிற்கும் ஒரே தரமான கட்டமைப்பை உருவாக்கப்பட்டமை ஒரு வரப்பிரசாதம் ஆகும். இது பயிற்றப்பட்ட பயிற்றப்படாத ஆசிரியர்களுக்கு ஒரு வழிகாட்டியாகும் .அத்தோடு அசெற் மொழிகள் ஸ்தாபனம் உருவாக்கப்பட்டதும் இந்நாட்டு மொழி ஸ்தாபனங்களான Our Languages திட்டம் (Cilt, SSAT, NRC) போன்றவையுடனும் முன்பள்ளிகளுடனும் பங்காளித்துவமாக முன் பள்ளிகளில் நடக்கும் தமிழ் பள்ளிகள் சமூகப் தமிழ் பள்ளிகள் கூட இணைந்து நடப்பதால் இன்று தமிழ் மொழி தனித்துவம் பெற்று இந்நாட்டு ஐரோப்பிய மொழிகளுக்குச் சமமாக **அங்கீகாரம் பெற்ற OCR** தராதரச் சான்றிதழ் கிடைக்கும் அளவிற்கு உயர்ந்து நிற்கிறது .அத்தோடு ஐக்கிய இராச்சியத்தில் )ஐஇ-UK) மொழிகள்

கற்கும் கற்பிக்கும் வரைபடத்தில் தமிழும் ஒரு மொழியாக இருக்கிறது. தொடந்து கற்றல் கற்பிப்பதில் மதிப்பீடு செய்து குறைபாடுகளை அடையாளம் கண்டு அவற்றை திருத்தி அமைத்து ஒரு தொடர் முன்னேற்றத்திற்கு வழி அமைக்க வேண்டும். தற்போது மாணவர்கள் கணினி மூலம் கற்பதையே விரும்புகின்றனர் ஏனைய பாடங்களை அவர்கள் கணினி உதவியுடன் கற்றறிவதால் தமிழ் மொழியையும் அவ்வாறு கற்கும் வழி முறைக்கு அவர்கள் உந்தப்படுகின்றனர் .அதிகமான மாணவர்கள் சொந்தமாகக் கணினி வைத்திருக்கிறார்கள் ஏனையோர் கணினியை பாவிக்கும் வாய்ப்பைப் பெற்றிருக்கிறார்கள் .ஆகவே மாணவர்கள் சுயமாக இணையத்தள இணைப்பு மூலம் தமிழை (E-learning)கற்கும் வாய்ப்பை பெற்றிருக்கிறார்கள். பக்கம் பக்கமாகப் புரட்டிப் படித்த மாணவன் இன்று கணினித் திரையில் தடவித்தடவி பக்கங்களை மாற்றுகிறான்.



IWB-Interactive White Board

கரும்பலகையில் வெண்கட்டியால் எழுதிப்படித்த மாணவன் இன்று மின்-வெண்பலகையில் **Interactive White Board (IWB)** மின்-பேனாவால் எழுதிப் படிக்கிறான் தொட்டுப் படிக்கிறான் நடைமுறையில் கணினித் தொழில்நுட்பங்கள் கற்றல் கற்பிப்பதில் பல மாற்றங்களை ஏற்படுத்தி உள்ளது .இங்கு மாணவன் தனக்கு விரும்பிய நேரத்தில் ,தனக்கு ஏற்ற சூழ்நிலையில் கற்பதற்கு வாய்ப்பு ஏற்படுகிறது . இதனால் அவன் சுதந்திரமாக கற்கும் ஒரு சூழ்நிலை உருவாகிறது .இன்று இணைய இணைப்பு கைஅடக்கக் கணினியில் தொலைபேசியில் இல்லாத மாணவர்கள் எண்ணிக்கை இந்நாட்டில் மிகக் குறைவு .ஏனைய பாடங்களில் இந்த வசதி கொண்டிருப்பதால் மாணவர்கள் தமிழையும் இவ்வழியில் கற்க முற்படுவார்கள் .இந்த மாற்றங்களுக்கு அமைய நாமும் தமிழ் கற்பிப்பதை மாற்றி அமைக்காவிடின் காலப்போக்கில் தமிழ் கற்றல் கற்பிப்பதில் மாணவர்களின் ஆர்வம் குன்றிப்போக வாய்ப்பு உண்டு

இணைய இணைப்பு வசதிகள் தான் இதில் முக்கிய பங்கு வகிக்கிறது .முன் பள்ளிகள் எல்லாவற்றிலும் இந்த இணைப்பு வசதிகள் உண்டு .முன்பள்ளிகளுடன் பங்காளித்தத்துவத்தை அரசாங்க முன் பள்ளிகளில் ,வார முடிவில் நடத்தும் தமிழ் பள்ளி ஆசிரியர்கள் ஏற்படுத்தி கணினி பாவிக்கும் வசதியை உண்டாக்கி தமிழ் கல்வி கற்பிக்கும் வளங்களை மேம்படுத்தின் இன்றைய மாணவர்கள் தமிழை ஆர்வமாகக் கற்பதற்கு அதிக வாய்ப்பு உண்டு .ஏனைய ஐரோப்பிய மொழிகளில் இவ்வாறான வளங்களைத் தாமே உருவாக்கிக் காட்டும் வாய்ப்பினையும் மொழி கற்கும் கற்பிக்கும் குறுந்திரைப் படங்களை ஆக்கும் திறமையையும் எமது மாணவர்கள் கொண்டுள்ளார்கள் .இவ்வாறான செயற்பாடுகள் பரந்த சிந்தனை ஆற்றலையும் ,கலை கலாச்சாரத்தையும் கண்டறிய உதவும் என்பதும் மொழி ஆராய்வாளர்கள் கண்டறிந்த உண்மை.

**Siva Pillai**

- Chief Examiner of Cambridge University ASSET Languages-Tamil Language
- Principal Examiner London - Edexcel Examination – iGCE-Tamil Language
- Winner of European Award for Languages 2007
- Visiting Lecturer, Goldsmiths, University of London, UK



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Facebook and Tamil Language in Singapore's Teacher Education

*Seetha Lakshmi*

*Asian Languages & Cultures*

*National Institute of Education, Singapore*

## **Abstract**

Media serves as an important motivator in the language teaching process(Brinton, Donna. 2001).Especially internet, which is accepted as the concrete compound of all the technologies(Kartal, E., and Arikan, A., 2010), enables us teachers/trainers to speak the same language as our students/trainees do. At the same time, they help us to connect to our customers in a personalized way of teaching language. In Singapore, Information and Technology has played a critical role in developing teachers and it is also the same for the Tamil teachers. At the National Institute of Education, the pre-service learning for the Tamil teachers was conducted this year with an instructional designer's technical guidance and moral support, I have embarked on with the pedagogical training through Facebook, Posterous and Voice-thread to develop writing skills, listening and speaking skills for the trainees in diploma in Education and Master in Education courses respectively. With the process based product approach, through these well-known social networks, we witnessed that there are many channels for our students to learn and construct knowledge in their lives and teacher is not the only person to provide knowledge and monopoly in the classroom (Lee Sing Kong, 2011). Through the shared, open concept based networks, the participants have enjoyed and constructed their knowledge on culture, identify the teenage related topics, interesting RJ (Radio Jockey) techniques, responsive, critiquing listenership and creative producers of the feature programmes. This paper practically shares the Tamil Language Teacher Training through Facebook in order to develop 21<sup>st</sup> century educators for teaching the 21<sup>st</sup> century learners in Singapore Tamil Classrooms. We chose Facebook, a social networking platform, for role playing by creating fictional characters and the project went well with the course participants and received invaluable responses. This paper also shares the importance of having the technical and instructional expertise with the same lingo as you have.

Key words: *Facebook, Tamil Language, Second Language, Social network*

## **Introduction**

In Education, technology especially information technology has its own trademark in instilling interest and motivation among students towards education. Also it is a right hand and tutoring assistant for the educators who are passionate on teaching or educating their students. In our National Institute of Education, we have a division to help the academics to embark on IT infused pedagogy. Although there were various kinds of support of the academics in their teaching through information technology, recently the formation of CeL which is the Centre for e-Learning is a boon for us. I myself

was engaged in the following pedagogical deliveries with the assistance of IT and they are: Vimba live, Web quest. However, from last year onwards I had a privilege to work with Instructional Designer to learn new technologies to enhance my pedagogical initiatives. Facebook is one of the innovative pedagogical methods which we felt are successful and we have received excellent feedback from our trainees from the two groups of diploma students.

Around April which was the end of 2010 January semester at our NIE, I had expressed my wish to the CeL colleagues, Ms Pratima Majal, Senior Instructional Designer and Ms Shamini Thilarajah, Instructional Designer to introduce social network based pedagogical training to the Diploma in Education year I and year II Tamil trainees (please see the annexe) in the forthcoming semester which was the September 2010 semester. That lead to a few discussions on preparing the ground work by Pratima and Shamini who are the senior instructional designer and Instructional designer respectively s at CeL. This project is about infusing social communication network modes in the teaching and learning of Tamil language. In this particular project, Facebook mode has been used in the teaching and learning of second language. As Facebook is familiar among youngsters, we would like to explore this network.

### **Learning and the Information Technology:**

For the past two decades, we could witness the infusion of technology in the teaching and learning fraternity. Also, the trainers/teachers are working closely with their students to come up with presentations to encourage and entertain one another. Here are some of the examples on the effective use of information technology and research initiatives on it. Nowadays, language learning is not only to communicate, but also to establish contacts, meet people and establish partnerships (Soontiens, 2004 in Sarah Elaine Eaton, 2010). In a collaborative online community, each student's ideas and knowledge are available and are a resource for everyone in the class (Hewitt and Scardamalia, 1998 in Kathryn I Mathew Emese Felvegi and Rebecca A Callaway, 2009). It gives an opportunity for collective knowledge and connections. Online information technology allows students to obtain information through their cognitive, emotional (Kartal, E., and Arikan, A., 2010), cultural, psychological experiences.

Judith Rance-Roney (2010) stated that the digital story telling technology for the English Language Learners to understand the cultural background, literacy skills and the language development to deal with the literary texts. At the same time, use of 3D Virtual Learning Environment for the students to understand the social words and personal learning (Jonathan Barkand and Joseph Kush, 2009). Waters (2008) and Sarah Elaine Eaton (2010) stated that the use of Skype which allows international connection between students and teachers from two or more countries proved good results in developing second or foreign language skills and teachers' professional development in pedagogy. At the same time, using blogs as an ICT tool in the language class is an effective tool and facilitator to develop reflective learning strategies among students (Hourigan T and Murray L., 2010). Faizah Mohamad, 2009 reiterated that an Internet based grammar instruction is useful in language class and facilitates understanding to the students instead of Conventional pen and board instruction Robert Hamilton 2010 encourages to provide curriculum support always to the lower proficiency students as well as the higher proficiency students through YouTube. Nowadays, computer technology enables students in their language learning and they are: 1. Experiential learning 2. Motivation 3. Enhance

student achievement 4. Authentic materials for study 6. Greater interaction 6. Individualization 7. Independence from a single source of information and 8. Global understanding (Lee, 2000 in Arif Bulut, 2004). Mobile Assisted Language Learning (MALL) and Computer Assisted Language Learning (CALL) for students to harness their creativity to express themselves and take ownership of their learning. In a class, with the teacher there are other inevitable factors to contribute the success of the curriculum: CD, DVD players, Blogs(Doris de Almeida Soares, 2008)scola Naval, Internet, Wiki, Web quests, Virtual fieldtrip, spreadsheet programme, software assisted writing, web 2.0, desktop publishing programmes, graphic organizers, recording devices, podcasting and media(Gwen Troxell, Castleberry and Rebecca B Evers, 2010),. Facebook is another feature in this line and it plays an essential and critical role in today's education, politics, social development and cognitive development.

Facebook which refers to the distributed authorship, collaborative and cooperative learning, openness, careful and purposeful usage of web .20, developing cultural awareness.

### **Tamil language and the information communication technology (ICT):**

In Singapore, Tamil has been taught as a mother tongue language at second language level. With the government support for the Tamil language, almost every school is equipped with necessary ICT infrastructure. Outside India, in Singapore with Tamil having the official language status, Tamil education has been developing its own curriculum and pedagogy. It is also contributing to the Tamil internet. In the Tamil speaking world, information technology has many facets and here are some of those initiatives. Tamil has been used for a variety of reasons with ICT for teaching, computational linguistics, mobile devices and providing assistance to the less privileged students. When we surface the presentations at the last year's Tamil internet conference(Vasu Renganathan, 2010), the papers are mainly in 9 categories. At the recent Tamil Internet Conference in India the following issues on Tamil and ICT were dealt with:

#### **Teaching and Learning of Tamil**

- Tamil Diaspora: Teaching Tamil as a second language and impact of Technology
- Technical Development
- Tamil in Mobile Phones and Handhelds
- Natural Language Processing: OCR Text to Speech Machine Translation Etc.,
- Tamil E-texts, Corpora and Digitization of Ancient Tamil Texts
- Morphological Tagger
- Electronic Dictionaries and Glossary of Technical Terms

Here in this project, we wish to enhance our trainees' knowledge through networking with one another on culture and upgrading their knowledge by interacting critically on culture with one another.

## Project Objectives:

There is a small story behind this project. Before I started this project, I thought of using this project as an effective platform to develop the training on the four critical language skills. Also based on my observations on my trainees, I found that they need more assistance in understanding the in-depth meaning of Indian culture especially Tamil culture and the traditional practices. Although many of them are from Tamil speaking homes, they had less time to use the language in schools and community domain as they studied Tamil as a second language and generally they have fewer opportunities to meet friends from their own ethnic group. It is also difficult to get them hooked towards the Tamil literary functions. In the internet, there is limited information for the Tamil Diaspora to read and understand. Quite a significant number of them haven't been to India to experience or immerse in that cultural world. Hence, I shared my wish to Pratima and Shamini about the project on Tamil culture.

## Objectives of the Tamil curriculum in Singapore:

Here, the objectives of the Tamil curriculum in Singapore on learning of Tamil Language are given below:

- Providing proper training to the students in the basic language skills in Listening, Speaking, Reading and writing
- Explaining the Tamil cultural and traditional features
- Helping them to acquire the characters which are essential for the formation of a country(MOE, 2008)
- For Tamil students in Singapore, the main initiative by the community is **to make Tamil a living language** by actively using it at home and community
- Nurturing Active Learners and Proficient Users of the mother tongue language (MOE, 2010).
- To make it happen, the Tamil learning has to be fun and cool. Students would like to enjoy the lessons while learning the language.

Based on the curriculum objectives, there is a clear understanding that to enable the 21<sup>st</sup> century students to be equipped well in their mother tongue/learning of Tamil language, we need to have well equipped teachers to teach them. Hence to create a 21<sup>st</sup> century teacher, he or she should know about the module in that semester:

- Infusing Tamil language teaching through Facebook network
- Equipping teaching and writing skills in Tamil language
- Developing positive social networking skills
- Learning from society and providing learning to others(peers) as well
- Responsible learning and teaching practices
- Understanding the culture of the Tamil community in Singapore and Diaspora countries

Based on the above mentioned needs, we designed the project with the following goals:

### Goals:

- Equip the trainee teachers to pick up the necessary teaching skills and to use Facebook in their schools
- Understanding the responsibilities of using Facebook as a teaching material
- niche areas on teaching through Facebook
- Engaging students in an enriching way
- Letting them understand that Facebook provides new avenues for teaching
- **Making Tamil learning as a fun and cool feature!**

With the above mentioned goals, we also have some niche areas to try and implement this project. They are the main reasons to embark on this project.

### The need for introducing this project:

- Tamil trainees need to work on new pedagogies
- They need to engage their students well and make teaching an interesting feature
- Tamil trainees need to provide assistance to their students to speak well in Tamil
- To make Tamil as a living language in Singapore
- To make Singapore a Hub for teaching Tamil as a second language

Although in my earlier modules, I have introduced the infusion of ICT in the teaching of Tamil language modules with web quest, student centered lesson learning package, vimba voice, video conferencing and distance learning through ICT. Although each and every initiative has its own unique features, this particular initiative is different from the previous ones. What are the differences?

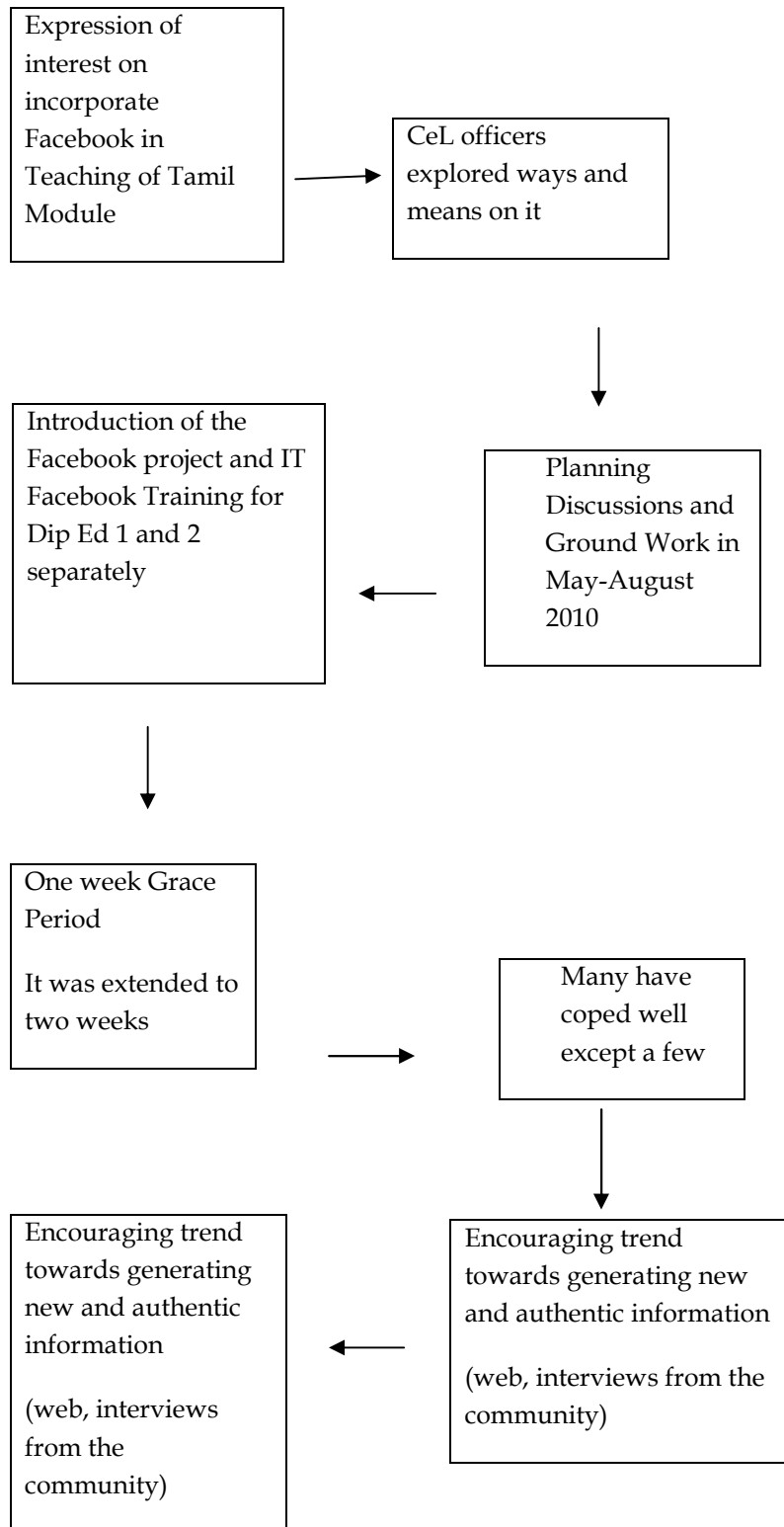
In this project the following are the potential different features while compared to the previous IT projects based Tamil teacher training:

- Harnessing Information Technology in teaching practices
  - Sharing and gaining information through their interactive postings
  - The use of Facebook by the trainees has responsibilities and is different from the normal Facebook usage which is for leisure.
  - They have to have regular updates of new information from their contacts.
  - They can play farm wheel game and can send pictures related to this
  - Evaluation based on weekly postings and reflections. Especially, on -Writing skills, depth of understanding,
- Questions are posed by them and responses are given by them as well
- They have to write reflections weekly for evaluation

In this project, we have given strong emphasis for the process than the product as the process is the feature to give value to educational software or an educational pedagogy. Let us view them here:



### Process of the project:



Facebook is used in Singapore regardless of their age, educational background, socio-economic status and professional background. But, we could not conclude that everybody understands the process or the potential positive and negative features of using it in their lives. Here, future leaders of education

and gatekeepers of the Tamil language in Singapore, we would like to ensure that these young trainees know well about the process of this project and to instill in them that the process is important in every phase of their professional life and teaching and learning of Tamil language in Singapore. Also, Mary Clarie (2010) stated that “It seems that the future of social media in the classroom will not reach its fullest potential until we can bridge the divide between new media and traditional academia”. Also both parties made it clear in this joint effort between the pedagogic and the ICT instructional designer. In this Facebook related ground work and ICT training to the students, we have make sure that continuous monitoring is there. At the same time, it is easy to come up with product. But it is difficult to come up with product based on a Process. So, the process is the important for us. Now let us go through on how the Process is important to us in this project.

## **How the Process is important?**

### **The Process:**

- It's a set of Lesson and Evaluation procedures
- Pre – preparation between the trainer and the ICT Instructional Designer
- Provide training to the trainees from two classes
- Tight timeline to be familiar within Facebook
- Each trainee needs to create a name and profile based on his / her country and context
- Sign undertaking on responsible use of Facebook
- Three postings per week for 12 weeks
- One reflection for every week
- Pictures, songs, video clips and artifacts in their postings
- Immersed in their Facebook interactions and reflections
- Closely monitored by the trainer for content, paraphrasing and by the ICT Instructional Engineer on the use of Facebook and the information technology
- Evaluation is there for their movement, entries, reflections, pedagogical and ICT knowledge
- Although there is a grace and transition period of 2 weeks, but a few of them performed well from the first week onwards.

We would like to express our thoughts that the above mentioned process went well in this project. At the same time, I have to mention that I have received excellent support from the IT Instructional Designer at each and every level of this journey. Also, both of us were able to speak, write and communicate well which is an additional advantage. We could easily share the nuances of the language, culture and context based information. At times, the trainer remember their ‘Facebook names’ and at times forget their real names.

### **Feedback:**

With the feedback on this project and as a trainer, I could say that this project creates a very good understanding and learning of Tamil culture, writing, paraphrasing and reading. It provides avenues to teach, vocabulary, Spoken Tamil, functional grammar and Listening and Speaking of Tamil. This Facebook contents will be a rich resource for the trainees understanding of authentic Tamil culture

and practices. It encourages effective search, note sharing and collaborative learning. The project has a number of features to add fun and cool in Tamil learning. It can be easily adapted for the upper Primary and Secondary class students.

In this project, students faced some challenges at the starting of the project as they have to get a suitable character role and create profile for themselves. After they have positioned themselves well in their roles, they had a constant challenge to prepare weekly postings and reflections for 12 weeks in the midst of their normal training and other modules they had to cover for that semester. Not only that, they also need to learn new information about their country and need to learn each other's cultural domains. Then, they had to ensure that there is good quality inputs and paraphrasing. This is to prevent plagiarism. In the whole process, each group of trainees (Dip Ed I and Dip Ed II) had to control themselves from interacting with the other group of trainees.

In the journey of this project, as a trainer and facilitator, I too faced a number of challenges as given below:

- Providing positive comments and suggestions during the first two weeks
- Familiarizing Tamil Terms and reading and replying to postings in the Facebook
- A hands-on session given for the Dip Ed II class proved to be informative and helped to clear and clarify many doubts
- Availability of Tamil software outside office is a good advantage
- First two weeks, there was a slow development and a certain level of hesitation among the trainees in following the procedures.

But most of the challenges were turned as happy developments in the middle and later part of the project. They are given below:

- Trainees did their postings regularly
- They were Creative and Critical in their reflections
- There was more interaction between themselves in knowledge creation
- Not much questions were posed to Veerasamy and Muthusamy

### **Facebook information, conversations and thoughts: A Midterm review:**

During the 12 weeks of journey time, the facilitator tried to encourage their efforts and provided needed explanations on their cultural domains based postings. The trainees themselves appreciated and critically analysed their classmates' postings and raise awareness with additional questions. This encourages the whole group to move to a level up and working hard to come up with more additional information. The instructional designer provided the updates on the students' postings and advice on their queries regarding the technical issues. At the same time, the facilitator engaged and encouraged the participants to provide their insights on Indian culture especially Tamil culture and traditions.

In October 2010, we had come up with a table to know their understandings of Tamil culture and traditions. Here the results are given below:

**Diploma in Education I trainees on their understanding of content in this project:**

**N = 13**

| எண்<br>No | வகை<br>Category            | பொருள்கள்/ தகவல்கள்<br>Information   | முன்பே<br>தெரியும்<br>I knew<br>already | இப்போதுதான்<br>தெரியும்<br>Now only I<br>know |
|-----------|----------------------------|--|---|---|
| 1.        | Food                       | சக்குநீர்(Dry Ginger Tea)  | 2                                       | 11  |
| 2.        | Food                       | சமையலில் மிளகின் பயன்பாடு(Use of pepper in Tamil cooking)  | 9                                       | 4   |
| 3.        | Food                       | அஞ்சறைப் பெட்டி(Spices box with five rooms)  | 2                                       | 11  |
| 4.        | Costumes                   | திருக்குறள் பட்டுச்சேலை(Thirukkural Silk Saree)  | 1                                       | 12  |
| 5.        | Costumes                   | ராஜ்புத் ராணியின் சேலை<br>விருப்பம்(Rajput Queen's Like on Sarees)                                 | 1                                       | 12  |
| 6.        | Traditional Arts           | பரதநாட்டியத்தில் உள்ள பலவகை<br>நடனபாணிகள்(Various Dance Forms in Bharathanatyam Dance)             | 8                                       | 5   |
| 7.        | Traditional Arts           | தட்டடவு(A dance form called, <i>thattadavu</i> )   | 3                                       | 10  |
| 8.        | Traditional Arts           | கோலாட்டம்(Kolaattam- Dance with two sticks)  | 9                                       | 4   |
| 9.        | Traditional Arts           | கரகாட்டம்(Karagaattam -Dance with a pot on the Head)   | 11                                      | 2   |
| 10.       | Traditional Arts           | குச்சுப்புடி(Kuchupudi- A dance of Andhrapradesh, India)   | 9                                       | 4   |
| 11.       | Traditional Arts           | 108 நடனக்கரணங்களில் வல்லவர்<br>நடராஜர்(Nadarajaa's dance talents)                                  | 5                                       | 7   |
| 12.       | Ancient Tamils' Lifestyles | சீயக்காயின் தனித்தன்மை,<br>குளிர்ச்சி(Seeyakkaai -Shampoo vegetable's uniqueness and coolness)     | 6                                       | 7   |
| 13.       | Ancient Tamils' Lifestyles | சந்தனத்தின் தனித்தன்மை,<br>பாரம்பரியச்சிறப்பு(Sandal wood's uniqueness and traditional speciality) | 4                                       | 9   |
| 14.       | Ancient Tamils' Lifestyles | தமிழ் வாஸ்து சிறப்பு(Special features of Tamil <i>Fengsui</i> )                                    | 3                                       | 10  |
| 15.       | Ancient Tamils' Lifestyles | புகுமனைப் புகுவிழாவின் சிறப்பு(Special meaning of the housewarming function/celebration)           | 2                                       | 11  |
| 16.       | Ancient Tamils' Lifestyles | தொட்டில் துணியின் சிறப்பு(Specialities of the cradle cloth)  | 2                                       | 11  |
| 17.       | Ancient Tamils' Lifestyles | முகப்பராமரிப்பில் இயற்கை<br>மூலிகைகளின் பயன்பாடு(Use of Natural Herbs in Facial Care)              | 3                                       | 6   |
| 18.       | Contemporary Art           | திரைப்பட உலகின் சாதனை(Tamil Film Industry's achievement)   |   |   |

Diploma in Education II trainees on their understanding of content in this project:

N=12 (3 absent)

| எண் | Category                       | Information<br>பொருள்கள்/ தகவல்கள்   | முன்பே<br>தெரியும்<br>I knew<br>already | இப்போதுதான்<br>தெரியும்<br>Now only I<br>know |
|-----|--------------------------------|--|---|---|
| 1.  | Food                           | ஆரஞ்சுநிறச் சர்க்கரையின் பயன்பாடு(Use of Orange Sugar)   | 4                                       | 8   |
| 2.  | Food                           | பலவகை லட்டு(Laddu)   | 4                                       | 8   |
| 3.  | Food                           | வாழைப்பழப் பருப்பு தோசை(Banana Nuts Thosai)  | 0                                       | 12  |
| 4.  | Ancient Tamils' Lifestyles     | பனை ஓலை விசிறியின் பயன்பாடு(Use of the Fan made up of Plam Leaves)   | 0                                       | 12  |
| 5.  | Mythological Stories & Values  | கிருஷ்ணா மற்றும் நரகாசுரா படம்(A video clip on Krishna and Narakasura)   | 6                                       | 6   |
| 6.  | Traditional Arts               | தஞ்சைப் பெரிய கோவில் கொண்டாட்டம் (Arts Festival at the Tanjore Big Temple)   | 4                                       | 8   |
| 7.  | Traditional Arts               | கரகாட்டத்தின் பலவகைகள்(Varieties in karagattam Dance)  | 5                                       | 7   |
| 8.  | Traditional Arts               | விஜயா மோகனின் கோலச் சாதனை(Guinness Record of Vijaya Mohan's kolam)   | 2                                       | 10  |
| 9.  | Overseas Tamils and Traditions | நமஸ்தே ஃபிரான்ஸ் விழா(A specially organized event titled, Namaste France)  | 1                                       | 11  |
| 10. | Overseas Tamils and Traditions | பொன்ஜோர் இந்தியா(A specially organised event titled, Bonjour India)  | 2                                       | 9   |
| 11. | Overseas Tamils and Traditions | அமெரிக்கா, லண்டன், இலங்கை, ஆஸ்திரேலியா ஆகிய நாடுகளில் பொங்கல் கொண்டாட்டம் (Pongal festival celebrations at USA, London, Sri Lanka and Australia) | 6                                       | 6   |
| 12. | Overseas Tamils and Traditions | ஆஸ்திரேலியத் தமிழ்ச் சங்கங்கள் (Australian Tamil Associations and Activities)  | 2                                       | 10  |

Two groups of trainees gave their comments on their projects at the mid of the semester in October 2010 and they are given below:

### **Diploma in Education Year 1 Trainees' comments:**

#### **Advantages**

- Although it is difficult to find answers for all the questions raised by our friends, there is a kind of happiness at the end as I know and I learnt new information.
- Easy to learn a variety of information at the same time. Know more information about our culture which I did not know before.
- It is easy to find the answers for the question. At the same time, can learn a number of rare details about our culture.
- Facebook is an excellent training ground to search, collate and store information for our communication of ideas.

#### **Challenges**

- There is a shortage of time. Because of this, it is difficult to read everybody's comments, postings and questions. At the same time, it is difficult to ask questions from everybody.
- I find it very difficult to type in Tamil. Hence it takes a lot time to upload the information.
- Time is a concern.

### **Diploma in Education Year II trainees' comments:**

#### **Advantages**

- FB has enabled me to see a new way in teaching and learning Mother Tongue.
- Has raised awareness and interest to learn,
- I have enjoyed FB interaction.
- FB has bridged people from different countries and also helps to inculcate our Indian traditions and cultures.
- The FB project has enabled me to take ownership of my learning.
- The friendly interactions between friends from various countries enabled me to share resources and information effectively.
- FB project has incorporated effective role play.
- ICT incorporated project which has enabled people of many countries to join in one network.
- This has enabled us to comment on each other's way of celebrating certain occasions and learn more about the cultures in other countries
- Having a Facebook account in Tamil has been quite fun.
- It is definitely an enriching experience as it provides opportunity to explore the laments of Facebook in Tamil.

#### **Challenges**

- The challenge however is that sometimes it is difficult to track previous comments made by other FB friends.

- If primary school students are to be engaged in this process, time would be a factor which they have to consider.
- We need more time to read, do research in order to post, comment and reflect effectively.
- Spend time commenting effectively, takes time and little research.
- Strong passion, commitment and love for the language are important. If a student just comes in to comment for the sake of doing, it defeats the whole process.

## Evaluation

For the Dip Ed I and II trainees, this project is part of their pedagogical module. Hence we had weightage of 45% of marks for the year 1 trainees in their teaching of Tamil language I module's main project which comprises 70% of marks for the whole semester. For the Dip Ed II trainees, this project focuses nearly 50% of their major project which forms 70% of my part of the module marks.

Generally, we could witness that the year 1 trainees were very enthusiastic and involved in this project with a number of comments and quality comments than the year 2. Although their groups are different, their topics are more or less same on culture. At the same time, technical expertise and experience based analysis; the year 1 trainees were well versed than the year 2. It also because of their IT orientation in the previous two Tamil projects in their first year modules. For year 1, this is their first IT based Tamil project for their very first Tamil pedagogical module.

### Evaluation on Facebook for the Diploma 1 and 2 trainees:

| No of trainees Year.1 | Marks (%)    | No of trainees Year.2 |
|-----------------------|--------------|-----------------------|
| 0                     | 91-100%      | 0                     |
| 2                     | 81-90        | 4                     |
| 5                     | 71-80        | 2                     |
| 5                     | 61-70        | 6                     |
| 1                     | 51-60        | 0                     |
| 0                     | 41-50        | 0                     |
| 2                     | 40 and below | 1                     |

There was a qualitative feedback collection done with both groups at different dates. That provided more insights on this projects and our planning of future projects.

Generally the pair of trainer and instructional designer team came up well and it is a milestone in the teaching of Tamil language and learning. We have learnt more and we will use our experiences in

planning our future modules in order to provide confidence to our younger teacher trainees who generally have learnt Tamil as a second language. They can be a role model to their trainees as well. Here some of our future plans:

- To share the development in this field with the Tamil communities in Singapore, India and Diaspora
- Publish a book with all the inputs
- Provide guidance to the trainees in their Teaching Apprenticeship and Teaching Practice and eventually in their future schools
- Able to do research on Singapore Tamil Trainees' writing, questioning and answering techniques.

In this January 2011 semester, we have embarked on a project to enable the lower primary students from primary 1 and 2 classes to speak up and use spoken Tamil in their oral presentations and conversations. It went very well and we received positive feedback from the trainees.

For the Master of Education course participants, we have introduced radio Jockey (RJ) style of oral features and presentations on teenage related hot topics. It also went well with the course's matured and senior teachers who are in service. We will share our experiences and lessons on them at different platforms.

## **Acknowledgements**

**I wish to record my sincere acknowledgements to:**

- Ms Shamini Thilarajah, Instructional Designer, Centre for e-Learning for her greater support in the whole process of this Facebook project
- Ms Pratima Malar, Senior Instructional Designer, Centre for e-Learning
- Dr Ashley Tan, Head, Centre for e-Learning for inviting us to share about this project with the NIE staff and highlight this project twice at the NIE seminars and giving constant encouragement to me
- My Tamil Trainees from Dip Ed I and II classes for their full support for our new initiatives in teaching and learning of Tamil language

## **References**

- Brinton, Donna. 2001. The Use of Media in language Teaching. In Marianne Celce Murcia. (Ed.). Teaching English as a Second or Foreign Language. Third Edition. New York: Heinle & Heinle. Pp. 459-476.
- Doris de Almeida Soares, 2008. Understanding class blogs as a tool for language development. Language Teaching Research October 2008 vol. 12 no. 4 517-533
- Gwen Troxell Castleberry and Rebecca B Evers, 2010. Incorporate Technology into the Mother Tongue classroom. Intervention in School and Clinic. January 2010 vol. 45 no. 3. Pp. 201-205



- Judith Rance-Roney (2010). Jump starting Language and Schema for English-Language Learners: Teacher – Composed Digital Jumpstarts for Academic Reading. *Journal of Adolescent & Adult Literacy*. 53(5). Pp.386-395.
- Kathryn I Matthew, Emese Felvegi & Rebecca A Callaway. 2009. Wiki as a Collaborative Learning Tool in a Language Arts Methods Class. *Journal of Research on Technology in Education* (2009). 51-72
- Kartal, E., and Arikan, 2010. A. "A recommendation for a new Internet-based environment for studying literature," *US-China Education Review*, 7(7), 93-100 (2010). Lee Sing Kong.2011. Guest of Honour's address at the NIE e-Feasta. Singapore: National Institute of Education.
- Mary Claire, 2010. Blog post in Digital Learning Trends . Website: <http://gradegurublog.com/tag/facebook-for-learning/>, Accessed on 05 May 2011.
- MOE. 2008. 2008 Syllabus Tamil Language Primary, Singapore: Ministry of Education
- MOE. 2010. Nurturing Active Learners and Proficient Users. 2010 Mother Tongue Languages Review Committee Report. Singapore: Ministry of Education. Website: <http://www.moe.gov.sg/media/press/files/2011/mtl-review-report-2010.pdf> Accessed on 05 May 2011.
- Sarah Elaine Eaton, Using Skype in the Second and Foreign Language Classroom. Presented at the Social Media Workshop: "Get your ACT (FL) together online: Standards based Language Instruction via Social Media on August 4, 2010.
- Triona Hourigan and Murray Liam, 2010. Using blogs to help language students to develop reflective learning strategies: Towards a pedagogical framework. *Australian Journal of Educational Technology*, 26(2), 209-225. Waters S., 2008. Skype Other Classrooms! <http://thedublogger.com/want-to-connect-with-other-classrooms>.
- Vasu Renganthan(Ed.).2010. Tamil Internet Conference Proceedings. Coimbatore.

**Annexe:****DIP ED I & II (Primary) Trainees' FB Profile Name Lists****Module: DCT100****Diploma in Education I class**

| No  | Name(English) | FB Profile Name                                      |
|-----|---------------|--|
| 1.  | Mdm XXXX      | மனீஷாராய் (Manisharai)                               |
| 2.  | Mr XXXX       | வேலு சாமிநாதன்(Velu Saminathan)                      |
| 3.  | Miss XXXX     | சங்கீதமேதை சர்விஷ்வாதினி(Music expert Sarvivaadhini) |
| 4.  | Miss XXXX     | சரோஜாதேவி(Sarojadevi)                                |
| 5.  | Miss XXXX     | நீலாம்பரி சரவணன்(Neelambari Saravanan)               |
| 6.  | Miss XXXX     | அன்னலெட்சுமி முத்து(Annaletchumi Muthu)              |
| 7.  | Mdm XXXX      | பெட்டிக்கடை மாதவன்(Sundry Shop Madhavan)             |
| 8.  | Miss XXXX     | அஞ்சலி ரகுராம்(Anjali Raguram)                       |
| 9.  | Miss XXXX     | கவிதா நாயர்(Kavitha Nair)                            |
| 10. | Miss XXXX     | தில்லானா மோகனா(Thillaana Mogana)                     |
| 11. | Miss XXXX     | வளையாபதி அன்னம்மா(Valayaapathy Annamma)              |
| 12. | Miss XXXX     | ஆரத்தீஸ்வரி ஆரத்தி(Aaratheeswari Aarathi)            |
| 13. | MR XXXX       | தண்ணிக்காட்டு ராஜா(Thannikaattu Raja)                |

| N0  | Name in English | FB Profile Name   |
|-----|-----------------|---|
| 1   | MISS XXXX       | வைஷ்ணவி ரகுராம்(Vaishnavi Raguraam)                         |
| 2   | MISS XXXX       | கரகாட்டக்காரன் மாங்குயிலு (Karagaaattaakkaaran Maanguiyilu) |
| 3   | MISS XXXX       | சிங்கக்குட்டி ஓமனக்குட்டி(Singakkutti Omanakutti)           |
| 4   | MISS XXXX       | மல்லிகை முல்லை(Malligai Mullai)                             |
| 5   | MR XXXX         | அவுட்டா ராக்கி(Avutta Raakki)                               |
| 6   | MISS XXXX       | ஷ்ரேயா சேகரன்(Shreyaa Segaran)                              |
| 7   | MISS XXXX       | முத்தழகு சிங்கவேலன்(Muthazhagu Singavelan)                  |
| 8   | MDM XXXX        | குண்டலகேசி சோனா(Kundalagesi Sona)                           |
| 9   | MISS XXXX       | ஓவியா சுந்தரி(Oviya Sundari)                                |
| 10  | MR XXXX         | சிங்கம் லியோ(Singam Leo)                                    |
| 11  | MISS XXXX       | சாமி சாலேகான்காஸ்(Samisalokangas)                           |
| 12  | MISS XXXX       | கயல்விழி பால்கேவா(Kayalvizhi Paalkova)                      |
| 13  | MISS XXXX       | விஷாலினி விஷ்வனாதன்(Vishaalini Vishwanathan)                |
| 14. | MRS XXXX        | ப்ரியாதர்ஷினி (Priyadarshini)                               |



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Virtual Environment as a Collaborative Platform to Enhance Pupils Information literacy skills

*Sivagouri Kaliamoorthy*  
*Beacon Primary School, Singapore*  
*sivagouri\_kaliamoorthy@moe.edu.sg*

## **Abstract**

The emergence of informative environment technology has made the ideology of learning, “anywhere, anytime”, a reality. The 21<sup>st</sup> century learners are equipped with readily available information at a mere click of a button. The Internet has erased international boundaries allowing our young charges the potential to develop as global citizens. This also demands that pupils develop a deeper understanding of the complex arrays of issues that involve them, now further complicated by the lack of traditional boundaries. Research studies have given good insights on the role of information literacy on the effectiveness of learning. However, very little studies demonstrate an effective implementation of programmes in virtual learning environments. The paper highlights how Beacon Primary School, one of the futuristic schools in Singapore, has implemented its Tamil Language programmes in a virtual learning environment thus providing a collaborative platform for pupils to meet and discuss issues. All our students, studying at the Primary four (P4) level (ten year old students), own their own personal learning device (notebook computer) which they bring to school daily. P4 Tamil curriculum and lesson packages are designed to infuse Information Communications Technology (ICT) meaningfully and make virtual learning a reality. Information literacy had been weaved into the P4 Tamil language curriculum with an online Web 2.0 software, wikispace, PBworkspace, as the platform for collaborative virtual learning environment. This paper presents how the virtual environment acts as a collaborative platform to enhance the pupil’s information literacy skills.

**Keywords:** Virtual Environment, Information Literacy Skills

## **1. Introduction & Purpose**

Pupils are surrounded by a wealth of knowledge. Today, at the click of a button, students have access to events occurring anywhere on the globe within seconds of it happening. Given this scenario, it is critical that our pupils are equipped with the skill to connect, construct and relate the information presented. The virtual environment provides the space for collaboration amongst pupils. The virtual environment eases and enriches the process out of which meaning is derived from the multitude of information presented. The virtual environment also presents a knowledge-based forum for pupils to build on each others’ contribution.

Today’s educational system has to respond to two seemingly contradictory demands: On one hand, it has to effectively transmit constantly evolving knowledge and know-how to a knowledge-driven civilization. On the other hand, it has to enable learners with the right skills to select pertinent information out of the explosion of available information. It also has to ensure that the personal and

social development of the young learner is catered for. Therefore 'education must ... simultaneously provide maps of a complex world in constant turmoil and the compass that will enable people to find their way in it' Delors. J.,(1996) This translates to a shift in focus for the amount and level of content taught in schools. It also calls for greater emphasis on equipping our pupils with relevant skills to pick out relevant information. This forms the basis of the nation-wide initiative of 'Teaching Less, Learning More'<sup>6</sup>. In today's context, the ability to access, evaluate, organize and use information in order to learn, problem-solve, make decisions in a formal and informal learning contexts are an integral part of their learning. A key characteristic of the lifelong learner is strongly connected with critical and reflective thinking.

Information communication technological tools are constructive tools that provide a collaborative platform for pupils to come on board and build on each other's knowledge. "Constructive tools are general-purpose tools that can be used for manipulating information, constructing one's own knowledge or visualizing one's understanding" Lim., C.P., & Tay, L.Y.,(2003). Jonassen, D. H., Carr, C. S., & Lajoie, S. P. (2000) purport the following constructivist approach- "ICT as mind tools for constructing evaluating, analysing, connecting, elaborating, synthesizing, imagining, designing, problem-solving, and decision-making." The term "constructive" stems from the fact that these tools enable students to produce a certain tangible product for a given instructional purpose. This paper takes a reflective, narrative approach in documenting our attempts to integrate the virtual environment as a collaborative platform in enhancing pupils' information literacy skills.

## **2. My Reflections**

One of the key themes in the P4 curriculum revolves around the topic of 'My Country'. The broad objectives include exposing students to the various issues that surround the country. The lesson design is tailored to educate on the various national issues, including the importance of tourism and consequently make logical connections to the implications and impact it poses to Singapore's economic growth. The lesson was planned and carried out via the virtual learning platform as a collaborative platform for pupils to virtually meet discuss and develop their knowledge on the issue.

The discussion began from an article on Tourism from the Singapore local Tamil newspaper, Tamil Murasu. The teacher posted questions adopting the Blooms Taxonomy to scaffold pupils skills up to the different stages. Relevant links for extended learning was also provided. These links however, was in the English language. Pupils were instructed to explore these links independently and gather pertinent information. They were subsequently asked to present them coherently in the Tamil language.

Pupils were taken through three main stages:

- 1) Connect - refers to the understanding of the article/ information presented.

---

<sup>6</sup> 'Teach Less; Learn More' (TLLM) is a call for schools and teachers to focus more on the active learning of students and the construction of their own knowledge.

- 2) Construct – refers to the pupils’ ability to comprehend the information, build on possible relationships and extend their knowledge and understanding from the information presented collectively in the platform.
- 3) Relate – relates to the presentation of collective information, analysis, synthesis, evaluation and creation of new perspectives from the issues presented. The following section details the activities conducted as part of each of the stages.

**1) Connect:**

- Pupils were asked to highlight the keywords and use the mind mapping technique to identify all the important points in the article.
- Each pupil is to contribute one finding from the article via online postings.
- Pupils also verified their friends’ understanding of the article and their related thoughts.
- If there was a misunderstanding of aspects in the article, the responsibility lay on fellow mates in the team to post a more accurate interpretation of the information.
- The teacher acts as a facilitator to ensure that pupils connect with their ideas.

**2) Construct**

- Pupils paraphrase, translate or give a short summary to express their comprehension of the article and the related issues.
- In response to the questions raised, other members in the class contribute and build on one another’s ideas via the platform.
- The pupils’ understanding of the content matter becomes apparent when they are able to identify relationships amongst ideas posted.
- Pupils also tap on prior knowledge to build on these ideas.

**3) Relate**

- Pupils are challenged with questions that require them to analyse available information and find logical patterns.
- Pupils then evaluate the information and relate it to the current situation and seek new perspectives and understanding.

Pupils were observed to be very engaged and used the language appropriately. However, there were instances where pupils used English language to express their ideas, instead of Tamil. Although pupils were strongly encouraged to use Tamil language, weaker pupils who needed to resort to code-switching to express their thoughts, were not discouraged. The other pupils in the subsequent postings helped to translate these ideas. This created a win-win situation for pupils to tap on and maximise each others’ strength and to learn collaboratively.

As part of the school ICT program, pupils were introduced to search engines and were guided in searching for the relevant information. Pupils were also taught principles of cyber-wellness and exercised civic respect in contributing ideas and in providing feedback and comments in the online platform. The contributions of students to the discussed topic and the postings of links leading to

other related information was motivating. Even students who were less proficient in the language, displayed interest in contributing to the discussion. Their posting displayed the collective understanding of the various points contributed in the platform. As all pupils had to work with their own personal learning, the learning was seamless.

The second extensive discussion took place after Japan's natural disaster. Pupils were exposed to this information during the morning assembly programme. As an extension an article from the newspaper was selected for online discussions. There was an intense discussion amongst pupils including the implications to the society and country. Pupils related the probable consequences. They were able to relate chain actions that would take place because of this disaster. Pupils used the Internet search engines to look up for latest update on the disaster such as on the British Broadcasting Corporation (BBC) news website. It was gratifying to note that the students took it upon themselves to update one another on the latest developments. In addition, they discussed and evaluated the situation and thought about the loss of those affected and the possible implications on their lives. It was heart warming to note pupils expressed concern and empathy for those affected.

### **3. Discussion & Conclusion**

Technology is used as a constructive tool to facilitate pupils' learning and making sense of their learning via a collaborative platform. Pupils' engagement was evident throughout the discussion. They were critical about their contributions and took great responsibility in actively using the net to search for information to enhance their learning. The project had benefited even pupils, less proficient in the Tamil language, who was observed to be actively contributing ideas. There was sincere commitment on the part of the students. They also showed initiative in providing additional links and support for others to make sense of the issue. This helped to bring out the best in each pupil. Pupils in addition, expressed positive feedback. Every pupil contributed and has equal share in collaboratively constructing the knowledge, thus the ownership was very strong amongst them. This was a demonstration that young age is not a barrier in understanding world issues if it is tailored to meet the needs of the young learners. What really matters is whether pupils are equipped with skill to understand the implication and impact of the issue discussed.

In terms of skills, all pupils were able to sieve out and decipher the main points from the information presented and build on this information. Through this communication, it was observed that pupils had tapped on prior knowledge and experience in developing their alternative perspectives. Pupils learned to use the information and ideas presented in a graphical organising format to organise ideas. Pupils exhibited strong bonding and collaboration during the various collaboration sessions. The usage of technology was pervasive and as Breivik., P. (2000) puts it "Information literacy (is not)... teaching a set of skills but rather a process that should transform both learning and the culture of communities for the better."

This paper is my attempt to share possible strategies in integrating information literacy into our daily lessons. It is through such sharing and exchanges where ideas could build upon ideas to further push the boundaries of our pursuit for pedagogical break-through in this fast changing world.

### **References**



- Burn., A. (2009). *Making New Media. Creative Productions and Digital Literacies*. New York: Peter Lang Publishing, Inc.
- Breivik., P.,(2000). *Foreword, Information Literacy Around the World*. Charles Sturt University
- Christine., B., (1997a). *The seven faces of information literacy*. Adelaide: Auslib Press.
- Delors, Jacques et al. (1996). *Learning: The Treasure Within*. Paris: UNESCO
- UNESCO. 2004. EFA Global Monitoring Report. Paris: UNESCO
- Lim., C.P., & Tay, L.Y.,(2003). *Information and Communication Technologies (ICT) in an Elementary School: Students' Engagement in Higher Order Thinking*. *Jl. Of Educational Multimedia and Hypermedia* (2003) **12**(4), 425-451
- Jonassen, D. H., Carr, C. S., & Lajoie, S. P. (2000). *Computers as cognitive tools*. Hillsdale, NJ: Lawlence Erlbaum Associates, Inc.
- Teach Less; Learn More- *Transforming Learning From Quantity To Quality*. Singapore Education Milestones 2004-2005 <http://www.moe.gov.sg/about/yearbooks/2005/pdf/teach-less-learn-more.pdf>
- Williams, M. D. (2000). *Integrating Technology into Teaching and Learning*. Singapore: Prentice Hall.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# இணையம் மற்றும் கணினி வழி தமிழ் கற்றல் கற்பித்தல்

நல்லாமூர் முனைவர் கோ. பெரியண்ணன்

இயக்குநர் , தமிழகக் கல்வி ஆராய்ச்சிவளர்ச்சி நிறுவனம், சென்னை

## முன்னுரை

அறிதொறு அறியாமை கண்டற்றால் என்பதற்கேற்ப இன்றைய கல்வியில் கணினித் தொழில் நுட்பம் நாளும் வளர்ந்து வருகிறது; மிகுந்த வரவேற்பைப் பெற்றுள்ளது. அனைத்துப் பாடப் பிரிவுகளிலும் இந்நுட்பம் பயன்படுகிறது. தமிழ் கற்றல் கற்பித்தலும் இந் நுட்பத்தின் பல்வேறு தளங்களைக் கண்டறிவது தமிழக் கல்வியாளர்களின் வேட்கையாகக் காணப்படுகிறது. கணினி நுட்பத்தோடு இணைந்த இணையவழி அணுகு முறையில் தமிழ் கற்றலுக்கும் கற்பித்தலுக்கும் பயன்படுத்தப் படவேண்டிய சூழல்கள் உருவாகி வருகின்றன. அவை நிறுவனம் சார் (Formal) நிறுவனம் சாரா (Non formal), இயல்பு நிலை (Informal) ஆகியனவற்றில் காணப்படுகின்றன. இவற்றிற்குக்கருத்தூட்டம் வழங்குவது தமிழக் கல்வி வல்லுநர்களோடு இணைந்த கணினி வல்லுநர்களின் கடமையாகும்.

## நிறுவனம் சார்ந்த கல்வி

நிறுவனங்கள் கற்பித்தல் முறைகளை வகுக்கவும் நடைமுறைப் படுத்தவும் பல அணுகு முறைகளை மேற்கொண்டு வருகின்றன. கணினி நுட்பத்தைப் பயன்படுத்தி மென்னுருக்களைத் (Software) தம் பாடப் பகுதிகளுக்கேற்ப ஆசிரியர்களே உருவாக்கி வருகின்றனர். உருவாக்கப்படும் மென்னுருக்களின் தன்மை அறிவியல் பாடங்களுக்கான மென்னுருக்களிலிருந்து தமிழ்ப்பாட மென்னுருக்கள் தன்மையாலும், தயாரிப்பு முறையாலும் வேறுபடுகின்றன. உயிரோட்டமுடைய தகவல்களைக் கொண்டு அமைவன தமிழ்ப் பாடங்கள். அதனைப் பவர் பாயிண்ட் (Power Point) முதலியவற்றில் வெளிப்படுத்தும்போது அவற்றுள் இயக்கமிகு காட்சிகளைப் பதிக்கவேண்டும். இவற்றிற்கான பயிற்சிகள் தமிழாசிரியர்களுக்கு - தமிழகத்திலுள்ள தமிழாசிரியர்களுக்குத் தேவைப் படுகின்றன.

## நிரல் வழிக்கற்றல் முறை:

வலைத்தளநுட்பமும் (Web Technology) தொடர் நிகழ்வுகளைக் கொண்ட மொழிப்பாடப் பகுதிகளைக் கற்பிப்பதற்கு மிகுதியும் பயன்படுவதாகக் காணப் படுகிறது. ஒரு வலைத்தளத்தில் தரப்படும் தகவல்களை விரித்துரைப்பதற்கும், ஆழ்ந்து கற்பதற்கும் வலைத்தள நுட்பம் சிறந்தது விளங்குகிறது. P.F. ஸ்கின்னர் வெளிப்படுத்திய நிரல்வழிக்கற்றல் (Programmed Learning) முறை, வலைத்தள நுட்பத்தால் திண்மையுறுகிறது. அக்கற்றல் முறையிலுள்ள கிளைவழித்திட்டத்தை (Branching) வலைத்தள நுட்பத்தால் நேர்த்தியாகப் பின்பற்றமுடியும். வலைத் தளத்தில் அமைக்க இயலுகின்ற ஒளிர் அல்லது மூன்றாம் தளப் பனுவல்கள் (Third dimension text) கிளைவழித் திட்டத்திற்கு மிகப் பொருத்தமானது.

## வலைத்தளமும் கற்பனை வளமும்:

தமிழ் இலக்கியங்கள் வலுவான கற்பனை வளத்தையும், நிறைவான பொருட்செறிவையும் கொண்டுள்ளன. எடுத்துக்காட்டுகள்

- 1 "காய்மாண்ட தெங்கின் பழம்வீழ கமுகின் நெற்றி..... எனத் தொடங்கும் சீவக சிந்தாமணிப் பாடல்"

2 "தண்டலை மயில்களாட..... தாமரை விளக்கம் தாங்க" எனத் தொடங்கும் கம்பராமாயணப் பாடல்

3 முல்லைப் பாட்டு - முழுமையான காட்சியமைப்பு இவற்றையெல்லாம் வலைத்தள நுட்பம் கொண்டு வகுப்பறையில் கற்பிக்க, கற்பித்தல் வளமுறும்.

கற்றல் வலுப்பெறும். நிறுவனம் சார் கல்வியில் இத்தகைய நுட்பங்களை மேற்கொள்ளவதற்கான நடவடிக்கைகளை 'உத்தமம்' போன்ற அமைப்புகளின் செயற்பாடாக வேண்டும்.

### **நிறுவனம் சாராக் கல்வி**

இன்றைய கல்வி, தொடக்கம், உயர்தொடக்கம், உயர்நிலை, மேனிலை, இளநிலைப்பட்டம், முதுநிலைப் பட்டம், ஆராய்ச்சி என 7 படி நிலைகளிலும் நிறுவனம் சாராது வளர்ந்து வருகிறது.

அச்சு ஊடகங்கள் சாதிக்கமுடியாத கற்றல் கற்பித்தல் முறைகளைக் கணினி இணைய நுட்பங்கள் சாதிக்கவியலும். ஆசுபெல் (Ausubul) பெஞ்சமின்புலும் (Benjamin Bloom) முதலானோர் எத்தகைய கடினமான கற்றல் பகுதியையும் கற்கவியலும் என்பதற்கு உளவியல் விளக்கங்களைத் தந்துள்ளனர்.

### **இலக்கியப் பனுவல்கள் (Literary text)**

மொழி இலக்கியப் பாடங்களில் மேற்குறிப்பிட்ட ஏழு படி நிலைகளுக்குரிய கற்றல் பொருள் அமைந்துள்ளது. படிநிலைக்கேற்றவாறு எளிமையிலிருந்து கடினம் என்னும் கோட்பாட்டின் அடிப்படையில் கற்றல் பனுவல்களை வழங்க இணைத்தள நுட்பம் வழிசெய்கிறது. வணிகத் துறையில் நிறுவனங்களுடைய வலைத்தளங்கள் அவற்றினுடைய பல்வேறு பிரிவுகளையும் செயல்பாடுகளையும் 'அவற்றின் முகப்புப் பக்கத்தில் குறித்துக் காட்டி மிகச் சிறப்பான முறையில் முழுவிவரங்களைத் தருகின்றன. அத்தகைய அணுகுமுறை தமிழ் இலக்கியங்களில் பொதிந்துள்ள

- 1 பின்னணித் தகவல் (Background information)
- 2 நேர்பொருள் (Direct meaning)
- 3 பொதி பொருள் (Implied meaning)
- 4 கலைக் கூறுகள் (Aesthetic features)
- 5 நுணுக்கங்கள் (Inferences)

முதலியனவற்றை வெவ்வேறு இணைப்புத் தளங்களில் அமைத்து எவரும் ஆசிரியர் உதவியின்றிக் கற்பதற்கேற்றவாறு வழிகாட்டவியலும்.

### **மொழி கற்றல்**

அரிச்சுவடி நிலையிலிருந்து ஆராய்ச்சிப் படிப்பு வரை மொழியைக் கற்பதற்கான தளங்கள் உள்ளன. அவற்றைக் கற்பதற்குரிய அகராதி கலைக்களஞ்சியம், பார்வைநூல்கள் முதலியனவற்றை இணைத்துப் பன்னோக் குடைய சொற்பொருள் அகராதியை வலைத்தள நுட்பத்தால் மட்டுமே உருவாக்கமுடியும். எடுத்துக்காட்டாக, 'வா' என்றும் வினையினை ஏழு படிநிலைகளிலும் கற்கத்தக்கப் பரிமாணங்கள் உள்ளன. இவற்றையெல்லாம் உள்ளடக்கிய பன்னோக்குப் பேரகராதியை மொழி இலக்கிய வல்லுநர்கள் கணினி வல்லுநர்களோடு இணைந்து உருவாக்கவியலும். இத்தகைய முயற்சி மொழிக்கல்வி வளர்ச்சிக்கு இன்றியமையாததாகும்.

### **இயல்புக்கல்வி:**

இன்றைய சமுதாயச் சூழலில் கணினி அறிவு இளைய தலைமுறை யினரிடத்துப் பெருகி வருகிறது. நாள்தோறும் கணினியின் முன்னமர்ந்து தகவல்களைத் திரட்டுதலில் இவர்கள் மிகுந்த ஆர்வம்

காட்டுகின்றனர். அவர்கள் ஆர்வத்திற்கேற்றவாறு மொழி இலக்கிய வலைத்தளங்களைக் கற்பித்தல் நுட்பங்களோடு உருவாக்குவது தமிழ்க் கல்வியாளர்களின் பொறுப்பாகும். இவர்களின் கல்வி நுட்பத்தைக் கணினி வல்லுநர்கள் வலைத்தள நுட்பத்துள் இணைக்க வேண்டும்.

### இலக்கியங்களில் பொதிந்துள்ள

"யாதும் ஊரே யாவரும் கேளிர்

தீதும் நன்றும் பிறர்தர வாரா"

"செல்வத்துப் பயனே ஈதல்

துய்ப்போம் எனினே தப்புந பலவே"

முதலிய பண்பாடுகளை அடையாளப் படுத்தல், அவற்றில் காணப்படும் விழுமங்களை வெளிப் படுத்துதல், இயற்கையோடியைந்த கற்பனை வளங்களை உணர்த்துதல், செறிவான சொற்பொருள் நயத்தை உள்ளூடச் செய்தல், தமிழுக்கே சொந்தமான சந்தநயத்தையும் இசைப் பெருக்கினையும் அறியச் செய்தல், எண்வகை மெய்ப்பாடுகளையும் காட்சிப் படுத்தல், மெய்ப்பாடுகளுக்கேற்ப முத்தமிழ் வளத்தில் மூழ்கச் செய்தல், என்பனவற்றிற்கெல்லாம் வலைத்தளமே பொருத்தமானதாக அமையும்.

### முடிவுரை

தமிழ் கற்பித்தலுக்கும் கற்றலுக்கும் இயல்பாகவே கணினி நுட்பங்கள் பொருத்தமாக விளங்குகின்றன. நிறுவனம்சார், நிறுவனம் சாரா, இயல்புநிலை ஆகிய மூவகைக் கல்விக்கும் அரிச்சுவடி முதல் ஆராய்ச்சி படிப்பு வரையிலான ஏழு படிநிலைகளுக்கும் கணினி சார்ந்த இணைய நுட்பத்தைக் கடைப்பிடித்து, தமிழ் கற்றல் கற்பித்தலைச் செம்மையாக்கவும் மேம்படுத்தவும் வழிகள் உள்ளன. தமிழ்க் கல்வியாளர்கள், கணினி வல்லுநர்கள், ஆசிரியர்கள் ஆகியோரின் இணைந்த செயல்பாடுகள் இன்றையத் தேவைகளாகக் கருதப்படுகின்றன.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011



# இணையம் வழித் தேர்வுகளில்லாக் கல்வி

## (Education without Examination)

### Through e-learning

முனைவர் ப. அர. நக்கீரன்  
இயக்குநர், தமிழ் இணையக் கல்விக்கழகம்

#### முன்னுரை

கல்வி என்பது அனுபவங்களைப் படித்து அறிவுபெற்று ஆக்கபூர்வ தனிமனித, சமுதாய வளர்ச்சிகளுக்குப் பயன்படுத்துவதாகும். இன்று கல்வி கற்றல் - கற்பித்தல் அமைப்பில் ஆசிரியர், மாணவர், கல்வித்திட்டம், மதிப்பீடு என்ற கூறுகள் அடங்கியுள்ளன. கற்றுப் பெற்ற அறிவைச் சோதித்தறியத் தேர்வுகள் பயன்படுகின்றன.

கற்பதை ஊக்கப்படுத்தவும், கற்ற அறிவை மதிப்பீடு செய்து தரம்பிரிக்கவும் உருவாக்கப்பட்ட தேர்வு முறைகள், 'மதிப்பெண்' என்ற வட்டத்துக்குள் முடங்கிக் கிடக்கின்றன. அறிவு என்ற நோக்கம் என்பதற்கு மாறாக மதிப்பெண் பெறுவதே நோக்கம் என்ற நிலை மாறிப் போய்விட்ட காரணத்தால், புரிந்து படித்தல் என்பதற்கு மாறாக, மனப்பாடம் செய்து ஒப்புவித்து மறந்துபோதல் என்ற நிலை இன்று உருவாகியிருக்கிறது.

இதனால் ஆக்கபூர்வ சிந்தனையைத் தூண்டி அறிவியல் அறிஞர்களையும், தொழில்நுட்ப, மருத்துவ வல்லுநர்களையும் உருவாக்குவதற்கு மாறாக அலுவலகக் கூலிகளை இந்தக் கல்வி உருவாக்கிக் கொண்டிருக்கிறது.

இந்நிலை மாறவேண்டுமானால் இன்றைய கல்வி முறைகளில் புதிய சிந்தனை வரவேண்டும். ஆண்டாண்டுகளாக தொடர்ந்து கொண்டிருக்கும் கற்றல் - கற்பித்தல் முறைகள் மாறவேண்டும்.

இதற்காக முன்மொழியப்படும் புதிய அணுகுமுறை தான் 'பள்ளியில்லாக் கல்வி - தேர்வுகளில்லா தேர்ச்சி' என்பது.

இன்றைய கல்வித்திட்டத்தில் உள்ள நிறைகளை ஆராய்ந்து, வளர்ந்து வரும் கணிப்பொறிக் காலத்திற்கேற்ப, இணையம் வழியாக எளிதாக அறிவு பெறும் மாணவர்களை உருவாக்கும் புதிய வழிகளை எடுத்துக் கூறுவதே இக்கட்டுரையின் நோக்கம்.

#### இன்றைய கல்வி முறை

இன்றைய கல்வி முறையில் பெற்றோர், மாணவர், பள்ளிக்கூடம், ஆசிரியர், கல்வித் திட்டம், வகுப்பறை, கற்பித்தல் - கற்றல், தேர்வு - தேர்ச்சி என்ற பயணம் வகுப்பு மாறித் தொடர்ந்து சென்று கொண்டிருக்கிறது. உயர் வகுப்பு என்பது உயர் அறிவைக் குறிக்கும்.

நடத்தப்படும் பாடங்களில் கேள்விகள் கேட்டு, அதற்குச் சொல்லப்படும் விடைகளைப் பொருத்து மதிப்பெண் தரப்படும். அதன் அடிப்படையில் தேர்ச்சி கணிக்கப்படும்.

ஆண்டாண்டுக் காலமாக நடைபெறும் இந்த மதிப்பீட்டு முறையில் கேள்விகள் ஏறக்குறைய நிலைத்தன்மை பெற்றுவிட்டன. கேள்விகளின் தோரணை சற்று மாறுபடலாம், அவ்வளவுதான்.

எனவே ஒரு குறிப்பிட்ட, கேள்விகளுக்கு மட்டும் படித்து மனப்பாடம் செய்து விட்டால் போதும்; எளிதில் தேர்ச்சி பெற்று விடலாம். மனப்பாடம் செய்வதற்குப் புரிந்து கொள்ள வேண்டும் என்ற தேவையில்லை; நினைவுத் திறம் கூட அவ்வளவாகத் தேவையில்லை. தேர்வுக் கூடத்தில் விடை எழுதும் வரை நினைவு இருந்தால் போதும்.

இப்படிப்பட்ட கல்வி முறையால் சிந்தனை வளர வாய்ப்பில்லை. சொன்னதைச் சொல்லும் கிளிப்பிள்ளைகளைத் தான் உருவாக்க முடியும்.

தமிழ் நாட்டின் தலைநகரம் சென்னை என்று சொல்லிக் கொடுத்து விட்டால்,

தமிழ் நாட்டின் தலைநகரம் எது? என்றுதான் கேள்வி கேட்க வேண்டும்.

சென்னை எந்த மாநிலத்தின் தலைநகரம் என்று கேள்வி கேட்டால், விடை தெரியாது. அந்தக் கேள்வியே பாடத்திட்டத்திற்கு அப்பாற்பட்டது என்று கூறப்பட்டு, சொல்லாத விடைக்கும் மதிப்பெண் கொடுக்கப்படும்.

இப்படித் தேர்ச்சி பெற்று வருபவர்களால் இந்தச் சமுதாயத்திற்கு என்ன பயன் விளையும்? பட்டம் பெற்ற பாமரர்களைத் தான் உருவாக்க முடியும்.

இன்றைய தேர்ச்சி முறையை சோதனைத் தரம் (Check in Quality) என்று கூறுவர். சோதனையில் தேர்ச்சி பெறவில்லை என்றால் அதனால் ஒரு ஆண்டு வீணாகி விடும்.

ஒரு காலத்தில் தொழிற் சாலைகளில் இந்த முறை தான் பயன்பாட்டில் இருந்தது. ஒரு பொருள் உருவான பின்னர், சோதனை செய்து சரியாக இருந்தால் ஏற்றுக் கொள்வர். குறையாக இருந்தால் நீக்கி விடுவர்.

### **கல்வியில் கட்டமைத் தரம்: (BUILD IN QUALITY IN EDUCATION)**

இன்று இந்த முறை மாறியிருக்கிறது. அதற்குக் கட்டமைத் தரம் (Build-in-Quality) என்று பெயர். ஒரு பொருள் உருவாவதற்குக் காரணமான வடிவமைப்பு, உலோகம், பணியாளர், உற்பத்தி என்று எல்லாக் கூறுகளையும் தரமானதாக அமைத்தால், அதிலிருந்து வரும் பொருள் தரமானதாக இருக்கும் என்பதே இதன் அடிப்படை. முழுத்தர மேலாண்மை (Total Quality Management) என்பதின் அடிநாதம் இது.

இந்த அடிப்படையில் ஒரு கல்வித் திட்டத்தை உருவாக்க முடியுமா? முடியும். அதற்குத் தேவை:

1. நல்ல ஆசிரியர்கள்
2. நல்ல கல்வித்திட்டம்
3. நல்ல கற்றல் - கற்பித்தல் சூழல்
4. நல்ல நிர்வாகம்
5. நல்ல தேர்ச்சி முறை

கல்வி என்றால் கொடுப்பது அல்ல; தோண்டி எடுப்பது. ஆசிரியர் என்பவர்கள் குற்றம் குறைகளை நீக்குபவர்கள் (ஆசு=குற்றம்) எனவே அவர்கள் சிந்தனைத் தூண்டிகளாக (Knowledge facilitator) இருக்க வேண்டும்.

தாம் பெற்ற அறிவுசார் அனுபவங்களை மாணவர்களுக்குக் கொண்டு சேர்க்க வேண்டும். அதற்கு அவர்களும் மாணவர்களாகத் தொடர வேண்டும். அறிவுக் கடலைத் தேடி அலைய வேண்டும். அந்தத் திசையை மாணவர்களுக்குக் காட்டவேண்டும்.



ஆனால் மாணவர்கள் தான் அத்திசையில் முயன்று ஓடி முன்னேறவேண்டும். ஞானம் என்பது கடைசியில்தான் வரும். இந்த மாணவர் முயற்சிகளுக்கு ஏற்ப ஆசிரியர்களும் திட்டமும், நிர்வாகமும், மதிப்பீட்டு முறைகளும் அமைய வேண்டும்.

இன்றைய கல்விமுறையில் ஆசிரியர் கேள்வி கேட்கிறார்- மாணவர்கள் விடை கூறுகிறார்கள்.

புதிய முறையில்,

மாணவர்கள் கேட்கும் கேள்விகளுக்கு ஆசிரியர்கள் விடை கூற வேண்டும்.

கேள்வி கேட்பவன் சிந்திக்கத் தொடங்கி விட்டான் என்று பொருள்.

எனவே மாணவர்களைக் கேள்வி கேட்கத் தூண்டுவதே ஆசிரியர்களின் கடமையாகும்.

வழிதேடும் விழிகளுக்கு வெளிச்சமாய் ஆசிரியர்கள் மாறவேண்டும்.

300 நாட்கள் கற்றதை மூன்று மணிகளில் மதிப்பீடு செய்வது என்ற முறை மாற வேண்டும். 300 நாட்களும் தேர்வுகளாக இருக்கவேண்டும்.

தரவுகள் சேர்ந்தால் தகவல்

தகவல்கள் சேர்ந்தால் அறிவு

அறிவு கூடினால் ஞானம்

ஞானத்தை அடையும் தவமாகக் கல்வி இருக்க வேண்டும்.

## இணையம் வழிக்கல்வி

சமையல் செய்வது எப்படி என்று கூறும் புத்தகங்களைப் போலத் தான் இன்றைய கல்வி இருக்கிறது. புத்தகங்களை வைத்துக் கொண்டு சமைக்க முடியாது? சமைத்தாலும் சாப்பிட முடியாது.

தோசை சுடுவது எப்படி? என்றால் விடை தெரியும் - ஆனால் தோசை சுடு என்றால்?

ஊசி செய்யும் சிறு தொழிலின் நுட்பத்தை கூறுவதற்கு மாறாக, ஒரு ஊசி செய்ய வேண்டும்.

எனவே அறிவும் அனுபவமும் சேர்ந்த கல்வித்திட்டத்தை உருவாக்க வேண்டியது இன்றைய இன்றியமையாத தேவை. இந்தத் தேவையை நிறைவேற்ற வந்திருக்கும் அட்சய பாத்திரம் தான் கணிப்பொறி. கணிப்பொறி ஊடான இணையவழிக் கல்வி.

இணையக் கல்வி முறையில் பாடங்கள் அனைத்தும் பல்லாடகத் தொழில்நுட்பக் கூறுகளுடன், படக்காட்சிகள், பேச்சு, இசை, செய்முறை ஆகியவற்றோடு இணையம் மூலமாகவே ஒரே ஆசிரியர் உலகில் உள்ள எல்லா மாணவர்களுக்கும் பாடம் நடத்துவார்.

ஆசிரியரைத் தேடி மாணவர்கள் போனது ஒரு காலம். ஆனால் இன்று மாணவர்களைத் தேடி அவர்கள் வீடுகளுக்கே ஆசிரியர்கள் போகிறார்கள்.

வீடுகளே வகுப்பறை, விரல் நுனியில் அறிவுப் புதையல்

கணிப்பொறியில் மட்டுமல்லாது செல்பேசிகளிலும் இந்தப் பாடங்களைக் கேட்கலாம்; படிக்கலாம்; பார்க்கலாம் - பள்ளிக்குப் போகாமலே.

ஆனால் இணைய வழிக் கல்வியிலும் மதிப்பீடு என்பது எழுத்துத் தேர்வு என்பதாகவே அமைந்திருக்கிறது.

## தேர்வுகள் இல்லாத தேர்ச்சி:

ஒரு ஊரில் உள்ள மாணவர்களை ஒன்று கூட்டி ஒரு இடத்தில் உட்காரவைத்து, ஒரு ஆசிரியர் மேற்பார்வையில் தேர்வுகள் நடத்தி விடலாம். உலகில் உள்ள மாணவர்களை எங்கே உட்கார வைப்பது? யார் மேற்பார்வையிடுவது?

இதற்கு ஒரே வழி தேர்வுகள் இல்லாத கல்விதான்? அப்படியென்றால் மதிப்பீடு செய்வது எப்படி? தேர்ச்சி முடிவுகள் தருவது எங்ஙனம்?

ஒரு பாடத்தில் ஒரு மாணவன் என்ன கற்கவேண்டுமோ, அதற்கேற்ப ஒரு திட்டப்பணியை ஆண்டுத் தொடக்கத்திலேயே வழங்கிவிடலாம். அத்திட்டப் பணியை ஏற்றுச் செயல்படுத்தும்போது ஏற்படும் கேள்விகளுக்கு ஆசிரியர்கள் விளக்கம் சொல்லலாம். தேவை என்பதால் மாணவர்கள் கூர்ந்து கேட்பார்கள்; புரிந்து கொள்ள முயற்சிப்பார்கள்.

இதற்குத் தேவைப்படும் ஆசிரியர்கள் இருக்கிறார்களா? பாட நூல்கள் உள்ளனவா?

இணையத்தில் ஆசிரியர்கள் இருக்கிறார்கள்! அவர்களின் மின்முகவரி தெரிந்தால்போதும். எளிதில் தொடர்பு கொள்ளலாம். மாணவரின் கேள்விகளை ஒரு வலைப்பூவில் பதியும் போது பல்நோக்குப் பார்வையில் விடைகள் கிடைக்கும்.

ஒரு பணித்திட்டத்தை ஒரு மாணவன் முடித்து இணையதளத்தில் இட்டு அதைப் பற்றிய கருத்துகளைக் (feed back) கேட்கலாம். ஒரு பணித்திட்டச் செயலாக்கமே அறிவு தரும் என்பதால், இதற்கான தனியான மதிப்பீடு தேவையில்லை. மின்னூட்ட கருத்துகளை வேண்டுமானால் மதிப்பீட்டுக் காரணியாக வைத்துக் கொள்ளலாம்.

இதற்கு ஏராளமான பணித்திட்டங்கள் தேவைப்படுமே? எங்கே போவது? ஏழு சுரங்களை வைத்துக் கொண்டு ஏராளமான இசைகளை உருவாக்குவதுபோல், சற்று மாறுதல் செய்து ஏராளமான பணித்திட்டங்களை உருவாக்கி விடலாம். இணையம் என்ற அறிவுக் கருவூலம் இதற்கு துணை செய்யும். இதனால் ஏன் என்று கேள்வி கேட்டு அறிவுபெறும் புதிய வழியைக் காட்டி, ஆக்க பூர்வச் சிந்தனைத் திறமுள்ள மாணவர்களை இம்முறை மூலம் உருவாக்கலாம்.

## முடிவுரை

மாற்றம் ஒன்றே மாறாதது. மாற்றமில்லாமல் வளர்ச்சி இல்லை. எதிர்ப்பு இல்லாத மாற்றமும் இல்லை. எனவே கல்வி முறையில் தேவைப்படும் ஒரு மாற்றத்திற்கான திட்டம் இங்கே கொடுக்கப்பட்டிருக்கிறது. இது தொடக்கம் தான் இதைப் பற்றிய விரிவான விவாதம் இனித் தொடங்க வேண்டும்; பயனுள்ள முடிவை எட்ட வேண்டும்.





## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Tamil Video Retrieval Based on Categorization in Cloud

**V.Akila, Dr.T.Mala**

*Department of Information Science and Technology,*

*College of Engineering, Guindy,*

*Anna University, Chennai*

*veeakila@gmail.com, malanehru@annauniv.edu*

## Abstract

Tamil Video retrieval based on categorization in cloud has become a challenging and important issue. Video contains several types of visual information which are difficult to extract in common information retrieval process. Tamil Video retrieval for query clip is a high computation task because of the computation complexity and large amount of data. With cloud computing infrastructure, video retrieval process has some scope and is flexible for deployment. The proposed method categorize the Tamil video into subcategories, splits the video into a sequence of shots and extracts a small number of representative frames from each shot and subsequently calculates frame descriptors depending on the edge and color features. The color histogram is computed for all the key frames based on hue, saturation and intensity values. Edge features are extracted using canny edge detector algorithm. The features extracted are stored in feature library in cloud. The features are tagged with Tamil text in cloud in order to satisfy Tamil query clip. Also, Videos are retrieved based on the Tamil audio information. The EUCALYPTUS cloud computing environment is setup within academic settings and the similarity matching of the Tamil video query is performed. The similar videos are displayed based on the similarity value and the performance is evaluated. Eucalyptus cloud platform is setup in Linux OS and the Tamil video retrieval process is deployed within the cloud. The efficiency of cloud computing technology improves the Tamil video retrieval process and increases the performance.

Keywords – video retrieval, categorization, cloud computing, Tamil query, Eucalyptus

## 1. Introduction

The need for intelligent processing and analysis of multimedia information has been increasing on a regular basis.

Researchers have found numerous technologies for intelligent video management which includes the shot transition detection, key frame extraction, video retrieval and more. Content based retrieval is considered to be the most difficult and significant issue of practical value amongst all the others. It assists the users in the retrieval of favored video segments from a vast video database efficiently based on the video contents. This paper aims at presenting the process of Tamil video retrieval in cloud environment. Video contains both visual and audio information. Audio contains natural language information which can be used to retrieve similar video content. The Tamil text processing is performed for user Tamil query.

The video retrieval system can be divided into two principal constituents: a module for the extraction of representative characteristics from video segments and defining a retrieval process to find similar

video clips from video database. A large number of approaches use a wide variety of features to symbolize a video sequence of which color histogram; shape information and text analysis are a renowned few. Application that requires a large number of computational resources might have to contact several different resource providers in order to satisfy its requirements. Cloud computing systems provide a wide variety of interfaces ranging from the ability to dynamically provision entire virtual machines. The feature database is stored in the cloud and the users query is compared. As based on cloud computing infrastructure, video retrieval process can be easily extended.

The rest of this paper is organized as follows: Section 2 deals with literature survey in the domain related to the project. It gives the different techniques adopted in the domain. Section 3 deals with system architecture. It includes detailed design of various phases involved in the project. It describes the internal working of the system. Section 4 deals with simulation and results of video retrieval process in cloud for Tamil videos. Section 5 deals with performance evaluation and result analysis. Section 6 focuses on conclusion and future enhancement.

## **2. Related Works**

Nurmi describes the basic principles of the EUCALYPTUS design, that allow cloud to be portable, modular and simple to use on infrastructure commonly found within academic settings [3]. EUCALYPTUS is an open source software framework for cloud computing that implements what is commonly referred to as Infrastructure as a Service. It allow users the ability to run and control entire virtual machine instances deployed across a variety physical resources.

Takagi explains a method for video categorization based on the camera motion[5]. Camera motion parameters in the video sequence contain very significant information for categorization of video, because in most of the video, camera motions are closely related to the actions taken. Camera motion parameters can be extracted from video sequence by analyzing motion information. Camera motion parameter has many advantages for categorization of video. Camera motion parameters like pan, fix are obtained using motion vector. Motion vectors are classified into 8 directions and histogram is calculated in each category. By analyzing characteristics of this histogram, camera motion parameters are extracted for each video [2].

The video shot segmentation system uses mathematical characterization of cuts and dissolves in the video [1]. Different kinds of transitions may occur. An abrupt transition is found in a couple of frames, when stopping and restarting the video camera. A gradual transition is obtained based on effects, such as fade in i.e. a gradual increase (decrease) in brightness or dissolves i.e. a gradual superimposition of two consecutive shots. Abrupt transitions are obtained for two uncorrelated successive frames. In gradual transitions, the difference between consecutive frames is reduced. Considerable work has been reported on the detection of abrupt transitions

A method for key frame extraction [6] which dynamically decides the number of key frames depending on the complexity of video shots and requires less computation. Priya and Shanmugam describe a method for feature extraction which provides the steps for extracting low level features [4]. The spatial distribution of edges is captured by the edge histogram with the help of sobel operators. Color histogram is the most extensively used method because of its usage in various fields. The color histogram value are recognized using hsv color space. Texture analysis algorithms are used in random

field models to multi resolution filtering techniques such as the wavelet transform. Several factors influence the utilization of Gabor filters for extracting textured image features. The feature library stores the extracted features.

### 3. System Overview

The architecture of our proposed system is shown in Fig 1. In the offline process, set of videos are given as input and features are extracted from the video. In the online process, the features are extracted from the query clip and matched against the feature vectors stored.

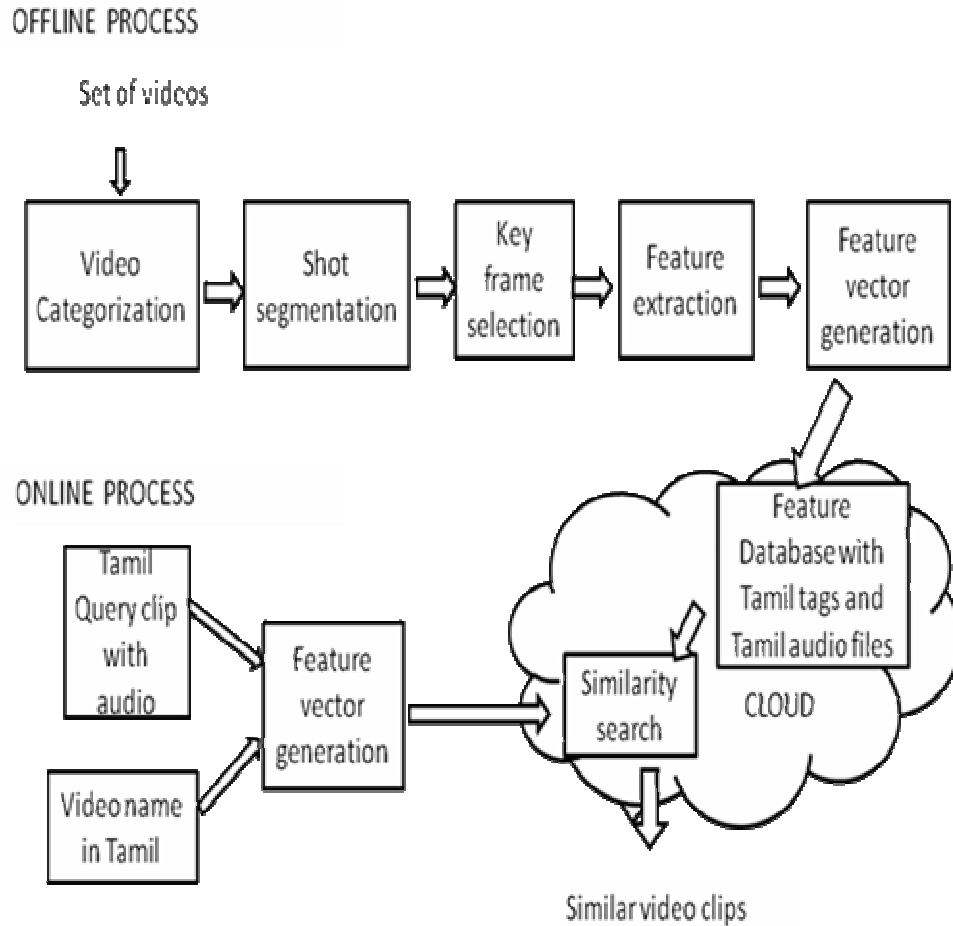


Fig 1 System Architecture

#### A. Video Categorization

The first process to be carried out is video categorization which is shown in Fig 2. The content based video categorizing method uses camera motion parameters. This parameter helps to categorize the sports videos for identifying different sports types. Camera motion parameters are changing the state among 2 types (Fix and Pan) along with time scale in video sequence. Here, motion vector are classified and histogram is calculated. By analyzing the characteristics of this histogram, camera motion parameters are extracted for each MPEG video.

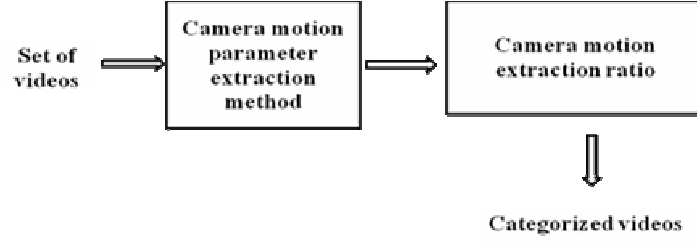


Fig 2 Video Categorization Process

In a video, panning is the sweeping movement of a camera across a scene and Fix means the static position of the camera. For this parameter, camera motion extraction ratio is calculated.

$$\text{camera motion extraction ratio } w[x]$$

$$w[x] = (\text{Num.appear} / \text{Num.total}) * 100\%$$

$$x = \{\text{FIX, PAN}\}$$

where,

Num.appear -> number of times of an appearance for camera work x.

Num.total -> total number of frames in the given video.

### B. Shot Segmentation

To segment the shots, the video has to be split into video shots prior of conducting any video object analysis. Scene change detection, either abrupt scene changes or transitional (e.g. dissolve, fade in/out, wipe) is employed to achieve the video shot separation.

