

# 21<sup>வது</sup> தகிழ் இணைய டாநாடு

## 21<sup>st</sup> Tamil Internet Conference

15<sup>th</sup> - 17<sup>th</sup> December, 2022



Central Institute of  
Indian Languages  
(CIIL) Mysore

  
**Veda Publications**  
vedapub@gmail.com



15<sup>th</sup> - 17<sup>th</sup> December, 2022

21<sup>வது</sup> தகிழ் இணைய டாநாடு

# 21<sup>வது</sup> தகிழ் இணைய டாநாடு

## 21<sup>st</sup> Tamil Internet Conference

15<sup>th</sup> - 17<sup>th</sup> December, 2022



Organized by

Tamil University, Thanjavur, Tamilnadu, India

Periyar Maniyammai Institute of Science and Technology, Thanjavur, Tamilnadu, India

International Forum for Information Technology in Tamil (INFITT)

Central Institute of Indian Languages (CIIL), Mysore, India



Central Institute of  
Indian Languages  
(CIIL) Mysore

# 21<sup>வது</sup> தமிழ் இணைய மாநாடு

21<sup>st</sup> Tamil Internet Conference

15<sup>th</sup> - 17<sup>th</sup> December, 2022

Organized by

Tamil University, Thanjavur,  
Tamilnadu, India

Periyar Maniyammai Institute of Science and Technology,  
Thanjavur, Tamilnadu, India

International Forum for Information Technology in Tamil (INFITT)

Central Institute of Indian Languages (CIIL), Mysore, India



Central Institute of  
Indian Languages  
(CIIL) Mysore



**21<sup>st</sup> Tamil Internet Conference 2022**

Edited by

Dr.R.Ponnusamy, Dr.Subhalalitha, Dr.Nimmala.K, Dr. Prem Kumar,  
Dr.Parameswari Krishnamurthy

© Veda Publications

ISBN: 978-93-91930-28-8

Published by

**VEDA PUBLICATIONS**

G2, Kasthuri Flats,  
Dhenupuri Housing Colony,  
Madambakkam, Chennai - 600 126

M.No: 70928 22277

vedapublications.in

vedapub@gmail.com

L<sup>A</sup>T<sub>E</sub>X Typeset & Printing by

**Xalent Graphixs**

Chennai

M.No: 87544 39300

xalentgraphixs@gmail.com

All rights reserved

*No part of this publication maybe reproduced, stored in a  
retrieval system, or transmitted, in any form or by any means,  
electronic, mechanical, photocopying, recording or otherwise,  
without the prior permission of the publishers.*

**Publishers Disclaimer**

*The Publisher of this book states that the Author(s) of this book  
has taken the full responsibility for the content of this book,  
any dispute and copyright violation arising based on the content of this book  
will be addressed by the author(s), furthermore, the authors indemnify  
the publisher from damages arising from such disputes and  
copyright violation as stated above.*

## **Conference Program Committee**

### **Conference Patrons:**

Dr. V. Thiruvalluvan, The Vice Chancellor, Tamil University, Thanjavur

Dr. S. Velusamy, The Vice- Chancellor, Periyar Maniammai Institute of Science and Technology (Deemed to be University), Thanjavur.

Mr. Thavaruban Thangaraja, Jaffna, Srilanka, Chairman, INFITT

### **Conference Program Committee (Computational Linguistics, literature, and Epigraphy)**

Dr. Appasamy Murugaiyan, University of Paris – Chair

Dr. Vasu Renganathan, University of Pennsylvania – Co-Chair

Dr. P.Vijayalakshmi Dean FHSM

Periyar Maniammai Institute of Science and Technology

### **Members:**

Dr. Ma. Ganesan, Annamalai University, Chidambaram

Dr. Umarani Pappusamy, Central Institute of Indian Languages

Dr. K. Kalyanasundaram, Federal Institute of Technology in Lausanne, Switzerland

Dr. N. Deivasundaram, Madras University

Dr. Dhanalakshmi, Pondicherry University

Dr. Rajendran, Amrita University

Dr. Subathini Ramesh, University of Jaffna

Dr. K. Rajan, Annamalai University

### **Conference Program Committee (Natural Language Processing, Speech/Text and Artificial Intelligence)**

Dr. Sobha Lalitha Devi, AU-KBC, Anna University – Chair

Dr. Kengatharaiyer Sarveswaran, Jaffna University, Srilanka – Co-Chair

### **Members:**

Dr. Muthu Annamalai, USA

Dr. K. Parameswari, University of Hyderabad

Dr. Pattabhi RK Rao, AU-KBC, Anna University

Dr. Vijay Sundar Ram, AU-KBC, Anna University

Dr. Uthayasanker Thayasivam, Moratuwa University, Srilanka

Mr. Jan Kučera, University College, London

**Publication Committee**

Dr. Ramalingam Ponnusamy, Chennai Institute of Technology – Chair  
Dr. Subalalitha, SRM Institute of Science and Technology (SRMIST)  
Dr. Nimala. K, SRM Institute of Science and Technology (SRMIST), Chennai.  
Dr. L. R. Prem Kumar, CIIL, Mysore  
Dr. K. Parameswari, University of Hyderabad

**Local Organizing Committee:**

Dr. M. Jayakumar, Bharathiyar University – Chair  
Dr. Mangaiyarkarasi, Department of Linguistics, Tamil University – Co-Chair &–  
Organising Secretary  
Ms.V.Saranya HoD/ Languages, Periyar Maniammai Institute of Science and  
Technology  
Dr. R.Ponnusamy, Chennai Institute of Technology – Organising Secretary

**Members:**

Dr. C. Thiyagarajan, The Registrar, Tamil university  
Dr. S. Kavitha, Dean, Language Faculty, Tamil University  
Dr. L. Ramamoorthy, Central University of kerala, Kasaragod  
Dr. R. Neelakandan, Dean, Faculty of Science, Tamil University  
Dr. Ula. Balasubramanian, Dean, Faculty of Developing Tamil, Tamil University  
Dr. T. Kannan, Dean, Faculty of Palmleaf Manuscripts, Tamil University  
Dr. Elaiyappilai, Dean, Faculty of Arts, Tamil University  
Dr. K. Ravikumar, Department of Computer Science, Tamil University  
Mr. Rama Suganthan, Chennai  
Mr. T. Srinivasan  
Dr. L. R. Prem Kumar, Central Institute of Indian Languages  
Dr. M. Rameshkumar, Department of Linguistics, Tamil University  
Dr. K. Perumal, Department of Linguistics, Tamil University

**International Organizing Committee:**

Mr. Thavaruban Thangaraja – Jaffna, Srilanka – Chair  
Mr. Elantamil, University of Malaya – Co-Chair

**Members:**

Dr. Arul Veerappan, New York University, USA.  
Dr. R. Sakthivel, IIT, Chennai, India

-----



தங்கம் தென்னரசு  
தொழில்துறை அமைச்சர்



தலைமைச் செயலகம்,  
சென்னை-600 009

நாள் 05.12.2022

### வாழ்த்துரை

"வாணையாளப்போம் கடல்மீனை யாளப்போம்  
சந்திர மண்டலத்தியல் கண்டு தெளிவோம்"

என்று அறிவியல் தொழில்நுட்ப வளர்ச்சியின் தேவையை எடுத்துரைத்தார்  
மகாகவி பாரதியார்.

முத்தமிழாய் முகிழ்த்த கணித்தமிழ் இன்று காலவோட்டத்திற்குத் தக்க  
தன்னைத் தகவமைத்துக்கொண்டு "கணித்தமிழாய்" மிளிர்கின்ற இக்காலத்தில்,  
திசம்பர் 15-17 ஆகிய நாள்களில் உத்தமம் நிறுவனம், தஞ்சைத் தமிழ்ப்  
பல்கலைக்கழகம், பெரியார் மணியம்மை அறிவியல் (ம) தொழில்நுட்ப நிறுவனம்  
ஆகியவற்றின் சார்பில் "தமிழ் இணைய மாநாடு - 2022" நடைபெறவுள்ளதை  
அறிந்து நான் மகிழ்ச்சி கொள்கிறேன்.

வளர்ந்துவரும் அறிவியல் தொழில்நுட்பத்தின் பயன்பாட்டைத்  
தவிர்த்துவிட்டுத் தொழில்துறையிலும், கற்றல் - கற்பித்தலிலும்,  
மொழிவளர்ச்சியிலும் எந்தவொரு மேம்பாட்டையும் எய்துவிட முடியாது என்பது  
காலமறிந்த உண்மை. அத்தகைய சூழலில், இந்த 21ஆம் உலகத் தமிழ் இணைய  
மாநாடானது "தொழில்துறை 4.0 மற்றும் 5.0இல் தமிழின் பங்கு" என்ற மையப்  
பொருண்மையில் நடைபெறவிருப்பது காலப்பொருத்தமும் சாலப்பொருத்தமும்  
ஆகும்.

இம்மாநாட்டையொட்டி, கல்வி மற்றும் தமிழ் ஆய்வுகளில் கணினியின்  
பயன்பாடு, தமிழில் தானியங்கி எந்திரன் (ரோபோ) ஆய்வு, நுண்பகுப்பாய்வு (ம)  
அறிவுசார் இலக்கியம், இக்கால உரைநடை தேடல் உள்ளிட்ட பல ஆய்வுப்  
பொருண்மைகளில் கருத்துநிறை கட்டுரைகள் அடங்கிய மாநாட்டு மலர்  
வெளியிடப்படவுள்ளதும் பாராட்டிற்குரியது. இந்நூல் படிப்போர்க்குப் பயன்மிக  
நல்கும் என்பது என் திண்ணமான எண்ணம்.

தங்கம் தென்னரசு  
தொழில்துறை அமைச்சர்



தலைமைச் செயலகம்,  
சென்னை-600 009

நாள்.....

. 2 .

தமிழ்க்கணினி ஆய்வை இளம் தலைமுறை மாணவர்களிடம் எடுத்துச் செல்லும் நோக்கில், பள்ளி - கல்லூரி மாணவர்களையும், ஆய்வாளர்களையும், பேராசிரியர்களையும் இம்மாநாட்டிற்கு அழைக்கவிருப்பதை நான் மனதாரப் பாராட்டுகிறேன்.

"தமிழ் இணைய மாநாடு - 2022" வெற்றிபெற அருந்தமிழ் அன்னை அருள்புரியட்டும் என்று கூறி, உத்தமம் நிறுவனத் தலைவர் திரு. தவருபன் தங்கராஜா, தமிழ்ப் பல்கலைக்கழகத் துணைவேந்தர் முனைவர் வி. திருவள்ளுவன், பெரியார் மணியம்மை அறிவியல் (ம) தொழில்நுட்ப நிறுவனத் (நிகர்நிலை பல்கலைக்கழகம்) துணைவேந்தர் முனைவர் செ. வேலுசாமி, உத்தமம் நிறுவன (இந்தியா) நிருவாக இயக்குநர் முனைவர் இரா. பொன்னுசாமி உள்ளிட்ட அனைவருக்கும் அன்பு வாழ்த்துகளைத் தெரிவித்துக்கொள்கிறேன்.

அன்புடன்,

(தங்கம் தென்னரசு)

த. மனோ தங்கராஜ்  
அமைச்சர் தகவல் தொழில்நுட்பவியல்  
மற்றும் டிஜிட்டல் சேவைகள் துறை



தலைமைச் செயலகம்,  
சென்னை-600 009

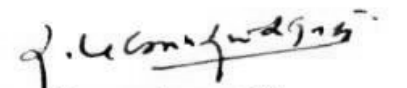
நாள் - 01.12.2022

### வாழ்த்துரை

உத்தமம் (உலகத்தமிழ் தகவல் தொழில்நுட்ப மன்றம்) ஏற்பாடு செய்துள்ள 21-வது தமிழ் இணைய மாநாடு (TIC2022) 2022 ம் ஆண்டு டிசம்பர் திங்கள் 15 முதல் 17 வரை தஞ்சாவூரில் நடைபெற உள்ளது என்பதை அறிந்து மகிழ்ச்சி அடைகிறேன். மேலும் இந்த மாநாட்டினை தமிழ்ப் பல்கலைக்கழகமும், பெரியார் மணியம்மை அறிவியல் மற்றும் தொழில்நுட்பக் கழகமும் (நிகர் நிலை பல்கலைக்கழகம்) தஞ்சாவூரில் நடத்துகின்றன, முக்கியமாக தொழில்துறையில் தமிழின் தாக்கம் மற்றும் தொழில்துறை 4.0 தொழில்துறை 5.0 என்ற குழுக் கருப்பொருளில் இந்த மாநாட்டை நடத்துகிறது என்பது மிகவும் சிறப்பு வாய்ந்த ஒன்றாகும்.

உத்தமம், தமிழ்ப் பல்கலைக்கழகம் மற்றும் பெரியார் மணியம்மை அறிவியல் மற்றும் தொழில்நுட்ப நிறுவனம் (நிகர்நிலைப் பல்கலைக்கழகம்) ஆகியவற்றின் உன்னதமான, பொருத்தமான மற்றும் தனித்துவமான முயற்சிகளை நான் பாராட்டுகிறேன், இது போன்ற ஒரு சர்வதேச நிகழ்வை ஏற்பாடு செய்ததற்காக, முக்கியமாக இந்நிகழ்ச்சியில் தமிழ் மொழியை இணையத்தில் உருவாக்குவதில் கவனம் செலுத்துகிறது. தொழில்துறை 4.0 மற்றும் தொழில்துறை 5.0 மற்றும் அதன் தொழில்நுட்ப தாக்கங்கள், உண்மையில், இந்த இலக்குகள் புதிய தொழில்நுட்ப வளர்ச்சியை வெளிக்கொணரும் மற்றும் இணையம் மற்றும் சமூக ஊடகங்களை அணுகுவதில் தமிழ் மொழியை பயனுள்ளதாக மாற்றும் என நம்புகிறேன்.

தமிழ் இணைய மாநாட்டில் பங்குகொள்ளும் அனைவருக்கும் பாராட்டுக்களை தெரிவித்துக்கொள்வதுடன், மாநாடு சிறப்பாக நடைபெற எனது மனமார்ந்த வாழ்த்துக்களை தெரிவித்துக்கொள்கிறேன்.

  
(த. மனோதங்கராஜ்)





முனைவர் **வி.திருவள்ளுவன்**  
துணைவேந்தர்  
தமிழ்ப் பல்கலைக்கழகம்  
தஞ்சாவூர் - 613 010  
தமிழ்நாடு, இந்தியா



அலுவலகம் : 04362-227040  
இல்லம் : 04362-226741  
கைபேசி : 9443480649  
மின் அஞ்சல் : vtvalluvan@gmail.com  
tamilunivc@gmail.com  
இணையதளம் : www.tamiluniversity.ac.in



### வாழ்த்துரை

தஞ்சாவூர், தமிழ்ப் பல்கலைக்கழகம், தஞ்சைப் பெரியார் மணியம்மை அறிவியல் மற்றும் தொழில்நுட்ப நிறுவனம், உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம் மற்றும் இந்திய மொழிகளின் நடுவண் நிறுவனம் (மைசூர்) ஆகியவை இணைந்து தொழில்நுட்பம் 5.0-இல் தமிழின் பங்கு என்னும் பொருளை மையக்கருத்தாகக்கொண்டு இந்த 21-வது தமிழ் இணைய மாநாடு, 2022 திசம்பர் திங்கள் 15, 16, 17 ஆகிய தேதிகளில் வெகுசிறப்பாக நடக்கவிருக்கிறது என்பதை எண்ணி மனம் மட்டற்ற மகிழ்ச்சி அடைகின்றது.