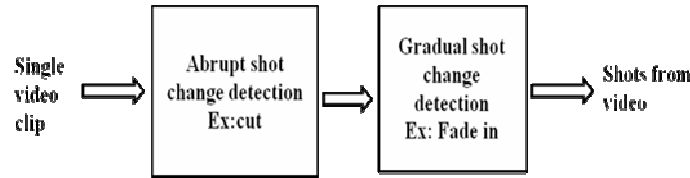


Fig 3 Shot segmentation Process

The proposed algorithm is based on the computation of an arbitrary similarity measure between consecutive frames of a video. The first phase of the algorithm detects the abrupt shot-change detection and second phase detects the gradual transition.

### C. Key frame Extraction

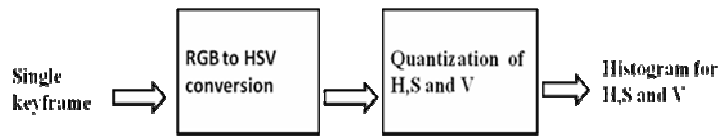
A key frame is a frame that represents the content of a shot. This content is the most representative one. In the large amount of video data, first reduce each video to a set of shots and find the representative frames. Each shot obtained by video segmentation algorithm contains a set of frames.



These segments are represented by two dimensional representative images called key frames that greatly reduce amount of data that is searched. Key frames from each shot are obtained by comparing the color information between adjacent frames. A frame will be chosen as key frame if the value exceeds certain threshold.

#### *D. Feature Extraction*

Feature extraction is an area of image processing which involves using algorithms to detect and isolate various desired portions of a digitized image or video stream. Different kinds of video features, including edge and color for each key frame is being extracted. To minimize the dimensionality of the data, feature extraction is employed which extracts discriminative features of data.



*Fig 4 Color Histogram Process*

Fig 4 shows the process of color histogram creation. Color histogram is the most extensively used method because of its robustness to changes due to scaling, orientation, perspective, and occlusion of images, which are recognized by using the HSV color space.

Edges in the key frames are detected based on the canny edge detector. The Canny operator works in a multi-stage process. First of all the image is smoothed by Gaussian convolution. Then a simple 2-D first derivative operator is applied to the smoothed image to highlight regions of the image with high first spatial derivatives. Edges give rise to ridges in the gradient magnitude image.

#### *E. Similarity matching*

The query video is categorized and key frames are extracted. The color and edge features extracted are matched against the features in the repository. The color features are matched based on the naive similarity algorithm and edge features are matched based on region based histogram.

The algorithm first calculates the color histogram for the query clip and compares with the video set. Each key frame feature vector of query clip is matched with all the feature vectors in the repository and most similar match is retrieved. The histogram values contain mean, entropy and standard deviation of color. From the mostly matched key frames the edge histogram is calculated and matched against query clip. The edge histogram contains region information. The key frames which give the most similarity values are selected and the corresponding videos are retrieved as the similar videos for user query clip.

#### *F. Audio Processing*

The next way of Tamil video retrieval focuses on audio processing. The audio track is extracted from the Tamil video as the first step. The audio files are segmented in order to remove the silence and noise. The audio files of each video are processed and the words are extracted and stored as .wav files.

These .wav files are called as features of the audio content.

The user gives the query Tamil video clip as input. This input file contains both audio and video information. The audio data will be segmented to remove silence and extract key words. These key word files are pattern matched against all the .wav files in the feature set. The most matched patterns are found and the corresponding videos are extracted.

The pattern matching of wav files are performed and the results which exceed certain threshold are taken as the result.

#### *F. Text Processing*

The next way of video retrieval is based on Tamil text. The wav files of audio input are chosen and are tagged with Tamil text. The user input of Tamil text is transliterated and is searched against the feature set. The matched results corresponding video are retrieved and given as result to user. Transliteration is the practice of converting a text from one language into another language phonetically. Transliteration is different than translation. The Table 1 shows some transliterated English word for tamil word.

| Tamil word | Transliterated English word |
|------------|-----------------------------|
| கடினம்     | Kadinam                     |
| பூக்கள்    | Pookkal                     |
| குழந்தை    | Kuzhandhai                  |
| பாப்பா     | Paappa                      |
| மழை        | mazhai                      |

*Table 1: Transliteration of Tamil to English*

#### *F. Cloud setup*

Eucalyptus is an open source cloud computing system.

The eucalyptus open source cloud environment is setup in Linux cluster. The eucalyptus software is installed. The cloud controller, cluster controller, walrus and storage controller are installed.

The cloud controller is the entry point into the cloud for users and administrators. It asks node managers for information about resources, makes scheduling decisions and implements them after requesting to cluster controller.

The cluster controller executes on a cluster front end machine, or any machine that can communicate to both the nodes running Node controllers and to the machine running cloud controller. Cluster

controllers gather information about and schedules virtual machine execution on specific node controller and also manages virtual instance network.

The Node controller is executed on every node that is used for hosting virtual machine instances. They control the execution, deployment and termination of virtual machine instances on the host where it runs.

The storage controller is capable of interfacing with various storage systems. It is a block device and it is attached to an instance file system. Walrus allows user to persistent data, organized as buckets and objects. It provides a unique mechanism for storing and accessing virtual machine images and user data.

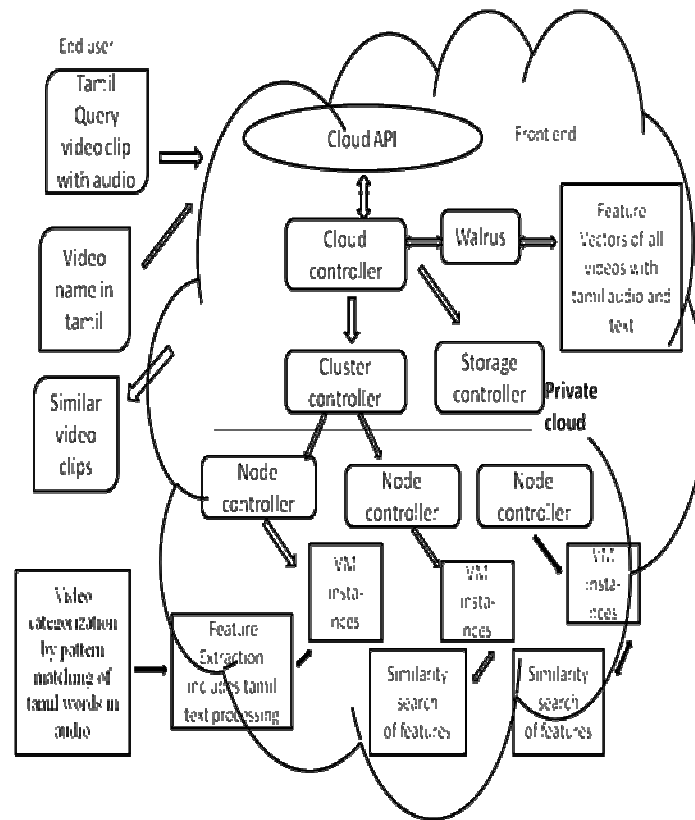


Fig 5 Video Retrieval process in Eucalyptus cloud

The Tamil video retrieval process is developed as application and this application is bundled to the virtual machine instance. The application bundled virtual machine image is uploaded and registered to the eucalyptus cloud. The instances are communicated and the application is run over the cloud. The query video clip is given as input in the cloud front end. The videos are categorized, the key frames are extracted and the similarity search is performed in separate parallel instances. The retrieved video result are given as output to the user.

#### 4. Simulation and Results

The video retrieval process includes video categorization, key frame extraction, feature extraction and similarity matching. The process is carried out in Java media framework and Java advanced imaging.

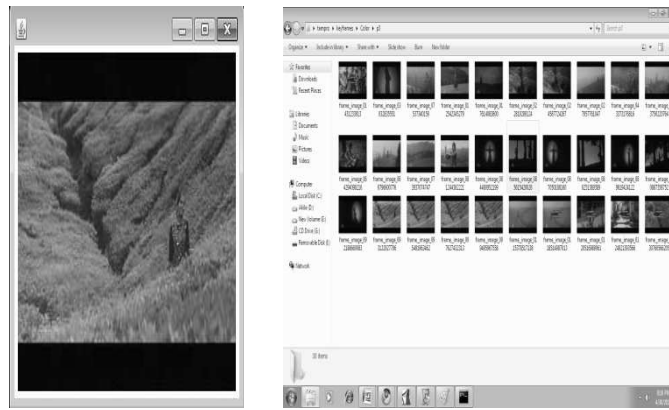


Fig 6 Tamil Quer Video and Key frame extracted from videos

Fig 6 shows the key frame extracted for a given tamil video and Fig 7 shows the categorization and similar video result



Fig 7 Similarity result of query video

The user gives the query video name as input and based on the commands the videos will be categorized, extracts key frames and features. The similar video will be retrieved if they give search command

## 5. Performance Evaluation

The video retrieval process is performed in cloud and the performance is evaluated while running in two instance.

| No. of videos in dataset | Execution time in 2 instances | Execution time in 1 instance |
|--------------------------|-------------------------------|------------------------------|
| 10                       | 5206.2                        | 18369.69                     |
| 15                       | 5522.63                       | 18924.41                     |
| 20                       | 6202.2                        | 21731.45                     |
| 25                       | 7291.7                        | 25319.52                     |
| 30                       | 8575.6                        | 28904.35                     |

Table 2:Relation between execution time in one instance and two instance

The application is run in EUCALYPTUS private cloud and the execution time is calculated while running in single instance and two instances. The execution time is much less when we run in two instances. This shows that the video retrieval process shows better performance in cloud.

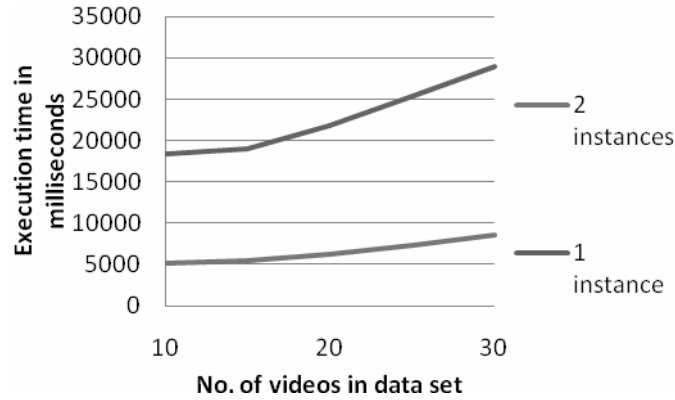


Fig 8 Performance graph in cloud environment

The performance of video retrieval process is checked by precision recall graph.

$$\text{Recall} = \text{DC}/\text{DB} \text{ and } \text{Precision} = \text{DC}/\text{DT}$$

Where DC is the number of similar clips which are detected correctly, DB is the number of similar clips in the database and DT is the total number of detected clips.

| Query video | Recall | Precision |
|-------------|--------|-----------|
| Q1          | 0.1    | 0.9       |
| Q2          | 0.35   | 0.78      |
| Q3          | 0.39   | 0.69      |
| Q4          | 0.6    | 0.4       |
| Q5          | 0.8    | 0.2       |

Table 3:Precision and recall for query video clips

The precision and recall for various query video clips are computed. The efficiency of the video retrieval process is improved as the retrieval process includes categorization process.

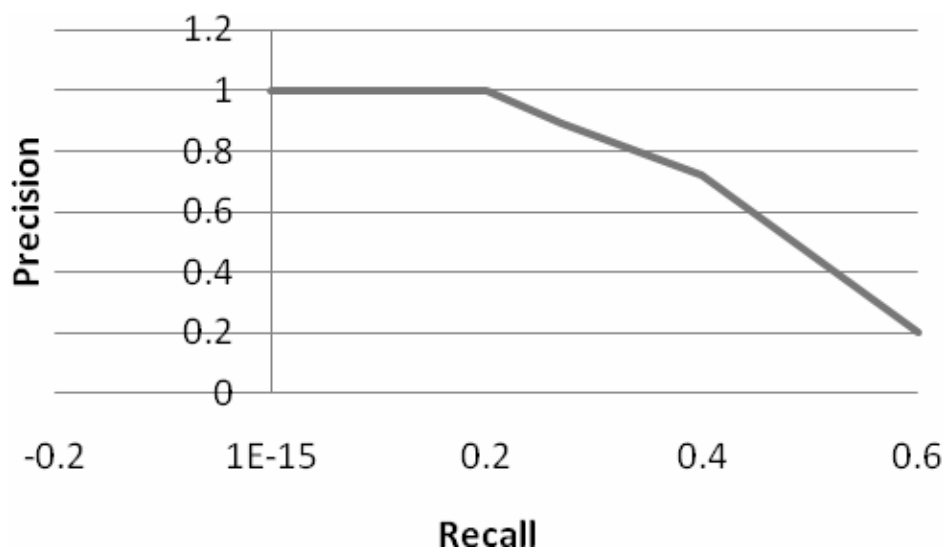


Fig 8 Performance graph for video retrieval

The performance of Tamil video retrieval shows that the most similar videos are retrieved. Also the application in cloud environment shows that the cloud computing provides better performance through execution time and resource sharing.

## 6. Conclusion and Future work

The proposed video retrieval categorizes the video into different category based on camera motion parameters. It facilitates the segmentation of the elementary shots in the video sequence proficiently. Then the key frames are extracted from the video shots. Subsequently, the extraction of features like edge and color histogram of the video sequence is performed and the feature library is employed for storage purposes.

Then Video retrieval system based on query video clip is incorporated within the cloud. Cloud computing, due to its high performance and flexibility, is under high attention around the industry and research and reduces the computation complexity of Video retrieval process based on visual, audio and text input.

## References

- Albanse M., Chianese A., Moscato V. and Sansone L., "A Formal Model for Video Shot Segmentation and its Application via Animate Vision", In Proceedings of Multimedia Tools and Applications, Vol 24, 2004, pp. 253-272.
- Dobashi K., Kodate A. and Tominaga H., "Camera Working Parameter Extraction for Constructing Video Considering Camera Shake", In Proceedings of International Conference on Image Processing (ICIP), Vol.III, 2001, pp.382-385.

- Nurmi D., Zagorodno D., Youseff L. and Soman S., “ *The Eucalyptus Open source Cloud-computing System*”, In Proceedings of International Symposium on Cluster Computing and the Grid, 2009.
- Priya R. and Shanmugam T.N., “*Enhanced content-based video retrieval system based on query clip*”, In Proceedings of International Journal of Research and Reviews in Applied Sciences ,Vol 1, 2009.
- Takagi S., Hattori S., Yokoyama K., Kodate A. and Tominaga H., “*Sports video categorizing method using camera motion parameters*”, In Proceedings of Visual communications and Image processing, Vol 5150, 2003, pp.2082-2088.
- Zeng X., WeimingHu , Wanqing Li, Zhang X. and Xu B., “ *Key frame extraction using dominant set clustering*”, In Proceedings of International conference on Multimedia & Expo(ICME), 2008.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011



# Animated Story Visualizer for Tamil Text

**M. Janani, Dr. T. Mala**

*Department of Information Science and Technology  
Anna University, Chennai, India  
janani\_mrl@yahoo.com, malanehru@annauniv.edu*

## Abstract

Natural language is a straightforward and efficient medium for describing visual facts and mental images. The System uses a novel approach to generate animation from Tamil texts such as stories. Tamil text is pre-processed and the necessary features like named entities, environmental constraints, temporal and emotion constraints for the given stories are extracted and placed in the database. The system automatically generates a query based on the users input and compare it with features stored in the database. Finally animation is dynamically generated using an external motion synthesis system. Using this system, even greenhorn users can generate animation quickly and easily by giving the Tamil text.

**Keywords— Computer animation, Natural Language Processing, Pre-processing, Feature Extraction, Motion synthesis**

## I. INTRODUCTION

These days, animations are widely used in many applications, such as cartoons, web graphics, games, and so on. Computer animation is one of the best methods for depicting the dynamic content. A medium is necessary for the animation to be created in a convenient and natural manner. It should be possible to describe the scenes directly from natural language. NLP is an easy and effective way to analyze, understand and generate languages that humans use naturally.

The aim of this work is to generate an animation from Tamil texts such as movie scripts or stories. Training input text is given to the pre-processing module. Here tokenization is performed and the tokens are given to the morphological analyser which is used to convert the tokens into a POS tags. Information related to named entities, temporal constraints, emotion and environment inference features are extracted. A query is generated automatically from the input text which contains information for the search process and compares it with the information already stored in the database. Finally motion synthesis generates an animation. The interactions between characters are handled by this module based on the information provided in the database.

## II. RELATED WORK

Generating animation from natural language texts has been a challenge. WordsEye developed by Coyne and Sproat [1] converts natural language texts to a scene. WordsEye focuses on generating a still image, when a character motion is specified in a given text, the system simply prefer to pose for

the action generated from the database. The Carsim system [2] describes a new version of text-to-scene converter that handles texts describing car accidents using computer program and it is visualized in the 3D environment. Storytelling System [3] illustrates a system called Interactive e-Hon, which provides storytelling in the form of animation and conversation translated from original text. A Constraint based scene conversion system [4] describes a Text2Scene conversion method which automatically converts text into 3D scenes. A large database of 3D models is used by this method to depict entities and actions.

### III. SYSTEM OVERVIEW

In this section, overview of our system (Figure 1) is given, where the major components are identified. When the Tamil text is given to the system, Tamil text is pre-processed and the information are extracted and stored in the database along with the objects created. When an input text is given to the system it automatically generates the query from the input text and compares it with the information stored in the database. An animation is then generated using an external motion synthesis system.

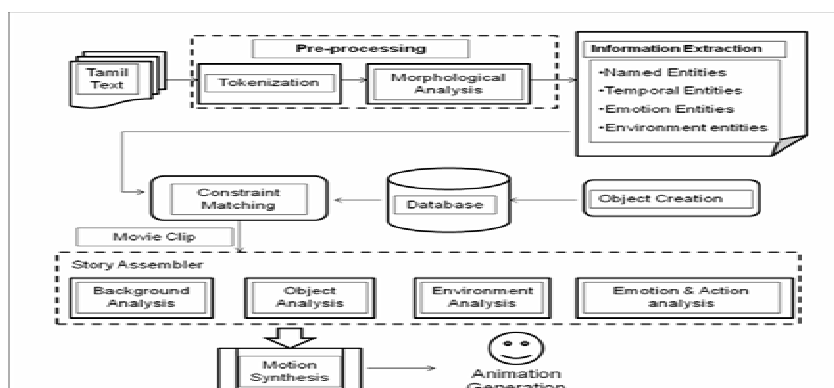


Figure 1. SYSTEM OVERVIEW

#### PRE-PROCESSING

When the Tamil text is given to the system, natural language processes (Tokenization and morphological analysis) are applied first.

##### Tokenization

The first step in NLP is to identify tokens, which decomposes the delimiters like punctuation and whitespaces. Here Tamil text is given as the input to the tokenizer which breaks the text into meaningful tokens. The tokens generated by the tokenizer are passed to the analysis engine.

##### Morphological Analysis

RCILTS [5] developed a tool called Atcharam, an analyser which performs Morphological Analysis for Tamil text. The Morphological analyser takes a derived word as input and separates it into root word and associated morphemes. It is the basic tool used in spell checker, grammar checker, parser and machine translation systems. It has two major modules noun analyzer and verb analyzer.

|              |                     |           |                         |
|--------------|---------------------|-----------|-------------------------|
| வண்டுகளுக்கு | வண்டு< noun >       | படித்தான் | படி< verb >             |
|              | கள்< plural >       |           | த்த< past tense marker> |
|              | உக்கு<case marker > |           | ன்< gender >            |

*Tamil Noun and Verb classification example*

By this method morphemes are generated and given to the learning process where the necessary informations are extracted.

## INFORMATION EXTRACTION

### OBJECT IDENTIFIERS

Object Identifiers recognize named entities in text by Named Entity Recognition (NER). “Rule based approach” is used to extract named entities from the given text. Initially, root words say Noun, verb, adjective, pronoun, adverb from text file are extracted. Rules are created based on prefix and postfix of noun, i.e. noun that occurs between verb and noun, noun that occurs between noun and noun, noun that occurs between noun and verb and so on. If any of the above rules satisfies the input text named entities are extracted. Here is an example,

Input: ஒரு குளத்தில் ஏறும்பு தத்தளித்து

Given input text is pre-processed and the root words are extracted.

குளம்<Noun>

ஏறும்பு<Noun>

தத்தளி<Verb>

Now the rules are applied to this extracted root words. Here ஏறும்பு comes between noun and verb which satisfies the rule is extracted.

### TEMPORAL AND EMOTION EXTRACTION

Temporal reasoning in NLP involves extraction, representation and reasoning with time and events in the natural language text. Here to extract temporal constraints, “manually created dictionary” is used. The root words are compared with the manually created dictionary and temporal constraints are extracted if the input text satisfies the inferences present in the dictionary. Similarly different emotion present in the text is also extracted using manually created dictionary.

Figure 2 shows the different emotional constraints to be depicted.

|          |   |
|----------|---|
| HAPPY    | இன்பம்,மகிழ்ச்சி,குதூகலம்,சந்தோஷம்,ஆனந்தம்,களிப்பு,பெருமிதம்,சிரிப்பு   |
| ANGER    | அகங்காரம்,சினம்,கோபம்,வெறுப்பு,எரிச்சல்,சீற்றம் தாபம்   |
| SURPRISE | அதிசயம்,ஆச்சரியம், பிரமிப்பு,மலைப்பு, வியப்பு   |
| FEAR     | அச்சம்,பயம்,பீர்,பொருமல்,விதிர்ப்பு,கவலை,அஞ்சு,கலக்கம்,நடுக்கம்   |
| SADNESS  | சோகம்,அழுகை,சோர்ந்த, துயரம்,வருத்தம், துன்பம், கண்ணீர், கூச்சல், அலறு, கதறு,புலம்பு,கத்து, முழக்கம்,கூக்குரல், துக்கம், வாட்டம், விசனம் |

*Figure 2 Emotion Constraints*

Finally, environmental constraints that specify location and actions are extracted.

## ANIMATION GENERATION

Movie clips for the extracted information are created using Adobe Flash professional and stored onto the database.

### CONSTRAINT MATCHING

When an input text is given, information constraint should match with the information present in the database to generate animation. String matching algorithm is used to compare the information extracted and information stored in the database. Let  $P[1..M]$  and  $T[1..N]$  be the character array for the given string. Pattern  $P$  is said to occur with shift  $s$  in text  $T$ . To find all valid shifts or possible values of  $s$  so that  $P[1..m] = T[s+1..s+m]$ ; There are  $n-m+1$  possible values of  $s$ .

Procedure String Matcher( $T, P$ )

1.  $n \leftarrow \text{length}[T]$ ;
2.  $m \leftarrow \text{length}[P]$ ;
3. for  $s \leftarrow 0$  to  $n-m$
4. do if  $P[1..m] = T[s+1..s+m]$
5. then shift  $s$  is valid

Find first match of a pattern of length  $M$  in a text stream of length  $N$ .

The extraction of Pattern கா க ம் is done by,

கா க ம்       $M = 4$

க ழு தை ஆ டு கா க ம் கு ர ங் கு

கா க ம்

கா க ம்

கா க ம்

கா க ம்

கா க ம்

கா க ம்

By this method exact string is matched from the database for the given information.

### STORY ASSEMBLER

Storyboards are the only way to convey rich information, viewing a particular order of events in a most appealing way. Basically the system searches for noun and verb from the given input text then automatically assemble and analyse the subsequence like background, named entities, temporal, emotion and action movie clip from a database that matches the constraints.

### MOTION SYNTHESIS

Animation is generated by motion synthesis by efficiently connecting the movie clips that are assembled by the story assembler from the database. The character and objects interactions are handled by this module based on the information that the movie clips have. Timeline specifies what

kind of action occurs at particular time. Once the timeline has been set, animation is generated for the given Tamil text. Figure 3 shows the animation generated for the given Tamil Text

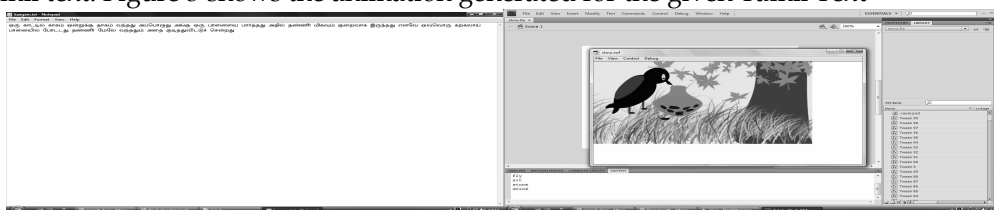


Figure 3 Animation is generated from Tamil text

## PERFORMANCE ANALYSIS

The performance analysis is used to monitor the functioning, efficiency, accuracy and other such aspects of a system. For the analysis performed for the learning process, the overall accuracy obtained is 83%. Figure 4 shows the Performance analysis for Animation generated from the Tamil text.

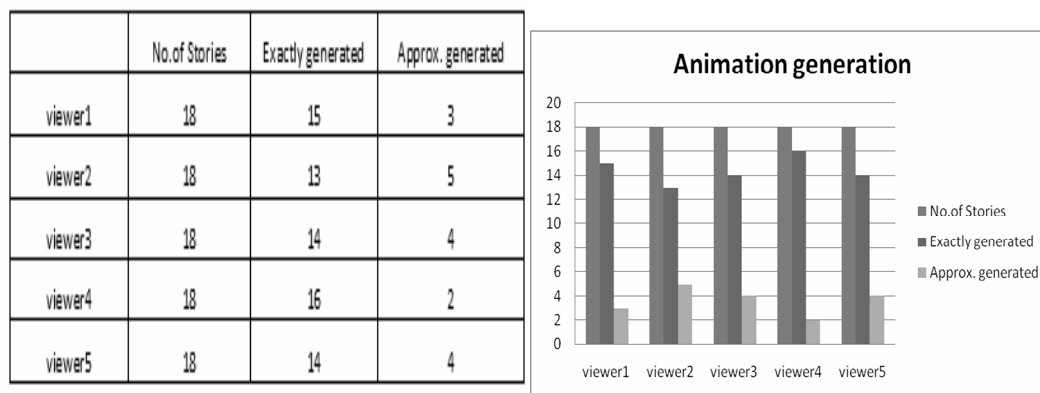


Table 1 Test case for Animation Generation      Figure 4 Performance analysis for Animation generation

The overall accuracy obtained for the generation of animation is 80%. The performance can be further improved by generating rules and optimizing the learning process.

## CONCLUSION AND FUTURE WORK

The system provides automated generation of animation from Tamil text which provides a new approach for users to create animation quickly. The proposed method takes Tamil text as the input and it is pre-processed and features like named entities, temporal constraints, emotion and environmental constraints are extracted and animation is generated dynamically by motion synthesis. Even non-professional people can rely on this system and they can generate animation quickly and easily by giving the Tamil text. Future work can be extended by generating animation via automatic speech recognition rather than text.

## REFERENCES

- Coyne.B and Sproat.R, *“Wordseye: an automatic text-to-scene conversion system”*, In Proceedings of SIGGRAPH 2001, pp. 487-496.
- Johansson.R, Berglund.A, Danielsson.M and Nugues.N, *“Automatic Text-to-Scene Conversion in the Traffic Accident Domain”*, In Proceedings of The Nineteenth International Joint Conference on Artificial Intelligence, 2005, pp. 1073–1078.
- Kaoru Sumi, Mizue Nagata, *“Animated Storytelling System via Text”*, In Proceedings of SIGCHI International Conference on Advances in computer entertainment technology,2006.
- Liuzhou Wu and Zelin Chen, *“A Constraint-based Text-to-Scene Conversion System”*, In Proceedings of International Conference on Computational Intelligence and Software Engineering 2009.
- Anandan, P., Ranjani Parthasarathy & Geetha, T.V., *“Morphological Analyser for Tamil”*, ICON 2002, RCILTS-Tamil, Anna University, India.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Popularity Based Scoring Model for Tamil Word Games

*Elanchezhian.K, Karthikeyan.S, T V Geetha,  
Ranjani Parthasarathi and Madhan Karky*

*chezhiyank@gmail.com, tv\_g@hotmail.com, rp@annauniv.edu, madhankarky@gmail.com*

*Tamil Computing Lab (TaCoLa),*

*College of Engineering Guindy, Anna University, Chennai.*

## Abstract

In this paper we propose a Popularity based Scoring model for Tamil word games. Games are one of the effective means to teach a language. There exist very few online word games for Tamil. Scoring is one of the key aspects of a game that nurtures interest in the player apart from the interface and logic. The Popularity Based Scoring Model proposed in this paper, uses a word statistics crawler to periodically collect the statistics of word usage in popular blogs, news articles and social nets. The popularity of every word is thus modeled in comparison with every other word in the language. The model was successfully implemented in a simple unscramble game titled 'Miruginajambo'. Over three lac root words from Agaraadhi Project were crawled for statistics and 20000 words were obtained for the game based on threshold levels for increasing levels of complexity in the game. The paper concludes providing the results and analysis of implementing the model and also discusses various word games where this model can be used.

## 1. Introduction

Word based games can serve as an effective tool to teach vocabulary in any language. The complexity levels, the score achieved motivates a player to play more and there by learn more words. One key challenge in designing such games is the scoring model. A scoring model for a game means a lot more than just a value associated with the game level or complexity. An effective scoring model has to motivate a player to play more and there by retain the player's interest to come back again.

Tamil language has very few online games available online. These games are mostly flashcard-based games. With new words being introduced in various domains such as medicine, computer science and other disciplines, such games can be the most effective way to teach words. The main reason for popularity of word games in English is their scoring models apart from the user interface they build around their games.

In this paper we propose a popularity-based scoring model for word-based games in Tamil. Providing this scoring model we test the scoring model over two games namely 'miruginajambo', a unscramble game and 'thoekkuthookki', a Tamil equivalent of hangman. Popularity Based Scoring Model generates score based on the combination of the word's popularity, length of the word and Tamil alphabet popularity.



This paper is organized into four major sections. The following section gives the background about scoring models and other relevant literature. The third section gives the popularity based scoring model and the components of the scoring model. The final section concludes discussing the results.

## 2. Background

A scoring model calculates scores based on performance of any system on various domains like Credit Scoring Model in banking application, Fuzzy Logic Approaches in game etc. In the new generation mobile multiplayer games, scoring was generated by using Fuzzy Logic Approach [1]. Automatic Target-scoring System of Shooting Game Based on Computer Vision [2], Online Score System Using Hierarchical Colored Petri Nets [3] to evaluate the outer and inner performances of the system, such as scan, score, and resource utilization. The Credit scoring models [4] are developed to classify the loan applicants as accepted or rejected. The decision is based on the information of each applicant such as age, income and debit ratio. First time we proposed The Popularity Based Scoring Model proposed for Word games, uses a word statistics crawler to periodically collect the statistics of word usage in popular blogs, news articles and social nets. This word Popularity was implemented in the Agaraadhi Online Dictionary [5]. The Word from agaraadhi is given to web and found the frequency distribution of the word across the popular blogs, news articles, social nets etc. The Scoring Model uses the Frequency Analysis of Tamil Alphabet [6]. The Frequency Analysis is done by splitting the Tamil word into alphabet, splitted alphabet are added to their corresponding counter, frequency of each alphabet was identified individually.

## 3. A Game Framework based on Popularity-Based Scoring Model

We propose a simple game framework for an unscramble game in this paper, depicted in figure1. The framework can be basically divided into two major divisions, online and offline, in terms of the time of processing. This section describes the various scoring generator in detail.

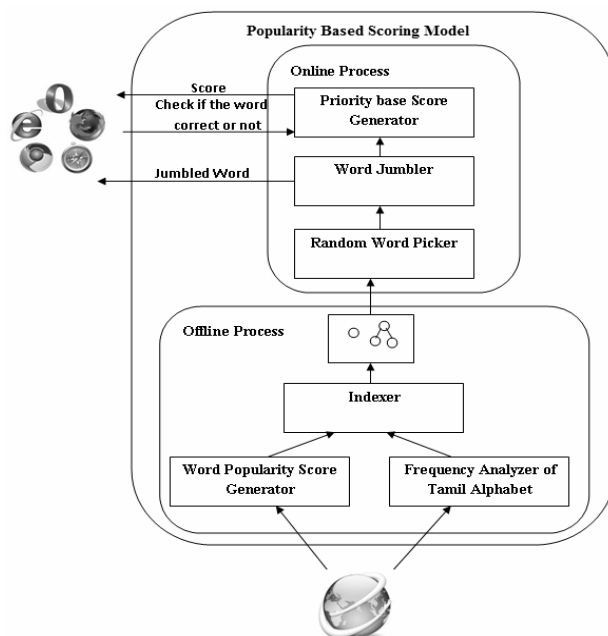


Fig 1: Popularity Based Scoring Model

### **3.1 Offline Processing**

The Offline Process takes a Set of Tamil words as input, Word Popularity Scoring, and Tamil Alphabet Frequency Analysis are the key tasks here.

#### **3.1.1 Bag of Tamil Words**

Tamil words are obtained for analysis from the Agaraadhi project comprising over 3 lac words on various domains such as General, Engineering Technology, Literature, Medicine and Computer Science.

#### **3.1.2 Word Popularity Scoring**

Word Popularity shows the word usage in the web. The words from agaraadhi are crawled over popular news sites; blog articles and social networking sites periodically and the frequency distribution of the word across sites are identified and recorded. This overall information is then used to compute the popularity score for each word.

#### **3.1.3 Tamil Alphabet Frequency Analysis**

The words obtained from the set of words are split into alphabet that constitute the word and the split alphabets are added to their corresponding counter, frequency of each alphabet is identified individually. Alphabet Frequency Analysis generates score based on the frequency value of the alphabets. This Analysis result will be used later to find the popularity of a letter and thus the complexity of a word. So, a low frequency alphabet contained word gets a higher score.

### **3.2 Online Processing**

The Online Process Comprises of Random word Picker, Word Jumbler and Popularity Based Score Generator

#### **3.2.1 Random word Picker**

Random word Picker fetches a random word based on the domain specified by the user and the Word Popularity. It will be possible now to decrease the word popularity score on every higher level to increase the complexity of the game.

#### **3.2.2 Word Scrambler**

Word Scrambler module scrambles the Tamil alphabets in a word such that all alphabets are not placed in their actual correct spots. The jumbler is randomized such that the next time the same word will be jumbled in another combination.

#### **3.2.3 Popularity Based Score Generator**

Score Generator generates the score based on the proposed Popularity Based Scoring Model using the parameters such as word popularity, level of the game, Low frequency scored occurrences of Tamil Alphabets in a word, total number of swaps to complete a level and time taken to complete a level. Let  $w$  be the word,  $l_w$  be the length, Let  $P_w$  be the popularity score of the word, Let  $P_a$  be the average alphabet popularity frequency of the alphabets in  $w$ , Let  $t$  denote the time taken to solve the word in

seconds, Let  $s$  denotes the number of swaps needed to solve the word, Let  $\alpha$  be the score for a perfect answer.

The score for a solving,

$$\text{Score } S(w) = \alpha * (1 - P_w) * (1 - P_a) * (1 - \frac{t}{60}) * (1 - \frac{s}{lw})$$

#### 4. Results and Discussion

The framework depicted in figure 1 was implemented with a simple web interface. The snapshots of the working game with the popularity based scoring model are given in figure 2. The same game was developed with and without the popularity based scoring model with the earlier version giving a score just based on the level and number of letters swapped. The version with popularity based scoring model was found to receive more users playing for a longer time and repeatedly as they find their current score rapidly increase if they identify a less popular word.

#### 5. Conclusion

In this paper we proposed a Popularity-Based Scoring Model for computer based word games in Tamil. The popularity of words was identified by their usage over the internet by news, blog and micro blog writers. This information is then converted to scoring every word in a dictionary. This score is used in the Scoring model to compute the score for the user at different levels. The scoring model is compared to a traditional level based scoring model. Analyzing user behavior over the two models we conclude that the popularity based scoring model creates more interest in user to play the game for long time and repeatedly compared to the traditional level based scoring.



Fig 2: Snapshots of the working game with the popularity based scoring model

## References

- Alan Graf, Siemens d.d, Fuzzy Logic Approach for Modelling Multiplayer Game Scoring System, ConTEL 2005.
- Design of Automatic Target-Scoring System of Shooting Game Based On Computer Vision, Xinnan Fan, Qianqian Cheng, Changzhou, Jiangsu Province, IEEE International Conference on Automation and Logistics Shenyang, China August 2009.
- Yang Xu , Xiaoyao Xie, Daoxun Xia<sup>1</sup>, Zhijie Liu, Lingmin Chen<sup>1</sup>, Modeling and Analysis of an Online Score System Using Colored Petri Nets, Anti-counterfeiting, Security, and Identification in Communication, 2009. ASID 2009. 3rd International Conference, Aug. 2009.
- Entropy multi-hyperplane credit scoring model, Wasakorn Laesanklang, Krung Sinapiromsaran Boonyarit Intiyot, International Conference on Financial Theory and Engineering, 2010.
- Agaraadhi – An Online Tamil-English Dictionary <http://www.agaraadhi.com>, Last Accessed Date on 28th April 2011.
- Karthika Ranganathan, Elanchezhiyan.K, T.V Geetha, Ranjani Parthasarathi & Madhan Karky, The Frequency Analysis of Tamil Alphabet, National Seminar on Computational Linguistics and Language Technology, March, 2011.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Multilingual Cross - Domain Classification of Tamil Web Documents based on Neural Network with Dimension Reduction

*M.Balaji Prasath<sup>1</sup>, Dr.D.Manjula<sup>2</sup>*

*itprasath@gmail.com<sup>1</sup>, manju@annauniv.edu<sup>2</sup>*

*Department of Computer Science and Engineering,  
Anna University, Chennai.*

## Abstract

Automatic classification of web document increases in the regional languages, because of amount of information available in the regional languages (like Tamil, Telugu, Hindi) is huge in the internet in the form of e-Book, news, articles and other type of formats. It is difficult to categorize those documents based on the subject of interest. Tamil is a Rich Dravidian language, it have a millions of documents in the Web Repository, due to growth of digital documents, categorization needed to classify document. Too much classification techniques are present for the English documents classification like SVM, K-NN, Decision trees, Neural Network technique, but classification in regional languages like Tamil, it's new and emerging. So that our proposed work involves first, genetic algorithm will be employed to reduce dimension of document .Second, Multilingual Cross- domain classification, involves the predefined labels in the English language will be used to classify the Tamil Corpus, because pre-defined labels in the source domain is expensive to create, so that look for other domain of same interest to classify the documents. Third Back Propagation Technique applied to classify Corpus.

Key Terms: Classification, Multilingual, Cross-Domain, Dimension Reduction

## Introduction

Today most of the documents exist in the electronic repository like e-books, journals, news articles and other sources of information in form of English only. This electronic document exists in other regional languages also (like Tamil). To classify those Region documents lot of research going on.

Tamil<sup>[1]</sup> is an oldest regional language present in the world. Around billion of people speaking Tamil and lot of documents present in the Tamil language. Natural Language processing of Tamil is difficult, because of little bit research is taken place. To analyze their keywords, linguistics plays an important role. Lot of research already taken place to classify English documents based on supervised and unsupervised learning. I.e. two types learning is their (i) supervised leaning means of classification documents based on the pre-defined label categorization. It first train the training document based upon the pre-defined labels, and test the test documents and classify based upon the training set. (ii) Unsupervised learning is a clustering.

Many machine learning technique available like SVM, KNN Classifier, Neural Network, Bayesian Classifier based on mathematical approaches. For Pre-label is expensive, to avoid that other domain

label is used for classification purposes, but in English document collection lot of Cross-Domain<sup>[10]</sup> Labels are available, in order to classify the documents in other domain but in Classification based on rare. So that use labels in the domain of English Document, to the corresponding labels in Tamil documents, it reduce the classification effort, and also produce better results.

Before using those approaches dimension reduction plays an important role to minimize the no of keywords present in the document. Our proposed approach use the genetic algorithm for reduction of no of key attributes present in the documents.

Dimension reduction carried out based on feature selection and feature reduction methods.

Feature selection<sup>[2]</sup> means that, it selects the keywords based on attributes, which contribute reduction of no of words in the documents. Selection plays an important role here. Two types present (i) filter method Separating the feature selection from the classifier learning, and relay on general characteristics of data, no bias over any learning algorithm, generally it fast. (ii) Wrapper model, relaying on predefined classification algorithm, and computationally expensive.

Genetic algorithm<sup>[3]</sup> will be used as a dimension reduction technique, it takes the set of keywords as a population of terms, and neural network will be employed as a classifier, which train and classify the training documents and classify testing documents after that training phase.

Tamil corpus will be generated automatically, by using a web crawler. Crawler return the set of document pages (Tamil) particularly news articles. These collectively articles used to form the corpus. Further classification will be done using those Tamil news articles (Corpus).

This paper section 2 describes the web crawler section 3 describes the Tamil corpus, section 4. describes the dimension reduction using the genetic algorithm, section 5, describes the classification using neural network.

## 2. Web Crawler

Crawler is a software program, which can fetch the WebPages based on the seed URL given to the Crawler. Here seed URL will be “Tamil news article” site URL. This crawler crawls only site given to the input to the crawler, it doesn’t navigate to other site. It uses a muti threaded downloader to down load the web pages , based on that seed URL given to the crawler, this crawler, crawls the pages only within that link. Suppose it should news.goole.com means, it crawl the link fully, retrieve the document within that.

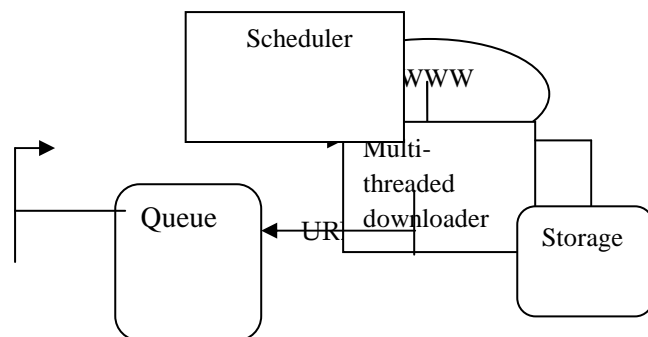


Figure 1. Architecture of Web Crawler

### 3. Tamil Corpus

Many research going on to build a corpus in Tamil. Central institute of Indian language (CIIL)<sup>[4]</sup>, Mysore actively involved in building the corpus in different regional languages.

Here, corpus build by using a web crawler, it crawl a web pages and stored it in a local database. After that it will be edited in order to make and formed as a corpus

Tamil has 12 vowels and 18 consonants. This are combined with together 217 composite characters and 1 special characters counting to the total of 247 characters. To build a corpus for that rich type of grammar is too difficult. So that crawler will used to retrieve web content, and it edited to form a corpus.

### 4. Dimension Reduction using genetic algorithm

Generally dimension reduction used to reduce the no of words in a corpus. Because corpus have a huge collection of words, but few collections of words in a corpus makes the document meaningful, So that to identify those words, dimension reduction plays a vital role.

Genetic algorithm used as an optimization technique. Here it plays as a dimension reduction, it choose a set of attributes like content name, sub-content title, and other.

Genetic algorithm uses a input as a set of population of attributes, instead of choosing a single attributes, it reduce the no of words from a thousand to hundred, each attribute like a gene, group of attribute forms a chromosome, uses a various operation like,

1. Crossover: single or multi-point
2. Mutation
3. Reproduction

#### 4.1 Algorithm of GA for dimension reduction

Step1: form the set of attribute as a chromosome.

Step2: generate the fitness function for each gene in the population.

Step3: apply the genetic operator, and evaluate the fitness once again.

Step4: stop, if attain the terminating condition, else generate new population and go to step2.

### 5. Classification of Web Document

After the Identification of set of key attributes, need to classify the training documents using a neural network. Neural network <sup>[9]</sup> have a set of input nodes, and hidden nodes, and a corresponding output nodes. Training documents taken as a input to the system, that should be trained and classified accordingly based on the attributes generated by the genetic algorithm, theoretically says that, dimension reduction after that classification improve the result future

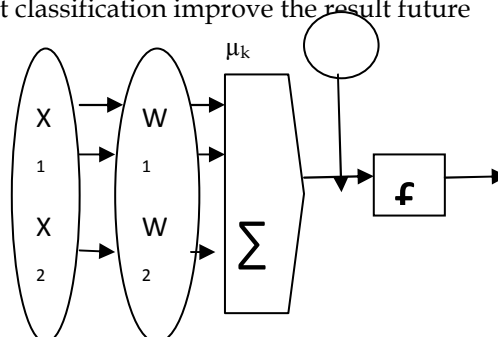


Figure 2. Architecture of Neural Network.



Back propagation technique employed in the classification of web documents. Feedback given to the neural network with every set of documents trained. After that documents tested against the network, whether it should be classified correctly. Theoretical performance is better than other classification technique.

## Conclusion and Future Work

Automatic classification of Tamil web content increase the need for separate classification approaches, for that genetic algorithm employed as a dimension reduction technique, and classified accordingly based on the selected attributes, it improve the precision and recall after the dimension reduction.

Future improvement in the neural network, will be use of winnow/preceptor technique with no hidden layer improve classification technique.

## References

- K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel, B. Palaniappan, Automatic classification of Tamil documents using vector space model and artificial neural network, *Expert Systems with Applications* 36 (2009) 10914–10918.
- Nan Du, Hong Peng, Wenfeng Zhang, Application of Modified Genetic Algorithm in Feature extraction of the Unstructured Data, *International Conference on Advanced Computer Control*, IEEE 124-128.
- Philomina Simon, S. Siva Sathya, Genetic Algorithm for Information Retrieval
- M. Ganesan, Tamil Corpus Generation and Text Analysis
- M. Selvam, and A. M. Natarajan, Language model adaptation in Tamil language using cross-lingual latent semantic analysis with document aligned corpora, *CURRENT SCIENCE*, VOL. 98, NO. 7, 10 APRIL 2010
- Thair Nu Phyu, Survey of Classification Techniques in Data Mining, *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009*, March 18 - 20, 2009, Hong Kong
- S.Kohilavani, T.Mala and T.V.Geetha, Automatic Tamil Content Generation, *IEEE IAMA* 2009.
- Chih-Ming Chen, Hahn-Ming Lee, Yu-Jung Chang, Two novel feature selection approaches for web page classification, *Expert Systems with Applications* 36 (2009) 260–272
- Cheng Hua Li, Soon Choel Park, An efficient document classification model using an improved back propagation neural network and singular value decomposition, *Expert Systems with Applications* 36 (2009) 3208–3215.
- Sinno Jialin Pany, Xiaochuan Niz, Jian-Tao Sunz, Qiang Yangy and Zheng Chen, Cross-Domain Sentiment Classification via Spectral Feature Alignment, *WWW 2010*, April 26–30, 2010, Raleigh, North Carolina, USA.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# On Emotion Detection from Tamil Text

*Giruba Beulah S E, and Madhan Karky V*

*Tamil Computing Lab (TaCoLa),*

*College of Engineering, Anna University, Chennai.*

*(ljcsegb@gmail.com) (madhankarky@gmail.com)*

## Abstract

Emotion detection from text is pragmatically complicated than such recognition from audio and video, as text has no audio or visual cues. Techniques of emotion identification from text are, by and large, linguistic based, machine learning based or a combination of both. This paper intends to perceive emotions from Tamil news text, in the perspective of a positive profile, through a neural network. The chief inputs to the neural network are outputs from a domain classifier, two schmalzky analyzers and three affect taggers. To aid precise recognition, tense affect, inanimate/animate case affect and sub-emotion affect dole out as the supplementary inputs. The emotion thus recognized by the generalized neural net is displayed via a two dimensional animated face generator. The performance and evaluation of the neural network are then reported. Index Terms—Emotion detection, Machine learning, Neural network, Tamil news text.

## I. Introduction

Emotion detection from text attracts substantial attention these days as this if realized, could result in the realization of a lot of fascinating applications like emotive android assistants, blog emotion animators etc. However, emotion detection from text is not trouble-free, as one cannot, with a single glance get a hold of the emotions of the people about whom the text is based. Further, what appears to be sad for a person may perhaps appear fear for someone. Emotion detection from text is thus influenced by the empathy of the readers. Also, as there may be no background information, to shore up the emotion of a person in text, it is better if emotion detection is based on some model empathy.

The aim of this paper is to detect emotion from Tamil news text by a self learning neural network which takes in linguistic and part of speech emotive features. The emotion is identified by assigning weights for features based on their affective influence.

## II. Background

Tamil is a Dravidian language as old as five thousand years and enjoys classical language of the world status along with Hebrew, Greek, Latin, Chinese and Sanskrit. Unlike Sanskrit, Latin and Greek, which are very rarely in use, it is a living language and has fathered many Dravidian languages like Malayalam, Telugu etc. It is morphologically very rich and has a partial free word order. It groups noun and verb modifiers, (adjectives and adverbs) under a single category, *Urichols*. Noun participles are equally affective as the verb participles. Thus, apart from parts of speech, morphological entities like cases and participles are also affective.

Among the methods of emotion detection, [1] observes that the Support Vector Machines using

manual lexicons and Bag Of Words approach perform better. The Support Vector Regression Correlation Ensemble is an album of classifiers, each trained using a feature subset tailored to find a single affect class. However, as support Vector machines suffer from high algorithmic complexity and requirement of quadratic programming and do not efficiently model non-linear problems, Neural networks has been opted as they provide greater accuracy than Support Vector Machines.

Nevertheless, Neural networks sport disadvantages like local minima and over fitting. The former can be handled by adding momentum and the latter is usually solved by early stopping. Since the neural net could memorize all training examples if over the need hidden neurons are present, three promising prototypes are constructed, trained and tested to find the optimal one. To this optimal neural net, the affective feature tags from Tamil news text are given, to recognize the inherent emotion, based on their respective affective strength.

Kao, Leo, Yahng, Hsieh and Soo use a combinatory approach of dependency trees, emotion model ontology and Case based reasoning [2] where cases are manually annotated. Such an approach may work for languages of few cases. However, not all cases in a language may be affect sensitive. Cases in Tamil are eight and only two of these are affect sensitive, namely the instrumental and the accusative case. Thus manually annotating cases in Tamil would normally fail as cases aid in increasing or decreasing the affect of the verb nearby.

The separate sub-class networks [10] for Emotion recognition with context independence in speech seem to be promising but for the low precision of 50 even when the learning is supervised using a simple backpropagation network. However, this project is similar to the above in one aspect; in assuming model empathy to support context independence, thereby avoiding the bias on a particular individual involved in the text.

Sugimoto and Masahide [9] split the text into discourse units and then into sentences identifying the emotion of discourse and sentences separately. The language of the text considered is Chinese which is monosyllabic partially with nouns, verbs and adjectives being largely disyllabic. Tamil is partial free word order and does not have such phonological restrictions. Hence, application of such an approach to Tamil may not be fitting intuitively.

Soo Seoul, Joo Kim and Woo Kim propose to use keyword based model for emotion recognition when keywords are present and to use Knowledge based ANN when text lacked emotional keywords [6]. The KBANN uses horn clauses and example data. However, the accuracy of emotion recognition using KBANN is less ranging from 45%-63% compared to the recognition range of 90% where emotional keyword affect analysis is incorporated.

Most of the research in emotion recognition is directed toward audio and video from which one can infer so many cues even without understanding the narration. For text, the features of the language of text has a major emphasis on emotion recognition. Tamil is a morphologically rich language and hence its affect sensitive features would drastically vary with the usually chosen languages for emotion recognition like Chinese which is character oriented and English which is least partially inflected.

### III. An architecture for Emotion Recognition

This section throws light on each module and its functions. Tamil Morphological analyzer, a product of the Tamil Computing lab of Anna University is used as the tool to retrieve part of speech like nouns, verbs, modifiers and cases. The 2D animated face generator is another tool to display the found emotion via a two dimensional face.

Following is the architecture diagram of the Neural net framework.

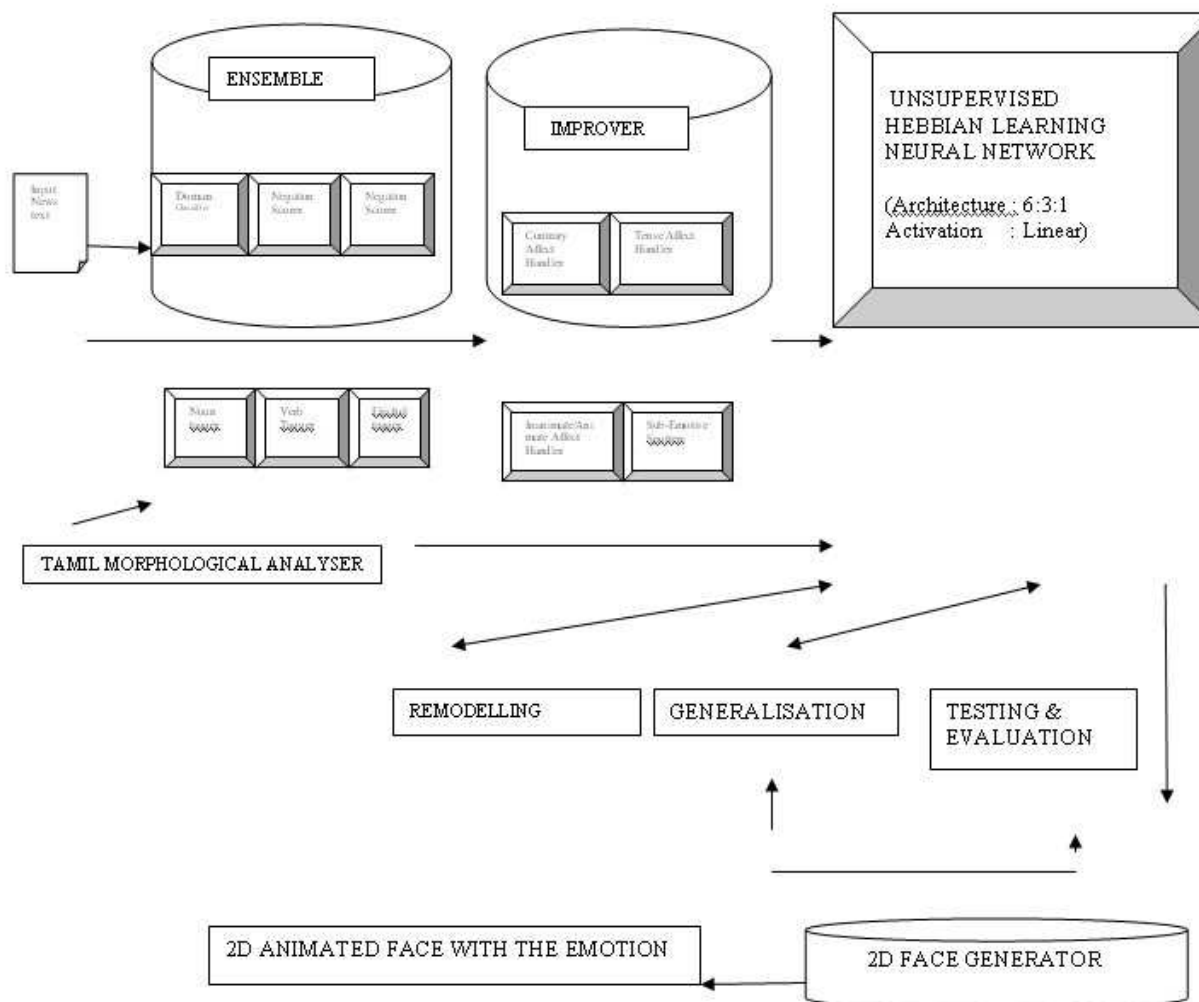


Fig 1: The Overall Architecture

#### A) The Domain Classifier

The Domain Classifier takes in documents which are already classified under five domains namely, politics, cinema, sports, business and health. Training process involves retrieving nouns and verbs using the Tamil Morphological Analyzer, calculating file constituent terms' domain frequencies and inserting them into the respective domain hash tables with terms as keys and term frequencies as values. Training reports a document's domain, based on the highest test domain frequency counter.

The algorithm is as follows

i) *Preprocessing:*

Let  $F_D$  represent the set of files in a domain and  $f$  be a file such that  $f \in F_D$ . Let  $T$  be the bag of words in a domain without stop words and  $t \in T$ .

1.  $\forall f \in F_D$ , remove stop words and tokenize
2. Get the nouns and verbs using the Tamil Morphological analyzer.
3.  $\forall t \in T, Ft = \sum f_t$  where  $f_t$  is the frequency of  $t$  in  $f$ .

ii) *Training*:

Let  $H_D$  represent a domain hash table

$\forall t \in T$ , insert  $(t, H_D)$  where  $t$  being the key and the value  $v$  being

$$v = I_{occ} + \sum f_{dup(v)}$$

where  $I_{occ}$  is the first occurrence of the term in the domain and  $f_{dup(v)}$  is the frequency of its duplicate.

iii) *Testing*:

**Initialise domain frequency counters  $P_c, C_c, B_c, S_c$  and  $H_c$ . Given a document, remove the stopwords and tokenise.**

**Let Tok be the bag of nouns and verbs in the document, retrieved using the Tamil Morphological Analyser.**

**Convert Tok to a set by removing the duplicates.**

**$\forall t \in Tok$ , get the frequencies of  $t$ , from all the domain hash tables. Increment the domain frequency counter of the domain that yields the highest frequency for  $t$ .**

**Report the domain of the test document as the domain that has the highest counter value.**

#### B) *The Negation Scorer*

The Negation classifier gets the documents, analyses whether positive words occur in the neighborhood of negative words or whether likes come close to dislikes and vice versa and assigns a score.

i) *Training*:

Let  $F$  denote set of all files in the corpus and let  $f \in F$ . Let  $T$  be the bag of words in a file  $f$  without stop words and  $t \in T$ .

- a)  $\forall f \in F$ , remove stop words and tokenize.
- b)  $\forall f \in F$ , let  $Neg$  denote the negation score of  $f$  primarily initialised to 1.
- c)  $\forall t \in T$ , let  $i$  be the current position. Check whether  $t$  is a positive/negative word or like/dislike word.  
 if  $t$  is positive/like, and a negative/dislike word occurs in a window of three places to the left or right, calculate  $Neg$  as  
 $Neg = Neg - 0.01$   
 if  $t$  is negative/dislike, and a positive/like word occurs in a window of three places to the left or right, calculate  $Neg$  as  
 $Neg = Neg - 0.1$
- d) If  $Neg < -0.5$ , report the document as negative.

### C) The Flow Scorer

The Flow scorer assigns a score to each document based on the pleasantness of the words it has. The score is set in view of the phonetic classification in Tamil alongside with place and manner of articulation. Let  $F$  denote set of all files in the corpus and let  $f \in F$ . Let  $T$  be the bag of words in a file  $f$  without stop words and  $t \in T$ . Let  $t_g$  be the English equivalent of a Tamil word  $t$ .

- $\forall f \in F$ , remove stop words and tokenize.
- $\forall t \in f$ , convert  $t$  to  $t_g$
- $\forall t_g \in f$ , compute the flow score as the sum of the Maaththirai counts diminishing when Kurukkams appear.

**Table 1:** Kurukkams and Maaththirai

| Rule             | Context   | Final Maaththirai      |
|------------------|---|------------------------|
| KuttriyaLukaram  | One among these கு சு ண து பு று at the end of the word | 1,(Decrease is 0.5)    |
| Aukaarakkurukkam | ஒள in the beginning                                     | 1.5,(Decrease is 0.5)  |
| Aikaarakkurukkam | ஐ in the beginning and middle                           | 1.5,(Decrease is 0.5)  |
|                  | ஐ in the end  | 1,(Decrease is 1)      |
| Maharakkurukkam  | வ before ம்   | 0.25,(Decrease is 1/4) |

**Table 2.** Phonemes under Manner of articulation

| Category      | Manner of Articulation   | Phoneme     |
|---------------|--------------------------|-------------|
| Greater Rough | Retroflex, Trill         | ட ற         |
| Rough         | Tap, Dental, Bilabial    | க ச த ப ர   |
| Intermediate  | Semivowels, Approximants | ய ல ள ழ வ   |
| Soft          | Nasal                    | ங ஞ ண ந ம ன |

Let  $t$  be the Tamil word,  $t_g$  be the Grapheme form and  $P_t$  be the bag of phonemes in  $t$  with  $p \in P_t$ . Let  $GRscore$  be the score of the greater rough category,  $Rscore$  be the score of the rough category,  $Iscore$  be the intermediate score and  $Sscore$  be the score of the soft category.

- Calculate  $GRscore$ ,  $Rscore$ ,  $Iscore$ ,  $Sscore$  as

$$GRscore = \sum f(p)_{GR}.$$

$$Rscore = \sum f(p)_R.$$

$$Iscore = \sum f(p)_I.$$

$$Sscore = \sum f(p)_S.$$

where  $f(p)_{GR}$  is the frequency of a greater rough category phoneme,  $f(p)_R$  is the frequency of a rough category phoneme,  $f(p)_I$  is the frequency of a intermediate category phoneme and  $f(p)_S$  is the frequency of a soft category phoneme.

- ii) Calculate the *FinalRoughScore* and *FinalSoftScore* as

$$FinalRoughScore = GRscore + Rscore .$$

$$FinalSoftScore = Iscore + Sscore .$$

- iii) If *FinalRoughScore* > *FinalSoftScore*  $T \rightarrow Pleasant$
- iv) Else if *FinalSoftScore* > *FinalRoughScore*  $T \rightarrow Unpleasant$
- v) Else  $T \rightarrow Neutral$

#### D) *The Taggers*

At the start four taggers were constructed specifically, the noun tagger, verb tagger, *Urichol* tagger and case tagger. But for the instrumental and accusative cases, the left behind cases are not affective. For this reason, the constructed case tagger is unused as an input to the neural network. Nonetheless, the affect sensitive cases are incorporated in the affect specificity improver, namely the animate/inanimate affect handler.

Each tagger, picks up the respective part of speech in the document, analyses the major affect and tags the document with that affect. As a consequence, noun tagger reports the affect of nouns, verb tagger reports the affect of verbs and *Urichol* tagger reports the affect of *Urichols*. The generalized algorithm is as follows.

Let  $F$  denote set of all files in the corpus and let  $f \in F$ . Let  $T$  be the bag of words in a file  $f$  without stop words and  $t \in T$ . Let  $N$  denote nouns,  $V$  denote the verbs and  $U$  denote the *Urichols* in a file. Let  $n \in N$ ,  $v \in V$ ,  $u \in U$ .

- a)  $\forall f \in F$ , remove stop words and tokenize.
- b)  $\forall f \in F$ , get  $N$  for Noun Tagger,  $V$  for Verb tagger and  $U$  for *Urichol* tagger, using the Tamil Morphological analyzer.
- c)  $\forall n \in N$ , use the noun affect lexicon to determine the noun affects. Initialize respective noun affect counters for the basic six emotions and increment them depending on the affect of each  $n$ .
- d)  $\forall v \in V$ , use the Verb affect lexicon to finalize the verb affects. Initialize respective affect verb counters for the basic six emotions and increment them depending on the affect of each  $v$ .
- e)  $\forall u \in U$ , find the *Urichol* affect using the *Urichol* affect lexicon. Initialise respective affect counters for the basic six emotions and increment them depending on the affect of each  $u$ .
- f) For each tagger, report the final affect of a file  $f$ , as the affect of the respective affect counter that has the maximum value.

#### E) *The Tense Affect Handler*

- a)  $\forall f \in F$ , get the verbs under each of the three tenses. Let  $Pr$  represent the present tense,  $Pa$  denote the past tense and  $F$  denote the future tense.
- b) Analyze the affect of each tense category verbs.
- c) Prioritize Present, Future and then Past tenses.
- d) Report the high frequent affect of the Present tense. If present tense verbs are absent, report the maximum frequent affect of future tense, else report the high frequent affect category of the past tense.

#### F) *The Inanimate/Animate Handler*

- a)  $\forall f \in F$ , get the verbs using the Morphological analyzer.



- b) *Restrict a window size of one to the left to find whether affect sensitive cases are present.*
- c) *Analyze the affect of the verbs and categorize them under mild and dangerous.*
- d) *Analyze the cases and see if they add to the affect of the verbs or nullify it.*
- e) *If affect intensity is increased report danger, else report the affect implied.*

#### **G) The Contrary Affect Handler**

- a)  *$\forall f \in F$ , check whether there are multiple entities in a sentence.*
- b) *Using the Profile monitor, check whether the entities are in like list or dislike list.*
- c) *Analyze the nouns next to the entities to see whether they are favorable to the preferred entities or not.*
- d) *Report the affect as the affect of the nouns with reference to the preferred entity.*

#### **H) The Sub-emotion Spotter**

- a)  *$\forall f \in F$ , Get all nouns, verbs and urichols using the Tamil morphological analyzer.*
- b) *Use the constructed sub-emotive affect lexicon to identify the high frequent sub emotion.*
- c) *Report the maximum occurring affect based on the sub- affect.*

#### **I) The Neural Network**

Initially a supervised Backpropogation network was constructed with the architecture 6:3:1. However, since emotion recognition is to do with emotional intelligence, unsupervised learning was preferred and Hebbian learning was incorporated. The forget factor is fixed as 0.02 and the learning factor as 0.1. Following is the pseudo code.

- a)  *$\forall f \in F$ , get the outputs from Domain Classifier, Negation Scorer, Flow Scorer and the three affect taggers.(The Improver modules fail to aid emotion recognition by getting eclipsed toward certain domains and are hence overlooked.)*
- b) *Initialize the network weights and threshold values which are set in the range of 0 to 1.*
- c) *Start the learning for each file using Hebb's rule.*
- d) *Report the affect as the affect of the output activation that approximately equals an affect value.*
- e) *Stop the learning when steady state is embarked or sufficient number of epochs has been elapsed.*

## **IV. Results and Evaluation**

Both Batch and Sequential learning are supported. The precision of the Neural network is 60% . Other datasets used were lyrics and stories. Stories usually have sub plots and hence overall affect usually gets eclipsed towards neutral. Hence, news (400 files) and lyrics (230) were used for training. Validation set included 50 news files with equal contribution from five domains and 10 lyrics.

For lyrics and stories, the neural net has a better precision of 70%. The precision gets increased when epochs increase for all the datasets. Simple supervised Back propogation network was implemented and used as the Baseline whose precision was 30% more than the constructed even with minimum epochs.

Following is a graph that depicts how precision of emotion found varies with respect to the number of epochs, in the case of a single news text. Values nearer to zero indicate the drastic variation of the reported affect from the actual one (ie joy instead of sad) and values near 100 indicate the closeness towards the actual affect.

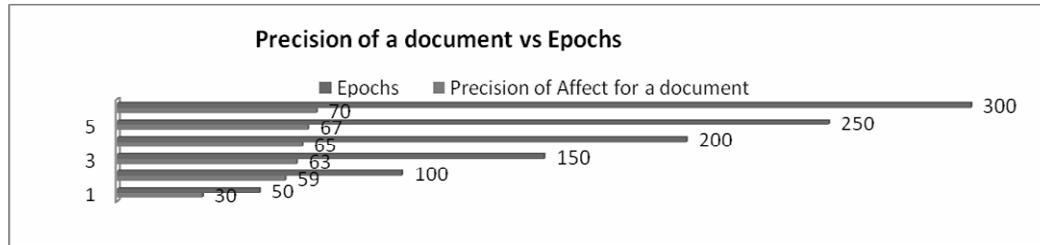


Fig 2: Epochs vs Precision of affect of a document.

The neural network suffers from ambiguity problems in the case of closely related categories like love-joy and sad-fear irrespective of dataset. The supervised Back propogation network is used as the baseline which is 30% more in precision than the Unsupervised Hebbian Learning.

Besides, the Unsupervised Hebbian learning Neural net has exponential complexity and consumes two thirds of the physical memory. The affect reporting of a single file amounts to three minutes where approximately one and a half minute is taken for indexing and loading of the domain hash tables by Domain Classifier. The CPU usage during the learning varies roughly from 11% to 72%.

## V. Conclusion

Thus, an unsupervised Hebbian learning neural network is constructed which fetches its major inputs from a domain classifier, two sentimental scorers and three part of speech taggers to figure out the affect in the presented text. As is the case with almost all natural language processing applications, ambiguities do exist in emotion recognition; here among closely related affective categories like love-joy and sad-fear. However, a competitive strategy in unsupervised learning can be opted to resolve this issue, as such a learning is concerned with demarcating one from the rest. Identification of other affect sensitive features could further aid in precise emotion detection.

## VI. References

- Ahmed Abbasi, Hsinchun Chen, Sven Thomas, Tianjun Fiu, "Affect Analysis of Web forums and blogs using correlation ensembles", IEEE Transactions on Knowledge and Data Engineering, Vol.20, No.9, pp.1168-1180, Sepetember 2008.
- Edward Chao-Chun Kao, Chun Chieh Liu, Ting-Hao Yong, Chang Tai Hseih, Von Wun Su, "Towards text-based Emotion detection", International Conference on Information Management and Engineering, 2008.
- Sowmiya, Madhan Karky V, "Face Waves : 2D Facial Expressions based on Tamil Emotion Descriptors", World Classical Tamil Conference, June 2010.

- Ze-Jing Chuang and Chung-Hsien Wu, "Multimodal Emotion Recognition from Speech and text", Computational Linguistics and Chinese Language Processing, Vol.9,No.2, pp. 45-62, August 2004.
- Maja Pantic, Leon J.M. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-Computer Interaction", Proceedings of the IEEE, Vol.91, pp.1370-1390, September 2003.
- Yong-Soo Seol, Dong-Joo Kim and Han-Woo Kim, "Emotion Recognition from Text Using Knowledge-based ANN", The 23<sup>rd</sup> International Technical Conference on Circuits/Systems, Computers and Communication 2008.
- Donn Morrison, Ruili Wang, W.L.Xu, Liyanage C.De Silva, "Voting Ensembles for Spoken Affect Classification", Elsevier Science, February 6,2006.
- Futoshi Sugimoto, Yoneyama Masahide, "A method for classifying Emotion of Text based on Emotional Dictionaries for Emotional Reading".
- J.Nicholson, K.Takahashi, R. Nakatsu, "Emotion recognition in Speech using Neural networks", Neural Computing and Applications, Springer-Verlag London limited, pp.290-296, 2009.
- Lyle N.Long, Ankur Gupta, "Biologically -Inspired spiking Neural networks with Hebbian learning for vision processing", AIAA 46<sup>th</sup> Aerospace Sciences meeting, Reno, NV, Jan 2008.
- Lei Shi, Bai Sun, Liang Kong, Yan Zhang, "Web forum Sentiment analysis based on topics", IEEE Ninth International Conference on Computer and Information Technology, 2009.
- Changrong Yu, Jiehan Zhou, Jukka Rieki, "Expression and analysis of Emotions - A Survey and Experiment", IEEE Symposia and Workshop on Ubiquitous, Autonomic and Trusted Computing, 2009.
- Irene Albrecht, Jorg Haber, Kolja Kahler, Marc Shroeder, Hans Peter Siedel, "May I talk to you :- ) Facial Animation from Text", IEEE Proceedings of the 10<sup>th</sup> Pacific Conference on Computer Graphics and Applications, 2002.
- Aysegul Cayci, Selcuk Sumengen, Cagatay Turkey, Selim Balcisoy, Yucel Saygin, "Temporal Dynamics of User interests in Web search queries", International Conference on Advanced Information Networking and Application Workshops, 2009.
- Tim Andersen, Wei Zhang, "Features for Neural net based region identification for Newspaper documents", Proceedings of the seventh International Conference on Document Analysis and recognition, 2003.
- Taeho Jo, Malrey Lee, Thomas M Gatton, "Keyword Extraction from Documents using a Neural network model", International Conference on Hybrid Information Technology, 2006.
- Azam Rabiee, Saeed Setayeshi, "Persian Accents Identification Using an adaptive Neural network", IEEE Second International Workshop on Education Technology and Computer Science, 2010.
- Changua Yang, Kevin Hsin-Yih Lin, Hsin-His Chen, "Writer meets Reader: Emotional Analysis of Social Media from both the Writer's and Reader's perspectives", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Workshops", 2009.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Tamil Online handwriting recognition using fractal features

*Rituraj Kunwar and A G Ramakrishnan*

*MILE Lab, Dept of Electrical Engineering, Indian Institute of Science, Bangalore.*

## Abstract

We present a fractal coding method to recognize online handwritten Tamil characters and propose a novel technique to increase the efficiency in terms of time while coding and decoding. This technique exploits the redundancy in data, thereby achieving better compression and usage of lesser memory. It also reduces the encoding time and causes little distortion during reconstruction. Experiments have been conducted to use these fractal codes to classify the online handwritten Tamil characters from the IWFHR 2006 competition dataset. In one approach, we use coding and decoding process. A recognition accuracy of 90% has been achieved by using DTW for distortion evaluation during classification and encoding processes as compared to 78% using nearest neighbor classifier. In other experiments, we use the fractal code, fractal dimensions and features derived from fractal codes as features in separate classifiers. Whereas the fractal code was successful as a feature, the other two features are not able to capture the wide within-class variations.

## Introduction

Fractal codes are the compressed representation of patterns, based on iterative contractive transformations in metric spaces proposed by Barnsley. A simplified version of the fractal block coding technique for digital images has been used to encode the 1-D ordered online handwritten character patterns. A novel partitioning algorithm has been proposed to reduce the computation complexity of encoding and decoding, with a minor fall in recognition accuracy.

## Building fractal codes for handwritten characters

We need to find the collection of affine transforms of the online handwritten character. The raw online handwritten character is first preprocessed using three steps: (i) smoothing (ii) re-sampling the variable number of points in each character to 60 points. (iii) normalizing the x and y coordinates between 0 and 1. The handwritten character locus is divided into non-overlapping range segments. Each range segment has a fixed number of points ( $R$ ) in it. Last point of each range becomes the first point of the next range, except in the case of the last range.

## Creating a pool of domain segments

The domain pool is formed for each character locus. Domain pool is the collection of all possible domain segments. The number of points in each domain segment is chosen to be double that in each range segment,  $D = 2R$ . Domain pool can be obtained by sliding the window containing  $D$  points at a time. The window is first located at the beginning of the stroke. The window is moved along the

stroke by  $\delta$  points, in such a way that it does not cross the end point of the stroke. The step  $\delta$  has been chosen as  $R/2$  in our experiments.

## Constructing transformed Domain pool

Transformed domain pool is constructed by multiplying each domain segment with the eight isometries that involve reflection and rotation about different axes. To begin with, each domain is translated to its centroid and scaled down by the contractivity factor ( $s=0.5$ ). The following transformations are then applied to each of the candidate domain segment.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

The above transformations produce a whole family of geometrically related domain segments. In domain pool, matching blocks will be looked for encoding the online handwritten character.

**Searching for the most similar domain segment for each range:** Each affine transformed domain segment is re-sampled into  $R$  points and then its centroid is translated to that of the concerned range segment. Distance between them is found. Similarly, distances w.r.t to the all the domain segments is calculated. The most similar domain segment corresponding to each range segment is identified and fractal code is stored corresponding to the particular range. The fractal codes are similarly obtained for all the online handwritten characters.

Fractal codes corresponding to each range segment consists of (1) the range segment index, (2) the range segment centroid, (3) index of the most similar domain segment and (4) the index of transformation used out of the 8 transformations.

### Issue related to constructing fractal codes

The whole character is divided into range segments of equal number of points. Smaller the number of points in each range segment, the more minutely we can capture the complexity in any region of the character. The number of range segments per character is thus inversely proportional to the number of points in each range. Again, the encoding speed is inversely proportional to the number of range segments per character. It has been noted that there are certain region in a character where the curliness is minimal so in those area the range segment size could be increased still encoding the region precisely.

**Steps to encode a handwritten character where the number of points in each range is variable:**

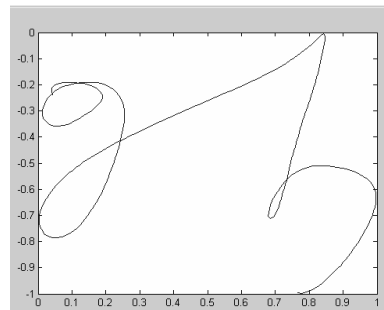


Fig 1. Tamil handwritten character 'aa'

Cumulative angle ' $\theta_c$ ' is calculated starting from the first point and traversing the character stroke till it crosses an empirically set threshold of  $\theta_T$ . Smaller the threshold, finer is the encoding. Figure 1 shows a sample of the handwritten character /aa/ in Tamil. Figure 2 shows the effect of the choice of the angle threshold on the reconstruction error.

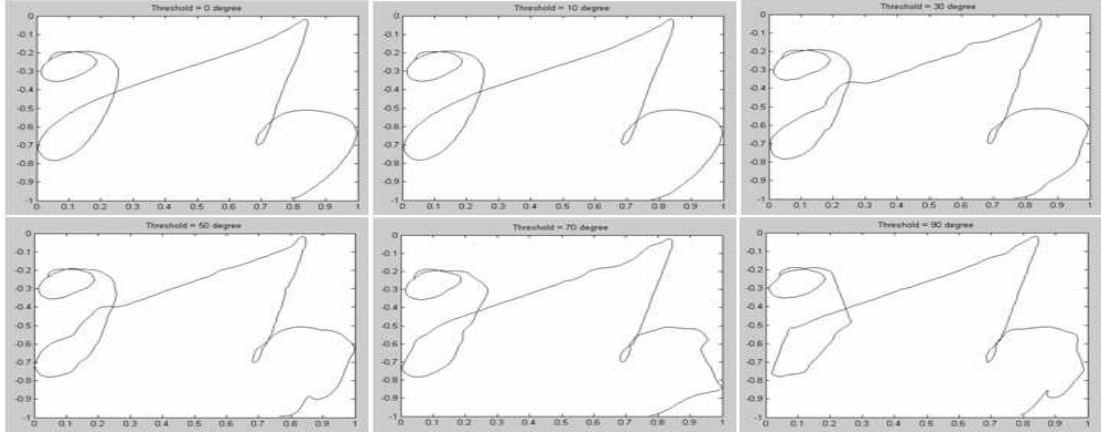


Fig 2. Illustration of the distortion in reconstruction with different angle thresholds  $\phi_T$ . Here reconstruction is performed from the fractal codes of the character /aa/ shown in Fig 1. For encoding, different values of range sizes are used, namely, 4, 8, 12, 16 and 20.

#### Algorithm for partitioning

1. Domain pool of different sizes (namely 8, 16, 24, 32, 40) are constructed corresponding to the range sizes of 4, 8, 12, 16 and 20. By size, we mean the number of points in each domain.
2. Start from the first point and move along the character from one point to the next and calculate the cumulative change in angle  $\theta_c$ .
3. The No. of points (K) till the point penultimate to the one, where  $\theta_c$  crosses the threshold  $\phi_T$  is noted.
4. The range size closest to and less than K is chosen. The most suitable domain is chosen from the corresponding domain pool and the fractal codes are stored.
5. The last point of the present range is then considered as the first point of the new range and the process repeats starting from step 2.

Along with the fractal codes, the size of the range chosen is also stored. If the end point is reached with  $\theta_c < \phi_T$ , then step 4 is followed, where K includes the last point also since  $\theta_c < \phi_T$ . If at the end, few points are left which is less than the smallest range, they are discarded else step 4 is repeated.

**Algorithm for reconstruction:** Banach's contractive mapping theorem states: "If a contractive mapping 'W' (which are the fractal codes here) is defined, then iterative application of the mapping on any sequence of the same space will lead to a Cauchy's sequence which will converge to a fixed, unique point.

#### CASE I: Range having fixed number of points

A random initial pattern having the same number of points is taken or generated. A domain pool of size double that of the range is created in a manner similar to the encoding process. First fractal code is

taken corresponding to the first range of the pattern, and operations are performed on the corresponding domain indicated by the domain index in the code of first range. The indicated domain segment's origin is shifted to its origin and then it is scaled down by the contractivity factor (0.5 here). Then the affine transformation as indicated in the code is applied on the scaled domain segment. Finally the transformed domain's centroid is shifted to the range segment centroid as present in the fractal code. The above steps are repeated to decode all the range segments. Then the whole decoded locus is smoothened. The above steps are repeated till the termination condition is satisfied to finally converge to a fixed and unique pattern. Termination condition could be (i) an empirically set fixed number of iterations, sufficient for convergence or (ii) minimal or no distortion between two consecutive patterns produced by 2 consecutive iterations.

## CASE II: Range having variable number of points in each range

In this case, multiple domain pools are created out of the random pattern taken for the reconstruction. Using the extra information given in the fractal code pertaining to the range size to be chosen so that domain segment is picked up from the appropriate domain pool. The rest of the steps are same as the case for the reconstruction with fixed range size. Using above two methods, fractal codes of any give pattern can be created and the same pattern could be decoded using any random pattern after applying this reconstruction algorithm iteratively for few times.

### 1. Classification by using fractal codes in construction and reconstruction:

The above fractal encoding and decoding method has been used for classification of characters. The assumption behind this classification is that if a sample of a class (say /a/) is encoded and fractal codes are obtained. The following process of reconstruction if started in 2 ways i.e. firstly by applying the reconstruction algorithm iteratively on a random pattern of any different class (anything other than 'A') and secondly doing the same on any random pattern of the same class (i.e. 'A') then the distortion between the initial pattern and the pattern obtained after first iteration of reconstruction is relatively much more in the first case than in the second. The reason behind this is that reconstruction process leads to convergence to the pattern whose code is used for reconstruction. And since the class of the initial pattern in the second case and the fractal code is same, the distortion in the second case is smaller than the first case.

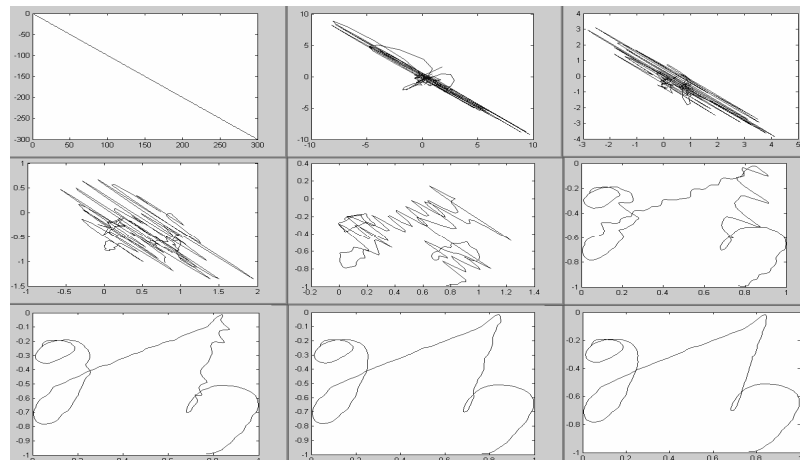




Fig 3. In the above image, reconstruction process is shown which starts from a random straight line and finally converges to a pattern which is very close to the original pattern after 8 iterations. The original pattern (in Fig 1) was encoded using different range sizes of 4, 8, 12, 16 and 20 with the threshold angle of 30 degrees.

### Character classification using fractal codes:

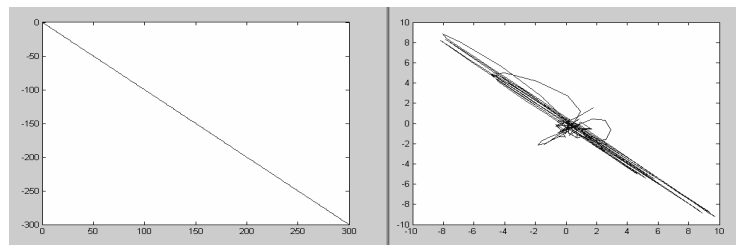


Fig. 4. The above image shows the distortion created, when an iteration of reconstruction was performed. This shows that the distortion is huge if the starting pattern (above left) is very different from the original pattern, whose fractal codes are used for reconstruction (in this case Fig 1).

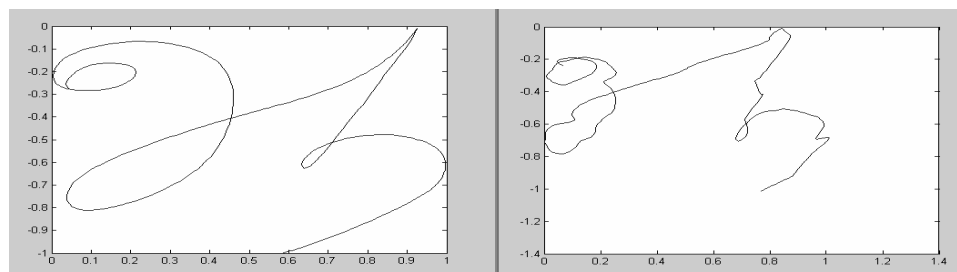


Fig. 5. The above image shows the distortion created, when an iteration of reconstruction was performed. This shows that the distortion is much less if the starting pattern (above left) is not very different from the original pattern, whose fractal codes are used for reconstruction (in this case Fig 1).

### Classification Algorithm

In the present research, fractal codes of 'N' samples of entire 156 classes are computed and stored. To classify a test sample, following steps are taken.

- a) An iteration of the reconstruction algorithm is applied on the test sample using the fractal code of each sample of each class.
- b) The distortion 'D' is calculated between the initial pattern and the pattern obtained after one step of reconstruction. Thus the distortion matrix of size  $156 \times N$  is obtained.
- c) The row number of the minimum value of the distortion is found from the distortion matrix and assigned to the test sample.

### Distortion evaluation:

The distortion between two patterns can be evaluated by finding the distance between them. The distance between 2 patterns is measured by Nearest Neighbor (NN) method. The issue with NN is that the matching is done point by point which increases the distance unusually and thus decreases the classification accuracy (evident from the result table). This drawback in distance evaluation of NN

is addressed by DTW pattern matching which is more intuitive. This intuitive matching takes place because DTW matches similar subsections of the patterns thus producing a reasonable distance between patterns. This method hence produces a remarkable increase in accuracy as shown Table 1.

### **Classification using the fractal codes as a features in the Nearest Neighbor (NN)**

In this method, the fractal codes are used as features and fed into a NN classifier. An accuracy of approximately 65% is obtained.

### **Classification using fractal dimension or features derived from the fractal codes**

Fractal dimension is a unique identity of any pattern or object. However, because of the mere nature of handwritten character recognition (i.e. large variation within every class), it fails completely to classify any random sample. The features derived from the fractal codes like MMVA and DRCLM, though successful in problems like face recognition and signature verification, fail in recognizing handwritten characters.

## **Results**

Table 1: Results show how the DTW comparison during reconstruction impacts the recognition accuracy

| Fixed Range Size | DTW used? | No. of Training samples | No. of Testing samples | Accuracy (in %) |
|------------------|-----------|-------------------------|------------------------|-----------------|
| 4                | No        | 20                      | 50                     | 78.9            |
| 4                | Yes       | 20                      | 50                     | 90.4            |

Table 2: Results show the efficacy of the Partitioning algorithm in improving the efficiency of the recognition system (in terms of time) with marginal drop in accuracy. Variable range sizes used (4, 8, 16, 24 and 32). No. of training and testing samples used are 5 and 30, respectively, No. of classes used is 156.

| Threshold angle (degree) | DTW used for distortion evaluation? | Encoding time per sample (sec) | Accuracy (in %) |
|--------------------------|-------------------------------------|--------------------------------|-----------------|
| 0                        | Yes                                 | 31                             | 90.44           |
| 10                       | Yes                                 | 22                             | 86.43           |
| 30                       | Yes                                 | 12                             | 85.04           |
| 50                       | Yes                                 | 10                             | 83.55           |
| 70                       | Yes                                 | 8                              | 81.60           |
| 90                       | Yes                                 | 6                              | 80.87           |

**Acknowledgment:** The authors thank Technology Development for Indian Languages (TDIL), DIT, Government of India for funding this research, as part of the research consortium on Online handwriting recognition of Indian languages.

## **References**

- M. F. Barnsley, Fractals everywhere, New York: Academic, 1988.
- T. Tan and H. Yan, Face recognition by fractal transformations, IEEE ICASSP, 6:3405-3408, 1999.
- Mozaffari S., Faez K. and Faradji F, One Dimensional Fractal Coder for Online Signature Recognition, IEEE ICPR, 2:857- 860, 2006.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Neuroscience inspired segmentation of handwritten words

*A G Ramakrishnan and Suresh Sundaram*

*MILE Lab, Dept of Electrical Engineering, Indian Institute of Science, Bangalore.*

The challenge of segmenting online handwritten Tamil words has hardly been investigated. In this paper, we report a neuroscience-inspired, lexicon-free approach to segment Tamil words into its constituent symbols (recognizable entities). Based on a simple dominant overlap criterion, the word is grossly segmented into candidate symbols (stroke groups). However, this segmentation is not fully reliable because of varying writing styles resulting in varying levels of overlap. Taking cues from vertebrate visual perception, we utilize both feature based attention and feedback from the classifier to detect possible wrong segmentations. This attention-feedback segmentation (AFS) strategy splits or merges the stroke groups to correct the segmentation errors and forms valid symbols. This maiden attempt on segmentation is tested on 10000 handwritten words collected from hundreds of writers. The efficacy of AFS in segmentation and improving the recognition performance of the handwriting system is amply demonstrated. Our results show a segmentation accuracy of over 99% at symbol level.

## Need for segmenting handwritten words

Since attempts to segment cursively handwritten English words have largely failed, researchers working on Indic scripts too feel that it is not advisable to try to segment individual characters from handwritten documents. However, we firmly believe that it is not only possible, but also something that ought to be done, if one is interested in recognizing words such as proper names appearing in the name and address fields of handwritten forms. Thus, this opens up the possibility of developing a recognizer that can handle unrestricted vocabulary, including any unusual word of foreign origin, such as names of people or places from other countries. Thus, we believe that our work is the first of its kind in proposing an approach for handwriting recognition that does not limit the writer from writing any text of any origin.

## Motivation for Attention-Feedback Segmentation

Traditional pattern recognition [1, 3-5, 8, 10, 13-15] primarily follows a feedforward architecture, whereas the same in mammalian brain involves complex feedback structures. Studies on visual perception in primates demonstrate the effect of attention on the response of the visual neurons [2]. Feature based attention biases the neuronal responses as though the attended stimulus was presented alone. Also, shifting spatial attention from outside to the inside of the receptive field increases the neuronal responses. Motivated by these observations, we incorporate local feature based attention to correct and improve segmentation [9]. Further, studies on visual pathways show extensive feedback from the cortex to the lateral geniculate nucleus (LGN), which have both inhibitory and facilitatory effects on the responses of LGN relay cells. In our work, we use feedback based on features as well as from the classifier posterior probabilities to rectify any incorrect segmentation by regrouping the strokes. Thus, we call our approach as 'attention-feedback' strategy for segmentation.

Further, studies on scene perception by humans [6] indicate that visual processing follows a top-down approach. The global cues characterizing the visual object, that appear within the visual span, are perceived before the local features. The human perceptual system treats every scene as if it were in the process of being focussed or zoomed in on, whereas initially, it is relatively less distinct. Moreover, the human perceptual processor has the capability to select parts of the input stimulus that are worth to be paid attention to. Motivated with these observations from the field of neuroscience, we present a segmentation strategy that first works on the global feature of overlap to output candidate Tamil stroke groups for the given input strokes. By analyzing local features characteristic to the given input pattern, we reevaluate the segmentation and modify the segmentation when found necessary. The localized features are derived by zooming on paying attention to specific parts of the online trace. Essentially, we adopt a multi-pass system, wherein fine grained processing is guided by the prior cursory (global) processing.

### **Data used for the study**

The 155 distinct Tamil symbols (comprising 11 vowels, 23 base consonants, 23 pure consonants, 92 CV combinations and 6 additional symbols) are presented in Appendix A. The publicly available corpus of isolated Tamil symbols (IWFHR database) is used for learning various statistics about Tamil symbols. The primary focus of this work is to address the challenges of segmentation. Towards this purpose, Tamil words are collected using a custom application running on a tablet PC and saved using a XML standard [7]. High school students from across 6 educational institutions in Tamil Nadu contributed in building the word data-base of 100, 000 words, referred to as the ‘MILE Word Database’ in this work [12]. Out of these, 10,000 words are used for this study. The words have been divided into 40 sets, each comprising 250 words. Owing to the comparable resolution of our input device to that used in the IWFHR dataset, statistical analysis performed on the symbols in the IWFHR database are applicable to the Tamil symbols in the MILE word database.

### **Dominant Overlap Criterion Segmentation**

An online word can be represented as a sequence of  $n$  strokes  $W = \{s_1, s_2, \dots, s_n\}$ . In the case of multi-stroke Tamil symbols, strokes of the same symbol may significantly overlap in the horizontal direction. The word is first grossly segmented based on a bounding box overlap criterion, generating a set of stroke groups. In this ‘Dominant Overlap Criterion Segmentation’ (DOCS), the heavily overlapped strokes are merged. A stroke group is defined as a set of consecutive strokes merged by the DOCS step, which is possibly a valid Tamil symbol.

For the  $k$ -th stroke group  $S_k$  under consideration, its successive stroke is taken and checked for possible overlap. Significant overlap necessitates the successive stroke to be merged with the stroke group  $S_k$ . Otherwise, the successive stroke is considered to begin a new stroke group  $S_{k+1}$ . The algorithm proceeds till all the strokes of the word are exhausted.

### **Neuroscience-inspired segmentation**

The stroke groups obtained from the above dominant overlap criterion segmentation are preprocessed by smoothing, normalization and resampling into standard number of equi-arc length spaced points.

The x and y coordinates of these processed stroke groups and their first and second derivatives are used as features for recognition using a support vector machine (SVM) classifier that outputs class labels and their posterior probabilities. Obviously, DOCS being simple, does not always result in correct segmentation. Sometimes it results in over segmentation of a single multi-stroke character into two stroke groups; other times, two distinct characters get combined into a single stroke group, due to the way they are written.

## Attention Features

Figure 1 shows the complete block schematic of the proposed segmentation scheme. An over-segmented symbol is usually small and hence results in low aspect ratio as well as has very few dominant points (points where the curvature is high). By paying attention to these features extracted from the stroke groups output by DOCS block, one can suspect wrong segmentation. Further, the symbols that result from over- or under-segmentation are classes that the classifier has not come across. Thus, these symbols usually result in a low confidence level of the classifier. Thus, the posterior probability of the classifier, when fed back to the input stages, can be used to invoke the computation of the attention features. The feedback, together with the attention features suggest possible resegmentation of the input strokes, resulting in new possible stroke groups. These modified stroke groups based on merger or splitting of original stroke groups, are once again recognized by the classifier after preprocessing and extraction of recognition features. An improved posterior probability of the new stroke group confirms right segmentation. Thus, the refinement in segmentation is caused based on memory, attention and feedback mechanisms prevalent in human perception. We call this as “attention-feedback segmentation (AFS)”.

## Commonly found segmentation issues

The two Tamil characters that ought to have a minimum of three strokes are the long /i/ (nedil) and the aydam. Since in both of these cases, in general there is no overlap between the final dot and the rest of the character, they always are over segmented into two or more stroke groups.

Pure consonants (*mey ezhuthu*), when they are written with the dot (*pulli*) beyond the base consonant, result in over segmentation too.

Characters such as /ka/, /nga/ and /ra/, which start with an initial vertical segment, are written by many with multiple strokes, with the first stroke being a simple down-going vertical line. These characters have a potential to be over segmented, if the following part of the character does not clearly overlap with the vertical line.

All CV combinations of /i/ and /I/ and the CV combinations of /u/ and /U/ with borrowed consonants such as /ja/ and /sha/ also have a tendency to be over-segmented, if the vowel matra is written with no horizontal overlap with the consonant.

Under segmentation occurs if the ending part (usually bottom extensions of /ta/ or /Ra/) of the following character goes far left below the previous character, causing significant horizontal overlap between them. At other times, people write two successive characters so closely, that there is significant overlap between them.

Naturally, in all the above cases, the simple segmentation (DOCS) is likely to result in wrong segmentation leading to erroneous recognition results.

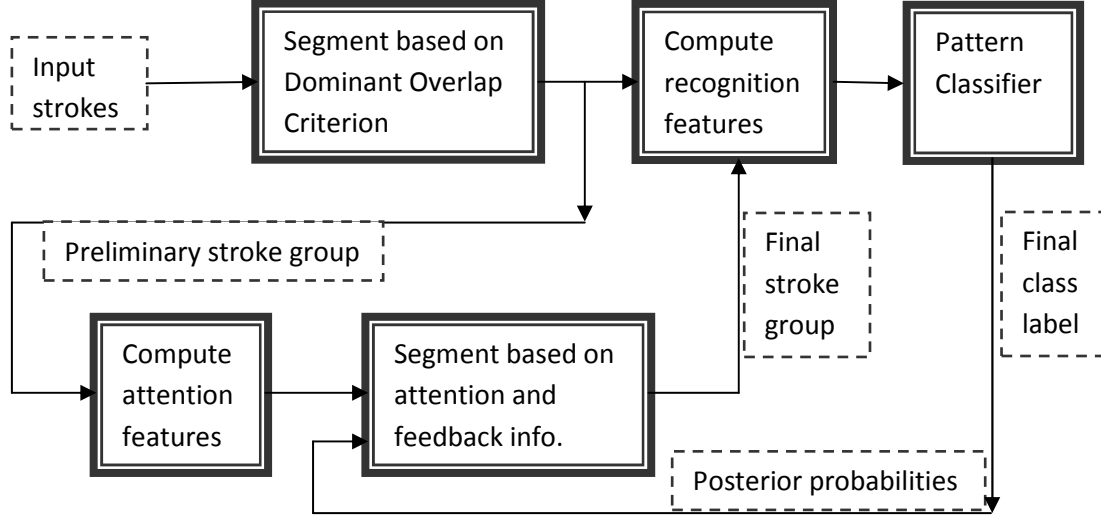


Fig. 1. Block diagram of neuroscience inspired segmentation of Tamil handwritten word [9].

## Segmentation results on the MILE Tamil Word Database

The proposed techniques are tested on the subset of 10,000 words. However, to start with, we evaluate the performance on a set of 250 words (denoted as DB1), that has a significant number of errors resulting from the DOCS. Of the 103 errors, 89 (or 86%) correspond to the merging of valid symbols, and the rest, to broken symbols. The AFS module aids in properly detecting and correcting 91 (or 90%) of these errors. In addition, the methods proposed effectively merge 11 (or 78%) of the over-segmented stroke groups to valid symbols. The improvement in character segmentation rate in turn reduces the number of wrongly segmented words. It is observed that only 7 of the total 250 words remain wrongly segmented after the AFS scheme, as against 67 words after the DOCS scheme. On evaluating the performance across the database of 10000 words, we obtain a 86% reduction in character segmentation errors.

## Recognition results on the MILE Database

We report experimental results demonstrating the impact of the proposed AFS strategy on the recognition of symbols in the MILE word database. Since a significant percentage of DOCS errors are corrected by AFS, a drastic improvement of 16% (from 70.5 % to 87.1 %) in symbol recognition is observed. In computing the symbol recognition rate, apart from the substitution errors, we take into account the insertion and deletion errors, caused by over-segmentation and under-segmentation, respectively. The edit distance is used for matching the recognized symbols with the ground truth data. Moreover, 11.6 % of the words, (29 additional words) wrongly recognized after DOCS, have

been corrected by the proposed technique. Across the 10000 words in the MILE Word database, an improvement of 4% (from 83 to 87%) in symbol recognition rate has been obtained.

## Conclusion

In this paper, we present a maiden attempt based on significant feedback from the classifier to the input blocks such as feature extraction and segmentation, as well as the use of memory (prior knowledge) to result in a very effective segmentation of online handwritten Tamil words. This approach being general, can be extended to any other Dravidian script, as well as any other script where cursive writing is not practiced. To our knowledge, there is no reported systematic research work on segmenting the individual characters or recognizable standard symbols from online handwritten words for any Indic language that does not have the *shiro rekha* (head line). Thus, we are unable to compare the performance of our work with any other technique. However, the results are promising and have also led to improved recognition of the handwritten words [16], thus confirming the possibility of proper segmentation of online Tamil words. We intend to extend this work very soon to online Kannada handwritten words.

## Acknowledgment

We thank Ms. Nethra Nayak, Mr. Rituraj Kunwar, Ms. Archana C P, Mr. Shashi Kiran, Ms. Chandrakala, Mrs. Shanthi Devaraj, Ms. Saranya and Ms. Sountheriya for their efforts in data collection and annotation, which made these experiments possible. We thank Dr. Arun Sripathi of the Centre for Neuroscience, IISc for the very useful discussions we had on visual perception. Special thanks to Technology Development for Indian Languages (TDIL), Department of Information Technology (DIT), Government of India for funding this research, as part of a research consortium on online handwriting recognition in several Indic scripts. We thank Prof. Deivasundram (University of Madras), AVM Matriculation Higher Secondary School Virugambakkam, Chennai, Govt. Boys Higher Secondary School, Sulur, Presidency College, Triplicane, Chennai and IIT Madras for contributing to the data set. We also thank Dr. Anoop Namboodiri for giving us the word level annotation tool. We thank CDAC, Pune for being a partner in finalizing the XML standard for handwritten data collection in Indic languages.

## References

- N Joshi, G Sita, A G Ramakrishnan, S Madhavanath, "Comparison of elastic matching algorithms for online Tamil handwritten character recognition", Proc. IWFHR (2004) 444-449.
- G M Boynton, "Attention and visual perception. Current Opinion in Neurobiology", (15) (2005) 465-469.
- H Swethalakshmi, C Chandra Sekhar, V S Chakravarthy, "Spatiostructural features for recognition of online handwritten characters in Devanagari and Tamil scripts", ICANN (2) (2007) 230-239.
- A Bharath, S Madhvanath, "Hidden markov models for online handwritten Tamil word recognition", Proc. ICDAR (2007) 506-510.
- Amrik Sen, G. Ananthakrishnan, Suresh Sundaram, A. G. Ramakrishnan, "Dynamic space warping of strokes for recognition of online handwritten characters. IJPRAI (2009) 23(5): 925-943.



- Arun P Sripathi and Carl R Olson, "Representing the forest before the trees: a global advantage effect in monkey inferotemporal cortex", *The Journal of Neuroscience*, June 17, 2009, 29(24):7788-7796.
- Swapnil Belhe, Srinivasa Chakravarthy, A. G. Ramakrishnan, XML standard for Indic online handwritten database, *ACM - Proceedings of the International Workshop on Multilingual OCR*, 2009.
- M. Mahadeva Prasad, M. Sukumar, A. G. Ramakrishnan, Divide and conquer technique in online handwritten Kannada character recognition, *ACM - Proc International Workshop on Multilingual OCR*, 2009.
- Suresh Sundaram and A G Ramakrishnan, "Verification based segmentation approach for online words", *Indian Patent Office Ref. No. 03974/CHE/2010*.
- Shashi Kiran, Kolli Sai Prasada, Rituraj Kunwar, A. G. Ramakrishnan, "Comparison of HMM and SDTW for Tamil handwritten character recognition", *Proc. 2010 IEEE International Conf Signal Processing & Communication*.
- Rituraj Kunwar, Mohan P., Shashi Kiran, A. G. Ramakrishnan, "Unrestricted Kannada online handwritten akshara recognition using SDTW, *Proc. 2010 IEEE International Conf Signal Processing & Communication*.
- B Nethravathi, C P Archana, K Shashikiran, A G Ramakrishnan, V Kumar, "Creation of a huge annotated database for Tamil and Kannada OHR", *Proc. IWFHR (2010)* 415-420.
- M. Mahadeva Prasad, M. Sukumar, A. G. Ramakrishnan, "Orthogonal LDA in PCA transformed subspace", *Proc. 12th International Conf Frontiers in Handwriting Recognition (ICFHR 2010)*, Nov 2010.
- Venkatesh N, A G Ramakrishnan, "Choice of classifiers in hierarchical recognition of online handwritten Kannada and Tamil aksharas", *Jl. Universal Computer Science*, 2011, Vol. 17, No. 1, pp. 94-106.
- Rakesh R, A G Ramakrishnan, "Fusion of complementary online and offline strategies for recognition of handwritten Kannada characters", *Journal of Universal Computer Science*, 2011, Vol. 17 (1), pp. 81-93.
- Suresh Sundaram and A G Ramakrishnan, "Attention-feedback based robust segmentation of online Tamil words", under review, *Pattern Recognition*, 2011.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Improving Tamil-English Cross-Language Information Retrieval by Transliteration Generation and Mining

*A Kumaran, K Saravanan & Ragavendra Udupa*

*Multilingual Systems Research*

*Microsoft Research India*

*Bangalore, India.*

*{kumarana, v-sarak, raghavu}@microsoft.com*

## **Abstract**

While state of the art Cross-Language Information Retrieval (CLIR) systems are reasonably accurate and largely robust, they typically make mistakes in handling proper or common nouns. Such terms suffer from compounding of errors during the query translation phase, and during the document retrieval phase. In this paper, we propose two techniques, specifically, transliteration generation and mining, to effectively handle such query terms that may occur in their transliterated form in the target corpus. Transliteration generation approach generates the possible transliteration equivalents for the out of vocabulary (OOV) terms during the query translation phase. The mining approach mines potential transliteration equivalents for the OOV terms, from the first-pass retrieval from the target corpus, for a final retrieval. An implementation of such an integrated system achieved the peak retrieval performance of a MAP of 0.5133 in the monolingual English-English task, and 0.4145 in the Tamil-English task. The Tamil-English cross-language retrieval performance improved from 75% to 81% of the English-English monolingual retrieval performance, underscoring the effectiveness of the integrated CLIR system in enhancing the performance of the CLIR system.

## **1. Introduction**

With the exponential growth of non-English population in the Internet over the last two decades, Cross-Language Information Retrieval (CLIR) has gained importance as a research discipline and as an end-user technology. While the core CLIR system is fairly robust and accurate with sufficient training data, it's handling of proper or common nouns (or, more generally, those query terms that could occur in their transliterated form in the target corpus in cross language environments) is far from desirable in most implementations. In essence, the name translations are not typically part of translation lexicons used for query translations, and hence do not get translated properly in the target language. Note that, from the CLIR point of view, any un-translated word is an out-of-vocabulary word, which typically include words that are literally transliterated into local language words (such as, *computer*, *corporation*, etc.), specifically in those countries where English is spoken as a second language. This phenomenon is called as *code-mixing*. Such words are not available in the translation lexicon, but are typically part of the local language corpora. Also, given that a name may be spelled differently in the target corpus – particularly for those names that are not native to the target language – the retrieval performance suffers further, as the errors in query translation and spelling variations compound. Given that proper and common nouns form a significant portion of query terms, it is

critical that such query terms are handled effectively in a CLIR system. This is precisely the research theme that we explore in this paper.

Evaluation of Tamil-English cross-language information retrieval systems started first in the Forum for Information Retrieval Evaluation (FIRE) [1], modeled after the highly successful CLEF [2] and NTCIR [3] campaigns. In 2010, FIRE organized several ad hoc monolingual and cross-language retrieval tracks, and we participated in the English monolingual and cross-language Hindi-English and Tamil-English ad hoc retrieval tracks. This paper presents the details of our participation, specifically, in the Tamil-English cross-language tasks and present the performance of our official runs.

## 2. Cross-language Retrieval System

In this section, we outline the various components of our CLIR system integrated with the two techniques for handling OOV words.

Our monolingual retrieval system is based on the well-known Language Modeling framework to information retrieval. In this framework, the queries as well the documents are viewed as probability distributions. The similarity of a query with a document is measured in terms of the likelihood of the query under the document language model. We refer interested readers to [6, 7] for the retrieval model and the details of this framework. In our CLIR model, the query in a source language is translated into the target language – English – using a probabilistic translation lexicon, learnt from a given parallel corpora of about 50,000 parallel sentences between English and Tamil. Such learnt translation dictionary included ~107 K Tamil words and ~45 K English words. From this dictionary, we used only top 4 translations for every source word, an empirically determined limit to avoid generation of noisy terms in the query translations.

Like any cross-language system that makes use of a translation lexicon, we too faced the problem of out of vocabulary (OOV) query terms. To handle these OOV terms, we used two different techniques, (i) generation of transliteration equivalents, and (ii) mining of transliteration equivalents:

- In Transliteration Generation, the transliterations of the OOV terms in the target language are generated using an automatic Machine Transliteration system, and used for augmenting the query in the target language.
- In Transliteration Mining, the transliteration equivalents of the OOV terms are mined from the top-retrieved documents from the first pass, which are subsequently used in the query for a final retrieval [8].

### 2.4 Generating Transliterations

We adopted a conditional random fields based approach using purely orthographic features, as a systematic comparison of the various transliteration systems in the NEWS-2009 workshop [9] showed conclusively that orthography based discriminative models performed the best among all competing systems and approaches. In addition, since the Indian languages share many characteristics among them, such as distinct orthographic representation for different variations – aspirated or un-aspirated, voiced or voiceless, etc. – of many consonants, we introduced a word origin detection module (trained with about 3000 hand-classified training set) to identify specifically Indian origin names. All other

names are transliterated through an engine that is trained on non-Indian origin names. Manual verification showed that this method about 97% accurate. We used CRF++, an open source implementation of CRF model, trained on about 15,000 parallel names between English and Tamil. The transliteration engine was trained on a rich feature set (aligned characters in each direction within a distance of 2 and source and target bigrams and trigrams) generated from this character-aligned data.

## 2.5 Mining Transliteration Equivalents

The mining algorithm issues the translated query minus OOV terms to the target language information retrieval system and mines transliterations of the OOV terms from the top results of the first-pass retrieval. Hence, in the first pass, each query-result pair is viewed as a “comparable” document pair, assuming that the retrieval brought in a reasonably good quality results set based on the translated query without the OOV terms. The mining algorithm hypothesizes a match between an OOV query term and a document term in the “comparable” document pair and employs a transliteration similarity model to decide whether the document term is a transliteration of the query term. Transliterations mined in this manner are then used to retranslate the query and issued again, for the final retrieval. The details of transliteration similarity model may be found in [6, 7], and the details of our training are given in [8].

## 3. Experimental Setup & Results

In this section, we specify all the data used in our experiments, both that were released for the CLIR experiments for FIRE task, and that used for training our CLIR system.

### 3.1 FIRE Data

The English document collection provided by FIRE was used in all our runs. The English document collection consists of ~124,000 news articles from “The Telegraph India” from 2004-07. All the English documents were stemmed. Totally 50 topics were provided in each of the languages, each topic having a title (T), description (D) and narrative (N), successively expanding the scope of the query. Table 1 shows a typical topic in Tamil, and the TDN components of the topic, for which relevant English documents are to be retrieved from the aforementioned English news corpus. It should be noted that FIRE has also released a set of 50 English (i.e., target language) topics, equivalent to each of the source language topics..

**Table 1.** A Typical FIRE Topic in Tamil.

| Type        | Topic   |
|-------------|---|
| Title       | குட்கா தயாரிப்பாளர்களுடனான தாதாக்களின் மறைமுகத் தொடர்பு.  |
| Description | கோவா மற்றும் மாணிக்கசுந்த் குட்கா தயாரிப்பாளர்களுடனான தாலுத் இப்ராஹிமினின் மறைமுகத் தொடர்பு.  |
| Narration   | கோவா மற்றும் மாணிக்கசுந்த் குட்கா தயாரிப்பாளர்களுடன் பிரபல தாதா தாலுத் இப்ராஹிமின் மறைமுகத் தொடர்பு பற்றிய செய்திககள் இந்த ஆவணத்தில் இடம்பெறலாம். தாலுத் இப்ராஹிமின் மற்ற தயாரிப்பாளர்களுடனான செய்திகள் இதில் இடம்பெறத் தேவையில்லை. |

**Table 2.** A Typical FIRE Topic in English.

| Type        | Topic   |
|-------------|---|
| Title       | Links between Gutkha manufacturers and the underworld.  |
| Description | Links between the Goa and Manikchand Gutkha manufacturing companies and Dawood Ibrahim.   |
| Narration   | A relevant document should contain information about the links between the owners of the<br>Manikchand Gutkha and Goa Gutkha companies and Dawood Ibrahim, the gangster. Information<br>about links between Dawood Ibrahim and other companies is not relevant. |

### 3.2 Metrics & Performance Data

The standard measures for evaluating our tasks were used, specifically, Mean Average Precision (MAP) and Precision at top-10 (P@10). As shown in Table 1, each of the 50 topics in Tamil has a title (T), description (D) and narrative (N), successively expanding the scope of the query. We ran our experiments taking progressively each of (title), (title and description), and (title, description and narrative), calibrating the cross-language retrieval performance at each stage, to explore whether expanding the query adds useful information for retrieval or just noise. Table 3 shows the notation used in our description of various configurations to interpret the results presented in Tables 4 and 5.

**Table 3.** Notations used

|                |   |
|----------------|---|
| T/TD/TDN       | Title/ Title and Description / Title, Description and Narration |
| M              | Transliteration Mining  |
| G <sub>D</sub> | Transliteration Generation                                      |

Tables 4 and 5 show the results of our monolingual as well as cross-language official runs submitted to FIRE 2010 shared task. The format of the run ids in the results table is ‘Source-Target-Data-Technique’, where ‘Data’ indicates the data used for topic, and is one of {T, TD, TDN} and ‘Technique’ indicates the technique and from the set {M, G<sub>D</sub>, G<sub>T</sub>, M+G<sub>D</sub>, M+G<sub>T</sub>}. The ‘+’ refers to the combination of more than one approach. The symbols double star (\*\*) and single star (\*) indicate statistically significant differences with 95% and 90% confidence respectively according to the paired t-test over the baseline. The best results achieved are highlighted in bold.

### 4.4 Monolingual English Retrieval

We submitted 3 official runs for the English monolingual track, as shown in the Table 4. For these runs, the English topics provided by the FIRE 2010 organizers were used.

**Table 4.** English Monolingual Retrieval Performance

| Run                 | MAP           | P@10         |
|---------------------|---------------|--------------|
| English-English-T   | 0.3653        | 0.344        |
| English-English-TD  | 0.4571        | 0.406        |
| English-English-TDN | <b>0.5133</b> | <b>0.462</b> |

With the full topic (TDN), our system achieved a peak MAP score of 0.5133. Generally this performance is thought to be the upper bound for cross-language performance, presented in Table 5.

#### 4.6 Tamil-English Cross-Language Retrieval

We submitted totally 12 official Tamil-English cross-language runs, as shown in Table 5. As discussed for the Hindi-English runs, the first run under each of the ‘T’, ‘TD’ and ‘TDN’ sections in Table 5 present the results of the runs without handling the OOV terms, and hence provide a baseline for measuring the incremental performance due to transliteration generation or mining, provided subsequently.

**Table 5.** Tamil-English Cross-Language Retrieval Performance

| Run                                  | MAP             | P@10         |
|--------------------------------------|-----------------|--------------|
| Tamil-English-T                      | 0.2710          | 0.258        |
| Tamil-English-T[G <sub>D</sub> ]     | <b>0.2891*</b>  | <b>0.268</b> |
| Tamil-English-T[M]                   | 0.2815**        | 0.258        |
| Tamil-English-T[M+G <sub>D</sub> ]   | 0.2816*         | 0.268        |
| Tamil-English-TD                     | 0.3439          | 0.346        |
| Tamil-English-TD[G <sub>D</sub> ]    | 0.3548*         | 0.35         |
| Tamil-English-TD[M]                  | <b>0.3621**</b> | 0.346        |
| Tamil-English-TD[M+G <sub>D</sub> ]  | 0.3617**        | <b>0.362</b> |
| Tamil-English-TDN                    | 0.3912          | 0.368        |
| Tamil-English-TDN[G <sub>D</sub> ]   | 0.4068**        | 0.378        |
| Tamil-English-TDN[M]                 | <b>0.4145**</b> | 0.368        |
| Tamil-English-TDN[M+G <sub>D</sub> ] | 0.4139**        | <b>0.394</b> |

From the results presented in Table 5, we observe that the usage of all of the components of the topic, namely T, D and N, produced the best retrieval performance. The basic Tamil-English cross-language run ‘Tamil-English-TDN’ (without transliteration generation or mining), achieved the MAP score

0.3912, and our best cross-language run ‘Tamil-English-TDN[M]’ with mining achieved a MAP score of 0.4145. We observe, in general, similar trends in the other runs that use only the title, or title and description sections of the topics. While the cross-language performance of Tamil-English achieves ~81% of our monolingual English retrieval performance, we observe that this is not as high as the Hindi-English retrieval, perhaps due to the highly agglutinative nature of Tamil.

Given that the mining technique performed generally above the other techniques, we focus on addition of mining to the base CLIR, for subsequent analysis in the following sections.

#### 4.7 Mining OOV terms and its effect on CLIR performance

In this section, we analyze the volume of the OOV terms in FIRE topics, and to what extent they are handled by our mining technique, which clearly emerged as the better technique for boosting the retrieval performance. Also, we show the effect of handling the OOVs on the cross-language retrieval performance, for both the Hindi-English and Tamil-English CLIR runs. Table 6 enumerates the number of unique OOV terms in Tamil FIRE 2010 topics.

**Table 6.** OOV terms in Tamil-English CLIR.

| Topic Config. | OOV terms | Translitteratable<br>OOV terms | Translitteratable<br>OOV terms handled | % of Translitteratable<br>OOV terms handled |
|---------------|-----------|--------------------------------|--|---|
| T             | 24        | 13                             | 5                                      | 38.46                                       |
| TD            | 58        | 29                             | 15                                     | 51.72                                       |
| TDN           | 129       | 47                             | 24                                     | 51.06                                       |

As shown in the third line in Table 6 in the TDN configuration of the 50 Tamil topics of the FIRE2010 shared task, there were totally 129 unique OOV terms, out of which 47 were proper or common nouns, handling of which may help improving the cross-language retrieval performance. Our mining technique did find at least one transliteration equivalent for 24 of these OOV terms (that is, 51%). As shown in Table 5, handling these OOV’s resulted in nearly 6% improvement in the MAP score over the baseline where no OOV’s were handled.

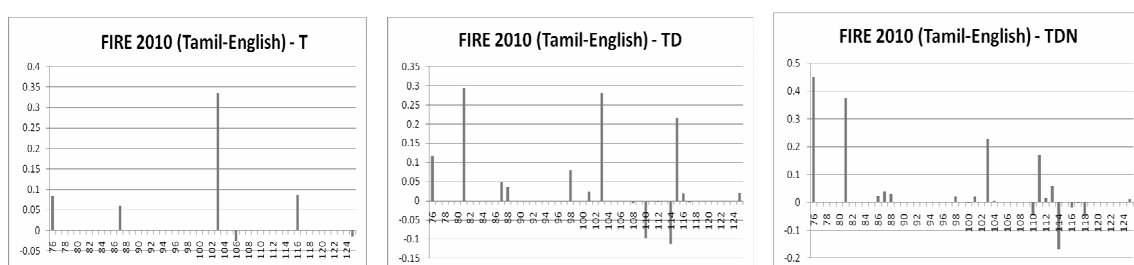
Note that Tamil OOV’s pose specific challenges as outlined below: First, the translitteratable terms mentioned in the third column of the Table 6 excludes some terms whose equivalents are multiword expression in English; mining such multiword transliteration equivalents is beyond the scope of our work and hence they were not handled. Second, 26 out of the 47 terms that are translitteratable were inflected or agglutinated. While our mining algorithm could mine some of them, many of the terms were missed at our parameter settings for mining. By relaxing the mining parameters settings we could mine more such terms, but such relaxation introduced many more noisy terms, affecting the overall retrieval performance. We believe that the use of a good stemmer for inflectional languages like Tamil may help our mining algorithm and, compositionally, the cross-language retrieval performance.



## 4.8 Mining OOV terms and its effect on individual topic performance

In this section, we discuss effect of our approaches on individual topics of the FIRE 2010 shared task, both for Hindi-English and Tamil-English tasks. Figure 1 shows the difference in the Average Precision – topic-wise – between the baseline CLIR system and that integrated with our mining technique. Individual figures provide the differences for each topic, in each of the three configurations T, TD and TDN, for the Tamil-English tasks. We see that many more topics benefitted from the mining technique; for example, in the Tamil-English language pair, in TDN configurations, 11 topics were improved (with 3 of them with an improvement of  $\geq 0.2$  in MAP score) whereas only 4 topics were negatively impacted (all of them dropping  $< 0.2$  in MAP score). Similar trends could be seen for all configurations, in the Tamil-English tasks.

**Fig. 1.** Differences in Average Precision between the baseline and CLIR with mining



From Table 6, we note that the retrieval performance in Tamil-English test collection mining brings maximum improvement of 6% over the baseline in TDN setup.

## 5. Conclusion

In this paper, we underscored the need for handling proper and common nouns for improving the retrieval performance of cross-language information retrieval systems. We proposed and outlined two techniques for handling out of vocabulary (OOV) words – using transliteration generation, and transliteration equivalents mining – to enhance a state of the art baseline CLIR system. We presented the performance of our system under various topic configurations, specifically for English monolingual task and Tamil-English cross-language tasks, on the standard FIRE 2010 dataset. We show that the performance of our baseline CLIR system is improved significantly by each of the two techniques for handling OOV terms, but consistently more so by the mining technique. Significantly, we also show empirically that the performance of the CLIR system enhanced with transliteration mining is close to that of monolingual performance, validating our techniques for handling OOV terms in the cross-language retrieval.

## References

- Forum for Information Retrieval Evaluation. <http://www.isical.ac.in/~fire/>
- The Cross-Language Evaluation Forum (CLEF). <http://clef-campaign.org>
- NTCIR: <http://research.nii.ac.jp/ntcir/>
- Peters, C.: Working Notes for the CLEF 2006 Workshop (2006)

- Majumder, P., Mitra, M., Pal, D., Bandyopadhyay, A., Maiti, S., Mitra, S., Sen, A., Pal, S.: Text collections for FIRE. In: Proceedings of SIGIR (2008)
- Jagarlamudi, J., Kumaran, A.: Cross-Lingual Information Retrieval System for Indian Languages. In: Working Notes for the CLEF 2007 Workshop (2007)
- Udupa, R., Jagarlamudi, J., Saravanan, K.: Microsoft Research India at FIRE2008: Hindi-English Cross-Language Information Retrieval. In: Working notes for Forum for Information Retrieval Evaluation (FIRE) 2008 Workshop (2008)
- Udupa, R., Saravanan, K., Bakalov, A., Bhole, A.: "They Are Out There, If You Know Where to Look": Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval. In: 31th European Conference on IR Research, ECIR (2009)
- Li, H., Kumaran, A., Pervouchine, V., Zhang, M.: Report of NEWS 2009 Machine Transliteration Shared Task. In: ACL 2009 Workshop on Named Entities (NEWS 2009), Association for Computational Linguistics (2009)
- Kumaran, A., Khapra, M., Bhattacharyya, P.: Compositional Machine Transliteration. ACM Transactions on Asian Language Information Processing (TALIP) (2010)





## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# A Package for Learning Negations in Tamil

***Dr. G. Singaravelu***

*Reader, UGC-Academic Staff College, & B.Ed Coordinator*

*Bharathiar University, Coimbatore-641 046.*

## **Introduction**

Tamil is an important language to learn different cultures of Tamilnadu and India. Teaching of Tamil is difficult to the teachers of Tamil due to the more letters in Tamil and learning Tamil grammar is difficult to the learners of primary and upper primary schools due to ineffective methods of teaching. Grammar is indispensable for learning any language. Maximum teaching methods of grammar is adopting formal grammar. Less concentration is transacted in the class room of the Tamil language in functional grammar. Conventional methods of teaching of functional grammar are ineffective and it lead the learners towards aversion in learning grammar. Negations have unique place in communicative competency. Conventional methods discourage the students to learn negation effectively at school level. Students are able to use it inappropriately. This study investigates the effectiveness of learning package of Negations in Tamil among the learners of standard VI.

## **Need of the study**

Conventional methods are unable to create the appropriate learning atmosphere for scoring more marks in Tamil grammar of the mother tongue of the learners and also for the learners of the second language as Tamil. Traditional methods did not help the learners to learn Negations in Tamil. It was a challenging task to the learners of standard VI. An innovative Learning package can be encouraged the young learners to learn more negations in limited time. Hence the researcher endeavoured to prepare a learning package for acquiring more negations in Tamil for the young learners.

## **Objectives**

The researcher has framed the following objectives of the study:

1. To find out the problems of conventional methods in learning Negations in Tamil at Government school.
2. To find out the problems of conventional methods in learning Negations in Tamil at Aided school.
3. To find out the significant difference in achievement mean score between the pre test of control group and the post test of control group in Government school.
4. To find out the significant difference in achievement mean score between the pre test of control group and the post test of control group in Aided school.
5. To find out the significant difference in achievement mean score between the pre test of Experimental group and the post test of Experimental group in Government school.

6. To find out the significant difference in achievement mean score between the pre test of Experimental group and the post test of Experimental group in Aided school.
7. To find out the impact of innovative Learning package in Negations of Tamil at standard VI in Government school and Aided school.

## **Hypotheses**

The research has framed the following hypotheses

1. Students of standard VI have problems of conventional methods in learning Negations in Tamil at Government school.
2. Students of standard VI have problems of conventional methods in learning Negations in Tamil at Aided school.
3. There is no significant difference in achievement mean score between the pre test of control group and the post test of control group in Government school.
4. There is no significant difference in achievement mean score between the pre test of control group and the post test of control group in Aided school.
5. There is no significant difference in achievement mean score between the pre test of Experimental group and the post test of Experimental group in Government school.
6. There is no significant difference in achievement mean score between the pre test of Experimental group and the post test of Experimental group in Aided school.
7. To find out the impact of innovative Learning package in Negations of Tamil at standard VI in Government school and Aided school.

## **Method of study**

**Methodology:** Equivalent group Experimental method was adopted in the study.

### **Sample selected for the study**

Sixty pupils of studying in standard VI from Government Higher Secondary school, Kalveeranpalayam, Coimbatore and another Sixty pupils of studying in standard VI from Maruthamalai Devasdanam Subramanian swamy Higher secondary school, Vadavalli ,Coimbatore were selected as sample for the study. Sixty students were considered as Controlled group and another Sixty were considered as Experimental group.

### **Instrumentation**

Researcher's self-made achievement test was used as a tool for the study.

### **Reliability of the tool**

Test- retest method was used for the study .The co-efficient correlation was found 0.85 in the tool through test-retest method.

### **Validity of the tool**

Content validity was established for the test through expert suggestions. Hence reliability and validity were properly established for the study.

### **Statistical Technique**

Percentage, mean, SD and t test were adopted in the study for analyzing the tabulated data.

### **Procedures of the study:**

Phase 1: Assessing the problems of the students in acquiring competency in learning Tamil Negations for both schools of Govt and Aided in existing methods through administering pretest.

Phase 2 Pre-production stage..

Phase 3: Production stage.

Phase 4: Preparation of package

Phase 5: Execution of activities through using the learning package

Phase 6: Adminstrating pretest and post test to the control group and tabulated the scores.

Phase7: Adminstrating pre test and post test to the Experimental group and tabulated the scores.

Phase 8: Finding the effectiveness of the Package for Negation.

### **Data collection:**

The researcher administered a diagnostic test to identify the problems of the students in learning Tamil with permission of Principals of the schools. Pretest –Treatment-Posttest was used in the study.

### **Hypothesis testing**

#### **Hypothesis 1&2**

1. Students of standard VI have problems of conventional methods in learning Negations in Tamil at Government school.
2. Students of standard VI have problems of conventional methods in learning Negations in Tamil at Aided school.

In the pre-test, students of Govt schools and Aided schools score 19%, 28% marks respectively in acquiring Negation in Tamil through conventional method and the Experimental group students score 49 %, 56% marks respectively..It shows the problems of acquisition of Negation in Tamil through conventional methods among the students.

#### **Hypothesis 3:**

There is no significant difference in achievement mean score between the pre test of control group and the post test of control group in Government school.

| Stages                         | N  | Mean  | S.D. | df | t- value | Result        |
|--------------------------------|----|-------|------|----|----------|---------------|
| <b>Pretest control group</b>   | 30 | 10.63 | 3.23 | 58 | 0.08     | insignificant |
| <b>Post test control group</b> | 30 | 10.78 | 3.21 |    |          |               |

The calculated t value is (0.08) less than table value (1.96). Hence null hypothesis is accepted at 0.05 levels. Hence there is no significant difference between the pre test of control group and post test of control group in achievement mean scores of the teachers in learning Tamil Negations in Govt school.

#### **Hypothesis 4:**

There is no significant difference in achievement mean score between the pre test of control group and the post test of control group in Private school.

| Stages                         | N  | Mean  | S.D. | df | t- value | Result        |
|--------------------------------|----|-------|------|----|----------|---------------|
| <b>Pretest control group</b>   | 30 | 10.53 | 3.23 | 58 | 0.29     | insignificant |
| <b>Post test control group</b> | 30 | 10.28 | 3.28 |    |          |               |

The calculated t value is (1.85) less than table value (1.96). Hence null hypothesis is accepted at 0.05 levels. Hence there is no significant difference between the pre test of control group and post test of control group in achievement mean scores of the teachers in learning Tamil Negations in private school.

#### **Hypothesis 5:**

There is no significant difference in achievement mean score between the pre test of Experimental group and the post test of Experimental group in Government school.

| Stages                              | N  | Mean  | S.D. | df | t- value | Result      |
|-------------------------------------|----|-------|------|----|----------|-------------|
| <b>Pre test Experimental group</b>  | 30 | 10.62 | 3.23 | 58 | 7.14     | significant |
| <b>Post test Experimental group</b> | 30 | 16.56 | 3.21 |    |          |             |



### **Achievement mean scores between pre test of Experimental and posttest of Experimental group.**

The calculated t value is (7.14) greater than table value (1.96). Hence null hypothesis is rejected at 0.05 levels. Hence there is significant difference between the pretest of experimental group and post test of experimental group in achievement mean scores of the students in learning Negation in Tamil.

### **Hypothesis 6:**

There is no significant difference in achievement mean score between the pre test of Experimental group and the post test of Experimental group in Aided school.

| Stages                       | N  | Mean  | S.D. | df | t- value | Level of significance |
|------------------------------|----|-------|------|----|----------|-----------------------|
| Pretest Experimental group   | 30 | 13.70 | 3.30 | 58 | 7.08     | P>0.05                |
| Post test Experimental group | 30 | 19.65 | 3.20 |    |          |                       |

### **Achievement mean scores between pretest of experimental group and posttest of Experimental group.**

The calculated 't' value is (7.08) greater than table value (1.96). Hence null hypothesis is rejected at 0.05 levels. Hence there is significant difference in achievement mean score between the pre test of Experimental group and post test experimental group in achievement mean scores of the students in Tamil Negation.

### **Hypothesis 7.**

### **Learning package is more effective than conventional learning in learning Negation in Tamil**

The above two tables prove and confirm the Learning Package is more effective than traditional approaches in developing Negation in Tamil.. Mean scores in pre-test of Experimental group is (10.62 and 13.70) greater than the mean score of post test of Experimental group by using **Learning Package** in acquiring Negation in Tamil (**16.56 and 19.65**).

### **Findings:**

1. Students of standard VI have problems of conventional methods in learning Negations in Tamil at Government school.
2. Students of standard VI have problems of conventional methods in learning Negations in Tamil at Aided school..
3. There is no significant difference in achievement mean score between the pre test of control group and the post test of control group in Government school.

4. There is no significant difference in achievement mean score between the pre test of control group and the post test of control group in Aided school.
5. There is significant difference in achievement mean score between the pre test of Experimental group and the post test of Experimental group in Government school.
6. There is significant difference in achievement mean score between the pre test of Experimental group and the post test of Experimental group in Aided school.
7. Learning package in Negations of Tamil is more effective than conventional methods in learning Tamil Negation at standard VI in Government school and Aided school.

#### **Educational Implications:**

1. Learning package can be prepared for other subjects also.
2. It can be encouraged to implement to use in adult education
3. It may be implemented in Higher education

#### **Conclusion**

The study reveals that the students have problems in learning Negation in Tamil by using traditional approaches. Learning Package is more effective in Learning Tamil Negation. Hence it will be more supportive to promote the learners in learning Tamil.

#### **References**

- **Vasu Renganathan(2009)** Enhancing the process of learning Tamil with synchronized Media, Tamil internet conference, INFILL: Germany.
- **Sampath. K, Paneerselvem. A and Santhanam. S (1998)** Introduction to Educational technology, sterling publication Pvt Lit. Pg: no:103
- INFIT (2009) Conference papers, Tamil internet 2009, University of Cologne: Germany
- INFIT (2010) Conference papers , Tamil internet 2010, Coimbatore.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Morphology based Factored Statistical Machine Translation

(F-SMT) system from English to Tamil

Anand Kumar M<sup>1</sup>, Dhanalakshmi V<sup>1</sup>, Soman K P<sup>1</sup>, Rajendran S<sup>2</sup>

<sup>1</sup>Computational Engineering and Networking

Amrita Vishwa Vidyapeetham Coimbatore, India

{m\_anandkumar,v\_dhanalakshmi, kp\_soman}@cb.amrita.edu

<sup>2</sup>Tamil University, Thanjavur, India

## Abstract

This paper presents a novel preprocessing methodology in factorized Statistical Machine Translation system from English to Tamil language. SMT system considers the translation problem as a machine learning problem. Statistical machine translation system for morphologically rich languages is a challenging task. Moreover it is very complex for the different word order language pair. So a simple SMT alone would not give good result for English to Tamil, which differs in morphological structure and word order. A simple SMT system performs only at the lexical level mapping. Because of the highly rich morphological structure of Tamil language, a simple lexical mapping alone will suffer a lacuna in collecting all the morphological and syntactic information from the English language. The proposed SMT system is based on factored translation models. The factored SMT uses machine learning techniques to automatically learn translation patterns from factored corpora. Using the learned model FSMT predicts the output factors for the given input factors. Using the Tamil morphological generator the factored output is synthesized.

## Introduction

Statistical approach to machine translation learns translation patterns directly from training sentences and generalized them to handle new sentences. When translating from simple morphological language to the rich morphological language, the SMT baseline system will not generate the word forms that are not present in the training corpora. For training the SMT system, both monolingual and bilingual sentence-aligned parallel corpora of significant size are essential. The corpus size decides the accuracy of machine translation. The limited availability of parallel corpora for Tamil language and high inflectional variation increases a data sparseness problem for phrase-based SMT. To reduce the data sparseness, the words are split into lemma and their inflected forms based on their part of speech. Factored translation models [Koehn and Hoang, 2007] allow the integration of the linguistic information into a phrase-based translation model. These linguistic features are treated as separate tokens during the factored training process.

$$P(T|E) = P(T) P(E|T) / P(E)$$

$$T^* = \operatorname{argmax}_T P(T) P(E|T)$$

T

SMT works on the above equation. Where T represents Tamil language and E represents English language. We have to find the best Tamil translation sentence ( $T^*$ ) using  $P(T)$  and  $P(E|T)$ , Where  $P(T)$  is given by the Language model and  $P(E|T)$  is given by the translation model.

### Factored SMT for Tamil

Tamil language is morphologically rich language with free word order of SOV pattern. English language is morphologically simple with the word order of SVO pattern. The baseline SMT would not perform well for the languages with different word order and disparate morphological structure. For resolving this, we go for factored SMT system (F-SMT). A factored model, which is a subtype of SMT [Koehn and Hoang, 2007], will allow multiple levels of representation of the word from the most specific level to more general levels of analysis such as lemma, part-of-speech and morphological features. A preprocessing module is externally attached to the SMT system for Factored SMT.

The preprocessing module for source language includes three stages, which are reordering, factorization and compounding. In reordering stage the source language sentence is syntactically reordered according to the Tamil language syntax using reordering rules. After reordering, the English words are factored into lemma and other morphological features. A compounding process for English language is then followed, in which the various function words are removed from the reordered sentence and attached as a morphological factor to the corresponding content word. This reduces the length of English sentence. Now the representation of the source syntax is closely related to the target language syntax. This decreases the complexity in alignment, which is also a key problem in SMT from English to Tamil language.

Parallel corpora and monolingual corpora are used to train the statistical translation models. Parallel corpora contains factored English sentences (using Stanford parser) along with its factored Tamil translated sentences (using Tamil POS Tagger [V Dhanalakshmi et.al, 2009] and Morphological analyzer [M Anand kumar et.al,2009]). Factorized monolingual corpus is used in the Language model.

The parsed source language is reordered according to the target language structure using the syntax based reordering system. A compounding process for English language is then followed, in which the various function words are removed from the reordered sentence and attached as a morphological factor to the corresponding content word. This reduces the length of English sentence. Now the representation of the source syntax is closely related to the target language syntax. This decreases the complexity in alignment, which is also a key problem in SMT from English to Tamil language.

The factored SMT system's output is post processed, where the Tamil Morphological generator is pipelined to generate the target sentence. Figure.1 shows the architecture of the prototype factored SMT system from English to Tamil.

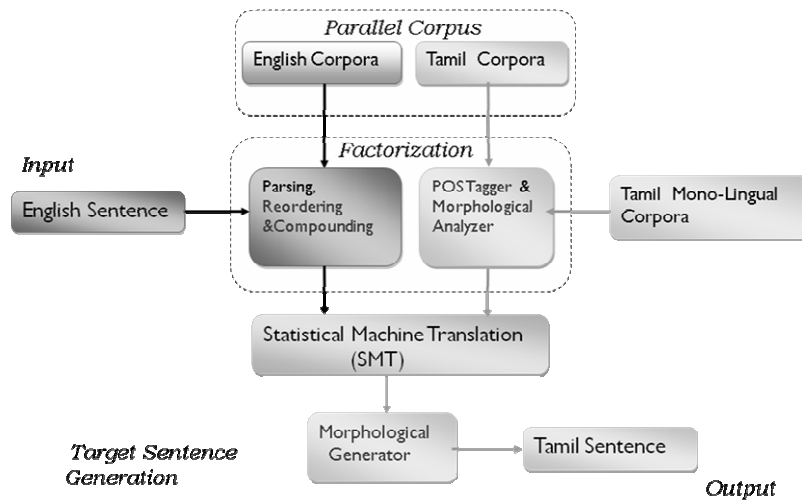


Figure.1 Architecture of the prototype factored SMT system from English to Tamil

## Morphological models for Tamil language

Morphological models for target language Tamil are used in preprocessing as well as post processing stage. In preprocessing, Tamil POS tagger and Morphological analyzer are used to factorize the Tamil parallel corpus and monolingual corpus. Morphological generator is used in the post processing stage to generate the Tamil words from Factored SMT output.

### Tamil POS tagger

Parts of speech (POS) tagging means labeling grammatical classes i.e. assigning parts of speech tags to each and every word of the given input sentence. POS tagging for Tamil is done using SVM based machine learning tool [V Dhanalakshmi et.al, 2009], which make the task simple and efficient. The SVM Tool[] is used for training the tagged sentences and tagging the untagged sentences. In this method, one requires Part of speech tagged corpus to create a trained model.

### Tamil Morphological Analyzer

The Tamil morphological analyzer is based on sequence labeling and training by kernel methods. It captures the non-linear relationships and various morphological features of natural language in a better and simpler way. In this machine learning approach two training models are created for morphological analyzer. These two models are represented as Model-I and Model-II. First model is trained using the sequence of input characters and their corresponding output labels. This trained model-I is used for finding the morpheme boundaries [M Anand kumar et.al, 2009].

Second model is trained using sequence of morphemes and their grammatical categories. This trained Model-II is used for assigning grammatical classes to each morpheme. The SVMTool is used for training the data. Generally SVMTool is developed for POS tagging but here this tool is used in morphological analysis.

### Tamil Morphological Generator

The developed morphological generator receives an input in the form of lemma+word\_class+Morpho-lexical Information, where lemma specifies the lemma of the word-form to be generated, word\_class

specifies the grammatical category (POS category) and Morpho-lexical Information specifies the type of inflection. The morphological generator system needs to handle three major things; first one is the lemma part, then the word class and finally the morpho lexical information. By the way the generator is implemented makes it distinct from other morphological generator[M Anand kumar et.al,2010].

The input which is in Unicode format is first Romanized and then the paradigm number is identified by end characters. For sake of easy computation we are using romanized form. A Perl program has been written for identifying paradigm number, which is referred as column index. The morpho-lexical information of the required word class is given by the user as input. From the morpho-lexicon information list the index number of the corresponding input is identified, this is referred as row index. A verb and noun suffix tables are used in this system. Using the word class specified by the user the system uses the corresponding suffix table. In this two-dimensional suffix table rows are morpho-lexical information index and columns are paradigm numbers.

## Conclusion

In this paper, we have presented a morphology based Factored SMT for English to Tamil language. The morphology based Factored SMT improves the performance of translation system for morphologically rich language and also it drastically reduces the training corpus size. So this model is suitable for languages which have less parallel corpus. Tamil morphological models are used to create a factorized parallel corpus. Source language reordering module captures structural difference between source and target language and reorder it accordingly. Compounding module converts the source language structure to fit into the target language structure. Initial results obtained from the Factored SMT are encouraging.

## References

- Philipp Koehn and Hieu Hoang (2007), "Factored Translation Models", Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic, June 2007.
- V Dhanalakshmi, M Anand kumar, K P Soman, S Rajendran (2009),"POS Tagger and Chunker for Tamil language", Proceedings of Tamil Internet Conference 2009, Cologne, Germany, October 2009.
- M Anand kumar, V Dhanalakshmi, K P Soman, S Rajendran (2009),"A Novel Approach For Tamil Morphological Analyzer", Proceedings of Tamil Internet Conference 2009 , Cologne, Germany, Page no: 23-35, October 2009.
- M Anand kumar, V Dhanalakshmi, R U Rekha, K P Soman, S Rajendran (2010), "Morphological Generator for Tamil a new data driven approach", Proceedings of Tamil Internet Conference 2010, Coimbatore, India, 2010.
- Jes'us Gim'enez and Llu'is M'arquez.(2004), "SVMTool: A general pos tagger generator based on support vector machines", Proceedings of the 4th LREC Conference, 2004.
- Fishel,M (2009), "Deeper than words : Morph-based Alignment for Statistical Machine Translation ", Proceedings of the conference of the pacific Association for Computational Computational Linguistics (PacLing 2009 ) Sapporo, Japan.





## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011



# Tamil Shallow Parser using Machine Learning Approach

*Dhanalakshmi V<sup>1</sup>, Anand Kumar M<sup>1</sup>, Soman K P<sup>1</sup> and Rajendran S<sup>2</sup>*

*<sup>1</sup>Computational Engineering and Networking*

*Amrita Vishwa Vidyapeetham Coimbatore, India*

*{m\_anandkumar,v\_dhanalakshmi, kp\_soman}@cb.amrita.edu*

*<sup>2</sup>Tamil University, Thanjavur, India*

## Abstract

This paper presents the Shallow Parser for Tamil using machine learning approach. Tamil Shallow Parser is an important module in Machine Translation from Tamil to any other language. It is also a key component in all NLP applications. It is used to understand natural language by machine and also useful for second language learners. The Tamil Shallow Parser was developed using the new and state of the art machine learning approach. The POS Tagger, Chunker, Morphological Analyzer and Dependency Parser were built for implementing the Tamil Shallow Parser. The above modules gives an encouraging result.

## Introduction

Partial or Shallow Parsing is the task of recovering a limited amount of syntactic information from a natural language sentence. A full parser often provides more information than needed and sometimes it may also give less information. For example, in Information Retrieval, it may be enough to find simple NPs (Noun Phrases) and VPs (Verb Phrases). In Information Extraction, Summary Generation, and Question Answering System, information about special syntactico-semantic relations such as subject, object, location, time, etc, are needed than elaborate configurational syntactic analyses. In full parsing, grammar and search strategies are used to assign a complete syntactic structure to sentences. The main problem here is to select the most possible syntactic analysis to be obtained from thousands of possible analyses a typical parser with a sophisticated grammar may return. This complexity of the task makes machine learning an attractive option in comparison to the handcrafted rules.

## Methodology

Machine learning approach is applied here to develop the shallow parser for Tamil. Part of speech tagger for Tamil has been generated using Support Vector Machine approach [Dhanalakshmi V e.tal., 2009]. A novel approach using machine learning has been built for developing morphological analyzer for Tamil [Anand kumar M e.tal., 2009]. Tamil Chunker has been developed using CRF++ tool [Dhanalakshmi V e.tal., 2009]. And finally, Tamil Dependency parser, which is used to find syntactico-semantic relations such as subject, object, location, time, etc, is built using MALT Parser [Dhanalakshmi V e.tal., 2011].

### General Framework and Modules

- The general block diagram for Tamil Shallow parser is given in Figure 1.

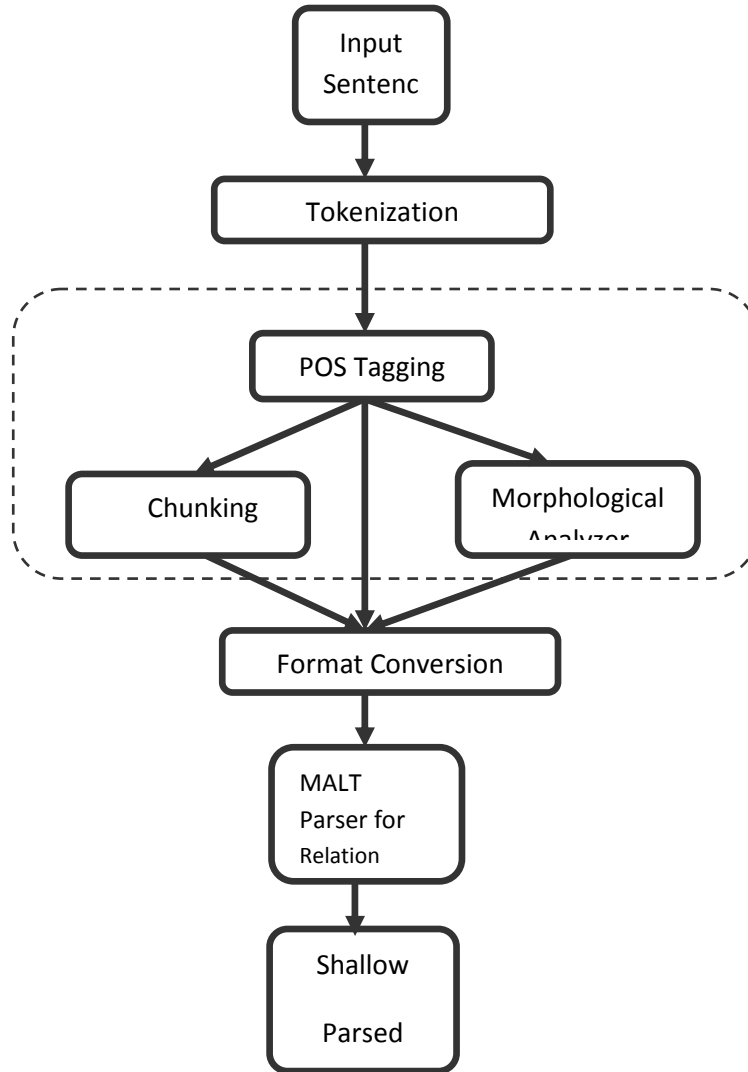


Figure.1. General Framework for Tamil Shallow Parser

- **Tamil Part-of-Speech Tagger** [Dhanalakshmi V e.tal., 2009]: The Part of Speech (POS) tagging is the process of labeling a part of speech or other lexical class marker (noun, verb, adjective, etc.) to each and every word in a sentence. POS tagger was developed for Tamil language using SVMTool [Jes´us Gim´enez and Llu´is M´arquez, 2004].
- **Tamil Morphological Analyzer** [Anand Kumar M e.tal., 2009]: Morphological Analysis is the process of breaking down morphologically complex words into their constituent morphemes. It is the primary step for word formation analysis of any language. Morphological Analyzer was developed using a novel machine learning approach and was implemented using SVMTool.
- **Tamil Chunker** [Dhanalakshmi V e.tal., 2009]: Chunks are normally taken to be non recursive correlated group of words. Chunker divides a sentence into its major-non-overlapping phrases

(noun phrase, verb phrase, etc.) and attaches a label to each chunk. Chunker for Tamil language was developed using CRF++ Tool [Sha F and Pereira F, 2003].

- **Tamil Dependency Parser for Relation finding** [Dhanalakshmi V e.tal., 2011]: Given the POS tag, Morphological information and chunks in a sentence, this decides which relations they have with the main verb (subject, object, location, etc.). Dependency parser was developed for Tamil language using Malt Parser tool [Joakim Nivre and Johan Hall, 2005].

## Dependency Parsing using Malt Parser

MALT Parser Tool is used for dependency parsing, which uses supervised machine learning algorithm. Using this tool dependency relations and position of the head are obtained for Tamil sentence. There are 10 tuples used in the training data that can be user define. For Tamil dependency parsing, the following features are defined and others are set as NULL and are mentioned as '\_' in the training data format.

**WordID:** Position of each word in the input sentence.

**Words:** Each word in the input sentence.

**CPos Tag and Pos Tag:** Defines the Parts Of Speech of each word.

**Head:** The position of the parent of each word.

**Lemma:** The lemma of the word.

**Morph Features** The Morphological features of the word.

**Chunk** The chunk information of the word.

**Dependency Relation:** The terminology given for each parent – child relation.

### Sample Training Data

- 1 அவள் \_ <PRP> <PRP> 8 <N.SUB> \_ \_
- 2 முட்டைகளை \_ <NN> <NN> 3 <D.OBJ> \_ \_
- 3 வாங்கி \_ <VNAV> <VNAV> 4 <ATT> \_ \_
- 4 சமைத்து \_ <VNAV> <VNAV> 6 <VNAV.MOD> \_ \_
- 5 தட்டில் \_ <NN> <NN> 6 <NST.MOD> \_ \_
- 6 போட்டு \_ <VNAV> <VNAV> 8 <V.COMP> \_ \_
- 7 உனக்கு \_ <PRP> <PRP> 8 <I.OBJ> \_ \_
- 8 கொடுக்கின்றான் \_ <VF> <VF> 0 <ROOT> \_ \_
- 9 . <DOT> <DOT> 8 <SYM> \_ \_

For Tamil language, a corpus of three thousand sentences is annotated with dependency relations and labels using the customized tag set (Table.1). The corpus is trained using the MALT Parser tool which generates a model. Using this model the new input sentences are tested.

| S.No | Tags  | Description     | S.No | Tags    | Description           |
|------|-------|-----------------|------|---------|-----------------------|
| 1    | ROOT  | Head word       | 5    | NST-MOD | Spatial Time Modifier |
| 2    | N-SUB | Subject         | 6    | SYM     | Symbols               |
| 3    | D-OBJ | Direct Object   | 7    | X       | Others                |
| 4    | I-OBJ | Indirect Object |      |         |                       |

*Table.1 Shallow Dependency Tagset*

## Application of Shallow Parser

Shallow parsers were used in Verbmobil project [Wahlster W, 2000], to add robustness to a large speech-to-speech translation system. Shallow parsers are also typically used to reduce the search space for full-blown, 'deep' parsers [Collins, 1999]. Yet another application of shallow parsing is question-answering on the World Wide Web, where there is a need to efficiently process large quantities of ill-formed documents [Buchholz and Daelemans, 2001] and more generally, all text mining applications, e.g. in biology [Sekimizu et al., 1998].

The developed Tamil Shallow Parser can be used to develop the following systems for Tamil language.

- Information extraction and retrieval system for Tamil.
- Simple Tamil Machine Translation system.
- Tamil Grammar checker.
- Automatic Tamil Sentence Structure Analyzer.
- Language based educational exercises for Tamil language learners.

## Conclusion

Shallow Parsing has proved to be a useful technology for written and spoken language domains. Full parsing is expensive, and is not very robust. Partial parsing has proved to be much faster and more robust. Dependency parser is better suited than phrase structure parser for languages with free or flexible word order like Tamil. Fully functional Shallow Parser for Tamil gives reliable results. The Shallow Parser system developed for Tamil is an important tool for Machine Translation between Tamil and other languages.

## References

- Anand kumar M, Dhanalakshmi V , Soman K P and Rajendran S (2009) , "A Novel Approach for Tamil Morphological Analyzer", Proceedings of the 8th Tamil Internet Conference 2009, Cologne, Germany.
- Buchholz Sabine and Daelemans Walter (2001), "Complex Answers: A Case Study using a WWW Question Answering System", Natural Language Engineering.
- Collins M (1999), "Head-Driven Statistical Models for Natural Language Parsing", Ph.D Thesis, University of Pennsylvania.

- Dhanalakshmi V, Anand Kumar M, Vijaya M S, Loganathan R, Soman K P, Rajendran S (2008), "Tamil Part-of-Speech tagger based on SVMTool", Proceedings of the COLIPS International Conference on natural language processing(IALP), Chiang Mai, Thailand.
- Dhanalakshmi V, Anand kumar M, Soman K P and Rajendran S (2009), "POS Tagger and Chunker for Tamil Language", Proceedings of the 8th Tamil Internet Conference, Cologne, Germany.
- Dhanalakshmi V, Anand Kumar M, Rekha R U, Soman K.P and Rajendran S (2011), "Data driven Dependency Parser for Tamil and Malayalam" NCILC-2011, Cochin University of Science & Technology, India.
- Jes´us Gim´enez and Llu´is M`arquez.(2004) SVMTool: *A general pos tagger generator based on support vector machines*.In Proceedings of the 4th LREC Conference, 2004.
- Joakim Nivre and Johan Hall, MaltParser: A language-independent system for data-driven dependency parsing. In Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT), 2005.
- Sekimizu T, Park H and Tsujii J (1998), "Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts", Genome Informatics, Universal Academy Press.
- Sha F and Pereira F (2003), "Shallow Parsing with Conditional Random Fields", Proceedings of Human Language Technology Coference'2003, Canada.
- Wahlster W (2000), "VERBMOBIL: Foundations of Speech-to-Speech Translation", Springer-Verlag.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# கணினிவழித் தமிழ்மொழியாய்வில் பொருள் மயக்கம்

## Ambiguities in Computer Assisted Tamil Language Processing

இல. சுந்தரம்

துணைப்பேராசிரியர், ஒருங்கிணைப்பாளர்,

கணினித்தமிழ்க் கல்வி தமிழ்ப்பேராயம், SRM பல்கலைக்கழகம். மின்னஞ்சல்: [sundarbaskar@gmail.com](mailto:sundarbaskar@gmail.com)

### முன்னுரை

கணினியில் தமிழ்மொழியின் பயன்பாடு பெருகியுள்ளது. தமிழ்மொழியின் வளர்ச்சிக்குக் கணினியின் பங்களிப்பு தவிர்க்கமுடியாத ஒன்றாகிவிட்டது. மொழி ஆய்வுக் கருவியாகக் கணினியைப் பயன்படுத்தி வருகிற நிலையில் தமிழ்மொழித் தரவுகளை அதற்கு ஓர் ஒழுங்கமைவுடன் கற்றுத்தரவேண்டியுள்ளது. அதாவது கணித அடிப்படையில் மொழியில் உள்ள மொழியியல் கூறுகளைக் கணினிக்கு ஏற்ற வகையில் மாற்றித்தரவேண்டியுள்ளது. இத்தகைய வழிமுறைகளைக் கொடுப்பதே கணினி மொழியியல் என்பதாகும். மொழி செயல்படுவதில் உள்ள ஒழுங்குமுறையின் தொகுப்புதான் இலக்கணம். இத்தகைய ஒழுங்குமுறை நவீன, தொழில்நுட்ப வளர்ச்சிகளினாலும் மொழி உலகமயமாக்கச் சூழலினாலும் சிதைந்தும் மாறுபட்டும் வருகிறது. மொழியை இத்தகைய சிதைவுகளிலிருந்து மீட்டெடுக்க மொழியியல் கூறுகளை முறையாகக் கற்று, பயன்படுத்தவேண்டிய கட்டாயம் ஏற்பட்டுள்ளது.

பொருள் மயக்கம் தமிழ்மொழிப் பயன்பாட்டில் உருவாக்குகின்ற நிலைப்பாடுகளையும் ,கணினிவழி ஆய்வு செய்யும்போது ஏற்படுகிற மொழியமைப்புச் சிக்கல்களையும் ,அவற்றைத் தவிர்ப்பதற்கான வழிமுறைகளையும், மொழியியல் கூறும் வகைப்பாட்டு நெறிமுறைகளையும் எடுத்துக்கூறுவதாக இக் கட்டுரை அமைகிறது.

### இயற்கைமொழியாய்வு; கணினிமொழியியல்; மொழித்தொழில்நுட்பம்:

தமிழ்மொழியின் இயல்புகளைத் தெளிவாக அறிந்துகொள்ள ஒலியனியல், உருபனியல், தொடரியல் மற்றும் பொருண்மையியல் போன்ற மொழியியல் அறிவு இன்றியமையாதன.

மனித மூளையைப் போன்று கணினியையும் இயற்கைமொழி அறிவைப் பெறவைத்து ,மொழித் தொடர்களைப் புரிந்துகொள்ளவும், உருவாக்கவும், செய்யவைக்கும் முயற்சியே இயற்கை மொழியாய்வு (Natural Language Processing). இத்தகைய இயற்கைமொழியாய்வை மேற்கொள்ள உருவாக்கப்படுகிற வழிமுறைகளும் முறைப்படுத்தலுமே கணினி மொழியியல் (Computational Linguistics). கணினி மொழியியலின் துணையோடு மொழிக்குத் தேவையான மின்னணு மொழிக் கருவிகளை உருவாக்க உதவும் நுட்பமே மொழித்தொழில்நுட்பம் (Language Technology). இவை மூன்றும்தான் தமிழ் மென் பொருள்களை உருவாக்குவதற்கு மேற்கொள்ளப்படுகிற படிமுறை வளர்ச்சிப் பணிகள்.

கணினித்தமிழ் வளர்ச்சி என்பது தமிழ்த் தொடர்களைப் புரிந்துகொள்ளவும் (Understanding), அவற்றை உருவாக்கவும் (Generate) தேவையான தமிழ்மொழி அறிவைக் கணினிக்கு அளிப்பதற்காக நாம் மேற்கொள்ளவேண்டிய பணிகளைக் குறிக்கிறது. தமிழ்த் தரவுகளைக் கணினி புரிந்துகொள்ளும் வகையில் கொடுப்பதற்கு மொழியியல் விதிகளும் கோட்பாடுகளும் துணைபுரிகின்றன. கணினி மொழியியல் கோட்பாடுகளைக்கொண்டு மொழியின் அமைப்பை, இலக்கணத்தைக் கணினிக்கேற்ற வகையில் நிரலிகளாக )Programs(, மின்னணு இலக்கணமாக மாற்றிக் கொடுத்து, தமிழ்மொழியின் தேவையை நிறைவுசெய்ய வேண்டும். இவ்வாறு தமிழ்மொழியின் அமைப்பை ஒழுங்கமைவுடன்,

விதிகளாக மாற்றும்போது தமிழ்மொழியின் தற்கால எழுத்து வழக்கில் பல்வேறு முறைகள் பயன்படுத்தப்படுவதால் சொற்களைப் பிரிக்கும்போதும்(Parsing) வரிசைப்படுத்தும்போதும் (Sorting) பல்வேறு மொழிப் பயன்பாட்டுச் சிக்கல்கள் எழுகின்றன. இத்தகைய மொழிப் பயன்பாட்டுச் சிக்கல்களில் ஒன்றுதான் பொருள் மயக்கம்(Word Sense Ambiguity).

தமிழில் சந்திப் பிழைதிருத்தி (Sandhi Checker), உருபனியல் பகுப்பாய்வி (Morphological Parser), தொடரியல் பகுப்பாய்வி (Syntactic Parser), அடைவி (Indexing)(சொல்லடைவு, தொடரடைவு, பொருளடைவு), தானியங்கி பேச்சு அறிவான் (Automatic Speech Recognizer-ASR), இயந்திர மொழிபெயர்ப்பு (Machine Translation) ஆகிய மொழியாய்வு மென்பொருள் கருவிகளை உருவாக்குவதில் இத்தகைய பொருள் மயக்கம் இடையூறாக அமைகின்றன. இவற்றைச் சரிசெய்ய, பொருள் மயக்கச் சொல்லகராதியை உருவாக்கவேண்டியது அவசியம்.

### பொருள் மயக்கம் - விளக்கம்

‘Word Sense Ambiguity’ என்னும் ஆங்கிலச் சொல் தமிழில் தெளிவின்மை, குழப்பம், கருத்துமயக்கம், பொருள்மயக்கம், இருபொருள்படுநிலை, தெளிவற்ற நிலை எனப் பல்வேறு நிலைகளில் பொருள்கொள்ளப்படுகின்றது. எனினும், கணினிமொழியியலில் பொருள் மயக்கம் என்றே கையாளப்படுகின்றது. இத்தகைய பொருள் மயக்கங்களைக் களைவதைக் கணினிமொழியியலில் ‘Word Sense Disambiguation )WSD(’ என்று கூறுவர்.

ஒரு தொடர் தன் அமைப்பில் வெளித்தோற்றத்திலும் உள்தோற்றத்திலும் வெவ்வேறு பொருள்தருகிறது . இத்தகைய பொருண்மை மாறுபாடு ஏற்படுவதற்குரிய சில சொற்களும் சில சூழ்நிலைகளும் இங்கு நோக்கப்படுகின்றன. தமிழ் மரபிலக்கணத்தில் ஒருசொல் குறித்த பல பொருள், பல பொருள் குறித்த ஒருசொல் என்ற வகைப்பாடும் காணப்படுகிறது. அகராதி நிலையில் ஒரு சொல்லுக்குப் பல பொருள்கள் இருக்கலாம் .ஆனால், இவற்றிலிருந்து பொருள் மயக்கம் என்பது மாறுபட்டது.

### பொருள் மயக்கம் ஏற்படுவதற்கான நிலைப்பாடுகள்

தமிழ்மொழித் தரவுகள் உலகளாவிய பொதுமொழியின் தன்மைகளைக் கொண்டிருப்பதோடு தமக்கெனச் சில தனித்தன்மைகளைக் கொண்டிருக்கின்றன. வழக்கிழந்த கூறுகளும் புத்தாக்கங்களும் தமிழில் காலங்காலமாக நிகழ்ந்துகொண்டுள்ளன. சாதி, தொழில், வட்டாரம் போன்றவை சார்ந்த வழக்குகளும், துறைசார்ந்த வழக்குகளும் பேச்சு, எழுத்து என்னும் நிலைப்பாடுகளும் தமிழ்மொழித் தரவினைக் கணினியின் ஏற்புத்திறனுக்கு ஏற்றாற்போல் ஒருமைப்படுத்துவதற்கும் பொதுவிதிகளை உருவாக்குவதற்கும் இடையூறுகளாக அமைகின்றன.

சொற்களின் இலக்கண வகைப்பாட்டை நாம் நுண்மையான இலக்கண அறிவு (Grammatical Knowledge) மற்றும் உலகியல் அறிவின் (Pragmatic Knowledge) துணையோடு அறிகிறோம். ஆனால் அவற்றைக் கணினிக்குக் கற்றுத்தருவதில் பல்வேறு மொழியமைப்புச் சிக்கல்கள் எழுகின்றன. அவற்றைச் சரிசெய்வதற்கு உருபனியல், தொடரியல் பகுப்பாய்வுகள் துணைபுரிகின்றன. ஒரு தொடரில் ஒன்றுக்கு மேற்பட்ட அமைப்புகள் காணப்படலாம். அதாவது குறிப்பிட்ட தொடரில் இடம்பெறும் சொற்கள் தங்களுக்குள் வெவ்வேறு வகையில் இணையலாம். அப்போது பொருள் மயக்கம் ஏற்படுகிறது.

ஆங்கிலத்தில் ஒலிபெயர்த்து (Transliterate) எழுதும்போது முறைப்படுத்தப்பட்ட ஒலிக்குறிப்பு எழுத்துக்களைப் பயன்படுத்தவேண்டும். ஆனால் குறில், நெடில், ல,ழ,ள, ற,ர போன்ற எழுத்துக்கள் வேறுபாடுகளின்றிப் பயன்படுத்தப்படுவதால் பொருள் குழப்பமும் அவற்றை உச்சரிக்கும்போது தெளிவில்லாத சூழ்நிலையும் காணப்படுகிறது. எடுத்துக்காட்டாக, பாடம் என்று எழுதுவதைப் ‘padam’ என்று எழுதினால் படம் என்று படிப்பதற்கும் வாய்ப்பிருக்கிறது. எனவே மக்களின் பெயர், ஊர்ப்பெயர்,



முகவரி, பொருள்களின் பெயர் போன்றவற்றைத் தவறாக உச்சரிக்கிற நிலை ஏற்படுகிறது. எனவே, இவற்றை ஓர் ஒழுங்குமுறைக்குக் கொண்டுவரவேண்டும்.

பொருள் வேறுபாட்டிற்கு வேற்றுமை உருபுகளும், சந்தி மாற்றங்களும், ல,ழ,ள, ற,ர வேறுபாடுகளும் முக்கியப் பங்காற்றுகின்றன. மேலும் சாரியைகள், இரட்டித்தல் போன்றவையும் துணைசெய்கின்றன.

பாடல்களைப் படிக்கும்போது எளிமையாகப் புரிந்துகொள்ளவேண்டுமென்னும் நோக்கில் சொற்களைப் பிரிப்பதாலும் உரைநடை எழுதும்போது பொருள் மயங்குவது தெரியாமல் சொற்களைப் பிரிப்பதாலும் பொருள் மயங்குகிறது. பொருள் மயக்கம் ஏற்படாதவாறு பிரிக்கவேண்டும் என்பதைக் கவனத்தில் கொள்ளவேண்டியது அவசியம். பொருள் உணரும் திறன் குறைந்த இக் காலத்தில் பாடல்களில் எல்லாச் சொற்களையும் பிரித்தே எழுதுதல் வேண்டும், எளிமைப்படுத்தவேண்டும், சாதாரணப் பேச்சுவழக்கில் இருக்கவேண்டும் என்பது போன்ற தன்மைகள் கடைபிடிக்கப்படுகின்றன. மேலும், எழுத்துநடையில் மற்றவர்களிடமிருந்து தங்களை வேறுபடுத்தவேண்டும் என்பதற்காகவும் இத்தகைய நிலை இருக்கின்றது.

## 1. தனிச்சொற்களால் ஏற்படுகிற பொருள்மயக்கம்

சில தனிச் சொற்கள் தொடர்களில் பயன்படுத்தும்போது இருவேறு பொருள்களைத் தந்து நிற்கின்றன. தமிழில் தனித்த சில சொற்களைத் தொடர்களில் பயன்படுத்தும்போது அவை தோற்றத்தில் ஒன்று போலவும் பொருளில் இருவேறு நிலைகளிலும் காணப்படுகின்றது. ஒரு தொடரில் வேலை என்ற சொல் காணப்படுகிறது. அது 'வேலையைக்' குறிக்கிறதா? அல்லது 'வேல்' என்னும் ஆயுதத்தைக் குறிக்கிறதா? என்ற மயக்கம் ஏற்படுகிறது. தொடர் நிலையில் அதற்கு அடுத்து அல்லது அதற்கு முன் அமைந்த சொல்லை வைத்தே, இந்தச் சொல் இதைத்தான் குறிக்கிறது என்று அறியமுடிகிறது. நான் வேலை வாங்கினேன்.

[அவரை - அவர் + ஐ அவரைச் செடி], [வருட - வருடம், தலையை வருட],

[காலை - கால் + ஐ காலைப்பொழுது], [பாத்திரம் - கதாப்பாத்திரம், சமையல்பாத்திரம்]

[ஆறு - ஆறு(River) எண்(Number)], [எண்ண - எண்ணம்(Thinking) எண்ண(Counting)]

மேற்குறித்த சில சொற்களுடன் இரண்டாம் வேற்றுமை உருபு வந்துள்ளதா அல்லது தனிச்சொல்தானா என்ற குழப்பமே இந்தப் பொருள்மயக்கத்திற்குரிய காரணமாகும். இத்தகைய குழப்பமின்றி வேறுபடுத்துவதற்குச் சில இடங்களில் 'இன்' சாரியை பயன்படுத்தப்படுகிறது.

காது + ஐ = காதை => காது + இன் + ஐ = காதினை.

காடு + ஐ = காடை => காடு + ட்(இன்) + ஐ = காட்டை, காட்டினை.

## 2. தொடரமைப்பு நிலையில் ஏற்படுகிற பொருள் மயக்கம்

ஒரு தொடர் அமைப்பில் எல்லாச் சொற்களும் சரியான பொருளையே தந்துநின்றாலும் அவை பொருள்கொள்ளும் முறையில் மயக்கம் ஏற்படுகின்றன. 'முட்டாள் குமரனின் மனைவி' என்னும் தொடரில் முட்டாள் என்பது குமரனுக்குப் பெயரடையாக வருகிறதா அல்லது அவன் மனைவிக்குப் பெயரடையாக வருகிறதா என்கிற குழப்பம் ஏற்படுகிறது. இத்தகைய நிலையில் வேற்றுமை உருபு மறைந்து வருவதாலும் முட்டாள் என்பதற்கு அடுத்து, காற்புள்ளி இட்டு எழுதாததாலும் இத்தகைய குழப்பம் ஏற்படுகிறது. இதனை அமைப்புப் பொருள் மயக்கம்(Structural Ambiguity) என்று மொழியியல் அறிஞர்கள் கூறுவர். தொடரின் புறநிலையிலும் அகநிலையிலும் மாறுபடாமல் குழப்பமின்றி இருந்தாலும் அவை எடுத்துக்கொள்ளும் முறையிலும் சூழல் தரும் பொருளிலும் வேறுபடுகின்றன.

### 3. சொற்களைப் பிரித்தும் சேர்த்தும் எழுதுகின்ற நிலையில் ஏற்படுகிற பொருள் மயக்கம்

தமிழில் வேர்ச்சொல்லுடன் பல்வேறுபட்ட ஒட்டுகள் இணைகின்றன. அவ்வாறு இணையும்போது அவற்றுக்குள்ளேயே ஓர் இயைபு விதி உருவாகின்றது. இவ்வாறு சொற்களுடன் ஒட்டுகளை இணைக்கும்போது சொற்களைப் பிரித்தும் சேர்த்தும் எழுதுகின்ற வழக்கம் காணப்படுகின்றது.

தமிழில் மொழியியல் விதிப்படி தனித்து நின்று பொருள்தராத துணைவினைகள் (Auxiliary Verb), ஒட்டுகள் (Affixes) மிதவை ஒட்டுகள் (Clitic) போன்றவற்றைப் பிரித்து எழுதக்கூடாது என்பதை மீறுவது பொருள் மயக்கத்திற்கு முக்கியக் காரணமாகும்.

பொதுவாக ஒரு சொல்லைப் பிரித்தோ சேர்த்தோ எழுதும்போது கூறவந்த கருத்தின் அடிப்படையே மாறுகின்ற நிலை ஏற்படுகிறது. எடுத்துக்காட்டாக, அவனுடனே என்று சேர்த்து எழுதினால் with him என்று பொருள்படும். அவன் உடனே என்று பிரித்து எழுதினால் he at once என்று பொருள்படும். எனவே மிகக் கவனத்தோடு இடமறிந்து பொருள்மயக்கம் ஏற்படாதவாறு சேர்த்தோ பிரித்தோ எழுதவேண்டும். பல்கலைக்கழகம், தொழில்நுட்பம் போன்ற சில கலைச்சொற்களையும் பிரித்து எழுதுதல் கூடாது. இதுபோல மொழிப் பயன்பாட்டு விதிகளை முறையாகப் பயன்படுத்தினால் கணினிவழி மொழியாய்வுக்கும் பொருள் மயக்கமின்றி வாசிப்பதற்கும் பயன்தரும்.

- **துணைவினைகள்**

**விடு**(வந்துவிடு, போய்விடு, படித்துவிடு, தூங்கிவிடு). **படு**(பாடுபடு, வேதனைப்படு, ஆசைப்படு). **இரு**(பார்த்துக்கொண்டிரு, படித்துக்கொண்டிரு). **இடு**சேர்த்திடு, காட்டிடு, பார்த்திடு). **கொண்டு** (தெரிந்துகொண்டு, பார்த்துக்கொண்டிரு). **கொள்ள** (பார்த்துக்கொள்ள, பேசிக்கொள்ள, அறிந்து கொள்ள). **விட்டு**, **விட்டது**(பார்த்துவிட்டு, பேசிவிட்டு, பார்த்துவிட்டது, போய்விட்டது). **பட்டு**, **பட்டது**(அறியப்பட்டு, விளக்கப்பட்டு, கூறப்பட்டது, சேர்க்கப்பட்டது). **வேண்டும்** (பார்க்க வேண்டும், செல்லவேண்டும், எழுதவேண்டும்). **உள்ளது**(தெரியவந்துள்ளது, பாடப்பட்டுள்ளது).

கொள், உண், ஆம், போடு, வரு, தரு, உள் இதுபோன்ற ஐம்பதுக்கும் மேற்பட்ட துணைவினைகள் எழுத்து வழக்கிலும் பேச்சு வழக்கிலும் காணப்படுகின்றன. ஒரு தொடரில் ஒன்றுக்கு மேற்பட்ட துணைவினைகளும் இணைந்து வரும்.

அவர்கள் படித்துவிட்டுச் சென்றனர். அவர்கள் படித்து விட்டுச் சென்றனர்.

பிரித்து எழுதியதால் இவ்விரு தொடர்களுக்கிடையே பொருள் வேறுபாடு தெளிவாகத் தெரிகிறது.

- **மிதவை ஒட்டு**

தான் - அதைத்தான், அவன்தான், அப்போதுதான், அதனால்தான்.

- ❖ **பின்னொட்டு**

**கீழ், மேல்** - துறையின்கீழ், தலைமேல். **வழி** - கணினிவழி, அதன்வழி.

**விட** - அவனைவிட, பேசியதைவிட.

- ❖ **வினை விகுதி**

**போது** - சொன்னபோது, பார்த்தபோது. **படி** - அதன்படி, சொன்னபடி.

- ❖ **பொதுநிலை**

**கண்** - அதன்கண். **காலம்** - இடைக்காலம், சங்ககாலம்.

**வர** - சென்றுவர, நடந்துவர.

## உருபனியலும் பொருள்மயக்கமும்

ஒரு சொல் ஓர் உருபன் கொண்டதாகவோ அல்லது அதற்கு மேற்பட்ட உருபங்களாகவோ இருக்கலாம். பல்வேறு உருபங்களால் உருவான சொற்களைக் கணினிவழிப் பகுப்பாய்வு செய்வது 'உருபனியல் பகுப்பாய்வு' என்பதாகும். இதற்காக உருபனியல் பகுப்பாய்விகள் (Morphological Parsers) உருவாக்கப் பட்டுவருகின்றன. இவ்வாறு உருவாக்கும்போது பொருள்மயக்கச் சொற்களின் சிக்கல்கள் நோக்கத் தக்கதாக உள்ளன.