**"கேடில் விழுச்செல்வம் கல்வி ஒருவர்க்கு**

**மாடல்ல மறையவை"**

(திருக்குறள்:400)

என்ற வள்ளுவனின் வாக்குக்கிணங்க, ஒரு மனிதனுக்குக் கல்வி என்பது இன்றியமையாதது; இருப்பினும் இன்றைய காலக்கட்டத்தில் ஒரு மொழியினைக் கணினித் தொழில்நுட்பத்தில் புகுத்தி அந்தக் கல்வியையும் கன்னித்தமிழையும் வளர்ப்பது நமது தலையாய கடமை. ஏனெனில், இன்றைய காலக்கட்டத்தில் கணினித் தொழில்நுட்பம் இன்றி ஓரணுவும் அசையாது என்ற நிலைக்கு உலகம் தள்ளப்பட்டுள்ளது என்பதை நாம் அறிவோம். அந்தவகையில், மனிதனுடைய மூளை எந்த அளவிற்கு மொழியில் வேலை செய்யுமோ; அந்த அளவிற்கும் மேலாகக் கணினித் தொழில்நுட்பம் தமிழ்மொழியில் உலா வந்துகொண்டிருக்கிறது என்பதில் சிறிதும் ஐயமில்லை. இருந்தாலும், ஒரு சிலவற்றில் தொய்வு ஏற்பட்டுள்ளது என்ற உண்மையை நாம் ஏற்றுக்கொள்ளத்தான் வேண்டும்.

**"தேமதுரத் தமிழோசை உலகமெல்லாம் பரவும் வகை செய்தல் வேண்டும்"**

என்னும் பாரதியின் பாடலுக்கிணங்க, உலகில் பல்வேறு கணினித்தமிழ் அமைப்புகளும் தமிழ்க் கணினி ஆர்வலர்களும் பல்வேறு நிகழ்வுகளை நடத்திக்கொண்டு, கணினித் தொழில்நுட்பம் மூலமாகச் சங்கத் தமிழ் முதல் தற்காலத் தமிழ்வரை உலகமெங்கும் எடுத்துச் சென்றுகொண்டிருக்கின்றனர் என்பதில் ஐயப்பாடு இல்லை. இருப்பினும், தமிழ் வரலாற்றையும், தமிழ்ப் பண்பாட்டையும், பழங்கால இலக்கண இலக்கியங்களையும் கணினித் தொழில்நுட்பம் மூலமாக, இன்னும் தமிழின் வளர்ச்சியை மெருகூட்டும் வகையில், உலகத்தாருக்குக் கொண்டு செல்ல வேண்டுமென்று இம்மாநாட்டில் கலந்துகொள்ளும் அறிஞர் பெருமக்களை அன்புடன் கேட்டுக்கொள்கிறேன்.

உலகின் பல்வேறு பகுதிகளிலிருந்து தமிழ்க் கணினியில் ஆர்வம் கொண்ட பல்வேறு அறிஞர்கள் இம்மாநாட்டில் கலந்து கொள்ள இருக்கின்றனர் என்பது மிகவும் மகிழ்ச்சி அளிக்கின்றது. அதுமட்டுமல்லாது, கூகுள், மைக்ரோசாப்ட் போன்ற பெரும் நிறுவனங்களிலிருந்து வரும் கணினி வல்லுநர்களின் சிறப்பு உரைகள் இம்மாநாட்டிற்கு ஒரு மகுடமாக விளங்கும்.

**"தனித்தியங்கும் உயர்ந்த தகுதி தமிழுக்கு உண்டு; தமிழ் பிறமொழித் துணையின்றியே வளரும் ஆற்றல் பெற்றது"** என்று டாக்டர் இராபர்ட் கால்டுவெல் அவர்கள் குறிப்பிடுவதுபோல, கணினித் தமிழின் வளர்ச்சி அகர பலத்தில் உள்ளது. இருந்தாலும், ஆங்கில மொழிக்கு இணையாக தொழில்நுட்பத்தோடு இணைந்த தமிழ் மொழியின் வளர்ச்சி உலகமெங்கும் பரவவேண்டும். அதற்கு இம்மாதிரியான மாநாடு துணைநின்ற கணினித் தமிழின் சவால்களை சமாளிக்கும் என்று நம்புகிறேன்.

இம்மாநாடு சிறக்க அரும்பணியாற்றிய அனைத்துக் குழுவினருக்கும் என் மனமார்ந்த நன்றியை உரித்தாக்குகிறேன். இம்மாநாட்டு நிகழ்வுகள் சீரும் சிறப்புடன் நடைபெறவும், மாநாட்டு மலரில் இடம்பெறும் கணினித் தமிழ்க் கட்டுரைகள் உலகளவில் ஒரு தாக்கத்தை ஏற்படுத்தவும் எனது உளமார்ந்த நல்வாழ்த்துக்களைத் தெரிவித்துக்கொள்கிறேன்.

  
(வி. திருவள்ளுவன்)







# திராவிடர் கழகம்

தலைமை நிலையம், பெரியார் திடல், 84/1 (50), ஈ.வெ.கி.சம்பத் சாலை, வேப்பேரி, சென்னை-600 007.  
தமிழ்நாடு, இந்தியா. • தொலைப்பேசி: 91-44-26618161/62/63 • தொலைப்பதிவி: 91-44-26618866  
• மின் அஞ்சல்: dkheadquarters@gmail.com • இணையதளம் : www.dravidarkazhagam.org

நிறுவனத் தலைவர்  
**தந்தை பெரியார்**

தலைவர்  
**கி.வீரமணி, எம்.ஏ., பி.எல்.,**

நாள் : 12.12.2022

பெறுநர்:

முனைவர் இரா. பொன்னுசாமி அவர்கள்,  
நிர்வாக இயக்குநர்,  
உத்தமம் நிறுவனம்

பேரன்புடையீர்,

வணக்கம்.

உத்தமம் நிறுவனத்தின் இருபத்தியொன்றாவது தமிழ் இணைய மாநாடு - 2022 டிசம்பர் 15, 16, 17 ஆகிய மூன்று நாட்கள், தஞ்சைத் தமிழ்ப் பல்கலைக்கழகமும், பெரியார் மணியம்மை அறிவியல் மற்றும் தொழில்நுட்ப நிறுவனமும் (நிகர்நிலைப்பல்கலைக்கழகம்) இணைந்து நடத்துவது மிகுந்த மகிழ்ச்சி அளிக்கிறது.

உலகின் மூத்த மொழி தமிழ் செம்மொழியை இன்றைய காலகட்டத்திற்குக்கேற்ப தொழிற்நுட்பத்தில் அதிகளவில் பயன்படுத்துவதற்காக மேற்கொள்ளும் முயற்சி பாராட்டுக்குரியதாகும்.

இம்மாநாட்டில் அயல்நாடுகளிலிருந்து தமிழ்க் கணினி ஆய்வாளர்கள் கலந்துக் கொள்ள இருப்பது பொருத்தமானதாகும்.

சிறப்பு வாய்ந்த மாநாட்டினை ஏற்பாடு செய்து நடத்தும் தங்களுக்கும், மாநாட்டு குழுவினருக்கும் எமது நெஞ்சம் நிறைந்த பாராட்டுகள்.

மாநாடு சிறக்க அன்பான வாழ்த்துகள்!

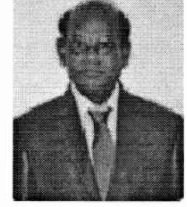
அன்புடன்,

(கி. வீரமணி)

தலைவர், திராவிடர் கழகம்



## வாழ்த்துமடல்



வணக்கம் !

உத்தமம் நிறுவனத்தின் 21 ஆவது உலகத் தமிழ் இணைய மாநாடு வருகிற 2022 - ஆம் ஆண்டு டிசம்பர் மாதம் 14 ஆம் தேதி முதல் 17 தேதி வரை பார் போற்ற உலகத் தமிழ் மாநாடு நடத்தி புகழ்பெற்ற தஞ்சை மாநகரில் உள்ள தமிழின் முழு வளர்ச்சிக்காக பாடுபடும் தஞ்சை தமிழ் பல்கலைக்கழகமும் மற்றும் தமிழ் எழுத்துக்கள் சீர்த்திருத்தத் செம்மல் தமிழர்கள் வளர்ச்சிக்காக தன் வாழ்வை அர்பணித்த தந்தை பெரியாரின்பெயர் தாங்கி புகழ்பெற்று விழங்கும் மக்கள் பல்கலைக்கழகமாம் பெரியார் மணியம்மை அறிவியல் மற்றும் தொழில்நுட்ப நிகர்நிலைப் பல்கலைக்கழகமும் இணைந்து நடத்துவது அனைவருக்கும் மட்டற்ற மகிழ்ச்சி அளிக்கிறது. இந்த மாநாட்டை சிறப்பாக ஏற்பாடுகளை செய்து நடத்திக்கொடுக்கும்படி எங்களிடம் முழுப் பொருப்பினையும் நம்பிக்கையுடன் ஒப்படைத்த உத்தமம் நிறுவனத்தினருக்கு, இம்மாநாட்டு குழுவினர் சார்பாக நன்றியினைத் தெரிவித்துக் கொள்கிறோம்.

தமிழ் மற்றும் கிரேக்கம்; லத்தீன் மற்றும் சமஸ்கிருதம் போன்ற சில மொழிகள் கல்வி உலகில் செம்மொழி தகுதியை அனுபவித்தாலும், அவற்றின் தொண்மை மற்றும் வளமான இலக்கிய பாரம்பரியத்திற்கு பெயர்போன செம்மொழி தகுதியை அதிகாரப் பூர்வமாக பெற்ற முதல் வாழும் மொழி தமிழ் மொழியே ஆகும்.

செம்மொழி தமிழ் எழுத்தினால் காலங்காலமாக மாற்றமில்லாத தனித்துவம் பெற்ற மொழியாக விளங்குகிறது. காலத்திற்கேற்ப பேச்சுவழக்கு, எழுத்துவடிவம், எழுத்து குறைப்பு மற்றும் வேற்று மொழி வார்த்தைகள் சேர்ப்பு மிக அவசியமாகும். தமிழ் மொழி சில மாற்றங்களை ஏற்க வேண்டும். அப்பொழுதுதான் அம்மொழி வளரும், உலக மொழியாக ஒருநாள் மலரும். தந்தை பெரியார் அவர்கள் அந்த மாற்றத்தைத்தான் விரும்பினார். தமிழ்மொழியை, தாய்மொழியாக இல்லாதவர்கள் சுலபமாக எழுதவும், படிக்கவும், அச்சடிக்கவும் மற்றும் தமிழ் அரசு மொழியாக வேண்டும் என்று முயன்றவர், இந்த புரட்சிகரமான, சிந்தனையையும் மற்றும் செயலையும் அனைவரும் வியக்கும் வண்ணம் செய்த தந்தை பெரியார் அவர்களை தமிழ் மொழி சீர்த்திருத்த செம்மலாகவும் மேலும் அவர் தமிழ் சமுதாய சீர்த்திருத்தவதியாகவும் திகழ்ந்தார் என்பது மிகையாகாது.

ஆகையால் இன்றைய சூழலுக்கு ஏற்ப, தமிழ் மொழி வளர்ச்சியடைய, இந்த மாநாட்டின் மையக் கருத்தாக தொழில் துறை 5.0 இல் தமிழின் பங்கு (Tamil Industry 5.0) என்னும் பொருளில் தமிழ் மாநாட்டுக் கட்டுரைகள் படைப்பும் மற்றும் சிறப்பு பேச்சுகளும் நடத்தத் திட்டமிட்டுள்ளது மிகவும் பாராட்டுக்குரியது.

இந்த தமிழ் இணைய மாநாட்டில், தமிழ் ரோபாட்டிக்ஸ், செயற்கை நுண்ணறிவு மற்றும் இயற்கை மொழி ஆய்வுகள் ஆகிய தலைப்புகளில் இந்தியா, இலங்கை, மலேசியா, அமெரிக்கா



மற்றும் உலகெங்கிலும் உள்ள பல்கலைக்கழகங்களின் தமிழ் மற்றும் கணினி அறிவியில் துறைகளைச் சேர்ந்த மாணவர்கள் மற்றும் ஆராய்ச்சியாளர்கள் தங்கள் கட்டுரைகளைப் படைக்க வருகிறார்கள் என்பது சிறப்புக்குரியதாகும்.

மேலும் பல்வேறு தலைப்புக்களில் சிறப்பு பேச்சாளர்கள் பல்வேறு நாடுகளில் இருந்து அழைக்கப்பட்டுள்ளதால், அத்துடன் மாணவர்கள் மற்றும் பயிற்றுனர்களுக்காக திறன் மேம்படுத்த உதவும் மென்பொருள்கள் குறித்த பயிற்சி பட்டறைகளும் மாநாட்டின் ஓர் அங்கமாக செயல்படவுள்ளது. இம்மாநாட்டின் கூடுதல் சிறப்பாகும்.

தமிழில் உருவாக்கப்பட்ட மென்பொருள்கள் மற்றும் அவை சம்மந்தமான படைப்புகள் குறித்த கண்காட்சியும் ஏற்பாடு செய்யப்பட்டுள்ளது தமிழ் ஆர்வலர் அனைவருக்கும் பயன்படும் என்பதில் அய்யமில்லை.

மொழி அழிந்தால் மக்களின் பண்பாடு அடையாளம் அழியும். பண்பாட்டு அடையாளம் இல்லாத மக்கள் வரலாறு அற்றவர்களாகப் போவார்கள். தன்மானத்துடன் வாழ, நாம் நம் தாய்மொழி காப்போம். நாம் அனைவரும் விழி போல் எண்ணி நம்மொழிகாக்க வேண்டும் என்று உறுதி மொழியோடு அதற்கான ஆக்க பூர்வமான சிந்தனையிலும் மற்றும் செயலிலும் நம்மை முழுமையாக ஈடுபடுத்திக் கொள்வதே செம்மொழி தமிழுக்கு நாம் செய்யும் நன்றிக் கடனாகும்.

- ❖ தாய்மொழியே பயிற்றுமொழி
- ❖ தாய்மொழியே ஆட்சிமொழி
- ❖ தாய்மொழியே நீதிமன்ற மொழி
- ❖ தாய் மொழியே வாழ்வியல் மொழி

என்றென்றும் வாழ்வில் அனைத்து துறையிலும் நம் மொழியைப் பயன்பாட்டு மொழியாகவும் பார்போற்றும் மொழியாகவும் ஆக்குவோம்.