இயந்திர மொழிபெயர்ப்பில் (Machine Translation) கணினிமொழியியல் விதியான இருநிலை உருபனியல் (Two Level Morphology) என்ற மொழித்தன்மைகுறித்து ஆராய்வர். ஒரு தொடரில் அடிநிலை (Deep Structure), புறநிலை (Surface Structure) ஆகிய இரண்டும் காணப்படும். இவற்றுள் புறநிலையில் எந்தவித மாறுபாடும் ஏற்படுவதில்லை. ஆனால், பொருள் மயக்கச் சொற்கள் வரும்போது அகநிலையில் குழப்பம் ஏற்படுகிறது.

தமிழில் காணப்படும் தொடர்களில் வேர்ச்சொற்கள் தனித்தும் விகுதிகளேற்றும் காணப்படுகின்றன. தனித்த சொற்களைக் கண்டறிவதற்கு அகராதிகளைப் பயன்படுத்தலாம். மற்றவற்றை உள்ளீடு செய்து ஆய்வுசெய்தே பகுத்தறிய முடியும். வேர்ச்சொற்களையும் ஒட்டுகளையும் பகுத்து, பொருள் மயக்கமின்றி வகைப்படுத்துவதற்கு உருபனியல் பகுப்பாய்வு அவசியமாகிறது.

## மொழியியல் வகைப்பாட்டில் பொருள்மயக்கம்

மொழியியல் அடிப்படையில் பொருள் மயக்கத்தை, ஒலியனியல் (Phonology), உருபனியல் (Morphology), தொடரியல் (Syntax), சொற்பொருண்மையியல் (Semantics), கருத்தாடல் (Discourse) ஆகிய நிலைகளில் வகைப்படுத்தலாம்.

ஒலியனியல் (சந்தி) நிலையில், 'வேலை செய்தான்', 'வேலைச் செய்தான்' என்பவற்றில் முதலாவது வேலை பணியைக் குறிக்கிறது, இரண்டாவது வேலை கருவியைக் குறிக்கிறது. உருபனியல் நிலையில், 'நான் கத்தி விற்பேன்' என்ற தொடரில் கத்தி என்ற பெயரைக் குறிக்கிறதா அல்லது வினையைக் குறிக்கிறதா என்பதில் குழப்பம் ஏற்படுகிறது. தொடரியல் நிலையில், 'நான் இராமனோடு சீதையைப் பார்த்தேன்' என்ற தொடரில் இரண்டு வகையாகப் பொருள்கொள்ளலாம். நானும் இராமனும் சீதையைப் பார்த்தோம் என்றும் நான் இராமனும் சீதையும் சேர்ந்திருக்கும்போது பார்த்தேன் என்றும் பொருள் படுகிறது. சொற்பொருண்மை நிலையில், 'பச்சைக் காய்கறி', 'பச்சைப் பொய்', 'பச்சை உடம்பு' ஆகிய தொடர்களில் பச்சை என்ற சொல் மூன்று வேறுபட்ட பொருள்களைக் குறித்து நிற்கிறது. மூன்றில் எந்தப் பொருளை எடுத்துக்கொள்வது என்பது அதன் அடுத்த சொல்லைப் பொறுத்தது. கருத்தாடல் நிலையில், ஏற்படுகிற பொருள் மயக்கத்தைக் கணினிக்குக் கற்றுத்தரமுடியாது. அவற்றை உலகியல் அறிவின் (Pragmatic Knowledge) வாயிலாகவே உணர முடியும்.

மேற்குறித்த பொருள் மயக்கங்களைத் தீர்த்துவைக்கக்கூடிய அறிவை - வழிமுறைகளை எவ்வாறு கணினிக்கு அளிப்பது குறித்து, பல்வேறு நிலைகளில் ஆராயப்பெறுகின்றன.

## பொருள் மயக்கத்தைத் தவிர்ப்பதற்குரிய பொதுவான சில வழிமுறைகள்

கணினிவழித் தமிழ்த் தொடர்களை ஆய்வு செய்யும்போது ஏற்படுகிற பொருள் மயகத்தை நீக்கிப் பொருளைத் தெளிவாகப் புரிந்துகொள்வதற்கு உருபொலியனியல் மாற்றங்கள் துணைபுரிகின்றன. பொருள் மயக்கத்தை இலக்கண வகைப்பாட்டின் வாயிலாகவே தெளிவுபடுத்த முடியும். பெயர், வினை அடிப்படையில் உருவாகும் சொற்களாக உருக்பனியல், தொடரியல் பகுப்பாய்வுகளைக் கொண்டு அடிச்சொல், விகுதிகள் ஆகியவற்றைப் பகுத்துத்தான் இவற்றைச் சரிசெய்ய முடியும்.

‘அவன் நெய்தான் விற்பான்’ என்ற தொடரில், அவன் நெய்யைத்தான்(நெய்+தான்) விற்பான் என்று வேற்றுமை மறைந்துநின்று பொருள்தருகிறதா? அல்லது அவன் துணியை நெய்தான் (நெய்+த்+த்+ஆன்) பிறகு விற்பான் என்ற பொருள்படுகிறதா? என்ற ஐயம் ஏற்படுகிறது. இத்தகைய நிலையில் தொடரியல் ஆய்வின் அடிப்படையிலேயே தெளிவுபெற முடியும்.

அடிச்சொல்லால் ஏற்படுகிற பொருள்மயக்கத்தை விசுவகளைக்கொண்டு தெளிவுபெறலாம். விசுவகளைல் ஏற்படுகிற பொருள் மயக்கத்திற்கு அடிச்சொல்லைக்கொண்டு தெளிவுபெறலாம். எடுத்துக்காட்டாக, ‘படித்தான்’ என்ற சொல்லில் படி என்பது பெயராக வரும்போது படித்தான் என்றும் வினையாக வரும்போது படித்தான் என்றும் வரும் என்பதனை அடிச்சொல் வாயிலாகப் பெறமுடிகிறது. ‘ஆல்’ என்னும் விசுவ ‘அவனால் நான் வந்தேன்’ என்னும் தொடரில் பெயருக்குப் பின் வந்ததால் வேற்றுமை விசுவி என்றும், ‘வந்தால் நான் வருவேன்’ என்னும் தொடரில் வினைக்குப் பிறகு வந்ததால் ஆல் என்பது நிபந்தனை விசுவி என்றும் பகுத்துக் கண்டறியமுடிகிறது.

‘இரு’ என்ற சொல் இருவேறு பொருள்தருகின்றன .அவற்றை இடப்பொருள் அடிப்படையிலேயே சேர்த்தோ பிரித்தோ எழுதமுடியும். விட்டிசைப்பிற்காகவும், வகைப்படுத்துவதற்காகவும், பொருள் தெளிவிற்காகவும் காற்புள்ளி ‘,’ இட்டு எழுதுவது கட்டாயமாகிறது. இதுபோன்ற பல்வேறு மொழிப் பயன்பாட்டு நெறிகள் தமிழ்மொழி இலக்கணங்களிலும் மொழியியல் விதிகளிலும் காணக்கிடைக்கின்றன.

## நிறைவாக

பொருள் மயக்கத்திற்கான அடைப்படைக் காரணங்கள், பொருள் மயக்கம் ஏற்படுவதற்குரிய நிலைப்பாடுகளை மூன்றாகப் பகுத்தும் மொழியியல் வகைப்பாட்டிலும் தகுந்த எடுத்துக்காட்டுகளுடன் ஆராயப்பெற்றன. மேலும், பொருள் மயக்கத்தைத் தவிர்ப்பதற்குரிய பொதுவான சில வழிமுறைகள், கணினிவழித் தமிழாய்வு செய்யும்போது ஏற்படுகிற சிக்கல்களும் ஆராயப்பெற்றன. ஒரு தொடரை எழுதும்போது பெயர், வினை, துணைவினை போன்ற அடிப்படை வேறுபாடுகளை அறிந்து , பயன்படுத்தினால் பல்வேறு மொழிப் பயன்பாட்டுச் சிக்கல்கள் சரிசெய்யப்படும். அனைவரும் ஒரேவிதமான மொழிப் பயன்பாட்டுக்கொள்கையைப் பயன்படுத்துவதன்வழி ,கணினிவழி மொழியாய்வு செய்வதற்கு எளிமையாக இருக்கும். இதுபோன்ற பல்வேறு மொழியமைப்புக் கூறுகளை முறைப்படுத்த வேண்டிய கட்டாயம் ஏற்பட்டுள்ளது என்பதை இக் கட்டுரை சுட்டிக்காட்டுகிறது.

## தேர்ந்தெடுக்கப்பட்ட துணைநூற்பட்டியல்

1. முனைவர் ச .அகத்தியலிங்கம் ,தமிழ்மொழி அமைப்பியல் ,மெய்யப்பன் தமிழாய்வகம் ,சிதம்பரம்.
2. டாக்டர் பொற்கோ, (2006), இக்காலத் தமிழ் இலக்கணம், பூம்பொழில் வெளியீடு ,சென்னை.
3. எம்.ஏ. நுஃமான், (2007), அடிப்படைத் தமிழ் இலக்கணம், அடையாளம் ,திருச்சி.
4. பேரா. கலாநிதி அ. சண்முகதாஸ், (2008), தமிழ்மொழி இலக்கண இயல்புகள், நியூ செஞ்சுரி புக்ஹவுஸ்.
5. முனைவர் செ. வை. சண்முகம், (2004), தொல்காப்பியத் தொடரியல், உலகத்தமிழாராய்ச்சி நிறுவனம்.
6. முனைவர் அ. தாமோதரன், துணைவினைகள் ,ஆய்வுக் கட்டுரை .
7. தமிழ் இணையம் 2010, மாநாட்டுக் கட்டுரைகள்.
8. Dr. M. Suseela, (2001), A Historical Study of Old Tamil Syntax, Tamil University.
9. Thomas Lehman, (1993), A Grammar of Modern Tamil, Pondichery Institute of Linguistics and Culture.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# கணினியில் ரோமன் வரிவடிவ ஒலிபெயர்ப்பு

## முனைவர் இராதா செல்லப்பன்

பேராசிரியர் (ஓய்வு)

பாரதிதாசன் பல்கலைக்கழகம், திருச்சிராப்பள்ளி

ரோமன் வரிவடிவப் பெயர்ப்பு என்பது ரோமன் எழுத்துக்களைப் பயன்படுத்தித் தமிழ் உரைகளை ஒலியன் வடித்தில் எழுதுவது ஆகும். அவ்வாறு எழுதும்போது தமிழ் எழுத்துக்களுக்கான தற்பவ ஒலிகள் கிடைக்காதபோது அணுக்கமான மற்றொரு ஒலியனைப் பயன்படுத்தி எழுதுவதாகும். தமிழ் மொழியில் இத்தகைய முயற்சி புதியதல்ல. தமிழில் கிடைக்கும் முதல் மொழிபெயர்ப்பு நூல் பட்டனார் கீதை என்பர். அந்நூலில் சமஸ்கிருதப் பாடல்களைத் தமிழில் ஒலிபெயர்த்துள்ளனர். தமிழில் இல்லாத ஒலிகளான ஸ,ஷ, ஜ, ஹ, ஸ்ரீ ஆகியவற்றை ஒலிபெயர்த்துள்ளபோது கிரந்த எழுத்துக்களைப் பயன்படுத்தினர். தமிழில் இல்லாத வர்க்க எழுத்துக்களை எழுத எண்முறை பயன்படுத்தப்பட்டுள்ளது.

பரித்தாரணாய ஸாதா4னாம் வினாசாய ச1 து3ஷ்க்ருதாம்!

த4ர்மஸம்ஸ்தா2பனார்த்தா2ய ஸம்ப4வாமி யுகே3 யுகே3 !!

ரோமன் வரிவடிவத்தில் தமிழ்ப் பனுவல்களை எழுதும் வழக்கம் பலராலும் பன்னெடுங்காலமாகக் கையாளப்பட்டு வருவதாகும். ஆங்கிலத்திலே கட்டுரை எழுதும் தமிழறிஞர்கள் தமிழ்ச் சிறப்புப் பெயர்களையும் (ஆட்பெயர், ஊர்ப்பெயர் முதலானவை) சொற்களையும் ரோமன் எழுத்தில் எழுதுகின்றனர். தமிழருக்கே உரிய நாகரிகப் பண்பாட்டுச் சொற்களையும் ஒலிபெயர்த்து எழுதுகின்றனர். கல்வி நிலையில் தமிழை ஆங்கிலம்வழிக் கற்க விரும்புவோர் ரோமன் வடிவில் தமிழை எழுதிக் கற்றனர். தமிழ் வரிவடிவம் தெரியாதவர்களுக்குத் தமிழ் கற்பிக்கவும் ரோமன் வரிவடிவத்தைப் பலரும் பயன்படுத்தினர். அவர்களுள் முக்கியமாக, வீரமாமுனிவர், போப் ஆகியோரைக் குறிப்பிடலாம். மொழியியல் அறிஞர்கள் ரோமன் வரிவடிவத்தை மிகவும் அதிகமாகப் பயன்படுத்துகின்றனர். ஒலிச்சான்றுகளையும் ஒலியன் சான்றுகளையும் காட்டவும் சொற்களைச் சான்றுகளாகக் காட்டவும் பிற மொழிச் சொற்களோடு ஒப்பிட்டுக் காட்டவும் எனப் பல்வேறு நிலைகளில் ரோமன் வரிவடிவத்தைப் பயன்படுத்துகின்றனர். அவர்களுள் குறிப்பிடத்தக்க மேனாட்டவர்களாக, ராஸ்மஸ் ராஸ்க், கால்டுவெல் ஆகியோரைக் குறிப்பிடலாம். தமிழாய்விலே ஆர்வமுடைய வெளிநாட்டவரும் பிற திராவிட மொழியாராய்ச்சியாளர்களும் தமிழை ரோமன் எழுத்தில் எழுதிப் பயன்படுத்தினர். தமிழர் பிற நாடுகளில் இரண்டு, மூன்று அல்லது நான்கு பரம்பரைகளுக்கு முன் குடியேறி வாழ்ந்து வரும் நிலையும் உள்ளது. அவர்கள் தம் தாய்மொழியாம் தமிழை நன்கு பயன்படுத்த இயலாதவர்களாக உள்ளனர். பேசத் தெரிந்த அளவிற்கு அவர்களுக்கு எழுதவோ படிக்கவோ பயிற்சி கிடைப்பதில்லை. அச்சூழலில் வாழும் அவர்கள் நமது பாரம்பரிய பக்திப் பாடல்களையும் பனுவல்களையும் தமிழ் எழுத்துக்களாலன்றி ரோமன் எழுத்துக்களாலேயே அறிகின்றனர். எனவே அவர்களுக்கு ரோமன் வரிவடிவ ஒலி பெயர்ப்பு மிகவும் தேவைப்படுகிறது. சான்றாக, மொரீஷியஸ் தென்னாப்பிரிக்கா முதலிய நாடுகளில் வழிபாட்டுப் பாடல்களை ரோமன் எழுத்துக்களில் எழுதிப் பயன்படுத்துகின்றனர். மேலைநாட்டு இந்தியவியல் ஆய்வாளர்கள் ரோமன் எழுத்துகளில் எழுதுவதில் ஆர்வம் காட்டி வந்தனர். அக்காலம் தற்காலக் கணினிக்கு முன்னர் உள்ள காலம். அவற்றைத் தரப்படுத்த வேண்டும் என்ற நோக்கமும் அவர்களிடையே 1888 களிலேயே விவாதிக்கப்பட்டதாக அறிகிறோம்.

ரோமன் எழுத்துக்களைப் பயன்படுத்தித் திராவிட மொழிகளை எழுதும் முறை ஏறத்தாழ ஒரு நூற்றாண்டிற்கு முன்பிருந்தே பயன்படுத்தப்பட்டது. இன்றையக் கணினி யுகத்திலும் ரோமன்

வரிவடிவம் மிகவும் தேவைப்படும் ஒன்று. தமிழ் இலக்கியங்களைப் பாரறியச் செய்யும் நோக்கில் பலரும் முயன்று வருகின்றனர். காட்டாக, மதுரைத்திட்டம் தமிழ் இலக்கியங்களை ரோமன் வடிவப் பெயர்ப்பிலும் தருகிறது. லைப்ரரி ஆப் காங்கிரசும் அமெரிக்க ஐரோப்பிய நாடுகளிலுள்ள அவர்களுடைய நூலகங்களும் வலையத்தில் தம் நூல் பட்டியலை, தேடிப்பார்க்கும் வசதியுடன் இட்டுள்ளன. Digital Dictionaries of South India என்ற தளத்தில் பல மொழி அகராதிகள் உள்ளன. இவற்றில் சொற்களைத் தேட ரோமன் எழுத்துக்களும் பயன்படுத்தப்படுகின்றன. தமிழில் எழுதுவதற்கு அதிகப் பழக்கமில்லாத தமிழர்களும் மற்றும் மயங்கொலிகளைப் பற்றிய தெளிவில்லாதவர்களும் ரோமன் வரிவடிவத்தைப் பயன்படுத்துகின்றனர். இவ்வாறான தேடலுக்கு வசதியாக பிளெய்ன் ஆஸ்க்கி வடிவங்கள் அவற்றிற்கான சிறப்புக் குறியீடுகளின்றிப் பயன்படுத்தப்படுகின்றன. கணினியிலே தமிழ்த் தட்டச்சு செய்யப் பயிற்சி இல்லாதவர்கள் தங்கள் கருத்துக்களை ரோமன் வடிவத்தில் எழுதி அவற்றைக் கணினியிலிருந்து தமிழ் எழுத்துக்களில் பெறுகின்றனர். கணினி வழி அச்சிற்கும் பயன்படுகிறது. தேடு பொறிகளில் தமிழ்ச் சொல்லைத் தேடுபவர்களுக்கும் அகராதிகளில் சொல்லைத் தந்து பொருள் தேடுபவர்களுக்கும் ரோமன் வரிவடிவம் ஒரே சீராக இருக்க வேண்டுவது இன்றியமையாதது.

ரோமன் எழுத்துக்களில் எழுதி அவற்றைத் தமிழ் எழுத்துக்களில் பெறுதல் என்ற நிலையில் பல மென்மங்கள் தற்போது உருவாக்கப்பட்டுள்ளன. இதில் தமிழில் பெறும் உரையில் திருத்தங்கள் செய்ய வேண்டுமென்றால் ஆங்கில உரைக்குச் சென்று திருத்த வேண்டும். தற்போது திரு என்ற ஒரு மென்மானது இரு திரைகளை உருவாக்கி மேல் திரையில் ஆங்கிலமும் கீழ்த் திரையில் தமிழும் எழுத வழி வகுத்துள்ளது. மற்றொரு மென்மம் ஆதவின் என்பது. இத்தகைய ரோமன் வடிவ உள்ளீடு மென்மங்களுள் அழகி என்பதும் குறிப்பிடத்தக்க ஒன்று. இதில் தமிழ் எழுத்துக்களை ரோமன் எழுத்துக்களில் பெறும் வசதி உள்ளது. சினிமா துறையினர் இதனை அதிகமாகப் பயன்படுத்துகின்றனர். Universal Digital Library என்பதில் நூல்களின் பெயர்கள் ரோமன் எழுத்துக்களிலும் தமிழ் எழுத்துக்களிலும் தரப்பட்டுள்ளன.

ரோமன் வடிவப் பெயர்ப்பில் முதன்முதலில் திட்டம் வகுத்த நிறுவனம் லைப்ரரி ஆப் காங்கிரஸ் என்பதாகும். முக்கியமான தமிழாய்வு நிறுவனங்கள் பலவும் லைப்ரரி ஆப் காங்கிரஸின் வரிவடிவ முறையையே ஏற்றுப் பயன்படுத்தின. அவற்றுள் முக்கியமாக ஆசியவியல் நிறுவனமும் ரோஜா முத்தையா தமிழாய்வு நூலகமும் குறிப்பிடத்தக்கன.

- 1926-36 களில் சென்னைப் பல்கலைக்கழகத்தால் வெளியிடப்பட்ட தமிழ்ப் பேரகராதி பிளென் ஆஸ்கி வடிவத்தை அடிப்படையாகக் கொண்ட வரிவடிவத்தைப் பயன்படுத்தியது. கொலோன் பல்கலைக்கழகத்தின் இந்திய மற்றும் தமிழாய்வு நிறுவனம் இந்த வரிவடிவத்தையே ஏற்று சங்க இலக்கியம் முதலான பழங்கால இலக்கியங்களை ஒலிபெயர்த்துள்ளது. ஆதவின், மதுரைத் திட்டம் ஆகியவை ஒலியிணைகளைப் பயன்படுத்துகின்றன.
- ITRANS- ஆங்கிலச் சிறிய எழுத்துக்களையும் சில சிறப்புக் குறியீடுகளையும் பயன்படுத்தின . இந்தத் திட்டம் 1912-இல் ஏதென்சில் நடந்த கீழைத்தேயத்தாரின் பன்னாட்டுக் கழக மாநாட்டின் பரிந்துரையை ஒட்டியது.
- ISO 15919 இந்திய மொழிகளுக்கான ரோமன் குறியீட்டு முறையைத் தந்துள்ளது. அதுவே பின்னர் மேல் கீழ் குறியீடுகளைத் தவிர்க்கும் முறையில் பக்கவாட்டில் குறியீடுகளை அமைத்து மாற்று முறையைத் தந்தது.
- பன்னாட்டு ஒலி நெடுங்கணக்கு முறையிலும் ரோமன் ஒலிபெயர்ப்புகள் நடைபெறுகின்றன.

- பென்சில்வேனியா பல்கலைக்கழகம், மதுரைத் திட்டம், கோலன் பல்கலைக்கழகம், கூகில் நிறுவனம் தயாரித்த கூகில் இண்டிக் டிரான்ஸ்லிட்டரேஷன் ஆகியனவும் ரோமன் வரிவடிவத்தில் தமிழ் இலக்கியங்கள், அகராதிகள் முதலானவற்றை உருவாக்கி வருகின்றன.

இவ்வாறு கணினிப் பயன்பாட்டில் ரோமன் வரிவடிவம் பயன்படுத்தப்படும் நிலையில், ரோமன் வரிவடிவ ஒலிபெயர்ப்பில் பலவகையான வேறுபாடுகள் காணப்படுவதைக் காணலாம். ஒலிபெயர்ப்பு முறையில் ஒரு சீர்மை இல்லை ஒரே எழுத்து பல்வேறு வகைகளில் பெயர்க்கப்படுவதைக் காண முடிகிறது. அவற்றை ஒட்டுமொத்தமாக ஆராய்ந்தால் அவற்றிற்கிடையே சில வடிவங்களில் ஒருமைப்பாடும் சிலவற்றில் வேறுபாடும் உள்ளதை அறிய முடிகிறது. இதற்கான முக்கியமான காரணங்களாக இருப்பவை வருமாறு. தமிழிலுள்ள எழுத்துக்களில் சில ரோமன் எழுத்து முறையில் இல்லை. உயிரெழுத்துக்களைப் பொறுத்த வரையில் ஆங்கிலம் குறில் நெடில் ஆகிய இரண்டு எழுத்துக்களுக்குமே ஒரு வடிவத்தையே பயன்படுத்துகின்றன. ண, ன, ந, ஆகியவற்றை வேறுபடுத்தும் வகையிலும், ல, ழ, ள ஆகியவற்றை வேறுபடுத்தும் வகையிலும், ர, ற ஆகியவற்றை வேறுபடுத்தும் வகையிலும் தனித்தனி எழுத்துக்கள் இல்லை. ங, ஞ ஆகிய எழுத்துக்களுக்கும் தனி வரிவடிவம் இல்லை. எனவே இவற்றை ரோமன் எழுத்தில் குறிக்கப் பல உத்திகளை ஒலிபெயர்ப்பாளர்கள் கையாண்டனர். முதலிலே தாளில் எழுதியவர்கள் அல்லது தட்டச்சு செய்தவர்கள் அந்த நோக்கத்திலேயே இந்த வேறுபாடுகளை நீக்கும் வகையில் குறியீடுகளைக் கையாண்டனர். அதிலும் ரோமன் வடிவத்தைக் கையாண்டு ஒலிபெயர்த்த வீரமாமுனிவர் கையாண்ட முறைகளுள் சில பின்னோரால் எடுத்துக் கொள்ளப்பட்டன; சில மாற்றம் பெற்றன; சில புதிய குறியீடுகள் உருவாக்கப்பட்டன.

குறில் உயிரெழுத்துக்கள்

- குறில் உயிர்களில் பெரும்பாலும் வேறுபாடு இல்லை. a i u e o இம்முறை பெஸ்கி, கால்டுவெல் தொடங்கிப் பல அறிஞர்களாலும் பயன்படுத்தப்பட்ட முறை. வீரமாமுனிவர் முறை ISO (1) 1591 மற்றும் KOELN-ஆல் பயன்படுத்தப்பட்டது. google Indic பெரியெழுத்தையும் சின்னவெழுத்தையும் சேர்த்துப் பயன்படுத்தியது.

|   | வீர<br>மாமு<br>னிவர் | கால்<br>டுவெ<br>ல் | போ<br>ப் | தெ.<br>பொ<br>மீ | ப<br>ரோ | ISO<br>(1)<br>1591<br>9 | ISO<br>(2) | Penn.U<br>ty | TL | L<br>C | Madu<br>rai | KOE<br>LN | google<br>Indic | IT<br>RA<br>NS |
|---|----------------------|--------------------|----------|-----------------|---------|-------------------------|------------|--------------|----|--------|-------------|-----------|-----------------|----------------|
| அ | A                    | a                  | a        | a               | a       | A                       | a          | a            | a  | a      | a           | A         | a/Aa            | a              |
| இ | I                    | i                  | i        | i               | i       | I                       | i          | i            | i  | i      | i           | I         | i/I             | i              |
| உ | U                    | u                  | u        | u               | u       | U                       | u          | u            | u  | u      | u           | U         | u/U             | u              |
| எ | E                    | e                  | e        | e               | e       | e                       | e          | e            | e  | e      | e           | E         | e/E             | e              |
| ஓ | O                    | o                  | o        | o               | o       | o                       | o          | o            | o  | o      | o           | O         | o/O             | o              |

- ஐ-ஒரு சீரான பெயர்ப்பு என்றாலும் கால்டுவெல் தவிர ஏனையோர் அனைவருமே ai என எழுதினர். கால்டுவெல் மட்டுமே ei என்று எழுதினார்.



google Indic அதனுடன் பெரிய எழுத்துக் குறியீட்டையும் சேர்த்துப்

பயன்படுத்தியது

|   | வீர<br>மா<br>முனி<br>வர் | கால்<br>டு<br>வெ<br>ல் | போ<br>ப் | தெ.<br>பொ<br>.மீ | ப<br>ரோ | ISO<br>(1)<br>1591<br>9 | ISO<br>(2) | Penn.<br>Uty | TL | L<br>C | Mad<br>urai | KOE<br>LN | Goog<br>le<br>Indic | ITRAN<br>S |
|---|--------------------------|------------------------|----------|------------------|---------|-------------------------|------------|--------------|----|--------|-------------|-----------|---------------------|------------|
| ஐ | ai                       | ei                     | ai       | ai               | ai      | ai                      | ai         | ai           | ai | ai     | ai          | ai        | ai/AI               | ai         |

- ஒள- au எனப் பயனாளிகளால் ஒரு சீர்மையுடன் பயன்படுத்தப்பட்டுள்ளது.

google Indic அதனுடன் பெரிய எழுத்தையும் சேர்த்துப் பயன்படுத்தியது.

|    | வீர<br>மா<br>மு<br>னி<br>வர் | கால்<br>டுவெ<br>ல் | போ<br>ப் | தெ.<br>பொ<br>.மீ | ப<br>ரோ | ISO<br>(1)<br>1591<br>9 | ISO<br>(2) | Penn.<br>Uty | TL | LC | Mad<br>urai | KOEL<br>N | google<br>Indic | ITRAN<br>S |
|----|------------------------------|--------------------|----------|------------------|---------|-------------------------|------------|--------------|----|----|-------------|-----------|-----------------|------------|
| ஒள | au                           | au                 | au       | au               | au      | au                      | au         | au           | au | au | au          | au        | au/A<br>U       | au         |

- நெட்டுயிர்கள்

|   | வீர<br>மாமு<br>னிவர் | கால்<br>டுவெ<br>ல் | போப் | தெ.<br>பொ<br>.மீ | ப<br>ரோ | ISO<br>(1)<br>159<br>19 | ISO<br>(2) | Pen<br>n.U<br>ty | TL | LC | M<br>a<br>d<br>u<br>r<br>a<br>i | KOE<br>LN | goog<br>le<br>Indic | ITRA<br>NS |
|---|----------------------|--------------------|------|------------------|---------|-------------------------|------------|------------------|----|----|---------------------------------|-----------|---------------------|------------|
| ஆ | Ā                    | ā                  | ā    | ā                | ā       | ā                       | -a         | aa/<br>A         | ā  | ā  | A                               | A         | aa/<br>Aw           | aa, A<br>ā |
| ஈ | Ī                    | ī                  | ī    | ī                | ī       | Ī                       | -i         | ii/ I            | ī  | ī  | I                               | I         | ii/II               | ii,I<br>ī  |
| உ | ū                    | ū                  | ū    | ū                | ū       | ū                       | -u         | uu/<br>U         | ū  | ū  | U                               | U         | uu/<br>UU           | uu,U<br>ū  |
| ஏ | Ē                    | ē                  | ē    | ē                | ē       | ē                       | -e         | ee/<br>E         | ē  | ē  | E                               | E         | ee/E<br>E           | E          |
| ஓ | ō                    | ō                  | ō    | ō                | ō       | ō                       | -o         | oo/<br>O         | ō  | ō  | O                               | O         | oo/<br>OO           | O          |

- ஆ--- ā a: aa -a A Aw
- ஈ--- ī i: ii -i I II
- ஊ-- ū u: uu -u U UU
- ஏ--- ē e: ee -e E EE
- ஓ-- ō o: oo -o O OO

மதுரைத் திட்டமும் கோலோனும் நெட்டுயிர்களைக் குறிக்கப் பெரிய எழுத்துக்களையே பயன்படுத்துகின்றன என்பது குறிப்பிடத்தக்கது.

மெய்யெழுத்துக்களில்

- க ச த ப - இவற்றின் ஒலிப்பெழுத்துக்களும் மூச்சொலிகளும் பெரும்பாலும் ஒரு சீராகப் பெயர்க்கப்பட்டுள்ளன.
  - க ஒருசீரான மொழிபெயர்ப்பு- K. அதுவே ஒலிப்பொலியாக வரும்போது g, இரு உயிர்களுக்கிடையே வரும்போது h மூச்சொலியாக வரும்போது kh, gh.
  - ச - c அதுவே ஒலிப்பொலியாக வரும்போது j, இரு உயிர்களுக்கிடையே வரும்போது s மூச்சொலியாக வரும்போது ch.
  - த - t, d அதுவே மூச்சொலியாக வரும்போது th, dh
  - ப - ஒருசீரான மொழிபெயர்ப்பு- p. அதுவே ஒலிப்பொலியாக வரும்போது b, மூச்சொலியாக வரும்போது ph, bh.

|   | வீர<br>மா<br>முனி<br>வர் | கால்<br>டு<br>வெ<br>ல் | போ<br>ப் | தெ.<br>பொ<br>.மீ | ப<br>ரோ | ISO<br>(1)<br>159<br>19 | ISO<br>(2) | Pen<br>n.Ut<br>y | T<br>L | L<br>C | Mad<br>urai | KO<br>EL<br>N | google<br>Indic | ITRA<br>NS     |
|---|--------------------------|------------------------|----------|------------------|---------|-------------------------|------------|------------------|--------|--------|-------------|---------------|-----------------|----------------|
| க |                          |                        |          |                  |         | k                       | k          | k                | k      | K      | K           | K             | k/g             | k,kh,g<br>,gh  |
| ச | s                        | ch,j                   | c,j      | c                | C       | c                       | c          | c□               | c      | c      | c           | C             | ch              | c,ch           |
| த | t                        | t, d                   | t,th     | t                | t       | t                       | t          | t th             | t      | t      | t           | T             | th              | ta,th,d<br>,dh |
| ப | p                        | p, b                   | p, b     | p                | p       | p                       | p          | p                | p      | p      | p           | P             | p               | p,ph,b<br>,bh  |

- ம, ய, ர, ல, வ ஆகியவை பெரும்பாலும் ஒரு சீராக முறையே m, y, r, l, v

எனப் பெயர்க்கப்பட்டுள்ளன. KOELN நிறுவனம் இவற்றை பெரிய எழுத்துக்களாக அதாவது M, Y, R, L, V எனப் பெயர்த்துள்ளது.

|   | வீர<br>மா<br>மு<br>னிவர் | கா<br>ல்டு<br>வெ<br>ல் | போ<br>ப் | தெ.<br>பொ<br>.மீ | ப<br>ரோ | IS<br>O<br>(1)<br>159<br>19 | ISO (2) | Pe<br>n<br>n.<br>Ut<br>y | TL | L<br>C | Ma<br>dur<br>ai | KO<br>EL<br>N | googl<br>e<br>Indic | ITR<br>AN<br>S |
|---|--------------------------|------------------------|----------|------------------|---------|-----------------------------|---------|--------------------------|----|--------|-----------------|---------------|---------------------|----------------|
| ம | m                        | M                      | m        | m                | m       | m                           | m       | m                        | m  | m      | m               | M             | m                   | m              |
| ய | y                        | Y                      | y        | y                | y       | y                           | y       | y                        | y  | y      | y               | Y             | y                   | y              |
| ர | r                        | R                      | r        | r                | r       | r                           | r       | r                        | r  | r      | r               | R             | r                   | r              |
| ல | l                        | L                      | l        | l                | l       | l                           | l       | l                        | l  | l      | l               | L             | l                   | l              |
| வ | v                        | V                      | v        | v                | v       | v                           | v       | v;<br>W                  | v  | v      | v               | V             | v                   | v              |

○ ங ஞ ட ண ந ன ழ ள ற -பல்வேறு பெயர்ப்புகள் காணப்படுகின்றன. இவ்வேறுபாடுகளுக்குக் காரணம் இவ்வெழுத்துக்களான இணைகள் ரோமன் வரிவடிவத்தில் இல்லை என்பதே. இவற்றைக் குறிக்க கீழ்க்காணும் முறைகள் பயன்படுத்தப்படுகின்றன.

- நெருங்கிய தொடர்புடைய ரோமன் எழுத்தின் முன்னும் பின்னும் பக்கவாட்டிலும் சிறப்புக் குறியீடுகளை அமைத்தல்

|   | வீர<br>மாமு<br>னிவர் | கால்<br>டுவெ<br>ல் | போ<br>ப் | தெ.<br>பொ<br>.மீ | ப<br>ரோ | ISO<br>(1)<br>1591<br>9 | ISO (2) | Penn.U<br>ty | TL | LC | M<br>ad<br>ur<br>ai | KO<br>ELN | goo<br>gle<br>Indi<br>c | ITR<br>ANS |
|---|----------------------|--------------------|----------|------------------|---------|-------------------------|---------|--------------|----|----|---------------------|-----------|-------------------------|------------|
| ண | n                    | □                  | □        | □                | □       | η                       | #n      | N            | η  | .  | N                   | N         | N                       | Na<br>.    |
| ன | n                    | N                  | .        | .                | .       | .                       | _n      | n            | n  | .  | n_<br>/n<br>2       | n_<br>/n2 | nZ                      | n          |
| ந | n                    | N                  | n        | n                | n       | n                       | n       | nd; n;<br>n^ | .  | n  | n'                  | N         | n                       | n          |

○ ண,ன ஆகிய தமிழ் எழுத்துக்களுக்குரிய இணைகள் இல்லை. அவை இரண்டும் ஒரே எழுத்தான **n** என்பதில் மேல், கீழ் , முன், பின் என பல்வேறு சிறப்புக் குறியீடுகளால் குறிக்கப்பெறுகின்றன.

- ண -n □ #n □ N Na
- ன - n □ \_n n2 nZ

- ல, ள, ழ - 1 என்ற எழுத்துடன் சில குறியீடுகளைச் சேர்த்துக் குறிக்கப்படுகின்றன. அத்துடன் பெரிய சிறிய எழுத்துக்களும் பயன்படுத்தப்படுகின்றன.

|          | வீர<br>மாழு<br>னிவர் | கால்<br>டுவெ<br>ல் | போ<br>ப் | தெ.<br>பொ<br>.மீ | ப<br>ரோ | ISO<br>(1)<br>1591<br>9 | ISO<br>(2) | Penn.<br>Uty | TL | LC | Mad<br>urai | KO<br>EL<br>N | googl<br>e<br>Indic | ITRA<br>NS                      |
|----------|----------------------|--------------------|----------|------------------|---------|-------------------------|------------|--------------|----|----|-------------|---------------|---------------------|---------------------------------|
| <b>ட</b> | d                    | □, □               | d, □□    | □                | □       | □                       | #t         | T; d         | □  | □  | T           | T             | T                   | T,Th,D<br>,Dh<br>□, □h<br>d, dh |
| <b>ள</b> | l                    | □                  | □        | □                | □       | □                       | #1         | L            | □  | □  | L           | L             | L                   | L, □                            |
| <b>ழ</b> | l                    | □                  | □        | □                | □       | □                       | _1         | z; zh        | □  | □  | z           | Z             | LZ/z                | z                               |

- ல - l, L
- ள - □, #1 L
- ழ - l, □, □ □, \_1 zh, z, Z LZ

○ ர,ற

|   | வீர<br>மாழு<br>னிவர் | கால்<br>டுவெ<br>ல் | போப்  | தெ.<br>பொ<br>.மீ | ப<br>ரோ | ISO<br>(1)<br>1591<br>9 | ISO<br>(2) | Penn<br>.Uty | TL | LC | Mad<br>urai | KOE<br>LN | google<br>Indic | IT<br>R<br>A<br>N<br>S |
|---|----------------------|--------------------|-------|------------------|---------|-------------------------|------------|--------------|----|----|-------------|-----------|-----------------|------------------------|
| ர | r                    | R                  | r     | r                | r       | r                       | r          | r            | r  | r  | r           | R         | r               | r                      |
| ற | rr                   | ·, ·r              | ·, ·· | ·                | ·       | ·                       | _r         | R            | ·  | ·  | R           | R         | Ra              | R                      |

○ - r

- ர - r, R
- ற - rr □, □r □ r, t, d, \_r, R, Ra

○ ங, ஞ n எழுத்துடன் சிறப்புக் குறியீடுகளைச் சேர்த்தோ பெரிய சிறிய எழுத்துக்களாலோ அவை குறிக்கப்படுகின்றன.

|   | வீர<br>மாழு<br>னிவர் | கால்<br>டுவெ<br>ல் | போப் | தெ.<br>பொ<br>.மீ | ப<br>ரோ | ISO<br>(1)<br>1591<br>9 | ISO<br>(2) | Penn<br>.Uty | TL | LC | Mad<br>urai | KOE<br>LN | goog<br>le<br>Indic | ITRA<br>NS     |
|---|----------------------|--------------------|------|------------------|---------|-------------------------|------------|--------------|----|----|-------------|-----------|---------------------|----------------|
| ங | ng                   | ·                  | ñ    | Ñ                | ñ       | ·                       | ^n         | ng           | ·  | ·  | ng          | G         | NG                  | ~N,<br>N^<br>· |
| ஞ | nj                   | Ñ                  | ñ    | ñ                | ñ       | ñ                       | ~n         | nj           | ñ  | ñ  | n·          | n^/j<br>n | NY                  | JN<br>ñ        |

- ங - ng, ·, Ñ, ^n, η, G, NG, ~N, N^
- ஞ -nj, ñ, ~n, n·, n· n^/jn, NY, JN

• ஆங்கிலப் பெரிய எழுத்துக்களையும் சிறிய எழுத்துக்களையும் பயன்படுத்துதல்  
மேலே குறிப்பிட்டபடி, கணினியில் பயன்படுத்தப்படும் ரோமன் வரிவடிவங்களில் ஒருமைப்பாடு இல்லாத காரணத்தால் பயனாளிகள் சிரமத்திற்காளாகின்றனர். காட்டாக, மணி என்ற சொல்லை லைப்ரரி காங்கிரசின் தளத்திலும் கலிவொர்னியா பல்கலைக்கழகத் தளத்திலும் **mani** என எழுதித் தகவலைப் பெறலாம். ஆனால் கொலான் பல்கலைக்கழகத் தளத்தில் **maNi** என எழுதினால்தான் தகவலைப் பெற முடியும். எனவே பயனாளிகள் ஒரு தகவல் தேடலைச் செய்யும்போது தாம் எந்த வடிவ முறையைப்

பயன்படுத்தும் தளத்தில் இருக்கிறோம் என்பதை அறிந்த பிறகுதான் செயலாற்ற முடியும். எனவே அவற்றைத் தரப்படுத்தவது இன்றைய கால கட்டத்தின் தேவையாக ஆகி உள்ளது.

|                     |                      |               |                |
|---------------------|----------------------|---------------|----------------|
| பாற்கடல் p āka al   | -ISO (1) 15919 தமிழ் | tami          | -ISO (1) 15919 |
| p a_rka#tal         | - ISO (2)            | tami_l        | - ISO (2)      |
| pɑː rkatal          | - IPA                | tami[ɹ]       | - IPA          |
| paaRkaDal, pARkaDal | - Penn Uty           | tamiz/tamizh  | - PennUty      |
| p ā · ka · al       | - TL                 | tami ·        | - TL           |
| p ā · ka · al       | - LC                 | tami ·        | - LC           |
| pARkaTal            | - Madurai            | tamiz         | - Madurai      |
| pARkaTal            | - KOELN              | tamiz         | - KOELN        |
| pARkaTal, paaRkaTal | - Adami              | thamiz/tamizh | - Adami        |

எனவே பல்வேறு வடிவ முறையில் இருக்கும் அவற்றைத் தரப்படுத்துவது என்பது மிகவும் தேவையான ஒன்றாக ஆகியுள்ளது .இவ்வாறு தரப்படுத்தும் முயற்சியில் கவனம் செலுத்தவேண்டியவை வருமாறு; ஏற்கனவே இருக்கும் குறியீடுகளுள் பொருத்தமானவற்றைத் தேர்ந்தெடுத்துத் தரப்படுத்தலாம்.

1. புதியதானதொரு முறையை உருவாக்கிப் பரிந்துரைக்கலாம்.
2. பின்னோக்கு ஒலிபெயர்ப்பிலும் சரியான சொல் கிடைக்கிறதா என்று உறுதி செய்துகொள்ள வேண்டும். **Universal Digital Library** தளத்தில் நூல்களின் பெயர்கள் ரோமன் எழுத்துக்களிலும் தமிழ் எழுத்துக்களிலும் தரப்பட்டுள்ளன. அருமுக நாவலர் ஸரித்திரம்... by அருணசல கவிராயர் **L: Tamil, Y: 1898, S: ARUMUGA NAVALAR SARITHIRAM, 98 pgs.** ஆங்கிலத்தில் முதலில் உள்ளீடு செய்து அதனைத் தமிழ்ப்படுத்தியபோது ஆறுமுக நாவலர் சரித்திரம், அருணாசலக் கவிராயர் ஆகியவை புரிந்துகொள்ள இயலாத அளவிற்கு வடிவம் மாறியிருப்பது புலனாகும்.
3. ரோமன் எழுத்துக்கு முன்னும் பின்னும், மேலும் கீழும் கோடுகள் அல்லது புள்ளிகளைப் பயன்படுத்தும் முறையை மறு ஆய்வு செய்ய வேண்டும். அத்துடன் பெரிய எழுத்துக்களும் இடம்பெறுகின்றன. இவ்வாறு அமையும்போது வரிகள் அழகின்றி ஒரு சீராக அமைவதில்லை. அத்துடன் அவற்றைத் தட்டச்சு செய்யும் வகையில் கணினி விசைகள் அமையவில்லை என்பது மிக முக்கியமானதொரு பிரச்சனை. எனவே இந்த ரோமன் வரிவடிவத்தில் தட்டச்சு செய்ய ஒரு தனி மென்பொருளோ தனிப்பயிற்சியோ தேவை.
4. கணினிப் பயன்பாட்டில் ரோமன் வடிவ ஒலிபெயர்ப்பைப் பற்றி முடிவு செய்வது தனிமனித முடிவாக அமையாமல் குழு முயற்சியாக அமைய வேண்டும். இதற்கென ஒரு பணிக்குழு செயலாற்றித் ஒலிபெயர்ப்பு முறைகளைத் தரப்படுத்த வேண்டும்.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# **Electronic Tamil Dictionary**

(மின் அகராதி)





# Agaraadhi: A Novel Online Dictionary Framework

*Elanchezhiyan.K, Karthikeyan.S, T V Geetha,*

*Ranjani Parthasarathi & Madhan Karky*

*{chezhiyank@gmail.com, sethuramankarthikeyan@gmail.com, madhankarky@gmail.com}*

*Tamil Computing Lab (TaCoLa),*

*College of Engineering Guindy, Anna University, Chennai.*

## Abstract

This paper describes Agaraadhi, a dictionary framework for indexing and retrieving Tamil words, their meaning, analysis and related information. With a database of over 3 lack root words and their corresponding meaning in English and Tamil, this paper proposes a framework to encompass various features such as morphological analysis, morphological generation, word usage statistics, word pleasantness analysis, spell checking, similar word finder, word usage in literature, picture dictionary, number to text conversion, phonetic transliteration, live usage analysis from micro blogs and more. Describing various components of the framework the paper concludes with a discussion over dictionary statistics and possible features for future extension of the framework.

## 1. Introduction

Most of the Tamil dictionaries are synonym based and they do not give enough information such as morphological analysis of the word, possible case endings for requested word, pleasantness score, word usage in the web and social networks, equivalent words or meaning etc. To overcome these issues we propose Agaraadhi, a framework. Agaraadhi Framework consists of a Morphological analyser, Morphological generator, Word pleasantness and Word usage score finder as well as analysis of current usage in Social Networks, Picture dictionary, equivalent Tamil words, Generator (Word suggestions), Spell checker, Phonetic transliteration, Number to Text Converter, Rare-Word of the day and Social Network sharing.

Agaraadhi dictionary has more than 3 lac words in various domains such as General, Literature, Medical, Engineering, Computer Science, etc. The Agaraadhi framework dictionary is a Tamil English bilingual dictionary. The following sections describe the framework and list the benefits of such a framework over traditional online Tamil-English dictionaries. A few features proposed in this framework such as popularity score for a word, to best of our knowledge, are not present in any other world dictionaries.

## 2. Agaraadhi Framework

Agaraadhi Dictionary Framework was designed to provide additional information to the user regarding the word that they query about. Agaraadhi framework presented in figure 1 can be divided

into two major divisions, online and offline, in terms of the time of processing. This section describes the various components used in the Agaraadhi framework in detail.

## 2.1 Online Process

Any user query is sent sequentially to dictionary and literature, to retrieve corresponding data from those indices, fetching phonetic transliteration from transliteration modules, morphological information from morphological analysis and generator module and fetching live usage analysis from micro blogs. All those Information are sent to user interface pages, shown in fig 1.

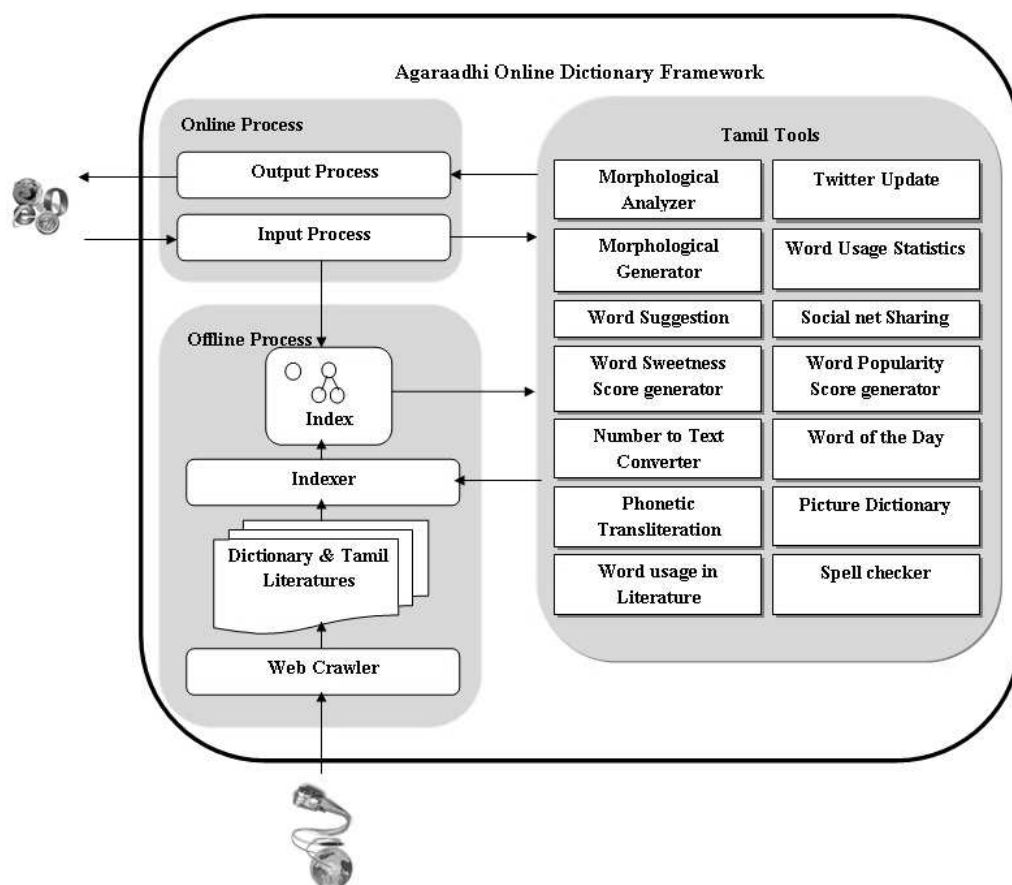


Fig 1: Agaraadhi Online Dictionary Framework

## 2.2 Offline Process

Tamil words and their meanings are entered manually and stored in text files. Those words are sequentially sent to modules such as popularity score generator, pleasantness score generator, picture dictionary, phonetic transliteration module and the resulting information is abstracted as a word object. Tamil literature such as Bharathiyaar songs, Avvaiyar songs, Thirukkural and lyrics are crawled using a static web crawler and are indexed in hash table as key value pairs.

## 2.3 Features of Agaraadhi Framework

Agaraadhi dictionary framework consists of more than twenty features such as Morphological Analysis Morph Generation, pleasantness scoring, popularity scoring, spell error suggestions etc.

### 3.1 Morphological Analyser

Morphological analyser [1] chunks the query word and gives the morphological features of the query word such as root word, parts of speech, gender, tense and count. If the Query word is *padithaan*, Morphological Analyser gives as *padi* as root, word represents male gender and query word is past tense and so on.

### 3.2 Morphological Generator

A Tamil morphological generator[2] needs to tackle different syntactic categories such as nouns, verbs, post positions, adjectives, adverbs etc. separately, since the addition of morphological constituents to each of these syntactic categories depends on different types of information. The generator is used to generate possible morphological variations of the query word.

### 3.3 Spell Checker

Spell Checker is used to check the spelling of Tamil words and to provide alternative suggestions for the wrong words. It uses the Morphological Analyzer. The Morphological Analyzer is used to split the given Tamil word into the root word and a set of suffixes. If the word is fully split by the analyzer and its root word is also found in the Agaraadhi dictionary, the given word is termed as correct. Otherwise, the correction process is invoked to generate all the possible suggestions with minimum variations from the given word.

### 3.4 Word Suggestions

Word Suggestion gives the list of equivalent or related words for the given query word.

### 3.5 Word Pleasantness and Word Popularity Score

Word Pleasantness score generator provides how easy to pronounce the word.

Word Popularity shows the word usage in the web. The Word from agaraadhi is given to web and found the frequency distribution of the word across the popular blogs, news articles, social nets etc.

### 3.6 Word Usage in Literature

This feature finds the usage of words in popular literature such as Thirukural, Bharathiyar Padalgal, Avvai songs and Lyrics.

### 3.7 Number to Text Converter

It converts a number to Tamil word equivalent as well as in English text. For example in Tamil we represent *oru Arpputham* (அற்புதம்) for 100 million, *Kumbam* (கும்பம்) for 10 billion and finally up to *Anniyan* (அந்நியம்) for one zillion.

### 3.8 Phonetic Transliteration

The pronunciation of words in Tamil and English language, as distinct from their written form based on the phonology and it can also vary greatly among dialects of a language. Phonetic transliteration module splits the word into syllables and gives the transliteration for each syllable.

### **3.9 Picture Dictionary**

Pictures, photos or line drawings to depict popular words have been included in the dictionary to enable efficient learning for children using this tool.

### **3.10 Social net Sharing and Twitter Update**

The framework also provides features to format results to be shared effectively on social networks. An Agaraadhi Bot was designed to post updates and word of the day on Twitter automatically.

### **3.11 Word of the Day and Word Usage statistics**

A rare word is randomly chosen and is displayed in the opening page to facilitate users to learn a new word every day.

Word Usage Statistics [3] shows the usage of the word in the social network over the past one week.

### **3.12 Tamil Word Games**

Games play a vital role in learning. Currently Agaraadhi has two Tamil word games namely Miruginajambo and Thookku Thookki. Miruginajambo is an unscramble game and Thookku Thookki is a Hangman game in Tamil.

## **4. Conclusion and Future Work**

This paper describes Agaraadhi, an Online Dictionary Framework. Agaraadhi online dictionary is a bilingual dictionary containing over 3 lac words on various domains like General, Medical, Engineering, Computer science, Literature etc. This Online Dictionary framework encompass various features such as morphological analysis, morphological generation, word usage statistics, word pleasantness analysis, spell checking, similar word finder, word usage in literature, picture dictionary, number to text conversion, phonetic transliteration, live usage analysis from micro blogs etc. Providing APIs for programmers and developing mobile apps for Agaraadhi framework will open a good platform for many researchers and developers working in Tamil Computing area.

## **References**

- Anandan, R. Parthasarathi, and Geetha, Morphological Analyser for Tamil. ICON 2002, 2002.
- Anandan, R. Parthasarathi, and Geetha, Morphological Generator for Tamil. Tamil Inayam, Malaysia, 2001.
- J. Jai Hari Raju, P. IndhuReka, Dr. Madhan Karky, Statistical Analysis and visualization of Tamil Usage in Live Text Streams, Tamil Internet Conference, Coimbatore, 2010.

# நவீன தமிழ் அகராதி

முனைவர் க. தமிழ்ச்செல்வன்,

இணைப் பேராசிரியர்,

சமூக அறிவியல் மொழிப்பள்ளி, வி.ஐ.டி. பல்கலைக்கழகம்,

வேலூர் - 632 014., வேலூர் மாவட்டம். தமிழ்நாடு, இந்தியா.

மின்னஞ்சல் : ktamilselvan@vit.ac.in

## முன்னுரை:

இன்றைய தொழில்நுட்ப யுகத்தில், புதிய புதிய கண்டுபிடிப்புகள் உருவான வண்ணம் உள்ளது. இதன் காரணமாகப் புதிய புதிய சொற்கள் உருவான வண்ணம் இருக்கிறது. உதாரணமாக, Cell Phone-ஐ செல்பேசி என்றும், கைப்பேசி என்றும் அழைக்கின்றோம். இதுபோல, தமிழில் உள்ள பழைய வார்த்தைகளுக்குப் பொருள் என்ன என்று அறிவதும் மிகுந்த தேவையாகும். உதாரணமாக (களிறு - யானை). மாணவர்களுக்கு இதுபோன்ற இலக்கியத் தரமானச் சொற்களுக்குப் பொருள் என்னவென்று தமிழ் அகராதி மூலம் விளக்குவது மிகுந்த தேவையாகும்.

இந்த நவீன தமிழ் அகராதி மென்பொருளில், ஒரு வார்த்தையைக் கணிப்பொறியில் தட்டச்சு செய்யத் தொடங்கும்போதே, முழு வார்த்தையை வழங்கும் வண்ணம் வடிவமைக்கப்படும். இதற்குப் புதிய தொழில்நுட்பம் பயன்படுத்தப்படும். இந்த நவீன தமிழ் அகராதியைத் தமிழ் - தமிழ் - ஆங்கிலம் என்றோ அல்லது ஆங்கிலம் - தமிழ் - தமிழ் என்றோ இருவேறு முறைகளில் எப்படி வேண்டுமானாலும் பயன்படுத்த முடியும். இதற்கான வசதியும் புதிய தொழில்நுட்பத்தைக் கொண்டு வடிவமைக்கப்படும்.

இந்த நவீன தமிழ் அகராதி மென்பொருள் கீழ்காணும் முறைகளில் வார்த்தைகளுக்குப் பொருளை வழங்கும்.

## நிலை - 1 :

முதலில் தேவையான மொழியைத் தேர்வு செய்ய வேண்டும். பின்னர் வார்த்தைகளைக் கணிப்பொறியில் தட்டச்சு செய்யும்போதே முழு வார்த்தையையும், அதற்கு அருகாமை வார்த்தைகளையும் வழங்கும். இந்த வசதியால், வார்த்தைகளை எளிதாகத் தேர்வு செய்யலாம்.

# நவீன தமிழ் அகராதி

◉ தமிழ்

◉ English

வார்த்தையைத் தட்டச்சு செய்யவும்

|      |    |
|------|----|
| அ    | Go |
| அ    |    |
| அழகு |    |
| அஃது |    |

உதாரணமாகத் தமிழ் மொழியைத் தேர்வு செய்து, "அ" என்ற எழுத்தைத் தட்டச்சு செய்தால், கீழ்க்கண்டவாறு அதன் அருகாமை வார்த்தைகளை வழங்கும். தேவையான வார்த்தையை தேர்வு செய்தவுடன் அதன் தமிழ்ப் பொருளையும், ஆங்கிலப் பொருளையும் பெறலாம்.

## நவீன தமிழ் அகராதி

◌ தமிழ்      ◌ English

வார்த்தையைத் தட்டச்சு செய்யவும்

பொருள் : தமிழ் நெடுங்கணக்கின் (அகரவரிசையின்) முதலெழுத்து

The first letter of Tamil alphabet

நிலை - 2 :

ஆங்கில மொழியைத் தேர்வு செய்து, "a" என்ற எழுத்தைத் தட்டச்சு செய்தால், கீழ்க்கண்டவாறு அதன் அருகாமை வார்த்தைகளை வழங்கும். தேவையான வார்த்தையைத் தேர்வு செய்தவுடன் அதன் ஆங்கிலப் பொருளையும், தமிழ்ப் பொருளையும் பெறலாம்.

## நவீன தமிழ் அகராதி

◌ தமிழ்      ◌ English

Enter your word

a  
abacus  
abbey

# நவீன தமிழ் அகராதி

தமிழ்

English

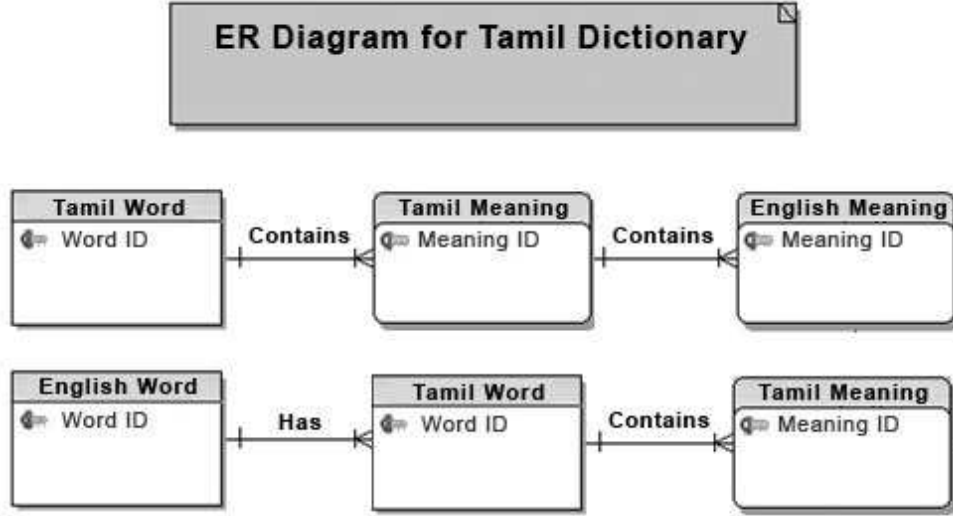
Enter your word

Go

பொருள் : The first letter of the English alphabet

ஆங்கில நெடுங்கணக்கின் (அகரவரிசையின்) முதலெழுத்து

தொழில்நுட்ப செயலாக்க வரைபடம் :



திட்ட உருவாக்கத்தின் படிநிலைகள் :

| வரிசை எண். | படிநிலைகள்   |
|------------|--|
| 1          | மென்பொருளின் மாதிரியை உருவாக்குதல் (Prototype)                             |
| 2          | இந்த மாதிரி வடிவத்திற்கு ஒப்புதல் பெறுதல் (Prototype Approval)             |
| 3          | மென்பொருளுக்கான கணிணி நிரல்களை உருவாக்குதல்                                |
| 4          | நவீன தமிழ் அகராதிக்கான டேட்டா பேசை உருவாக்குதல்                            |
| 5          | டேட்டாபேசை மென்பொருளுடன் இணைத்தல்  |
| 6          | மென்பொருளின் தரத்தை ஆய்வு செய்தல் (QA Testing)                             |
| 7          | நவீன தமிழ் அகராதியை இணையத்தில் நிறுவுதல்                                   |
| 8          | இணைய வசதி இல்லாத கணினியிலும் பயன்படுத்தும் Desktop Application வெளியிடுதல் |
| 9          | திட்டத்தை வெற்றிகரமாக முடித்தல்.   |



**உருவாக்கத்திற்கான மென்பொருள்கள் :**

| S.No. | Components         | Software   |
|-------|--------------------|--|
| 1     | Operating System   | Windows 2000 Professional or XP or above   |
| 2     | Application Server | ASP.NET 2.0, AJAX  |
| 3     | Web Tools          | HTML , Javascript  |
| 4     | Database           | MS SQL Server 2005   |
| 5     | Browser            | Internet Explorer, Mozilla Firefox, Netscape, Google chrome - with unicode support |

இந்த மென்பொருள் **Microsoft Visual Studio.NET 2005 with .NET Framework 2.0** என்ற நவீன தொழில்நுட்பத்தைக் கொண்டு உருவாக்கப்படும். இந்த மென்பொருள் **MS SQL 2005 Server** என்ற சிறப்பான, வேகமாகச் செயலாற்றக்கூடிய டேட்டாபேஸ் பயன்படுத்தித் தயாரிக்கப்படும்.

இந்த இரு மென்பொருள்களின் இணைப்பில் உருவாக்கப்படும் நவீன தமிழ் அகராதி அதிக வேகத்தில் தகவல்களைத் தரும்.

**தரம் :**

நவீன தமிழ் அகராதியானது 100% சதவீதம் கணிணி நிரல் குறைகள் அற்றதாக உருவாக்கப்படும். இதற்காக மிகச் சிறந்த வகையில் தரக்கட்டுப்பாட்டு ஆய்வு மேற்கொள்ளப்படும்.

மென்பொருள் உருவாக்கத்தில் ஒவ்வொரு படிநிலையிலும் தரக்கட்டுப்பாட்டு ஆய்வு மேற்கொண்டு **100% Bug Free** மென்பொருளாக உருவாக்கப்படும்.

# மொழிபெயர்ப்புக் கலையில் அகராதியின் பயன்பாடு

இளங்குமரன் த/பெ சிவநாதன்

சுல்தான் இட்ரிஸ் கல்வியியல் பல்கலைக்கழகம், மலேசியா

E-mail: s.ilangkumaran@gmail.com

மொழிபெயர்ப்புப் பணிகளில் அகராதிகளின் பயன்பாடு இன்றியமையாததாகின்றது. இருப்பினும் சிற்சில வேளைகளில் பெரிதும் அகராதிகளையே நம்பி செய்யப்படும் மொழிபெயர்ப்புப் பணிகள் அதன் இயற்கைத்தன்மையைக் காட்டத் தவறி, ஒருவித செயற்கை உணர்வை மொழிபெயர்ப்புப் பணிகளில் வெளிப்படுத்துகின்றன. சிறந்த மொழிபெயர்ப்பு பணி எனில், அவை மொழிபெயர்க்கப்பட்டவை என்ற உணர்வைத் தராது, குறிப்பிட்டதொரு மொழியிலேயே படைக்கப்பட்ட படைப்பு எனும் உணர்வையே படிப்போர்க்கு ஏற்படுத்த வேண்டும்.

இதன் அடிப்படையிலேயே அகராதியின் பயன்பாடு ஒரு மொழிபெயர்ப்புப் பணியில் எவ்வளவில் பயன்படுத்தப்படுதல் சிறப்பு; மற்றும், அதனைக் கையாளும் வழிமுறைகள் என்ன என்பது குறித்தே இவ்வாய்வு மேற்கொள்ளப்பட்டுள்ளது. இதன் மூலம் மொழிபெயர்ப்பாளர்களும், மொழிபெயர்ப்புத் துறை மாணவர்களும் கவனிக்கத் தவறவிட்ட சில விஷயங்களும், தகவல்களும் வெளிப்படுத்தப்பட முயற்சி செய்யப்பட்டுள்ளது.

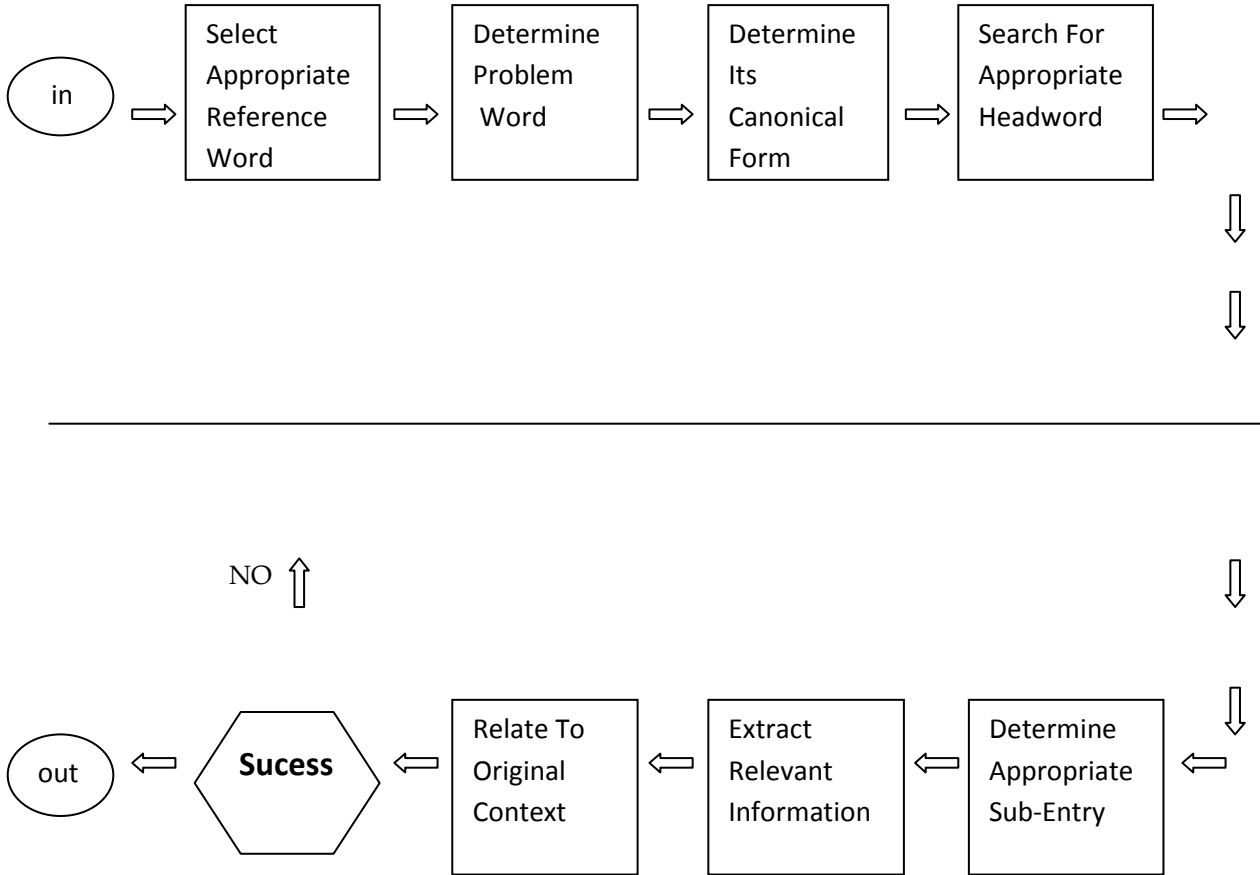
கால காலமாக நாம் பயன்படுத்தி வரும் புத்தக வடிவில் இருக்கும் மரபுவழி அகராதிகள் எந்த அளவிற்குத் இன்றைய நாட்களில் மொழிபெயர்ப்புப் பணிகளுக்கு உதவுகின்றன என்பது சிந்திக்க வேண்டிய விஷயமாகவே இருக்கின்றது. நாளுக்கு நாள் வளர்ந்து வரும் உலகில், பல புதிய கண்டுபிடிப்புகளுக்கும், ஆராய்ச்சிகளுக்கும் ஏற்ற அகர முதலிகள் மிகவும் இன்றியமையாததாக இருக்கின்றன. கையால், பழைய அகராதிகளையும், மரபு வழி அகராதிகளையும் தவிர புதிய அகராதிகளும், மின்னியல் அகராதிகளும் இன்றைய சூழலில் மிகவும் முக்கியத்துவம் வாய்ந்தவையாக இருக்கின்றன. இவை புத்தக வடிவில் மட்டுமின்றி இணையத்திலும், கையடக்க மின்னியல் அகராதிக் கருவியாகவும் உருமாற்றம் அடைந்து வருகின்றன; அது வரவேற்கத்தக்கதும் ஆகும்.

பல்வேறு ஆய்வுகளின் வாயிலாக, பொதுவாக அகராதிகளைப் பயன்படுத்துவோர் அதன் பயன்பாட்டை முழுமையாக அறியாமல் இருக்கின்றனர் என்று தெளிவுபடுத்தி இருக்கின்றன. உதாரணத்திற்கு, •பௌலி (Fawley) (1990) அவர்கள் கூறியதாவது, அகராதிகளைப் பயன்படுத்துவோர் மிக மிகக் குறைந்த அளவிலேயே அதன் பயன்பாட்டினை உணர்ந்துள்ளனர். அவர்கள் வெறுமனே சொற்களின் நேரடிப் பொருளை அறியவும் சரியான எழுத்துருக்களை அறிந்துக்கொள்ளவுமே அகராதிகளைப் பயன்படுத்தி வருகின்றனர். மாறாக சொல்லுருவாக்கம், உச்சரிக்கும் விதம், சரியான முறையில் பயன்படுத்திக் காட்டப்பட்டிருக்கும் வாக்கியங்கள், அச்சொல்லுக்கு ஏற்ற எதிர்ச்சொற்கள் போன்ற பல்வேறு குறிப்புகளில் மக்கள் அக்கறை கொள்வதே இல்லை என்பது அவரது குற்றச்சாட்டு. இதனாலேயே பெரும்பாலானோர் தாங்கள் தேடும் சொற்களுக்குச் சரிவர அல்லது போதிய தகவல்களைப் பெற தவறிவிடுகின்றனர். அதன் விளைவாக அவர்கள் தங்கள் படைப்புகளில் அவற்றைப் பிரயோகம் செய்யும்போது தவறானதொரு வார்த்தையைப் பயன்படுத்தி தொடர்ந்து வாசகர்களையும் குழப்பத்தில் ஆழ்த்தி விடுகின்றனர்.

இந்நிலை, ஒரு மொழியில் படைப்புகளை வெளியிடும் எழுத்தாளர்களுக்குச் சிரமத்தை விளைவிக்கின்றன என்றால், மொழிபெயர்ப்பாளர்களுக்கு அதைக்காட்டிலும் மிகப்பெரிய சுமையை ஏற்படுத்தி விடுகின்றன; ஏனெனில், படைக்கப்பட்டிருக்கும் மொழியில் பயன்படுத்தப்பட்டிருக்கும்

சொல்லுக்குச் சரியான பொருளை அறிந்துக்கொள்ளும் அதே வேளையில், தான் மொழிபெயர்க்க விரும்பும் மொழியில் அதற்குத் தகுந்த சொல்லைத் தெரிவு செய்ய வேண்டியவர்களாகவும், தொடர்ந்து அக்கட்டுரை படைக்கப்பட்டிருக்கும் சூழல், துறை ஆகியவற்றைக் கருத்தில் கொண்டு அந்தந்த துறைக்கும் சூழலுக்கும் ஏற்றாற்போல் தம் மொழிபெயர்ப்பைத் தரக்கடவர்களாகவும் மொழிபெயர்ப்பாளர்கள் இருக்கின்றனர்.

அகராதிகளின் பயன்பாடு குறித்து எழுந்துள்ள ஆய்வுகளில் மிக முக்கிய ஆய்வாகக் கருதப்படும் ஹார்த்மேன் (Hartmann) (1989) அவர்களின் ஆய்வு மொழிபெயர்ப்பாளர்கள் தாங்கள் மொழிபெயர்க்க விரும்பும் சொற்களுக்குச் சிறந்த முறையில் அகராதிகளில் பொருள்கொள்ள ஒரு கட்டமைப்பை உருவாக்கினார். அது :



Hartmann (1989) : Sociology of the dictionary user :Hypothesis and Empirical Studies, Worterbucher Dictionaries Dictionnaires [Art 12], Walter de Gruyter, Berlin, New York Vol. 1 : 102-111

### மொழிபெயர்ப்பாளர்களின் எண்ணங்களும் கருத்துக்களும்

1. எந்த மாதிரியான அகராதிகளையும் தேர்ந்தெடுத்து உபயோகிக்கலாம்.

- மொழிபெயர்ப்பாளர்களில் பெரும்பாலானோர் மிகவும் பிரசித்தி பெற்ற, மக்கள் மத்தியில் அதிகம் பேசப்படக்கூடிய அகராதிகளைப் பயன்படுத்துவதிலேயே ஆர்வம் காட்டுகின்றனர். மேலும் தங்களின் ஆசிரியர்கள் மற்றும் மொழிபெயர்ப்புத் துறை நண்பர்கள் அறிமுகப்படுத்தும் அல்லது ஊக்குவிக்கும் அகராதிகளைப் பயன்படுத்தத் தொடங்கும் மொழிபெயர்ப்பாளர்களில் பலர், கடைசி வரை தங்களைக்

காலத்துக்கேற்ப புதுபித்துக்கொள்ளாமலேயே கடைசி வரை மொழிபெயர்ப்புப் பணிகளில் தொடர்ந்து ஈடுபடுகின்றனர்.

2. கையடக்க அகராதிகளைப் பயன்படுத்துவது இலகுவானது.

- சில மொழிபெயர்ப்பாளர்கள் கையடக்க அகராதிகளைப் பயன்படுத்துவதில் பெரிதும் ஆர்வம் காட்டுகின்றனர். “மொழிபெயர்ப்பாளர்களாக விளங்கும் நாங்கள் எங்கு சென்றாலும் எங்களது அகராதிகளைக் கொண்டு செல்ல வேண்டியுள்ளது; ஏனெனில், அவ்வப்போது எங்களின் திறமைகளில் நம்பிக்கை வைத்து நேரிலும் தொலைபேசிகளிலும் அதிகமானோர் அணுகி தங்களது சந்தேகங்களுக்கு விளக்கம் கோருகின்றனர். அவர்களின் சந்தேகங்களை நிவர்த்திக்கும் பொருட்டு நாங்கள் எப்போதும் அகராதிகளுடனேயே இருக்கிறோம்” என சில தரப்பினர் கூறுகின்றனர். இன்னும் சிலர், குறிப்பாக மொழிபெயர்ப்புத் துறையில் நீண்ட காலம் பயிற்சி பெற்ற மொழிப்பெயர்ப்பாளர்களும் தங்களின் நற்பெயர் கலங்கப்படாதிருக்க மக்களின் சந்தேகங்களைக் களையும் நோக்கில் இவ்வாறு செயல்படுவதும் வருத்தமளிக்கின்றது. எல்லோருக்கும் எல்லா விஷயங்களும் தெரிந்திருக்க நியாயம் இல்லை என்பதை உணராதது, தெரியாதவற்றைத் தெரியவில்லை என பகிரங்கமாக ஒப்புக்கொள்ளும் தைரியம் இல்லாமல் போவது ஒரு புறமிருக்க, குறிப்பிட்ட வார்த்தைகளுக்குச் சரியான விளக்கங்கள் தான் அளிக்கிறோமா என்ற தெளிவும் அற்று ஒருவித குழப்பத்தையும் சமயங்களில் இது போன்றவர்கள் ஏற்படுத்துகின்றனர். இதுபோன்ற கையடக்க அகராதிகள் மாணவர்களுக்குப் பெருமளவில் பயன்படுகிறதேயொழிய மொழிபெயர்ப்பாளர்களுக்கு அந்த அளவிற்குப் பயன்படுவதில்லை. (இருப்பினும் கையடக்க மின்னியல் அகராதி இதிலிருந்து விதிவிலக்காகின்றது என்பதை அறிக)

2. அகராதிகளில் குறிப்பிடப்பட்டிருக்கும் சொற்களைத் தாராளமாகப் பயன்படுத்தலாம்

- பெரும்பாலான மொழிபெயர்ப்பாளர்கள் அகராதிகளில் குறிப்பிடப்பட்டிருக்கும் சொற்களையும், விளக்கங்களையும் தாராளமாகப் பயன்படுத்தலாம் என எண்ணம் கொண்டிருக்கின்றனர். இதனாலேயே சில சமயங்களில் நடைமுறைக்கு ஒவ்வாத தவறான மொழிபெயர்ப்புப் பணிகளை நாம் பார்க்க முடிகின்றது. மேலும் இதுபோல் அகராதிகளிலிருந்து எடுக்கப்பட்ட நேரடி வார்த்தைகள் சில வேளைகளில் சம்பந்தப்பட்ட கட்டுரை படைக்கப்பட்டிருக்கும் சூழலுக்கும், அவை படைக்கப்பட்டிருக்கும் துறைக்கும் சற்றும் பொருந்தாமல் போவது இங்கு குறிப்பிடத்தக்கது. உதாரணத்திற்கு இணையத்தில் பரவலாகப் பயன்படுத்தப்படும் Browse என்ற வார்த்தைக்கு அகராதியின் வாயிலாக நேரடிப் பொருள் கொள்ளும்போது, இளந்தளிர் உணவு, கிளை தழை, பசுந்தீவனம், தழை மேய்தல் மற்றும் புற்கறித்தல் என்ற பொருள்களைத் தருகின்றது. ஆனால், உண்மையில் இச்சொல் உணர்த்தவரும் பொருள் வலம் வருதல், அணுகுதல் போன்றவையாகும். இந்நிலையில் இச்சொல் பயன்படுத்தப்பட்டிருக்கும் சூழலையும் அதன் துறையையும் அறியாது மொழிபெயர்க்கப்பட்டிருக்கும் படைப்புகள் உகந்த பொருளைத் தர தவறுவதோடு அதைப் படிப்பவர்களுக்குப் பெருங்குழப்பத்தை ஏற்படுத்திவிடுகின்றது.

3. நமக்குத் தெரிந்த விஷயம்தானே என்ற போக்கு

- சில வேளைகளில் மொழிபெயர்ப்புப் பணிகளில் ஈடுபடும் சிலர் இது நமக்குத் தெரிந்த விஷயம்தானே, இதற்காகவெல்லாம் அகராதியைப் புரட்டவேண்டியதில்லை என்ற எண்ணமும் கொண்டு செயல்படுகின்றனர். பொதுவாக மொழிபெயர்க்கப்படப்போகும் மொழிகளில் பாண்டித்தியம் பெற்றவர்களே மொழிபெயர்ப்புகளைச் செய்வதால் இத்தகைய சிந்தனையால் பெரிதாகப் பிரச்சனை ஏதும் எழாது என்று எண்ணத் தோன்றுகிறது. இருப்பினும், சில வேளைகளில் நுண்ணிய விஷயங்களை மொழி பெயர்க்கும்போது பல கோணங்களில் அவற்றைப் பகுத்துப் பார்ப்பது இன்றியமையாததாகின்றது. உதாரணத்திற்கு 1996-ம் ஆண்டு மலேசிய விமானச் சேவையின் மொழிபெயர்ப்புப் பணியை ஏற்று

முடித்த மொழிபெயர்ப்பாளர் ஒருவர் பின்னர் அந்நிறுவனம் மேற்கொண்ட சட்ட நடவடிக்கையால் (மான நஷ்ட வழக்கு) திவாலாகும் நிலையை அடைந்தது குறிப்பிடத்தக்கது. விமானப் பயணத்தின்போது நெருக்கடி நிலை ஏற்படுமாயின் பின்பற்ற வேண்டிய இலகுவான வழிவகைகள் குறித்து சீன மொழியில் மொழிபெயர்க்க வேண்டியிருந்த பணியில், சற்றே கவனக்குறைவாக இலகுவாக நெருக்கடி நிலை ஏற்படக்கூடிய இவ்விமானப் பயணத்தில் பின்பற்ற வேண்டிய வழிவகைகள் என்று தவறுதலாக மொழிபெயர்த்து பின்னர் பெரும் சிக்கலுக்கு உல்லான அம்மொழிபெயர்ப்பாளர் அதற்கு முன்னர் ஏராளமான மொழிபெயர்ப்புப் பணிகளில் ஈடுபட்டு அவற்றைச் செவ்வனே முடித்தவர் என்பது வழக்கு விசாரணையில் தெரிந்தது. இதைவிடக் குறிப்பாக ஆரம்பக்காலங்களில் சின்ன சின்ன விஷயங்களுக்கும் அகராதியின் துணைகொண்டு பொருளை அறிந்த பின்னரே மொழிபெயர்க்கும் அவர் காலப்போக்கில் அகராதியின் பயன்பாடு குறைந்து போக, தனக்கு தெரிந்தது தானே என்று தன் அனுபவத்தை முற்றிலுமாக நம்பி செயல்பட்டதே இந்த தவறுக்குக் காரணம் என விசாரணையில் ஒப்புக்கொண்டதும் குறிப்பிடத்தக்கது.

4. எந்த வகையான மொழிபெயர்ப்புகளையும் செய்யலாம்.

- சில மொழிபெயர்ப்பாளர்கள் தங்களுக்குக் கிடைக்கும் எவ்வகையான பணிகளையும் செய்து விடலாம் என்ற எண்ணம் கொண்டுள்ளனர். இது சரியல்ல. சில குறிப்பிட்ட துறைகளில் மிகுந்த திறமை கொண்டுள்ள ஒருவர் மற்ற துறைகளிலும் விற்பன்னராக இருப்பார் என்று எண்ணுவது தவறு. மொழிபெயர்ப்புகளில் பல பிரிவுகள் உண்டு. அவை சட்டத்துறை மொழிபெயர்ப்புகள், மருத்துவ மொழிபெயர்ப்புகள், கணினி மொழிபெயர்ப்புகள், பொருளாதாரத்துறை மொழிபெயர்ப்புகள், விளம்பர மொழிபெயர்ப்புகள் போன்ற பல பிரிவுகளாலான துறைகளில் மொழிபெயர்ப்புப் பணிகளை மேற்கொள்ள பல வகையான திறமைகள் தேவைப்படுகின்றன. கையால் 'இதனை இதனால் இவன் முடிக்கும்' என ஆராய்ந்து அவற்றைச் சம்பந்தப்பட்டவர்களிடம் ஒப்படைப்பதே உசிதம். மொழி பெயர்ப்பாளர்களும் பணத்தை மட்டுமே குறியாகக் கொள்ளாது மொழிபெயர்ப்பின் தரத்தைக் காக்க ஆவன செய்ய கடமைப்பட்டவர்களாவர்.

5. பலதரப்பட்ட அகராதிகளைப் பயன்படுத்துதல்

- பலதரப்பட்ட அகராதிகளைப் பயன்படுத்துதல் ஒரு மொழிபெயர்ப்பாளரைப் பொருத்த வரையில் மிக மிக வரவேற்கக்கூடிய ஒன்றாக இருப்பினும், கவனக்குறைவு ஏற்படவும் இதில் பெரிய வாய்ப்பு உள்ளதை பெரும்பாலான மொழிப்பெயர்ப்பாளர்கள் உணர தவறுகின்றனர். எப்படி? பல துறைகளில் மொழிப்பெயர்ப்புப் பணிகளை மேற்கொள்ளும் மொழிபெயர்ப்பாளர்கள் ஒன்றுக்கும் மேற்பட்ட அகராதிகளைத் துணைக்கு வைத்திருப்பது இயற்கையே. நீண்ட காலத்திற்குச் செய்யப்படும் மொழிபெயர்ப்புப் பணிகளில் சில சமயங்களில் அப்பணியின் ஆரம்பத்தில் பிரயோகிக்கப்பட்ட குறிப்பிட்டதொரு வார்த்தை அப்பணி முடிவடையும்போது வேறு வார்த்தைப் பிரயோகத்தில் முடிவதைப் பார்க்க முடிகின்றது. உதாரணத்திற்கு, 100 பக்கங்களைக் கொண்ட ஒரு மொழிபெயர்ப்புப் பணியை ஒருவர் ஒரே நாளில் செய்து முடித்துவிடுவதென்பது அரிய காரியம். இவ்வாறு நான்கு அல்லது ஏழு நாட்களுக்குத் தொடரும் பணியில், ஆரம்பத்தில் பயன்படுத்திய அகராதியை விடுத்து பிரிதொரு அகராதியின் துணைகொடலின்போது பிரிதொரு வார்த்தையை அம்மொழிப்பெயர்ப்பாளர் பிரயோகிக்க வாய்ப்புண்டு. இது இணைய அகராதிகளைப் பயன்படுத்துவோரிடமும் அதிகம் நேருகின்றது; ஏனெனில், இவர்களைப் போன்றவர்கள் நான்கு அல்லது ஐந்து இணைய தளத்தினை உதவிக்குத் துணை கொள்பவர்களாக இருக்கின்றனர். இவ்வாறான தவறுகள் பொருளாதாரத்துறை மொழிபெயர்ப்புகளிலும் ண்டறிக்கைகளிலும் பெருமளவு காணப்படுகின்றது.

க, மொழிபெயர்ப்பாளர்கள் அகராதிகளைப் பயன்படுத்துவதில் மிகவும் விழிப்பாக இருக்க வேண்டும். பொருந்தாத அகராதிகளின் பயன்பாடும், அதிகமான மற்றும் குறைவான பயன்பாடுகளும் கூட தவறான மொழிபெயர்ப்புகளுக்கு வித்திட்டுவிடும். கையால் இவ்விஷயத்தில் மிகுந்த கவனம் தேவை. மேலும், முறையான அகராதிகளின் பயன்பாடுகள் குறித்து பொதுவாக பள்ளிகளிலிருந்தும், தவிர மொழிபெயர்ப்புக் கல்விகளைப் போதிக்கும் கல்விக்கூடங்களும் முறையாக போதிக்க வேண்டும். மாணவர்களுக்கு அகராதிகளின் முழுமையான பயன்பாட்டினைப் போதிக்கும் பட்சத்தில் வருங்காலங்களில் சிறப்பான மொழிபெயர்ப்புகள் மட்டும் அன்றி தரமான படைப்புகளை உருவாக்குவதற்கும் அது வழிவகுக்கும் என்பதில் ஐயமில்லை.





## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011



# Agaraadhi: A Novel Online Dictionary Framework

*Elanchezhian.K, Karthikeyan.S, T V Geetha,*

*Ranjani Parthasarathi & Madhan Karky*

*{chezhiyank@gmail.com, sethuramankarthikeyan@gmail.com, madhankarky@gmail.com}*

*Tamil Computing Lab (TaCoLa),*

*College of Engineering Guindy, Anna University, Chennai.*

## Abstract

This paper describes Agaraadhi, a dictionary framework for indexing and retrieving Tamil words, their meaning, analysis and related information. With a database of over 3 lack root words and their corresponding meaning in English and Tamil, this paper proposes a framework to encompass various features such as morphological analysis, morphological generation, word usage statistics, word pleasantness analysis, spell checking, similar word finder, word usage in literature, picture dictionary, number to text conversion, phonetic transliteration, live usage analysis from micro blogs and more. Describing various components of the framework the paper concludes with a discussion over dictionary statistics and possible features for future extension of the framework.

## 1. Introduction

Most of the Tamil dictionaries are synonym based and they do not give enough information such as morphological analysis of the word, possible case endings for requested word, pleasantness score, word usage in the web and social networks, equivalent words or meaning etc. To overcome these issues we propose Agaraadhi, a framework. Agaraadhi Framework consists of a Morphological analyser, Morphological generator, Word pleasantness and Word usage score finder as well as analysis of current usage in Social Networks, Picture dictionary, equivalent Tamil words, Generator (Word suggestions), Spell checker, Phonetic transliteration, Number to Text Converter, Rare-Word of the day and Social Network sharing.

Agaraadhi dictionary has more than 3 lac words in various domains such as General, Literature, Medical, Engineering, Computer Science, etc. The Agaraadhi framework dictionary is a Tamil English bilingual dictionary. The following sections describe the framework and list the benefits of such a framework over traditional online Tamil-English dictionaries. A few features proposed in this framework such as popularity score for a word, to best of our knowledge, are not present in any other world dictionaries.

## 2. Agaraadhi Framework

Agaraadhi Dictionary Framework was designed to provide additional information to the user regarding the word that they query about. Agaraadhi framework presented in figure 1 can be divided

into two major divisions, online and offline, in terms of the time of processing. This section describes the various components used in the Agaraadhi framework in detail.

## 2.1 Online Process

Any user query is sent sequentially to dictionary and literature, to retrieve corresponding data from those indices, fetching phonetic transliteration from transliteration modules, morphological information from morphological analysis and generator module and fetching live usage analysis from micro blogs. All those Information are sent to user interface pages, shown in fig 1.

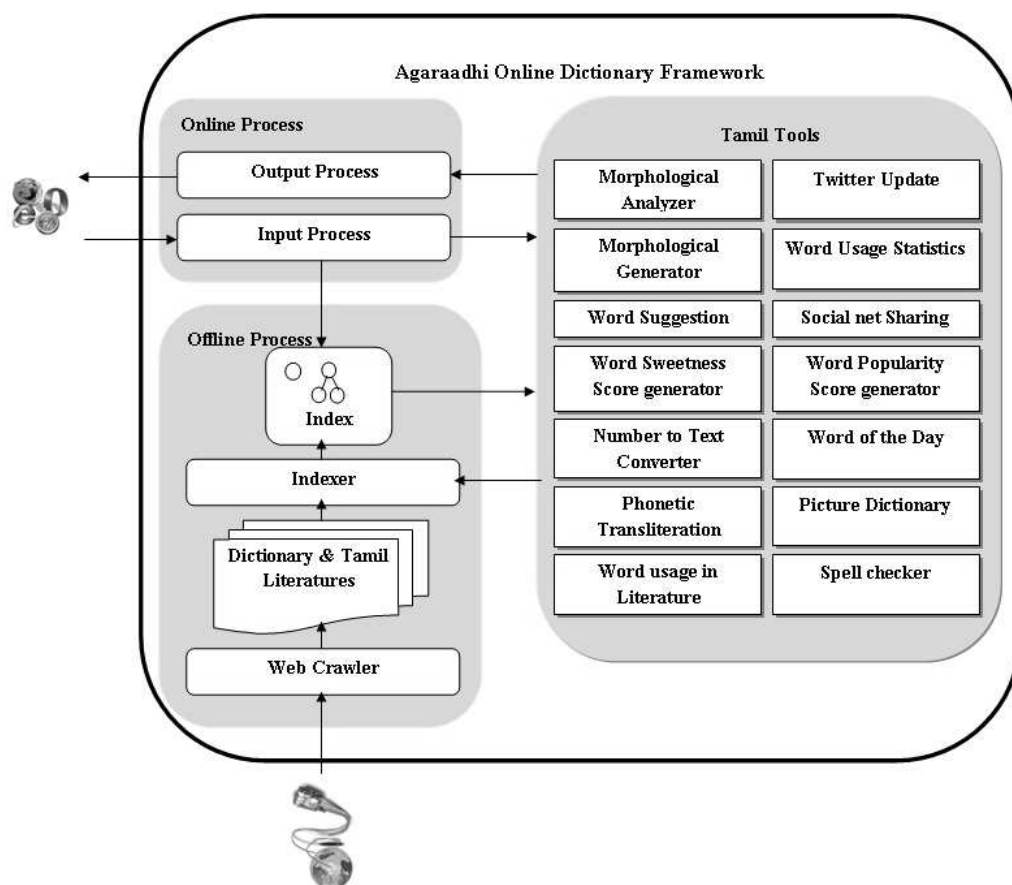


Fig 1: Agaraadhi Online Dictionary Framework

## 2.2 Offline Process

Tamil words and their meanings are entered manually and stored in text files. Those words are sequentially sent to modules such as popularity score generator, pleasantness score generator, picture dictionary, phonetic transliteration module and the resulting information is abstracted as a word object. Tamil literature such as Bharathiyaar songs, Avvaiyar songs, Thirukkural and lyrics are crawled using a static web crawler and are indexed in hash table as key value pairs.

## 2.3 Features of Agaraadhi Framework

Agaraadhi dictionary framework consists of more than twenty features such as Morphological Analysis Morph Generation, pleasantness scoring, popularity scoring, spell error suggestions etc.

### 3.1 Morphological Analyser

Morphological analyser [1] chunks the query word and gives the morphological features of the query word such as root word, parts of speech, gender, tense and count. If the Query word is *padithaan*, Morphological Analyser gives as *padi* as root, word represents male gender and query word is past tense and so on.

### 3.2 Morphological Generator

A Tamil morphological generator[2] needs to tackle different syntactic categories such as nouns, verbs, post positions, adjectives, adverbs etc. separately, since the addition of morphological constituents to each of these syntactic categories depends on different types of information. The generator is used to generate possible morphological variations of the query word.

### 3.3 Spell Checker

Spell Checker is used to check the spelling of Tamil words and to provide alternative suggestions for the wrong words. It uses the Morphological Analyzer. The Morphological Analyzer is used to split the given Tamil word into the root word and a set of suffixes. If the word is fully split by the analyzer and its root word is also found in the Agaraadhi dictionary, the given word is termed as correct. Otherwise, the correction process is invoked to generate all the possible suggestions with minimum variations from the given word.

### 3.4 Word Suggestions

Word Suggestion gives the list of equivalent or related words for the given query word.

### 3.5 Word Pleasantness and Word Popularity Score

Word Pleasantness score generator provides how easy to pronounce the word.

Word Popularity shows the word usage in the web. The Word from agaraadhi is given to web and found the frequency distribution of the word across the popular blogs, news articles, social nets etc.

### 3.6 Word Usage in Literature

This feature finds the usage of words in popular literature such as Thirukural, Bharathiyar Padalgal, Avvai songs and Lyrics.

### 3.7 Number to Text Converter

It converts a number to Tamil word equivalent as well as in English text. For example in Tamil we represent *oru Arpputham* (அற்புதம்) for 100 million, *Kumbam* (கும்பம்) for 10 billion and finally up to *Anniyan* (அந்நியம்) for one zillion.

### 3.8 Phonetic Transliteration

The pronunciation of words in Tamil and English language, as distinct from their written form based on the phonology and it can also vary greatly among dialects of a language. Phonetic transliteration module splits the word into syllables and gives the transliteration for each syllable.

### **3.9 Picture Dictionary**

Pictures, photos or line drawings to depict popular words have been included in the dictionary to enable efficient learning for children using this tool.

### **3.10 Social net Sharing and Twitter Update**

The framework also provides features to format results to be shared effectively on social networks. An Agaraadhi Bot was designed to post updates and word of the day on Twitter automatically.

### **3.11 Word of the Day and Word Usage statistics**

A rare word is randomly chosen and is displayed in the opening page to facilitate users to learn a new word every day.

Word Usage Statistics [3] shows the usage of the word in the social network over the past one week.

### **3.12 Tamil Word Games**

Games play a vital role in learning. Currently Agaraadhi has two Tamil word games namely Miruginajambo and Thookku Thookki. Miruginajambo is an unscramble game and Thookku Thookki is a Hangman game in Tamil.

## **4. Conclusion and Future Work**

This paper describes Agaraadhi, an Online Dictionary Framework. Agaraadhi online dictionary is a bilingual dictionary containing over 3 lac words on various domains like General, Medical, Engineering, Computer science, Literature etc. This Online Dictionary framework encompass various features such as morphological analysis, morphological generation, word usage statistics, word pleasantness analysis, spell checking, similar word finder, word usage in literature, picture dictionary, number to text conversion, phonetic transliteration, live usage analysis from micro blogs etc. Providing APIs for programmers and developing mobile apps for Agaraadhi framework will open a good platform for many researchers and developers working in Tamil Computing area.

## **References**

- Anandan, R. Parthasarathi, and Geetha, Morphological Analyser for Tamil. ICON 2002, 2002.
- Anandan, R. Parthasarathi, and Geetha, Morphological Generator for Tamil. Tamil Inayam, Malaysia, 2001.
- J. Jai Hari Raju, P. IndhuReka, Dr. Madhan Karky, Statistical Analysis and visualization of Tamil Usage in Live Text Streams, Tamil Internet Conference, Coimbatore, 2010.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# நவீன தமிழ் அகராதி

முனைவர் க. தமிழ்ச்செல்வன்,

இணைப் பேராசிரியர்,

சமூக அறிவியல் மொழிப்பள்ளி, வி.ஐ.டி. பல்கலைக்கழகம்,

வேலூர் - 632 014., வேலூர் மாவட்டம். தமிழ்நாடு, இந்தியா.

மின்னஞ்சல் : ktamilselvan@vit.ac.in

## முன்னுரை:

இன்றைய தொழில்நுட்ப யுகத்தில், புதிய புதிய கண்டுபிடிப்புகள் உருவான வண்ணம் உள்ளது. இதன் காரணமாகப் புதிய புதிய சொற்கள் உருவான வண்ணம் இருக்கிறது. உதாரணமாக, Cell Phone-ஐ செல்பேசி என்றும், கைப்பேசி என்றும் அழைக்கின்றோம். இதுபோல, தமிழில் உள்ள பழைய வார்த்தைகளுக்குப் பொருள் என்ன என்று அறிவதும் மிகுந்த தேவையாகும். உதாரணமாக (களிறு - யானை). மாணவர்களுக்கு இதுபோன்ற இலக்கியத் தரமானச் சொற்களுக்குப் பொருள் என்னவென்று தமிழ் அகராதி மூலம் விளக்குவது மிகுந்த தேவையாகும்.

இந்த நவீன தமிழ் அகராதி மென்பொருளில், ஒரு வார்த்தையைக் கணிப்பொறியில் தட்டச்சு செய்யத் தொடங்கும்போதே, முழு வார்த்தையை வழங்கும் வண்ணம் வடிவமைக்கப்படும். இதற்குப் புதிய தொழில்நுட்பம் பயன்படுத்தப்படும். இந்த நவீன தமிழ் அகராதியைத் தமிழ் - தமிழ் - ஆங்கிலம் என்றோ அல்லது ஆங்கிலம் - தமிழ் - தமிழ் என்றோ இருவேறு முறைகளில் எப்படி வேண்டுமானாலும் பயன்படுத்த முடியும். இதற்கான வசதியும் புதிய தொழில்நுட்பத்தைக் கொண்டு வடிவமைக்கப்படும்.

இந்த நவீன தமிழ் அகராதி மென்பொருள் கீழ்காணும் முறைகளில் வார்த்தைகளுக்குப் பொருளை வழங்கும்.

## நிலை - 1 :

முதலில் தேவையான மொழியைத் தேர்வு செய்ய வேண்டும். பின்னர் வார்த்தைகளைக் கணிப்பொறியில் தட்டச்சு செய்யும்போதே முழு வார்த்தையையும், அதற்கு அருகாமை வார்த்தைகளையும் வழங்கும். இந்த வசதியால், வார்த்தைகளை எளிதாகத் தேர்வு செய்யலாம்.

# நவீன தமிழ் அகராதி

◉ தமிழ்

◉ English

வார்த்தையைத் தட்டச்சு செய்யவும்

அ

Go

அ

அழகு

அஃது

உதாரணமாகத் தமிழ் மொழியைத் தேர்வு செய்து, "அ" என்ற எழுத்தைத் தட்டச்சு செய்தால், கீழ்க்கண்டவாறு அதன் அருகாமை வார்த்தைகளை வழங்கும். தேவையான வார்த்தையை தேர்வு செய்தவுடன் அதன் தமிழ்ப் பொருளையும், ஆங்கிலப் பொருளையும் பெறலாம்.

## நவீன தமிழ் அகராதி

தமிழ்

English

வார்த்தையைத் தட்டச்சு செய்யவும்

Go

பொருள் : தமிழ் நெடுங்கணக்கின் (அகரவரிசையின்) முதலெழுத்து

The first letter of Tamil alphabet

நிலை - 2 :

ஆங்கில மொழியைத் தேர்வு செய்து, "a" என்ற எழுத்தைத் தட்டச்சு செய்தால், கீழ்க்கண்டவாறு அதன் அருகாமை வார்த்தைகளை வழங்கும். தேவையான வார்த்தையைத் தேர்வு செய்தவுடன் அதன் ஆங்கிலப் பொருளையும், தமிழ்ப் பொருளையும் பெறலாம்.

## நவீன தமிழ் அகராதி

தமிழ்

English

Enter your word

a

Go

a

abacus

abbey

# நவீன தமிழ் அகராதி

தமிழ்

English

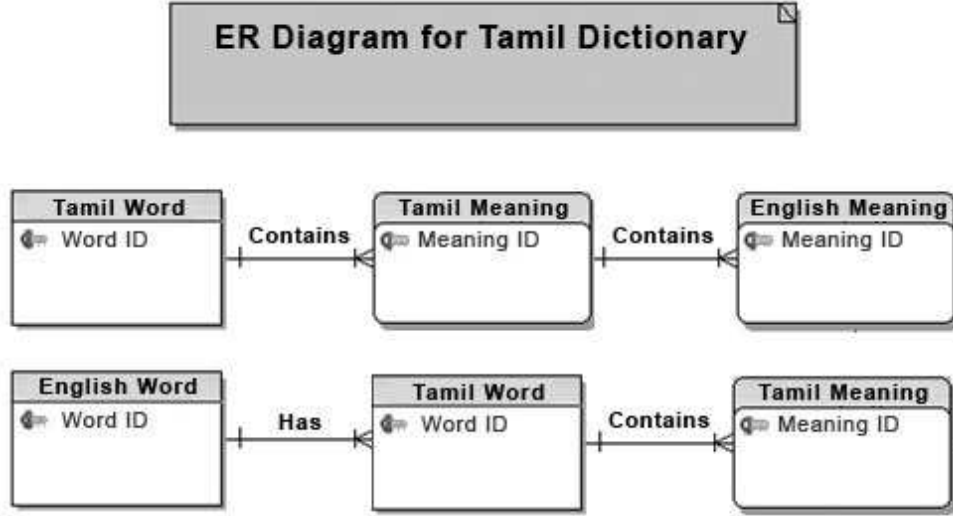
Enter your word

Go

பொருள் : The first letter of the English alphabet

ஆங்கில நெடுங்கணக்கின் (அகரவரிசையின்) முதலெழுத்து

தொழில்நுட்ப செயலாக்க வரைபடம் :



திட்ட உருவாக்கத்தின் படிநிலைகள் :

| வரிசை எண். | படிநிலைகள்   |
|------------|--|
| 1          | மென்பொருளின் மாதிரியை உருவாக்குதல் (Prototype)                             |
| 2          | இந்த மாதிரி வடிவத்திற்கு ஒப்புதல் பெறுதல் (Prototype Approval)             |
| 3          | மென்பொருளுக்கான கணிணி நிரல்களை உருவாக்குதல்                                |
| 4          | நவீன தமிழ் அகராதிக்கான டேட்டா பேசை உருவாக்குதல்                            |
| 5          | டேட்டாபேசை மென்பொருளுடன் இணைத்தல்  |
| 6          | மென்பொருளின் தரத்தை ஆய்வு செய்தல் (QA Testing)                             |
| 7          | நவீன தமிழ் அகராதியை இணையத்தில் நிறுவுதல்                                   |
| 8          | இணைய வசதி இல்லாத கணினியிலும் பயன்படுத்தும் Desktop Application வெளியிடுதல் |
| 9          | திட்டத்தை வெற்றிகரமாக முடித்தல்.   |



**உருவாக்கத்திற்கான மென்பொருள்கள் :**

| S.No. | Components         | Software   |
|-------|--------------------|--|
| 1     | Operating System   | Windows 2000 Professional or XP or above   |
| 2     | Application Server | ASP.NET 2.0, AJAX  |
| 3     | Web Tools          | HTML , Javascript  |
| 4     | Database           | MS SQL Server 2005   |
| 5     | Browser            | Internet Explorer, Mozilla Firefox, Netscape, Google chrome - with unicode support |

இந்த மென்பொருள் **Microsoft Visual Studio.NET 2005 with .NET Framework 2.0** என்ற நவீன தொழில்நுட்பத்தைக் கொண்டு உருவாக்கப்படும். இந்த மென்பொருள் **MS SQL 2005 Server** என்ற சிறப்பான, வேகமாகச் செயலாற்றக்கூடிய டேட்டாபேஸ் பயன்படுத்தித் தயாரிக்கப்படும்.

இந்த இரு மென்பொருள்களின் இணைப்பில் உருவாக்கப்படும் நவீன தமிழ் அகராதி அதிக வேகத்தில் தகவல்களைத் தரும்.

**தரம் :**

நவீன தமிழ் அகராதியானது 100% சதவீதம் கணிணி நிரல் குறைகள் அற்றதாக உருவாக்கப்படும். இதற்காக மிகச் சிறந்த வகையில் தரக்கட்டுப்பாட்டு ஆய்வு மேற்கொள்ளப்படும்.

மென்பொருள் உருவாக்கத்தில் ஒவ்வொரு படிநிலையிலும் தரக்கட்டுப்பாட்டு ஆய்வு மேற்கொண்டு **100% Bug Free** மென்பொருளாக உருவாக்கப்படும்.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# மொழிபெயர்ப்புக் கலையில் அகராதியின் பயன்பாடு

இளங்குமரன் த/பெ சிவநாதன்

சுல்தான் இட்ரிஸ் கல்வியியல் பல்கலைக்கழகம், மலேசியா

E-mail: s.ilangkumaran@gmail.com

மொழிபெயர்ப்புப் பணிகளில் அகராதிகளின் பயன்பாடு இன்றியமையாததாகின்றது. இருப்பினும் சிற்சில வேளைகளில் பெரிதும் அகராதிகளையே நம்பி செய்யப்படும் மொழிபெயர்ப்புப் பணிகள் அதன் இயற்கைத்தன்மையைக் காட்டத் தவறி, ஒருவித செயற்கை உணர்வை மொழிபெயர்ப்புப் பணிகளில் வெளிப்படுத்துகின்றன. சிறந்த மொழிபெயர்ப்பு பணி எனில், அவை மொழிபெயர்க்கப்பட்டவை என்ற உணர்வைத் தராது, குறிப்பிட்டதொரு மொழியிலேயே படைக்கப்பட்ட படைப்பு எனும் உணர்வையே படிப்போர்க்கு ஏற்படுத்த வேண்டும்.

இதன் அடிப்படையிலேயே அகராதியின் பயன்பாடு ஒரு மொழிபெயர்ப்புப் பணியில் எவ்வளவில் பயன்படுத்தப்படுதல் சிறப்பு; மற்றும், அதனைக் கையாளும் வழிமுறைகள் என்ன என்பது குறித்தே இவ்வாய்வு மேற்கொள்ளப்பட்டுள்ளது. இதன் மூலம் மொழிபெயர்ப்பாளர்களும், மொழிபெயர்ப்புத் துறை மாணவர்களும் கவனிக்கத் தவறவிட்ட சில விஷயங்களும், தகவல்களும் வெளிப்படுத்தப்பட முயற்சி செய்யப்பட்டுள்ளது.

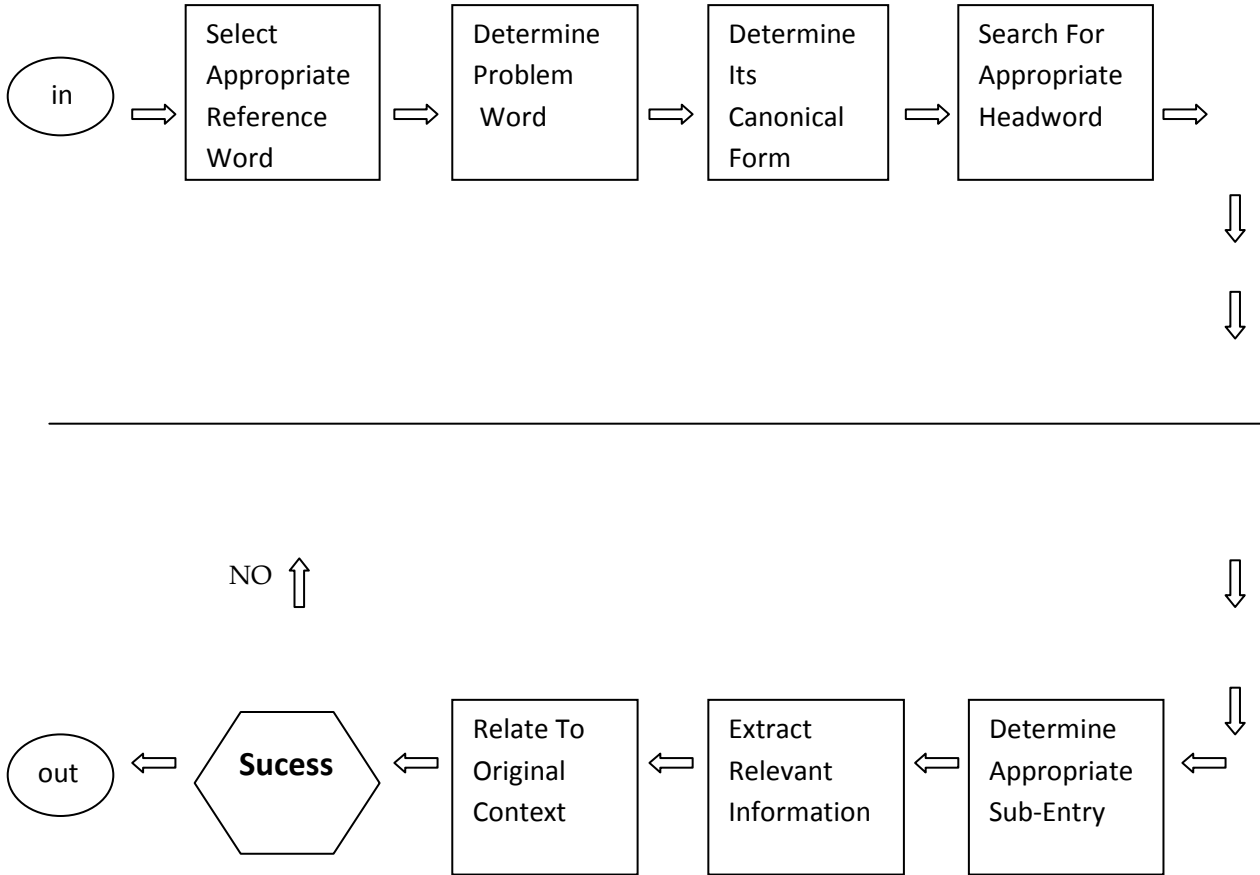
கால காலமாக நாம் பயன்படுத்தி வரும் புத்தக வடிவில் இருக்கும் மரபுவழி அகராதிகள் எந்த அளவிற்குத் இன்றைய நாட்களில் மொழிபெயர்ப்புப் பணிகளுக்கு உதவுகின்றன என்பது சிந்திக்க வேண்டிய விஷயமாகவே இருக்கின்றது. நாளுக்கு நாள் வளர்ந்து வரும் உலகில், பல புதிய கண்டுபிடிப்புகளுக்கும், ஆராய்ச்சிகளுக்கும் ஏற்ற அகர முதலிகள் மிகவும் இன்றியமையாததாக இருக்கின்றன. கையால், பழைய அகராதிகளையும், மரபு வழி அகராதிகளையும் தவிர புதிய அகராதிகளும், மின்னியல் அகராதிகளும் இன்றைய சூழலில் மிகவும் முக்கியத்துவம் வாய்ந்தவையாக இருக்கின்றன. இவை புத்தக வடிவில் மட்டுமின்றி இணையத்திலும், கையடக்க மின்னியல் அகராதிக் கருவியாகவும் உருமாற்றம் அடைந்து வருகின்றன; அது வரவேற்கத்தக்கதும் ஆகும்.

பல்வேறு ஆய்வுகளின் வாயிலாக, பொதுவாக அகராதிகளைப் பயன்படுத்துவோர் அதன் பயன்பாட்டை முழுமையாக அறியாமல் இருக்கின்றனர் என்று தெளிவுபடுத்தி இருக்கின்றன. உதாரணத்திற்கு, •பெளலி (Fawley) (1990) அவர்கள் கூறியதாவது, அகராதிகளைப் பயன்படுத்துவோர் மிக மிகக் குறைந்த அளவிலேயே அதன் பயன்பாட்டினை உணர்ந்துள்ளனர். அவர்கள் வெறுமனே சொற்களின் நேரடிப் பொருளை அறியவும் சரியான எழுத்துருக்களை அறிந்துக்கொள்ளவுமே அகராதிகளைப் பயன்படுத்தி வருகின்றனர். மாறாக சொல்லுருவாக்கம், உச்சரிக்கும் விதம், சரியான முறையில் பயன்படுத்திக் காட்டப்பட்டிருக்கும் வாக்கியங்கள், அச்சொல்லுக்கு ஏற்ற எதிர்ச்சொற்கள் போன்ற பல்வேறு குறிப்புகளில் மக்கள் அக்கறை கொள்வதே இல்லை என்பது அவரது குற்றச்சாட்டு. இதனாலேயே பெரும்பாலானோர் தாங்கள் தேடும் சொற்களுக்குச் சரிவர அல்லது போதிய தகவல்களைப் பெற தவறிவிடுகின்றனர். அதன் விளைவாக அவர்கள் தங்கள் படைப்புகளில் அவற்றைப் பிரயோகம் செய்யும்போது தவறானதொரு வார்த்தையைப் பயன்படுத்தி தொடர்ந்து வாசகர்களையும் குழப்பத்தில் ஆழ்த்தி விடுகின்றனர்.

இந்நிலை, ஒரு மொழியில் படைப்புகளை வெளியிடும் எழுத்தாளர்களுக்குச் சிரமத்தை விளைவிக்கின்றன என்றால், மொழிபெயர்ப்பாளர்களுக்கு அதைக்காட்டிலும் மிகப்பெரிய சுமையை ஏற்படுத்தி விடுகின்றன; ஏனெனில், படைக்கப்பட்டிருக்கும் மொழியில் பயன்படுத்தப்பட்டிருக்கும்

சொல்லுக்குச் சரியான பொருளை அறிந்துக்கொள்ளும் அதே வேளையில், தான் மொழிபெயர்க்க விரும்பும் மொழியில் அதற்குத் தகுந்த சொல்லைத் தெரிவு செய்ய வேண்டியவர்களாகவும், தொடர்ந்து அக்கட்டுரை படைக்கப்பட்டிருக்கும் சூழல், துறை ஆகியவற்றைக் கருத்தில் கொண்டு அந்தந்த துறைக்கும் சூழலுக்கும் ஏற்றாற்போல் தம் மொழிபெயர்ப்பைத் தரக்கடவர்களாகவும் மொழிபெயர்ப்பாளர்கள் இருக்கின்றனர்.

அகராதிகளின் பயன்பாடு குறித்து எழுந்துள்ள ஆய்வுகளில் மிக முக்கிய ஆய்வாகக் கருதப்படும் ஹார்த்மேன் (Hartmann) (1989) அவர்களின் ஆய்வு மொழிபெயர்ப்பாளர்கள் தாங்கள் மொழிபெயர்க்க விரும்பும் சொற்களுக்குச் சிறந்த முறையில் அகராதிகளில் பொருள்கொள்ள ஒரு கட்டமைப்பை உருவாக்கினார். அது :



Hartmann (1989) : Sociology of the dictionary user :Hypothesis and Empirical Studies, Worterbucher Dictionaries Dictionnaires [Art 12], Walter de Gruyter, Berlin, New York Vol. 1 : 102-111

### மொழிபெயர்ப்பாளர்களின் எண்ணங்களும் கருத்துக்களும்

1. எந்த மாதிரியான அகராதிகளையும் தேர்ந்தெடுத்து உபயோகிக்கலாம்.

- மொழிபெயர்ப்பாளர்களில் பெரும்பாலானோர் மிகவும் பிரசித்தி பெற்ற, மக்கள் மத்தியில் அதிகம் பேசப்படக்கூடிய அகராதிகளைப் பயன்படுத்துவதிலேயே ஆர்வம் காட்டுகின்றனர். மேலும் தங்களின் ஆசிரியர்கள் மற்றும் மொழிபெயர்ப்புத் துறை நண்பர்கள் அறிமுகப்படுத்தும் அல்லது ஊக்குவிக்கும் அகராதிகளைப் பயன்படுத்தத் தொடங்கும் மொழிபெயர்ப்பாளர்களில் பலர், கடைசி வரை தங்களைக்

காலத்துக்கேற்ப புதுபித்துக்கொள்ளாமலேயே கடைசி வரை மொழிபெயர்ப்புப் பணிகளில் தொடர்ந்து ஈடுபடுகின்றனர்.

2. கையடக்க அகராதிகளைப் பயன்படுத்துவது இலகுவானது.

- சில மொழிபெயர்ப்பாளர்கள் கையடக்க அகராதிகளைப் பயன்படுத்துவதில் பெரிதும் ஆர்வம் காட்டுகின்றனர். “மொழிபெயர்ப்பாளர்களாக விளங்கும் நாங்கள் எங்கு சென்றாலும் எங்களது அகராதிகளைக் கொண்டு செல்ல வேண்டியுள்ளது; ஏனெனில், அவ்வப்போது எங்களின் திறமைகளில் நம்பிக்கை வைத்து நேரிலும் தொலைபேசிகளிலும் அதிகமானோர் அணுகி தங்களது சந்தேகங்களுக்கு விளக்கம் கோருகின்றனர். அவர்களின் சந்தேகங்களை நிவர்த்திக்கும் பொருட்டு நாங்கள் எப்போதும் அகராதிகளுடனேயே இருக்கிறோம்” என சில தரப்பினர் கூறுகின்றனர். இன்னும் சிலர், குறிப்பாக மொழிபெயர்ப்புத் துறையில் நீண்ட காலம் பயிற்சி பெற்ற மொழிப்பெயர்ப்பாளர்களும் தங்களின் நற்பெயர் கலங்கப்படாதிருக்க மக்களின் சந்தேகங்களைக் களையும் நோக்கில் இவ்வாறு செயல்படுவதும் வருத்தமளிக்கின்றது. எல்லோருக்கும் எல்லா விஷயங்களும் தெரிந்திருக்க நியாயம் இல்லை என்பதை உணராதது, தெரியாதவற்றைத் தெரியவில்லை என பகிரங்கமாக ஒப்புக்கொள்ளும் தைரியம் இல்லாமல் போவது ஒரு புறமிருக்க, குறிப்பிட்ட வார்த்தைகளுக்குச் சரியான விளக்கங்கள் தான் அளிக்கிறோமா என்ற தெளிவும் அற்று ஒருவித குழப்பத்தையும் சமயங்களில் இது போன்றவர்கள் ஏற்படுத்துகின்றனர். இதுபோன்ற கையடக்க அகராதிகள் மாணவர்களுக்குப் பெருமளவில் பயன்படுகிறதேயொழிய மொழிபெயர்ப்பாளர்களுக்கு அந்த அளவிற்குப் பயன்படுவதில்லை. (இருப்பினும் கையடக்க மின்னியல் அகராதி இதிலிருந்து விதிவிலக்காகின்றது என்பதை அறிக)

2. அகராதிகளில் குறிப்பிடப்பட்டிருக்கும் சொற்களைத் தாராளமாகப் பயன்படுத்தலாம்

- பெரும்பாலான மொழிபெயர்ப்பாளர்கள் அகராதிகளில் குறிப்பிடப்பட்டிருக்கும் சொற்களையும், விளக்கங்களையும் தாராளமாகப் பயன்படுத்தலாம் என எண்ணம் கொண்டிருக்கின்றனர். இதனாலேயே சில சமயங்களில் நடைமுறைக்கு ஒவ்வாத தவறான மொழிபெயர்ப்புப் பணிகளை நாம் பார்க்க முடிகின்றது. மேலும் இதுபோல் அகராதிகளிலிருந்து எடுக்கப்பட்ட நேரடி வார்த்தைகள் சில வேளைகளில் சம்பந்தப்பட்ட கட்டுரை படைக்கப்பட்டிருக்கும் சூழலுக்கும், அவை படைக்கப்பட்டிருக்கும் துறைக்கும் சற்றும் பொருந்தாமல் போவது இங்கு குறிப்பிடத்தக்கது. உதாரணத்திற்கு இணையத்தில் பரவலாகப் பயன்படுத்தப்படும் Browse என்ற வார்த்தைக்கு அகராதியின் வாயிலாக நேரடிப் பொருள் கொள்ளும்போது, இளந்தளிர் உணவு, கிளை தழை, பசுந்தீவனம், தழை மேய்தல் மற்றும் புற்கறித்தல் என்ற பொருள்களைத் தருகின்றது. ஆனால், உண்மையில் இச்சொல் உணர்த்தவரும் பொருள் வலம் வருதல், அணுகுதல் போன்றவையாகும். இந்நிலையில் இச்சொல் பயன்படுத்தப்பட்டிருக்கும் சூழலையும் அதன் துறையையும் அறியாது மொழிபெயர்க்கப்பட்டிருக்கும் படைப்புகள் உகந்த பொருளைத் தர தவறுவதோடு அதைப் படிப்பவர்களுக்குப் பெருங்குழப்பத்தை ஏற்படுத்திவிடுகின்றது.

3. நமக்குத் தெரிந்த விஷயம்தானே என்ற போக்கு

- சில வேளைகளில் மொழிபெயர்ப்புப் பணிகளில் ஈடுபடும் சிலர் இது நமக்குத் தெரிந்த விஷயம்தானே, இதற்காகவெல்லாம் அகராதியைப் புரட்டவேண்டியதில்லை என்ற எண்ணமும் கொண்டு செயல்படுகின்றனர். பொதுவாக மொழிபெயர்க்கப்படப்போகும் மொழிகளில் பாண்டித்தியம் பெற்றவர்களே மொழிபெயர்ப்புகளைச் செய்வதால் இத்தகைய சிந்தனையால் பெரிதாகப் பிரச்சனை ஏதும் எழாது என்று எண்ணத் தோன்றுகிறது. இருப்பினும், சில வேளைகளில் நுண்ணிய விஷயங்களை மொழி பெயர்க்கும்போது பல கோணங்களில் அவற்றைப் பகுத்துப் பார்ப்பது இன்றியமையாததாகின்றது. உதாரணத்திற்கு 1996-ம் ஆண்டு மலேசிய விமானச் சேவையின் மொழிபெயர்ப்புப் பணியை ஏற்று

முடித்த மொழிபெயர்ப்பாளர் ஒருவர் பின்னர் அந்நிறுவனம் மேற்கொண்ட சட்ட நடவடிக்கையால் (மான நஷ்ட வழக்கு) திவாலாகும் நிலையை அடைந்தது குறிப்பிடத்தக்கது. விமானப் பயணத்தின்போது நெருக்கடி நிலை ஏற்படுமாயின் பின்பற்ற வேண்டிய இலகுவான வழிவகைகள் குறித்து சீன மொழியில் மொழிபெயர்க்க வேண்டியிருந்த பணியில், சற்றே கவனக்குறைவாக இலகுவாக நெருக்கடி நிலை ஏற்படக்கூடிய இவ்விமானப் பயணத்தில் பின்பற்ற வேண்டிய வழிவகைகள் என்று தவறுதலாக மொழிபெயர்த்து பின்னர் பெரும் சிக்கலுக்கு உல்லான அம்மொழிபெயர்ப்பாளர் அதற்கு முன்னர் ஏராளமான மொழிபெயர்ப்புப் பணிகளில் ஈடுபட்டு அவற்றைச் செவ்வனே முடித்தவர் என்பது வழக்கு விசாரணையில் தெரிந்தது. இதைவிடக் குறிப்பாக ஆரம்பக்காலங்களில் சின்ன சின்ன விஷயங்களுக்கும் அகராதியின் துணைகொண்டு பொருளை அறிந்த பின்னரே மொழிபெயர்க்கும் அவர் காலப்போக்கில் அகராதியின் பயன்பாடு குறைந்து போக, தனக்கு தெரிந்தது தானே என்று தன் அனுபவத்தை முற்றிலுமாக நம்பி செயல்பட்டதே இந்த தவறுக்குக் காரணம் என விசாரணையில் ஒப்புக்கொண்டதும் குறிப்பிடத்தக்கது.

4. எந்த வகையான மொழிபெயர்ப்புகளையும் செய்யலாம்.

- சில மொழிபெயர்ப்பாளர்கள் தங்களுக்குக் கிடைக்கும் எவ்வகையான பணிகளையும் செய்து விடலாம் என்ற எண்ணம் கொண்டுள்ளனர். இது சரியல்ல. சில குறிப்பிட்ட துறைகளில் மிகுந்த திறமை கொண்டுள்ள ஒருவர் மற்ற துறைகளிலும் விற்பன்னராக இருப்பார் என்று எண்ணுவது தவறு. மொழிபெயர்ப்புகளில் பல பிரிவுகள் உண்டு. அவை சட்டத்துறை மொழிபெயர்ப்புகள், மருத்துவ மொழிபெயர்ப்புகள், கணினி மொழிபெயர்ப்புகள், பொருளாதாரத்துறை மொழிபெயர்ப்புகள், விளம்பர மொழிபெயர்ப்புகள் போன்ற பல பிரிவுகளாலான துறைகளில் மொழிபெயர்ப்புப் பணிகளை மேற்கொள்ள பல வகையான திறமைகள் தேவைப்படுகின்றன. கையால் 'இதனை இதனால் இவன் முடிக்கும்' என ஆராய்ந்து அவற்றைச் சம்பந்தப்பட்டவர்களிடம் ஒப்படைப்பதே உசிதம். மொழி பெயர்ப்பாளர்களும் பணத்தை மட்டுமே குறியாகக் கொள்ளாது மொழிபெயர்ப்பின் தரத்தைக் காக்க ஆவன செய்ய கடமைப்பட்டவர்களாவர்.

5. பலதரப்பட்ட அகராதிகளைப் பயன்படுத்துதல்

- பலதரப்பட்ட அகராதிகளைப் பயன்படுத்துதல் ஒரு மொழிபெயர்ப்பாளரைப் பொருத்த வரையில் மிக மிக வரவேற்கக்கூடிய ஒன்றாக இருப்பினும், கவனக்குறைவு ஏற்படவும் இதில் பெரிய வாய்ப்பு உள்ளதை பெரும்பாலான மொழிப்பெயர்ப்பாளர்கள் உணர தவறுகின்றனர். எப்படி? பல துறைகளில் மொழிப்பெயர்ப்புப் பணிகளை மேற்கொள்ளும் மொழிபெயர்ப்பாளர்கள் ஒன்றுக்கும் மேற்பட்ட அகராதிகளைத் துணைக்கு வைத்திருப்பது இயற்கையே. நீண்ட காலத்திற்குச் செய்யப்படும் மொழிபெயர்ப்புப் பணிகளில் சில சமயங்களில் அப்பணியின் ஆரம்பத்தில் பிரயோகிக்கப்பட்ட குறிப்பிட்டதொரு வார்த்தை அப்பணி முடிவடையும்போது வேறு வார்த்தைப் பிரயோகத்தில் முடிவதைப் பார்க்க முடிகின்றது. உதாரணத்திற்கு, 100 பக்கங்களைக் கொண்ட ஒரு மொழிபெயர்ப்புப் பணியை ஒருவர் ஒரே நாளில் செய்து முடித்துவிடுவதென்பது அரிய காரியம். இவ்வாறு நான்கு அல்லது ஏழு நாட்களுக்குத் தொடரும் பணியில், ஆரம்பத்தில் பயன்படுத்திய அகராதியை விடுத்து பிரிதொரு அகராதியின் துணைகொடலின்போது பிரிதொரு வார்த்தையை அம்மொழிப்பெயர்ப்பாளர் பிரயோகிக்க வாய்ப்புண்டு. இது இணைய அகராதிகளைப் பயன்படுத்துவோரிடமும் அதிகம் நேருகின்றது; ஏனெனில், இவர்களைப் போன்றவர்கள் நான்கு அல்லது ஐந்து இணைய தளத்தினை உதவிக்குத் துணை கொள்பவர்களாக இருக்கின்றனர். இவ்வாறான தவறுகள் பொருளாதாரத்துறை மொழிபெயர்ப்புகளிலும் ண்டறிக்கைகளிலும் பெருமளவு காணப்படுகின்றது.

க, மொழிபெயர்ப்பாளர்கள் அகராதிகளைப் பயன்படுத்துவதில் மிகவும் விழிப்பாக இருக்க வேண்டும். பொருந்தாத அகராதிகளின் பயன்பாடும், அதிகமான மற்றும் குறைவான பயன்பாடுகளும் கூட தவறான மொழிபெயர்ப்புகளுக்கு வித்திட்டுவிடும். கையால் இவ்விஷயத்தில் மிகுந்த கவனம் தேவை. மேலும், முறையான அகராதிகளின் பயன்பாடுகள் குறித்து பொதுவாக பள்ளிகளிலிருந்தும், தவிர மொழிபெயர்ப்புக் கல்விகளைப் போதிக்கும் கல்விக்கூடங்களும் முறையாக போதிக்க வேண்டும். மாணவர்களுக்கு அகராதிகளின் முழுமையான பயன்பாட்டினைப் போதிக்கும் பட்சத்தில் வருங்காலங்களில் சிறப்பான மொழிபெயர்ப்புகள் மட்டும் அன்றி தரமான படைப்புகளை உருவாக்குவதற்கும் அது வழிவகுக்கும் என்பதில் ஐயமில்லை.







## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# **Blog**

(வலைப் பூக்கள்)



# Impact of SOA and Web 2.0 in Tamil Blogs and Social Networks

*Ferdin Joe J*

*PG Scholar/CSE, Einstein College of Engineering, Tirunelveli, India*

## **Abstract**

Social Networks and Blogs are the most powerful electronic media which reaches every nook and corner of the planet. They are being viewed and managed by even devices using the embedded technology. Service Oriented Architecture (SOA) and Web 2.0 are the technologies which are the base for social networks and blogs. Tamil blogs are also gaining popularity among readers with ample contribution from Unicode format of data management. Many Tamils all over the globe; especially the Non Residential Indians (NRIs) have a powerful base of reflecting their thoughts on the social issues of India using the Tamil Blogs. Social Networks like Twitter, Facebook etc., are also supporting the Tamil Unicode nowadays and the Tamil Internet User Community is benefited in many ways. The use of SOA and Web 2.0 plays a vital role in Tamil Unicode with Social Networks and Tamil Blogs. In this paper, various Social Networks and Tamil blogs from all over the globe are evaluated. This evaluation study is done to find the impact of SOA and Web 2.0 on the web. The views, target audience and significant features of various Tamil bloggers and Social Network users are also taken into account for this evaluation.

*Index Terms: Social Networks, Tamil Blogs*

## **Introduction**

Tamil is powered by Unicode in the web, especially in the social networking is in the increase. Policies were framed after the Tamil Internet Conference 2010 by the Government of Tamil Nadu, India. The usage of Tamil Unicode was approved officially for any communication with the Government offices. With the collaboration of Microsoft, Tamil Unicode was given a sort of recognition at the level of silverlight frameworks. Due to these policies, the use of Tamil Unicode has increased proportionally with the increase in the number of internet users. Lot many Tamil Language Researchers who were new to internet usage, utilized this technology as a boon for their research. Encouragement for the usage was enormous by the release of free and open source softwares like NHM writer released in the CD of CDAC, Govt of India, Azhagi word processor and many more. Nowadays Tamil Unicode is used to comment in various social networking sites like orkut, facebook, twitter etc and chat applications of gmail, yahoo etc. Tamil journalists and magazine column writers found this technology easy and quicker way to reach the editorial board. These areas are powered by the service oriented architecture (SOA) and web 2.0 using blogs and social networking sites. These sites reach the people with the Really Simple Syndication (RSS) and Atom feeds in a more effective way. Tamil Unicode is used by the users to express the culture, activities and emotions of Tamils around the globe.

In this paper, I have analyzed many social networks and Tamil blogs. The use of SOA and Web 2.0 is taken as the usage of these sites. From these websites, the impact of Tamil Users are analyzed for the past one year and the major issues of the Tamil Unicode users are listed out as conclusion of the study.

### **Study sites & Criteria**

Various sites like savukku.net, athirvu and many bloggers in blogspot and wordpress are taken into account. They were monitored throughout the year and the issues were listed. This listing was done until the release of wikileaks cables and the UNO report on Srilankan war. Most of the bloggers focused on the Srilankan Tamils issue by slamming the Government of India. Then the 2G spectrum scam which is currently blaming the ruling parties of India and the State Government of Tamil Nadu. Apart from these, the issues were raised on the byelections and the small incidents happened in and across Tamil Nadu.

### **Aim of writers**

The main aim of these writers was to change the minds of the people and they have succeeded partially by getting positive comments from exclusive and anonymous users. They use to write this stuff in Tamil Unicode and therefore the reader base has increased but the contents were not reliable. The quality of writing included many third rated comments on National leaders.

### **Impact on readers**

The impact of these articles made both positive and negative impact on the reader base. Though they got an increased reader base, they failed to give quality articles. The issues raised on the Srilankan war became a success with the advent of the UNO report on war crime. But in the case of 2G spectrum scam, it is not. In the case of 2G spectrum scam, no one has the rights to comment on either the Government or the politicians. The scam is still under the trial in the special court of Central Bureau of Investigation, India. Based on some filthy youtube released tapes and opposition party demonstrations, the real conclusion could not be made. But the Tamil content writers made their attempt in this issue with ghost referred knowledge. The Parliamentary Accounts Committee (PAC) is headed by Mr. Murali Manohar Joshi from the opposition party in the Parliament of India. The initial audit report gave a doubt of loss of money due to the spectrum auction. But the writers were just slamming on the Government for the illusion created. The Joint Parliament Committee (JPC) is formed as per the demand of the opposition party and the trial is going on. Recently, "The Hindu" has given a news stating that, the effect of JPC is becoming a boon to the ruling party itself and the opposition parties want the committee to retire as they couldn't achieve what they felt too. On April 28<sup>th</sup>, 2011, the Parliament Accounts Committee's report was rejected whole heartedly rejected by the committee and a new Head was appointed. These activities show that the scam is not the offence of only ruling party but there are some others who are behind. In this scenario, the content of Tamil writers in the web came to be known as just illusions and confused the reader base. Similarly, the Government of TamilNadu was criticized for freebies in the 2011 Assembly Elections Poll campaign. Elections are normally the reflection of the pulse of voters. Without knowing the this pulse, the content was published just by slamming the welfare schemes done. We need to wait till May 13<sup>th</sup> to see whether the voters have rejected the freebies or they welcomed with red carpet. Without this made clear, the media will never have rights to comment on the Government to change the mindset of people.

## **Conclusion**

In this scenario, as a neutral reader, the issues stated in the social networks and blogs need to be analyzed thoroughly by the readers and then the issues should be taken into account. It is really an offense to feed the people with wrong news. Tamil internet users need to take all these concepts in mind before they read any sort of sensitive news in the media. The links posted in the social networks need to be audited by any neutral organization upon appeal. The negative certificate on the content will make the users to understand the difference between reality and illusions. Currently the Tamil media in the form of blogs, social networks, TV channels are not reliable because of the backing of single person opinion being stuffed on reader base. The auditing on the reality of content and the certification of content will definitely lead the readers to get the correct picture on the issue.

## **References**

Nothing in specific. Most of the eminent social networks and blogs were taken for this study.

# Tamil Classical Literature in the age of blogging and social network

*Palaniappan Vairam Sarathy*

<http://karkanirka.org>

Tamil classical literature popularly known as the Sangam literature is hailed by scholars around the world as one best literary output of human civilizations. Scholars from various western countries have patronized Sangam literature and produced various studies and translation on this literature. Many Universities around the world have understood the literary significance of Sangam literature and have opened departments to study Sangam literature. With such literary greatness, Sangam literature is not much known to common people of Tamil Nadu. The common people of Tamil Nadu are more exposed to literature such as Thirukkural and Bakthi than Sangam literature.

Except for very few Puram poems much of Sangam literature is unknown to common people of Tamil Nadu. The only exposure most people have towards Sangam literature is the few poems which they had read in their school Tamil syllabus. There has been no popular appeal or drive to encourage people to read Sangam literature in the likes of Thirukkural.

Themes of Love and war essentially make most of the poems, and not well appreciated by children of young ages. Hence the only time people are exposed to this literature they had not developed any special interest for it. In such circumstances, the age of blogging and social network which are the mostly used by the youngsters in the prime of their life can be effectively used to popularize the Sangam literature. Sangam literature has strong emotional connection with present day lives. Hence if Sangam poems are presented in proper way, it is sure strike a chord with present generation.

## **Obstacles to a common man:**

The biggest obstacle for a common man is the archaic language of Sangam Literature. To understand and appreciate the Sangam poems, a person has to know lot of background details like themes, landscape etc. Dr.Kamil Zvelebil has once remarked that only trained readers can completely understand and appreciate Sangam literature. A common man requires considerable education before he can appreciate Sangam poems. More often than not, translating the poem or explaining the poem line by line doesn't give full experience of the poem. As an example I would like to take a poem and explain how much back ground detail he requires to appreciate the poem

மாரி ஆம்பல் அன்ன கொக்கின்  
பார்வல் அஞ்சிய பருவரல் ஈர் ளெண்டு  
கண்டல் வேர் அளைச் செலீஇயர், அண்டர்  
கயிறு அரி எருத்தின், கதழும் துறைவன்  
வாராது அமையினும் அமைக  
சிறியவும் உள ஈண்டு, விலைஞர் கைவளையே.

If the chief of maritime land

*Where,  
the wet crab  
seeing the stork,  
Which looked  
Like a water lily in rain ,  
Gets afraid and  
moves inside the hollow roots of  
the mangrove trees ,  
Like a bull which cut loose  
from the rope tied by the herdsman,*

Doesn't return

let it be so.

The merchant sells smaller bangles  
in this town.

In the following poem we expect the reader of this translation to know the following points to enjoy the poem completely

1. How the water lily looks when it rains - and how it is so similar to the crane
2. Wet crab being compared to Thalaivan
3. Crane compared to gossip of women
4. Running for safety into hollow roots, as imagery for Thalaivan hurrying to Thalaivi's home to save his love.
5. Bull cutting loses from the rope tied by the herdsman as imagery for Thalaivan breaking the shackles of the pressure from the society.
6. Small bangles to indicate the Tamil akam tradition of love sickness and thinning of hands and loosening of bangles. Small bangles will fit the smaller hand and hence they can throw away the memory of Thalaivan and be safe from the society.

It can be shown with further examples that traditional methods of translation and brief explanation cannot give a reader the full experience of a sangam poem. Hence experimental approaches should be considered to take this classical literature to the common reader.

### **Blogging as a medium:**

Blogging as medium of communication has developed over the last few years. This medium gives enough flexibility and freedom and in most cases free of cost. Blogging has become a popular medium for readers who like to gather knowledge on various topics. Most internet users are used to blog medium hence blogging overcomes one of the major pitfalls of Book. Books involve printing cost



hence there are severe restrictions on page counts and illustrations. Blogs therefore can afford to be as long as they want and can use illustrations as required. The blogging medium is also flexible enough to allow audio and video. The flexibility of this medium can be really helpful in presenting Sangam literature to the new age reader.

With no restrictions in page count, more explanation can be given on the theme, landscape and other required back stories can be given in addition to translation to understand a particular poem.

Sangam literature has very visual nature to it. The poets of the Sangam infused elements of nature-flora and fauna effectively in the poems to describe emotions. They play an integral part in understanding the poem. Most readers' knowledge of flora and fauna especially with names in archaic Tamil is pretty low. Hence pictures or illustration of flora and fauna mentioned in the poems helps the reader understand the poems better.

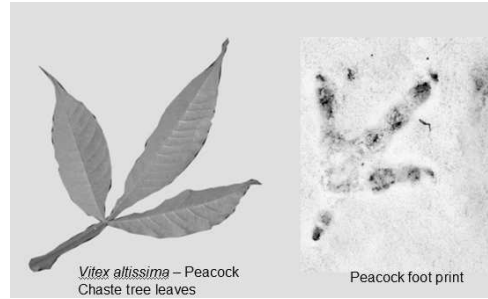
### Experimental approaches:

The visual nature of the Sangam nature can be utilized and the poems can be presented in new formats such as explaining poems with series of pictures or more experimental approach of comic book format (with illustrations and subtext). Audio files with rendering of the poem along with explanations can be uploaded. Lectures on the poem can be recorded and uploaded as video blogs. Small animations can be made to explain similes and imagery of the poem.

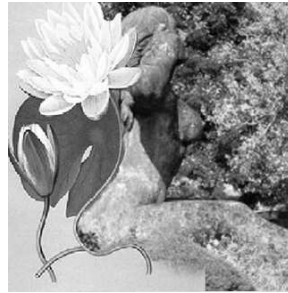
Few examples of illustrating visual elements of Sangam poems are given below



மாரி ஆம்பல் அன்ன கொக்கின்



மயில் அடி இலைய மாக் குரல் நொச்சி



தண் நறுங் கழுநீர்ச் செண் இயற் சிறுபுறம்

A more experimental approach of Sangam poem in comic book style



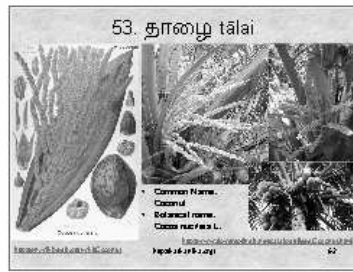
With blogging medium we have no restrictions to creativity. Experimental and creative approaches can help in popularizing the Sangam poems.

### Visual catalogs:

Visual catalogs of various flora fauna and smiles can be made into searchable online databases. In one such effort visual catalog of 99 flowers of Kurinchippattu was created 2 years ago in my blog karkanirka.org. This effort can be extended to include all flora and fauna of Sangam and also efforts can be put convert into a searchable online database. Such visual catalogs would help the common reader as well as researchers in understanding the Sangam literature in much better way.



61



62



63



66



67



68

Few slides from the visual presentation of 99 Flowers of Kurichippaatu.

### Relating to popular Culture:

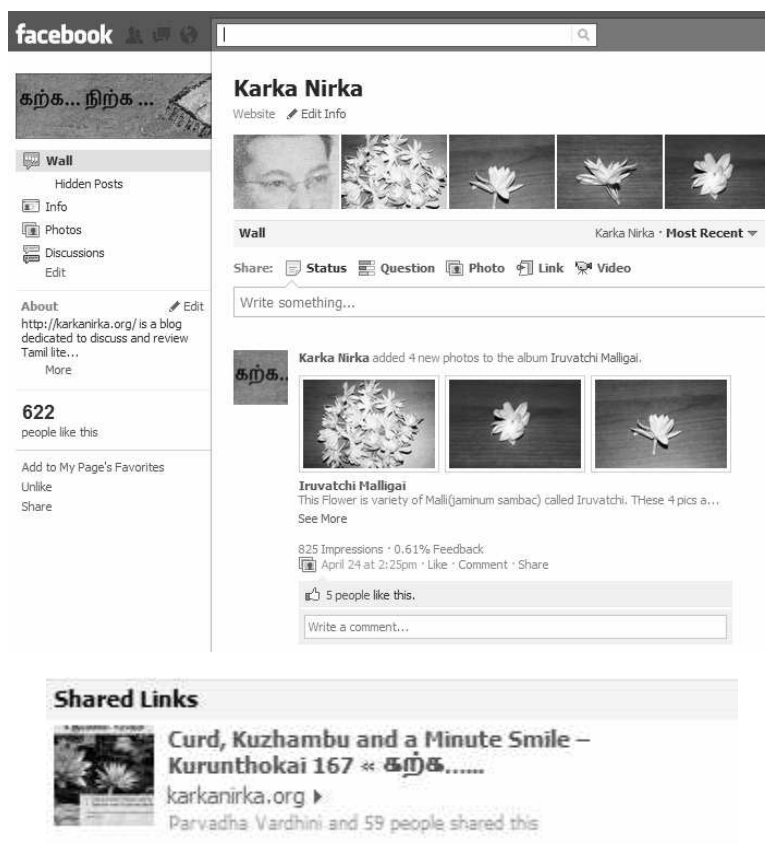
Most themes of Sangam poems are still prevalent in popular culture though Tamil Cinema, either as scenes or as lyrics for film songs. Relating the Sangam poems to prevalent popular culture helps common reader understand the classical poems better since they are able to relate to the emotion of the poem.

Themes such as love at first sight, opposition for love from parents, eloping of lovers, temporary separation of lovers, wife cooking for first time and waiting for approval of husband, girls parents mistaking girls love sickness and inviting priest to perform rituals are still prevalent as movie themes.

Relating the Sangam themes to popular culture makes them understand that Sangam literature isn't as alien as they had imagined and creates an interest in them to explore more in the classical literature.

### Use of Social Network:

Social networks like Orkut, Twitter and Facebook have become the most popular medium for knowledge transfer and sharing. Each of these social networks has hundreds of groups on Tamil Language and literature. Sharing the links of the blogs at such groups helps spread the reach of Sangam literature. This method popularly known as viral marketing is very useful tool. Many new readers are exposed to Sangam literature this way. When a reader likes a poem he shares it to his friends and any one in his friends list who likes the post passes on to his friends. This way with minimum effort the reach of the poems can be maximized. The social networks have played an important role in growth of Karka Nirka blog. In April 2010 Karka Nirka Facebook page was started and presently there are 622 followers and growing. On average one blog posted in Face book is passed on/shared by 50 other people.



## Conclusion:

New approaches are required to popularize the Sangam literature. There are many takers for such efforts. Fresh enthusiasm and creativity can help to revitalize the Sangam Literature and take it to the present generation.





## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Impact of SOA and Web 2.0 in Tamil Blogs and Social Networks

*Ferdin Joe J*

*PG Scholar/CSE, Einstein College of Engineering, Tirunelveli, India*

## **Abstract**

Social Networks and Blogs are the most powerful electronic media which reaches every nook and corner of the planet. They are being viewed and managed by even devices using the embedded technology. Service Oriented Architecture (SOA) and Web 2.0 are the technologies which are the base for social networks and blogs. Tamil blogs are also gaining popularity among readers with ample contribution from Unicode format of data management. Many Tamils all over the globe; especially the Non Residential Indians (NRIs) have a powerful base of reflecting their thoughts on the social issues of India using the Tamil Blogs. Social Networks like Twitter, Facebook etc., are also supporting the Tamil Unicode nowadays and the Tamil Internet User Community is benefited in many ways. The use of SOA and Web 2.0 plays a vital role in Tamil Unicode with Social Networks and Tamil Blogs. In this paper, various Social Networks and Tamil blogs from all over the globe are evaluated. This evaluation study is done to find the impact of SOA and Web 2.0 on the web. The views, target audience and significant features of various Tamil bloggers and Social Network users are also taken into account for this evaluation.

*Index Terms: Social Networks, Tamil Blogs*

## **Introduction**

Tamil is powered by Unicode in the web, especially in the social networking is in the increase. Policies were framed after the Tamil Internet Conference 2010 by the Government of Tamil Nadu, India. The usage of Tamil Unicode was approved officially for any communication with the Government offices. With the collaboration of Microsoft, Tamil Unicode was given a sort of recognition at the level of silverlight frameworks. Due to these policies, the use of Tamil Unicode has increased proportionally with the increase in the number of internet users. Lot many Tamil Language Researchers who were new to internet usage, utilized this technology as a boon for their research. Encouragement for the usage was enormous by the release of free and open source softwares like NHM writer released in the CD of CDAC, Govt of India, Azhagi word processor and many more. Nowadays Tamil Unicode is used to comment in various social networking sites like orkut, facebook, twitter etc and chat applications of gmail, yahoo etc. Tamil journalists and magazine column writers found this technology easy and quicker way to reach the editorial board. These areas are powered by the service oriented architecture (SOA) and web 2.0 using blogs and social networking sites. These sites reach the people with the Really Simple Syndication (RSS) and Atom feeds in a more effective way. Tamil Unicode is used by the users to express the culture, activities and emotions of Tamils around the globe.

In this paper, I have analyzed many social networks and Tamil blogs. The use of SOA and Web 2.0 is taken as the usage of these sites. From these websites, the impact of Tamil Users are analyzed for the past one year and the major issues of the Tamil Unicode users are listed out as conclusion of the study.

### **Study sites & Criteria**

Various sites like savukku.net, athirvu and many bloggers in blogspot and wordpress are taken into account. They were monitored throughout the year and the issues were listed. This listing was done until the release of wikileaks cables and the UNO report on Srilankan war. Most of the bloggers focused on the Srilankan Tamils issue by slamming the Government of India. Then the 2G spectrum scam which is currently blaming the ruling parties of India and the State Government of Tamil Nadu. Apart from these, the issues were raised on the byelections and the small incidents happened in and across Tamil Nadu.

### **Aim of writers**

The main aim of these writers was to change the minds of the people and they have succeeded partially by getting positive comments from exclusive and anonymous users. They use to write this stuff in Tamil Unicode and therefore the reader base has increased but the contents were not reliable. The quality of writing included many third rated comments on National leaders.

### **Impact on readers**

The impact of these articles made both positive and negative impact on the reader base. Though they got an increased reader base, they failed to give quality articles. The issues raised on the Srilankan war became a success with the advent of the UNO report on war crime. But in the case of 2G spectrum scam, it is not. In the case of 2G spectrum scam, no one has the rights to comment on either the Government or the politicians. The scam is still under the trial in the special court of Central Bureau of Investigation, India. Based on some filthy youtube released tapes and opposition party demonstrations, the real conclusion could not be made. But the Tamil content writers made their attempt in this issue with ghost referred knowledge. The Parliamentary Accounts Committee (PAC) is headed by Mr. Murali Manohar Joshi from the opposition party in the Parliament of India. The initial audit report gave a doubt of loss of money due to the spectrum auction. But the writers were just slamming on the Government for the illusion created. The Joint Parliament Committee (JPC) is formed as per the demand of the opposition party and the trial is going on. Recently, "The Hindu" has given a news stating that, the effect of JPC is becoming a boon to the ruling party itself and the opposition parties want the committee to retire as they couldn't achieve what they felt too. On April 28<sup>th</sup>, 2011, the Parliament Accounts Committee's report was rejected whole heartedly rejected by the committee and a new Head was appointed. These activities show that the scam is not the offence of only ruling party but there are some others who are behind. In this scenario, the content of Tamil writers in the web came to be known as just illusions and confused the reader base. Similarly, the Government of TamilNadu was criticized for freebies in the 2011 Assembly Elections Poll campaign. Elections are normally the reflection of the pulse of voters. Without knowing the this pulse, the content was published just by slamming the welfare schemes done. We need to wait till May 13<sup>th</sup> to see whether the voters have rejected the freebies or they welcomed with red carpet. Without this made clear, the media will never have rights to comment on the Government to change the mindset of people.



## **Conclusion**

In this scenario, as a neutral reader, the issues stated in the social networks and blogs need to be analyzed thoroughly by the readers and then the issues should be taken into account. It is really an offense to feed the people with wrong news. Tamil internet users need to take all these concepts in mind before they read any sort of sensitive news in the media. The links posted in the social networks need to be audited by any neutral organization upon appeal. The negative certificate on the content will make the users to understand the difference between reality and illusions. Currently the Tamil media in the form of blogs, social networks, TV channels are not reliable because of the backing of single person opinion being stuffed on reader base. The auditing on the reality of content and the certification of content will definitely lead the readers to get the correct picture on the issue.

## **References**

Nothing in specific. Most of the eminent social networks and blogs were taken for this study.



## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Tamil Classical Literature in the age of blogging and social network

*Palaniappan Vairam Sarathy*

<http://karkanirka.org>

Tamil classical literature popularly known as the Sangam literature is hailed by scholars around the world as one best literary output of human civilizations. Scholars from various western countries have patronized Sangam literature and produced various studies and translation on this literature. Many Universities around the world have understood the literary significance of Sangam literature and have opened departments to study Sangam literature. With such literary greatness, Sangam literature is not much known to common people of Tamil Nadu. The common people of Tamil Nadu are more exposed to literature such as Thirukkural and Bakthi than Sangam literature.

Except for very few Puram poems much of Sangam literature is unknown to common people of Tamil Nadu. The only exposure most people have towards Sangam literature is the few poems which they had read in their school Tamil syllabus. There has been no popular appeal or drive to encourage people to read Sangam literature in the likes of Thirukkural.

Themes of Love and war essentially make most of the poems, and not well appreciated by children of young ages. Hence the only time people are exposed to this literature they had not developed any special interest for it. In such circumstances, the age of blogging and social network which are the mostly used by the youngsters in the prime of their life can be effectively used to popularize the Sangam literature. Sangam literature has strong emotional connection with present day lives. Hence if Sangam poems are presented in proper way, it is sure strike a chord with present generation.

## **Obstacles to a common man:**

The biggest obstacle for a common man is the archaic language of Sangam Literature. To understand and appreciate the Sangam poems, a person has to know lot of background details like themes, landscape etc. Dr.Kamil Zvelebil has once remarked that only trained readers can completely understand and appreciate Sangam literature. A common man requires considerable education before he can appreciate Sangam poems. More often than not, translating the poem or explaining the poem line by line doesn't give full experience of the poem. As an example I would like to take a poem and explain how much back ground detail he requires to appreciate the poem

மாரி ஆம்பல் அன்ன கொக்கின்

பார்வல் அஞ்சிய பருவரல் ஈர் ளெண்டு

கண்டல் வேர் அளைச் செலீஇயர், அண்டர்

கயிறு அரி எருத்தின், கதழும் துறைவன்

வாராது அமையினும் அமைக

சிறியவும் உள ஈண்டு, விலைஞர் கைவளையே.

If the chief of maritime land

*Where,  
the wet crab  
seeing the stork,  
Which looked  
Like a water lily in rain ,  
Gets afraid and  
moves inside the hollow roots of  
the mangrove trees ,  
Like a bull which cut loose  
from the rope tied by the herdsman,*

Doesn't return

let it be so.

The merchant sells smaller bangles  
in this town.

In the following poem we expect the reader of this translation to know the following points to enjoy the poem completely

1. How the water lily looks when it rains - and how it is so similar to the crane
2. Wet crab being compared to Thalaivan
3. Crane compared to gossip of women
4. Running for safety into hollow roots, as imagery for Thalaivan hurrying to Thalaivi's home to save his love.
5. Bull cutting loses from the rope tied by the herdsman as imagery for Thalaivan breaking the shackles of the pressure from the society.
6. Small bangles to indicate the Tamil akam tradition of love sickness and thinning of hands and loosening of bangles. Small bangles will fit the smaller hand and hence they can throw away the memory of Thalaivan and be safe from the society.

It can be shown with further examples that traditional methods of translation and brief explanation cannot give a reader the full experience of a sangam poem. Hence experimental approaches should be considered to take this classical literature to the common reader.

### **Blogging as a medium:**

Blogging as medium of communication has developed over the last few years. This medium gives enough flexibility and freedom and in most cases free of cost. Blogging has become a popular medium for readers who like to gather knowledge on various topics. Most internet users are used to blog medium hence blogging overcomes one of the major pitfalls of Book. Books involve printing cost

hence there are severe restrictions on page counts and illustrations. Blogs therefore can afford to be as long as they want and can use illustrations as required. The blogging medium is also flexible enough to allow audio and video. The flexibility of this medium can be really helpful in presenting Sangam literature to the new age reader.

With no restrictions in page count, more explanation can be given on the theme, landscape and other required back stories can be given in addition to translation to understand a particular poem.

Sangam literature has very visual nature to it. The poets of the Sangam infused elements of nature-flora and fauna effectively in the poems to describe emotions. They play an integral part in understanding the poem. Most readers' knowledge of flora and fauna especially with names in archaic Tamil is pretty low. Hence pictures or illustration of flora and fauna mentioned in the poems helps the reader understand the poems better.

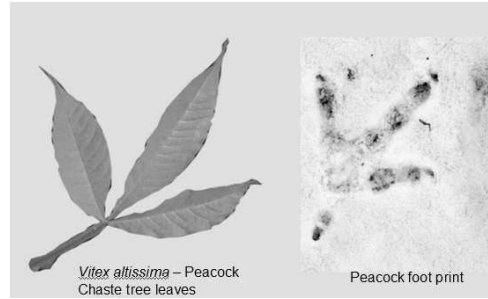
### Experimental approaches:

The visual nature of the Sangam nature can be utilized and the poems can be presented in new formats such as explaining poems with series of pictures or more experimental approach of comic book format (with illustrations and subtext). Audio files with rendering of the poem along with explanations can be uploaded. Lectures on the poem can be recorded and uploaded as video blogs. Small animations can be made to explain similes and imagery of the poem.

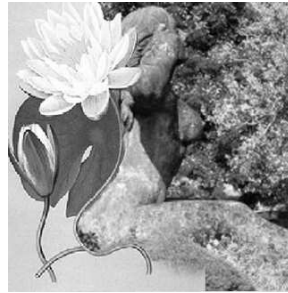
Few examples of illustrating visual elements of Sangam poems are given below



மாரி ஆம்பல் அன்ன கொக்கின்



மயில் அடி இலைய மாக் குரல் நொச்சி



தண் நறுங் கழுநீர்ச் செண் இயற் சிறுபுறம்

A more experimental approach of Sangam poem in comic book style



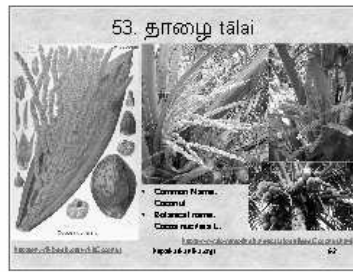
With blogging medium we have no restrictions to creativity. Experimental and creative approaches can help in popularizing the Sangam poems.

### Visual catalogs:

Visual catalogs of various flora fauna and smiles can be made into searchable online databases. In one such effort visual catalog of 99 flowers of Kurinchippattu was created 2 years ago in my blog karkanirka.org. This effort can be extended to include all flora and fauna of Sangam and also efforts can be put convert into a searchable online database. Such visual catalogs would help the common reader as well as researchers in understanding the Sangam literature in much better way.



61



62



63



66



67



68

Few slides from the visual presentation of 99 Flowers of Kurichippaatu.

### Relating to popular Culture:

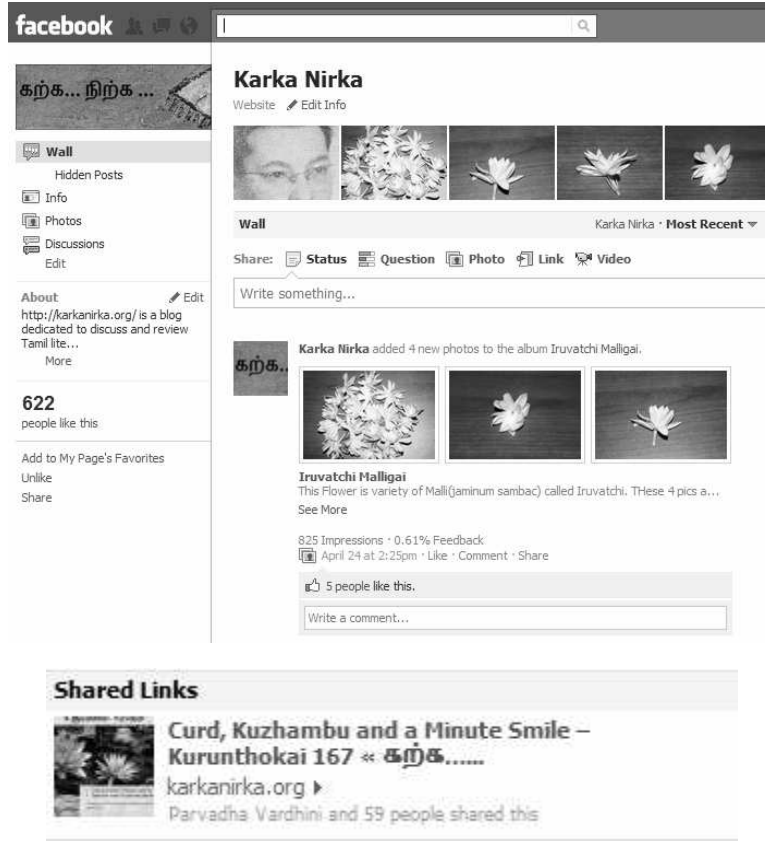
Most themes of Sangam poems are still prevalent in popular culture though Tamil Cinema, either as scenes or as lyrics for film songs. Relating the Sangam poems to prevalent popular culture helps common reader understand the classical poems better since they are able to relate to the emotion of the poem.

Themes such as love at first sight, opposition for love from parents, eloping of lovers, temporary separation of lovers, wife cooking for first time and waiting for approval of husband, girls parents mistaking girls love sickness and inviting priest to perform rituals are still prevalent as movie themes.

Relating the Sangam themes to popular culture makes them understand that Sangam literature isn't as alien as they had imagined and creates an interest in them to explore more in the classical literature.

### Use of Social Network:

Social networks like Orkut, Twitter and Facebook have become the most popular medium for knowledge transfer and sharing. Each of these social networks has hundreds of groups on Tamil Language and literature. Sharing the links of the blogs at such groups helps spread the reach of Sangam literature. This method popularly known as viral marketing is very useful tool. Many new readers are exposed to Sangam literature this way. When a reader likes a poem he shares it to his friends and any one in his friends list who likes the post passes on to his friends. This way with minimum effort the reach of the poems can be maximized. The social networks have played an important role in growth of Karka Nirka blog. In April 2010 Karka Nirka Facebook page was started and presently there are 622 followers and growing. On average one blog posted in Face book is passed on/shared by 50 other people.



## Conclusion:

New approaches are required to popularize the Sangam literature. There are many takers for such efforts. Fresh enthusiasm and creativity can help to revitalize the Sangam Literature and take it to the present generation.







## மாநாட்டுக் கட்டுரைகள் CONFERENCE PAPERS

TAMIL INTERNET 2011



தமிழ் இணையம் 2011

# Enriching Tamil and English Wikipedias

*N. Murugaiyan*

*Chief Resource Person, Central Institute of Classical Tamil,  
Chennai, Tamil Nadu, South India*

## Introduction

‘Tamil,’ as A.K.Ramanujan[i] says, ‘one of the two classical languages of India, is the only language of contemporary India which is recognizably continuous with a classical past’. Kamil V. Zvelebil[ii] says, ‘probably the most significant contribution of Tamil literature, which still remains to be ‘discovered’ and enjoyed by the non-Tamilians and adopted as an essential and remarkable part of universal heritage’. According to Harold Schiffman[iii] ‘most Tamils feel that their language and their linguistic culture really are different from most others in India’ Though information about Tamil literature in works by those referred to above is presented effectively, detailed information about Tamil literary works especially Sangam literature presented in one of the most common online reference sources, namely Wikipedia, is far from satisfactory.

The present paper aims at exploring problems connected with presentation of information about Classical Tamil Literary works in the on line encyclopedias namely, Tamil and English wikis. These two on line reference works being made free, people in different parts of the world have an easy access to them. But unfortunately some of the accounts about Tamil literary works found in them are often found to be highly skeletal, fragmentary, lacking in citations, disappointing the users remaining unverifiable and being incoherent. Hence the paper makes an attempt at analyzing the problems relating to presentation of information through online resources in them.

## Encyclopedias

There are the highly traditional encyclopedias comprising entries developed as per norms of an encyclopedic entry in standard encyclopedias such as Encyclopedia Americana and Encyclopedia Britannica which contain useful information on Tamil literature. However, they are not easily available to those who want to use them for updating their knowledge about Tamil literature in general, specific Tamil literary works in particular. Even though the encyclopedia Britannica is available as an online encyclopedia, the stipulation seeking the users’ credit card number even for a trial run for a few days serves as a factor inhibiting its use. In this context, the Wikipedia encyclopedias (Tamil and English Wikis), freely available for anyone who has an access to a computer with an internet connection, in any part of the world, come in handy or easy to reach. But the information about Tamil literature found in them is far from satisfactory as they are highly fragmentary or skeletal. Against some of the Wikipedia encyclopedic entries suffering from presentation problems carry instructions such as the following:

‘This article needs additional citations for verifications. Please help improve this article by adding reliable references. Un-sourced material may be challenged and removed’ or ‘This Tamil related article

is a stub. You can help Wikipedia by expanding it' or 'This article about the literature of India is a stub. You can help Wikipedia by expanding it'.

Given below is a specimen entry from English Wikipedia on Nānmanikkatikai one of the Eighteen Minor works of the Post-cankam period.

## Nānmaṇikkaṇikai

From Wikipedia, the free encyclopedia

**Nanmanikkatigai** is a Tamil poetic work of didactic nature belonging to the *Pathinenkilkanakku* anthology of Tamil literature. This belongs to the 'post Sangam period corresponding to between 100 – 500 CE. *Nanmanikkatigai* contains one hundred songs written by the poet Vilambi Naganaar. This poetic work is famous for its clarity and easy readability and is often a prescribed text for schools in Tamil Nadu. The poems of *Nanmanikkatigai* are written in the Venpa meter.

The poems of *Nanmanikkatigai* each contain four different ideas. The name *Nanmanikkatigai* denotes this fact comparing the four ideas to four well-chosen gems adorning each poem. The following poem describes four different groups of people who cannot sleep well at night, namely, a thief, a lovelorn person, someone who hankers after money and a miser who worries about losing his money:

கள்வம்என் பார்க்கு துயில் இல்லை, காதலிமாட்டு  
உள்ளம்வைப்பார்க்கும் துயில் இல்லை, ஒண்பொருள்  
செய்வம்என் பார்க்கும் துயில் இல்லை, அப்பொருள்  
காப்பார்க்கும் இல்லை துயில்.

## [edit] References

- Mudaliyar, Singaravelu A., Apithana Cintamani, An encyclopaedia of Tamil Literature, (1931) - Reprinted by Asian Educational Services, New Delhi (1983)
- <http://www.tamilnation.org/literature/>
- <http://www.tamilnation.org/literature/pathinen/pm0047.pdf> *Nanmanikkatigai* eText at Project madurai

புதுப் பயனர் உதவி | தட்டச்சு உதவி | Font help | ஆலமரத்தடி | ஒத்தாசை | அகரமுதலி | செய்திகள்

What is given below can be treated as an expanded encyclopedic entry version which is in no way complete<sup>[iv]</sup>. An entry relating to this post-cankam work can be attempted using various criteria insisted upon for a comprehensive encyclopedic entry.

## Nānmaṇikkaṇikai

Nānmaṇikkaṇikai is one of the Eighteen Minor Works known in Tamil as Patinenkil, 'kkaṇakkunūlkal. It comprises one hundred and six quatrains, the first two quatrains being invocation verses. The venpā in its variant forms is the metre of Nānmaṇikkaṇikai. However, the first verse and the other two namely the 30<sup>th</sup> and the 61<sup>st</sup> verses have five lines in each of them. As the first two quatrains are

in praise of lord Vishnu, people describe it as a Vaishṇavaite work.

Those who view it as the post-sangam work would place it at the later part of Buddhists and Jains, AD 100 to 600 AD.S. Vaiyapuri Pillai in his History of Tamil Language and Literature (From the Beginning to 1000 AD) assigns 750 AD as the period of Composition of Nāṇmaṇikkāṭikai but many scholars do not subscribe to this view. As certain lines from Kuruntokai, one of the works of Eight Anthologies, has been used in the Nāṇmaṇikkāṭikai, it is possible to surmise that the Nāṇmaṇikkāṭikai belongs to a later or post-sangam period. The lines that have a striking similarity in the two works referred to above are the following:

தாயுடன்று அலைக்குங் காலையும் வாய்விட்டு

அன்னாய் என்னும் குழவி (Kuruntokai – 397)

குழவி அலைப்பினும் அன்னே என்று ஓடும் (Nāṇmaṇikkāṭikai 23)

As the coincidence between certain lines of Nāṇmaṇikkāṭikai and the Tirukkural is striking, commentators believe that the former has'al certainly borrowed from the latter giving credence to the view that Nāṇmaṇikkāṭikai belongs to an age later than that of Tirukkural. The quotes cited below will illustrate the point made above.

இனிமையின் இன்னாதது யாதெனின் இன்மையின்

இன்மையே இன்னாதது (The Tirukkural - 1041)

இன்மையின் இன்னாதது இல்லையில் லென்னாத

வன்மையின் வண்பாட்டது இல் (Nāṇmaṇikkāṭikai)

One of the verses of the verses of Nāṇmaṇikkāṭikai and its translation version are given below:

கன்வம்என் பார்க்கு துயில் இல்லை, காதலிமாட்டு

உள்ளம்வைப்பார்க்கும் துயில் இல்லை, ஒண்பொருள்

செய்வம்என் பார்க்கும் துயில் இல்லை, அப்பொருள்

காப்பார்க்கும் இல்லை துயில்.

No sleep for those who are surreptitious; no sleep for

Those who have set their mind on their favorite women

No sleep for those who are keen on wealth creation

And those who safeguard such wealth sleep not.

## References

The translation of the Nāṇmaṇikkāṭikai verse into English quoted above is the one attempted by the presenter of this paper.

Tamil Wikipedia தேடல் முடிவுகள்

கட்டற்ற கலைக்களஞ்சியமான விக்கிப்பீடியாவில் இருந்து.

உங்கள் வினவலுக்கான முடிவுகள் எதுவும் இல்லை.

## "Nanmanikkatikai" பக்கத்தை இந்த விக்சியில் உருவாக்கவும்

As there is no entry relating to Nanmanikkatikai in the Tamil Wiki, an entry similar to the one given below can be made:

நான்மணிக்கடிகை பதினெண்கீழ்க்கணக்கு நூல்களுள் ஒன்றாகும். இது ஒரு கடைச்சங்ககால அல்லது சங்கம் மருவிய கால நூலாகும். இதன் ஆசிரியர் விளம்பினாகனார். இது நாலடி வெண்பாக்களால் ஆனது. ஒரு சில வெண்பாக்கள் ஐந்து அடிகளால் ஆனவை. இந்நூல் நூறு வெண்பாக்களால் ஆனது. சைவசித்தாந்த நூற்பதிப்புக்கழக வெளியீட்டில் 106 பாடல்கள் உள்ளன (1904). கி. ஆ. பெ. விசுவநாதம் வெளியிட்டுள்ள மும்மணிகளும். நான்மணிகளும் என்ற நூலில் 104 வெண்பாக்கள் மட்டுமே உள்ளன (பாரி நிலையம், சென்னை, 1954). இவரின் இன்னுற் பதிப்பில் கடவுள் வாழ்த்துச் செய்யுட்கள் இரண்டும் காணப்படவில்லை யாதலால் 104 செய்யுட்கள் மட்டுமே உள்ளன. இன்னாலின் ஆசிரியர் விளம்பினாகனார். இந்நூலின் கடவுள் வாழ்த்துப் பாடல்கள் திருமாலைப் பற்றி இருப்பதால் இந்நூலாசிரியர் வைஷ்ணவர் என்று கூறப்படுகிறார். இன்னாலாசிரியர் பெயர் விளம்பினாகனார், விளம்பி என்பது இவர் தொழிலைக்குறிப்பதாகவும், தமிழகதில்லுள்ள ஜைனர்கள் நயினார் என்று அழைக்கப்படுவதால் இவர் பெரில் உள்ள நாகனார் என்பது நயினாரரெனக்கொண்டு இவர் ஜைனர் எனக்கொள்வாறுமுள். மேலும் ஜைன சமயக்கருதுக்கள் பல இடங்களில் காணப்படுவதால் இந்நூலாசிரியர் ஜைனரே என்றும் கொள்ளப்படுகிறார்.

### குறிப்புகள்

பதினெண்கீழ்க்கணக்கு நூல்கள் :நான்மணிக்கடிகை, ஆசிரியர் டி. எஸ். பாலசுந்தரம் பிள்ளை என்கிற இளவழகனார், சைவசித்தாந்த நூற்பதிப்புக்கழகம் திருநெல்வேலி லிமிடெட், முதற்பதிப்பு 1904, ஏப்ரல் 1980, திருவரங்கனார் பதிப்பகம், சென்னை 600 018

கி. ஆ. பெ. விசுவநாதன், மும்மணிகளும் நான்மணிகளும், பாரிநிலையம், சென்னை - 600 0108, முதற்பதிப்பு 1954, 11 வது பதிப்பு2007

### Enriching Encyclopedias

An encyclopedic entry will be a long essay comprising a preview or introduction followed by treatment of each item mentioned in the preview in separate sections. Diagrams or maps or tables or charts are inserted wherever required. Use of photos or images or pictures as illustrative materials is attempted wherever possible. A dictionary entry restricts itself with whatever connected with that word such as phonological information, grammatical information, semantic information, idiomatic information connected with that word etc. While an encyclopedic entry deals with whatever connected with the subject referred to by the word. At the end of the article a brief summary of what has been dealt with in the article is presented. For the benefit of those who are interested in acquiring more information on the topic chosen for treatment in the entry, a short bibliography is presented. Certain entries may be as long as a few hundred pages, a table of contents is usually presented enabling easy reference and location of information needed in the entry.

On line encyclopedias offer the additional advantage of being dynamic : new information relating to the subject dealt with are made available in the encyclopedia as when they are available, not waiting for the next static format such as the disc or paper based publication to come out. The Tamil and English Wikis are as much dynamic as the encyclopedia Britannica online, The Wikipedia is one of the first user-generated content encyclopedia. The principles of democracy is enshrined in its making and it would never become obsolete as it is dynamic.

## Mismatch between Technical Know-how and Tamil Studies' Scholarship

Scholars who have a depth of knowledge in Tamil literary studies are not able to enrich the encyclopedias like the Tamil and English Wikipedia as they do not have the technical skills needed for serving as collaborative editors for them. But those who are well versed in using the computer for serving as a collaborative editor have only a superficial knowledge and understanding of Tamil language and literature in general, especially the Cankam or literature of the Academies in particular. As a result of the mismatch between possession of computing skills needed for serving effectively as a collaborative editor of Tamil and English Wikis and the ability to write convincingly on Tamil literature with suitable citations and making references necessary to make their accounts about Tamil literature authenticated and well documented. Even among the scholars who have a thorough knowledge of Tamil literature, only a small group of native scholars are capable of using the English language and the Tamil language for this purpose. Some of the foreign scholars who can use the English language effectively for describing the ancient Tamil literature they are unable to give a convincing account of the literary works as their understanding of the ancient Tamil literature. Superficiality of native and foreign scholars either in the ability to use the computer or in their understanding of the Tamil literature not only by the foreign scholars but also by the native scholars or in their mastery of using either the English language for English Wiki or the Tamil language for the Tamil wiki. As a result of this mismatch, some of the articles written for the wikis remain Stubs which need elaboration and citations for increasing their verifiability and coherence.