இந்த உலக தமிழ் இணைமாநாடு வெற்றியடைய இப்பல்கலைக்கத்தின் நிறுவனர் மற்றும் வேந்தர் சார்பாகவும் மேலும் இப்பல்கலைக்கழகத்தின் சார்பாகவும் வாழ்த்துக்களையும் மற்றும் பாராட்டுகளையும் அனைவருக்கும் தெரிவித்துக்கொள்கிறேன்.

வாழ்த்துக்களுடன்



வல்லம்  
தஞ்சாவூர்

நாள் : 01.12.2022

(செ.வேலுசாமி)  
துணை வேந்தர்



**த. தவருபன்  
தலைவர்  
உலகத்தமிழ் தகவல் தொழில்நுட்ப  
மன்றம்**




அதீத வேகத்துடன் வளர்ந்து வரும் தகவல் தொழில்நுட்பத்தில் தமிழ் மொழிக்கு ஒரு சிறப்பான இடத்தை உறுதிசெய்யவும் அந்த தகவல் தொழில்நுட்பத்தின் ஊடாக தமிழ் மொழியை அடுத்த தலைமுறைக்கு இட்டுச்செல்லும் நோக்குடன் தமிழ் இணையம் என்ற தொனிப்பொருளில் 1997 இல் ஆரம்பிக்கப்பட்ட தமிழ் இணையமாநாடு, 2000 ஆம் ஆண்டில் நடைபெற்ற மாநாட்டில் உலகத்தமிழ் தகவல் தொழில்நுட்ப மன்றம்( உத்தமம்) அமைப்பை தோற்றுவித்ததன் மூலம் தகவல் தொழில்நுட்பத்தில் தமிழின் வகிபாகத்தை உறுதிசெய்யக்கூடிய உலகளாவிய ரீதியில் தொழிற்படக்கூடிய ஒரு அமைப்பிற்கான இடைவெளியை இல்லாதொழித்திருந்தது.

அதன் பின் பல்வேறு நாடுகளில் சுழற்சி முறையில் நடைபெற்ற தமிழ் இணைய மாநாடுகளின் ஊடாக பல ஆய்வுகள் மென்பொருள்கள் மற்றும் தமிழ் தகவல் தொழில்நுட்ப செயற்பாடுகள் உலகளாவிய ரீதியில் ஆய்வாளர்களாலும் தொழில்நுட்ப வல்லுனர்களாலும் வெளிச்சத்துக்கு கொண்டுவரப்பட்டது.

இன்று நாம் அனுபவித்துக்கொண்டிருக்கும் அநேகமான தமிழ் சார்ந்த தகவல் தொழில்நுட்ப விளைவுகள் கடந்த காலங்களில் நடைபெற்ற மாநாடுகளின் ஆய்வுகளில் பேசப்பட்ட அவற்றினால் உந்தப்பட்ட முயற்சிகள் என்றால் அது மிகையல்ல . இந்த அமைப்புக்கு தலைவராக இருக்கும் நான் 2004 இல் உத்தமத்தில் இணைந்து தான் தகவல் தகவல் தொழில்நுட்பத்தில் தமிழ் குறித்து அறிந்து கொண்டேன் கிட்டத்தட்ட 20 ஆண்டுகளை அண்மித்த நிலையில் இன்று எனது தலைமையில் இருக்கும் உத்தமத்தின் 21 வது தமிழ் இணையமாநாடு தஞ்சை தமிழ்ப்பல்கலைக்கழகத்தில் பெரியார் மணியம்மை அறிவியல் மற்றும் தொழில்நுட்ப நிறுவனத்துடனும் இணைந்து இவ்வருடம் நடாத்தப்படுவது எனக்கு பெருமகிழ்வை தருகின்றது. கடந்த இரண்டு வருடங்களாக நாம் ஒருவரை ஒருவர் சந்திக்க முடியாதிருந்தோம் இம்முறை அனைவரும் அறிவியல் தமிழின் பெயரால் மீண்டும் ஒன்று கூடுகின்றோம் என்பது இரண்டிப்பு மகிழ்ச்சி .

காலத்துக்கேற்ற வகையில் மாநாட்டின் மையக்கருத்தை தொழிற்சூழ்நெறி 5.0 இல் தமிழின் வகிபாகம் என்று வைத்திருப்பது இம்மாநாட்டுக்கு மேலும் வலுச்சேர்க்கும் என நம்புகின்றேன். தொழில்நுட்பம் பல்வேறு விடயங்களில் புகுத்தப்பட்டு இறுதியில் தொழிற்சூழ்நெறியில் கூட பல்வேறு பரிணாமங்களை தாண்டி வளர்ந்து வருகின்ற வேளையில், அதற்கு நிகராக தமிழும் தன்னை அதற்குள் உட்புகுத்த வேண்டிய கட்டயத்தில் அதனை எவ்வாறு எதிர்கொள்ளப்போகின்றோம் என்பதற்கு இந்த மாநாடு பதில் சொல்லும் என உறுதியாக நம்புகின்றேன்

இம்மாநாடு வெற்றி பெற தமிழ் இன்றும் உயிர்ப்புடன் வாழும் இலங்கையின் யாழ்ப்பாண மண்ணில் இருந்து வாழ்த்துக்களை வழங்குவதில் பெருமையடையும் அதேவேளை அனைவரும் எமது தமிழ் மொழியை தொழில்நுட்பங்கள் ஊடாக மட்டுமல்லாது அன்றாட வாழ்க்கையிலும் செயற்பாடுகளிலும் வளர்ப்பதற்கு உறுதி எடுக்க அழைக்கின்றேன்.

(தலைவர், உத்தமம் நிறுவனம்).





**மாண்புமிகு தகவல் தொழில்நுட்பம் மற்றும் டிஜிட்டல் சேவைகள் துறை அமைச்சர் திரு மனோ தங்கராஜ்  
அவர்களின் உரை**

எல்லோருக்கும் வணக்கம்!

இங்குக் குழுமியிருக்கும் பேராசிரியப் பெருமக்களே!

கணினித்தமிழ் வல்லுநர்களே!

தமிழ் ஆர்வலர்களே!

இந்த இரண்டு பல்கலைக்கழகங்களின் மாணவ மாணவிகளே!

உங்கள் அனைவருக்கும் என்னுடைய இதயங்களிந்த காலை வணக்கத்தைத் தெரிவித்துக் கொள்கின்றேன்.

தஞ்சைத் தமிழ்ப் பல்கலைக்கழகம், தஞ்சைப் பெரியார் மணியம்மை அறிவியல் மற்றும் தொழில்நுட்ப நிறுவனம், உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம், மைசூரில் உள்ள இந்திய மொழிகளின் நடுவண் நிறுவனம் ஆகியவை இணைந்து, தொழில்நுட்பம் 5.0-இல் தமிழின் பங்கு என்னும் பொருளை மையக்கருத்தாகக்கொண்டு, இந்த 21-வது தமிழ் இணைய மாநாட்டை வெகு சிறப்பாக நடத்திக் கொண்டிருக்கின்றன. கன்னித் தமிழை இந்த உலகத்தாருக்குக் கொண்டு செல்லும் நோக்கில் செயல்படும் நிறுவனங்களை மனதாரப் பாராட்டுகிறேன்.

உத்தமம் (INFITT) என்று சுருக்கமாக அழைக்கப்படும், உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம் ஒன்றுவிட்ட ஆண்டுகளில், மலேசியா, சிங்கப்பூர், இலங்கை, அமெரிக்க, ஜெர்மன் போன்ற வெளிநாடுகளிலும் பிற ஆண்டுகளில் இந்தியாவிலும் இம்மாநாடுகளை நடத்திக் கொண்டிருக்கின்றது. இவ்வாறு வெளிநாடுகளிலும் தமிழ்நாட்டில் உள்ள பல்வேறு பல்கலைக்கழகங்களிலும் விடா முயற்சியாக, 2000 ஆண்டுமுதல் இதுவரை ஏறத்தாழ 21 மாநாடுகளை நடத்தியுள்ளது என்று எண்ணும்போது மனம் பூரிப்பு அடைகின்றது.

இன்றைய காலம் கணினி யுகம் ஆகிவிட்டது; உலகில் கணினித்தொழில்நுட்பம் இல்லையென்றால் ஓரணுவும் அசையாது என்ற நிலை தற்போது ஏற்பட்டுள்ளது. அந்தக்

கணினித் தொழில்நுட்பம் மூலமாக நாமும் நம்முடைய பாரம்பரியமிக்க தமிழ் மொழியை உலகுக்கு கொண்டு செல்ல வேண்டும்.

செழுமை மிக்க பாரம்பரியத்தைக் கொண்டு விளங்குவதும், மிகச் சிறந்த பழைமையான, செறிவான இலக்கண இலக்கியங்களைத் தன்னகத்தே கொண்டதுமாகிய தமிழ் மொழிக்குச் சொந்தக்காரர்கள் நாம். உயிர்த்துடிப்புள்ள மொழி, தமிழ் மொழி. அத்தகைய தமிழ் மொழி, தொழில்நுட்பத்தில் இன்னும் மிகச் சிறப்படைய வேண்டும்.

முன்னொரு காலத்தில் தமிழ் மொழியினை இணையத்தில் கொண்டு செல்வதற்கும் C, C++, JAVA போன்ற கணினி மொழிகளில் உள்ளீடு செய்வதற்கும் பல்வேறு சிக்கல்கள் எழுந்தன. ஆனால் அந்தத் தடைக்கற்களை எல்லாம் கடந்து இன்று ஒருங்குறி, அதாவது யுனிக்கோட் என்ற முறையில் தமிழை இணையத்திற்குக் கொண்டு செல்வதற்கும் கணினி மொழிகளில் உள்ளீடு செய்வதற்கும் பயன்படுத்தி நாம் வெற்றி கண்டுள்ளோம்.

கணினித் தொழில்நுட்பத்தில் தமிழின் வளர்ச்சியானது அபரிமிதமானது. உலகில் பல்வேறு நாடுகளில் பேசக்கூடிய மொழிகளுக்கு இணையாக நம்முடைய தமிழ் மொழியையும் கணினித் தொழில்நுட்பத்தின் உதவியால் முன்னெடுத்துச் சென்றுள்ளோம். எடுத்துக்காட்டாக, தமிழில் நாம் பேசும் ஒவ்வொரு சொல்லையும் / ஏன் வாக்கியங்களையும் / பனுவல்களையும்கூட எழுத்தாக்கம் செய்வதில் வெற்றி கண்டுள்ளோம். அதுபோலவே, எழுத்து வடிவில் கொடுக்கும் பனுவல்களை (OCR) முறைப்படி தட்டச்சு வடிவில் நகல் எடுத்து, திருத்தம் செய்துகொள்ளும் தொழில்நுட்பத்தைச் செயல்படுத்திக் கொண்டோம். அதே பனுவல்களைத் தமிழில் உரையாகக் கொடுப்பதற்குத் தொழில்நுட்பத்தைப் பயன்படுத்தி வெற்றியும் கண்டோம்; ஏன், கையால் எழுதும் கைப்பிரதியைக்கூட அச்சு வடிவில் கொடுக்கும் தொழில்நுட்பத்தைத் தமிழில் ஏற்படுத்திக் கொண்டோம்.

இருந்தாலும், இந்த வெற்றி என்பது 70 முதல் 80 விழுக்காடு தான் நிறைவேறியுள்ளது என்பதையும் இங்கு குறிப்பிட விரும்புகின்றேன். இவற்றில் 100 விழுக்காடு நாம் வெற்றி காண வேண்டும்; அதற்காக நம்முடைய தமிழ்க் கணினி வல்லுநர்கள் இன்னும் முயற்சிகள் எடுக்க வேண்டும்.

இணையத்தில் தமிழின் பயன்பாட்டை மேம்படுத்துவதற்குத் தேவையான மென்பொருள்களையும் கருவிகளையும் உருவாக்குவதை முதன்மை நோக்கமாகக் கொண்டு நம் தமிழ்க் கணினி ஆர்வலர்கள் பாடுபட்டுக் கொண்டிருக்கின்றனர்.

கணினித்தமிழ் வளர்ச்சியில் தமிழ்க் கலைச்சொல்லாக்கம், தமிழ் ஒருங்குறி எழுத்துரு, தமிழ் எழுத்துரு அறிதல், லினக்ஸில் தமிழ் போன்ற பல்வேறு கணினி தொழில்நுட்பம் சார்ந்த ஆய்வுப் பணிகளில் நாம் தன்னிறைவு பெற்றுள்ளோம்.

தமிழ் மொழி, அறிவியல் தொழில்நுட்பம், தகவல் தொழில்நுட்பம், மொழியியல், கல்வெட்டியல், சுவடியியல், மானுடவியல், அறிவுசார் துறைகள் அனைத்திலும் கணினித் தொழில்நுட்ப வளர்ச்சி தொடர்ந்து தமிழ்ப்பணிக்கு உறுதுணையாக விளங்குகின்றது. அந்த

வளர்ச்சிக்கு இந்தக் கணினி வல்லுநர் குழுவும், உலகில் பெரும் நிறுவனங்களாகத் திகழும் கூகுள், மைக்ரோசாப்ட் போன்ற கணினித்துறை சார்ந்த நிறுவனங்களும் உறுதுணையாக நிற்க வேண்டும்.

தமிழர் வரலாறு, தமிழ் இலக்கியம், இலக்கணம், தமிழ் பண்பாடு பற்றி அறிந்து கொள்ள, மிகப்பெரிய தமிழ் மின் நூலக்கப் பணிகள் பல்வேறு நிறுவனங்களில் நடந்த வண்ணம் உள்ளன. இவ்வாறான, கணினித் தொழில்நுட்பம் சார்ந்த தமிழ்ப் பணிகள், (வெளிநாட்டினரோடு ஒப்பிடும்போது) தமிழ்நாட்டில் பாமர மக்களைப் போய்ச் சேர்ந்து உள்ளதா; சொல்லப்போனால், நம் போன்ற படித்த மக்களைப் போய்ச் சேர்ந்துள்ளதா என்பதுகூட சந்தேகத்துக்கு இடமளிக்கிறது.

காரணம், கணினித் தொழில்நுட்பத்தில் தமிழுக்கென பல்வேறு கருவிகள் வந்துள்ளன. இருந்தாலும், நம்மில் எத்தனை பேர் அதனைப் பயன்படுத்துகிறோம் என்று எண்ணிப் பார்க்கும் தருணம் இது. எத்தனையோ பேர் இன்னும் அலைபேசியில் தமிழைத் தட்டச்சு செய்யக் கற்றுக்கொள்ளவில்லை. இன்னும் அலுவலகங்களில் எத்தனையோ பேர் தமிழ் ஒருங்குறி எழுத்துக்களைப் பயன்படுத்தித் தட்டச்சு செய்வதில்லை. தமிழ் பேசும் ஒவ்வொருவரும் கணினித் தொழில்நுட்பக் கருவிகளை நாம் பயன்படுத்த வேண்டும். அப்போதுதான், தமிழின் பெருமை கணினித் தொழில்நுட்பம் மூலமாக இணையத்திற்கு சென்று உலகம் முழுவதிலும் உள்ள அனைத்து தரப்பினரும் பயனடைவர்.