## Conclusion

The paper has focused on the sorry state of affairs prevailing in presenting information about Tamil literature in general, classical Tamil Literature in particular in world's most common dynamic reference books such as English and Tamil Wikipedia. Sample material with problems of presented along with improved versions of information. Certain procedures or practices relating to making encyclopedic entries are referred to in the section that deals with enriching encyclopedias. The chief reasons for information presented in the Wikis being incoherent are identified as the mismatch between the technologically savvy suffering from superficiality in Tamil studies and those who have a thorough knowledge in Tamil Studies but not being aware of the technical skills needed for presenting information in the Wikis as collaborative editors

## Notes

- *The Interior Landscape: Love Poems from a Classical Tamil Anthology*, (1967)
- *The Smile of Murugan : On Tamil Literature of South India*
- Language Policy and Linguistic Culture in Tamilnadu, **Chapter 6, on** Tamilnadu from *Linguistic Culture and Language Policy*, H. Schiffman, 1996.
- An entry complete with all the necessary components cannot be attempted because of space constraints prescribed for this paper

# கூட்டாசிரியப் படைப்பு: தமிழ் விக்கிப்பீடியா

செ. இரா. செல்வக்குமார்

மின்னியல் மற்றும் கணினியியல் துறை, வாட்டர்லூ பல்கலைக்கழகம், வாட்டர்லூ, ஒண்டாரியோ,  
கனடா N2L 3G1

selvakumar@uwaterloo.ca (OR) c.r.selvakumar@gmail.com

<http://valluvar.uwaterloo.ca/~selvakum/biop.html>

குறிச்சொற்கள்: தமிழில் இணையவழி கூட்டாசிரியப் படைப்பு, விக்கி தொழில்நுட்பம்

## சுருக்கம்

உலகில் முதன்முறையாகப் பெரிய அளவில் தமிழில் இணையவழி உருவாகிவரும் பல்துறைக் கலைக் களஞ்சியம் தமிழ் விக்கிப்பீடியா. இத்திட்டம் விக்கி (Wiki) என்னும் தொழில் நுட்பத்தால் வளர்ந்துவரும் ஒரு கூட்டாசிரியப் படைப்பு (content created by collaborative authoring). தற்பொழுது ஏறத்தாழ ஒரு கோடி (10 மில்லியன்) சொற்கள் அடங்கிய இக் கலைக்களஞ்சியத்தில் 31,000 கட்டுரைகளுக்கும் மேல் உருவாக்கப்பெற்றுள்ளன. கட்டுரைகளின் சராசரி பைட் (byte) அளவில் உலக மொழிகளின் வரிசையில் 10 ஆவது இடத்தில் உள்ள இத் தமிழ்க் கலைக்களஞ்சியம் எவ்வாறு உருவாக்கப்பட்டு வருகின்றது என்றும், பிற இந்திய மொழிகளிலும், உலக மொழிகளிலும் நிகழ்ந்து வரும் விக்கிப்பீடியா வளர்ச்சிகள் பற்றிய புள்ளிக்குறிப்புகளின் அடிப்படையில் ஒப்பிட்டு சில தரம் சார்ந்த கருத்தலசல்களும் இக் கட்டுரையில் வழங்கப்படுகின்றன. பல நாடுகளில் வாழும் பல பண்பாட்டுப் பின்னணியுடைய தமிழர்கள் ஒன்றிணைந்து அறிவு, தொழில்நுட்பக் கூட்டுழைப்புடன் உருவாக்கிவரும் இக் கூட்டாக்கத்தில் எதிர்கொண்ட சிக்கல்கள் பற்றியும், தீர்வுகள் பற்றியும், சிறப்புக் கூறுகள் பற்றியும் துய்ப்பறிவும் பட்டறிவும் இக் கட்டுரையில் வழங்கப்படும்.

27,000 பேர் பயனர்களாகப் பதிவு செய்துள்ள இத் தளத்தில் இதுகாறும் 778,600 தொகுப்புகள் (edits) செய்யப் பட்டு கட்டுரைகள் உருவாக்கப்பட்டுள்ளன. தமிழ் வளர்ச்சிக்கும், தமிழில் இணையம், கணினி, பொறியியல், கலை, அறிவியல், மருத்துவம் போன்ற அறிவுத்துறைகள் அனைத்துக்கும், பள்ளிப் பாடங்கள் முதல் ஆய்வு மட்ட நூல்கள் வரை பல்வகைப் படைப்புகளை உருவாக்கி பயன்பெருக்க விக்கித் திட்டம் எவ்வாறு துணை செய்யக்கூடும் என்றும் கருத்துகள் முன்வைக்கப்படும்.

## 1. அறிமுகம்

எளிய எழுதுகோல் முதல் வானூர்தி, ஏவுகணை, கணினி வரை ஏறத்தாழ அனைத்துமே பலருடைய கூட்டுழைப்பால் உருவாக்கப்படுவனவே. ஆனால் கதை, புதினம், கவிதை, புதினமல்லா உரைநடை நூல்கள் போன்ற எழுத்துப் படைப்பிலக்கியம் போன்றவற்றைத் தவிர, வேறுபல எழுத்துப்படைப்புகளும் ஒருவாறு கூட்டாக, பல ஆசிரியர்கள் இணைந்து உருவாக்குவன. என்றாலும், பலர் உருவாக்கும் உசாத்துணை நூல்கள் (reference works), கலைக்களஞ்சியம் போன்றவையும் தனித்தனியே பலர் எழுதி, பின்னர் பிணைத்துத் தொகுக்கப்படுவன. திருத்தங்கள் செய்யும் பொறுப்பாசிரியர்களின் பங்களிப்பைத் தவிர எழுத்தில் பெரிதாக கூட்டுழைப்பு இல்லை எனலாம். தகவல் திரட்டலில் கூட்டுழைப்பு இருக்கலாம். ஆய்வுக்கட்டுரைகளின் படைப்பில் பல ஆசிரியர்களின் கூட்டுப்பங்களிப்பும், “கூட்டு ஆசிரியராக” இருக்கும் நிலையும் வேறு ஒரு கூட்டுப் படைப்பு. 1993 ஆன் ஆண்டு நியூ இங்கிலாந்து செர்னல் (New England Journal) வெளியிட்ட ஆய்வுத்தாள் ஒன்றின் ஆசிரியராக 972 பேரைக் குறிப்பிட்டிருந்தது [1]. 2008 ஆண்டு, அணுத்துகள் பற்றிய ஆய்வுத்தாள் ஒன்றை செர்னல் ஆப் இன்சுற்றுமென்ட்ரேசன் (Journal of Instrumentation) வெளியிட்டது; அதில் 169 ஆய்வகங்களைச் சேர்ந்த 2,926 பேர் அக்கட்டுரையின் ஆசிரியர்களாக தெரிவிக்கப்பெற்றனர். ஆனால் இப்படியான



“கூட்டு ஆசிரியர்கள்” படைப்பு வேறு வகையானது. இவை போல் அல்லாமல், இக் கட்டுரையில் குறிப்பிடப்பெறும் கூட்டு ஆசிரியப் (collaborative authoring) படைப்பு என்பது அண்மையில் உருவான கணினி சார்ந்த தொழில்நுட்ப வசதியால், புதினங்கள் முதல் அறிவியல், வாழ்வியல் கலைக் களஞ்சியங்கள், பொறியியல் கையேடுகள், மருத்துவமனை தகவல் பராமரிப்பு ஒருங்கியம் (Information Management system) போன்ற பற்பல பயன்பாட்டுக்கும், தனித்தனியாக பிணைத்து சேர்க்காமல், பிரித்தறிய அரிதான வகையில் பலரும் சேர்ந்து எழுத்துருவாக்கம் செய்ய இயலும் கூட்டாசிரியப் படைப்பைப் பற்றியது [10]. இவ்வகைக் கூட்டாசிரியப் படைப்பு எழுத்துப் படைப்பாக மட்டும் அல்லாமல், மென்கலன் உருவாக்கம் போன்றவற்றுக்கும் பயன்படுகின்றது. பொதுவாக கூட்டாசிரியப் படைப்புகளுக்குப் பயன்படும் அறிவுத்தகவல்கள் பலவற்றையும் கலிபோர்னியா பல்கழகத்தைச் சேர்ந்த சிம் வொய்ட்ஃகெட் (Jim Whitehead) பராமரிக்கும் வலைத்தளத்தில் காணலாம் [3]. இவ்வகையான கூட்டாசிரியப் படைப்புக்கு அடிப்படையாக உள்ள தொழில்நுட்பங்களில் முகன்மையான ஒன்று விக்கி (“Wiki”) என்று அழைக்கப்படும் மென்கலன் (software). மென்கலத்துறை இலக்கியத்தில் அலசப்படும் குழுப்பயன்பாட்டு மென்கலன் (groupware) மற்றும் பதிப்புநிலை கண்காணிப்பு வகையான (Version Control) (இது மென்கல வடிவொழுக்கு மேலாண்மை (SCM, Software, Configuration Management) வகையை சேர்ந்தது) போன்ற மென்கல நுட்பங்களோடு தொடர்புடையது இந்த விக்கிநுட்பம். பொதுவாக இந்நுட்பத்தின் உதவியால் ஒரே நேரத்தில் பல இடங்களில் இருக்கும் பலர், இணையவழி ஒரு கட்டுரையையோ அல்லது ஆவணத்தையோ திருத்தவும் வளர்த்தெடுக்கவும், புதுக் கட்டுரைகளையும் உருப்பதிகளையும் உருவாக்கி சேர்க்கவும், எளிதாக வகைப்படுத்தவும், முன் பதிவுவடிவங்கள் எதுவும் அழியாமல், எல்லாக் கட்டங்களின் பதிவுகளையும் மீட்டெடுக்கும் வசதியும் படைத்தது.

இக்கட்டுரையில் முதலில் விக்கி என்றால் என்ன என்று விளக்கிய பிறகு, விக்கிநுட்பத்தால் எவ்வாறு பல்துறைசார்ந்த பல நாட்டுத் தமிழர்கள் ஒன்றிணைந்து கூட்டாசிரியப் படைப்பாக இத்தமிழ்க்கலைக் களஞ்சியத்தை உருவாக்கி வருகிறார்கள் என்றும், அதன் தரங்களைப் பற்றிய பார்வைகளும் ஒரு நிகழ் எடுத்துக்காட்டு (அல்லது case study) என்னும் அளவில் முன்வைக்கின்றேன். கூட்டாசிரியப் படைப்பில் ஏற்படும் நன்மைகள், சிக்கல்கள் பற்றியும், கடந்த 5 ஆண்டுகளாக பங்களித்த பட்டறிவையும் பகர்கின்றேன். இத்தொழில்நுட்பத்தைப் பயன்படுத்தி, இதன் நீட்சியாக செய்யத்தக்கவை பற்றியும் மிகச் சுருக்கமாக இறுதியில் கூறுகின்றேன்.

## 2. விக்கி என்றால் என்ன?

விக்கி என்ற சொல்லையும் அதன் கருத்தாக்கத்தையும் போ’ லியூஃவ் (Bo Leuf), வார்டு கன்னிங்காம் (Ward Cunningham) ஆகியோர் 1995 இல் அறிமுகப்படுத்தினர். இந்த விக்கி (Wiki) என்னும் சொல் அவாயி மொழியில் (Hawaiian) விக்கிவிக்கி (wikiwiki) என்றால் சட்டுசட்டென்று, கிடுகிடு, மளமள என்பது போன்ற இரட்டித்து வந்து அழுத்தத்தோடு விரைவைக் குறிக்கும் சொல்லில் இருந்து பெற்றது (ஆங்கிலத்திலும் பிறமொழிகளிலும்). விரைவாக (எளிதாகவும்) மாற்றங்கள் ஏற்படுத்த வல்ல தொழில்நுட்பம் என்னும் பொருளில் இச்சொல் இன்று அறியப்படுகின்றது. இதனை ஆக்ஃசுபோர்டு ஆங்கில அகரமுதலியர், தம் 2006 ஆம் ஆண்டுப் பதிப்பில் உள்வாங்கிக்கொண்டனர் [4]. இச்சொல்லின் பயன்பாடும், இக்கருத்துருவை செயற்படுத்திய தொழில்நுட்பமும், வார்டு கன்னிங்காம் 1995 இல் முதன்முதல் உருவாக்கிய விக்கிவிக்கிவெப்’ (wikiwikiweb) என்னும் மென்பொருளில் இருந்து தொடங்குகின்றது (<http://www.c2.com/cgi/wiki>). இன்று 200 வகைகளுக்கும் மேலான விக்கிமென்கலங்கள் உள்ளன. எனினும் மீடியாவிக்கி (MediaWiki) என்னும் விக்கி தொழிநுட்பத்தைப் பயன்படுத்தி ஆங்கிலத்தில் 2001 ஆம் ஆண்டு உருவாகத் தொடங்கி, இன்று பெருக வளர்ந்துள்ள, இலவசமாகக் கிடைக்கும், கட்டற்ற பல்துறை விக்கிப்பீடியா என்னும் கலைக்களஞ்சியத்தால் இத்தொழில்நுட்பம் பரவலாக அறியப்படுகின்றது [5]. இன்று ஆங்கிலமொழியில் 3.6 மில்லியனுக்கும்

கூடுதலான தலைப்புகள் (கட்டுரைகளும் குறிப்புரைகளும்) கொண்ட இக்கலைக்களஞ்சியம் மிகப் பரவலாக பயன்படுத்தப் படுகின்றது (எல்லா மொழிகளின் கட்டுரைகளும் சேர்ந்து 18 மில்லியன்). இன்று மொத்தம் 281 மொழிகளில் இவ் விக்கி தொழில்நுட்பத்தைப் பயன்படுத்தி விக்கிப்பீடியா (வகை) கலைக்களஞ்சியங்கள் உள்ளன [6]. எல்லா மொழிகளுக்கும்மாகச் சேர்ந்து, தனித்து அறியத்தக்க வருகையாளர் (unique visitors), 1.3 பில்லியனைத் தாண்டுகிறது [6]. எல்லா “விக்கி”களும் விக்கிப்பீடியா போன்ற கருத்துகளை உருவாக்கி, வளர்த்தெடுத்து வகைப்படுத்துவன அல்ல (பார்க்க: <http://c2.com/cgi-bin/wiki?ContentCreationWiki>).

### 3. தமிழ் விக்கிப்பீடியா

தமிழ் விக்கிப்பீடியா செப்டம்பர் 30, 2003 இல் தொடங்கப்பெற்றது. இதன் வரலாற்றைத் தமிழ் விக்கிப் பீடியாவில் காணலாம் [7]. விக்கிப்பீடியாவின் வரலாற்றை தேனி எம். சுப்பிரமணி தமிழில் நூலாகவும் எழுதியுள்ளார் [8] யாழ்ப்பாணத்தைச் சேர்ந்த இ. மயூரநாதன் என்பவர் நவம்பர் 20, 2003 முதல் தமிழ் விக்கிப்பீடியாவில் பங்கு பற்றி பணியாற்றத் தொடங்கிய பிறகே தமிழ் விக்கிப்பீடியா வளர்ச்சியடையத் தொடக்கியது. முதற்கட்டங்களில், தமிழ் எழுத்துகளில் எழுதுவதும், விக்கிப்பீடியாவுக்குத் தேவையான தமிழ் இடைமுகங்களை உருவாக்குவதிலும் பெரும் இடர்ப்பாடுகள் இருந்தன. இன்று கணிதச் சமன்பாடுகளில் தமிழ் எழுத்துகள் சேர்ப்பதைத் தவிர, ஏறத்தாழ எல்லா இடைமுகங்களும் தமிழில் உருவாக்கிப் பயன்படுத்த இயலுகின்றது. இன்று இந்திய மொழிகளில் முன்னணியில் இருக்கும் ஒரு விக்கிப்பீடியாவாகத் தமிழ் விக்கிப்பீடியா உள்ளது. தமிழ் விக்கிப் பீடியாவில் கட்டுரைகள் உருவாக்குவதிலும், பலர் சேர்ந்து கூட்டுழைப்பால் கூட்டாசிரியப் படைப்பாக உள்ளடக்கங்களை வளர்த்தெடுப்பதிலும் எதிர்கொண்ட நன்மைகளையும் சிக்கல்களையும் விரிக்கும் முன்னர், விக்கிப்பீடியாவில் உள்ள கட்டுரைகளின் சில எளிய தர அளவீடுகள் பற்றிக் குறிப்பிடல் வேண்டும். கட்டுரைகளை வெறும் ஏற்புபெற்ற (“official”) “கட்டுரை” எண்ணிக்கையைக் கணக்கில் கொண்டால் தமிழ் விக்கிப்பீடியா இந்திய மொழிகளில் நான்காவதாக உள்ளது (இந்தி, தெலுங்கு, மராத்தி ஆகிய மொழிகளுக்கு அடுத்து), ஆனால் குறைந்தது 200 எழுத்துகளாவது (characters) உள்ள கட்டுரைகள் என்று பார்த்தால் தமிழ் விக்கிப்பீடியா இந்திக்கு அடுத்து இரண்டாம் நிலையில் உள்ளது. கலைக்களஞ்சியத்தின் தரமானது எண்ணிக்கை, சராசரி பைட் அளவு (bytes), மொத்த பைட் அளவு, கட்டுரையின் நீளம் ஆகியவனவற்றில் மட்டும் இல்லை என்றாலும், இவை அனைத்திலும் தமிழ், இந்திய மொழிகளில் முதல் 2-3 இடங்களில் உள்ளது. இப்படியான “தர” அளவீடுகளை அட்டவணை-1 இல் காணலாம் (மே 2010 வரையிலான தரவுகள்).

### 4. சிறப்பான நன்மைகளும் எதிர்கொண்ட, கொள்ளும் சிக்கல்களும்:

நன்மைகள்: (1) தமிழ் எழுத்து வரலாற்றில், பல நாட்டில் வாழும் தமிழர்கள், பல்வேறு வகைப்பட்ட குழுமொழி (dialect), பண்பாட்டுப் பின்புலங்கள் உள்ளவர்கள், இப்படித் தாங்களாகவே தன்னார்வலர்களாக ஒன்றிணைந்து கூட்டாக உழைத்து ஒரு கோடி சொற்களுக்கும் கூடுதலானவற்றால் உருவாக்கிய 31,000 கட்டுரைகள் கொண்ட பல்துறை கருத்துகள் சார்ந்த ஒரு பொது கலைக்களஞ்சியம் உருவாக்கியது முதன் முறை. கூட்டாசிரிய முயற்சிகளில், அதுவும் விக்கிப்பீடியா போன்ற யாரும் பங்குகொள்ளக்கூடிய ஒரு முயற்சியில், வளர்முகமான இவ் வகையான வளர்ச்சி ஏற்பட்டது குறிப்பிடத்தக்கது. (2) அபிதான கோசம், அபிதான சிந்தாமணி, 1960களில் உருவான தமிழ்க்கலைக்களஞ்சியம் முதல் அண்மையில் வெளியான பிரிட்டானிகா தகவல் களஞ்சியம், தஞ்சாவூர் தமிழ்ப் பல்கலைக்கழகம் வெளியிட்ட 34 தொகுதிகள் கொண்ட அறிவியல், வாழ்வியல் கலைக்களஞ்சியம் வரை குறைந்தது 20 தமிழ்க்கலைக்களஞ்சியங்கள் சிறிதும் பெரிதுமாய் அச்சில் வெளிவந்துள்ளன [9]. ஆனால் தமிழ் விக்கிப்பீடியாவில், தகவல்கள் உடனுக்குடன் இன்றையநிலை ஆக்கப்பட்டு வழங்குவதோடு, பிற மொழி கலைக்களஞ்சியங்களோடு உடனுக்குடன் ஒப்பிடக்கூடியதாகவும் உள்ளது. இது இணையத்தில் கிடைக்கக்

கூடிய இலவச கலைக்களஞ்சியமாக, யாரும் தொகுக்கக்கூடியதாகவும் உள்ளது. (3) பிற மொழி விக்சிப் பீடியாக்களில் (எ.கா: ஆங்கிலம்) திருத்துவதிலும் தொகுப்பதிலும் பல்வேறு கடும் கருத்துப் போராட்டங்கள் (எதிர்-எதிர் திருத்தம்/தொகுப்புகள் edit wars) நிகழ்வதும் சிறுபான்மையான கட்டுரைகளில் உண்டு. ஆனால் அவை தமிழ் விக்சிப்பீடியாவில் மிகவும் குறைவாகவே நிகழ்ந்துள்ளன. தமிழ் விக்சிப்பீடியாவில் நிகழ்ந்த கருத்து உறவாட்டங்கள் மிக மிகப் பெரும்பாலும் வளர்முகமாகவே நிகழ்ந்துள்ளன. இவை விக்சிப்பீடியாவின் ஐந்து பெரும் தூண்கள் என்னும் கொள்கைப்படி ([http://en.wikipedia.org/wiki/Wikipedia:Five\\_pillars](http://en.wikipedia.org/wiki/Wikipedia:Five_pillars)) மிகப்பெரும்பாலும் நடந்துள்ளன. (4) இதுவரை தமிழில் எங்கும் பதிவாகாத கருத்துகள் பல்லாயிரக்கணக்கில் சிறிதும் பெரிதுமாக பதிவாகி உள்ளன. சில கருத்துகள் உலகில் வெளிப்பட்டவுடன் உடனுக்குடன் கலைக்களஞ்சிய நோக்கில் பதிவாகி உள்ளன (எ.கா: நினைவுகொள் மின்தடை (memristor)). சில கண்டுபிடிப்புகள் ஆங்கில விக்சிப்பீடியாவிலோ பிற மொழி விக்சிப்பீடியாக்களிலோ பதிவாகும் முன்னமே தமிழ் விக்சிப் பீடியாவில் பதிவாகி உள்ளன. (5) ஆயிரக்கணக்கான புதிய கலைச்சொல்லாக்கங்கள் இயல் சூழலில் உருவாக்கிப் பயன்படுத்தப் பட்டுள்ளன. (6) தமிழ் நாட்டில் சில பள்ளிகளிலும் மாணவர்கள் பாட நேரத்தில் தமிழ் விக்சிப்பீடியாவைப் பார்த்துப் பயன்கொள்ளுகிறார்கள் என்று அறிந்து இன்னும் பொறுப்பாக விக்சிப்பீடியர்கள் பங்களிக்கிறார்கள். (7) தமிழ் விக்சிப்பீடியா உருவாக்கும் சூழலில் புதிய அறிவுசார்ந்த, கூட்டுழைப்பானது உறவாட்டத்தில் புதிய விழுமிய வளர்முக நிலைகள் எட்ட வாய்ப்பு அளித்தது. தொழில்நுட்பத்தோடு குழு, குமுக (சமூக) உறவாட்ட, ஒத்துழைப்புப் பழக்கங்களும் அலசப்படும் ஒரு கருத்தாக பிறமொழி விக்சிகளும் உள்ளது.

சிக்கல்கள்: (1) விக்சி தொழில்நுட்பம் எளிதே ஆயினும், தமிழ் அறிந்தவர்களில் பலரும் இன்னமும் தமிழில் கணினிவழி தமிழில் உள்ளீடு செய்யப் பழகவில்லை. (2) ஏறத்தாழ 10,000 தமிழ் வலைப் பதிவர்கள் இருந்த போதும், அவர்களில் பலரும் அலங்கார நடை இன்றி, பொதுவாக நடுநிலை நின்று கருத்தை நடுவாக முன் வைத்து கட்டுரை நடையில் எழுதுவதில் போதிய பட்டறிவு இல்லை, அல்லது ஆர்வமுடன் முன்வருவதில்லை. தமிழ் விக்சிப்பீடியா ஆர்வலர்கள் தரும் பயிற்சிப் பட்டறைகளிலும் இது பற்றி அதிகம் பயிற்சி அளிக்கவில்லை. பொதுவாக நல்ல மொழி நடை பற்றிய போதிய விழிப்புணர்வு இல்லை. (3) யாரும் தொகுக்கக்கூடிய கலைக்களஞ்சியம் என்பதால், இதில் தரும் கருத்து களுக்கும் தகவல்களுக்கும் போதிய துல்லியமான சான்றுகோள்கள் தரவேண்டும் என்னும் பரிந்துரை இன்னும் பரவலாக எடுபடவில்லை. ஆனால் இத்தேவைகளைச் சுட்டும் குறிச்சொல் (tag, flag) இடும் வசதி இத்தொழில்நுட்பத்தில் இருப்பதால், முன்னேற வழிகளும் உள்ளன. (4) தன்னார்வர்களால் தொகுக்கப்படுவதால், சீராக எல்லா தலைப்புகளிலும் கட்டுரைகள் எழுதப்படுவதில்லை. எடுத்துக் காட்டாக திரைப்பட நடிக நடிகைகள் மீது ஆர்வம் உள்ளவர் அது பற்றியே நிறைய எழுதக்கூடும், ஆனால் நோபல் பரிசு பெற்றவர்களைப் பற்றியோ அவர்கள் கண்டுபிடிப்புகள் பற்றியோ அதிகம் எழுதப் படாமல் இருக்கலாம். (5) தமிழர்களின் மொழிப் பயன்பாட்டில் பொதுவாக இலங்கைத் தமிழர்களுக்கும் தமிழ்நாட்டுத் தமிழர்களுக்கும் இடையே பல இடங்களில் குறிப்பிடத்தக்க வேறுபாடுகள் இருக்கின்றன. இவை இரண்டு வகையானவை ஒன்று சொல்வழக்குகள், எழுத்து நடை முதலானவை, மற்றது ஆங்கிலம் போன்ற பிறமொழிச் சொற்களைத் தமிழில் ஒலிபெயர்க்கும் பொழுது ஏற்படும் பெரும் மாறுபாடு. எ.கா Toronto என்னும் சொல்லை யாழ்ப்பாணத் தமிழர்கள் ரொறன்ரோ என்னும் தமிழ்நாட்டுத் தமிழர்கள் டொரண்டோ என்னும் எழுதுதல். (இப்படியான சுழல்களில் இரண்டையும் வழங்குவதும், தேடுவோர் சரியான கட்டுரையை அடையுமாறும் வசதிகள் செய்யப்பட்டுள்ளன). இலங்கையிலும் தமிழ்நாட்டிலும் பாடநூல்களில் வழங்கும் சொற்கள் பல இடங்களில் மாறுபடுவதையும் இதே முறையில் தீர்க்கப் படுகின்றது. (6) சொற்கள், மொழிநடை, எழுத்துப் பெயர்ப்பு ஆகிய பலவற்றைப் பற்றியும் மிகப்பல நேரங்களில் சீர்மை நோக்கியோ, எது “சரி” என்னும் நோக்கிலோ எழும் எதிர்-எதிர் கருத்துகள்

சிக்கல்களை எழுப்பியுள்ளன,. ஆனால் அவை நடுநிலை நின்று எதிராளியைப் புரிந்துகொள்வதும், பொது நன்மைகருதி இணக்க முடிவுகள் எடுக்க முடிந்ததும் ஒரு புதிய விழிப்புணர்வு ஏற்படுத்தியது.

### கலைக்களஞ்சியம் அல்லாத விக்கி தொழில்நுட்பத்தின் பயன்கள்:

உள்ளடக்க உருவாக்கம் என்னும் அளவில் பிற பயன்பாடுகளும் இதைப் போன்றதே என்றாலும் பள்ளிப் பாடங்கள் முதல், உயராய்வுக் கல்வி வரை பாடங்களும் பல்வேறு அறிவுத் தொகுப்புக்குக்கும் பயிற்சி களுக்கும் இத் தொழில்நுட்பம் பயன்படும். முதன் முறையாக பன் மொழி அகரமுதலி தமிழ் விக்கிசனரி என்று விக்கிப்பீடியாவின் உறவுத்திட்டமாக உள்ளது. உலக மொழிகளில் முதல் 10 மொழிகளில் ஒன்றாகத் தமிழ் விக்கிசனரி உள்ளது (<http://meta.wikimedia.org/wiki/Wiktionary#Statistics> அணுகப் பட்ட நாள் ஏப்பிரல் 30, 2011). இது தவிர, மருத்துவம் சட்டம், ஆட்சி ஆவணங்கள், பொறியியல் கையேடுகள் போன்ற, பல வகையான பயன்பாட்டுகளுக்கும் இத் தொழில்நுட்பம் பயன்படும்.

இங்கு கருத்தில் எடுத்துக்கொள்ளப்படாதவை சில: ஏன் கூட்டாசிரியப் படைப்பு முக்கியம்? (பலர் பங்களிப்பதால் கருத்துகள் செம்மையாக **படைக்கப்படுகின்றனவா, அல்லது** கெடுகின்றனவா?) [12]. சட்டப்படி எழக்கூடிய எழுத்து உரிமச் சிக்கல் ஏதும் உள்ளனவா? படைத்தவருக்குப் போதிய நிறைவு ஏற்படுகின்றதா? வளர்முகமாகச் அணுகுவதில் மேலாண்மை செய்வதில் ஏற்படக்கூடிய சிக்கல்கள் (குழுக்களாக பிரிந்து வளர்ச்சியைத் தடுக்கக்கூடிய வாய்ப்புக்கூறுகள்) யாவை? மெட்காஃவ் விதிபோல் (Metcalf's law) பலர் பயன்படுத்துவதால் பயன் கூடுகின்றதா? (இணைக்கப்பட்டவர்களின் எண்ணிக்கையின் இருபடிய மதிப்பா?) முதலியன இங்கு கருதப்படவில்லை.

**அட்டவணை-1:** மே 2010 தர அளவீடுகள் ஒப்பீடு - இந்திய மொழிகள்

### துணைநூல், ஆவணப் பட்டியலும் குறிப்புகளும்:

1. Investigators, The Gusto (1993). "An International Randomized Trial Comparing Four Thrombolytic Strategies for Acute Myocardial Infarction". The New England Journal of Medicine 329 (10): 673. doi:10.1056/NEJM199309023291001. PMID 8204123
2. Collaboration, The Atlas; Aad, G; Abat, E; Abdallah, J; Abdelalim, A A; Abdesselam, A; Abidinov, O; Abi, B A et al. (2008). "The ATLAS Experiment at the CERN Large Hadron Collider". Journal of Instrumentation 3 (08): S08003. doi:10.1088/1748-0221/3/08/S08003.
3. <http://users.soe.ucsc.edu/~ejw/collab/>
4. wiki, ஆக்ஃசுபோர்டு ஆங்கில அகராதி (OED), மூன்றாம் பதிப்பு, 2006; Third edition, December 2006; online version March 2011. <http://www.oed.com:80/Entry/267577> ;
5. சிம்மி வேல்சு (Jimmy Wales), லாரி சாங்கர் (Larry Snger) ஆகிய இருவரும் இலாப நோக்கற்ற விக்கிப்பீடியாவை நிறுவும் முன், நியூப்பீடியா (Nupedia) என்னும் முயற்சி 2000 இல் தொடங்கி அதிகம் வெற்றி பெறவில்லை. 1999 இல் மார்க்கு கூசிடால் (Mark Guzidal) என்பவர் விக்கித் தொழில்நுட்பத்தைக் கொண்டு கோவெப் (CoWeb) என்பதை நிறுவினார்.
6. <http://stats.wikimedia.org/reportcard/>
7. [http://ta.wikipedia.org/wiki/தமிழ்\\_விக்கிப்பீடியா](http://ta.wikipedia.org/wiki/தமிழ்_விக்கிப்பீடியா)
8. தேனி. எம். சுப்பிரமணி, தமிழ் விக்கிப்பீடியா, மணிவாசகர் பதிப்பகம் வெளியீடு, நவம்பர் 2010..
9. [http://ta.wikipedia.org/தமிழ்க்\\_கலைக்களஞ்சியங்கள்](http://ta.wikipedia.org/தமிழ்க்_கலைக்களஞ்சியங்கள்)
10. Leuf, B, Cunningham, W, The Wiki Way. Quick Collaboration on the Web. Addison-Wesley, Boston, 2001.
11. Bordin Sapsomboon, Restiani Andriati, Linda Roberts and Michael B. Spring, "Software to Aid Collaboration: Focus on Collaborative Authoring"
12. Dillon A. How Collaborative is Collaborative Writing? An Analysis of the Production of Two Technical Reports., pages 69--86. Springer-Verlag, London, 1993.

அட்டவணை-1: மே 2010 தர அளவீடுகள் ஒப்பீடு - இந்திய மொழிகள்

| மொழி        | விக்னியியர்கள் | ஒதுப்புபெற்ற கட்டுரை எண்ணிக்கை | 200 எழுத்துகளுக்கும் கூடுதலாக உள்ள கட்டுரைகள் | சராசரி பைட் அளவு | நீளம் 500 ப் ட்டைத் தாண்டியவை | நீளம் 2 கி.பைட் ட்டைத் தாண்டியவை | சொற்கள் | மொத்த பைட் அளவு | படங்கள்  |
|-------------|----------------|--------------------------------|---|------------------|-------------------------------|----------------------------------|---------|-----------------|----------|
| தமிழ்       | 398            | 23 க்                          | 22 க்   | 3320             | 83%                           | 28%                              | 200 MB  | 8.6 M           | 15 க்    |
| இந்தி       | 460            | 58 க்                          | 35 க்   | 1476             | 42%                           | 11%                              | 213 MB  | 13.3 M          | 14 க்    |
| வங்காளி     | 319            | 21 க்                          | 16 க்   | 1536             | 60%                           | 16%                              | 94 MB   | 4.7 M           | 9.4 க்   |
| மராத்தி     | 231            | 29 க்                          | 11 க்   | 852              | 26%                           | 7%                               | 73 MB   | 3.3 M           | 6.1 க்   |
| தெலுங்கு    | 358            | 45 க்                          | 17 க்   | 1120             | 23%                           | 9%                               | 130 MB  | 6.3 M           | 10 க்    |
| மலையாளம்    | 458            | 13 க்                          | 12 க்   | 2699             | 84%                           | 34%                              | 97 MB   | 3.8 M           | 10 க்    |
| கன்னடம்     | 176            | 8.6 க்                         | 7.3 க்  | 4497             | 56%                           | 21%                              | 98 MB   | 4.6 M           | 7.8 க்   |
| குசராத்தி   | 79             | 15 க்                          | 14 க்   | 1267             | 34%                           | 5%                               | 50 MB   | 2.9 M           | 1.4 க்   |
| பஞ்சாபி     | 25             | 1.7 க்                         | 0.479 க்                                      | 1071             | 20%                           | 11%                              | 6.7 MB  | 0.335 M         | 0.645 க் |
| சமசுகிருதம் | 39             | 4.0 க்                         | 0.379 க்                                      | 201              | 5%                            | 1%                               | 6.9 MB  | 0.165 M         | 0.125 க் |

# விக்கிபீடியா - தமிழ் நிரலிகள்

## ச. சந்திரகலா

தமிழ்த்துறை முனைவர் பட்ட ஆய்வாளர்,  
அவினாசிலிங்கம் நிகர் நிலை பல்கலைக்கழகம், கோவை, தமிழ்நாடு, இந்தியா.  
chandrakala.vetrivel@gmail.com <mailto:chandrakala.vetrivel@gmail.com>

### முன்னுரை

விக்கிபீடியா என்பது 'விக்கிமீடியா' என்ற நிறுவனத்தால் உருவாக்கப்பட்ட தன்னலமற்ற, கட்டற்ற கலைக்களஞ்சியம் ஆகும். இது ஓர் பன்மொழி பனுவல்களை உள்ளடக்கியதாகும். இந்நிறுவனம் தமிழ் இலக்கியங்களின் சிறப்புக்களை பறைசாற்றும் வகையில் சங்க இலக்கியம் முதல் தற்கால இலக்கியம் வரையிலான கருத்துக்களைத் தொகுத்தளித்துள்ளது சிறப்பிற்குரியனவாகும். விக்கிபீடியா, முதற்பக்கம், சமுதாய வலைவாசல், நடப்பு நிகழ்வுகள், அண்மைய மாற்றங்கள், ஏதாவது ஒரு கட்டுரை, உதவி நன்கொடைகள், தூதரகம் போன்ற உட்பிரிவுகளில் அமைந்துள்ளன. தமிழ் அறிவியல் புவியியல் பண்பாடு, கணிதம், சமூகம், வரலாறு, தொழில்நுட்பம், நபர்கள் போன்ற தலைப்புகளில் அகர வரிசைப்படி செய்திகளை உள்ளடக்கியுள்ளது. நடயுத் நிகழ்வுகள் மற்றும் விக்கிபீடியர் அறிமுகம் போன்ற செய்திகள் முதற்பக்கத்தில் இடம்பெற்றுள்ளன.

### கணிப்பொறித்தமிழின் அவசியம்:

ஆய்வு நோக்கில் ஆராயும் ஆய்வாளர் மற்றும் தமிழ் ஆர்வலர்களுக்கு எண்ணற்ற ஆய்வுகளத்தினை விக்கிபீடியா அளித்துள்ளது. இத்தகைய சிறப்பு மிக்க தமிழ் இலக்கிய கருத்துச் செறிவினை சிலர் மட்டுமே பயன்படுத்துகின்றனர். குறிப்பாக சில மாணவர்கள் இணையத்தில் தமிழ் பற்றிய செய்திகளை அறியாதவர்களாகவே உள்ளனர். இக்குறையினை நீக்க மாணவர்களுக்கு 'கணிப்பொறித்தமிழ்' "இணையமும் தமிழும்" அவசியமாகும். அப்பாடத்திட்டம் நூல்வழிக் கற்றல் மட்டுமின்றி, செயல்வழிக் கற்றலுக்கு முக்கியத்துவம் அளித்தல்வேண்டும். விக்கிபீடியா, 'நீங்களும் எழுதலாம்' பகுதியில் எண்ணற்ற தலைப்பின் கீழ் கட்டுரைகளை வேற்கின்றது. ஆராய்ந்து கட்டுரைகளை சமர்ப்பிக்க, செயல்வழிக் கற்றல் முறை மிகவும் பயனுள்ளதாக அமையும்.

### தமிழ் இலக்கணப் பகுப்பு:

விக்கிபீடியாவில் தமிழ் இலக்கணம் பற்றிய செய்திகள் சிறப்பாக இடம் பெற்றுள்ளன. இருந்தபோதிலும், தொல்காப்பியம், இறையனார் அகப்பொருள், அவிநயம், போன்ற 53 இலக்கண நூல்களின் அறிமுகம் மட்டுமின்றி அவற்றினை பகுப்பு முறையில் கொடுத்திருந்தால், இலக்கணம் சுமையானதாக அன்றி, சுவையானதாகக் கருதப்படும். எ.காட்டாக, தொல்காப்பியம்

எழுத்து சொல் பொருள்

அகம் புறம் செய்யுள் உவமை

அணி, அலங்காரம்

தண்டியலங்காரம், மாறன் அலங்காரம் முதலியன

இறையனார் அகப்பொருள் புறப்பொருள்வெண்பாமாலை, யாப்பு

தமிழ் நெறி முதலியன யாப்பெருங்கலம்

யாப்பெருங்காரிகை பாட்டியல், வெண்பாட்டியல், முதலியன

## செவ்வியல் இலக்கியம்:

விக்கிரமபதியாவில் தமிழ் இலக்கியம் பற்றிய செய்திகள் மிகவும் சிறப்பாக இடம் பெற்றுள்ளன. இருந்த போதிலும் செவ்வியல் நூல்களான தொல்காப்பியம், எட்டுத்தொகை, பத்துப்பாட்டு முதலிய 41 நூல்களின் பெயர்களை குறிப்பிட்டு, சிரியர் பெயர், பாடல் எண்ணிக்கை, கருத்து போன்றவற்றை உள்ளடக்கிய வரிசைப்பட்டியல் இடம் பெற்றிருத்தல் இன்னும் சிறப்புமிக்கதாகக் கருதப்படும்.

### எண்ணிக்கை நூல் பெயர் ஆசிரியர் பாடல்களின் எண்ணிக்கை ஆமையக்கருத்து

தொல்காப்பியம் தொல்காப்பியர் 1610 தமிழரின் ஒழுக்கம், பண்பாடு, வாழ்க்கை முதலியவற்றைப் பிரதிபலிக்கும் கருத்துக்கருவூலமாகும்.

1. இறையனார் அகப்பொருள் இறையனார் ----- தமிழரின் ஒழுக்கம், பண்பாடு, வாழ்க்கை முதலியவற்றைப் பிரதிபலிக்கும் கருத்துக் கருவூலமாகும்.

8. எட்டுத்தொகை ஆசிரியர் பலர் 2348 அகம் மற்றும் புற வாழ்க்கையை வெளிப்படுத்துவது 10. பத்துப்பாட்டு சிரியர் பலர் 3552 அகம் மற்றும் புற வாழ்க்கையை வெளிப்படுத்துவது

18. பதினெண்கீழ்கணக்கு ஆசிரியர் பலர் 3254 மக்களுக்கு தேவையான கருத்துக்களை வலியுருத்துவது.

1 சிலப்பதிகாரம் இளங்கோவடிகள் 5001 (அடிகள்) 'அரசியல் பிழைத்தோர்க்கு அறம் கூற்றாவதும் உரைசால் பத்தினிக்கு உயர்ந்தோர் ஏத்தலும் ஊழ்வினை உருத்து வந்து ஊட்டும்' என்பதை வலியுறுத்தல்.

1 மணிமேகலை சீத்தலைச்சாத்தனார் 4286(அடிகள்) ' உண்டி கொடுத்தோர் உயிர் கொடுத்தோரே'

### தமிழ் இலக்கியத்தின் தனித்தன்மைகள்:

ஒரு மொழியை செம்மொழியாக ஏற்றுக்கொள்ள வேண்டுமெனில், அதற்கு 11 தகுதிகள் இருக்கவேண்டும் என்று மொழி இயல் வல்லுனர்கள் வரையறை செய்துள்ளனர். அவை (1)தொன்மை, (2) தனித்தன்மை (3)பொதுமைப்பண்பு (4)நடுவு நிலைமை (5)தாய்மைத்தன்மை (6)பண்பாடு, கலை,பட்டறிவு வெளிப்பாடு (7) பிறமொழி கலப்பில்லா தனித்தன்மை (8) இலக்கிய வளம் (9) உயர் சிந்தனை (10)கலை, இலக்கியத் தனித்தன்மை வெளிப்பாடு (11)மொழி கோட்பாடு.

உலகில் பழம்பெரும் மொழிகளாக அடையாளம் காணப்பட்டுள்ள எந்த ஒரு மொழிக்கும் செம்மொழிக்குரிய இந்த 11 தகுதிகளும் முழுமையாக இல்லை. சமஸ்கிருதத்துக்கு 7 தகுதிகளும் இலத்தின் மற்றும் கிரேக்க மொழிகளுக்கு 8 தகுதிகளும் மட்டுமே உள்ளன என்பது அறிஞர்கள் கருத்து. நம் தமிழ் மொழிக்கு மட்டுமே செம்மொழிக்கான தகுதிகள் 11-ம் முழுமையாக உள்ளன. மேல் நாட்டு மொழியியல் வல்லுனர்கள் வகுத்தமொழித்தகுதிகள் நம்முடைய தமிழ் மொழிக்கு முழுவதுமாக ஒத்துப்போவது மிகப்பெரிய வரலாற்று உண்மையாகும்.

அறிவியல், புவியியல்,பண்பாடு, கணிதம், சமூகம், வரலாறு, தொழில்நுட்பம், நிர்வாகம் போன்ற பிறதுறைகள் தமிழ் இலக்கண, இலக்கியங்களில் பொதிந்துள்ளன.அவற்றை வெளிக்கொணருதல் மிகவும் சிறப்பானதாகும்.

எ.கா. எந்த மொழியிலும் இல்லாத கணித எண்கள் தமிழ் மொழியில் இடம் பெற்றுள்ளன.(க-1, உ- 2....) இந்த எண்ணுருக்களை கணிப்பொறியில் உருவாக்கினால் சிறப்பாக இருக்கும். தொல்காப்பியர் பெரும்பலான அடிகளில் எண்ணிக்கையை வெளிப்படுத்தியுள்ளார்.

'ஆறுதலையிட்ட அந்நால் ஐந்து' என்று செய்யுள் உருப்புகளை குறிப்பிடும்பொழுது,

நால், ஐந்து  $4 \times 5 = 20$

அறு தலையிட்ட அந்நால் ஐந்து- $6+4 \times 5 = 26$ .

பரிபாடலின் அளவினை எடுத்துரைக்கும்போது,

அபரிபாட்டெல்லை

நாலீர் ஐம்பது உயர்பு அடியாக

ஐ ஐந்து கும் இழிபு அடிக்கு எல்லைஅ

அதாவது  $4 \times 2 \times 50 = 400$  அடி பேரெல்லை.

$5 \times 5 = 25$  அடி சிற்றெல்லையாகவும் வரும்.

இவற்றின் மூலம் தமிழ் மொழியில் கணிதம் கலந்திருந்ததை அறியமுடிகிறது.

தாவரவியல் மற்றும் விலங்கியல் பற்றிய செய்தியை அன்றே தொல்காப்பியர்,

'புல்லும் மரனும் ஓர் அறிவினவே'

'நந்தும் முரளும் ஈர் அறிவினவே'

'மக்கள் தாமே ஆற்றிவு ஆயிரே

பிறவும் உளவே அக்கிளைப் பிறப்பே'

எனக் குறிப்பிட்டுள்ளார்.

### முடிவு:

தேமதுரத் தமிழோசை உலகமெல்லாம் பரவ வழி செய்திடல் வேண்டும் என்ற பாரதியின் கனவு, தன்னலமற்ற விக்கிபீடியா போன்ற நிறுவனத்தால் நனவானது. மேலும் தமிழ் சிறக்க, தமிழ் மாணவர்கள் 'நீங்களும் எழுதலாம்' பகுதியில் தரமான செய்திகளை பதிவு செய்து தமிழரின் சிறந்த பண்பாட்டை உலகறிய செய்யலாம்.



# **E-Commerce**

(மின் வணிகம்)



# E-Governance Activities in Tamil Nadu

**E.Iniya Nehru**

Senior Technical Director,

National Informatics Centre Tamil Nadu State Centre, Chennai

(E-mail:nehru@nic.in)

A number of G2C services are being rendered electronically to the citizens through a single window mechanism. The list includes different types of certificates such as Land Ownership Certificate, Community Certificate, Birth Certificate, Encumbrance Certificate and Nativity Certificate etc. along with other services such as Scholarship portals, permits, passes, licenses to name a few.

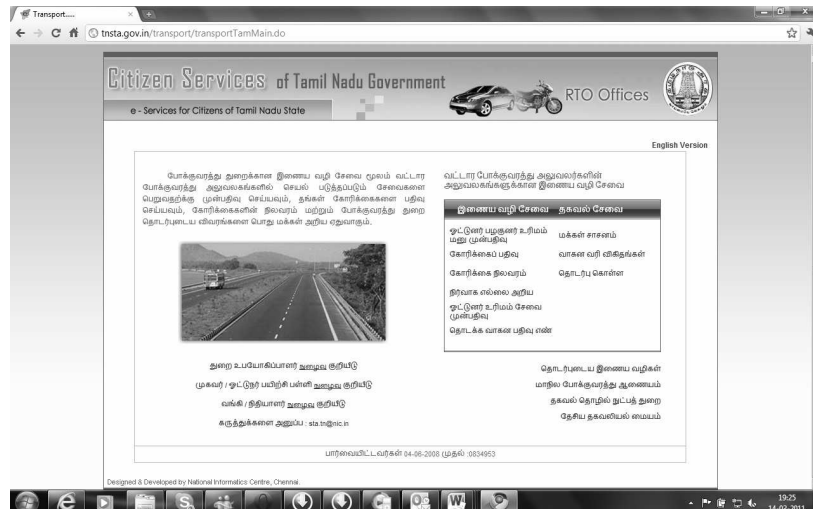
The Major E-Governance Projects implemented recently are:

## Tamil Nadu - E-Services of Transport Department:

A single portal which enables Citizens to file Learner's License application online, Register their Grievances, know the Status of their redressal, Appointments to visit RTOs and Know their RTOs has been implemented. It also provides the facility to Dealers to file the New Vehicle Registration applications online, generation of Heavy Vehicle Training Course attended certificate online, filing of applications by the Financier for endorsement of Hire Purchase agreement and hire purchase Termination online. 1577 Driving Schools and 1307 Dealers have already enrolled. More than 1,10,000 New Vehicle Registration applications, 60,000 Learner's License applications are filed through this system and 15,000 Heavy Vehicle Training Course attended certificates are being generated through this system every Month.

## Tamil Nadu - e-Services for Department of Commercial Taxes

To facilitate the Dealers of Commercial Taxes the Government of TamilNadu provides the anytime anywhere services like Online filing of VAT returns, Online payment of Taxes with 5 different Banks , Online submission of Form-W refund Claims, Online filing of e-Request for saleable forms, Fast Track Clearance system at Checkpost and Online submission of New Registration application.



The system also provides the facility to all the Citizens to search on the commodity code, Tax Rate, TIN number, VAT clarifications, GO's, Notifications and Circulars issued by the Government, contact Details, VAT Act , Rules and Auction details. It helps the Department to monitor the status of Returns filed by the Dealers and also to identify the Non-filers. More than 3,10,000 Dealers are enabled for filing e>Returns online. As of now, average of 2,30, 000 Dealers are filing their Returns online every month. More than 2 Crores of Sales and Purchase invoice data are being captured every month. More than Rs 1500 Crores of Taxes per Month are being paid through e-payment through the 5 Nationalized Banks.

**வணிகவரித் துறை**  
**Commercial Taxes**  
Department, Tamil Nadu

**Dealer Services**

- VAT Clarifications
- VAT Circulars
- GOs & Notifications
- Commodities Rates
- Commodities Search
- TIN Search
- Dealer Registration
- Refunds
- Auction
- Helpline
- Checkpost FTCS

**e-Payment**

**Tamil Nadu Value Added Tax**

Tamil Nadu Value Added Tax Act 2006 has come into effect from 1st January 2007.

VAT is a multi-stage tax on goods that is levied across various stages of production and supply with credit given for tax paid at each stage of Value addition.

VAT is the most progressive way of taxing consumption rather than business.

**General Information**

- Acts
- Rules
- Forms
- FAQs
- Citizen Charter
- RTI Act
- About Us
- Contacts
- Feedback

**Related Links**

- Tinxsy
- NIC Mail
- Tamil Nadu Govt. Site
- Traders Welfare Board

**What's New**

- CT Policy Note 2010-2011-Demand No.10 (English)
- CT Policy Note 2010-2011-Demand No.10 (Tamil)
- CST Return filing dealers mandated for e-

**FAQs on e>Returns?**  
Help File For Filing e>Returns

**To file e>Returns Choose Your Division**

- Coimbatore and Salem Division Dealers
- Chennai and Other Division dealers

Toll Free No : 1800 425 1959  
Call Centre No : 044 28290962

Total No. of Visitors (Since 31/12/2007): 15759643

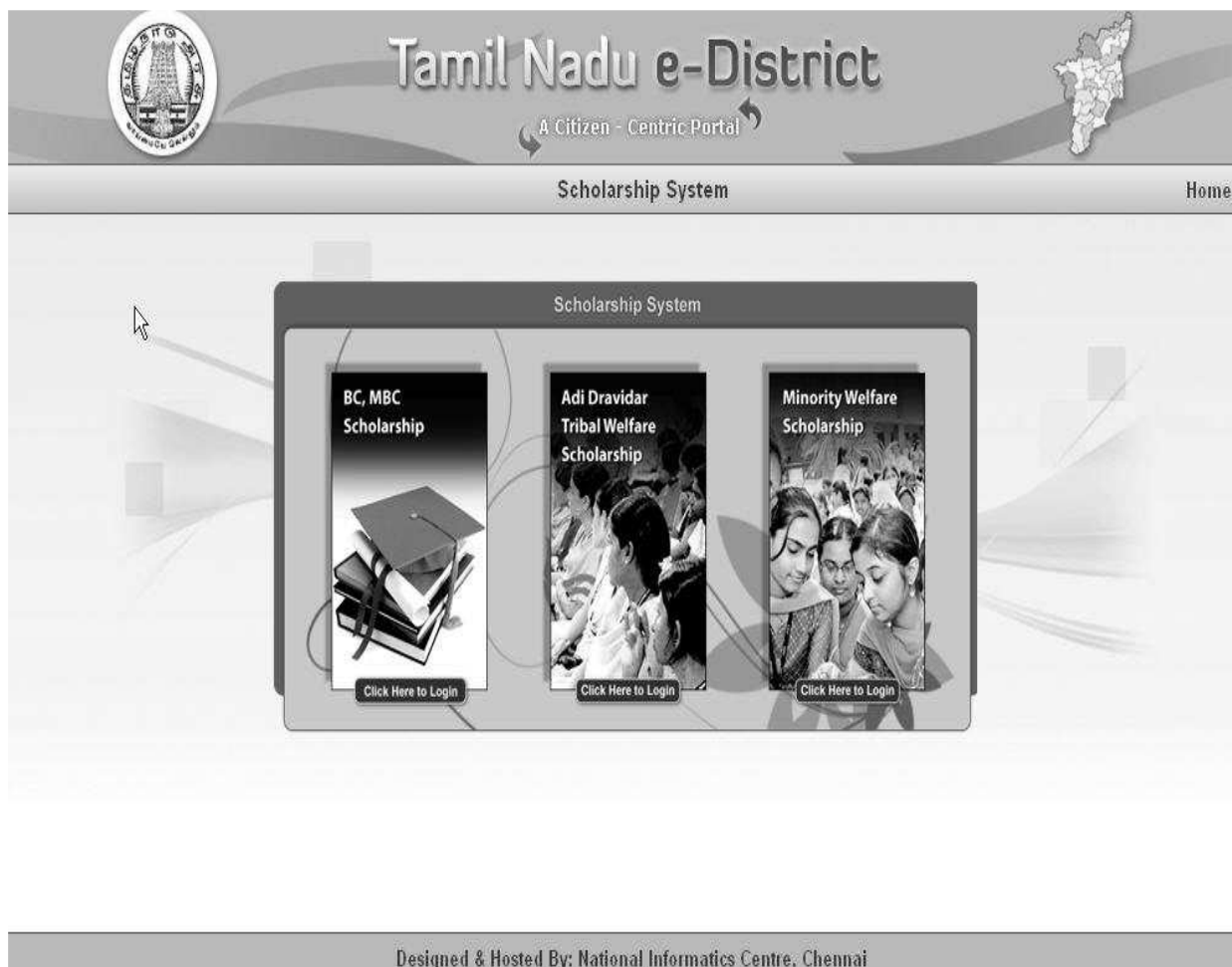
Commercial Taxes Department  
Ezhilagam, Chepauk, Chennai - 600 005.  
E-Mail: ccd@tn.nic.in

Designed, developed and hosted by:  
National Informatics Centre  
email: vatfeedback@tn.nic.in

## Tamil Nadu - eDistrict : Scholarship System

The Web based system implemented by the Government of TamilNadu, provides a facility for students to file application for scholarship online through their respective Institutions, as a first step towards bringing in transparency in the processing of the scholarship forms at various levels of the Government. The system has the necessary provision for requisite backoffice work flow for processing the application. **It facilitates quicker processing of scholarship applications of the student and also it provides the status of the scholarship application through the website/CSC/SMS** to the students. Automatic SMS messaging services is also sent to individual students who have provided the mobile number in the scholarship application. 1200 institutions dealing with the Scholarship of the BC/MBC and 54 institutions for SC/ST students are making use of

this facility in the state. About 3.7 lakh students for BC&MBC and 2.45 lakh students for Adi Dravidar scholarship have applied so far.



## **Tamil Nadu - Government e-Procurement System of NIC (GePNIC)**

The TamilNadu Government implemented the SKoch-Challenger 2000 and eIndia awarded GePNIC – the total automation of the process of physical tendering activity on internet in a faster, and secure environment adopting industry standard open technologies. It is highly generic in nature and can easily be adopted for all kinds of procurement activities such as Goods, Services & Works, by Government offices across the country. All the registered government departments on this application are double authenticated one with the Login-id and Password and other one with the Digital Signature Certificates and PIN for their ensured security transaction on the internet. The departments create Tender and Corrigendum and publish them on the site. Bidders bid against the eligible tender and bids are encrypted with the bid opener's public key. The same bids are opened only by the Bid openers DSC.

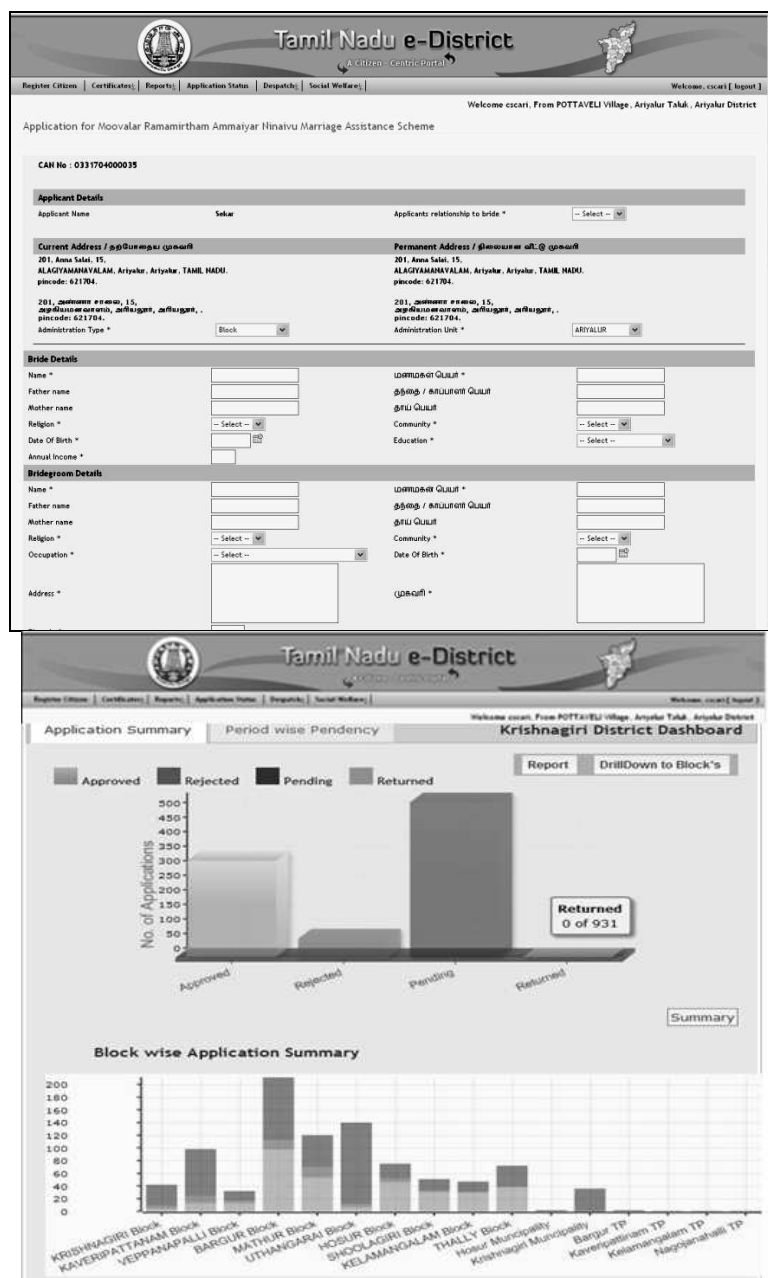
GePNIC has been implemented successfully in the State Governments of Orissa, Tamil Nadu, West Bengal, Uttar Pradesh, Haryana, Chandigarh UT Jharkhand, NICSI, Mahanadi Coalfields Limited (MCL) Orissa, PWD Punjab and Viskhapatnam Port Trust. It is also being implemented for

procurements under Pradhan Mantri Gram Sadak Yojana (PMGSY) of Rural Development Ministry in 21 states, covering the North Eastern states. Around 52,080 tenders, worth over Rs 85,089 Crores, have been processed successfully from 2008 to February 2011



## Tamil Nadu – eDistrict Project of Social Welfare Department


An initiative of Social Welfare department covering 215 Citizen Service Centres and 10 Block Offices of the Pilot district of Krishnagiri District to provide five different services of interest to the citizens. Services of Marriage Assistance to Widow Daughter, Orphan Girls, Widow Remarriage, Inter Caste and Child protection scheme are some of the services one can avail through the registered CSCs or facilitation centres at Block/District/Taluk. Being a workflow based application the transparency in processing the application for the marriage assistance is provided to the citizens who can verify the status of their submitted application using the ID number provided in the acknowledgement receipt of their application. Facility to provide the SMS messaging service to the citizens after the approval of their application is also implemented. More than 885 applications have been received to date and are in different stages of processing.




## Tamil Nadu - Pregnancy and Infant Cohort Monitoring & Evaluation System

An Online monitoring system that helps to monitor the health status of Pregnant Women registered with any PHC in the rural areas of Tamil Nadu has been successfully implemented from 2008 across 1500+ PHCs. All the 385 Block Medical Officers and 42 District Health Officers are making use of the system for the effective monitoring of the PHCs in monitoring the health conditions of the Pregnancy women and Infant cohort. More than 28 lakh records of Ante Natal checkup details are maintained.

The system captures the details of the pregnant women at various stages like ante-natal care, delivery, post-natal care etc. Similarly it captures details of infants like growth, immunization etc.



**Directorate of Public Health & Preventive Medicine**  
 Pregnancy and Infant Cohort Monitoring and Evaluation

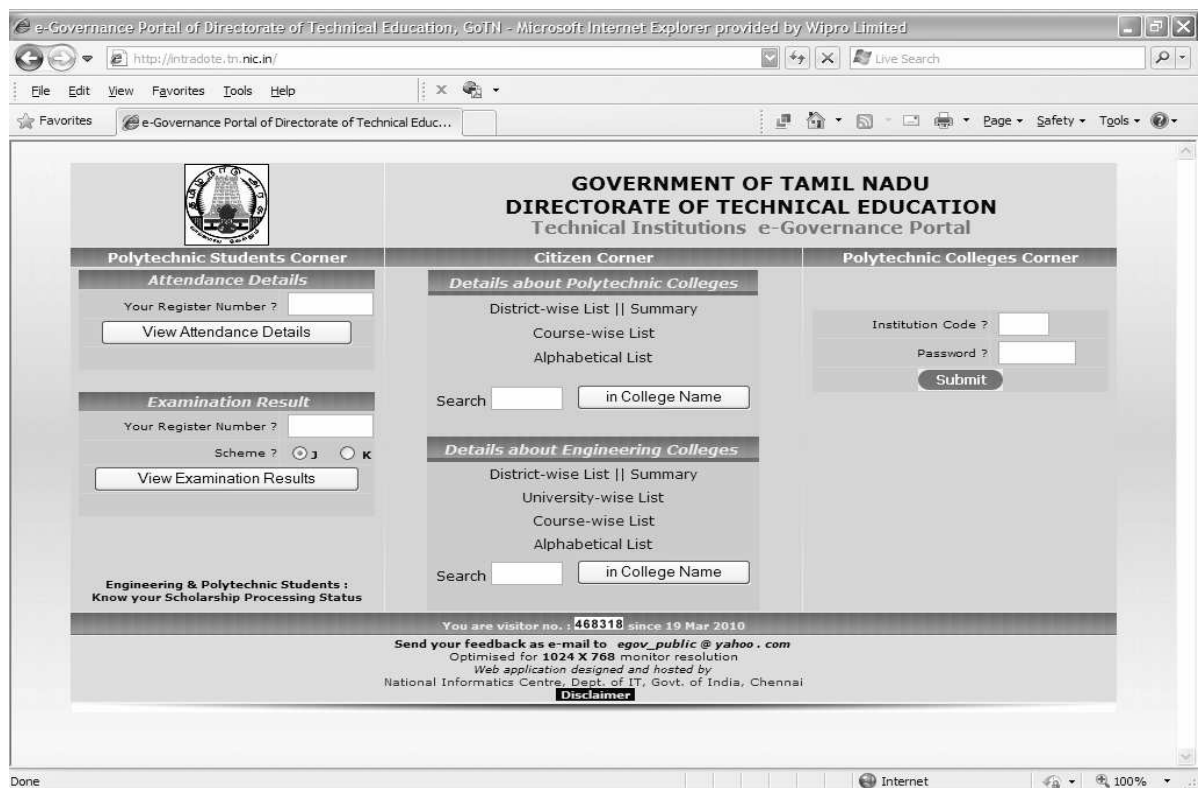


| PREGNANCY COHORT MONITORING SYSTEM       |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
|--|---------------|----------------|---------------------|--------------|------------|---------------------------------|----------|--------------------------|---------------|-------------|-----------|--------------|---------------------------|------------|----------------------------|------------------|--|
| AN Mother General Information            |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| Month & Year : August-2010               |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| PHC                                      |               |                | Pattarai Perumbudur |              |            |                                 |          | HSC                      |               |             |           | Kaivandur    |                           |            |                            |                  |  |
| Name of VHN                              |               |                | R Vajjyanthi        |              |            |                                 |          | VHN Phone No.            |               |             |           | 9445136927   |                           |            |                            |                  |  |
| AN Mother View - General Information     |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| ANC ID No.                               |               |                | 4009301155          |              |            |                                 |          | Para                     |               |             |           |              |                           |            |                            |                  |  |
| Name of AN mother                        |               |                | Parameswari         |              |            |                                 |          | LMP                      |               |             |           | 18/08/2009   |                           |            |                            |                  |  |
| Name of Husband                          |               |                | Silamban            |              |            |                                 |          | EDD                      |               |             |           | 25/05/2010   |                           |            |                            |                  |  |
| Phone                                    |               |                | 9787275635          |              |            |                                 |          | Date of AN Registration  |               |             |           | 13/10/2009   |                           |            |                            |                  |  |
| Address                                  |               |                | kupamal chataram    |              |            |                                 |          | Height ( in cm)          |               |             |           | 144          |                           |            |                            |                  |  |
| Age of mother                            |               |                | 25                  |              |            |                                 |          | Blood Group              |               |             |           | A+           |                           |            |                            |                  |  |
| Community                                |               |                | SC                  |              |            |                                 |          | VDRL                     |               |             |           | Non-reactive |                           |            |                            |                  |  |
| Education Status of AN Mother            |               |                | High(Std. 9-10      |              |            |                                 |          | HIV Status of AN Mother  |               |             |           | Negative     |                           |            |                            |                  |  |
| Education Status of AN Husband           |               |                | Hr.Sec(std. 11-12)  |              |            |                                 |          | HIV Status of AN Husband |               |             |           | Negative     |                           |            |                            |                  |  |
| Gravida                                  |               |                | 1                   |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| Remarks                                  |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| No.                                      | Date of visit | Place of visit | BP (mm Hg)          | Weight ( Kg) | Date of TT | Date of IFA                     | IFA Nos. | Date of Albendazole      | Urine Albumin | Urine sugar | Hb in gms | Blood sugar  | Height of uterus in weeks | FH Rate/mt | High risk factors detected | Treat ment given |  |
| Ultrasonogram done                       |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| Ultrasonogram done                       |               |                | Date                |              |            | Result                          |          |                          | Findings      |             |           |              |                           |            |                            |                  |  |
| 1st Time                                 |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| 2nd Time                                 |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| 3rd Time                                 |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| AN Referral                              |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| High Risk referred                       |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| If yes Specify the Complication          |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| Date & Time of Referral                  |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| Place of Referral                        |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| Compliance                               |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| Natal Referral, Delivery & Death Details |               |                |                     |              |            |                                 |          |                          |               |             |           |              |                           |            |                            |                  |  |
| High Risk referred                       |               |                | No                  |              |            | If yes Specify the Complication |          |                          |               |             |           |              |                           |            |                            |                  |  |

**Tamil Nadu - Technical Institutions e-Governance Portal for Directorate of Technical Education**

To enable the Citizens to get basic details of all the **430+ Polytechnic Colleges and 450+ Engineering Colleges** and to enable the Diploma Students studying in these Polytechnic Colleges to know the Attendance details and Semester Examination results for the current Academic Year, this web application was designed and hosted by the the DoTE of Government of Tamil Nadu. This portal has 3 different sections for the Citizen, for the Polytechnic Colleges and for the Student of Polytechnic Colleges to get to know the entire details of their interest of Polytechnic and Engineering colleges. Institutions and DoTE are provided with the staff profile and institution profile of all the institutions. More than 5.45 lakh of citizens visited the site from its launch in March 2010.





National Informatics Centre has been providing informatics support to Central Ministries, State Government and District Administration. E-Governance initiatives have already made a large impact in various sectors including agriculture, rural development, judiciary, health, education, transport and administration.

# New Media and Tamil - using softwares, tools and Technology

தமிழ் மற்றும் புதிய ஊடகம் - செயலிகள், மென்பொருள்கள், தொழில்நுட்பம் பயன்பாடு

**S. Gunasegaran**

*New Media Trainer*

*ACTA Trainer (Singapore WDA Approved), 2D animator/3D visual specialist*

*Temasek Polytechnic, Singapore*

## Abstract

Recent developments in new media and Tamil computing have changed the overview of creating Tamil contents and engaging applications especially using revolutionary iphone and ipad technology. The Internet has brought the world closer and helped creative industries to make realistic achievements in the e-learning environment. As with introduction of Unicode and technical capabilities to include Tamil as part of iPhone and iPad, the teaching and learning approaches are changing.

This paper looks into ways to create engaging contents, using digital music, animation and high definition audio and videos. It will showcase the usage of new learning environment using Adobe, Toon Boom, Second Life, Blackboard, Moodle and Eon Reality to engage with new Tamil around the globe. It will also examine the need to create more digital games and tools to address the learning needs of young learners of Tamils. It will focus on using blogging, wiki, chatting and tweeting to promote the language effectively.

## தொடக்கம்

காலத்தை வென்று நிற்கும் செம்மொழியான தமிழ்மொழி, இந்நூற்றாண்டின் விஞ்ஞான புரட்சியில் தனது பிம்பங்களைப் பிரதிபலிக்கத் தவறவில்லை. ஓலைச்சுவடிகளில் தொடங்கிய தமிழ் மொழியின் பரிணாமம், நொடிபொழுதில் நாம் வாழும் உலகை விரல் நுனியில் ஆளும் வல்லமையை மனித குலத்துக்கு வாரி வழங்கும் வல்லமை பெற்ற WWW(world wide web) எனும் மூன்றெழுத்துகள், நம் மொழியை தரம் மிக்க வாழும் மொழியாக வளர்வதற்கு வித்திட்டுள்ளது.

12 உயிரும் 18 மெய்யும் 1 ஆய்த எழுத்துக்களால் உருவான தேமதுரத்தமிழ், இணையத்தளத்தில் உலாவரும் மொழிகளில் ஒன்றாக உயர்வடைந்துள்ளது. சுமார் 30 வருடங்களுக்கு முன்பு சிங்கப்பூரில் கணிணி எனும் புதுச்சொல்லுக்கு அடையாளம் கண்ட அமரர் நா.கோவிந்தசாமி இட்ட அடித்தளம், தமிழகம், மலேசியா, இலங்கை என உலகம் முழுவதும் வியாபித்திருக்கும் தமிழர்களை ஒன்றிணைக்கும் பாலமாக உருவெடுத்துள்ளது.

20ம் நூற்றாண்டில் ஏற்பட்ட தொழில்புரட்சி, மனித வாழ்க்கையை ஏற்றம் மிக்கதாக மாற்றி உள்ளது. செய்திதாள், கடிதத் தொடர்பு, வானொலி, தொலைக்காட்சி, தொலைபேசி என உருவான அடிப்படை ஊடக வசதிகள், இணைய வசதிகள் வந்த பிறகு, அன்றாட வாழ்க்கை முறைகளை வழிநடத்தும் கூறுகளாக மாறி உள்ளன. முகநூல் (facebook), குறுந்தகவல்(SMS), you tube, twitter, e-mail என இணையத்தின் அனைத்து பிரிவுகளிலும் தமிழை பயன்படுத்தும் வாய்ப்புக்களை ஏற்பட்டுள்ளது. தகவல் பரிமாற்றங்கள் உடனுக்குடன் நடப்பதால் தமிழை பயன்படுத்தும் தேவைகள் அதிகரித்துள்ளன.

புதிய ஊடகங்கள், ஆர்வம் உள்ள அனைவரும் தமது கருத்துக்களையும், படைப்புக்களையும் உருவாக்க, மாற்ற, பகிர்ந்து கொள்ள வாய்ப்பளிக்கிறது. கணினி மற்றும் கைத்தொலைபேசி மூலம், அதிக பொருள் செலவில்லாமல், உலகின் எந்த மூலையில் இருந்தும் ஒருவர் மின்னூலகில் உலா வரலாம்.

podcasts, RSS feeds, social media, text messaging, blogs, wikis, virtual world, என தினந்தோறும் உருவாகி வரும் புதுப்புதுத் தொழில்நுட்ப முறைகளை பயன்படுத்தி, செய்திகளை, சேவைகளை பரிமாறிக்கொள்ள புதிய ஊடகம் வாய்ப்பு ஏற்படுத்தித் தந்துள்ளது.தொலைந்து போன உறவுகளை, மனித நேயப் பண்புகளை வளப்படுத்திக்கொள்வதற்கும்,ஒருமித்த கருத்துக்களைக் கொண்டவர் களோடு தொடர்பு ஏற்படுத்திக் கொள்வதற்கும் ,புத்தாக்க செயல்கள், சேவைகள், சமூகங்களை உருவாக்கி, ஒருவர் மற்றவர் தெரிந்து பயனடையக் கூடிய வழிமுறைகளை உருவாக்கவும் புதிய ஊடகம் உதவியாக உள்ளது.

தமிழ் வளர்ச்சியில் அண்மைய காலங்களில் பெரும் பங்காற்றி வருகிறது. அதிகமான வலைபூக்கள் (blogs) தமிழில் எழுதப்படுகின்றன.சுதந்தரமாக கருத்துப்பரிமாற்றம் செய்வதற்கு இம்முறை பயன்படுகிறது.

கல்வித்துறையில் தமிழ் கற்பிக்க சிங்கப்பூர்,மலேசியா நாடுகளில் E-Learning portals ஏற்படுத்தப் பட்டுள்ளன. இலங்கை மற்றும் வெளிநாடுகளில் வாழும் பல தமிழர்களை இணைக்கும் வகையில் பல இணையத் தளங்கள் செயல் பட்டு வருகின்றன. தமிழக அரசின் செயல் முறைகள் தமிழ் கணினி கட்டமைப்பில் உருவாக்கப்பட்டுள்ளது.Tamil

Virtual University எனும் இணையத்தளம் வழி தமிழ் இலக்கியங்களையும் மொழி சார்ந்த செய்திகளையும் கணினியில் கண்டு கேட்டு படித்து பயனடையலாம்.

செயலிகள், மென்பொருள்கள். தொழில்நுட்பம் மென்பொருள்கள் adobe Flash CS5.5 iPad .iphone மென்பொருள் உருவாக்கத்தில், dynamic font முறை தற்போது இடம்பெற்றுள்ளதால், unicode முறையில் அமைந்த எழுத்துருக்கள் தெளிவாகத் தெரிகின்றன. தொட்டுணர்தல், இருவழிப் பயனீடு போன்ற புதிய முறைகளைப் பயன்படுத்தி கணினி விளையாட்டுகள், கல்வி சார்ந்த மென்பொருள்கள் உருவாக்குவதற்கு வழி அமைத்துள்ளன.

## TamiliBooks and app in Iphone/iPad

கடந்த சில ஆண்டுகளில் மடிக்கணினி துறையில், ஏற்பட்டிருக்கும், தொழில் நுட்பப் புரட்சி, ஒவ்வொரு மனிதனின் வாழ்க்கைமுறைகளையும் அடியோடும் மாற்றி விட்டது.apple computers அறிமுகம் செய்த Iphone/ iPad ல் வழங்கப்பட்டிருக்கும் கட்டுப்படுத்தப்பட்ட மென்பொருள் உருவாக்கத்தைப் பயன்படுத்தி, சிலநூறு மென்பொருள்கள் இணைய தளத்தில் பதிவிறக்கம் செய்வதற்கு வழங்கப்படுகின்றன.

Certified Developer என்ற முறையில் ஒரு iBook, தமிழ் விளையாட்டுகள்,அனைவரும் பேச தமிழ் எனும் செயலியையும் உருவாக்கி உள்ளேன்.தற்போது, சிங்கப்பூர் அரசிந் கலாசார மன்ற ஆதரவில், அனைத்து இன மக்களும் பங்குபெரும் வகையில் தமிழ் மொழியும் இசையும் கலந்து Iphone/iPad மட்டுமே பயன்படுத்தி நடத்தப்படும் இசை நிகழ்ச்சியை நடத்தி வருகிறேன்.நமது மொழி, இசை பற்றி,அனைவரும் தெரிந்து கொள்வதற்கு இது போன்ற புத்தாக்க அங்கங்கள் பேருதவி புரிகின்றன.

## Microsoft Office in Window7

கணினியில் மட்டும் இன்றி,புதிதாக அரிமுகம் செய்யப்பட்டுள்ள Window7 கைத்தொலைபேசிகளில் ,இலவச SDK மூலம் பல செயலிகளை உருவாக்கும் வாய்ப்பு ஏற்பட்டுள்ளது. Open source முறையில் பதிவேற்றம் செய்வதால், கருத்து சுதந்திரத்துடன் செயல் பட முடிகிறது.

## Google Website

இங்கு மின்னஞ்சல், குந்துள்ளனந்தகவல் அனுப்பும் முறைகளில் நேரடியாகத் தமிழைப் பயன்படுத்தும் முறை அறிமுகம் செய்யப்பட்டுள்ளது. Romanised வடிவில் அச்சடிக்கப்படும் வார்த்தைகள் தமிழில் இடம்பெரும் கூறு இதில் இடம்பெற்றுள்ளது.

## Toon Boom animation

கேளிச்சித்திரம் மூலம் வரைகலைப் படங்களை உருவாக்கி,தமிழில் உரையாடல்,நடிப்பு போன்ற அம்சங்களை இணைத்து தமிழை சிருவர் முதல் பெரியவர் வரை சுவைக்கும் வகையில் வெளிக்கொணர இந்த மென்பொருள் வாய்ப்பளிக்கிறது.

## Eon Reality

முப்பரிணாம தொட்டு உணர்தல்(3d virtual reality) முறைகளைப் பயன் படுத்தி. பலதரப்பட்ட காட்சிகளையும்,கற்றல் அனுபவங்களையும் இந்த மென்பொருள் மூலம் நாம் செயல்முறையில் பயன்படுத்தி புதுமையான அனுபவத்தைப் பெறலாம்.

## podcasting

இணையத்தில் ஒலி,ஒளி வடிவில் நேரடி நிகழ்ச்சிகளையும் படைப்புக்களையும் உருவாக்குவதன் மூலம், தமிழில் அனைவரும் பேசுவதற்கும் உரையாடுவதற்கும் வழிமுறைகள் ஏற்பட்டுள்ளன. live streaming முறையில் தரமான தமிழ் வானொலி, தொலைக்காட்சி செயல்பாடுகளை இல்ல அறையில் இருந்து நடத்தலாம்

## New media and Tamil

தொலைதூரக் கல்வி,வாழ்நாள் தொடர்பயிற்சி, போன்ற முறைகளில் கற்பித்தல் முறைகளை கையாள்வதற்கு ,புதிய ஊடகங்கள் கைகொடுக்கின்றன. moodle, Blackboard LMS போன்ற இணையம் வழி கற்பிக்கும் கணினி கட்டமைப்பில்,தமிழில் பாடங்களை நடத்த முடியும்.கற்றல்,கற்பித்தல் வழிகளில், மாணவர்களும், பயிற்றுவிப்பாளர்களும்,அறிஞர்களும் ஒருசேர பெரும் அளவில் பயன் பெற முடியும்.

## Virtual Classrooms

சிங்கப்பூரில்,வாழ்நாள் கல்விக்கும்,தொலைதூரக் கல்வி, தொடர்பயிற்சி, போன்ற முறைகளில் கற்பித்தல் முறைகளை கையாள்வதற்கு ,புதிய ஊடகங்கள் கைகொடுக்கின்றன. Second Life, Elluminate, Adobe Connect Pro, you tube, Skype, slide share, google apps, animoto, voki animated Characters,போன்ற இணையம் வழி கற்பிக்கும் முறைகளை பயன்படுத்துவதன் மூலம் நேரத்தை மிச்சப்படுத்தி கற்பிக்கும் பாணியை,அனைவரும் உணரும் வண்ணம் படைக்க முடிகிறது. பெரும்பாலும்,இலவசமாக மென்பொருள்கள் இணையத்தில் இருந்து பதிவிறக்கம்

செய்ய இயலுவதால்,புது தகவல்களையும்,பாடத்திட்டங்களையும் அறிமுகம் செய்வதற்கு இவை பேருதவியாக இருக்கின்றன.பரபரப்பான வாழ்க்கையில்,சிறந்த கல்வியை வழங்குவதற்கு இணைய

வகுப்புகள் நல்ல தீர்வாக அமைகிறது.இது போன்ற புத்தாக்க நடவடிக்கைகளை ஊக்குவிக்க சிங்கப்பூர் அரசாங்கம் தமிழ்மொழி வளர்ச்சிக்கு 1.5 மில்லியன் வெள்ளியை செலவிடத் திட்டமிட்டுள்ளது.இது பல நல்ல திட்டங்களை தீட்டுவதற்கு வாய்ப்பாக அமைந்துள்ளது

### Social network sites

My space, facebook, றுtwitter,frienster,Orkut,bebo,wordpress,blogger,live spaces,yahoo,live journal,blackplanet,myyearbook,freewebs,Typepad,Xanga,multiply போன்ற புத்தாக்கம் மிக்க செயலிகளும்,மின்வலைத்தளங்களும் கல்வியை சுவைபட வழங்குவதற்கு பெரிதும் உதவியாய் அமைந்துள்ளன.Cloud computing முறையில், அதிக செலவில்லாமல்,நல்ல மென்பொருள்கள் பயன்படுத்த கிடைப்பதால்,தமிழ் சார்ந்த எண்ணங்களையும்,தொடர் வளர்ச்சி திட்டங்களையும் செயல்படுத்துவதற்கு நல்ல வாய்ப்புகள் நிறையவே அமைந்து தந்துள்ளன.

### Constrains and Remedies / இடர்பாடுகள், தீர்வுகள்

கணினித் துறையை பொருத்த வரை Microsoft Windows,Apple Mac OS என 2 பெரிய நிருவனங்களின் கணினி கட்டமைப்புக்களைச் சார்ந்து தமிழ் மென்பொருட்கள் பெரும்பாலும் உருவாக்கப் படுகின்றன.அப்பிள் கணினியில் இதுவரை ஒருசில மென்பொருள்களே தமிழில் வெளிவந்துள்ளன. iPad .iphone வெளியாகி 2 வருடங்கள் கடந்த பிறகும்,தமிழ் செயலியும், தட்டச்சு முறையும் வெளிப்படையாக அனைவரும் பயன்படுத்தும் வகையில் இணைக்கப் பாடாததால், இதுநாள் வரை ஒருசில எழுத்துருக்களை மட்டுமே பயன்படுத்த வேண்டிய கட்டாய நிலை. இதனால் பயன்மிக்க மென்பொருள்களை உருவாக்க முடியவில்லை. இந்நிலை மாற வேண்டும்.

பலநூறு மென்பொருள்கள் விற்பனைக்கு கிடைத்தாலும் அவை பெரும்பாலும் தமிழகம் சார்ந்து மட்டுமே இருப்பதால் சிங்கப்பூர் போன்ற வளரும் நாடுகளில், அன்றாட வாழ்க்கைச்சூழலில்,கல்வித்துறையில் பயன்படுத்துவதற்கும் பல்வேறு தொழில்நுட்பத் தடைகளைக் எதிர்கொள்ள வேண்டி இருக்கிறது.எழுத்துருவில் யூனிகோட் முறை வந்த பிறகு ஒரு சில மென்பொருள்கள் இலவசமாக கிடைத்தாலும், பொருளாதர அடிப்படையில் இதர மொழிகளுக்கு ஈடாக மென்பொருள்களை உருவாக்கி வெற்றி பெருவது பெரும்பாலும் எட்டாத கனியாகவே உள்ளது.

தமிழில் கணினி விளையாட்டுகள் உருவாக்கப் படுவது அரிதாக உள்ளது. Microsoft,Sony, Nitendo போன்ற நிருவனங்கள் மூலம் தரமான தமிழ் விளையாட்டு மென்பொருள்களை உருவாக்க கணினி வல்லுனர்கள் முயல் வேண்டும்.இன்றைய இளம் தமிழர்களை கவர்ந்திழுக்க இந்த முறை பயனளிக்கும்.

Tamil in the future / எதிர்காலத்தில் தமிழ் புதிய ஊடகங்கள் தற்போது தமிழுக்கு அணிசேர்க்கும் வகையில் சிறப்பான மாற்றங்கள் கண்டு வருகின்றன.கடந்த காலத்தில் எத்தனையோ மாற்றங்களை,சீர்திருத்தங்களை உள்வாங்கி செம்மொழியாய் உயர்ந்திருக்கும் தமிழ்,எதிர்காலத்தில் வாழும் மொழியாக உலகோடு ஒன்றித்து இருப்பதற்கு,புதிய ஊடகங்களும்,இணைய தகவல் தொழில் நுட்ப வளர்ச்சியும் இன்றி அமையாதவை.இவற்றை பயன்படுத்துவதன் மூலம்,தமிழின் ஓசையும்,மொழி நடையும்,எதிர்காலத்தில்,மேலும் ஏற்றம் கண்டு,பிரபஞ்சத்தின் கடைசி எல்லை இருக்கும்வரை புதுப்பொலிவுடன் வாழும் மொழியாக வீற்றிருக்கச் செய்யலாம். இணைய உலகில் தமிழுக்கென தனியிடம் என்றும் உண்டு என ஆணிதரமாக நம்பலாம்.

### Conclusion / முடிவுரை

தமிழ் நமக்கு வாழ்வியலை கற்றுத்தரும் உயரிய மொழியாக உலக மக்கள் பாராட்டும் இலக்கியங்களும்,காப்பியங்களும் நிறைந்த அறிவுச்சுரங்கமாக,இயல் இசை நாடகம் எனும் முப்பாலும்

ததும்பும் தெய்வீக மொழியாக என்றேன்றும் வாழ வைப்பது தமிழை சுவாசிக்கும் ஒவ்வொரு தமிழன் கையிலும் உள்ளது.புதிய ஊடகங்கள் மூலம், எழுத்தாலும்,இசையாலும் ,கணினி செயலிகளின் படைப்பாலும் நாம் அனைவரும் தமிழ் வளர்ச்சிக்கு உருதுணை புரியலாம்.

இப்பிறவியில்,சிங்கப்பூர் தமிழனாக பிறந்து,இங்கு அமெரிக்க மண்ணில் புதிய ஊடகம் மூலம் தமிழை பெருமைப் படுத்தும்,கட்டுரையை,10வது தமிழ் இணைய மாநாட்டில் படைக்க எனக்கு வாய்ப்பளித்த அனைவருக்கும் மனமார்ந்த நன்றி!

வாழிய தமிழ் மொழி! வெல்க நம் தாய் மொழி !