”திங்களொடும் செழும் பருதி தன்னோடும்

விண்ணோடும் உடுக்களோடும்

மங்குல் கடல் இவற்றோடும்

பிறந்த தமிழுடன் பிறந்தோம் நாங்கள்”

என்ற பாவேந்தர் பாரதிதாசனின் முழக்கத்திற்கு இணங்க நம்முடைய கன்னித் தமிழை உலக மக்களுக்கு எடுத்துச் செல்ல வேண்டும்; இந்தப் பணி என்பது கணினித் தொழில்நுட்பம் சார்ந்தது; ஆகவே, கணினித்துறை வல்லுனர்களும் தமிழ் அறிஞர்களும் தனித்தனியாக அமர்ந்து செயல்பட்டால் இந்த இலக்கை அடைய முடியாது; ”கூடி வாழ்ந்தால் கோடி நன்மை”, ”ஒன்று பட்டால் உண்டு வாழ்வு” என்ற தமிழ் வாக்கிற்கு இணங்க, இந்த இரு துறை அறிஞர்களும் இணைந்து செயல்பட்டால் நம்முடைய செந்தமிழின் தொழில்நுட்ப வளர்ச்சி எங்கோ சென்று நிற்கும் என்பதில் சிறிதளவும் ஐயமில்லை. ஆகவே பல்வேறு துறை அறிஞர்களும் கணினித் துறையோடு இணைந்து செயல்படுங்கள்; வெற்றி காண்போம்.

இங்குக் குழுமியுள்ள தமிழ் ஆர்வலர்கள், கணினித் துறை வல்லுநர்கள், கணினித் தமிழை அடுத்த கட்டத்திற்கு எடுத்துச் செல்லக்கூடிய இளைஞர்கள் இங்கு குழுமி இருக்கின்றீர்கள். தமிழ்த் தொழில்நுட்பம் சார்ந்த வல்லுநர்கள் பலர் இம்மாநாட்டில் கலந்துகொண்டு தம்முடைய ஆய்வுரைகளை வழங்க இருக்கின்றீர்கள். அவர்களுக்கு என்னுடைய மனமார்ந்த நன்றி.

தஞ்சாவூரில் இந்த இரண்டு கல்வி நிறுவனங்களோடு இணைந்து இந்த 21-வது தமிழ் இணைய மாநாட்டினை நடத்த முன்வந்த உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றத்தாருக்கும் இந்திய மொழிகளின் நடுவன் நிறுவனத்திற்கும் எனது பாராட்டுக்கள். இந்த மூன்று நாள் மாநாடு வெற்றி

பெறவும், மாநாட்டை ஒட்டி வெளியிட இருக்கும் கட்டுரைத்தொகுப்பு மலர் உலகெங்குமுள்ள தமிழ் ஆர்வலர்களுக்குப் பயனாக அமையவும் எனது உளம் நிறைந்த வாழ்த்துக்களைத் தெரிவித்துக் கொள்வதில் பெருமகிழ்ச்சி அடைகின்றேன்.

வாழ்க செந்தமிழ்!

வளர்க கணினித் தமிழ்!

## தந்தை பெரியாரும் தமிழ் எழுத்துச் சீர்திருத்தமும்

அனைவருக்கும் அன்பான வணக்கம்.

21ஆம் பன்னாட்டுத் தமிழ் இணைய மாநாட்டை உத்தமம் (INFITT) அமைப்புடனும், தஞ்சை தமிழ்ப் பல்கலைக்கழகத்துடனும் இணைந்து நடத்துவதில் பெரியார் மணியம்மை அறிவியல் மற்றும் தொழில்நுட்ப நிறுவனம் (நிகர்நிலைப் பல்கலைக்கழகம்) பேருவகையும் பெருமிதமும் கொள்கிறது.

தமிழ் அறிவியல் மொழியாக வேண்டும்; தொழில்நுட்ப வளர்ச்சியோடு இயைந்து வளர்ந்து செழுமைபெற வேண்டும் என்பதைத் தொடர்ந்து வலியுறுத்தி வருவதுடன், அதற்கான கருத்தாக்கங்களையும் முன்னெடுப்புகளையும் தொடர்ந்து மேற்கொண்டும் வரும் பெரியாரின் தளத்தில் இம் மாநாடு நடைபெறுவது சாலப் பொருத்தமாகும்.

இன்றைய தொழில்நுட்ப உலகில் தமிழ் வழி தமிழரும், தமிழர் வழி தமிழும் வெல்ல, உத்தமம் ஆற்றியிருக்கும் பணிகள் போற்றத்தக்கவையாகும். மொழியின் வளர்ச்சி என்று தந்தை பெரியார் கருதியதும், வலியுறுத்தியதும் இதைத் தான். பழம் பெருமையிலும், அறிவியலுக்கு ஒவ்வாதனவற்றிலும் மூழ்கிக் கிடக்காமல் முன்னேற வேண்டும். உலகின் போக்கு அறிவியலால் நெறிப்படுத்தப்பட்டுவருவதை உணர்ந்ததனாலும், எதிலும் பகுத்தறிவுப் பார்வையும், மனித வளர்ச்சிப் பற்றும் கொண்டதனாலும் தொலைநோக்காளரான தந்தை பெரியார், மொழித் துறையையும் அதே பார்வையுடன் தான் அணுகினார்.

தமிழ்மொழி தாய்மொழியாக உள்ள நாட்டில் இந்தியைப் புகுத்தக் கூடாது என்று கிளர்ச்சி செய்தது எதற்காக என்று தந்தை பெரியாரே கூறுகிறார்.

“அது என் தாய்மொழிப் பற்றுதலுக்காக என்று அல்ல. அது என் நாட்டு மொழி என்பதற்காக அல்ல. சிவபெருமானால் பேசப்பட்டது என்பதற்காக அல்ல. அகத்திய முனிவரால் திருத்தப்பட்டதென்பதற்காக அல்ல. மந்திர சக்தி நிறைந்தது; எலும்புக் கூட்டைப் பெண்ணாக்கிக் கொடுக்கும் என்பதற்காக அல்ல. பின் எதற்காக? தமிழ் இந்நாட்டுச் சீதோஷண நிலைக்கேற்ப அமைந்துள்ளது. இந்திய நாட்டுப் பிற எம் மொழியையும் விடத் தமிழ், நாகரீகம் பெற்று விள?ங்குகிறது. தூய தமிழ் பேசுதல் மற்ற வேறுமொழிச் சொற்களை நீக்கிப் பேசுவதால் நம்மிடையேயுள்ள இழிவுகள் நீங்குவதோடு மேலும் மேலும் நன்மையடைவோம் என்பதோடு நம் பழக்க வழக்கங்களுக்கு ஏற்ப நம் மொழி அமைந்திருக்கிறது. வேறு மொழியைப் புகுத்திக் கொள்வதன் மூலம் நம் அமைப்புக் கெடுவதோடு, அம் மொழியமைப்பிலுள்ள நம் நலனுக்குப் புறம்பான கருத்துக்கள் கேடு பயக்கும் கருத்துக்கள் நம்மிடைப் புகுந்து நம்மை இழிவடையச் செய்கின்றன என்பதால்தான்.

வடமொழியில் நம்மை மேலும் மேலும் அடிமையாக்கும் தன்மை அமைந்திருப்பதால்தான் அதையும் கூடாதென்கிறேன். நமது மேன்மைக்கு, நமது தகுதிக்கு, நமது முற்போக்குக்கு ஏற்ற மொழி- தமிழைவிட மேலான ஒரு மொழி இந்நாட்டிலில்லை என்பதற்காகவே தமிழை விரும்புகிறேனே தவிர, அது அற்புத அதிசயங்களை விளைவிக்கக்கூடியது என்பதற்காக அல்ல. இம்மொழியில் - அதிசயங்கள் விளைவித்ததாகப் புராணங்கள் கூறுகின்றன. ஆனால், அந்தக் கருத்து நமக்குத்தேவையில்லை. ஏன்? ஆயிரம் முதலையை வைத்துக் கொண்டுதான் பாடிப் பாருங்களேன்; அவற்றில் எதுவாவது, தான் தின்ற ஒரு மீனையாவது கக்குகிறதா என்று; தாழ்ப்பாளிட்ட சிற்றறையின் முன்னின்றுதான், இலக்கணப்படியே மனம் உருகிப் பாடுங்களேன் - சிறிதாவது தாழ்ப்பாள் அசைகிறதா என்று; அற்புத சக்திகள் நிறைந்த மொழி என்று பிடிவாதம்



செய்வது அறியாமைதான்; அது தமிழ்ப் பண்புகூட அல்ல. தமிழில் - அதிசயம், மந்திரம், சக்தி முதலிய சொற்களே இல்லை.”

தமிழை அழகு மொழி, அதிசய மொழி, மூத்த மொழி என்பதற்காகவெல்லாம் போற்றக் கூடியவரல்ல தந்தை பெரியார். மாறாக, அதன் பயன்பாட்டைக் கருதி தமிழை உயர்ந்த மொழி என்றவர். வடமொழி ஆதிக்கத்தினின்றும் தமிழ் விடுபட்டு, தூய தமிழாகத் துலங்க வேண்டும் என்று கருதியதற்கான காரணமும் வெறும் மொழித் தூய்மை பேசும் தூய்மைவாதம் அன்று.

வடமொழித் தொடர்பால் தமிழில் மாற்றுச் சொல்கூட இல்லாத மூடநடம்பிக்கைக் கருத்துகளும் (மோட்சம், திதி, சொர்க்கம், மோட்சம், நரகம், ஆத்மா), பேதமும் (ஜாதி, தாரா முகூர்த்தம், கன்னிகாதானம், பதிவிரதாத் தன்மை) புகுந்து தமிழ் மக்களின் அன்றாட வாழ்க்கையிலும் நிகழ்த்தியிருக்கும் கேடுகளையும் நீக்கிட வேண்டும் என்ற நோக்குடனும் தான். இத்தகைய பார்வையோடு மொழியை அணுகியதிலும் பெரியார் ஒருவர் தான் பெரியார்.

“நம் நாட்டு சீதோஷண நிலையைப் பொறுத்தும் கருத்துகளின் செழுமையைப் பொறுத்தும் நமக்குத் தமிழ்தான் உயர்ந்த மொழியாகும்” என்ற பெரியார், அது மேலும் செம்மைப்படுத்தப்பட வேண்டும், மக்கள் கற்க மேலும் இலகுவாக்கப்படவேண்டும், பயனுள்ள பரந்த மொழியாக்கப்பட வேண்டும் என்று விரும்பினார்.

எதையும் தனது சிந்தனையால் விருப்பு - வெறுப்பின்றி சமுதாய, மக்கள் பயனாகும் கண்ணோட்டத்திலேயே சிந்தித்து கருத்துக்களைக் கூறும் துணிச்சல் அவரது இயல்பு.

நமது மக்களில் மிகப்பெரும்பாலோர் ஏன் கல்வி அறிவற்றவர்களாக இருக்கிறார்கள்? அவர்கள் அனைவரும் எளிதில் எழுத, படிக்க, சிந்திக்க எப்போதுதான் பரவலாக வருவார்கள்? - இப்படி கவலையோடு கேள்வி கேட்டார் - விடையும் கண்டார்!

‘நோய்நாடி நோய் முதல்நாடி’ பரிகாரம் காணும் தன்மையே அவரது அணுகுமுறை.

தமிழ்நாட்டில் உள்ள நம் தமிழ் மக்களில் பெரும்பாலோருக்கு ஏன் எழுதப் படிக்க வாய்ப்பில்லை என்பதுபற்றி சிந்தித்த தந்தை பெரியார் அவர்கள், அதற்குரிய மூலகாரணங்களை மக்கள் முன் வைத்தார்!

1. குலதர்மம் என்ற வர்ணதர்ம முறைப்படி, மிகப் பெரும்பாலோரான சூத்திர, பஞ்சம மக்கள் படிக்க தடை இருந்தது. வேதம், இதிகாசங்கள், உயர்ஜாதியினருக்கு-பார்ப்பனருக்கு மட்டுமே படிக்க உரிமை தந்தன.

2. அத்தடைகளையும் தாண்டி படிக்க முனைந்தாலும், ஏராளமான எண்ணிக்கையில் உள்ள தமிழ் எழுத்துக்களை அம்மக்கள் கற்பது கடினமாகவே உள்ளது.

இவ்விரண்டு தடைகளையும் உடைத்தெறிய சூளுரை மேற்கொண்டார்.

“தமிழ்ப் பாஷை எழுத்துக்கள் வெகுகாலமாகவே எவ்வித மாறுதலும் இல்லாமல் இருந்து வருகின்றன.

உலகில் உள்ள பாஷைகள் பெரிதும் சப்தம், குறி, வடிவம் எழுத்துக்கள் குறைப்பு, அவசியமான எழுத்துக்கள் சேர்ப்பு ஆகிய காரியங்களால் மாறுதல் அடைந்துகொண்டே வருகின்றன.

காலதேச வர்த்தமானங்களுக்கு ஏற்ப சப்தங்களும் பாஷைகளும், உச்சரிப்புகளும், வடிவங்களும் மாறுவது இயல்பேயாகும்.

வார்த்தைகள் கருத்தை வெளியிடுவதற்கு ஏற்பட்டவைகள் என்பது போலவே, எழுத்துக்களும் சப்தத்தை உணர்த்த ஏற்பட்டவைகளேயாகும்.

ஆனால், நம் பண்டிதர்களுக்குத் தாராளமாய் அறிவைச் செலுத்த இடமில்லாமல், மதம், பழக்கவழக்கம் ஆகியவைகள் குறுக்கிட்டு விட்டதால், எழுத்துக்களுக்கும், அதன் கோடுகளுக்கும், வடிவங்களுக்கும் தத்துவார்த்தம் கற்பிக்க வேண்டிய நிலை நம் நாட்டில் ஏற்பட்டுவிட்டது.”

“இன்றைய தமிழ் மிகவும் பழைய மொழி, வெகுகாலமாகச் சீர்திருத்தம் செய்யப்படாதது, மற்ற மொழிகளைப்போல திருத்தப்படாதது” என்பதான இவைகள் ஒரு மொழிக்குக் குறையாகுமே தவிர, பெருமையாகாது என்பேன். ஏன்? பழமை எல்லாம் அநேகமாக மாற்றமாக இருக்கிறது; திருத்தப்பட்டிருக்கிறது. மாற்றுவதும், திருத்துவதும் யாருக்கும் எதற்கும் இழிவாகவோ, குற்றமாகவோ, ஆகிவிடாது. மேன்மையடையவும், காலத்தோடு கலந்து செல்லவும், எதையும் மாற்றவும், திருத்தவும் வேண்டும். பிடிவாதமாய்ப் பாட்டிக்காலத்திய, பண்டைக் காலத்திய பெருமைகளைப் பேசிக்கொண்டிருந்தால், கழிபட்டுப் போவோம்; பின்தங்கிப் போவோம்.

மொழி என்பது உலகப் போட்டிப் போராட்டத்திற்கு ஒரு போர்க்கருவியாகும். போர்க் கருவிகள் காலத்திற்கேற்ப மாற்றப்பட வேண்டும். அவ்வப்போது கண்டுபிடித்துக் கைக்கொள்ளவேண்டும். நம் பண்டிதர்கள் இந்த இடத்திலும் நம் மொழிக்கு மிக்க அநீதி விளைவித்துவிட்டார்கள். தமிழ் சிவனும், சுப்ரமண்யனும் பேசிய மொழி, உண்டாக்கிய மொழி என்று பண்டிதர்கள் கூறுகிறார்கள். அதே சிவனும் சுப்ரமண்யனும் உபயோகித்த போர்க்கருவிகள் இன்று நம் மக்களுக்குப் பயன்படுமா? அவைகளை நாம் இன்று பயன்படுத்துவோமா? அல்லது அவர்களே இன்று போரிட நேர்ந்தால் அவைகளைப் பயன்படுத்துவார்களா? சிந்தித்துப் பாருங்கள். கடவுள் உண்டாக்கினார் என்பது நமக்குத் தோன்றிய இயற்கைத் தத்துவம் ஆகும்.

இயற்கையின் தத்துவம் நமது அறிவு வளர்ச்சிகளுக்கேற்ப மாறுதல்களுக்கும், செப்பனிடுவதற்கும் வசதியளிக்கக்கூடியதேயாகும். சிவன் கோலும், மழுவும், கத்தியும், வேலும், ஆலமும் கொண்டுதான் சண்டைபிடித்து இருக்கிறாராம். விஷ்ணு வந்த பிறகே வில் வந்திருக்கிறது. அதன் பிறகே துப்பாக்கியும், அதிலிருந்து பீரங்கியும், மிஷின் பீரங்கியும் ஏற்பட்டு இன்று அணுகுண்டு வரை போர்க்கருவிகள் முன்னேற்றமாகியிருக்கின்றன. இன்று நாமோ, நம் கடவுள்களோ போரிட நேர்ந்தால், வில்லும் வேலுமா உபயோகிப்போம்? ஆகவே, போர்க்கருவிகள் மாற்றமடைந்திருப்பதுபோல் நமது மொழியும் மாற்றம் அடையவேண்டாமா?” என்று கேள்வி எழுப்பியதுடன், தமிழ் எழுத்துகளை எவ்வாறு சீர்திருத்தலாம் என்பதைப் பரிந்துரைத்து, அவற்றைச் செயல்படுத்தியும் காட்டினார்.

“பிறர் சுலபமாகத் தமிழ்மொழியைக் கற்றுக் கொள்வதற்காகவும், சுலபமாக அச்சுக் கோக்கவும், டைப் அடிக்கவும், தமிழ் எழுத்துகளில் சில சீர்திருத்தங்கள் செய்யப்படுவது நலம் என்று நினைக்கிறேன்” என்று இன்றைய சூழலில் தமிழ் எழுத்துச் சீர்திருத்தம் அவசியம் என்று தான் கருதும் நோக்கத்தையும் முன்வைத்து, அச் சீர்திருத்தங்களையும் காரண காரியங்களுடன்

விளக்கி, ஒவ்வொரு கட்டமாக முன்வைக்கிறார்.

மிக முக்கியமாக இதில் நாம் நோக்க வேண்டியதென்னவென்றால், ஆங்கிலத்திலும், பிற மொழிகளிலும் எழுத்துகள் குறைவாக இருக்கின்றன என்பதற்காக நாமும் குறைக்கலாம் என்று போகிற போக்கில் சொல்லிவிடவில்லை. ஆங்கிலத்தின் அமைப்பு முறையையும், தமிழ் இலக்கணத்தின் இயற்கைத் தன்மையையும் புரிந்து, போற்றி, அதன் இலக்கணமும், ஒலிப்பும் கெடாமல் எப்படி சமகாலத்திற்கேற்ப அந்த மாற்றங்கள் இருக்க வேண்டும் என்ற கருத்தின் அடிப்படையிலேயே எழுத்துச் சீர்திருத்தத்தைப் பரிந்துரைக்கிறார். ஆங்கில எழுத்துகளிலும், உச்சரிப்பிலும் செய்யப்பட்டுவந்த மாற்றங்களைச் சுட்டிக்காட்டி, “அமெரிக்காவில் சமீபகாலத்தில் எழுத்துக் கூட்டும் முறைகள், இவைகளில் மாற்றம் செய்வதால், இலக்கணத்தில், உச்சரிப்பில், பொருளில் மாற்றம் ஏற்படுவதாயிருந்தாலும்கூட துணிவாக மாற்றிக்கொண்டிருக்கிறார்கள். ஆனால் நான் சொல்லும் மாற்றங்களுக்கு அப்படிப்பட்ட குற்றங் குறைகள் இல்லையென்றே கருதுகிறேன்” என்று இலக்கணம் கெடாமல் சீர்திருத்தத்தைப் பரிந்துரைத்திருப்பதைக் காணலாம்.

“சாதாரணமாகத் தமிழ் உயிர் எழுத்துகளில் ஐ, ஓ ஆகிய இரண்டு எழுத்துகளைக் குறைத்துவிடலாம். இந்த இரண்டும் தேவையில்லாத எழுத்துகள். மேலும், இவை கூட்டெழுத்துகளே ஒழிய, தனி எழுத்துகள் அல்ல. இவை இல்லாமல் எந்தத் தமிழ்ச் சொல்லையும் எழுதலாம், உச்சரிக்கலாம். இவைகளை எடுத்துவிட்டால் சொற்களின் உச்சரிப்பிலோ, பொருளிலோ, இலக்கணத்திலோ எவ்விதக் குறையும் குற்றமும் ஏற்பட்டுவிடும் என்று தோன்றவில்லை, சுமார் 40 வருடங்களுக்கு முன்னால் இருந்தே நான் இதைக் கவனித்து வந்திருக்கிறேன், இந்தப்படி எழுத்து கோத்து அச்சடிக்கப்பட்டுள்ள ஒரு குறள் புத்தகத்தையும் நான் 40 வருடத்திற்கு முன்பே பார்த்திருக்கிறேன். இப்படிச் செய்வதில் மொத்தத்தில் 38 எழுத்துகள் (அதாவது, உயிரெழுத்து ஐ, ஓ ஆகிய 2ம் அவை ஏறும் மெய் எழுத்துகளில்  $2 \times 18 = 36$ ம் ஆக  $36 + 2 = 38$ ) ஞாபகத்திற்கும் பழக்கத்திற்கும் தேவை இல்லாத எழுத்துகள் ஆகிவிடும். (ஐ-அய்; ஓ-அவ் என எழுதலாம்) இவை தவிர, உயிர்மெய் எழுத்துகளில் தனிமாற்றம் பெற்றிருக்கிற ஐ, ஓ ஆகிய மூன்று எழுத்துகளுக்கும் தனி உருவம் தேவை இல்லாமல் ணா, நா, னா போல் ஆக்கிவிடலாம்.”

இவை போன்றே ‘ணை, லை, னை’ போன்ற எழுத்துகளில் இருந்த ‘துதிக்கை’யைத் தூக்கிவிட்டு, அவற்றையும் இயல்பாக்கிடுதலைப் பரிந்துரைத்தார்.

1935 முதலே தமது ஏடுகளான “பகுத்தறிவு”, “விடுதலை”, “குடிஅரசு”, “உண்மை” ஆகியவற்றில் தொடர்ந்து நடைமுறைப்படுத்தினார்-தான் கூறிய தமிழ் எழுத்துச் சீர்திருத்தத்தினை. நமது நாட்டில் பழைமை மோகமும், பிடிவாதமும் பல பண்டிதப் புலவர்களின் ரத்தத்துடன் கலந்தவைகளாகி விட்டன.

‘பழையன கழிதல், புதியன புகுதல்’ பற்றி படிப்பார்கள்; பதவுரை, பொழிப்புரை கூறுவார்கள். ஆனால், தாங்கள் மாறுவதற்கு தயங்குவார்கள். மாறத் துணிந்தவர்களையும் கேலி, கிண்டல் பேசுவர்.

இதுபற்றி கவலைப்படாது, ‘குடி செய்வார்க்கில்லை பருவம்’ என்ற போக்கோடு எழுத்துச் சீர்திருத்தத்தினை தனது ‘விடுதலை’ நாளேட்டில் தொடர்ந்து செயல்படுத்தி, தமிழ் எழுத்துப் புரட்சியை, மவுனப் புரட்சியாக நிகழ்த்தினார்.

1978-இல் தந்தை பெரியார் அவர்களது நூற்றாண்டு விழாவை ஓர் ஆண்டு முழுவதும் கொண்டாடிய அ.இ.அ.தி.மு.க. அரசும், அன்றைய முதலமைச்சருமான வள்ளல் எம்.ஜி.இராமச்சந்திரன் அவர்களும் தந்தை பெரியார் எழுத்துச் சீர்திருத்தத்தினை, அய்யாவுக்கு மரியாதை-‘காணிக்கை’ செலுத்தும் முகத்தான், ஆணையாக்கி, துணிந்து செயல்படுத்தினார்! (ஆணை பின்னிணைப்பில்)

வழக்கம்போல சில “தமிழ்ப் புலவர்கள்” சலசலப்புக் காட்டினர்.

அவர்கள் வாதம் அடிப்படையற்றவை என்பதைக் காட்டிட, பன்மொழிப் புலவர் தெ.பொ.மீனாட்சிசுந்தரனார் (மதுரை காமராசர் பல்கலைக் கழகத் துணைவேந்தர்) மற்றும் பேராசிரியர்கள் டாக்டர் அ.ச. ஞானசம்பந்தம், டாக்டர் சுப்புரெட்டியார் போன்ற தமிழ் மொழி வல்லுநர்களை அழைத்து சென்னையில் கருத்தரங்கம் (24-12-1978) நடத்தினோம். அவ்வுரைகளையும், டாக்டர் மு. வரதராசனாரின் கருத்துக்களையும் இணைத்து நூலாகவும் வெளியிட்டோம்.

தந்தை பெரியார் கல்லூரி காணா கிழவர்; என்றாலும் அவரது பகுத்தறிவு வீச்சாலும் தொண்டறத்தின் வீச்சாலும் பல பல்கலைக் கழகங்கள் அவரது சிந்தனைகளை ஆய்வு செய்கின்றன!

தட்டச்சு காலம் கடந்து கணினி மற்றும் கணினி அச்சுக் கோப்பு, ஒருங்குறி, செல்பேசிகளில் குரல் எழுத்து முறை வளர்ந்தோங்கிடும் இக்காலத்தில் தமிழ் எழுத்து மாற்றம் இல்லாமல் இருந்தால், உலகத்தின் வேகத்திற்கு நாம் ஈடுகொடுக்க முடியுமா? முடியாதே!

தமிழ்நாட்டு அரசுஆணை பிறப்பித்து, மாற்றம் செய்ததன் எதிரொலி நடவடிக்கையாக, மலேசியா மற்றும் சிங்கப்பூர் ஆகிய நாட்டுத் தமிழ் மக்களும் இதே தமிழ் எழுத்தினை அங்கே பின்பற்ற அந்தந்த நாட்டு அரசுகள் ஆணை பிறப்பித்துச் செயலாற்றி வருகின்றன!

தந்தை பெரியாரின் ‘மண்டைச் சுரப்பை உலகு தொழும்’ என்று பாடிய புரட்சிக் கவிஞரின் பொன் வரிகளுக்குச் சான்று பகருகின்றன!

அறிவியல் அடிப்படையில் தமிழ் எழுத்து, இலக்கியங்கள் எல்லாம் அமைந்தால்தான் என்றும் வாழும் தமிழ் எனப்படும் எமது மொழி, எங்கும் வாழும் தமிழாக ஆகும் - புதிய புரட்சி ஏற்படும்!

எட்டு கோடி தமிழர்கள் புதுமைகளைத் துய்க்க வேண்டும் என்றால், காலத்தால் ஏற்படும் மாற்றங்களை ஏற்றாக வேண்டும்!

எழுத்தாணி, ஓலைச் சுவடியை நம்முடைய பழைமைச் சொத்து என்பதால், அச்சுக்கலை கூடாது; இன்றைய கணினி, செல்பேசி கூடாது என்று புறக்கணிப்பது பெரும் பேதமை அல்லவா?”

தமிழில் 247 எழுத்துகள் என்று சொல்லப்பட்டாலும், அவற்றில் தனி எழுத்துகளையும், குறிகளையும் கணக்கிட்டு, ஒரே ஒலிப்பு முறையுடைய எழுத்துகளுக்கு ஒன்று போல் வடிவம் தந்து எளிமைப்படுத்த வேண்டும் என்பதைப் பெரியார் வலியுறுத்தினார்.

பெரியார், தன் கருத்து உள்பட அனைத்தும் வளர்ச்சிக்கும், மாறுதலுக்கும், பகுத்தறிவுக்கும், பயன்பாட்டுக்கும் உட்பட்டே இருக்க வேண்டுமென்று கூறினார். தன்னுடைய கருத்தையே மாற்றிக் கொண்டிருப்பதையும், அந்த மாற்றங்கள் சுயநலத்திற்காகவோ சுயலாபத்திற்காகவோ மேற்கொள்ளப்படுவன அல்ல என்பதையும் அவர் தெளிவுபடுத்தியிருக்கிறார்.



பெரியார் பரிந்துரைத்த எழுத்துச் சீர்திருத்தங்களில் நேரடியாக ஒரு சில எழுத்துகள் மட்டும் அரசு ஆணையின் படி ஏற்கப்பட்டிருந்தாலும், மிகப் பெரும்பாலானவை தத்துவத்தின் அடிப்படையிலும், தேவை, தொழில்நுட்ப வளர்ச்சி அடிப்படையிலும் இயல்பாகவே பயன்பாட்டுக்கு வந்துவிட்டன.

இகர, ஈகார உயிர்மெய் எழுத்துகளுக்குத் தனிக்குறிகள் தோற்றுவிக்கப்பட வேண்டும் என்றார். அச்சுக்கோப்பு நடப்பிலிருந்த அந்தக் காலகட்டத்தில் இகர, ஈகார உயிர் மெய்களுக்கு 36 எழுத்துகள் தனியாகத் தேவைப்பட்டன. இன்று அந்நிலை இல்லை. கணினி வழியோ, குரல் வழியோ தட்டச்சு செய்யப்படும் இக்காலகட்டத்தில் இகர, ஈகாரக் குறிகள் எளிமையாகிவிட்டன. உகர, ஊகாரத்திற்கான 36 எழுத்துகளுக்கும் இரண்டு குறிகளைக் கொண்டு எளிமையாக்கிட முடியும் என்று முனைவர் வா.செ.குழந்தைசாமி அவர்கள் தலைமையில் விவாதித்து, அதனை எளிதில் கணினியில் செய்துவிடலாம் என்பதை செயல்முறைப்படுத்தியும் காட்டியுள்ளோம். அவற்றை நடைமுறைக்குக் கொண்டுவருதல் பற்றி சிந்திப்பதும் இன்றியமையாததாகும்.