கணினித் தமிழாய் பார்புகழ் செம்மொழியாய்

வெல்க நம் தாய் மொழி !

## References

- Websites
- <http://tamilelibrary.org/teli/tlinks4.html>
- <http://namnaadi.edumall.sg>
- <http://sangamam.edumall.sg>
- <http://www.singai-tamil.org/main/index.html>
- <http://www.tamilvu.org/>
- <http://spp.moe.edu.my>
- <http://www.pazhahutamil.com/login/>
- <http://blangahrisetamil.blogspot.com/>
- <http://kidsone.in/tamil/>
- <http://www.sangapalagai.com/>
- <http://ta.wikipedia.org/wiki>
- <http://en.wikipedia.org/wiki/Tamil>
- [http://www.eonreality.com/news\\_releases.php?ref=news/news\\_releases&sid=475](http://www.eonreality.com/news_releases.php?ref=news/news_releases&sid=475)
- <http://beta.toonboom.com/>
- [www.microsoft.com](http://www.microsoft.com)
- [www.apple.com](http://www.apple.com)
- <http://developer.apple.com/>

## Reference Books

- Salmon, Gilly(2002) E-tivities: The Key to Active Online Learning, London, Kogan Page
- Brown, Sally, Bull,Joanna and Race.Phil(1999) Computer-Assited Assesment in Higher Education,London,Kogan Page

- Paul Chin, Using C&IT to Support Teaching
- Slater, Paul and Varney-Burch, Srah(2001) Multimedia in Language Learning, London: Language for Information on Language Teaching and Research
- Susan Ko, Steve Rossen : Teaching Online -A Practical Guide, 2nd Edition
- Reynol Junco & Jeanna Mastrodicasa : Connecting to the net.generation
- Les Lloyd, Editor : Technology and Teaching

# தமிழில் கணினிப் பாவனை

சிவா அனுராஜ்

ஆசிரியர் - கம்ப்யூட்டர் ருடே

(இலங்கையின் முதலாவது தேசிய தமிழ் கணினிச் சஞ்சிகை)

Email: tamilambu@yahoo.com

கணினி மற்றும் இணையம் போன்றவற்றின் பாவனையில் தமிழ் மென்பொருள்களின் இருப்பும் பாவனையும் தொடர்பாக 'தமிழில் கணினிப் பாவனை' என்ற தலைப்பிலான இந்த கட்டுரை ஆராயவிருக்கிறது. கடந்த 9 ஆண்டுகளாக புலத்திலும் மற்றும் உலகின் பல பகுதிகளில் பரந்துவாழும் தமிழ் மக்களிடையே தமிழ் கணினி தொடர்பான விழிப்புணர்வு நடவடிக்கைகள் பலவற்றினை மேற்கொண்ட அனுபவத்தினையும் 2000ஆம் ஆண்டுமுதல் இலங்கையில் இருந்து வெளிவரும் தமிழ் தகவல் தொழில்நுட்பச் சஞ்சிகையொன்றின் (நாட்டில் நிலவிய போர்ச்சூழல் காரணமாக சில காலம் தடைப்பட்டிருந்தது.) ஆசிரியராக இருந்து பணியாற்றிய அனுபவங்களின் அடிப்படையிலும் கணினி, இணையம் போன்றவற்றின் தமிழ் பாவனையாளர்களுடன் எனக்கு இருந்துவந்த நெருக்கமான தொடர்பும் இந்தக் கட்டுரையை வரைவதற்கு தூண்டுதலாக அமைந்ததுடன் இங்கே ஆராயும் விடயங்களின் உண்மைத்தன்மையையும் நியாயப்படுத்தும்.

கணினிப் பாவனை என்பது இன்றைய நிலையில் அனைவருக்கும் இன்றியமையாத ஒன்றாக இருக்கின்றது. இன்றைய காலகட்டமானது கணினியுக்ம் என்று அழைக்கக்கூடிய அளவிற்கு கணினிப் பாவனையும் கணினியின் தேவையும் எமது வாழ்க்கையில் ஒன்றாக மாறிப்போயுள்ளது. மனித வாழ்க்கையின் அன்றாட அடிப்படைத் தேவைகளில் ஆரம்பித்து அனு ஆராய்ச்சிவரை எல்லாமே இன்று கணினியை நம்பியே நடந்துவருகிறது. வேலைவாய்ப்பு, வேலைமுன்னேற்றம் என்பதில் கணினி அறிவு சிறப்புத்தகுதியாக இருந்த காலம் போய் கட்டாய தகுதியாக மாறிப்போயுள்ளது. இவை எல்லாவற்றையும் தாண்டி இன்றைக்கு சாதாரண மனிதர்கள்கூட கணினி அறிவு இல்லாமல் தமது அடிப்படைத் தேவைகளைக்கூட நிறைவுசெய்யமுடியாத நிலைக்கு தள்ளப்பட்டுள்ளனர்.

அடுத்ததாக, ஒரு மொழியின் இருப்பு, வளர்ச்சி ஆகியவற்றிலும் கணினி மற்றும் இணையத்தின் தாக்கம் மிகவும் அதிகமாகவே காணப்படுகிறது. அந்தவகையில் கணினியில் உள்ளீடு செய்யமுடியாத அழிந்து போகும் நிலையையே எதிர்நோக்கியுள்ளன. அதுமட்டுமல்லாமல் சரியான முறையில் கணினிமயப் படித்தப்படாத மொழிகள் அடுத்த சந்ததியினரிடம் போய்ச்சேராத நிலையும் ஏற்பட்டுள்ளது,

உண்மை மற்றும் யதார்த்த நிலைமை இப்படி இருக்க எமது தமிழ் பேசும் மக்களிடையே கணினிப் பாவனை குறிப்பாக தமிழில் கணினிப்பாவனை என்பது தொடர்பாகவே இந்தக் கட்டுரை ஆராய இருக்கிறது.

தமிழ் மக்கள் என குறிப்பிடும் பொழுது நாம் இரண்டு பிரிவினரைக் கருதவேண்டும். முதலாம் வகையினர் புலத்தில் உள்ள மக்கள். அதாவது பிரதானமாக இலங்கைமற்றும் இந்தியாவில் உள்ள தமிழர்கள். இரண்டாவது புலம்பெயர்ந்து உலகெங்கும் பரந்துவாழும் மக்கள் - புலம்பெயர் தமிழர்கள். இந்த இரு பிரிவினருக்குமான தேவைகள், பாவனைமுறை மற்றும் அவர்களது கணினி அறிவு என்பன வேறுபட்டிருக்கும். அதிலும் முக்கியமாக புலத்தில் உள்ள தமிழர்களிடத்தில் கணினிப்பாவனை குறைவாகவும் தமிழ்ப்பாவனை அதிகமாகவும் காணப்படுகிறது. புலம்பெயர் தமிழர்களை எடுத்தால் கணினிப்பாவனை அதிகமாகவும் தமிழ்ப்பாவனை குறைவாகவும் என தலைகீழாக உள்ளது. எனவே இந்த இரு பிரிவினர் தொடர்பாகவும் ஆராயவேண்டிய தேவை உள்ளது.



அந்த வகையிலே எனது கட்டுரையானது கீழ்வரும் பிரதானமான விடயங்கள் தொடர்பாக ஆராய இருக்கிறது.

1. மக்களிற்கு தமிழிலான மென்பொருட்களின் தேவை
2. தற்போது உள்ள மென்பொருட்கள்
3. தற்போது உள்ள மென்பொருட்கள் மக்களினை சென்றடைந்த வீதம்.
4. அவற்றினை தேவையுள்ள மக்களிற்கு கொண்டுசேர்க்கும் பொறிமுறை.
5. புதிய மென்பொருட்களின் உருவாக்கம்.

### **மக்களிற்கு தமிழிலான மென்பொருட்களின் தேவை**

தமிழ் மென்பொருட்களை பொறுத்தவரையில் மக்களின் உண்மையான தேவைகள் என்று பார்க்கும் பொழுது பிரதானமாக அவர்கள் சார்ந்துள்ள பிரதேசத்தைப்பொறுத்து இரண்டு வகைப்படும். புலத்தில் உள்ள மக்கள், அதாவது இலங்கை மற்றும் இந்தியாவில் உள்ள மக்களைப் பொறுத்தளவில் தமிழில் பாவிக்கும் மென்பொருட்களின் தேவையே அதிகமாகும். அதாவது தமிழில் கணினி. ஏனெனில் ஆங்கிலத்தில் உள்ள ஏனைய மென்பொருட்களை பாவிப்பதில் அவர்களின் மொழி அறிவு தற்பொழுது பெரும் தடையாக இருக்கிறது. ஆனால் புலம்பெயர் மக்களை பொறுத்தளவில் தமிழினைப்பாவிக்கும் மென்பொருட்களே பிரதான தேவையாக இருக்கின்றது. அதாவது கணினியில் தமிழ். ஏனெனில், அவர்கள் புலம்பெயர்ந்து வாழும் பெரும்பாலான நாடுகளில் அந்தந்த நாட்டு மொழிகளிலேயே (ஆங்கிலம் உட்பட) மென்பொருட்கள் கிடைப்பதால் அவர்களிற்கு மொழி ஒரு பிரச்சினை இல்லை. ஆனால் தமிழ் என்று வரும்போது அதனை உள்ளீடு செய்தல் மற்றும் அமிழினை கற்றல்புலம்பெயர்ந்து வாழும் எமது அடுத்த சந்ததிக்கு தமிழினை எடுத்துச்செல்லல் போன்றவற்றிற்குத் தேவைப்படுகிறது.

### **தற்பொழுது உள்ள தமிழ் மென்பொருட்கள்**

தற்பொழுது பாவனையில் பல தமிழ் மென்பொருட்கள் உள்ளன. இவற்றில் தமிழினை பாவிப்பதற்கான ஒருதொகுதி மென்பொருட்களும் தமிழில் பாவிப்பதற்கான ஒரு தொகுதி மென்பொருட்களும் அடங்கும். தமிழினை பாவிக்கும் மென்பொருட்களில் தமிழில் தட்டச்சு செய்வது, மின்னஞ்சல் அனுப்புவது, இணைய அரட்டையின்போது தமிழ் எனவும் தமிழ் சொல்திருத்தி, கையெழுத்து உணரி, தமிழ் அச்செழுத்து உணரி எனவும் பலவகையானவை உள்ளன. அடுத்து, ஏனைய அனைத்து மென்பொருட்களும் தமிழ் இடைமுகப்புடன் வரும்பொழுது அவை தமிழில் அமைந்த மென்பொருட்கள் என கருதப்படும்.

அதேவேளை தற்போது பாவனையில் உள்ள இவ்வாறான தமிழ் மென்பொருட்கள் மக்களின் தற்போதைய தேவையை பூர்த்திசெய்கிறதா என்றால், இல்லை என்பதே அதற்கான பதிலாக கிடைக்கும். இப்பொழுது உள்ள மென்பொருட்களில் பல பரிசோதனை நிலையிலும், மக்களின் உண்மையான தேவையை பூர்த்திசெய்வதாக இல்லாமலுமே காணப்படுகின்றன.

### **தற்போது உள்ள மென்பொருட்கள் மக்களினை சென்றடைந்த வீதம்.**

தற்பொழுது பல விதமான சாதாரண மற்றும் உயர் பாவனைத்திறன் கொண்ட தமிழ் மென்பொருட்கள் உள்ளபோதிலும் அவை பெரும்பாலான மக்களிடன் சென்றடையவில்லை என்பதே உண்மை. இதற்கு உதாரணமாக, சாதாரணமான தமிழ் தட்டச்சுக்கு உதவும் மென்பொருட்கூட பெரும்பாலான தமிழ் கணினி மற்றும் இணையப் பாவனையாளர்களுக்கு தெரிந்திருக்கவில்லை என்பதையே குறிப்பிடலாம்.

**இவ்வாறான மென்பொருள்கள் சரியான முறையில் பாவனையாளர்களிடம் சென்றடையாமைக்கான காரணங்களாக,** எமது தமிழ் மொழியானது பல பெருமைகளுக்கு உரிய மொழி. தமிழினை ஒரு மொழியாக மட்டும் கருதுவதோடு நிற்காமல் எமது கவுரவமாகவும் அதை பார்க்கிறோம். எனவே தமிழ் மொழியில் நாம் உருவாக்கும் மென்பொருள்களினை வர்த்தகரீதியாக பார்க்காமல் மொழிக்கு ஆற்றும் ஒரு சேவையாகவே கருதப்படுகிறது. இதனால் அந்த மென்பொருள்களினை உருவாக்குவதுடன் தமது கடமை முடிந்துவிடுவதாக பலர் கருதுகின்றனர். மக்களுள் இவ்வாறான மென்பொருள்களினை காட்சிப்பொருள்களாக பார்க்கிறார்களேயன்றி பாவனைக்கானதாக உணரவில்லை.

### **அவற்றினை தேவையுள்ள மக்களிற்கு கொண்டுசேர்க்கும் பொறிமுறை**

இந்த தமிழ் மென்பொருள்களை மக்களிடம் கொண்டுசேர்க்க வேண்டுமானால் முதலாவதாக அவற்றினை சரியானமுறையில் வரிசைப்படுத்தி அனைவரும் தெரிந்துகொள்ளும்வகையில் வைக்க வேண்டும். மேலும் மக்களின் தேவைகள் அறிந்து அவற்றினை பூர்த்திசெய்யக்கூடியான மென்பொருள்களினை உருவாக்கவேண்டும். அதுமட்டுமல்லாமல், கணினிப்பாவனை மற்றும் தமிழ் மென்பொருள்கள் தொடர்பாக மக்கள் விழிப்புணர்வினை ஏற்படுத்தவேண்டும். இவை எல்லாவற்றிற்கும் மேலாக, தற்பொழுதுள்ள மென்பொருள்களின் பாவனையாளர்களிடம் அவைதொடர்பான சரியான பின்னூட்டங்களை பெற்று பாவனையிலுள்ள மென்பொருள்களினை உரிய முறையில் மேம்படுத்துவதும் ஒரு பயனுள்ள செயற்பாடு.

### **புதிய மென்பொருட்களின் உருவாக்கம்**

புலத்தில் உள்ள தமிழர் (இலங்கை, இந்தியா), புலம்பெயர் தமிழர் என இரண்டு பிரிவுகளாக தமிழ் மென்பொருள் பாவனையாளர்களை பார்க்கும்போது தமிழ் மென்பொருள் தொடர்பில் இந்த இரண்டு பிரிவினருக்குமான தேவைகள், பாவனைமுறை என்பன மிகவும் மாறுபட்டவை. அதனை கருத்தில் கொண்டு எவ்வாறான புதிய மென்பொருள்களினை உருவாக்க வேண்டும் என்பதனை ஆராயவேண்டும்.

அடுத்ததாக, கடந்த காலங்களில் எமது தமிழ் இணைய மாநாடுகளில் தமிழ் மென்பொருள்களின் உருவாக்கம் தொடர்பாக பல கட்டுரைகள் படிக்கப்பட்டிருக்கின்றன. அவை அனைத்துமே மக்களிற்கு ஏதோ ஒரு வகையில் தேவையானவையே. ஆனாலும் அவற்றில் பெரும்பாலானவை ஆராய்ச்சி வடிவுடனேயா நின்றுபோகின்றன. அவ்வாறு நின்றுபோகாமல் அவற்றிற்கு முழு வடிவம் கொடுத்து மக்கள் பயன்பாட்டிற்கு உகந்ததாக மாற்றினால் எம் மக்கள் மத்தியில் பெரும் அதிசயங்களை நிகழ்த்தும் என்பதில் சந்தேகம் இல்லை.

மொத்தத்தில், கணினி மற்றும் இணையப் பாவனையில் தமிழ் மென்பொருள்களின் இருப்பும் பாவனையும் என்பது தொடர்பாக ஆராய்வதன்மூலம் தற்போது உள்ள தமிழ் மென்பொருள்களை சரியான மக்கள் பாவனைக்கு கொண்டுசெல்வதுடன் மக்களுக்கு தேவையான சரியான புதிய தமிழ் மென்பொருள்களை உருவாக்குவதிலும் கவனத்தை செலுத்தமுடியும்.

# Electronic Commerce

*Dr. B. Neelavathy*

**Electronic commerce**, commonly known as **e-comm**, **e-commerce** or **eCommerce**, consists of the buying and selling of products or services over electronic systems such as the Internet and other computer networks. The amount of trade conducted electronically has grown extraordinarily with widespread Internet usage. The use of commerce is conducted in this way, spurring and drawing on innovations in electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. Modern electronic commerce typically uses the World Wide Web at least at some point in the transaction's lifecycle, although it can encompass a wider range of technologies such as e-mail, mobile devices and telephones as well.

A large percentage of electronic commerce is conducted entirely electronically for virtual items such as access to premium content on a website, but most electronic commerce involves the transportation of physical items in some way. Online retailers are sometimes known as e-tailers and online retail is sometimes known as **e-tail**. Almost all big retailers have electronic commerce presence on the World Wide Web.

Electronic commerce that is conducted between businesses is referred to as business-to-business or B2B. B2B can be open to all interested parties (e.g. commodity exchange) or limited to specific, pre-qualified participants (private electronic market). Electronic commerce that is conducted between businesses and consumers, on the other hand, is referred to as business-to-consumer or B2C. This is the type of electronic commerce conducted by companies such as Amazon.com. Online shopping is a form of electronic commerce where the buyer is directly online to the seller's computer usually via the internet. There is no intermediary service. The sale and purchase transaction is completed electronically and interactively in real-time such as Amazon.com for new books. If an intermediary is present, then the sale and purchase transaction is called electronic commerce such as eBay.com.

Electronic commerce is generally considered to be the sales aspect of e-business. It also consists of the exchange of data to facilitate the financing and payment aspects of the business transactions.

## History

Originally, electronic commerce was identified as the facilitation of commercial transactions electronically, using technology such as Electronic Data Interchange (EDI) and Electronic Funds Transfer (EFT). These were both introduced in the late 1970s, allowing businesses to send commercial documents like purchase orders or invoices electronically. The growth and acceptance of credit cards, automated teller machines (ATM) and telephone banking in the 1980s were also forms of electronic commerce. Another form of e-commerce was the airline reservation system typified by Sabre in the USA and Travicom in the UK.

From the 1990s onwards, electronic commerce would additionally include enterprise resource planning systems (ERP), data mining and data warehousing.

In 1990, Tim Berners-Lee invented the WorldWideWeb web browser and transformed an academic telecommunication network into a worldwide everyman everyday communication system called internet/www. Commercial enterprise on the Internet was strictly prohibited until 1991. Although the Internet became popular worldwide around 1994 when the first internet online shopping started, it took about five years to introduce security protocols and DSL allowing continual connection to the Internet. By the end of 2000, many European and American business companies offered their services through the World Wide Web. Since then people began to associate a word "ecommerce" with the ability of purchasing various goods through the Internet using secure protocols and electronic payment services.

## **DOT-COMS**

Internet use gave a large jump toward the turn of the century, from being common in 26 percent of households in 1998 to 55 percent in 2003. Usage rates continue to climb in the United States and worldwide. This widespread use caused the rise—later followed by the collapse—of many Internet-based businesses, called “dot-coms” for their adoption of the suffix “.com” at the end of their names, referring to their Web site addresses. They used the three Cs method of business—commerce, content, and connection—offering one of the three to possible customers. Although the dot-coms formed the basis for today's e-commerce, inflated expectations and inexperience in online business transactions lead to the dot-com bubble of 2000 and 2001, when many purely online businesses imploded, costing investors millions. Some of the more famous dotcom busts include Flooz.com, 360Hiphop, Pets.com, Kibu.com, and GovWorks.com, which was featured in the documentary *Startup.com*.

After the dot-com bubble, the surviving companies dropped the coms from the end of their names and went on, some becoming successful businesses. For most companies, however, a combination of physical-based customer service and products with online components offering similar services has proven to be a more trustful method of incorporating e-commerce. In response to the dot-com bust and the continued growing interest in online trade, the Federal Trade Commission, or FTC, began to elaborate on their previous online business regulations.

Chief among the FTC's regulations is the policy that all online advertisement must tell the truth and not mislead customers. As in physical markets, all online claims must be substantiated. Disclaimers can be particularly complex on Web sites, and the FTC requires that all disclaimer information must be easily accessible and readable. In response to worries of online security issues such as account and identity theft, the FTC has also made it clear that online companies should notify customers when collecting personal data, and several Privacy Protection Acts created during the dot-com era were made to enforce that policy.

## **Strategies**

One of the first challenges involved in moving to online commerce is how to compete with other e-commerce sites. A common problem in addressing this challenge is that e-commerce is often analyzed from a technical standpoint, not a strategic or marketing perspective. E-commerce provides several technical advantages over off-line commerce. It is much more convenient for the buyer and the seller, as there is no need for face-to-face interaction and Web-based stores are open 24 hours a day. Also, e-commerce purchasing decisions can be made relatively quickly, because a vendor can present all

relevant information immediately to the buyer. These factors lend themselves to a transactional approach, where e-commerce is seen as a way to reduce the costs of acquiring a customer and completing a sale.

In contrast, most successful e-commerce Web sites take a relational view of e-commerce. This perspective views an e-commerce transaction as one step among many in building a lasting relationship with the buyer. This approach requires a long-term, holistic view of the e-commerce purchasing experience, so that buyers are attracted by some unique aspect of an e-commerce Web site, and not by convenience. Since consumers can easily switch to a competing Web site, customer loyalty is the most precious asset for an e-commerce site.

While the primary focus of most Internet activity is on the business-to-business and business-to-consumer facets of e-commerce, other transaction methods are included. The success of eBay and its consumer-to-consumer portal for auction-based transactions has dramatically changed how people and companies conduct business. In addition to having a significant effect on business-to-business transactions, retailers are beginning to tap into this new and dynamic approach to commerce.

## **Widgets and e-Commerce**

A widget is a transferrable piece of code that can move itself in and out of Web site data, collecting information or executing a particular function for a metadata program. Some of the most visible widgets are the advertisements seen on most Web sites. These are in fact pieces of code from a third-party business that are being used to communicate marketing messages.

Widgets are one of the most important tools for e-commerce, used most often for distribution of information and online promotional activities. A 2008 article by Ori Soen with *TechNewWorld* explores the new possibilities widgets offer companies interested in e-commerce. Not only are widget-advertisements inexpensive and relatively

easy to employ, they can be combined with present marketing efforts and visual productions with the added effect of animation, if desired.

The problem most cited with present-day widget use by e-commerce companies is that online users no longer pay attention to widget advertisement. Most business Web sites accessible today have a multitude of widgets, and the advertisements are often diluted. Like emerging problems with TV commercials, users have learned to simply stop paying attention. Soen, however, sees this as an opportunity for companies to develop more innovative marketing techniques, better online animations, and more effective branding strategies aimed at online users.

Still, e-commerce Web sites are often crowded, and Soen suggests a different focus for widget advertisement: social networks. Social applications, such as MySpace and Facebook, are another field open to creative widget use, but they also offer a more open demographic, namely, people who are more likely to be attracted to creative widgets and – more importantly – have the ability to spread the word to their friends about advertisements that have caught their eye, giving companies two ways to promote instead of one.

Widgets serve a third purpose for e-commerce companies: the ability to collect important data concerning what advertisements are most effective to customers. When widgets are combined with

analysis tools, they can be very useful gatherers of marketing information. They can judge how long a potential customer spends with the widget, and to what extent they interact with the animation. Promotional activities and marketing analysis can be effectively combined.

## **Personalization**

One of the key practices to a successful e-commerce company is personalization. Rachelle Crum's 2008 article, "Personalization: Telling E-tail Customers What They Really Want," lists several ways businesses can personalize their customers' online experiences.

When customers buy products online, they often receive a short list of other items they may be interested in. This is known as *recommendation*, and because of the ease and access in online business, it is relatively easy for businesses to include in their e-commerce activity. Customers are much more likely to order from the company when they receive a personalized list of products.

Tailored Web pages are another important part of personalization. Many companies have designed their e-commerce businesses to use the data from returning customers to create specific Web pages advertising new products the customer may be interested in, deals that may appeal particularly to the customer, and other information tailored especially for them.

## **Mobile Web**

Certain devices, such as the iPhone, are becoming popular for their ability to access the Internet remotely. New technology has allowed remote Web access through phone and other handheld devices to become faster and easier to use. Some e-commerce companies have begun developing Web applications specifically for mobile Web users. This entails creating streamlined Web pages that condense information and allow customers to find what they looking for quickly and without hassle. Since these streamlined applications can be easier to navigate than normal Web pages, and can be accessed from nearly any location, some predictions say the mobile Web will become a powerful tool in the e-commerce field.

There are several different ways companies can make e-commerce more available to mobile Web users. Web designers can simply remove graphics from the Web site for mobile applications, giving users a simplified, text-only site to navigate. Style sheets can also be used, to create other versions of online stores, tailored to specific devices. Or, if a company wanted to devote more time to the project, a second Web site could be created solely for mobile Web users.

## **Web Site Creation Tips**

Beyond technology tools such as widgets and mobile Web devices, there are simple ways to improve Web pages and how they read and look. Slight changes in the way a Web page integrates marketing, product information, and visuals can create an enormous difference in the perceptions of customers. Most Web surfers spend a very short time inspecting online stores before moving on, and the right words or the right information, displayed correctly, can make a great difference.

As David Needle says in his 2008 article for *Smallbusinesscomputing.com*, there are three different kinds of written cues or signposts that companies can place in their Web sites. The first type of signpost is navigation-oriented. This involves the way the online store is constructed—where the links to

products are and where they go, how far down the Web site pages scroll, and links to related sites. "Bread crumbs," or easy ways for customers to return to the sites they have previously seen, are an excellent tool to give online stores structure and usability.

The second type of signpost is microcontent-related. This refers to Web site headings, URLs, and titles that organize online information. All information in the company's online store should be clear and easy to read and understand.

The third type of signpost is metadata, which is data concerning the information of the Web site itself, such as how many users have accessed it and the keywords within the site that would come up in a search engine or analytical program.

## **Barriers to Success**

Despite the growing number of e-commerce success stories, plenty of e-commerce Web sites do not live up to their potential. There were two primary causes of e-commerce failures during the early 2000s.

First, most Web sites offer a truncated e-commerce model, meaning that they do not give Web users the capability to complete an entire sales cycle from initial inquiry to purchase. As analyzed by Forrester Research, the consumer sales cycle has four stages. First, consumers ask questions about what they want to buy. Second, they collect and compare answers. Third, the user makes a decision about the purchase. If the purchase is made, the fourth phase is order payment and fulfillment (delivery of the goods or services). The problem is that many Web sites do not provide enough information or options for all four phases. For example, a site may provide answers about a product, but not answers to the questions that the consumer has in mind. In other cases, the consumer gets to the point where he or she wants to make a purchase, but is not given an adequate variety of payment options to place the actual order.

The second problem occurs when e-commerce efforts are not integrated properly into the corporate organization. A survey by *InteractiveWeek* magazine found that in most companies e-commerce is treated as part of the information system (IS) staff's responsibility, and not as a business function. While sales and marketing staff generally assist in the development of e-commerce Web sites, final profit and loss responsibility rests with the IS staff. This is a major source of breakdowns in e-commerce strategy because the units that actually make products and services do not have direct responsibility for selling them on the Web. One promising trend is that more companies are beginning to decentralize the authority to create e-commerce sites to individual business units, in the same way that each unit is responsible for its part of a corporate intranet.

## **Success Factors**

After studying many aspects of electronic commerce, several consulting and analytic firms created guidelines on how to implement and leverage it successfully. In particular, two organizations have developed lists of critical success factors that seem to capture the state of thinking on this topic. First is the Patricia Seybold Group, which publishes trade newsletters and provides consulting services related to using information technology in corporations. This firm identified five critical e-commerce success factors:

**Support customer self-service.** If they so desire, Web users should be enabled to complete transactions without assistance.

**Nurture customer relationships.** Up-front efforts should focus on increasing customer loyalty, not necessarily on maximizing sales.

1. **Streamline customer-driven processes.** Firms should use Web technology to reengineer back-office processes as they are integrated with e-commerce systems.
2. **Target a market of one.** Each customer should be treated as an individual market, and personalization technology should be employed to tailor all services and content to the unique needs of each customer.
3. **Build communities of interest.** A company should make its e-commerce Web site a destination that customers look forward to visiting, not simply a resource people use because they have to conduct a transaction.

A quick review of two successful e-commerce sites, the [Amazon.com](http://Amazon.com) bookstore site and Dell Computer's Web site, illustrate how many of these principles combine to help develop a strategic e-commerce capability.

Amazon.com, which has one of the highest sales volumes of any Web-based business, has optimized its site for the nature of its products and the preferences of its customers. The site is highly personalized; each visitor to the site, once registered, is greeted by name. The site content also is customized. Using software based on pattern recognition, [Amazon.com](http://Amazon.com) compares a particular customer's purchase history to its overall record of transactions and generates a list of recommended books that seem to fit his or her interests and tastes. The company has a very integrated customer service support system, so that any customer service representative can access all data on the transactions, purchasing information, and security measures of each customer. The system also supports communications using e-mail, fax, and telephone.

Finally, Amazon.com helps to build a community of users through its Associates Program. Under this program, a Web site can host a hyperlink directly to the Amazon.com site. Any time that a visitor to that site buys books through Amazon.com, the Web site owner receives a share of the transaction revenues. This is a very inexpensive way for Amazon.com to extend its marketing and advertising reach across the Web. Dell Computer also uses personalization and customization tools. For every major corporate customer, Dell creates a special Premier Page, which shows all products covered under purchasing contracts with that firm, as well as the special pricing under those contracts. This ensures that employees of that firm always get the right price for each purchase. Ford Motor Company reports that by encouraging employees to buy PCs from its Premier Page, the company saved \$2 million in one year.

Dell also has integrated its e-commerce Web site with all back-office systems, so that when a customer orders a custom-configured PC, that information is automatically transferred to the production system to ensure that the unit is built according to specifications. This also improves customer service; Dell will proactively notify any customer if a production problem or inventory shortage will delay delivery.



Electronic commerce, as used by U.S. firms, has already undergone several generations of evolution. Early experiences helped to stabilize e-commerce technology and set the development path for more sophisticated and useful technologies. Later experiences provided guidelines on strategic approaches and operational models that will help to improve e-commerce success.

Three key issues will determine the long-term viability of electronic commerce. These are:

1. Technological feasibility, or the extent to which technology—bandwidth availability and information reliability, tractability, and security—will be able to sustain exponentially increasing demands worldwide.
2. Socio-cultural acceptability, or the extent to which different global cultures and ways of doing business will accommodate this new mode of transacting, in terms of its nature (not face-to-face), speed, asynchronicity, and unidimensionality.
3. Business profitability, or the extent to which this way of doing business will allow for profit margins to exist at all (e.g., no intermediaries, instant access to sellers, global reach of buyers).

As technology continues to develop and mature, the ability to assess the impact of electronic commerce will become more cogent. Moreover, the significance of privacy, security, and intellectual property rights protection as prerequisites for the successful worldwide diffusion, adoption, and commercial success of Internet-related technologies—especially in places with less democratic political institutions and highly regulated economies—is continually increasing. The differentiation between the Internet (the global network of public computer networks) and intranets (corporate-based computer networks that involve well-defined communities and potentially more promising technology platforms for fostering Internet-related commerce) became significant in the late 1990s and early 2000s. Intranet development has surpassed the Internet in terms of revenue—by 2005 more than half of the world's Web sites were commercial in nature.

### **Adverse Possibilities of e-Commerce**

Ned Kock, in his book *Encyclopedia of E-collaboration* (2008), gives several possible negative effects of e-commerce, if the trend continues at the same rate it is currently growing.

Global companies with highly developed online stores may already possess the extra edge to attract potential customers. This may leave beginning companies, eager to enter the online market, without much chance to make an impact. International competition may become skewed and lead to an unhealthy type of oligopoly in the e-commerce world.

Some also fear that e-commerce will allow companies to evade certain tax laws, especially when it comes to international trade. New regulations might need to be set for customs concerning online exchanges.

Others wonder how e-commerce will change the job market. While online business offers jobs to those with newer IT skills, it can also displace many traditional jobs.

### **Advantages of Electronic Commerce**

The greatest and the most important advantage of e-commerce, is that it enables a business concern or individual to reach the global market. It caters to the demands of both the national and the

international market, as your business activities are no longer restricted by geographical boundaries. With the help of electronic commerce, even small enterprises can access the global market for selling and purchasing products and services. Even time restrictions are nonexistent while conducting businesses, as e-commerce empowers one to execute business transactions 24 hours a day and even on holidays and weekends. This in turn significantly increases sales and profit.

Electronic commerce gives the customers the opportunity to look for cheaper and quality products. With the help of e-commerce, consumers can easily research on a specific product and sometimes even find out the original manufacturer to purchase a product at a much cheaper price than that charged by the wholesaler. Shopping online is usually more convenient and time saving than conventional shopping. Besides these, people also come across reviews posted by other customers, about the products purchased from a particular e-commerce site, which can help make purchasing decisions.

For business concerns, e-commerce significantly cuts down the cost associated with marketing, customer care, processing, information storage and inventory management. It reduces the time period involved with business process re-engineering, customization of products to meet the demand of particular customers, increasing productivity and customer care services. Electronic commerce reduces the burden of infrastructure to conduct businesses and thereby raises the amount of funds available for profitable investment. It also enables efficient customer care services. On the other hand, It collects and manages information related to customer behavior, which in turn helps develop and adopt an efficient marketing and promotional strategy.

### **Disadvantages of Electronic Commerce**

Electronic commerce is also characterized by some technological and inherent limitations which has restricted the number of people using this revolutionary system. One important disadvantage of e-commerce is that the Internet has still not touched the lives of a great number of people, either due to the lack of knowledge or trust. A large number of people do not use the Internet for any kind of financial transaction. Some people simply refuse to trust the authenticity of completely impersonal business transactions, as in the case of e-commerce. Many people have reservations regarding the requirement to disclose personal and private information for security concerns. Many times, the legitimacy and authenticity of different e-commerce sites have also been questioned.

Another limitation of e-commerce is that it is not suitable for perishable commodities like food items. People prefer to shop in the conventional way than to use e-commerce for purchasing food products. So e-commerce is not suitable for such business sectors. The time period required for delivering physical products can also be quite significant in case of e-commerce. A lot of phone calls and e-mails may be required till you get your desired products. However, returning the product and getting a refund can be even more troublesome and time consuming than purchasing, in case if you are not satisfied with a particular product.

### **Conclusion**

Thus, on evaluating the various pros and cons of electronic commerce, we can say that the advantages of e-commerce have the potential to outweigh the disadvantages. A proper strategy to address the technical issues and to build up customers trust in the system, can change the present scenario and help e-commerce adapt to the changing needs of the world.

# **Natural Language Processing**

(இயற்கை மொழி பகுப்பாய்வு)



# An Efficient Tamil Text Compaction System

*N.M..Revathi, G.P.Shanthi, Elanchezhian.K, T V Geetha,  
Ranjani Parthasarathi & Madhan Karky*

*Tamil Computing Lab (TaCoLa),  
College of Engineering Guindy, Anna University, Chennai.  
haisweety18@gmail.com, jijutodo@gmail.com, madhankarky@gmail.com*

## Abstract

Tamil is slowly becoming the online language and mobile text messaging languages for many Tamils around the world. Social networks and mobile platforms now extensively support Unicode and applications for keying Tamil text. The number of characters in a text message is limited in some social nets and mobile text messages. The need for compacting the text becomes essential as it translates to saving online storage space, cost and many more factors. The paper proposes a text compaction system for Tamil, a first of its kind in Tamil. The system proposed in this paper handles common Tamil words, acronyms/abbreviations and numbers. Morphological analyzer [1] and Morphological generator are used to stem inflexion words and replace them to compact using a mapping repository. The proposed work is tested with over 10,000 words and it is found that the final result is reduced to 40% of the original text. The paper concludes by discussing possible extensions to this system.

## 1. Introduction:

In all languages, using compact or short form of words in text messages, emails, and blogs is rapidly increasing. It is particularly popularly amongst young urbanities as it allows for voiceless communication, useful in noisy environment that would defeat a voice conversation and also buffered communication since the message the sender wants to convey can be accessed by the receiver at any time. Compacting text is thus necessary because of limited message length in blog sites and tiny user interface of mobile phone. Getting the shortest word has no rule and it is mainly aimed at understanding. That is, those words should be understood by everyone. We can obtain the compact words by omitting letters, replacing prefix and suffix of through suitable symbols and numbers. This causes the compacted system to be credited with creating a language. The paper proposes a Text Compaction system for Tamil, the primogenital in Tamil..

## 2. Background:

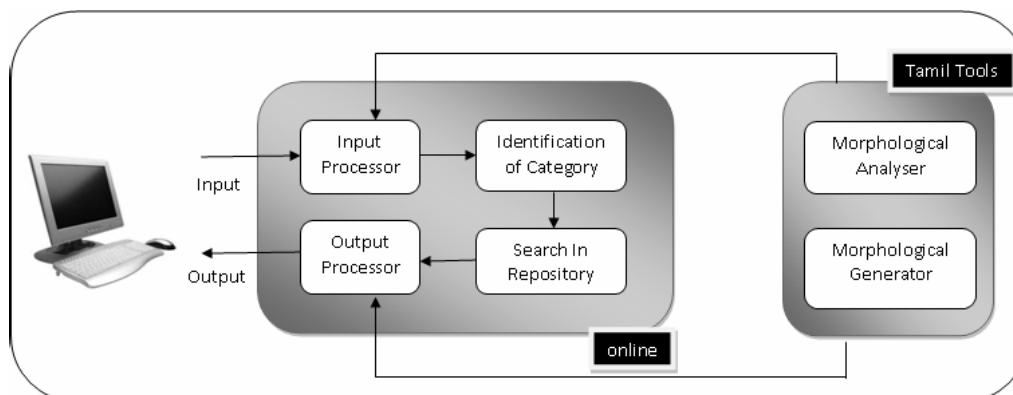
Tamil is perhaps the only classical language, whose glorious literatures date back to the pre-Christian era, has remained in continuous use for more than many millennia now. Due to the untiring efforts of scholars, researches and enthusiasts, it has also evolved creatively over the years to the extent that it is also used today profusely in computers, internet, mobile phone etc. Diverse creative efforts have been taking place that would pave the way for a quantum jump in the usage of Tamil in Information Technology. "Tamil Virtual University", "Centre for Research and Applications of Tamil in Internet",

“Tamil Software Development Fund” is to quote a few. These efforts paved the way for the motivation of proposing Tamil compaction system in Tamil.

Many compaction systems have been developed for English and other languages. Lee Ming Fung in [2] proposed a Short form Identification and Categorization model based on maximum entropy to identify short forms from actual words and acronyms/abbreviations and categorize the short forms into the short forms formed from letter omission and those formed through phonetic substitution of parts of words. In the proposed system the compact words are formed in a diverse variety of ways such as omission, truncation and phonetic substitution. Acronym Identification and detection has been much researched. Acrophile in [3] automatically searches acronyms from acronym-expansion pairs from domain specific databases. By acronyms expansion pairs, we refer to a pairs each containing acronyms and their full expanded form or meaning. The paper makes use of acronym expansion pairs to replace the full expanded form with the acronyms.

### 3. Text Compaction Framework:

The figure below presents the various components of the framework.



#### 3.1 Input Processing

The input text is tokenized based on a delimiter and is passed on to the Morphological Analyzer. The analyzer removes the suffix (if present) added to the word and delivers the root word (RW). For example if the input to the analyzer is கணிப்பொறியில் the output is given as கணிப்பொறி.

#### 3.2 Identification of the type

The proposed paper handles three categories of words; common Tamil words, Abbreviations /acronyms, numbers. Now, the category to which the RW belongs is to be identified. The RW is checked to decide the category of abbreviations /acronyms. This is done by comparing the root word with the keys of the hash map (2.3). If the comparison results are true then the RW is considered as the abnormal word (AW) i.e. it belongs to the category of acronyms/abbreviations, else, it is treated as the normal word (NW) i.e. it belongs to either the first or third category.

### 3.3 Extraction of the compact word

If the word is identified as a normal word, it is passed to a tree which is built dynamically from the set of words that has already been stored in the dictionary. The NW is then searched in the binary search tree. On finding the NW in the binary search tree, the compact word is retrieved with an efficient mapping algorithm that maps each of the normal word with its compact word.

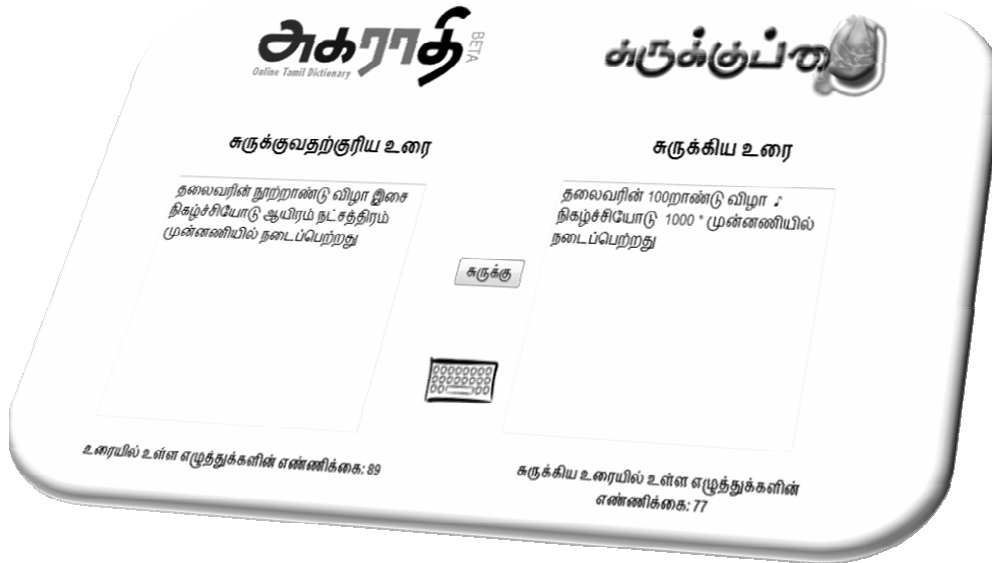
Say suppose the word is an abnormal word, its compact word is retrieved in the following manner. A linked hash map is built for all the abbreviated words. The hash map uses the first word the abbreviated word as its key. Again with the help of an efficient mapping algorithm, the compact word is retrieved. In case the NW is a number name it is replaced with the numerals based on the place value system.

### 3.4 Output Processing

The compact word that is being extracted is passed on the Tamil tool Morphological Generator to add the suitable suffix to cater to the rules of the language.

## 4. Results and Analysis:

The paper proposes the following layout for displaying the results to the user. It has two text areas: the one on the left is for entering the input text and the other on the right for displaying the output. The user can also view the no of characters that have been reduced in the output text.



Efficiency of the system can be calculated as (no of characters in the input text / no of characters in the output text) X 100%. The proposed work is tested with over 10,000 words and it is found that the final result is reduced to 40% of the original text.

## 5. Conclusion and Future work:

The paper describes the Tamil Compaction System, a framework for shrinking the text such that its meaning remains the same. Different subsystems and components of the framework are described in detail. Results from the implementation of this Tamil compaction system framework is provided and is compared against the compacting third party applications of social networking sites that are implemented for English language. Improving the mapping for words which are frequently used, conceptual reducing, integrating numerical analyser will take this system to its next level.

## References:

- Anandan, R. Parthasarathi, and T.V. Geetha, *Morphological Analyser for Tamil*. ICON 2002, 2002.
- Fung, L. M. (2005). *SMS short form identification and codec*. Unpublished master's thesis, National University of Singapore, Singapore
- *Acrophile* (LSLarkey, P Ogilvie, MA Price, B Tamilio, 2000) a system that automatically searches acronym expansion pairs.
- *Short Message Service (SMS) Texting Symbols: A Functional Analysis of 10,000 Cellular Phone Text Messages* by Robert E. Beasley, Franklin College.



# Tamil Summary Generation for a Cricket Match

*J. Jai Hari Raju, P. Indhu Reka, K.K Nandavi, Dr. Madhan Karky*

Tamil Computing Lab (TaCoLa),

College of Engineering Guindy, Anna University, Chennai.

jaihari1989@gmail.com, p.indhu@gmail.com, ashwathas@gmail.com, madhankarky@gmail.com

## Abstract

Cricket is one of the most followed sports in the Indian subcontinent. There is a wide requirement for natural language descriptions, which summarize a cricket match effectively. The process of generating match summaries from statistical data is a manual process. The objective of this paper is to propose a framework for automatic analysis and summary generation for a cricket match in Tamil, with the scorecard of the match as the input. Data analytics is performed on the statistical match data, to mine all frequently occurring patterns. The paper proposes a parameter called *Interestingness*, which quantifies the interestingness of the match. The paper also proposes a customization model for the summary. We propose an evaluation parameter called *humanness*, which quantifies the similarity between the output and a manually written summary. Discussing the results and analyzing the summaries generated for matches based on scorecards, this paper concludes with proposing some extensions for future developments.

## 1. Introduction

The number of websites which facilitate people to follow and analyze sports has increased manyfold. Among them, there are an exceptionally large number of sites devoted to Cricket. Mostly these involve participation of experts, who present their views and summaries in English about cricket matches. There are no such sites which provide similar services in Tamil. In this case it is also desirable if there is an alternative for human creativity. As a solution the paper proposes an automated Tamil summary generation framework which is capable of analyzing and generating a Tamil summary about a cricket match, provided the score card as the input. This paper discusses the overall architecture and implementation details of such a framework.

The large amount of data in this domain makes it possible to apply data mining and data analytics techniques. The input scorecard is analyzed to construct feature vectors, which are then subjected to data mining. Based on the various parameters identified, the interestingness of the match is quantified.

The summary generation part involves extraction of key players and events from a match. Appropriate sentences are then synthesized to express these selected events. The sentence constructs and the vocabulary used are chosen based on the linguistic ability specified by the user. Then the sentences are combined in to a meaningful summary.

The results of the system, i.e. the summaries, are evaluated based on the *Humanness* parameter. This parameter gives the degree of similarity between the generated summary and the manually written summary, with which it is compared. This value helps us decide, the level of creativity achieved by

the system. In section 2 we provide an overview of the literature survey conducted. In section 3 we discuss the design of the various modules of the framework. In section 4 we discuss the implementation of the proposed framework and the results obtained from the analysis. Finally we conclude in section 5 with extensions to the current framework and directions for further studies in the field of Tamil Summary Generation Systems.

## 2. Background

In the literature there are existing works on summary generation from statistical data. Alice Oh et al. generated multiple stories about a single baseball game based on different perspectives using a reordering algorithm [1]. Ehud Reiter et al. in their book building natural language generation systems explain the difference between natural language generation and natural language processing and also describe the various steps involved in the natural language generation process with examples [2]. Jacques Robin et al. presented a system (called STREAK) for summarizing data in natural language. It focuses on basketball game to design and evaluate the system [3]. L. Bourbeau et al. came up with the FoG (Forecast Generator) using the streamlined version of the Meaning-Text linguistic model. This system was capable of generating weather forecasts in both English and French [4].

## 3. Summary Generation Framework

The Tamil Cricket Summary Generator consists of the following major components:

- Data Gathering and Modeling module
- Data Mining and Data Analytics module
- Summary Generator
- Evaluator

Figure (1) given below depicts the Summary Generation framework.

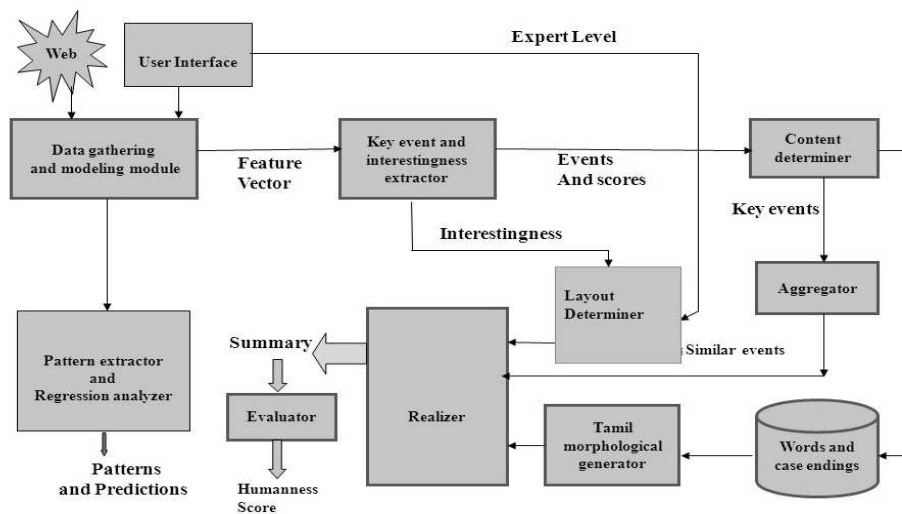


Figure 1: Tamil Cricket Summary Generator Framework

### **3.1 Data Gathering and Modeling Module**

Data gathering is the first step of the system. The data to be gathered is present in internet. This module has a custom designed parser, for the tag structure of the site. The user must provide the URL from where the particular match's data can be obtained. The module checks whether the match has already been processed. If not the parser parses the page and retrieves the statistical data. The statistical data is then modeled in the form of the predefined feature vectors.

### **3.2 Data Mining and Analytics Module**

Modified version of Apriori algorithm is used to find the association rules from the feature vectors. After performing mathematical analysis using correlation of variance (CoV), CoV is plotted against average to give an idea about how consistent the player is. The interestingness of the match is calculated based on the weighted average of the scores assigned to the factors identified, they include the Winning margin, Team history, Individual records made, High run rate, Series state, Relative position in international ranking, Reaction in social networks etc.

### **3.3 Summary Generator**

The summary generator part of the framework consists of the following sub modules Content Determiner, Aggregator, Tamil Morphological Generator and Layout Determiner. The events to be included in the summary are not predefined and are not the same for every match. Based on the interestingness of the total match, the interestingness of the individual events and the expert level chosen by the user, particular events are chosen to be included in the summary. The content determiner is responsible for identifying those facts which are worth mentioning in the summary.

Aggregation of relevant events from other matches in the summary will make it more readable and interesting. The aggregator performs this function. It chooses events based on their similarity and coherence and aggregates them with the key events selected in the content determiner module.

As a next step, the sentences used to describe the events are synthesized. The sentence which is the most apt to the current event under consideration is selected. The vocabulary used in the sentence and the depth to which an event is discussed is also varied based on the expert level of the user. The nouns in the key events are passed to the morphological generator along with the desired case endings and the generated variants are added to the sentences.

The layout determiner module chooses the layout of the summary to be generated. The layout is varied based on the interestingness of the match. The sentences are aggregated in the fashion of the layout selected and the final output summary is passed to Evaluator.

### **3.4 Evaluator**

The summary generated by the system is evaluated based on its degree of similarity with human written summaries. The summaries are compared based on two parameters, the Nouns Mentioned and the Events Mentioned.

The nouns and the events in the summaries are extracted along with their absolute positions. The events in the summary are modeled as a set consisting of, one or more Performers (the persons who takes part in the event), Numeral (the numeric part involved in the event e.g. 4 wickets) and a

Descriptor (the action connecting the Performer and the Numeral). Their absolute positions refer to the sentence number in which they are mentioned. Then these absolute positions are normalized based on the total number of sentences present in the summary. Three different scores are calculated they are,

- Similarity Score: The ratio of the number of nouns and events mentioned in both the summaries to the total number of nouns and events mentioned at least in one summary.
- Count Score: The ratio of the number of nouns and events mentioned in the system generated summary to the number of nouns and events mentioned in the human written summary
- Closeness Score: The degree of closeness, in terms of the normalized positions of the nouns and events mentioned in both the summaries.

A weighted average of these three scores yields the final humanness score.

## **4. Implementation**

To implement the proposed framework, *espnricinfo* a reliable and prominent site for Cricket data is chosen as the source of input. The frame work was implemented in java. The URL of the match for which the summary is to be generated is obtained from the user. The feature vectors designed for modeling a match are stored as rows with unique identities, in the back end oracle database. The patterns which are generated as a result of data mining are validated based on the support and confidence parameters. As a design decision all nouns are stored in English and are translated on the fly using a constantly updated look up database. This decision was taken to allow interoperability and easy extension of the system to other languages in future. The sentence pattern files are stored external to the system, so as to allow modifications without changes in the system. The summary generated for the match is stored in the back end, indexed with the unique identity assigned already. The user interface is designed to be simple and robust. It allows the users to search matches based on various parameters and also to save their preferences.

### **4.1 Results**

Score cards of 90 One Day International matches were retrieved and their summaries were generated. These include matches between 9 countries. Both individual matches and series were considered. A large number of hidden patterns in cricket domain have been retrieved based on the algorithm used. The patterns have been validated and the ones which are interesting have been reported. The factors contributing to the interestingness of the match have been identified and the weights associated with them have been found. The consistency of a player has been modelled and consistency analysis of a player is done to analyse his performance.

The difference in the language used and the events mentioned in the summary is pronounced when the user opts for an expert level. Similar facts occurring in the past have been identified and added to the summary. Each summary was compared with two human written summaries, one an expert summary and other an average summary, their cumulative scores were considered. The humanness score of the summaries tend to be in the range of 70% to 85%. The recurrence of layouts is also minimal, which reflects the fact that the summaries generated are not monotonous.

## 5. Conclusion and future work

In this paper we have proposed the framework for an Automated Tamil Cricket Summary Generator. The current implementation of the system can be enhanced by adding machine learning capabilities to make the summaries more human and interesting. The system can be extended to produce summaries in multiple languages apart from Tamil. The system can be enhanced to generate summaries about the match in real time. As a next level the system can be modified for summary generation in other sports too.



Figure 2: Screenshot of the Tamil Cricket Summary Generation System

The frame work can be used as a guideline to develop summary generation systems, which can be applied for any domain where frequent numerical reports are used. (Weather Prediction, Industrial Quality Testing etc)

## References

- Alice Oh and Howard Shrobe, "Generating baseball summaries from multiple perspectives by reordering content," in Proc. 5th International Natural Language Generation Conference, 2008, pp. 173-176.
- Ehud Reiter and Robert Dale, "Building natural language generation systems," Cambridge: Cambridge University Press, 2000.
- Jacques Robin and Kathleen McKeown, "Empirically Designing and Evaluating a New Revision-Based Model for Summary Generation," Department of Computer Science, Columbia University, 1996, vol. 85, pp.135-179.
- L. Bourbeau, D. Carcagno, E. Goldberg, R. Kittredge and A. Polguere. "Bilingual generation of weather forecasts in an operations environment," In Proc. 13<sup>th</sup> International Conference on Computational Linguistics, Helsinki University, Finland, 1990. COLING

# Lyric Mining: Word, Rhyme & Concept Co-occurrence Analysis

*Karthika Ranganathan, T.V Geetha, Ranjani Parthasarathi & Madhan Karky*

*Tamil Computing Lab (TaCoLa),*

*College of Engineering Guindy, Anna University, Chennai.*

*karthika.cyr@gmail.com, madhankarky@gmail.com*

## ABSTRACT

Computational creativity is one area of NLP which requires extensive analysis of large datasets. Laalalaa [1] framework for Lyric analysis and generation proposed a lyric analysis subsystem that required statistical analysis of Tamil lyrics. In this paper, we propose a data analysis model for words, rhymes and their usage in Tamil lyrics. The proposed analysis model extracts the root words from lyrics using a morphological analyzer [2] to compute the word frequency across the lyric dataset. The words in their unanalyzed form are used for computing the frequent rhyme, alliteration and end-rhyme pairs using adapted apriori algorithm. Frequent co-occurring concepts in lyrics are also computed using Agaraadhi, an on-line Tamil dictionary. Presenting the results, this paper concludes by discussing the need of such an analysis to compute freshness, pleasantness of a lyric and using these statistics for Lyric Generation.

**Keywords :** Tamil Lyrics, Morphological Analyser, Apriori algorithm.

## I. INTRODUCTION

Tamil is one of the world's oldest languages and has a Classical status. Numerous forms of literature exist in Tamil language of which, lyrics play a vital role in taking the language to every house hold in form of original film soundtracks, jingles, private albums, and commercials. With over thousands of lyrics being created every year, we do not have proper tools to model and analyse lyrics. Such an analysis framework would enable one to see various patterns of words, combinations and thoughts used over time. The analysis framework will also make it possible to generate fresh lyrics where the freshness can be associated with the concepts and thoughts associated with the lyric.

In this paper, we discuss about Tamil lyric Analysis on the basis of word usage, rhyme usage and the co-occurrence of word. The frequency of word usage is identified by considering a morphological root of the word using morphological analyser, instead of considering terms. For analysing the frequent rhyme, alliteration and end-rhyme pairs, we adapted Apriori algorithm [4]. To identify the co-occurring concepts in lyrics, we used "Agaraadhi", an on-line Tamil dictionary Framework [3] and a new algorithm has been proposed to compute the frequent usage of co-occurring concepts in lyrics.

The rest of this paper has been organized as follows. In Section 2, we explain about the algorithm and tools. In Section 3, we explain the methodology and in Section 4, we discuss our results. Conclusions and future extensions to this work are presented in section 5.

## 2. MORPHOLOGICAL ANALYSIS AND APRIORI ALGORITHM

Morphological analysis is the process of segmenting words into morphemes and identifying its grammatical categories. For a given word, morphological analyser (MA) generates its root word and its grammatical information. The role of morphological analyser in the proposed work is to identify the noun and verb morphology of a given word, examples are as follows

Example 1, for noun morphology:

**இராமனை (Ramanai)**  
**இராமன் (Raman) + ஐ (ai)**  
Entity + Accusative Case

Example 2, for verb morphology:

**சென்றான் (Senraan)**  
**செல் (sel) + ற் (R) + ஆன் (Aan)**  
Verb + Past Tense Marker + Third Person Masculine Singular Suffix

In the proposed work, the frequency of a word is identified by considering the variations of a noun in terms of its morphology. For instance, the variations of Ramanai such as Ramanaal, Ramanukku, Ramanin, Ramanadhu are also counted for the word Raman.

The Apriori Algorithm is an influential algorithm for mining frequent item sets for Boolean association rules [4]. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. It has objective measures: support and confidence. The support of an association pattern is the percentage of task-relevant data transactions for the apparent pattern. Confidence can be defined as the measure of certainty or trustworthiness associated with each discovered pattern.

## 3. LYRIC ANALYSIS

### (i) Word Analysis

The frequency of words is used to associate a popularity score for each word. This score is proposed to be used for lyric generation part of the frame work proposed in [1]. In this work, the popularity score of a word has been identified from lyrics. In lyrics, the words are mainly attached with the suffix. So, the root words are taken into consideration for determining its frequency count. The root words are identified using morphological analyser. The algorithm to find the word usage is illustrated below:

#### Algorithm

Let  $L_D$  is the set of lyric dataset and  $L_S$  denotes the set of sentences of lyric dataset and  $L_W$  denotes the set of words in the lyric dataset.  $W_C$  denotes the word count across all lyric dataset. Let  $m$  be the total number of sentences in lyrics and  $n$  be the total number of words in lyrics.

a) Given a Lyric dataset  $L_D$

b) For each  $L_{Si} \leftarrow 1$  to  $m$

Split the sentence  $L_S$  into words  $L_W$

c) For each  $L_{Wj} \leftarrow 1$  to  $n$

$R_W \leftarrow \text{ProcessMorphAnalyser}(L_{Wj})$

d) Let  $R_W$  be the root word

if  $R_W$  exist, then add into the word count list (  $W_C$  )

else add into the word count list (  $W_C$  )

e) Return  $W_C$ .

Here  $\text{ProcessMorphAnalyser}(L_{Wj})$  returns the root of the given word.

### (ii) Rhyme Analysis

Alliteration (Monai) is the repetition of the same letter at the beginning of words. The rhyme (Edhugai) is defined as the repetition of the same letter at the second position of words. The end rhyme (iyaibu) is defined as the repetition of the same letter at the last position of words. The example of alliteration, rhyme and end rhyme is given below:

#### Examples:

**உயிர்** and **உன்** rhyme in alliteration (monai) as they start with the same letter.

**இதயம்** and **காதல்** rhyme in rhyme (edhugai) as they share the same second letter.

**யாக்கை** and **வாழ்க்கை** rhyme in end - rhyme (iyaibu) as they share the same last letter.

We have adapted apriori algorithm to find the frequency count of rhyme, alliteration and end rhyme pairs of Tamil lyrics which has been illustrated below:

#### Algorithm:

Let  $L_D$  be the set of lyric dataset and  $L_S$  denote the set of sentences of Lyric dataset. Let  $m$  be the total number of sentences in lyrics. Let  $P_{C1}$  denote the count of alliteration and  $P_{C2}$  denote the count of rhyme and  $P_{C3}$  denote the count of end - rhyme.

a) Given lyric dataset  $L_D$ .

b) For each  $L_S \leftarrow 1$  to  $m$

Join the pair of sentences ( $L_P$ )

c) For each  $L_P$

Consider the first ( $L_{P1}$ ) and last words ( $L_{P2}$ )

d) Rhyme ( $L_{P1}, L_{P2}$ )

e) return  $P_C$



**Algorithm : Rhyme ( $L_{P1}, L_{P2}$ )**

a) Let  $k$  denote the  $i$ <sup>th</sup> character. Let  $M_L$  denote the alliteration (monai) list and  $R_L$  denote the rhyme (edhugai) list and  $E_L$  denote the end – rhyme (iyaibu) list.

b) For  $\forall i$ ,

if  $k = 1$ , if  $L_{P1(k)} = L_{P2(k)}$ , then add into  $M_L$  list and increment  $P_{C1}$

if  $k = 2$ , if  $L_{P1(k)} = L_{P2(k)}$ , then add into  $R_L$  list and increment  $P_{C2}$

if  $k = i - 1$ , if  $L_{P1(k)} = L_{P2(k)}$ , then add into  $E_L$  list and increment  $P_{C3}$

**(iii) Co-occurrence concept Analysis**

Co-occurrence is defining the frequent occurrence of two terms from a text corpus on the either side in a certain order. This word information in NLP system is extremely high. It is very important for cancelling the ambiguous and the polysemy of words to improve the accuracy of the entire system [5].

In this method, to improve the efficiency of co-occurrence, we have been considering the concept of each word. The concept for each word has been identified using the Agaraadhi, an on-line Tamil dictionary. The example for concept word which has been in lyric is given below:

**Example:** The word "நிலவு" which has the concept வெண்ணிலா, மதி, மாதம், துணைக்கோள், வெண்ணிலவு, அம்புலி, அம்புலிமான்.

By considering these concepts, we have been determining the co-occurring words using our own algorithm is described below:

**Algorithm :**

Let  $L_D$  denote the set of Lyric dataset and  $W$  denote each word in lyric. Let  $C_W$  denote the set of concepts for each word. Let  $W_C$  denote the word count.

a) If the word  $C_W$  identify, then consider the next word.

Increment the count  $W_C$

b) Else the word  $W$  and consider the next word.

Increment the count  $W_C$

c) return  $W_C$

**4. RESULTS**

The lyric corpus of more than two thousand songs were analysed for the word usage, rhyme usage and Co-occurrence concepts usage. The analysed results are given below:

Table 1 shows the list of top 10 usage words in lyrics.

| WORDS | USAGE | WORDS | USAGE |
|-------|-------|-------|-------|
| நீ    | 2009  | வா    | 1062  |
| என்   | 1941  | ஒரு   | 987   |
| நான்  | 1645  | கண்   | 965   |
| உன்   | 1556  | பூ    | 857   |
| காதல் | 1153  | இல்லை | 793   |

Table 2 shows the list of top 10 rhyme words in lyrics.

| EDHUGAI    | USAGE  | MONAI      | USAGE | IYAIBU      | USAGE  |
|------------|--------|------------|-------|-------------|--------|
| என்,உன்    | 107975 | என்,என்னை  | 33492 | என்,உன்     | 111552 |
| நான்,என்   | 80125  | உன்னை,உன்  | 26289 | நான்,என்    | 83435  |
| நான்,உன்   | 61204  | எந்தன்,என் | 16478 | நான்,உன்    | 63543  |
| என்,உன்னை  | 33731  | என்,என்ன   | 15405 | என்,உந்தன்  | 18411  |
| என்,என்னை  | 32570  | உந்தன்,உன் | 14001 | எந்தன்,என்  | 16478  |
| உன்னை,உன்  | 25747  | உயிர்,உன்  | 11640 | உந்தன்,உன்  | 14001  |
| என்னை,உன்  | 24867  | இந்த,இது   | 10993 | எந்தன்,உன்  | 12524  |
| நான்,உன்னை | 19297  | எனது,என்   | 9985  | நான்,உந்தன் | 10524  |
| நான்,என்னை | 18935  | எந்த,என்   | 9976  | எந்தன்,நான் | 9367   |
| என்,என்ன   | 14486  | நீ,நீயும்  | 9962  | இந்த,அந்த   | 4818   |

Table 3 shows the list of top 10 co-occurring concept words in lyrics.

| CO-OCCURRING WORDS | USAGE |
|--------------------|-------|
| அன்பே,அன்பே        | 638   |
| சின்ன,சின்ன        | 530   |
| வா,வா              | 506   |
| என்,காதல்          | 478   |
| ஒரே,ஒரு            | 469   |
| நீயும்,நானும்      | 434   |
| உன்னை,நான்         | 419   |
| தமிழ்,எங்கள்       | 367   |
| ஒரு,நாள்           | 302   |
| நீ,என்னை           | 287   |

By analysing those data, this shows that the most of lyrics which predicts the emotion of happiness and love. In the adapted apriori algorithm, the support which represents the total number of pair words with the total number of combination of sentences and the confidence which described the total number of pair words with the total number of pair of sentences. The results may vary if the number of lyrics used for the analysis is increased.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we discussed the usage of words and rhymes in the lyrics dataset. The adapted apriori algorithm has been used to detect the frequency count for rhyme, alliteration and end word pairs. This analysis has been mainly used in the lyric generation and computing freshness scoring for lyrics. Frequent co-occurring concept is also been identified for the development of lyric extraction, semantic relationship, word sense identification and sentence similarity. Possible extensions of this work could be the detection of emotions in lyrics by genre classification and identify genre specific rhymes and concept co-occurrence.

## REFERENCES

- Sowmiya Dharmalingam, Madhan Karky, "LaaLaLaa - A Tamil Lyric Analysis and Generation Framework" in World Classical Tamil Conference - June 2010, Coimbatore.
- Anandan P, Ranjani Parthasarathy, Geetha, T.V. "Morphological analyzer for Tamil". ICON 2002.
- Agaraadhi Online Tamil Dictionary. <http://www.agaraadhi.com>, Last accessed date 25<sup>h</sup> April 2011.
- HAN Feng, ZHANG Shu-mao, DU Ying-shuang, "The analysis and improvement of Apriori algorithm", Journal of Communication and Computer, ISSN1548-7709, USA, Sep. 2008, Volume 5, No.9 (Serial No.46).
- EI-Sayed Atlam, Elmarhomy Ghada, Masao Fuketa, Kazuhiro Morita and Jun-ichi Aoe, "New Hierarchy Technique Using Co-Occurrence Word Information", International Journal of Information Processing and Management, Volume 40 Issue 6, November 2004.

# Template based Multilingual Summary Generation

Subalalitha C.N, E.Umamaheswari, T V Geetha,

Ranjani Parthasarathi & Madhan Karky

subalalitha@gmail.com

Tamil Computing Lab (TaCoLa)

College of Engineering Guindy Anna University, Chennai.

## Abstract

Summarization of large text documents becomes an essential task in many Natural Language processing (NLP) applications. Certain NLP applications deal with domain specific text documents and demand for a domain specific summary. When the essential facts are extracted specific to the domain, the summary proves to be more efficient. The proposed system builds a bilingual summary for an Information Retrieval (IR) system named CoRee, which tackles Tamil Language and English Language text documents [1]. As the input documents are tourism domain specific documents, the summary is extracted based on specially designed seven tourism specific templates 7 both for Tamil and English. The templates are filled in with the required information extracted from the UNL representation and a bilingual summary is generated for each text document irrespective of the language of input text document. The efficiency of the summary has been tested manually and it has achieved 90% efficiency. This efficiency depends on factors other than summary generation such as conversion accuracy and dictionary entry coverage. The proposed system can be extended for many languages in future.

## 1. Introduction

Automatic summary generation has been a research problem for over 40 years [2]. Summarizing the texts helps in avoiding information overload and also saves time. Multi lingual Natural Language applications have emerged in great number in recent years. This makes the need for a multi lingual summary generation a quintessential task. Alkesh patel et al have come up with a multi lingual summary generation by using structural and statistical factors [2]. David Kirk Evans has generated multi lingual summary using text similarities existing in the sentences [3]. Dragomir Radev et al have developed a multi lingual summary generation tool named MEAD using centroid and query based methods. They have also used many learning techniques such as decision trees, Support Vector Machines (SVM) and Maximum Entropy [4].

All the above works on multi lingual summarization have not used a interlingua document representation . We propose that a multi lingual summary can be generated with much more ease by using a interlingua document representation language called, "Universal Networking Language" (UNL) [5]. UNL converts every term present in a natural language text document into a language independent concept, thereby making the applications built using it a language independent one. The proposed work extracts a domain specific summary, as the UNL documents used are tourism domain specific. Tourism specific templates are framed and the sentences fitting the templates are chosen and formed as a summary.

The rest of the paper is organized as follows. Section 2 gives a brief introduction about UNL. Section 3 describes the proposed summarization technique. Section 4 discusses the evaluation of the proposed work. Section 5 reveals the enhancements needed to the proposed work and Section 6 gives the conclusion of the paper.

## 2. Universal Networking Language

UNL is an intermediate language that processes knowledge across language barriers. UNL captures the semantics of the natural language text by converting the terms present in the document to concepts. These concepts are connected to the other concept through UNL relations. There are 46 UNL relations like plf(Place From), plt(Place To), tmf(Time from), tmt(Time to) etc [1]. This process of converting a natural language text to UNL document is known as Enconversion and the reverse process is known as Deconversion. The UNL document is normally represented as a graph where the nodes are concepts and edges are UNL relations. An example UNL graph is shown for the example 1.

Example 1: John was playing in the garden .

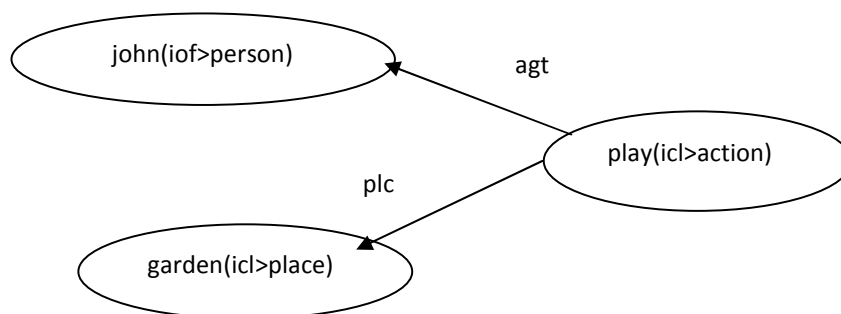


Figure 1: UNL graph for Example 1

The nodes of graph namely, "John(iof>person)", "Play(icl>action)" and garden(icl>place) represent the terms John, playing and garden present in the example 1. The semantic constraints in the concepts, "iof>person", "icl>action" and "icl>place" denotes the context in which the concepts occur. The edges namely, "agt" and "plc" indicates that, the concepts involved are agents and place. From the above discussion, it is shown that the UNL inherits many semantic information from the natural language text and portrays in a language independent fashion.

The proposed work uses Tamil language text documents and English language documents enconverted to UNL for summary extraction which is described in the next section.

## 3. Template based Information Extraction

As discussed earlier, the summary is generated using the tourism specific templates. Figure 2 shows the over view of the proposed summary generation framework. The Framework consists of both language dependent and independent parts. The functionalities involving UNL are language independent and the inputs supplied to the framework to generate bi lingual summary are the language dependent parts. The bilingual summary generation is explained in the coming sections.

The seven templates describe about the tourism specific information of a place such as, god, food, flora and fauna, boarding facility, transport facility, place and distance. The correct information for these templates are extracted as discussed below. The usage of semantics helps greatly in eliminating the ambiguities that may arise while picking up a concept to fill the slot. For instance, the word, “bat” may denote a cricket bat or the mammal bat.

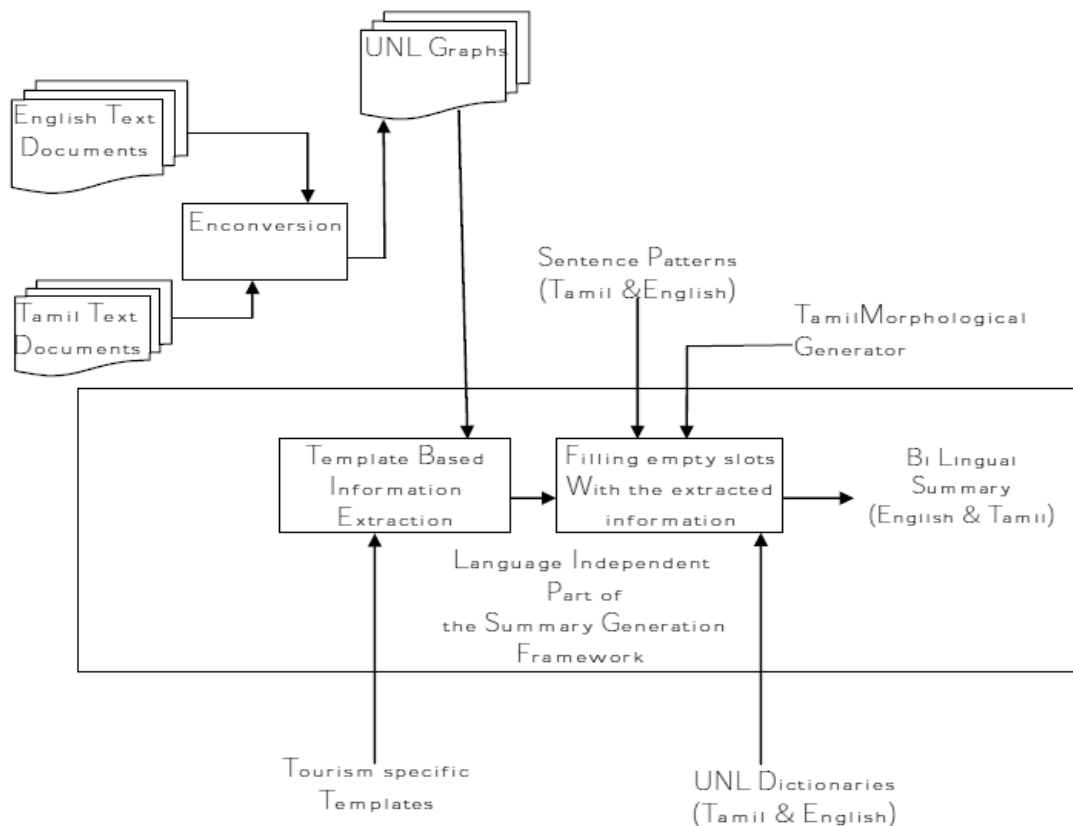


Figure 2: Overview of the Summary Generation Framework

This type of ambiguity is resolved by the semantic constraint, as the cricket bat will get the semantic constraint, “obj<thing (object thing)”, whereas the mammal gets the semantic constraint, “icl>mammal”. **Table 1** displays few semantics used for the respective templates.

The extracted tourism specific concepts are converted to the target language terms for building a summary using the sentence patterns which is explained in the next section.

#### 4 Multi Lingual Summary generation

The information (concepts) extracted from the UNL graph using the templates are converted to the target language term using the respective UNL dictionary. For instance, to generate the English summary, the concepts comprising the semantic constraints are converted to English terms using the English UNL dictionary which consists of mapping between English terms and UNL concepts. These terms which when filled into the appropriate English sentence patterns, gives a English summary . The same procedure is done for building a Tamil summary. For each UNL graph irrespective of its source language, a summary in Tamil and English are generated.

| Template           | Semantics                                       |
|--------------------|---|
| God                | iof>god, iof>goddess, icl>god                   |
| Food               | icl>food, icl>fruit                             |
| Flaura and Fauna   | icl>animal, icl>reptile, icl>mammal, icl> plant |
| Boarding facility  | icl>facility                                    |
| Transport facility | icl>transport                                   |
| Place              | icl>place, iof>place, iof>city, iof>country     |
| Distance           | icl>unit , icl>number                           |

Table 1 :Semantics used for each templates

The terms obtained from the UNL dictionary will be a root word. For instance, the term, “eating” will be entered as eat (icl >action) in the UNL dictionary. So the terms obtained from the UNL dictionary needs to be generated to its original form using Morphological generator. The summary generation requires only tourism specific concepts, so the generation is almost not required . But we have used a morphological generator for Tamil, as the place information and distance information in Tamil with the case suffixes இல் (il), இலிருந்து(ilirunthu), உக்கு (ukku) etc needs to be generated. For the example UNL graph shown in figure 3, the generated transport template in Tamil which is part of the summary is given in example 2.

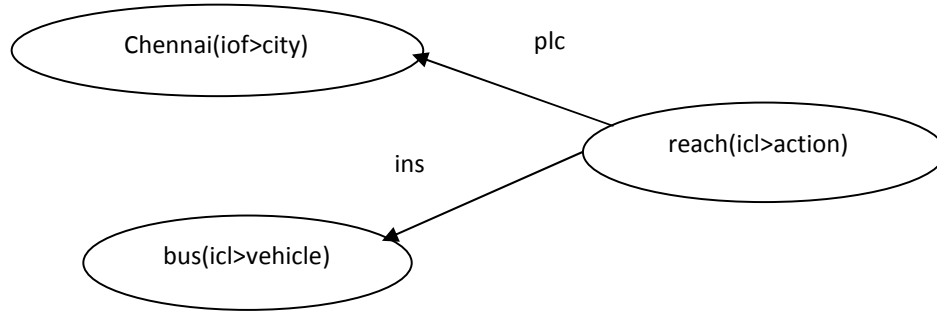


Figure 3:UNL graph given as input for example 2

Example 2: சென்னைக்கு பேருந்தில் செல்லலாம்

The concept chennai(iof>city) in the above graph, is generated as "சென்னைக்கு" by adding the case suffix “உக்கு ” and the concept bus(icl>vehicle) is generated as

"பேருந்தில்" by adding the case suffix, “இல்”.

## 5. Performance Evaluation

The proposed work has been tested with 33,000 Tamil and English text documents converted to UNL graphs. The performance of the methodology proposed has been evaluated using human judgement. The accuracy of the summary generated has achieved 90% . Apart from the summary generation factors such as tourism specific concept extraction , the accuracy also depends on the quality of conversion and dictionary entry. By improving these factors, the accuracy can further be improved.

## 6. Conclusion and Future work

The proposed work generates a tourism specific bilingual summary using the intermediate document representation, UNL and tourism specific templates. The bilingual summary is generated in a simple and efficient manner compared to the earlier work done for multi lingual summary generation. The only overhead involved is developing a converter framework.

As future enhancements, sentence patterns can be replaced by selecting the sentences having high sentence score based on its sentence position and the frequency of concepts. Query specific summary can also be generated on line, as the summary discussed here is a tourism specific generated off line using the templates. The evaluation of the generated summary can also be done by comparing it with the human generated summary. By doing this, many factors to make the machine generated summary compatible with human generated summary may evolve.

## Reference

- Elanchezhian K, T V Geetha, Ranjani Parthasarathi & Madhan Karky, CoRe - Concept Based Query Expansion, Tamil Internet Conference, Coimbatore, 2010.
- Alkesh Patel , Tanveer Siddiqui , U. S. Tiwary , "A language independent approach to multilingual text summarization", Conference RIAO2007, Pittsburgh PA, U.S.A. May 30-June 1, 2007
- David Kirk Evans, "Identifying Similarity in Text: Multi-Lingual Analysis for Summarization ", Doctor of Philosophy thesis, Graduate School of Arts and Sciences , Columbia University, 2005
- Radev, Allison, Blair-Goldensohn et al (2004), *MEAD - a platform for multidocument multilingual text summarization*
- The Universal Networking Language (UNL) Specifications Version 3 Edition 3, UNL Center UNDL Foundation December 2004.
- Jagadeesh J, Prasad Pingali, Vasudeva Varma, " Sentence Extraction Based Single Document Summarization" Workshop on Document Summarization, March, 2005, IIT Allahabad.
- Naresh Kumar Nagwani, Dr. Shrish Verma , "A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm " International Journal of Computer Applications (0975 - 8887) Volume 17- No.2, March 2011 .
- Prof. R. Nedunchelian, "Centroid Based Summarization of Multiple Documents Implemented using Timestamps " First International Conference on Emerging Trends in Engineering and Technology, IEEE 2008



# Special Indices for LaaLaLaa Lyric Analysis & Generation Framework

*Suriyah M, Madhan Karky, T V Geetha, & Ranjani Parthasarathi*

*{suriyah.cse@gmail.com, madhankarky@gmail.com,*

*tv\_g@hotmail.com, rp@annauniv.edu}*

*Tamil Computing Lab (TaCoLa),*

*College of Engineering Guindy, Anna University, Chennai.*

## **Abstract**

With the advent of computational tools for creativity, it becomes inevitable to design data structures which cater to the specific needs of the creative form considered. A lyric generator has to retrieve words fast based on Part of Speech and rhyme. This paper aims at building special indices for the LaaLaLaa Lyric Generator framework based on POS and rhyme to facilitate faster retrieval. The retrieval times of the proposed model and the conservative word indexed model are compared. The indexing is based on the KNM(Kuril, Nedil, Mei) pattern and the letters that occur in the rhyming spots of the words. The data structure is organized as hash tables to ensure best retrieval complexity. Separate hash tables for each POS and rhyming scheme are created and populated. Here, the key would be the meter pattern with the letters occurring at the rhyming spots and the value would be the list of all those words which fall under the key's constraint. When the word indexed and meter rhyme indexed retrievals were compared, the latter reduced the average retrieval time drastically. There were not steep variations in the retrieval times as was in the former approach. This remarkable efficiency was traded-off with space.

## **1. Introduction**

Tamil, one of the oldest languages, has a very rich literary history dating back to two thousand years. We have more than two thousand lyrics being written in this language in the form of film songs, advertisements, jingles, private albums etc. With the advent of computational tools for creativity, it becomes inevitable to design data structures which cater to the specific needs of the creative form considered. Tools for poem generation, story generation and lyric generation have been proposed.

A poem has to have three qualities – meaningfulness, poeticness and grammaticality [1]. A lyric is a poem which has constraints of having to satisfy a tune and a theme. This paper talks about building special indices for the LaaLaLaa lyric generator framework[2] to aid faster retrieval of words based on POS and rhyme.

A lyric generator would require the retrieval of words of a particular meter and particular letters at rhyming spots. This would maximize the poeticness of the lyric generated with the increase in rhymes. This retrieval process, when carried out on an un-indexed word database, is too expensive as it would take separate processing for meter, and each of the three rhymes in Tamil. To facilitate faster retrieval of words satisfying these constraints, the word database has to be indexed based on

meter-pattern and rhyme. This makes indexing of the word database based on the abovesaid constraints necessary.

This paper is organized as eight sections. The second section discusses about a few existing works in this area. The third section gives an overview of the Rhyme schemes in Tamil. An overall view of the system is given by the fourth section while the fifth section talks about the approach proposed for indexing. This work concludes with the results obtained and scope for future work in this area.

## 2. Background

Though significant number of works has been done in the arena of poetry generation in other languages, there is only less number in Tamil.

The “Automatic Generation of Tamil Lyrics for Melodies” [3] identifies the required syllable pattern for the lyric and passes this to a sentence generation module which generates meaningful phrases that match the pattern. This system generates rhyme based on maximum substring match and fails to make use of the three rhyming schemes that are specific to Tamil language. “LaaLaLaa - A Tamil Lyric Analysis and Generation Framework” [2] generates Tamil lyrics for POS tagged pattern with words from a rhyme finder according to rhyming schemes in Tamil. Nichols et al[5] investigate the assumption that songwriters tend to align low-level features of a song’s text with musical features. K. Narayana Murthy[4] suggests having a non-dense TRIE index in main memory and a dense index file stored in secondary memory.

Anna Babarczy et al[6] suggested a hypothesis that a metaphoric sentence should include both source-domain and target-domain expressions. This assumption was tested relying on three different methods of selecting target-domain and source-domain expressions: a psycholinguistic word association method, a dictionary method and a corpus-based method. Hu, Downie and Ehmann[7] examine the role lyric text can play in improving audio music mood classification. Mahedero et al[8] argue that a textual analysis of a song can generate ground truth data that can be used to validate results from purely acoustic methods. Mayer et al[9] present a novel set of features developed for textual analysis of song lyrics, and combine them with and compare them to classical bag-of-words indexing approaches and results for musical genre classification on a test collection in order to demonstrate our analysis. “Semantic analysis of song lyrics” studies the use of song lyrics for automatic indexing of music. Netzer et al explore the usage of Word Association Norms (WANs) as an alternative lexical knowledge source to analyze linguistic computational creativity. Logan et al. use song lyrics for tracks by 399 artists to determine artist similarity[12].

## 3. Rhyme Schemes and Rhyme Patterns

**Rhyme Schemes:** English has number of rhyme effects like assonance, consonance, perfect, imperfect, masculine, feminine etc. This arises from the variation in stress patterns of words, lack of clear cut description about the spots where rhymes can occur.

In Tamil, the grapheme and phoneme are bound stronger than in English. There are 3 characteristic rhyme schemes in Tamil – Monai (மோனை), Edhugai (எதுகை) and Iyaibu (இயைபு).

Two words are said to rhyme in monai if their first letters are the same, in edhugai if their second letters are the same and in iyaibu if their last letters are the same.

Examples: பறவை and பச்சை rhyme in monai as they start with the same letter.

அருவி and விருப்பு rhyme in edhugai as they share the same second letter.

யாக்கை and வாழ்க்கை rhyme in iyaibu as they share the same last letter.

As one may infer, two words can rhyme in more than one pattern also.

Examples: அருவி and குருவி rhyme in edhugai and iyaibu.

கவிதைகள் and கவிஞர்கள் rhyme in all the three schemes.

**Meter Pattern:** One way of classifying alphabets of the Tamil language is based on the time interval (மாத்திரை - maathirai) for which they are pronounced. One maathirai corresponds to the time taken to wink the eyelid. The types of letters in this classification are

Nedil (N) (நெடில்) - Those alphabets which are pronounced for the time interval of 2 maathirai.

Kuril (K) (குறில்) - Alphabets which take 1 maathirai to be pronounced.

Mei (M) (மெய்) - Alphabets which are pronounced for 0.5 maathirai.

Meter pattern of a word refers to its Kuril Nedil Mei pattern.

For example, the meter pattern of the word பாலல் is NKM as ப is a Nedil(N), ல is a Kuril(K) and ல் is a Mei(M).

The indexing methodology proposed needs to take care of both meter pattern and rhyme to facilitate faster retrieval of words for the LaaLaLaa Lyric Generation and Analysis framework[2].

#### 4. Overview of the System

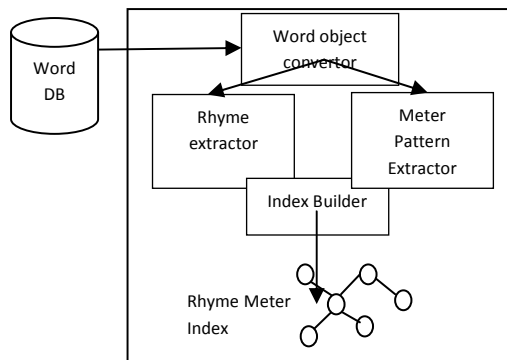


Figure 1. Overview of the system.