அச்சுக்கோப்புக் காலத்தில் மெய்யெழுத்துகளுக்குத் தனியாக (க், ங், ச்) 18 கட்டைகள் தேவைப்பட்டன. எனவே அவற்றுக்குப் பதிலாக அகர உயிர் மெய்களுக்கு அருகில் ஒரு குறியைப் பயன்படுத்தி அவற்றை மெய்யெழுத்துகளாகக் கருதலாம் என்று பெரியார் பரிந்துரைத்தார். ஆனால், அதன் பின்னான தொழில்நுட்ப வளர்ச்சியில் எழுத்துகளின் மீது புள்ளி வைத்து எழுதுவது எளிமையாக நிறைவேறிவிட்டது.

இப்படி பயன்பாட்டுப் பிரச்சினையில் என்னென்ன நோக்கங்களுக்காக பெரியார் எழுத்துச் சீர்திருத்தத்தில் பல்வேறு கட்டங்களைப் பரிந்துரைத்தாரோ, அவை நடைமுறைக்கு வந்துகொண்டிருப்பதைப் பார்க்கிறோம். எழுத்துகளைக் கற்கும் முறையில் சிரமம் இருக்கக் கூடாது என்னும் நோக்கில் இன்னும் நாம் தொடர்ந்து பயணம் செய்தாக வேண்டும்.

நம் கல்வியாலும், சமூகநீதியாலும், அறிவியல் வளர்ச்சியாலும், மாதத் தொலைவில் இருந்த தேசங்கள், நேரத் தொலைவில் வந்திருக்கின்றன. மனிதனின் கால்கள் நீண்டு, உலகம் சுருங்கியிருக்கிறது. பெயர் அறியா நாடுகளுக்கும் நம் தமிழ் இளைஞர்கள் சென்று சேர்ந்திருக்கிறீர்கள். அயல்நாடுகளில் வாழும் நம் அடுத்தடுத்த தலைமுறைகள் தங்களுக்கு அவசியமில்லை என்று தமிழில் ஆர்வம் செலுத்தாமல் போய்விடக் கூடாது என்பதற்காக தமிழ்ப் பள்ளிகளை உருவாக்கி, இணையம், கணினி இவற்றின் வழியே பயிற்சிகளை நடத்தி தமிழைக் கொண்டு போய்ச் சேர்க்கிறோம். ஆனால், அப்படிச் கற்பதற்கும் தமிழ் எளிமையாக இருக்க வேண்டும் என்பது அடிப்படையல்லவா? அதற்கு தமிழ் எழுத்துச் சீர்திருத்தம் அவசியப்படுகிறதல்லவா?

இதோ பெரியார் முடிக்கிறார்: “அறிஞர்களும், பண்டிதர்களும் தீர்க்கமாய்ச் சிந்தித்து ஒரு முடிவுக்கு வர வேண்டும். எப்படியும் தமிழ்மொழி எழுத்துகள் குறைக்கப்பட்டாக வேண்டும். அச்சுக் கோப்பதற்கும், டைப் அடிப்பதற்கும் ஆங்கிலத்தைப் போல் இலகுவாக்கப்பட வேண்டும் என்பதும், கற்கும் பிள்ளைகளுக்கும் 3 மாதத்தில் படிக்கத் துவக்கலாம் என்பதும் தான் நமது ஆசை.”

நாம் தொடர்வோம்!

- **கி.வீரமணி** எம்.ஏ., பி.எல்., டி.லிட்.,

வேந்தர், பெரியார் மணியம்மை அறிவியல் மற்றும் தொழில்நுட்ப நிறுவனம்  
(நிகர்நிலைப் பல்கலைக்கழகம்), வல்லம், தஞ்சாவூர்.

=====

Paper Number	Title of the Paper	Page Number
1.	Tamil Talking Box: An Introduction to the Conversational Partner for Language and Literature Learners  தமிழ்ப் பேசுப் பெட்டி வழி இரண்டாம் மொழி மற்றும் இலக்கியம் கற்போருக்கான உரையாடல் வடிவமைப்பு Vasu Renganathan, University of Pennsylvania	1
2.	Corpus Development for Malaysian Tamil  மலேசியத் தமிழ்த் தரவக மேம்பாடு C.M. Elantamil, Saravanan Ramachindran, Neelavathi Samykanu University Malaya (UM), Malaysia	8
3.	Emotionality in Suicide Notes N. Nirmeen, N. Vijayan, Department of Linguistics, Bharathiar University (BU), India.	16
4.	Opinion Mining on Tamil Movie Reviews Using BERT – A Study Syam Mohan E, R. Sunitha, Amudha T. K	22
5.	Memory Based Learning of Tamil Morphology K.Rajan	29
6.	Computer-Assisted Learning System for the Tamil Grammar Punarial Senthamizh Selvi S, Anitha R	34
7.	An Exotic Natural Language Processing Technique for Ancient Tamil Inscriptions: A Linguistic Approach Ezhilarasi S, UmaMaheswari P	41
8.	The Efficacy of the Jamboard Virtual Learning Strategy in Overcoming Grammatical Errors in Tamil  இலக்கணப் பிழைகளின்றி தமிழ் எழுதிட ஜேம்போர்ட் (JAMBOARD) வழி மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறை புஷ்பராணி சுப்ரமணி செல்வன்	49
9.	பைத்தான் நிரல்மொழி மூலம் விக்கிமூலம் இயங்கும் முறைகள் (Ways to run a Tamil wikisource through the Python programming language) சத்தியராஜ் தங்கச்சாமி, க. சண்முகம், தகவலுழவன்	55

10.	Automatic Question Generation using Centrality-based Keyword Extraction Approach for Tamil Text Senthilkumar P, Nandhini K	63
11.	Local Binary Pattern based Feature extraction and Recognition of Ancient Tamil Palm Leaf Manuscripts Characters using Neural Networks S Uma Maheswari , P Uma Maheswari	71
12.	Named Entity Recognition for Gynecological domain in Tamil using Machine Learning Algorithms M. Rajasekar, Angelina Geetha	82
13.	Diffusion of Meaningful Information in Tamil on OSN by Forming Communities with the Aid of HetNet Formation in 5G G.Ramasubramanian, S. Rajaprakash	92
14.	விஷேட தேவையுடைய மாணவர்களின் கற்றலை மேம்படுத்துவதில் கணினி வழிக் கற்பித்தலை மேற்கொள்வதில் ஆசிரியர்களின் பங்களிப்பு N.Koventhan, R.Thakshaayini	99
15.	Deep Analyses of the Evolution of Tamil characters from Stone Inscriptions: Digital Conservation Perspective Karishma V R, P Uma Maheswari	108
16.	மகிழ்வூட்டும் கற்றலுக்கான மின்னிலக்கப் புதிர் அறை Using Digital Escape Rooms to Make Learning Fun Raman Vimalan	119
17.	கம்பயில்லா உணரி வலைதளத்தில் பாதுகாப்பான மற்றும் திறமையான தரவு பரிமாற்றத்திற்கான மூன்று நிலை பாதுகாப்புகள் இரா.நேசமலர், முனைவர் கா.இரவிக்குமார்	124
18.	Mapping Comparative constructions in Tamil and English for Machine Translation  Dhanalakshmi V, Rajendran S	130
19.	An Efficient Approach for Computer Aid Teaching and Learning Using DEEDs Lab Mr.Dharmaraj, Kirubakaran, Anastraj	137

20.	Teachers' Willingness to Use Instructional Technology For Teaching and Learning	145
	கற்றல் கற்பித்தலில் தொழில்நுட்ப அறிவுறுத்தலைப் பயன்படுத்துதலில் ஆசிரியர்களின் முனைப்பு Tasaratha Rajan Anamalai , Maizatul Hayati Mohamad Yatim	
21.	Efficient Technique for Corpus Creation of Code-Mixed Data on Tamil and English Text from Social Forum	153
	M.Sangeetha, K.Nimala	
22.	Digital Library and its Features	165
	J.Lingeswaran, P.Mangayarkarasi	
23.	Tools for constructing AI/ML solutions in Tamil	170
	Abdul Majed Raja RS, Muthiah Annamalai	
24.	கணினிவழி அகரநிரல் உருவாக்கல் நீச்சல்காரன் சே. இராஜாராமன்	179
25.	சித்த மருத்துவக் களத்தின் சமூக-மூலப்ப ாருண்மமயியல் அமமப்பும் தகவவல் மீட்எஸ். வீர அழகெரி, எஸ். இரொசசந்திரனபும்	185
26.	கல்வியில் விளையாட்டுகள்: தமிழ்க் கற்றல் கற்பித்தலில் சொல் விளையாட்டுகளை ஒரு துணைக் கருவியாகப் பயன்படுத்துதல் முகிலன் முருகன் - Muhelen Murugan	196
27.	Corpus Analysis of ditransitive verb “koṭu” in Modern Tamil for Information Retrieval System	197
	A.Murugaiyan	
28.	இயந்திர மொழிபெயர்ப்புக்காகன அகரொதி உருவாக்கம் (Lexicon for Machine Translation)	198
	P.Kumar	
29.	Migrating TamilPesu to Cloud based Deployment	199
	Surendhar Ravichandran, T. Shrinivasan, Muthiah Annamalai	
30.	Natural Language Resources in Tamil	200
	Mohammed Afsal	
31.	Symmetries in Number Forms of Tamil and Dravidian Languages	201
	Muthiah Annamalai	
32.	Linguistic issues in machine translation in Tamil	202
	தமிழ்மொழியில் இயந்திர மொழிபெயர்ப்பில் மொழியியல் சிக்கல்கள் Selvajothi Ramalingam	





## Tamil Talking Box: An Introduction to the Conversational Partner for Language and Literature Learners

தமிழ்ப் பேசுப் பெட்டி வழி இரண்டாம் மொழி மற்றும் இலக்கியம்  
கற்போருக்கான உரையாடல் வடிவமைப்பு

Vasu Renganathan, University of Pennsylvania

[vasur@sas.upenn.edu](mailto:vasur@sas.upenn.edu)

<https://www.sas.upenn.edu/~vasur/project.html>

### சுருக்கம்

#### Keywords:

- A ரோவர்
- B பேச்சு அறியும் தொழிற்நுட்பம்
- C கணினி மொழியியல்
- D இயற்கைமொழித் தொழிற்நுட்பம்
- E தமிழ் கற்றல் மற்றும் கற்பித்தல்

தமிழ்க் கணினி ஆய்வு என்பது தமிழ்த் தொழிற்நுட்பம் 5.0 என்னும் ஆய்வாக மாறுவதை இக்காலக்கட்டத்தில் காணலாம். கணினி வழி என்பதை கணினி மனிதன் என்னும் நோக்கு இப்பொழுது பரவலாக நடைபெறுகிறது. இவ்வகையில் [robot.tamilnlp.com](http://robot.tamilnlp.com) என்னும் தளத்தில் தமிழ் இயந்திர ரோவர் வழி தமிழ் மொழிக் கற்றல் மற்றும் கற்பித்தலுக்கான உரையாடல் வடிவமைத்தல் மற்றும் தமிழ் இலக்கியங்களைக் கற்றுக்கொள்ளல் ஆகியவற்றுக்கான வழிமுறைகள் விளக்கப்பட்டுள்ளன. இவ்வகை ரோவர் இயந்திரங்களைப் பற்றிய தொழிற்நுட்ப விளக்கத்தினை [https://uttamam.org/papers/21\\_32.pdf](https://uttamam.org/papers/21_32.pdf) என்னும் கட்டுரை வழியாக அளிக்கப்பட்டுள்ளது. இக்கட்டுரை இவ்வாராய்ச்சியின் தற்போதைய நிலையை விளக்குகிறது. மனிதர்களின் பேச்சை இயந்திரம் புரிந்துகொண்டு அப்பேச்சு வழி இடும் ஆணைகளைச் செய்யவும் கேட்கப்படும் விளக்கங்களுக்கு பதில் கூறுவதையும் மனித பேச்சைப் புரிந்துகொள்ளும் நிரலிகள் எனலாம். மனிதப் பேச்சை அப்படியே புரிந்துகொண்டு மனிதர்களுக்கு ஈடாக இயந்திரம் பேச்சு வழியாகச் செயல்படுவது என்பது முழுவதும் இயலாத நிலையில் ஒரு வரையறைக்குள் இயந்திரம் செயற்படுவது சாத்தியமே. இவ்வகையில் இக்கட்டுரை அத்தகைய சாத்தியங்கள் என்ன என்பதை விளக்குகிறது. இரு வகை இயந்திரங்களை இங்கு விளக்குகிறோம். முதல்வகையில் கட்டளைக்கேற்ப நகரும் திறமை கொண்ட ரோவர். மற்றொரு வகை பேச்சைப் புரிந்துகொண்டு தமிழ் உரையாடலில் ஈடுபடும் ஒரு இயந்திரப் பெட்டி. இவ்விரு வகை அணுகுமுறைகள் வழி தமிழ் மொழி மட்டுமல்லாது இலக்கியங்களைப் படிப்பதற்குமான வழிமுறைகளைப் பற்றி இக்கட்டுரை விளக்குகிறது. இக்கட்டுரை வழி முக்கியமாக இரண்டாவது வகை தமிழ்ப் பேசுப் பெட்டி ஒன்றை அறிமுகம் செய்கிறோம்.