Each word from the word database is converted to an object. Meter pattern and the alphabets at the Rhyming positions of each word are found out by Meter Pattern Extractor and Rhyme Extractor respectively. Using the data obtained, the Index builder builds the Rhyme Meter Index aiding faster retrieval than the normal un-indexed retrieval.

## 5. Indexing Algorithm

| Part of Speech |               |         |       |
|----------------|---------------|---------|-------|
| மோனை           | MeterPattern1 | Letter1 | Words |
|                |               | Letter2 | Words |
|                | MeterPattern2 | Letter1 | Words |
|                |               | Letter2 | Words |
| எதுகை          | MeterPattern1 | Letter1 | Words |
|                |               | Letter2 | Words |
|                | MeterPattern2 | Letter1 | Words |
|                |               | Letter2 | Words |
| இயைபு          | MeterPattern1 | Letter1 | Words |
|                |               | Letter2 | Words |
|                | MeterPattern2 | Letter1 | Words |
|                |               | Letter2 | Words |

Figure 2. Indexing Logic

This indexing has been designed to facilitate fast retrieval of words specifically for lyric generation. For instance, the system would need a word of a particular meter pattern with a particular letter rhyming in monai scheme.

The data structure is organized as hash tables with separate tables for each Part Of Speech and Rhyming Scheme. For example, there is a separate table for Nouns' monai, Nouns' edhugai, Nouns' iyaibu and so on. Here, the keys are the Meter Pattern and the Letter at the particular Rhyming spot. The values are the words corresponding to the particular Meter pattern and letter.

**The algorithm :** The word database is scanned word by word and is indexed based on meter pattern and rhyme. Here,

$\alpha \leftarrow$  number of words in the word database;

$w_i \leftarrow$  ith word in the word database;

$\beta \leftarrow$  meter pattern of  $w_i$ ;

$monaiKey \leftarrow$  key of  $w_i$  corresponding to monai;

$edhugaiKey \leftarrow$  key of  $w_i$  corresponding to edhugai;

$iyaybuKey \leftarrow$  lkey of  $w_i$  corresponding to iyaybu;

for

$i = 1, 2, 3 \dots \alpha$  do

$\beta \leftarrow$  MeterPattern of  $w_i$ ;

$monaiKey \leftarrow$  firstLetter ( $w_i$ ) +  $\beta$ ;

$edhugaiKey \leftarrow$  secondLetter ( $w_i$ ) +  $\beta$ ;

$iyaybuKey \leftarrow$  lastLetter ( $w_i$ ) +  $\beta$ ;

Add the word to the list of words with the respective keys in the respective Hashtables.

end;

For each word in the word database, meter pattern is extracted first followed by letters at the rhyming spots namely, first, second and last positions (for monai, edhugai and iyaibu respectively). Keys for monai, edhugai and iyaybu are found out using the abovesaid equations. The word is added to the monai, edhugai and iyaibu hashtables with keys monaiKey, edhugaiKey and iyaybuKey respectively if those keys don't appear previously in the tables. Else, the word is added to the list of words with that key.

Hash-tables are chosen for the implementation as they have the retrieval complexity of  $O(1)$ .

## 6. Results

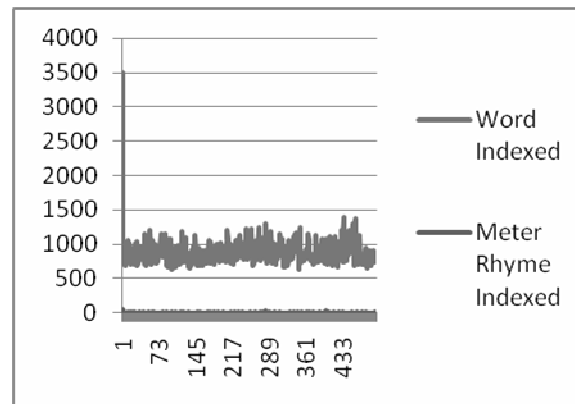


Figure 3. Word Indexed Vs Meter Rhyme Indexed Approach

Indexing has brought about a drastic increase in the speed of retrieval of the words rhyming with a given word. Using the Word Indexed approach, the time complexity was  $O(\alpha)$  where  $\alpha$  was the total number of words. But after Meter Rhyme indexing, the complexity has become  $O(1)$  due to the use of hash table which is a very efficient data structure for retrieval. Rhyming words for a sample of 500 words were retrieved using both the approaches and the above mentioned graph was obtained. The Word indexed system took 875.47 millisecond in an average while the Meter Indexed system took 1.90 millisecond only. From the graph it can also be inferred that there are steep variations in the Word-Indexed approach while the Meter-Rhyme Indexed approach does not show such steep variations and is consistent. In terms of time efficiency, Meter-Rhyme indexed approach is evidently superior compared to Word-Indexed approach. In terms of space, it is not so efficient as each word will occur not once, but nine times in various Hash tables.

## 7. References

- Hisar Maruli Manurung: "An evolutionary algorithm approach to poetry generation", Thesis for Doctor of Philosophy, University of Edinburgh, 2003.
- Sowmiya Dharmalingam., Madhan KarKy. "LaaLaLaa - A Tamil Lyric Analysis and Generation Framework" in World Classical Tamil Conference - June 2010, Coimbatore

- RamaKrishnan, A., S. Kuppan, and S.L. Devi. "Automatic Generation of Tamil Lyrics for Melodies" in NAACL HLT Workshop on Computational Approaches to Linguistic Creativity. 2009. Colorado.
- K. Narayana Murthy "An Indexing Technique for Efficient Retrieval from Large Dictionaries", National Conference on Information Technology NCIT-97, 21-23 December 1997, Bhubaneswar.
- Eric Nichols, Dan Morris, Sumit Basu, Christopher Raphael, "Relationships between lyrics and melody in popular music", ISMIR 2009, October 2009, Japan.
- Anna Babarczy, IldiKó Bencze, István FeKete, Eszter Simon, "The Automatic Identification of Conceptual Metaphors in Hungarian Texts: A Corpus-Based Analysis", Proceedings of The seventh international conference on Language Resources and Evaluation (LREC), 2010, Malta.
- Xiao Hu, J. Stephen Downie, Andreas F. Ehmann, "Lyric Text Mining in Music Mood Classification", ISMIR 2009, Japan.
- Jose P. G. Mahedero, Alvaro Martinez, Pedro Cano, "Natural Language Processing of Lyrics", Proceedings of the 13th annual ACM internationalconference on Multimedia, New York, NY, USA, 2005.
- Rudolf Mayer, Robert Neumayer, Andreas Rauber, "Rhyme and style features for musical genre classification by lyrics", Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08), Philadelphia, PA, USA, September 14-18, 2008.
- Beth Logan, Andrew KositsKy, Pedro Moreno, "Semantic analysis of Song Lyrics", IEEE International Conference on Multimedia and Expo (ICME), June 2004.
- Yael Netzer, David Gabay, Yoav Goldberg, Michael Elhadad, "GaiKu : Generating HaiKu with Word Associations Norms", Workshop on Computational Approaches to Linguistic Creativity, CALC-2009 in conjunction with NAACL-HLT 2009, Boulder, Colorado.
- B. Logan, A. KositsKy, and P. Moreno, "Semantic Analysis of Song Lyrics", in Proc IEEE ICME, 2004.

# Tamil Document Summarization Using Latent Dirichlet Allocation

N. Shreeya Sowmya<sup>1</sup>, T. Mala<sup>2</sup>

<sup>1</sup>*Department of Computer Science and Engineering, Anna University*

<sup>2</sup>*Department of Information Science and Technology, Anna University  
Guindy, Chennai*

<sup>1</sup>shreeya.mel@gmail.com

<sup>2</sup>malanehru@annauniv.edu

## Abstract

This paper proposes a summarization system for summarizing multiple tamil documents. This system utilizes a combination of statistical, semantic and heuristic methods to extract key sentences from multiple documents thereby eliminating redundancies, and maintaining the coherency of the selected sentences to generate the summary. In this paper, Latent Dirichlet Allocation (LDA) is used for topic modeling, which works on the idea of breaking down the collection of documents (i.e) clusters into topics; each cluster represented as a mixture of topics, has a probability distribution representing the importance of the topic for that cluster. The topics in turn are represented as a mixture of words, with a probability distribution representing the importance of the word for that topic. After redundancy elimination and sentence ordering, summary is generated in different perspectives based on the query.

**Keywords-** Latent Dirichlet Allocation, Topic modeling

## I. Introduction

As more and more information is available on the web, the retrieval of too many documents, especially news articles, becomes a big problem for users. Multi-document summarization system not only shortens the source texts, but presents information organized around the key aspects. In multi-document summarization system, the objective is to generate a summary from multiple documents for a given query. In this paper, summary is generated from the multi-documents for a given query in different perspectives. In order to generate a meaningful summary, sentences analysis, and relevance analysis are included. Sentence analysis includes tagging of each document with keywords, named-entity and date. Relevance analysis calculates the similarity between the query and the sentences in the document set. In this paper, topic modeling is done for the query topics by modifying the Latent Dirichlet Allocation and finally generating the summary in different perspectives

The rest of the paper is organized as follows. **Section 2** discusses with the literature survey and the related work in multi-document summarization. **Section 3** presents the overview of system design. **Section 4** lists out the modules along with the algorithm. **Section 5** shows the performance evaluation. **Section 6** is about the conclusion and future work.

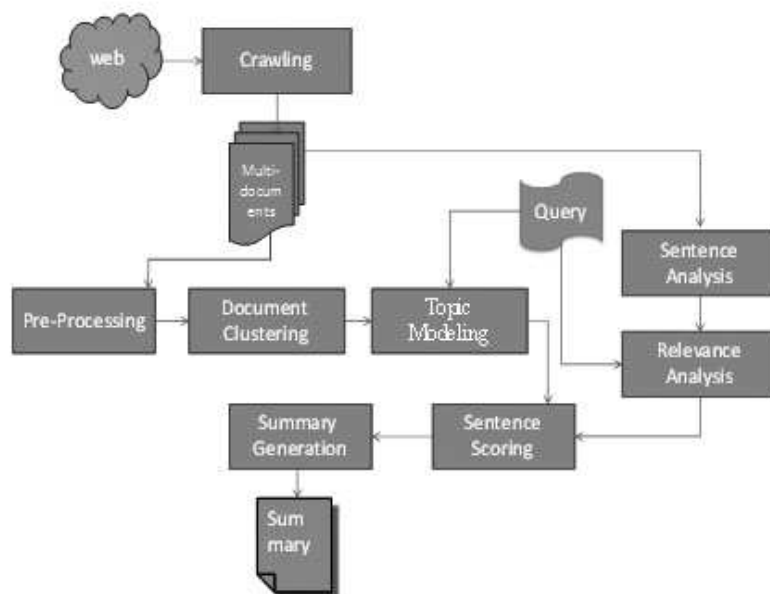
## II. Literature Survey

Summarization approaches can be broadly divided into extractive and abstractive. A commonly used approach namely extractive approach was statistics-based sentence extraction. Statistical and linguistic features used in sentence extraction include frequent keywords, title keywords, cue phrases, sentence position, sentence length, and so on [3]. Cohesive links such as lexical chain, co-reference and word co-occurrence are also used to extract internally linked sentences and thus increase the cohesion of the summaries [2, 3]. Though extractive approaches are easy to implement, the drawback is that the resulting summaries often contain redundancy and lack cohesion and coherence. Maximal Marginal Relevance (MMR) metric [4] was used to minimize the redundancy and maximize the diversity among the extracted text passages (i.e. phrases, sentences, segments, or paragraphs).

There are several approaches used for summarizing multiple news articles. The main approaches include sentence extraction, template-based information extraction, and identification of similarities and differences among documents. Fisher et al [6] have used a range of word distribution statistics as features for supervised approach. In [5], qLDA model is used to simultaneously model the documents and the query. And based on the modeling results, they proposed an affinity propagation to automatically identify the key sentences from documents.

## III. System Design

The overall system architecture is shown in the Fig. 1. The inputs to the multi-document summarization system are multi-documents which are crawled based on the urls given and the output given by the system is a summary of multiple documents.



*Fig. 1 System Overview*

### System Description

The description of each of the step is discussed in the following sections. The architecture of our system is as shown in Fig. 1.



## 1. Pre-processing

Pre-processing of documents involves removal of stop words and calculation of Term Frequency-Inverse Document Frequency. Each document is represented as feature vector, (ie.,) terms followed by the frequency. As shown in Fig. 1, the multi-documents are given as input for pre-processing, the documents are tokenized and the stop words are removed by having stop-word lists in a file. The relative importance of the word in the document is given by

$$Tfidf(w) = tf(w) * (\log(N) / df(w)) \quad - (1)$$

where,  $tf(w)$  – Term frequency (no. of word occurrences in a document)

$df(w)$  – Document frequency (no. of documents containing the word)

$N$  – No. of all documents

## 2. Document clustering

The pre-processed documents are given as input for clustering. By applying the k-means algorithm, the documents are clustered for the given k-value, and the output is the cluster of documents containing the clusters like cricket, football, tennis, etc., if the documents are taken from the sports domain.

## 3. Topic modeling

Topic models provide a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together. In this paper, Latent Dirichlet Allocation is used for discovering topics that occur in the document set. Basic Idea- Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

### Sentence analysis

Multi-documents are split into sentences for analysis. It involves tagging of documents by extracting the keywords, named-entities and the date for each document. Summary generation in different perspectives can be done from the tagged document.

### Query and Relevance analysis

The semantics of the query is found using Tamil Word Net. The relevant documents for the given query are retrieved. The relevance between the sentences and the query is calculated by measuring their similarity.

### Query-oriented Topic modeling

In this paper, both topic modeling and entity modeling is combined [3]. Based on the query, the topic modeling is done by using Latent Dirichlet Allocation (LDA) algorithm. Query is given as prior to the LDA and hence topic modeling is done along with the query terms. Query may be topic or named-entity along with date i.e. certain period of time.

#### 4. Sentence scoring

The relevant sentences are scored based on the topic modeling. For each cluster, the sum of the word's score on each topic is calculated, the sentence with the word/topic of high probability are scored higher. This is done by using the cluster-topic distribution and the topic-word distribution which is the result of the Latent Dirichlet Allocation.

#### 5. Summary generation

Summary generation involves the following two steps

##### 5.1 Redundancy elimination

The sentences which are redundant are eliminated by using Maximal Marginal Relevance (MMR) technique. The use of MMR model is to have high relevance of the summary to the document topic, while keeping redundancy in the summary low.

##### 5.2 Sentence ranking and ordering

Sentence ranking is done based on the score from the results of topic modeling. Coherence of the summary is obtained by ordering the information in different documents. Ordering is done based on the temporal data i.e. by the document id and the order in which the sentences occur in the document set.

### IV. Results

Table 1 shows the topic distribution with number of topics as 5, the distribution includes the word, count, probability and z value. The topic distribution is for each cluster.

Table 1 Topic model with number of topics as 5

TOPIC 0 (total count=1061)

| WORD ID | WORD     | COUNT | PROB  | Z   |
|---------|----------|-------|-------|-----|
| 645     | அயோத்தி  | 42    | 0.038 | 5.7 |
| 2806    | தீர்ப்பு | 38    | 0.035 | 5.4 |
| 2134    | இந்து    | 36    | 0.033 | 5.2 |
| 589     | அமைப்பு  | 32    | 0.029 | 4.8 |
| 1417    | அரசு     | 27    | 0.025 | 2.9 |
| 2371    | லக்னோ    | 27    | 0.025 | 4.5 |
| ...     |          |       |       |     |

### V. Conclusion and Future Work

In this paper, a system is proposed to generate summary for a query from the multi-documents using Latent Dirichlet Allocation. The multi-documents are pre-processed, clustered using k-means

algorithm. Topic modeling is done by using Latent Dirichlet Allocation. The relevant sentences are retrieved according to the query, by finding the similarity between the sentences and the query. Sentences are scored based on the topic modeling. Redundancy removal is done using MMR approach.

Topic modeling can be extended to find the relationship between the entities, i.e. the topics associated with the entity as a future work.

## References

- Arora.R and Ravindran.B, "Latent dirichlet allocation based multi-document summarization". In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, 2008, pp. 91-97.
- Azzam.S, Humphrey.K and Gaizauskas.R, "Using coreference chains for text summarization". In *Proceedings of the ACL Workshop on Coreference and its Applications*, 1999, pp. 77-84
- Barzilay.R and Elhadad.M, "Using lexical chains for text summarization". In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, 1997, pp.10-18.
- Carbonell.J, and Goldstein.J. "The use of MMR, diversity-based reranking for reordering documents and producing summaries". In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 24-28 August, Melbourne, Australia, 1998, pp. 335-336.
- Dwei Chen, Jie Tang, Limin Yao, Juanzi Li, and Lizhu Zhou. Query-Focused Summarization by Combining Topic Model and Affinity Propagation. In *Proceedings of Asia-Pacific Web Conference and Web-Age Information Management (APWEB-WAIM'09)*, 2009, pp. 174-185.
- Fisher. S and Roark.B, "Query-focused summarization by supervised sentence ranking and skewed word distributions," *Proceedings of the Document Understanding Workshop (DUC'06)*, New York, USA, 2006 pp.8-9.

# **Language Ideology, Inscriptions, Spoken Tamil and Technology**

(மொழிக் கொள்கை, கல்வெட்டுத் தமிழ்,  
பேச்சுத் தமிழ் – தொழில் நுட்பத்தின் பங்கு)



# Mapping Language Change in Tamil:

## Corpus analysis and Computer Database Making

*Appasamy Murugaiyan*

EPHE- UMR 7528 Mondes iranien et indien, Paris

### Introduction

In this paper, I would like to illustrate how 'corpus analysis' can help us mapping the process of language change and language use. The vast Tamil language corpora, dating more than two millennia, make it very envious to any outsider. One of the most challenging debates that we all have been witnessing in the last fifty years or so in the Tamil milieu have to do with Tamil LANGUAGE. However, so many questions arise related to, the language of Tamil Sangam anthologies, language of inscriptions, language of bhakti literature and then the emergence of modern Tamil. What is the relationship between these different varieties of Tamil? How they distinguish from each other? How to analyse and understand the grammatical structure of each of the genre of Tamil varieties- are some of the issues that deserve a definitive answer. The problem is that most of the studies that have been done so far give a vague idea mostly based on what may be called a 'native speakers' intuition. Yet, a corpus based empirical picture is due to be made.

### 1. Corpus analysis and Computational methodology in Tamil:

Mapping language change has been a major concern of corpus linguists. Historical corpora provide evidences for language change in many ways. Such a corpus-based study would not only enrich the history of Tamil language but also would contribute to the theoretical models for language change. I will illustrate the importance of the corpora based empirical studies in mapping language change and use in Tamil primarily based on few citations (Murugaiyan 1993, 2004 and 2011).

The field of corpus linguistics has proved beyond doubt that the corpus is a fundamental tool for any type of research on language and particularly if we want to find answers to some relevant issues on the language use and language change. At the outset, it is important to make a distinction between computational methodology for linguistic analysis and computational linguistics (CL). CL is concerned by fields like Artificial Intelligence and aims at developing formal models based on aspects of human cognition and implement them as computer programmes. Our concern here is how computer can be used in linguistic analysis on a specific Tamil corpus through a series of research questions. Computer-based methodology allows us 1) to work on vast corpora, which would otherwise be impossible and 2) to seek answers to many questions on different diachronic and synchronic aspects of Tamil linguistics. In many instances, CL relies on the results of corpus linguistic analysis using computer methodologies.

The paper is structured as follows: §2 introduces four / three questions from Tamil linguistics, §3 surveys the currently existing and or used POS Tag for (Indian languages) Tamil and raises in §4 the

question of structure of the database and the type of corpus and §5 concludes on a Tamil inscription database.

## **2. Few issues in Tamil linguistics:**

I will illustrate the importance of the corpora based empirical studies in mapping language change and use in Tamil primarily based on few citations (Murugaiyan 1994, 2004 and 2011). Of which the first two have to do with the historical linguistics of Tamil and the last two with the modern Tamil linguistics.

### **2.1. Word order variation in inscriptional Tamil**

Languages vary widely in many ways, including their canonical word order. It is known that some word orders are much more common than others are. Change or variation in word order type is one of the most important areas in the study of historical linguistics and language change.

We can roughly identify two different views on the word order in Old Tamil: (1) SOV is considered as the basic order and 2) view suggests a free word order. However, scholars have overtly recognized variation in the SOV order but have not made a detailed attempt to discuss this variation (Zvelebil 1967:71, 1997:43), except Herring (Herring 2001). Nevertheless, a corpus based empirical analysis of inscriptional Tamil, conducted by me, shows a third possibility: the constituent order is neither free nor strictly of SOV type and the variation in constituent order is motivated by pragmatic factors (Murugaiyan A 2011). In other words, the position of different constituents in a sentence is conditioned by information structure and many other contextual factors. In such cases, the word order is not used to encode grammatical relations within a sentence (like -subject vs. object-). For this pilot survey, I analysed a total number of 35 Hero stone inscriptions dating from 450 to 650 CE.

### **2.2. Experiencer (or Dative subject) constructions in Sangam corpus**

In most of the Modern South Asian languages, the verbs of physical, psychological and cognitive processes, known also as affective verbs, mark the principal actant (experiencer) in dative, accusative or genitive cases. This type of constructions known as dative subject constructions (or oblique-experiencer) since 1960's is considered as a major feature of the Indian linguistic area (Masica C. 1976). During the last three decades, there have been numerous studies on this major areal feature, which is shared by Indo Aryan, Dravidian, Munda and Tibeto-Burman family of languages (Verma and Mohanan 1990, Shibatani and Pardeshi 2001, Bhaskararao Peri, 2001).

Despite the great attention paid to experiencer constructions, we are yet to know about their origin and their diachronic developments in the different language families of South Asia. It is still commonly suggested that it has spread from the Dravidian family of languages. In order to bring in new data on the historical development of experiencer constructions in the South Asian typological and areal linguistics, I worked on a small corpus of Sangam literature. My survey showed that the dative experiencer construction is absent in classical Tamil (Murugaiyan, A. 2004) even though it is widely attested in modern Tamil (*yā viyartta* in Sangam, compared to *eṭṭakku viyarkki* in modern Tamil).

### 2.3. Non-dative dative subjects

We notice in Modern Tamil, and in some of the Dravidian languages, the experiencer or the dative subject is not always marked in dative case. Some notions like *virumba* 'to desire', *porāmai* '(be) jealous' depict constructions that range from a typical nominative-accusative structure (N1<sub>Nom</sub> - N2<sub>Acc</sub>) to nominative-oblique (N1<sub>Nom</sub> - N2<sub>Obl</sub>) structure. This type of Nominative and Accusative case marking represents higher degree of transitivity and is characteristic of agent role but not that of experiencer.

The major question here is how to define or identify a "(proto) typical experiencer" (for instance, absence of volition -control -contact with the second participant (effectiveness) and capable of receiving physical, psychological, cognitive experiences). Experiencer is a scalar notion like transitivity or agent. According to the syntactic structure of the construction, it seems possible to identify different kinds of experiencer like "agent-like", "patient-like" and "typical experiencer". The morphological case marking of the first and second participants (arguments) of these constructions correlates to some extent with the 'variation' in the semantic role of the experiencer.

### 2.4. Differential Object Marking and the accusative case in Tamil

Tamil is a morphologically rich language and many of the major grammatical functions are marked by case markers. For example, in Tamil the direct object is marked in accusative case. In general, it is held both by the traditional and modern Tamil grammars that the accusative marker "ai" is present obligatorily with animate nouns and is not obligatory with inanimate nouns. The case marking of direct objects is typologically one of the major concerns in the theoretical setting of Differential Object Marking (DOM) and goes beyond the animate / inanimate dichotomy. A number of studies on Indian languages have shown that several parameters like humanness, definiteness, individuation and affectedness of the noun on the one hand and the semantics of the verb on the other, contribute to the marking or non-marking of the direct objects (A. Murugaiyan 1993).

The four cases mentioned above raise few interesting questions. How these morphosyntactic, syntactic and semantic features in Tamil can be successfully encoded in a computer-aided corpora analysis? What would be the adequate standard of data annotation for Tamil? The point to be remembered here is that our objective is to map language change and use but not machine learning.

## 3. Parts-of-Speech Tags for Tamil and Indic languages

Over the past decades, large number of electronic language data have been created and annotated in the area of NLP for Indian languages. The POS Tag set and the construction of database are closely related to the objectives fixed by the researcher.

Several IL POS (Indian languages Parts of Speech) tagsets are often designed by a number of research groups working on Indian languages. The existing tagsets differ considerably from one another as they have been motivated by specific research agenda. They follow in general the PENN Tree Bank model and or the recommendations made by EAGLES. The number of tagsets varies from 34 up to 64 and 123 (Evaluating SKT tagsets). They differ considerably in terms of the selection of morphosyntactic categories and the scale of granularity. These different works, in course of time, have led to the creation of standard POS Tag set for Indian languages based on the EAGLES model. The proposal made by Baskaran et al. (2008) aims at fixing standard coding frame for Indian languages in



general. This model is expected to “capturing the shared linguistic structures in a methodical way [...] and ensure cross-linguistic compatibility”. This is no doubt part of computational linguistics focusing on machine learning.

This being so, in computer aided linguistic analysis we face different challenges. In most cases, we need language specific, fine-grained multi level frame. The picture becomes still more complex while we study a specific aspect of language change and use (cf. §2) on a large corpus. We need to decide, among other things, setting up various types of corpora: synchronic and diachronic on the one hand and according to the type of literature under scrutiny (for example Sangam anthology, Bhakti literature, Tamil inscriptions, Modern Tamil and so on).

#### 4. Linguistic corpora and Database Making

If we wish to answer empirically to questions raised in §2 above and similar others, the corpus based study is a fundamental requisite. The first phase of corpus creation is data entry, which involves rendering the text in electronic form. The Sangam corpus is electronically available in standard format. On the contrary, the inscriptional corpus is not available electronically and the printed versions cannot be digitized via OCR. We still need to standardize the complex orthographic rules and the different scripts used in the inscriptions. The whole corpus of Tamil inscriptions has to be captured manually.

The corpus annotation as we have seen early is language oriented and should reflect the type of questions we are addressing in the corpora. Now turning back to the sample questions raised in §2, it is important to know precisely what information should we need to encode in a linguistic database? Should a coarse tag set be enough or should we need detailed set of information at morphosyntactic, syntactic, semantic and discourse levels.

For instance, to account for the word order variation and its correlation with pragmatics or discourse structure, one should include pragmatic information in the corpus annotation and the type of word order for each ‘sentence’.

As for experiencer (dative subject) constructions in Classical Tamil, a large variety of lexical and grammatical devices are used to express physical, biological, cognitive and mental perceptions in classical Tamil. In the following examples, we have a compound verb composed of a noun and support or light verb.

|                                     |             |             |
|-------------------------------------|-------------|-------------|
| <i>pasi + pa□u</i> (hunger+endure)  | feel hungry | [pu□.260.6] |
| <i>pasi+kūra</i> (hunger+grow)      | hungry grow | [na□.29.3]  |
| <i>pasi+te□u</i> (hunger+burn).     | Hungry burn | [aka.291.3] |
| <i>añar u□a</i> (suffering+to have) | to suffer   | [ku□.76.6]  |

The database should on the one hand provide all lexical bases, both noun and verb, and on the other hand all light verbs and the syntactic and semantic rules of combination that licence the instantiation of the predication. A large historical corpus should be able to trace the evolution from ‘nominative’ to ‘dative’ structure.

The case marking variation noticed in dative subject constructions in modern Tamil is due to the lexical and semantic nature of the experiencer predicates:

- 1) *nā* *Sandōśa-p* *pa* *ugi* *ē* 'I am happy'  
 I joy feel.pres.1s
- 2) \**nā* *magi* *cci* *pa* *ugi* *ē* 'I am happy' (not accepted)  
 I joy feel.pres.1s
- 3) *vīra* *semmāttiyai* *virumbupiki* *ā*  
 pn.nom pn.acc like.pres.3m.s  
 Viran likes Semmathi
- 4) *vīra* *ukku* *semmāttimīdu* *viruppam*  
 pn.dat pn.loc liking

Viran like Semmathi

In (1) & (2) the two lexical items with almost identical meaning do not fall under the word formation rule. In (1) the noun *sandōśam* is an Indo-Aryan loan word. In (3) and (4), the predicates are of two different grammatical categories. The finite verbal predicate in (3) marks the direct object in accusative case and shows agreement with experiencer. On the contrary, in (4) with a nominal predicate, the direct object is marked in locative and the experiencer is marked in dative case.

These predicates, nominal and verbal, are divided into three groups: cognitive, psychological and physiological. There is a close correlation between the experiencer construction and, among other things, (1) the grammatical category of the predicate and (2) the semantic nature of the predicate. Even though we notice some generalisation between the semantic role and the morphological case marking, this relation is a language specific phenomenon. The point here is that at what level we incorporate the various lexico-semantic features in the corpus annotation.

Finally, the differential object marking is noticed in Tamil and in many other languages. In order to account for the marking or non-marking of the direct object in accusative case we have to distinguish several features: human / non-human, definite / indefinite, referential / non-referential, degree of affectedness of the direct object on the one hand and the semantics of the verb on the other. In examples (5) and (6) the mere presence or absence of the accusative case change completely the meaning. In (6), the object, 'loan' known as object of creation, exists prior to the utterance of the sentence. On the contrary, in (5), the speaker is trying to contract a loan, and it does not exist really in the discourse context.

- 5) *avari* *am* *ka* *a* *kē* *ēn* *sa* *aikku* *vandu* *vittār*  
 he.locative loan ask.past.1.s quarrel.dative come.adp.aux.past.3.s  
 I requested him to lend me some money, but he started to pick on me
- 6) *avari* *am* *ka* *a* *ai* *kē* *ēn* *sa* *aikku* *vandu* *vittār*  
 he.locative loan.acc ask.past.1.s quarrel.dative come.adp.aux.past.3.s  
 I asked him to repay my money back, but he started to pick on me

## 5. Conclusion

As mentioned earlier, nature of database and the inventory of tagsets depend on the type of research agenda. This type of multi-layered analysis on large corpora is possible only with computer-aided methodology. The four cases mentioned in §2 and the few examples given in §4 are a few among

hundreds of other questions that we have to account for in Tamil. Most of the existing POS tag sets do not provide us with a fine-grained annotation scheme. But in compute-aided corpora-based linguistic research, a fine-grained annotation paradigm is essential. In my experimental database on Tamil Inscriptions, I have about 120 tagsets. Three types of rule-based annotations- morphosyntactic, syntactic and semantic- are done manually. I have aimed at a fine-grained analysis and so I have opted for a high number of tags. I have also included information like word order types, verbal valency and other minute details that would help to map different changes at morphological, syntactic and semantic levels. I am using, for each sentence the interlinear glossed text (IGT) format, which includes source language text, a morpheme-by-morpheme gloss, and a translation into French or English.

The ongoing research is to illustrate how linguistic corpora can be used as readily available evidence for mapping language development and language variation in time and space. A huge computerized historical corpus would certainly allow a comparative view of the Tamil language at different moments in the history. These data would help us not only to capture different stages of linguistic developments but will also help to test modern theories about variation and change. The construction of huge electronic corpora in Tamil presents many constraints related to linguistic theories. However, a close collaboration with computational specialists would certainly help to overcome many of the shortcomings and certainly would enhance computational methodology for corpus-based linguistic analysis in Tamil.

## References

- Baskaran Sankaran, Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha, and K. V. Subbarao, A Common Parts-of-Speech Tagset Framework for Indian Languages, Proceedings of LREC 2008.
- Ide, N., L. Romary, et al. (2003). International standard for a linguistic annotation framework. Proceedings of the HLT-NAACL 2003 workshop on Software engineering and architecture of language technology systems - Volume 8, Association for Computational Linguistics: 25-30.
- Madhav Gopal, Diwakar Mishra, and Devi Priyanka Singh, (2010) Evaluatin Tag set for Sanskrit, Sanskrit Computational Linguistics, LNCS.
- Murugaiyan, A (1993) Marquage différentiel de l'objet et variation actancielle en tamoul, CNRS-Actances n° 7, p. 161-183.
- Murugaiyan, A (1999) « De l'agent affecté à l'expérient en tamoul », *Cahiers de linguistique de l'INALCO* n° 1/2, INALCO, Paris, p. 147-160.
- Murugaiyan, A (2004) Note sur les prédications expérientielles en tamoul classique, *Bulletin de la Société de Linguistique de Paris*, 99, p. 363-382.
- Murugaiyan, A (Forthcoming) Identifying Basic Constituent Order in Old Tamil: Issues in historical linguistics with Special Reference to Tamil Epigraphic texts (400-650 CE), Proceedings of the World Classical Tamil Conference, 2010 July, India.
- Vasu, R. (2001) Development of Morphological tagger for Tamil, INFITT, Kuala Lumpur.

# Language Ideology and Technology

*E. Annamalai*

*The University of Chicago*

Technology, which is a device that increases the ratio of output to the input, works for the language also as it does for any human production. The first technology applied to language was writing invented around 3500 BCE in Mesopotamia (modern Iraq). It made possible to enormously increase the reach of the language –i.e. the content in the language– in space and time. It also provided relative stability to language, which reduced variation. These added characteristics of the written language were necessary for trade, which was the cultural context for the emergence of writing. The idea of the language changed from being a tool of cooperative communication to that of record keeping to eliminate mistrust. From this, the ideology emerged that the written word is more trust worthy than the spoken.

The next major technology is the invention of woodblock printing in China developed to print language on paper around 650 CE. Its improvement into movable metal types around 1440 CE in Germany made copying of printed texts possible at a low cost and in less time, which in turn made their reach increase exponentially in space, physical and social. This ushered in the era of print capitalism (Anderson 1991), one aspect of which was production and dissemination of language materials as commodities of the market. The printed texts became the private property of printing establishments or individual authors. The content and the particular form of the language in which it is coded came to be owned by the producers of it while the abstract language remained the public domain. When the written language through print lent itself to be controlled by the institutions of the society such as school, media, courts, it was possible to shape the language according to their ideologies such as purism, precision (as in law or science).

When the audio recording and replay technology was improved to use electrical and magnetic devices in the early decades of the 20<sup>th</sup> century, the added characteristics of the written and printed language could be transferred to the spoken language. This did not bring about any new ideology to the language except that the ideologies such as purism and precision were difficult to carry out in the spoken language, as it was not controlled by societal institutions through their system of rewarding the adherents of their language ideology.

The latest in technology that bears on language is the digital technology, which equally applies to the spoken and the written language. While the earlier technologies moved from their use for non-linguistic content such as pictures and music to language, the digital technology moved from numbers to language. While writing and printing technologies represent the language unit, viz. the letters, directly, the digital technology converts language units, viz. letters and phones, into digits for processing. Digital technology gives enormous scope for editing while composing. The implication of this is that the characteristic of finality and permanence, which the earlier technologies gave to texts, turns out to be fragile in this technology. Digital technology is useful not only for composing new

language content, but also for copying and storing the old. It reduces drastically the time and cost for transmitting the materials globally. It takes away the control of language from social institutions and gives it back to individuals.

What this new technology is doing and can do to Tamil? The writing technology gave new documentation register and literary code to Tamil, as seen in the emergence of Sangam literature and contemporaneous Tamil Brahmi inscriptions of record. The printing technology stabilized the Tamil alphabet both in their graphic form (which was evolving over 2000 years) and number (the five grantha letters were added and three contextually dependent letters (caarpezuttu) were dropped with the advent of this technology). The digital technology has brought, and is capable of bringing more, new effects on Tamil. Decontrolling of Tamil brings the written Tamil closer to the spoken Tamil with regard to the effects and gives legitimacy to this convergence. The societal institutions lose their commanding role in shaping Tamil. As anyone with a vocal cord could speak at will, any one with literacy skill and access to digital technology could write at will. Anyone can be a writer without vetting by a teacher or an editor. The spontaneous writing could be more effective on language than spontaneous speaking because the language as written at will by any and all individuals gets into the public domain accessible to any, unlike the language spoken.

It follows that there will be more language ideologies at play in shaping Tamil, which are subscribed to by the individuals; they will not be just the ones promoted, and penalized for non-compliance, by the elite in control of the societal institutions. Purism ideology is also prevalent among the individual practitioners of the new technology. Purism of Tamil includes elimination of loan words and acquired letters, avoidance of spoken forms in vocabulary and grammar and, to a lesser extent, reclaim of historically antedated grammatical constructions. This ideology is also enforced using the same digital technology on the writings of others with a different ideology in communally created content such as Wikipedia. Nevertheless, the multiplicity of ideologies in shaping Tamil cannot be excommunicated from this technology.

Any technology is not in itself and by itself an aid to modernize a language. It depends on its user, who decides what it is used for. The print technology helped not only the use of prose as a language of literature, but also helped recoding of the oral literature in the visual medium on paper. The books of folk literature printed exceeded numerically the books of modern fiction and poetry (Blackburn 2005) when print technology came into being in Tamil Nadu. So is the religious literature compared to the secular literature. The digital technology used for astrological predictions is in demand as it is for weather predictions. While technology cannot be appropriated for what is valued as modern, its potential for this task should not be wasted away.

An ideology that digital technology is capable of implementing on Tamil is parity of written and spoken Tamil and reduction of distance between them. I shall not go into discussing here the rationale for this ideology and its importance for the survival of Tamil as a modern language with vitality (Annamalai 2011a). There are many areas to implement this ideology. I shall mention some of them, many of which relate to teaching Tamil to learners of different backgrounds as to their exposure to Tamil before, during and after learning.

Inclusion of spoken Tamil in Tamil pedagogy is on the increase. It is an important part in the Tamil teaching programs designed for non-native adults outside Tamil Nadu and Jaffna, especially in the Universities in the U.S., where the goal of Tamil learning is as much oral communication as it is philological and literary inquiry. This extends to children of Tamil ethnicity, who have lost their heritage language and want to revive it, as in Mauritius. Their interest is communicative within the community as well as cultural and political for reasons of identity. The traditional Tamil language curriculum in schools under the rubric of learning the mother tongue as a minority in countries like Singapore focuses on literacy skills and introduction to literature. This is being modified in Singapore to include speaking of Tamil within the curriculum in order to make Tamil more relevant for students in their lives (Seethalakshmi). Like these students, the second generation Tamils of post-colonial diaspora has limited listening comprehension in Tamil and they want to add spoken skill to their Tamil competence. Even in a curriculum that focuses on the reading skill, reading modern fiction and magazines will be hard without the knowledge of spoken Tamil, where conversations between characters, jokes etc are written in this variety.

The basic need when spoken Tamil becomes part of the Tamil curriculum is identification of standard spoken Tamil, which is spoken in inter-dialect communicative situations by the schooled. To identify it empirically, we need a searchable database of spoken Tamil. Such a database could be built relatively easily using digital technology from dialogues in movies and television shows. This database is a necessary, if not sufficient, tool to compile the grammatical, lexical, semantic and phonetic parameters of the standard spoken Tamil. The data needs to be processed to sift out dialect and formal features, which are mixed up in the standard speech as well as the mixing of English in it; the phonetic data needs to be brushed up to upgrade the non-standard pronunciation found commonly in the public programs used to build the database. It should be possible to write algorithms to do these jobs mechanically. Pedagogy requires not speech as it, but as standardized. I shall not go here into discussing what the standardized spoken Tamil is (Annamalai 2011b).

The standard spoken Tamil data needs to be transcribed in Tamil script. The publically available Google tool for speech recognition and instant transcription needs to be improved substantially for Tamil. When this is developed, it should be possible to go into making popular, inexpensive tools that instantly convert the utterance of a student into a written sentence, and conversely a written sentence of spoken Tamil into an utterance. This will help the student recognize speech visually and aurally, which facilitates the recall of the language in the learning situation.

Converting speech into writing assumes a standard spelling system of the standard spoken Tamil. This is yet to be developed by linguists to be used in the pedagogical context to begin with. A fundamental principle of any spelling system is that it is not an exact phonetic transcription of speech, but a convention of writing from which speaking could be deduced with straightforward rules of correspondence. The spelling system of every language has such rules. In the case of Tamil, an additional requirement is that the spelling system of the spoken Tamil also serves as the base for the student to migrate to the conventional spelling of the written Tamil with straightforward rules. No such spelling system exists now. The way the spoken Tamil written is notoriously inconsistent between authors and in the same author at different places. The spelling system used by on Tamil teacher is consistent but it is not the same across teachers, though each is relatable. The spelling

system for spoken Tamil needs to be standardized urgently. I have developed a system as the starting point to initiate this process (Annamalai 2009). If the standardized spelling necessitates some additional letters in the Tamil alphabet, we need to get cultural acceptance of it. Then they will need to be provided in Unicode.

The other need is to relate the spoken and written Tamil at the lexical level with regard to the spelling of words. This should be bidirectional from the spoken to the written and from the written to the spoken. This tool will reduce the distance between the two by facilitating migration between the two. There are now tables available to relate the written to the spoken with general rules of deletion and change. They need to be fine tuned to cover complex relations, variations and exceptions and then algorithms are to be written so that one can get the spoken form from the written on any device, hand held or desk top, just like checking the meaning of a word. The converse of relating the spoken with the written is more difficult primarily because of the fact that one form of the spoken will relate to two of the written. A solution to this problem has to be found such as using the absence of one of the two possible forms in a built-in dictionary to reject it.

There is no dictionary of standard spoken Tamil available now. It is possible to compile one using its standardized spelling for the head entries followed by the written word in conventional spelling with gloss in English for learners who are not proficient in Tamil. The converse of it would be to have parallel head entries first in conventional spelling followed by spoken spelling. It is possible to produce mechanically one version of the dictionary from the other and rearrange the entries alphabetically. An alternative to this has the additional advantage of using the spoken input when referring to a dictionary. In this, a spoken word keyed in in its spelling will first identify its written equivalent, which will lead to its meaning in the digital dictionary. It should also be possible to speak the word a student, who is not competent in writing Tamil, hears an unknown word when watching a video or touring Tamil Nadu and wants to know its meaning . A digital dictionary should be able to point out the meaning either directly from the spoken cue or through the written word coded in the dictionary by automatically linking the spoken word to the written form.

Technology is available to do the above things if the language ideology with regard to spoken Tamil mentioned above is a driving force. Using digital technology for the spoken Tamil is not just for doing research on it. It is also for teaching Tamil effectively and attractively. The hope is that it will attract more students to learn Tamil, who are now put off because of the perceived difficulty of learning separately the two varieties of Tamil. It is the desire of all of us to mitigate the difficulties of learning Tamil by the younger generation of Tamils and also the non-native speakers of Tamil

## References

- Anderson, Benedict. 1991. *Imagined Communities: Origin and Spread of Nationalism*. Verso: London. Revised edition.
- Annamalai, E. 2011a. Diglossic Convergence. In Annamalai, E. *Social Dimensions of Modern Tamil*. CreA: Chennai

- Annamalai, E. 2011b. Standardization of Speech. In Annamalai, E. *Social Dimensions of Modern Tamil*. CreA: Chennai
- Annamalai, E. 2009. Spelling System for Standard Spoken Tamil. <http://www.crea.in/esources.html>
- Blackburn, Stuart. 2005. *Print, Folklore and Nationalism in Colonial South India*. New Delhi: Permanent Black
- Seethalakshmi and Vanitamani Saravanan. 2009. தரமான பேச்சுத்தமிழும் ஆசிரியவியலும் . Singapore: Author (at National Institute of Education)



# Tamil: A Family of Languages

*Harold Schiffman*

*South Asia Studies. University of Pennsylvania*

## Abstract

As is well-known, the Tamil language is one of the oldest languages on the Indian subcontinent, dating from the early centuries of the Common Era, if not before. It is commonly divided into a number of 'stages', beginning with the earliest period (Sangam Tamil), followed by Medieval Tamil (or Middle Tamil), and then the Modern Literary Language, which dates from about the 13<sup>th</sup> century. I would also add a fourth stage, that represented by the spoken language of today, which differs sometimes quite radically from its written form. This presents a formidable challenge to non-Tamils who wish to learn to both read and speak Tamil, since Tamil society offers little help to those wishing to speak, even though most authentic communication between live speakers goes on in Spoken Tamil, and learners who wish to learn something about Tamil culture will not get far without a knowledge of the spoken language. Ignoring the modern spoken language also hides the tremendous diversity among dialects of Tamil, especially those that differ radically from what I call "Standard Spoken Tamil" or SST (Schiffman 1998). Sri Lanka Tamil is one of those dialects that are not mutually intelligible to many other speakers. For this reason, I propose that we should cease treating Tamil as one language, and begin to think of Tamil as a *family of languages*, related of course through history, from the oldest stages to the most modern. In historical linguistics terms, we would treat the oldest stage as Proto-Tamil, and later stages as 'daughter languages' or even 'granddaughters'.

## The Accessibility Problem

Foreigners who wish to learn Tamil are confronted with enormous challenges. Tamil culture tends to value the study of Classical Tamil and its 'daughter' languages (Medieval Tamil, modern Literary Tamil) but not its 'grand-daughters' i.e. the spoken dialects used by all Tamils for most of their interpersonal communication. Foreigners who attempt to learn spoken Tamil are discouraged from doing so in various ways:

- *scolding* the learner for 'corrupting' the language
- 'correcting' the spoken form by repeating the LT form
- Ridiculing the learner by laughing at him/her for using spoken forms

As an example of the first strategy, when I was doing research on spoken Tamil in my first visit to India in 1965-66, students who were influenced by DMK came to me and asked me to cease and desist from studying spoken Tamil, because it 'contributed to the downgrading' of the language. An example of the third type, ridicule, happened to me while passing through customs from Singapore into Malaysia—I spoke Tamil to the customs agent, who had a Tamil name and looked to be of Indian descent. Her response was "*You talk just like my Granny!*"

## The Sociology of Language

One of the things that Tamils are famous for is their 'love of their language' which, however it can be measured, has got to be more intense than any other expression of 'language loyalty' found on the Indian subcontinent. More Tamils have died for their language than any other language group, and this intense loyalty has of course attracted attention by scholars who study other parts of India. I came to India to study Tamil syntax and write a dissertation on that subject, but I soon found myself being asked to comment on the Tamils' language loyalty, which of course in 1965 had resulted in various forms of extreme (and sometimes violent) resistance to the imposition of Hindi as the national language.

When asked to write something about this topic, I did so, but soon found myself inadequately prepared to approach this topic without preparation in a field that was far from what the discipline of Linguistics had prepared me for. Fortunately for me, I was drawn into what is known as the Sociology of Language by the appearance of a book on the German language in America (Kloss 1963), which drew me into the topic of my own linguistic heritage as an American of German descent, and with more reading, I decided to offer a course on the topic of 'Language Policy.' My experience with Tamil helped widen my approach to this topic, and I taught the course both at the University of Washington and the University of Pennsylvania for almost 35 years. I soon discovered that there was an extensive body of literature on this subject, including but not limited to the work of Fishman, Ferguson, Haugen, Hymes, and many others, which helped me to understand topics such as 'language loyalty' and to present them to students.

The study of language policy soon became my primary research interest, and because I had once concentrated in Slavic Linguistics and had visited the Soviet Union and gotten a taste of its linguistic diversity and its language policy, coupled with the fact that I had also lived in France for two years, which has its own kind of linguistic chauvinism, gave me a range of experience to deal with various kinds of language policy issues.

**What keeps a language alive?** The point I want to make about this is based on research about what keeps a language vital and alive, and what leads to language shift, and language death. Much has been written about what the effective methods for language vitality and preventing language death, but one of the most important claims for effectiveness was made by Fishman (1991) who proposed that 'intergenerational transfer' is the most crucial factor in keeping a language alive. That means that all other strategies, such as using the language in education, or recording the words of the last viable speaker, or any other strategy that may be proposed, are all *ineffective and useless*, unless **intergenerational transfer** takes place. For the case of Tamil (and in fact for any 'threatened' language), this means that Tamil must be learned and spoken *in the home*. If it is not spoken in the home, but only learned at school, it will not survive, and some other language, probably English, will replace it.

**The Case of Singapore.** One of the crucial cases that illustrates this problem most clearly is the case of Singapore. In Singapore, as is well known, Tamil is one of the 'official' languages and receives support from the educational system along with other 'mother tongues' of the Singapore population. But Tamil is increasingly not the language spoken at home in many Singapore Tamil families, and many researchers now fault the educational system, which until recently insisted on teaching only Literary Tamil, but giving no support to the spoken language. Without this, English takes over as the dominant language (or its Singapore variant, Singlish). In research I conducted in Singapore in 1994, Tamil students in the system revealed to me that they did not feel that Tamil was *their language*; it belonged to someone else, they said, and they saw no use in learning it, especially since it had no economic value in Singapore. They also voiced the complaint that no matter how hard they tried, they could never satisfy their teachers, who always faulted them on their poor knowledge of Tamil. This research is buttressed by research by others who have studied the Tamil system in Singapore, e.g. Gopinathan, Seethalakshmi, Saravanan, and others.

One might argue that Singapore is different from Tamil, and that in Tamilnadu, where Tamil is the ambient language, this is not a problem. But increasingly, I find that Indian Tamils who have been educated in English medium schools do not handle Literary Tamil well, and speak a kind of spoken Tamil that is heavily mixed with English.

What I am saying is that we make an error if we assume that supporting the study of *one* kind of Tamil, such as Classical Tamil or modern Literary Tamil, will solve all our problems, and keep Tamil alive. We need in fact to treat Tamil as a *family of languages*, and support all of the members of that family. The branch of the family that gets the least support typically is that of spoken Tamil and all its variants. This is the true 'mother tongue' of all Tamils, the one they learn first, the one in which their emotions are centered.

This notion is often ridiculed by Tamil experts, but in the discipline of Linguistics, we know that by the age of six, about the time when children start school and start to acquire literacy, they have solidified their knowledge of their mother tongue, the one they learn at home at their mother's knee. Our educational systems tend to assume that their only responsibility is to teach the literary language, but in fact the literary language will not be learned if the spoken language is not used as a *resource*, instead of treated as a *liability*. Educational systems that try to 'kill off' the spoken language (whether English, Tamil, French or any other language) will also probably kill the language that they are attempting to teach.

**Spoken Tamil as Resource.** Let me give an example from my own experience of teaching Tamil. One of the hardest things for learners of Tamil who have no home background in Tamil is the syntax of relative clauses in Tamil. In English, we can take the example of two sentences, that are combined using a relative pronoun such as 'that' or 'which' or 'who' and make one sentence:

1. That boy came yesterday.
2. I saw the boy.

→I saw the boy **WHO/THAT** came yesterday.

In Tamil, of course, this is done differently. There is no relative pronoun, but instead, the verb of one of the sentences is converted into an adjective, and placed before the co-referential noun:

1. anda payyan neettu vandaan.                      That boy came yesterday.
2. naan payyane paatteen.                              I saw that boy.

→naan **neettu vanda** payyane paatteen I saw the boy who came yesterday.

These are obviously two very different kinds of syntactic structures, and what I have found in my thirty years of teaching Tamil to both Americans with no background in Tamil, and some Tamils whose parents were born in India, is that Americans with no background have a very hard time with these structures—they have to be drilled over and over to master them, whereas students with a Tamil background at home have no problem with these sentences, either in spoken Tamil or in Literary Tamil.

In other words, knowledge of the spoken language is not only not useless, it is an asset, and needs to be built-upon, rather than exterminated.

### **So where shall we start?**

First of all, we need data in the form of a database of spoken Tamil materials, in order to be able to conduct some much-needed research on what is the most appropriate form of Tamil to teach. I suggest that we organize to create this database using the following sources:

- Tamil ‘social’ films. There are hundreds if not thousands of Tamil films whose soundtracks could be digitized, transliterated, and made searchable for examples of all kinds of spoken Tamil, but mostly what I call ‘standard’ Tamil. (Schiffman 1998)
- Tamil radio plays and television sitcoms also constitute a source for more spoken Tamil.
- **Linguistic Survey of India.** The LSI, compiled early in the last century, has transcription of spoken Tamil of various sorts, and recently gramophone records of some of these samples have been digitized and will soon be made available for study.
- The *English Dictionary of the Tamil Verb* has sound files for more than 9,700 spoken Tamil example sentences. These sentences could be easily adapted for use in a ‘matched-guise’ type study.
- **Field recordings.** Many researchers, myself included, have made tape recordings of Tamil speakers in various dialect areas, and could pool these resources to add to the data-base.

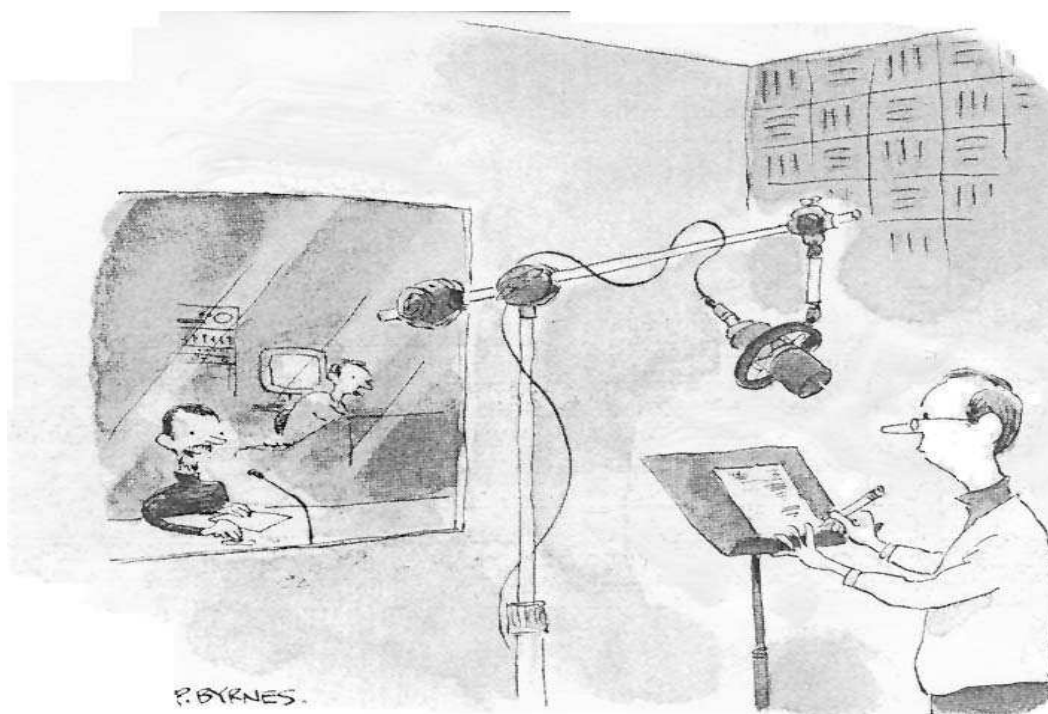
**The Matched Guise Test:** One of the effective ways to study spoken Tamil and discover how mother-tongue speakers conceive of various forms, including which examples of spoken Tamil constitute the most useful and ‘acceptable’ form to be used as models for students to imitate and learn, is the research methodology known as ‘Matched Guise’ testing. Matched guise tests originated in Canada in the 1960’s as a way to determine Canadians’ attitudes towards the ‘other’ language, i.e. attitudes of Anglo-Canadians towards French, and attitudes of Franco-Canadians towards the English of Anglo-Canadians. Over 100 such studies have been done on languages around the world, comparing the language attitudes of bilinguals and bidialectal people, and recently some studies have also been carried out in Singapore as a way to study attitudes of Tamils towards various forms of Tamil (Seethalakshmi et al. 2005, 2006). More needs to be done on this among Tamils in Tamilnadu, since attitudes toward Tamil among Tamils in Singapore seem to be different from those found in Tamilnadu.

The effectiveness of matched-guise testing comes from the fact that subjects are asked to evaluate, not the *language* of other speakers, but the *speaker* him/herself.

Matched-guise tests are constructed utilizing bidialectal or bilingual speakers, who are recorded speaking in each of their two variants. Then the recordings, typically of five bilinguals, are scrambled—they are mixed with those of other speakers, and played to subjects who are themselves bilingual/bidialectal. Typically, the subjects fail to recognize that the same speaker has been recorded twice, so when they hear the samples, they think they are hearing *ten* different speakers. Asked to judge the speakers on variables such as level of education, what kind of job they might have, what kind of earning power they might have, as well as other social variables, the subjects rate the ‘guise’ of one of the bilinguals more highly than the other guise.

In Canada, English ‘guises’ are ranked higher than French guises, even by French speakers; subjects even rank English speaker-guises as *taller*, although French guises are usually ranked as ‘more friendly.’ In all the matched-guise studies I have looked at, there is always a differential in these areas—there is never ‘equality’ of rank, despite any attempts in various societies to create social equality.

In case the notion that speakers can tell if someone is tall by listening to their voice, consider the following cartoon, which seems to assume that speakers can imagine all kinds of physical characteristics of other speakers by their voices!



***“Great! O.K., this time I want you to sound taller,  
and let me hear a little more hair.”***

## Conclusion

So who will volunteer to join me in this endeavor? There are a number of problematical issues we will need to deal with:

1. We will need to convince many people in the Tamil establishment that Spoken Tamil is something worth studying and collecting data for.
2. We will need to seek funding to support the collection of data, its transcription, and a method for accessing forms on-line. This will involve 'tagging' of forms, since Spoken Tamil is (in some ways) morphologically more complex and less transparent than Literary Tamil.
3. We will need to get legal permissions to copy the sound tracks of Tamil films, and other material that already exists 'out there.'

## Bibliography

- Ferguson, Charles A. (1959a) 'Diglossia'. *WORD*, 15(2): 325-40.
- Ferguson, Charles A. (1959b) Myths about Arabic. In Joshua Fishman (ed.) (1972) [1968] *Readings in the Sociology of Language*. Mouton, The Hague, pp. 375-81.
- Ferguson, Charles A. and Gumperz, John J. (eds) (1960) *Linguistic Diversity in South Asia: Studies in Regional, Social and Functional Variation*, vol. 26(3): part 2 of *International Journal of American Linguistics*.
- Fishman, Joshua (ed.) (1966) *Language Loyalty in the United States*. Mouton, The Hague.
- Fishman, Joshua (1967) 'Bilingualism with and without diglossia; diglossia with and without bilingualism'. *Journal of Social Issues*, 23(2): 29-38.
- Fishman, Joshua (ed.) (1972) [1968] *Readings in the Sociology of Language*. Mouton, The Hague.
- Fishman, Joshua. (1991). *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Clevedon: Multilingual Matters.
- Haugen, Einar (1956) *Bilingualism in the Americas: A Bibliography and Research Guide*. Alabama University Press, Montgomery.
- Haugen, Einar (1957) The semantics of Icelandic orientation. *WORD*, 13: 447-60.
- Hymes, Dell (ed.) (1964) *Language in Culture and Society*. Harper & Row, New York.
- Kloss, Heinz (1963) *Das Nationalitätenrecht der Vereinigten Staaten von Amerika*. Braumüller, Vienna.
- Saravanan, V., Seetha Lakshmi, S., and Caleon, I. "Attitudes towards Literary and Standard Spoken Tamil in Singapore." *The International Journal of Bilingual Education and Bilingualism* Vol. 10, No. 1, 2007
- Schiffman, Harold F. (1998) "Standardization and Restandardization: the case of Spoken Tamil." *Language in Society*, Vol. 27 (3) 359-385.

- Seetha Lakshmi, S. (2001) The contribution of the mass media to the development of Tamil language and literature in Singapore. Unpublished doctoral dissertation,
- National Institute of Education, Nanyang Technological University, Singapore.
- Seetha Lakshmi, S. (2005) Discussion on mother tongue issues in multilingual
- communities: Tamil language varieties Standard Spoken Tamil (SST). *SAAL Quarterly. SAAL 20th Anniversary Special Issue.*
- Seetha Lakshmi, S., Vaish, V. and Saravanan, V. (2006) A critical review of the Tamil
- language syllabus and recommendations for syllabus revision. (Technical Report)
- Institute of Education Singapore. CRP 36/03 SL.

**Website:** <http://ccat.sas.upenn.edu/~haroldfs/bibliogs/MACHGUIS.HTM>

# **Tamil Literature from Sangam to Modern Period: A Continuum with colorful changes: What does a search of the Tamil Electronic data reveal us?**

*Vasu Renganathan*

University of Pennsylvania (vasur@sas.upenn.edu)

## **Introduction**

A systematic study of the Tamil language from Sangam to Modern period from a historical perspective may reveal that there does exist a continuum of changes that occurred from one stage to another in Tamil language. Without such a study, any synchronic description of Tamil would only reflect its complexity in an overwhelming way. In other words, The Tamil language, the way it is now with a museum of complex forms, expressions and grammatical constructions, both in written and spoken variety, demonstrates a vast number of linguistic characteristics at phonological, morphological and syntactic levels, that require a comprehensive diachronic study to fully understand them in a coherence way. In this respect an extensive electronic database of Tamil texts from all of the stages along with a powerful query tool to search texts from various dimensions is indispensable. This paper is an attempt to illustrate how such an electronic database for Tamil (<http://www.thetamilanguage.com/sangam>) can be used extensively to study some of the morphological and syntactic behaviors of Tamil from a historical point of view.

Upon exploring the Tamil electronic database consisting of a variety of data ranging from the Sangam to Modern Tamil, especially by employing the principles of historical linguistics, one may immediately be able to notice that the changes that underwent throughout the history of Tamil language exhibit a systematic, regular and what one may attribute as a set of colorful changes in it. Phonological, morphological and syntactic changes that took place to this language one after another in a sequential manner contributed to the dearth of complexity as we see now as modern Tamil (both spoken and written) – a language that many have attempted to study it using many grammars and dictionaries in many different points of views! What may one illustrate it in a minuscule is that when words or combination of words and suffixes undergo all possible phonological rules on them, either successively at one point of time or periodically at different stages, what results is a set of the most complex forms that can be understood in terms of many dichotomies such as social versus regional dialect; spoken versus written variety; high versus low register; casual versus platform speech and so forth. Thus, attempting to learn this language that contains such a complex set of shades of variations does indeed pose a greater level of difficulty than normal for any second language learner. Not only does it become a big challenge to any second language learner in having to comprehend and use these multiple facets of this language, but it also becomes an immense task for an instructor/evaluator as to how one can judge the competency of a learner who attempts to master it! Thus, by not familiarizing oneself with the myriad of complexities within the Tamil language, either from a historical or purely from a synchronic point of view, one may tend to attribute each of these varieties as belonging to a separate language; and subsequently consider the variations therein as haphazard and random. Upon



studying the Tamil language variations from a historical point of view, one may easily note that such variations are vibrant and quite regular, and notably they conform to a logical sequence of changes. In this respect, no form of Tamil, either it is spoken or written, is neither random nor spontaneous in nature. Not to mention the fact that any study of diaspora Tamil of any region, without such a systematic account from a historical point of view, would only result to provide an unscientific description of the language of the respective region.

Language change occurs as a result of both internal as well as external causes. Internal causes are a) application of more than one phonological rule on agglutinative words; b) undergoing many naturally occurring linguistic processes such as, grammaticalization, reanalysis, metaphorization etc., in the language (See Renganathan 2010). The external causes, on the other hand, can be attributed to such factors like ‘foreign language contact’, ‘bilingualism’, ‘language dominance’, and so on to name a few. Not to mention the fact that over the period of a long history, Tamil language did undergo many changes both due to internal as well as external causes. Prakrit, Sanskrit, Persian, Portuguese, and more recently the English language contributed enormously to the development/distortion of Tamil language in a number of different ways. Interestingly, many Tamil language movements, both conscious as well as unconscious, such as ‘language purism’, ‘official language planning’, ‘language standardization’, ‘Tamilization’, ‘coining new vocabularies’ and so on contributed to the retention of many of these variations within it without having to undergo any extinction in any subtlest manner possible. Many of the so called indigenous and historically relevant Tamil words and morphological and syntactic forms - although not all of them - from the Sangam era are still extant in modern Tamil in one way or another: in one dialect or another, in one speech form or another, or in one register or another. This particular behavior of the Tamil language poses as a big threat not only for its continued consideration as an individual language, but also for its continued use of indigenous and historically significant forms under various labels as ‘pure Tamil’, ‘Sangam Tamil’, ‘Chastised Tamil’ and so on. Ironically, the major threat comes mostly from the judgments of second language learners for whom these historically relevant changes and existence of complex variations pose as a major hurdle in learning the language in a casual manner.

### **Delving into the complexity - A case in point is the use of the verb *en* ‘to say’:**

Almost all of the grammatical categories in Tamil have a systematic history behind them, and accounting all of them may require enormous amount of time and energy. An attempt is made in this section to trace the various use of the Tamil quotative marker *en* ‘that’ and its historical development, especially by making use of the electronic data extensively. Use of the verb *en* ‘to say’ can be taken as one of the instances for the contribution of complex forms in Tamil. This verb has underwent a wide range of alterations throughout the history of the Tamil language, but yet, it is still in use in the modern language the way it was during the Sangam period - perhaps with more number of characteristics which were not prevalent at its earlier stages. Unlike any other verb, this verb exhibits many structural gaps in modern written variety, but, significantly, not in spoken Tamil. Learning to master all of the uses of this verb, especially in spoken Tamil, is definitely one of the major challenges to any second language learner for the main reason that it not only underwent the process of grammaticalization, but also shows an agglutinative structure that is very difficult to comprehend and use by any non-native speaker of the language. This verb was used both as a regular lexical form

as well as a grammatical form representing the ‘complementizer’ in the Tamil language. Both the forms of *en* ‘that’ and *enpatu/enal* ‘the fact that’ had their equivalents both in old and modern Tamil.

A search of the database using a number of combinations, including **என், என்று, என்றிட், என்கொள்** and so on would reveal that besides the many of the finite forms of this verb, what underwent a significant change at a later period are the forms of negative adverbial (*ennātu* ‘without telling’) and nominal derivative (*ennāmai* ‘not saying’), which do not show any equivalent in modern literary Tamil.

அரும் படர் அவல நோய் ஆற்றுவள் **என்னாது** (Kali. 28)

arum pa<sup>ar</sup> avala nōy ā<sup>u</sup>va<sup>u</sup> ennātu

‘Without revealing the fact that she would experience the contagious love disease...’

அரிய ஆகும் **என்னாமை**... (Aham. 191)

ariya ākum ennāmai ..

‘Not saying that s.t. would be intricate to accomplish...’

Notably , the Modern Tamil equivalents of the suffixes *-ātu* and *-āmai* such as *-āmal* (eg. *collāmal* ‘without saying’ \**ennāmal*) and *-ātatu* (*collātatu* ‘that which was not said’ \**ennātatu*) respectively tend to occur with the verb *en* only in spoken Tamil but not in the corresponding literary variety. What turns out to be the crux of the issue here is the obscure nature of the spoken Tamil equivalents of the verb *en* ‘say’ in present, past and future forms, which normally occur as a single or clustered consonant: **ங்** ‘<sup>ng</sup>’ (**பாக்றேங்றேன்** pāk<sup>ē</sup>ē<sup>ē</sup> ‘I say that I see’, **சொல்றாங்றான்** col<sup>ā</sup>ā<sup>ā</sup> ‘he says that he tells’); **ண்ண்** ‘<sup>nn</sup>’ (**பாக்றேண்ணேன்** pāk<sup>ē</sup>ē<sup>ē</sup> ‘I said that I see’); and **ம்ப்** ‘<sup>mp</sup>’ (**கொடுப்பேம்பேன்** ko<sup>u</sup>ppēmpēn ‘I will say that I would give’) respectively (Cf. Search: ngr, **என்கிற**) . The obscure form of this suffix, its complex clause construction in an agglutinative form, along with the non-existence of some of the conjugations of this verb in written Tamil contribute enormously to the complexity of spoken Tamil.

**வருவாங்காமெ** varu-vān-<sup>ā</sup>me

‘without saying that he would arrive..’

**வரமாட்டேங்காமெ** vara-mā<sup>ā</sup>-ēn-<sup>ā</sup>m<sup>ā</sup>

‘without saying that I won’t come...’

**ஆகுங்காதது** āku(m)-<sup>ā</sup>ta-tu

‘saying that s.t. wouldn’t happen’

**கேப்பாங்காதது** kē(<sup>ā</sup>)-pp-ā<sup>ā</sup>-āt-atu

‘saying that he wouldn’t ask’

Notably, these, supposedly, commonly occurring forms in spoken Tamil do not have any parallel in written Tamil, as a result it generates a structural gap in the corresponding written variety of Tamil. What one can attribute to this phenomenon is that the spoken Tamil exhibits a perfect continuum from Sangam to the present time as it continues to retain the structure that one can attest from old Tamil, but this is not the case with the corresponding written variety of Tamil, which exhibits a structural gap in terms of not exhibiting the equivalents of *āmal* and *āmai* with the verb ‘en’.

\*varuvēn enkāmal (\*வருவேன் என்காமல்)  
 \*varamā□□ēn enkāmal (\*வரமாட்டேன் என்காமல்)  
 \*ākum enkātatu (\*ஆகும் என்காதது)  
 \*kē□pān enkātatu (\*கேட்பான் என்காதது)

If the form **என்காமல்** 'enkāmal' is nonexistent in modern written Tamil, but only found in spoken Tamil as in -□kāma, a question arises as to when and how the form *enka* as an infinitive form of this verb lost its use in the history of Tamil language? The other alternative point of view would be to consider this form as an innovation in spoken Tamil but not in modern Tamil. Note that the Sangam Tamil form **என்க** *enka* occurs as an 'optative form' to mean 'let it be said', but not as infinitive form of the verb 'en'.

நாடன் **என்கோ?** ஊரன் **என்கோ?** (Puram 49).

nā□an enkō? ūran enkō?

'Would I call him a country person or a town person?'

பின்னாளில் தன் மனைவியைக் காணும் மகிழ்ச்சியாற் பாசறையில் இனிய துயில்

கொள்கின்றான் **என்க**. (Mullaip paattu 11).

'pinnā□il tan manaiviyaik kā□um makir□cciyā□ pāca□aiyil iniya tuyil ko□kin□ān enka'

'Assume that he takes a comfortable nap at the jail with the prevailing thought that he would see his wife in the future!'

However, neither the Sangam Tamil forms such as *ennāmai* or *ennātu*, nor the relatively more recent forms such as *enāmal* or *enātu* do not seem to have any parallels in written Tamil, but as we noticed above, they do occur in spoken Tamil with their root forms of the verb **ங்** '□', **ண்ண** '□□' and **ம்ப்** 'mp' in a relatively large number of conjugations.

This is particularly true for the fact that one can observe from the search results of the electronic database using the forms such as **என்றிட்** and **என்றுகொள்**, which especially use of the aspectual auxiliaries such as **இடு** i□u (definitive auxiliary) and **கொள்** ko□ (reflexive auxiliary). Along with the progressive auxiliary form **கொண்டிரு** ko□□iru, these forms seemed to have been attested only starting from the medieval bhakti literature, especially from Tirumular's Tirumantiram, as sited below.

அறிவே அறிவை அறிகின்றது **என்றிட்டு** (Tirum. 2033)

a□ivē a□ivai a□ikin□atu en□i□□u

'having said that Knowledge knows the knowledge...'

ஈவ பெரும்பிழை **என்றுகொள்** ளீரே (Tirum. 506)

īva perumpi□ai en□uko□□irē

'Assume that s.t. would result to a great fault'

Surprisingly, like in the earlier cases of negative verbal participle and verbal derivative form, these constructions also do not exhibit in parallel in modern written variety, but only found widely in spoken Tamil.

எல்லாரும் வருவோம்னுட்டாங்க! (எல்லாரும் வருவோம் \*என்றுவிட்டார்கள்)

ellārum varu-v-ōm-u-āka! 'Eveyone proclaimed affirmatively that they would come)

என்ன நீ என்னெ மாடுகீடுண்ணுகிட்டிருக்கே? (cf.

<http://www.thetamilangauge.com/spokentamil> search: என்று)

(\*என்ன மாடு கீடு என்கொண்டிருக்கிறாய்?)

enn mā u kī u-ki-iru-kk-? 'Why do you keep calling me a water buffalo?'

நீயே எடுத்துக்குவேண்ணுக்கோ!

nīyē e-u-tt-u-kku-v-ē-u-kkō

'Proclaim that you would take everything for yourself'

What do these exceptional forms imply is that 'spoken Tamil' and 'written Tamil' seem to have followed two different historical paths from Sangam to modern Tamil and in this respect the spoken Tamil seems to show a richer structure than the written Tamil, especially in terms of retaining more number of archaic forms than the corresponding written version. This is in opposition to those instances of modern Tamil where new structures evolved and no traces of which can be found either in Sangam or in medieval Tamil. An example may be the case of experience subject construction, which is new to modern Tamil, but not in Sangam Tamil, as in *yā viyartta* 'I was sweat' as opposed to *e akku viyarkki* (cf. Murugaiyan 2004). Yet another feature from a historical point of view is loss of medieval and Sangam forms which do not have any trace in modern Tamil. A case in point is the use of imperative suffixes *-min* (*kēlmin* 'listen') and *-anmin* (*kū anmin* 'do not utter') etc., which do not have any occurrence in any identical forms in modern Tamil (cf. Renganathan 2010). Identifying the point of time in which these changes occurred is an endeavor that requires analyses of text of different genre in a thorough manner.

Yet another advantage of studying word forms that underwent many changes historically using electronic data is that it is possible for one to trace the trajectories of the cause of certain changes over the period of time in a systematic manner. One of such phenomena is authors' handling of a particular style causing the development of new categories. One of them that may be cited here is the formation of the modern Tamil modal auxiliary *lām*. It may be stated that various use of the combination of the infinitive suffix *-al* with the neuter future form of verb *āku* 'become' in ancient Tamil later caused the formation of *lām*. An extensive search using the keys such as *லாகுமே*, *லாமே*, *லும் ஆமே*, *லும் ஆகும்* etc., one may notice that the modal auxiliary *lām* came into existence in modern Tamil by the linguistic process of reanalysis due to various use of this structures by poet saints. Consider for example the expression *kē u en al tu intu colal ākum* - Manimekalai and its modern Tamil equivalent *kē irukkum enki-atu tu intu collalām* (Modern Tamil) 'One may say for sure that there would be a devastation', where the syntactic construction *colal ākum* 'saying is possible' is found to be occurring with many different combinations synonymously, as in *colal ām* - after phonological reduction of *ākum* to *ām*; *colla lām* with a reanalysis of verb forms and so on (see Renganathan 2010: pp. 171-73 for a detailed study of this change). By toggling between the selections of the bhakti, Sangam and modern literatures using the above search keys, one can notice the various use of this combinations more in bhakti texts than in Sangam texts.

## Search techniques and need for a tagged Corpora:

Perhaps an advance search technique with many combinatory possibilities is needed to successfully derive all of the intended and unattested forms from all of the genres of Tamil language. Ideally, one may want to search text in many complex ways, like ‘words that end in particular suffix (-vi□u; ki□; ko□ etc.)’, sentences with a particular combination of words (dative subject and psychological verbs; subject with the suffix *āl* and modal verbs like -o□□, -iyal etc.) and so on. Even though such sophisticated search possibilities is yet to be made available for Tamil using any conceivable tagged corpus as discussed in detail in Renganathan(2001), Baskaran et al (2008) etc., with the current database, however, storing text in Unicode does offer some work around. For example, if one intends to retrieve all of the word forms with the suffix *-i□u*, *āl*, *-ukku* and so on, one can use the Unicode glyphs of the initial vowels, as in *இடு* , *ஶல்* , *ஶக்கு* respectively to accomplish this task. This method can be considered as a substitute for any equivalent method of information retrieval using tagged corpus, which would normally contain all of the affixes parsed and stored separately in a more systematic manner. Absence of any such tagged corpus and an intelligent parser for all of the genres of Tamil texts from Sangam to Modern Tamil, one requires to use this kind of alternative search methods to accomplish the task. Among many others, the other significant advantages of using electronic data may be making dialect geographies from a historical point of view, attempting to find the chronology of authors and texts and so on.

## References

- Baskaran Sankaran, Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha, and K. V. Subbarao, A Common Parts-of-Speech Tagset Framework for Indian Languages, Proceedings of LREC 2008.
- Murugaiyan, A (2004) Note sur les prédications expérientielles en tamoul classique, *Bulletin de la Société de Linguistique de Paris*, 99, p. 363-382.
- Renganathan, Vasu. (2010). The Language of Tirumūlar’s Tirumantiram, A Medieval Saiva Tamil Religious Text. Unpublished Ph.D. thesis, University of Pennsylvania, Philadelphia.
- Renganathan, Vasu. (2001). Development of Morphological Tagger for Tamil, INFITT Conference, Kuala Lumpur.

# Open Source Tamil Computing

*S. Gopinath and E.I. Nehru*

National Informatics Centre, Chennai

## Abstract

For many of us English is the natural choice for commodity Computing such as Internet Web browsing, email, Instant Messaging, Word Processing etc. As computer proliferates in every aspects of our daily life, it is a need to have some kind of Multilingual Computing. For example an individual would like to send an email to his friend in a local language (like sending a greeting email in Tamil), an e-Governance application needs to be benefited for persons who are not familiar with English, or a product brochure to be made available on Web in a local language. More importantly the multilingual data should be interoperable and to be long lived across computer systems. The fundamental requirements for the Language Computing are

1. Coding system for Local Language Script Set,
2. Input Methods, Output Methods,
3. Software Libraries, APIs, Fonts
4. Applications

The aim of this article/paper is to summarize the various aspects of Multilingual Computing in OpenSource Platform with emphasis on Linux and Tamil.

Contemporary Linux systems comes with Multilingual features and in most cases they can be enabled very easily.

## 1. Unicode

Internally, coding of character system to be done to make meaningful processing of characters. A decade ago Standard ASCII (7bit) is a very popular encoding system for English characters. Later extended ASCII (8 bit/1 byte) is used for representing English characters and several Latin characters, Indic characters etc. But a true multilingual system needs to represent all possible script set (including Math Symbols, braille characters etc.) independently so that any electronic text shall contain all sorts of characters.

In late 1980s, Joe Becker (Xerox), Lee Collins and Mark Davis of Apple conceptualized multilingual character set encoding. Joe Becker published a draft for International/Multilingual Character Encoding System and tentatively called Unicode. This proposed 16 bit character model (each character is to be represented by 16 bits). This proposal is popularly called as Unicode88. Later when Unicode 2.0 was proposed, the character width is no longer restricted to fixed 16 bits and hence its possible to represent many ancient character sets like Egyptian Hieroglyphs. As of 2011 Unicode 6.0 standard prevails.

Unicode defines codespace containing code points in the range 0x0 to 0x10FFFF (represented in Hexadecimal) which means there are 1,114,112 code points in the Unicode Codespace. Each character is assigned a code-point. Unicode codespace is divided into 17 planes and numbered from 0 to 16. The Plane 0 is called "Basic Multilingual Plane" which contains codepoints numbered from 0x0 to 0xFFFF. Code Points for Tamil Character Set, other Indic Character Set, Cyrillic, Geometric Shapes, Math Symbols, Arabic etc are allocated in this plane (hence this Plane-0 is called Basic Multilingual Plane). Other planes are

1. Plane 1 - Supplementary Multilingual Plane
2. Plane 2 - Supplementary Ideographic Plane
3. Plane 14- Supplementary Special Purpose Plane
4. Plane 15,16- Private Area Use

The Tamil Language is allocated codepoints from 0x0B80 to 0x0BFF (128 codepoints) in BMP Plane. Please remember, that the codepoints have nothing to do with physical representation of characters in computer binary bits (or qubits in Quantum Computing). The physical representation also called as Unicode Character Encoding is done through a methodology called Unicode Transformation Format and it is defined by the Unicode Consortium. In Posix Systems and in Internet, UTF-8 Encoding Scheme is popular. In this article/paper, by default we frequently refer UTF-8 encoding scheme.

The UTF-8 was initially proposed for Plan9 Operating System. It is a multibyte character encoding system which uses one byte to 4 bytes depending upon the codepoint. The encoding for the code point from 0 to 127 is same as ASCII encoding for ASCII space 0 to 127 and hence it is already compatible with ASCII for the code points 0-127. The main advantage for the UTF-8 is its self synchronising and does not depend on endianness of the computer system. No special marking in the data stream is required for UTF-8. The main disadvantage is that it needs more bytes and it requires extra processing.

One of the basic question that arises is this. We have only 128 codepoints allotted for Tamil in the Unicode and we have 247 number of characters in Tamil Character Set. Is this a right way ?. How do we manage to represent all characters in Tamil Character Set?. This is very debatable. The brief explanation for the question how to accommodate 247 characters in 128 code point is as follows.

The most of the characters in Tamil are conjunct characters and share the same set of glyphs (or visual representation) in most of the characters and they are highly structured. Hence, it is possible to represent the Tamil Characters with very little compromise within 128 code point. But conjunct characters takes more bytes space than vowels or consonants (consonant equivalents).

## 2. Input System

It is quite obvious that keyboard is the de-facto input device for humans (Yes.. we have voice and other input systems for Humans!). In PC based architecture, keyboard is a raw device, and has no language centric mechanisms expect that in most of the keyboards we have "English" letters inscribed on keys. One of the questions beginners of Tamil Computing asks, is there any keyboard where I can find all 247 characters ?. The answer is similar to that of in the previous section that we indeed do not require individual keys for all the 247 characters (infact one can make that either physical or virtual

onscreen keyboards, but it would be expensive and bulky ) as many are conjunct characters and many share common glyphs (diacritics) and hence the normal general purpose contemporary keyboard is suffice. This methodology requires some kind of software popularly called Input Methods to assemble the characters when multiple keystrokes are required to form conjunct characters.

Pioneering works has been done by the creators of m17n libraries which is quite popular in Posix systems. The m17n library is a multilingual text processing library for C language. In Linux the popular Input Methods available are Smart Common Input Method (SCIM) and iBUS (Intelligent Input Bus). Most of the current Linux uses iBUS by default.

### **3. Output Methods**

The whole fanfare in Multilingual Computing is on the output. Unless one sees the output in appropriate script, the multilingual computing is not fulfilled. It's the most obvious part of Multilingual computing. Normally, we talk about fonts whenever we discuss about Multilingual Computing. Fonts contain information about how to draw the shapes on the screen. The renderer engine which is part of the output system takes the font and renders the shapes. The fonts are usually distributed in a file (often called as a font file). There are 3 types of fonts. They are Bitmap fonts, Outline fonts and Stroke based fonts. The bitmap fonts or raster fonts contain pixel images of the glyphs. The Outline fonts or vector fonts contain the vector images and mathematical formulae can be applied to get the different sizes of a glyph. The popular font types such as TrueType, OpenType, Adobe fonts are Outline fonts.

One of the functions of m17n library is displaying and rendering multi-lingual texts. It requires complex processing to render scripts such as Tamil, Devanagiri etc. A character may have a single glyph or few glyphs spaced at one another. A re-ordering may also be required. The technology for such rendering is known as "Complex Text Layout" or CTL. The m17n library database contains what is called Font Layout Tables (FLT) which bridges the fonts and the rendering engine.

The present day Linux distribution such as Debian, RedHat are available with TrueType and OpenType Font rendering engine. The Lohit fonts are quite popular in Linux Community. Lohit Fonts Project is sponsored by RedHat and is a Fedora Hosted. CDAC distributes OpenType Fonts for Indic Languages. By default most of the Linux distros have Tamil fonts by default or can be added very easily.

### **4. Software Libraries/APIs**

The GNU C Library (glibc) supports UTF8 encoding and contains functions for multilingual computing. The utilities such as sort, grep which are based on glibc can be used on multilingual texts. The holy grail M17N C library is the flag ship for multilingual computing. The link <http://unicode.org/resources/libraries.html>

contains links for various multilingual libraries.

### **5. Applications**

The end user experience on multilingual computing based on the Applications. The simple commodity applications like Internet surfing, Word Processing and e-mail etc. can be done using



existing applications like Mozilla Firefox, ThunderBird, OpenOffice, Libre office etc. The user has to only choose the appropriate Input Method (so that he can type in Tamil). Applications like Yahoo Messenger is multilingual and once can chat in clear Tamil (instead of transliteration) Empathy which is an GNOME application can be used for Instant Messaging and its multilingual. Simple editor like gedit can be used to create multilingual files.

Web Static Pages can be created by including a http header Content Type. This can be done by adding a meta tag between <head> and </head> as follows.

```
<head>  
  
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8"/>  
  
</head>
```

In the body section, the content can be multilingual. The multilingual content can be created even with simple editor like gedit. Of course the browser and the system in which it runs must have UTF-8 support with appropriate language fonts installed. Enterprise and backend applications use Database Management Systems and in Open Source Community one of the most popular DBMS is Postgres. The current version of Postgres supports UTF-8 encoding by default on Posix systems. This facilitates the storage of data in multilingual form.

## 6. Requirements and Suggestions

Although it is possible to carry out most of the multilingual computing, the following are the suggestion by the authors for further enhanced functionality.

1. The simpler keyboards can be made to meet the various requirements of the end user. For example to use TamilNet99 keyboard requires practice and considerable understanding of the scripts, the people who just knows only to read or write the scripts could not easily use these keyboard layouts or definitely not as trivial as someone who only knows English and use English keyboard. A much simpler keyboard may be prototyped after doing sufficient study on the above aspects. As the input methods are implemented in a software, these initiatives can be done without changes in underlying software.
2. As of now Input Methods are implemented in software and they run in main system/CPU. It is possible to design and implement an Intelligent keyboard, which directly sends the UTF-8 sequences instead of raw keyboard scan codes. As Unicode integrates across language/scripts, culture etc, the Universal Intelligent Keyboard is appropriate. The keyboard can be made programmable such that the end user can configure the keyboards to function either as raw scancode mode, or direct UTF-8 mode. If the same kind of keyboard is mechanically designed foldable/wrappable, it can be used in mobile devices such as smart phones, tablets universally (which can bring down the costs).
3. The authors could see that more works to be done on Optical Character Recognition, Speech to text conversion, Accessibility in the context of Multilingual computing.
4. Better Tamil outline fonts needs to be forged for High Resolution equipments.

## 7. Conclusion

Multilingual computing in Open Source Systems like Linux is matured. Works like M17N C libraries, UTF-8 encodings really makes Multilingualization possible in Linux like systems which carefully eliminates need for any special magic numbers. Due to glibc and m17n libraries, large number of applications are already multilingual enabled.

## 8. Acknowledgements

I would like to thank Mr.T.N.C. VenkataRangan of Vishwak Solutions Ltd, for spending hours in answering my novice questions when I started to learn Multilingual Computing.

My special thanks to CDAC Chennai officials (Mr.C.Sudhakar, Mr.Nagesh, Mr.Gowri Ganesh) for assisting me in learning some of the concepts in this area.

My thanks to my seniors (Mr.E.I.Nehru,Mr.JD Prince) for encouraging as well answering my doubts on the subject. My thanks to NIC's FMG Engineers (Mr.Emthias and others) for their Technical Support.

I thank NIC for providing enough resources.

Thanks to IITM Students for providing help in preparing the document in LaTeX.

Not but not least, my thanks to my parents and my brother who patiently accepts my late arrivals and patiently serve late dinner as I had to spend late hours while preparing this article.

## 9. Few References

- <http://www.cl.cam.ac.uk/~mgk25/unicode.html>
- <http://www.cl.cam.ac.uk/~mgk25/ucs/UTF-8-Plan9-paper.pdf>
- <http://code.google.com/p/ibus/>
- <http://www.m17n.org>
- <http://www.postgresql.org/docs/9.0/static/sql-createdatabase.html>
- Ofcourse, Wiki and Google pages.