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

#### Corresponding Author:

Vasu Renganathan  
Department of South Asia Studies  
University of Pennsylvania, USA  
Email: [vasur@sas.upenn.edu](mailto:vasur@sas.upenn.edu)

### 1. அறிமுகம்

இயந்திர மொழியியலும் தொழிற்நுட்பம் ஐந்தாம்

இருபத்தோராம் நூற்றாண்டில் இயந்திர மொழியியலும் அதற்கான தொழிற்நுட்பமும் தொழிற்நுட்பம் ஐந்தாம் நிலை என்னும் போக்கில் நகர்ந்துவருகிறது. இதுகாறும் தமிழ் எழுத்துருக்களை மின்வழிப் பயன்படுத்தல், அதற்கான தரம் மட்டும் கட்டுப்பாடு போன்ற ஆய்வுகளில் சிறப்பான வளர்ச்சியைக் கண்ட நிலையில் தரவு மொழியியல், இயந்திரம் வழி மொழிக் கற்றல் மற்றும் கற்பித்தல் போன்ற தொழிற்நுட்பங்களில் பெரிய மாற்றத்தைக் கடந்த நூற்றாண்டில் தமிழ் மொழி உட்பட பல மொழிகளில் கண்டோம். எழுத்தை உணர்ந்து பேச்சுக்கு மாற்றும் தொழிற்நுட்பமும் பெரிய அளவில் முன்னேற்றம்

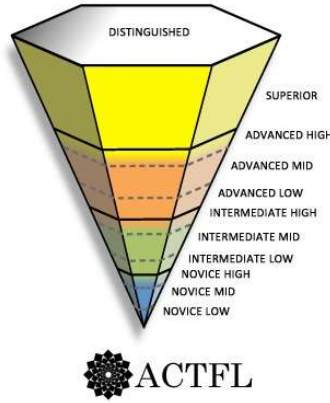
கண்டுள்ளது எனலாம். ஆனால் பேச்சை உணர்ந்து உரையாகவோ பேச்சாகவோ மாற்றும் தொழிற்நுட்பம் முழுமையான வளர்ச்சியை இன்னமும் நாம் பெறவில்லை எனவே கூறலாம். இவ்வகை ஆய்வுகள் இன்னமும் தொடர்ந்து வளர்ந்துவரும் நிலையில் இக்கட்டுரை வழி தமிழுக்கான தமிழ்ப் பேசுப் பெட்டி ஒன்றை அறிமுகம் செய்து விளக்குகிறது இக்கட்டுரை. பேச்சு உணர் என்னும் தொழிற்நுட்பத்தை முதன்மையாகக் கொண்டு செய்யப்படும் இத்தொழிற்நுட்பம் வழி நம்மால் செய்யமுடிகிற சாத்தியக் கூறுகளை அறிந்துகொள்ளல் அவசியம். இத்தொழிற்நுட்பம் வழி சிறு சிறு வாக்கியங்களை இயந்திரம் புரிந்துகொண்டு அதற்கு ஏற்றவாறு செயல்படும் வழிவகை இரண்டாம் மொழியாகத் தமிழ்மொழியைக் கற்பது மற்றும் இலக்கியங்களைப் பற்றித் தெரிந்துகொள்ளல் ஆகிய முயற்சிகள் பற்றியும் அலசுவது இன்றியமையாததாகிறது. தமிழ் இலக்கியங்கள் பலவற்றை வாய்மொழியாக அறிந்துவந்த முறைகளை அறிவோம். ஆத்திசூடி, திருக்குறள், சைவ மற்றும் வைஷ்ணவ இலக்கியங்கள் பலவற்றை வாய்மொழியாக அறிந்துவந்துள்ளமை பற்றி அறிவோம். ஆனால் இவற்றை முறையாக மனப்பாடம் செய்யும் முறை ஒன்றை இதுவரை நாம் வழிவகுத்ததில்லை. இவ்வகையில் இத்தொழிற்நுட்பம் வழி இதற்கான வழிமுறையை விளக்க முற்படுகிறது இக்கட்டுரை.

## 2. பின்னணி

“என் பேரு தமிழு: (en pēru tamīlu) A speech recognition for Tamil” (Renganathan 2021) என்ற கட்டுரை வழியும் [robot.tamilnlp.com](http://robot.tamilnlp.com) என்னும் இணையப் பக்கம் வழியும் ஆர்டுவினோ மற்றும் ஈசிவியார் ஆகியத் தொழிற்நுட்பம் வழி எவ்வாறு தமிழ்ப் பேச்சைப் புரிந்துகொண்டும் செயற்படும் ரோவர்களைப் பற்றி விளக்கினோம். இவ்வாராய்ச்சியின் தொடக்கமாக இக்கட்டுரையும் இக்கட்டுரை வழி கொடுக்கப்படும் விளக்கமும் செயற்படுகிறது. முக்கியமாக இக்கட்டுரை வழி தமிழ் இலக்கியங்களை மனப்பாடம் செய்து கற்றுக்கொள்ளும் வழிமுறைகளை விளக்க முற்படுகிறோம். ஒலி நாடாக்கள் வழி தமிழ் மொழி உரையாடல் மற்றும் இலக்கியங்களைப் பதிவு செய்து திரும்பத் திரும்பக் கேட்டு மனப்பாடம் செய்துகொள்ளும் வழிமுறைகளை அறிவோம். ஆனால் ஒலிநாடா வழி உள்ள தொழிற்நுட்பங்கள் பேச்சைப் புரிந்துகொண்டு செயற்படும் வழிகள் இல்லை. இந்த ஒரு குறிப்பிட்ட தொழிற்நுட்பம் இருந்தால் எவ்வாறு இலக்கியங்களை மனப்பாடம் செய்தல் மற்றும் புரிந்துகொள்ளல் ஆகிய வழிமுறைகள் மேம்படும் என்பதோடு எவ்வாறு தமிழ்த் திறமையை வளர்த்துக்கொள்ள விரும்பும் தமிழ்ப் பயில்வோர் மற்றும் தமிழ் ஆர்வலர்கள் இத் தமிழ்ப் பேசுப் பெட்டியைப் பயன்படுத்திக்கொள்ளலாம் என்பதையும் விளக்குகிறது இக்கட்டுரை.

## 3. மொழித் திறன் மற்றும் தமிழ்த் திறன்

மொழித் திறன் என்பதை நான்கு நிலைகளில் விளக்குவர் மொழியியல் அறிஞர்கள் (காண்க <https://www.actfl.org/resources/actfl-proficiency-guidelines-2012>). முதல் நிலை, தொடக்க நிலை, இடைநிலை, முதுநிலை மற்றும் சிறப்பு நிலை என்னும் நான்கு நிலைகளைப் பற்றி விளக்கும் மொழியியல் அறிஞர்கள் முதல் நிலையிலிருந்து சிறப்பு நிலைக்கு நாம் நம் மொழித் திறமையைப் பல்வேறு வழிகளில் வளர்த்து வருகிறோம். குறிப்பிட்ட தாய்மொழிச் சூழலிலேயே கல்வி கற்று வளர்ந்து வரும் அனைவருக்கும் தாய்மொழியில் சிறப்பு நிலைத் திறனைப் பெறுவது மிகவும் எளிது. ஆனால் தாய்மொழிச் சூழலை விட்டு வேறு மொழிச் சூழலுக்கு கல்வி, பணி மற்றும் வேறு ஒரு காரணத்துக்காக இடம்பெயர்வோரால் மொழித் திறனை முறையாக வளர்த்துக்கொள்ள இயலாமல் போகும். ஆனால் தொழிற்நுட்பம் வழி சிறப்பு மொழித் திறனை வளர்த்துக்கொள்ளுதல் இக்காலக்கட்டத்தில் இயல்வதாக இருக்கிறது. கீழ்க்காணும் தலைகீழான பிரமிட் எனக் கூறப்படும் படத்தின் வழி ஆக்ட்பெல் நிறுவனத்தினர் இந்த நான்கு நிலைகளின் திறமையை விளக்குகின்றனர்.



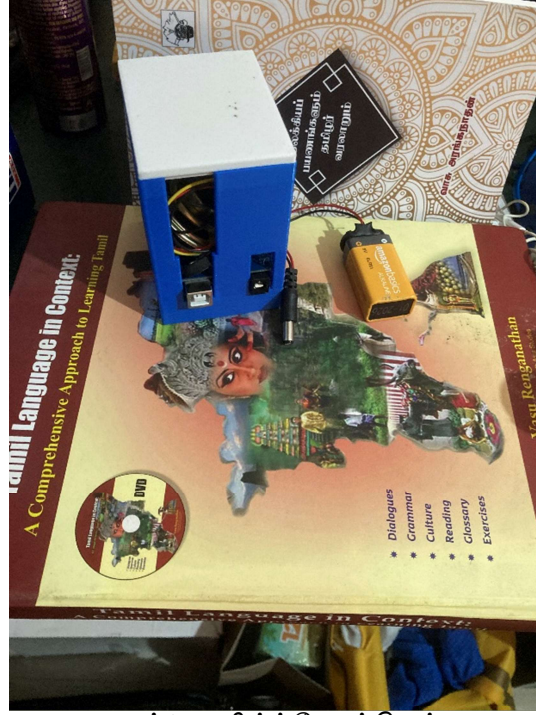
படம் 1: ஆக்ட்பெல் நிறுவனம்

நான்கு நிலைகளை வெவ்வாறு மொழித் திறமைகளின் நுண்ணிய திறமைகளின் அடிப்படையில் ஒவ்வொரு திறமையையும் கீழ்நிலை, மத்திய நிலை மற்றும் உயர்நிலை எனவும் பிரித்தறிகின்றனர். இக்கருத்தின் வழி ஒவ்வொருவரின் மொழித் திறமையையும் இந்தப் பத்து நிலையில் ஒன்றில் அறியலாம். தாய்மொழிப் புலமையைப் பொறுத்தவரையில் ஒருவரின் பேச்சு மற்றும் எழுதும் திறன் ஆகிய செயற்பாட்டுத் திறனையும் கேட்டல் மற்றும் படித்தல் ஆகிய மந்தநிலைத் திறனையும் வெவ்வேறு உத்திகளின் அடிப்படையில் அளவிடலாம். இந்த வகை அளவிடும் முறைகளையும் ஆக்ட்பெல் நிறுவனத்தினர் வடிவமைத்து அமெரிக்காவில் தனி ஒருவரின் மொழித்திறனைப் பல்வேறு காரணங்களுக்காக அளவிட்டு வருகின்றனர். இவ்விளக்கங்களுக்குக் காண்க <https://www.actfl.org/assessment-research-and-development/actfl-assessments>. மாணவர்கள் தங்களின் பள்ளி, இளநிலைக் கல்லூரி மற்றும் முதுநிலைக் கல்லூரி வகுப்புகளில் வேற்று மொழியில் இடைநிலை மற்றும் உயர்நிலையில் திறமை பெற்றிருக்கவேண்டும் என்ற கட்டாய நிலை அமெரிக்காவிலும் வேறு சில நாடுகளிலும் இருக்கிறது. அதோடு இராணுவம், மருத்துவம் போன்ற பணிகளுக்கும் வேற்று மொழியில் குறிப்பிட்டத் திறன் இருக்கவேண்டும் என்றும் கட்டாயம் இருக்கிறது. இவர்கள் இந்த காரணத்துக்காக ஆக்ட்பெல் நிறுவனத்தி் அணுகுவர்.

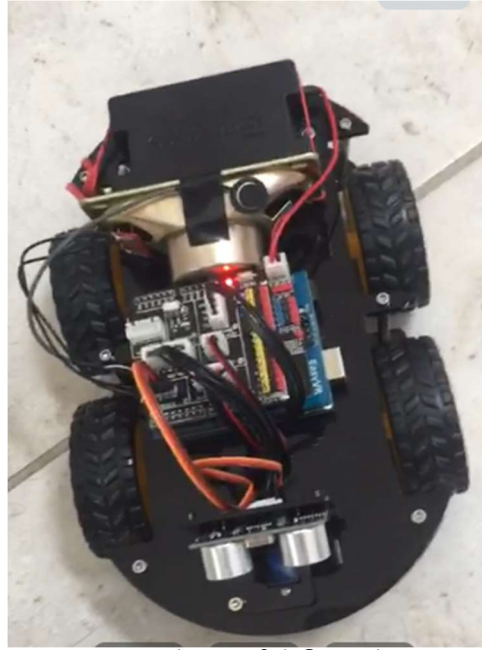
இச்சூழலில் தமிழ்த் திறன் என்பதை நாம் எவ்வாறு கணிக்கிறோம் இதற்கான உத்திகள் என்ன என்பதை அறியும் நிலையில் மேற்படி ஆக்ட்பெல் நிறுவனத்தையும் அணுகலாம். அவர்கள் தமிழுக்கான திறன் அளவிடு உத்தியையும் மற்ற மொழிகளின் அடிப்படையில் வகுத்துள்ளனர். <http://www.thetamilanguage.com> என்னும் வலைப்பக்கத்தின் வழி மேற்படி திறமைகளின் அடிப்படையில் முதல் நிலையிலிருந்து முதுநிலை வரையிலான பாடங்களை ஒலி மற்றும் ஒளி ஆகியவற்றின் வழி வடிவமைக்கப்பட்டுள்ளது. இவ்வலைப் பக்கம் மற்றும் <http://tamilnlp.com> என்னும் வலைப் பக்கங்களின் வழி தமிழ் மொழிக்கான வெவ்வேறு வசதிகளைப் பெற வாய்ப்பு ஏற்படுத்தப்பட்டுள்ளது.

#### 4. தொழிற்றுட்பம் வழி தமிழ்த் திறனை வளர்த்துக்கொள்ள வாய்ப்புகள்: தமிழ்ப் பேசுப் பெட்டி மற்றும் தமிழ் ரோவர்.

தொழிற்றுட்பக் காலக்கட்டம் ஐந்து என்னும் இக்காலக்கட்டத்தில் இயந்திரங்களின் வழி அனைத்து மொழித் திறனையும் வளர்த்துக்கொள்ளும் வாய்ப்புகள் இருக்கிற தருணத்தில் இக்கட்டுரையின் வழி தமிழ்ப் பேசுப் பெட்டி மற்றும் தமிழ் ரோவர் ஆகிய இரு தொழிற்றுட்பங்களை அறிமுகப்படுத்தப்படுகிறோம்.



படம் 2 – தமிழ்ப் பேசுப் பெட்டி



படம் 2 – தமிழ் ரோவர்

முன்னரே அறிமுகம் செய்யப்பட்ட தமிழ் ரோவர் பற்றிய விளக்கங்களை [robot.tamilnlp.com](http://robot.tamilnlp.com) என்னும் இணையப் பக்கம் மற்றும் Renganathan (2021) கட்டுரைகளின் வழி அறியலாம். தமிழ்ப் பேசுப் பெட்டி என்பது தொழிற்றுட்பம் வழி தமிழ் உரையாடலை மனிதனுக்கும் இயந்திரத்துக்கும் இடையே ஏற்படுத்தப் பயன்படுத்தப்படும் கணினி இயந்திரமாகும். இது ஆர்டுவினோ மற்றும் ஈசிஸீயார் ஆகிய இரு தொழிற்றுட்பங்களைக் கொண்டு வடிவமைக்கப்பட்டிருக்கிறது. இக்கருவியின் முக்கிய அங்கமாக ஒருவரின் பேச்சை ஒலிப்பதிவு செய்துகொண்டு அதன் வழி இயந்திரத்தோடு உரையாடும் வழிவகை செய்யப்பட்டிருப்பதே, பேச்சு உணர்த் தொழிற்றுட்பத்தின் வழி வடிவமைக்கப்பட்டிருக்கும் இவ்வியந்திரம் வழி தனி ஒருவரின் பேச்சை உணரும் திறனைக் கொண்டு மொழியில் பல்வேறு வகையான உரையாடல் வகைகளை அமைப்பது, இலக்கியங்களை மனப்பாடம் செய்வது என்னும் இருவகை நோக்கங்களை மனதில் கொண்டு இத்தொழிற்றுட்பம் வடிவமைக்கப்பட்டுள்ளது. எடுத்துக்காட்டாக அன்றாடச் சூழலில் ஒருவருக்கொருவர் அறிமுகம் செய்துகொள்ளும் உரையாடல், ஒரு இடத்திலிருந்து இன்னொரு இடத்துக்கு வழி கேட்டுத் தெரிந்துகொள்ளல் போன்ற சிறு சிறு உரையாடல்களை ஏற்படுத்துவது என்பது ஒரு புறம். இரண்டாவதாக ஆத்திசூடி, திருக்குறள், புறம், அகம் ஆகிய பாடல்களை

ஒவ்வொரு வாக்கியமாகப் பதிவு செய்து கொண்டு அவ்வாக்கியங்களை பயனாளரை உச்சரிக்க வைத்துப் பதிவு செய்துகொண்டு அவற்றைத் திரும்பத் திரும்ப சொல்லவைத்துப் பயிற்சி செய்வது என்பது என்பது இக்கருவியின் அடுத்தப் பயன்பாடு.

#### 4.1. அறிமுக உரையாடலும் தமிழ்ப் பேசுப் பெட்டியும்

தமிழ்ப் பேசுப் பெட்டியை உரையாடலில் ஈடுபடும் இன்னொருவராகக் கருத்தில் கொண்டு பின்வரும் உரையாடலில் ஈடுபடுவது:

தமிழ்ப் பேசுப் பெட்டி (தபெ): வணக்கம்  
 மாணவர்: வணக்கம்  
 தபெ: வணக்கம். எப்படி இருக்கிறீர்கள்?  
 மாணவர்: நன்றாக இருக்கிறேன்.  
 தபெ: ரொம்ப மகிழ்ச்சி.  
 மாணவர்: நீங்கள் எப்படி இருக்கிறீர்கள்?  
 தபெ: நானும் நன்றாக இருக்கிறேன்.  
 மாணவர்: உங்க பேரு என்ன?  
 தபெ: என் பேரு தமிழு  
 மாணவர்: உங்க ஊரு பேரு என்ன?  
 தபெ: என் ஊரு பேரு பிலடல்பியா  
 மாணவர்: கொஞ்சம் காப்பி குடிக்கிறீர்களா?  
 தபெ: வேண்டாம். நான் காப்பியெல்லாம் குடிக்கமாட்டேன்.  
 மாணவர்: நீங்கள் என்ன குடிப்பீர்கள்?  
 தபெ: நான் ஆரஞ்சு சூஸ் குடிப்பேன்.  
 மாணவர்: பின்னால் வாங்க  
 தபெ: சரி பின்னால் வறேன்.  
 மாணவர்: சுத்துங்க  
 தபெ: சரி சுத்துறேன்.  
 (உரையாடல் 1)

மேற்படி உரையாடலில் மாணவர் பயன்படுத்தப்படும் சொற்றொடர்க் கூறுகளை முன்னதாகவே பதிவு செய்துகொண்டு அமைக்கப்பட்டிருக்கிறது. மாணவரின் அத்தைகைய உரையாடலைக் கேட்கும்போது அதைப் பேச்சு உணரி வழியாக உணர்ந்து அதற்கான சரியான உரையாடலை இயந்திரம் பதிவு செய்யப்பட்டிருக்கும் ஒலிக்கோப்பு வழிக் கூறுகிறது. இதில் குறிப்பாக ஒரு தனி மாணவனின் உரையாடலை பயன்படுத்தப்படும் நிலையைக் காணலாம். எல்லோரும் இவ்வியந்திரத்தோடு உரையாடலில் ஈடுபடவேண்டும் என்றால் ஒவ்வொருவரும் இவ்வியந்திரத்துக்கு “சொல்லிக்கொடுங்கள்” என்னும் கட்டளையைக் கொடுத்துத் தங்களின் பேச்சைப் பதிவுசெய்துகொள்ளும் வாய்ப்பு உள்ளது. இதைக் கீழ்க்காணும் உரையாடலிலிருந்து அறியலாம்.

மாணவர்: சொல்லிக்கொடுங்கள்.  
 தபெ: சரி. நான் சொல்றதெ திருப்பிச்சொல்லுங்கள். பீப் கேட்டபிறகு சொல்லுங்கள்.  
 1) வணக்கம். (மாணவர் திருப்பிச் சொல்லவேண்டும்)  
 2) நன்றாக இருக்கிறேன்  
 3) நீங்கள் எப்படி இருக்கிறீர்கள்?  
 4) கொஞ்சம் காப்பி குடிக்கிறீர்களா?  
 5) நீங்கள் என்ன குடிப்பீர்கள்?  
 6) உங்க பேரு என்ன?  
 7) உங்க ஊரு பேரு என்ன?  
 8) .....

(உரையாடல் – 2: பதிவுசெய்துகொள்ளும் உரையாடல்)

இந்தப் பதிவுசெய்துகொள்ளும் உரையாடல் வழி மாணவர் உரையாடலைப் பதிவுசெய்துகொண்டு இந்த இயந்திரத்தோடு உரையாடலில் ஈடுபடுவதே இதன் நோக்கமாகும். குறிப்பாக இங்கு நோக்கவேண்டுவது என்னவெனில் இரண்டாம் மொழியாகத் தமிழ் மொழியைக் கற்கும் மாணவர்களுக்கென பல்வேறு திறனில் உரையாடல்களை ஏற்படுத்தி அவற்றை தனிதனி ஒலிக்கோப்புகளாகப் பதிவு செய்துகொண்டு மாணவர்கள் பேசவேண்டியதை அவர்களிடம் பேசுசொல்லிப் பதிவு செய்துகொண்டு பின்னர் அந்த உரையாடலில் திரும்பத் திரும்ப ஈடுபட்டு மாணவருக்கு அவ்வுரையாடலில் மிகுந்த திறனை வளர்த்துக்கொள்ள வாய்ப்பளிப்பதே இம்முறையின் முக்கிய நோக்கமாகும்.



#### 4.2. தமிழ்ப் பேசுப் பெட்டியோடு இலக்கியங்களை அறிந்துகொள்ளல்

இம்முறையிலேயே எந்த ஒரு இலக்கியத்தையும் ஒரு மாணவர் பழக்கப்படுத்திக்கொள்ளவேண்டுமென்றால் அதற்கான ஒலிக்கோப்புகளைப் ஒவ்வொரு வாக்கியமாகப் பதிவு செய்துகொண்டு மாணவரிடம் அவற்றைப் பதிவுசெய்துகொண்டு பின்னர் பேச்சு உணர் மூலம் அவ்விலக்கியங்களைப் பழக்கப்படுத்திக்கொள்ளும் வாய்ப்பை ஏற்படுத்துவதே இத்தொழிற்றுட்பத்தின் நோக்கமாகும். ஆத்திசூடியையும், புறநானூறு பாடல்களையும் இத்தகைய தொழிற்றுட்பத்தில் தமிழ் ரோவர் மூலம் பயன்படுத்தியுள்ளமையை <http://robot.tamilnlp.com/index1.php>, <http://robot.tamilnlp.com/index3.php>, <http://robot.tamilnlp.com/index4.php> ஆகிய பக்கங்களின் வழி அறியலாம்.

இத்தொழிற்றுட்பம் வழி இலக்கியங்களைப் பழக்கப்படுத்திக்கொள்ளும் வழிமுறைகளில் முதலில் பதிவு செய்துகொள்ளும் நிலையின் வழி மாணவருக்கு இலக்கியம் அறிமுகப்படுத்தப்படுகிறது. பின்னர் ஒவ்வொரு வரியாக உரையாடல் வழி மாணவரை ஈடுபடுத்தும்போது மாணவருக்கு அவ்விலக்கியத்தைப் பழக்கப்படுத்திக்கொள்ளும் வாய்ப்பு இருக்கிறது. எடுத்துக்காட்டாக ஆத்திசூடியைக் கற்றுக்கொடுக்கும் முறையைக் கீழ்க்காணும் உத்தி வழி அறியலாம்.

மாணவர்: ஆத்திசூடி சொல்லிக்கொடுங்கள்  
தபெ: சரி. நான் சொல்றதெ திருப்பிச்சொல்லுங்க. பீப் கேட்டபிறகு சொல்லுங்க  
தபெ: அறம் செய விரும்பு (மாணவர் கேட்டுத் திருப்பி சொல்லவேண்டும்)  
தபெ: ஆறுவது சினம்  
தபெ: இயல்வது கரவேல்

.....

இவ்வரையாடல் பதிவுக்குப் பிறகு மாணவரும் தமிழ்ப் பேசுப் பெட்டியும் ஆத்திசூடியை ஒதும்போது கீழ்க்கண்ட சூழற்களை நோக்கலாம்.

தபெ: அறம் செயவிரும்பு  
மாணவர்: அறம் செயவிரும்பு (முன்னால் பதிவு செய்தது போலவே உச்சரிக்கவேண்டும். இல்லையெனில் தபெ மறுபடியும் அறம் செய விரும்பு என்ற வரியையே திருப்பிச்சொல்லிக்கொண்டிருக்கும். இங்கு நோக்கவேண்டியது என்னவெனில் பதிவு செய்யும் போது மாணவர்கள் கவனமாகக் கேட்டுச் சரியாகப் பதிவுசெய்யவேண்டும். ஒருவாறு இதில் புலமை பெற்ற பின் பின்வரும் உரையாடற் சூழலை ஏற்படுத்திக்கொள்ளலாம்.

தபெ: அறம்செயவிரும்பு  
மாணவர்: ஆறுவது சினம்  
தபெ: இயல்வது கரவேல்  
மாணவர்: ஊக்கமது கைவிடேல்  
தபெ: எண் எழுத்து இகழேல்  
மாணவர்: ஒப்புறவு ஒழுகு  
தபெ: அறம் செய விரும்பு

இப்பயிற்சியில் முக்கியமாக பின்வரும் ஆத்திசூடி வரிகளை முதலில் ஒலித்துக்காட்டி மாணவர்களைப் பதிவுசெய்துகொள்வது முதற்கட்டத்தில் நடக்கவேண்டும்.

- 1) அறம் செய விரும்பு
- 2) ஆறுவது சினம்
- 3) இயல்வது கரவேல்
- 4) ஈவது விலக்கேல்
- 5) உடையது விளம்பேல்
- 6) ஊக்கமது கைவிடேல்
- 7) எண் எழுத்து இகழேல்
- 8) ஏற்பது இகழ்ச்சி
- 9) ஐயம் இட்டு உண்
- 10) ஒப்புறவு ஒழுகு
- 11) ஒதுவது ஒழியேல்
- 12) ஔவியம் பேசேல்
- 13) அஃகம் சுருக்கேல்

இவ்வரிகளைப் பதிவு செய்துகொண்டபின் பேச்சு உணரி ஒவ்வொரு வரியைக் கேட்டதும் அடுத்த வரி எது என அறிகிறது. மாணவர் ஒவ்வொரு வரியையும் சரியாக உச்சரித்துப் பழகியபின் முதல்வரியை தமிழ்ப் பேசப் பெட்டி கூற அடுத்த வரியை மாணவர் கூறத் திரும்பத் திரும்ப தமிழ்ப் பேசப் பெட்டியும் மாணவரும் ஆத்திசூடி வாசித்துக்கொண்டிருக்கும் நிலையை ஏற்படுத்தலாம். கேட்டல், உச்சரித்தல் மற்றும் மனப்பாடம் செய்தல் என்னும் இந்த மூன்று முக்கிய நிலையில் இத்தொழிநுட்பம் வழி இலக்கியங்களை மாணவர்கள் அறிந்துகொள்ளும் வாய்ப்பை ஏற்படுத்தலாம். இம்முறை ஏறக்குறைய வாய்வழி இலக்கியத்தை பழங்காலத்தில் பரப்பிய முயற்சி போன்றதே. குருவின் வழி சீடர்கள் வாய்மொழி இலக்கியத்துக்கு அறிமுகம் பெற்றுப் பின்னர் அவ்விலக்கியத்தைத் தினமும் தொடர்ந்து பயன்படுத்தி அதில் புலமை பெற்றனர். இதே உத்தியை இங்கு தொழிநுட்பம் வழி அறிமுகப்படுத்துகிறோம். இங்கு தொழிநுட்பம் குருவாக அறிமுகப்படுத்தப்படுகிறது என்பதை நோக்கவேண்டும்.

##### 5. முடிவுரை

குறிப்பாக இதற்கான நிரலிகளையும் இயந்திரத் தொழிநுட்பமும் உருவாக்கப்பட்டிருப்பதால் எந்தவொரு உரையாடலையும் இலக்கிய வரிகளையும் குறிப்பிட்ட mp3 ஒலிக்கோப்பு வழி இணைக்க வாய்ப்புள்ளது. இவ்வகையில் ஆசிரியரின் பங்கு இவ்வகை உரையாடல்களைத் தயார் செய்துகொள்வதும் அவற்றுக்கான உரையாடல்களில் மாணவர்களை ஈடுபட செய்யவேண்டும் என்பதே. இத்தொழிநுட்பம் வழி தமிழ் மொழி மட்டுமல்லாது மற்ற மொழிகளுக்கான பயிற்சிகளையும் ஏற்படுத்திக்கொள்ளும் வாய்ப்பு உள்ளது என்பதை அறியலாம். குரு மற்றும் சீடர்கள் வழி வாய்மொழியாக வளர்ந்துவந்த இலக்கியங்கள் இக்காலக்கட்டத்தில் தொழிநுட்பம் வழி வளர பல்வேறு வாய்ப்புகள் இருக்கின்றன என்பதைச் சுட்டிக்காட்டும் நோக்கமே இக்கட்டுரை எனலாம். பல்வேறு நிலைகளில் உரையாடல்கள் மற்றும் இலக்கியங்களை ஒலிக்கோப்பாகப் பதிவு செய்துகொள்ளும் நிலையில் இத்தொழிநுட்பம் குருவாகச் செயல்பட்டு இதைப் பயன்படுத்தும் சீடர்கள் பலரை உருவாக்கும் நிலையை ஏற்படுத்தலாம்.

##### மேற்கோள் கட்டுரைகள் மற்றும் நூற்கள்

Arden. A. H. 1942. A progressive grammar of the Tamil language. Madras: Christian Literature Society.

அரங்கநாதன், வாசு (2019). “சுழற்சி முறை ஒலியனியல் மாற்றங்கள்.” மொழியியல் 1:1 2019, மொழியியற் கழகம், கோயம்புத்தூர்.

அரங்கநாதன், வாசு (2020). இக்காலத் தொல்காப்பிய மரபு. நியூ செஞ்சரி புத்தக நிலையம், சென்னை.  
Kiparsky, Paul (1982). ‘From Cyclic Phonology to Lexical Phonology’, in Harry van der Hulst and Norval Smith (eds), The structure of phonological representations (vol. 1). Dordrecht: Foris, 131-75.

Renganathan, Vasu (2021). என் பேரு தமிழு: (en pēru tamīlu) A speech recognition for Tamil” (Renganathan 2021), 20தமிழ் இணைய மாநாடு, உத்தமம் நிறுவனம். [http://www.uttamam.org/papers/21\\_32.pdf](http://www.uttamam.org/papers/21_32.pdf)

Renganathan, Vasu (2016). Computational Approaches to Tamil Linguistics, Cre-A. Chennai, India.

Renganathan, Vasu (2016). Computational Approaches to Tamil Linguistics: Scopes and Prospects. In the proceedings of the 15th Tamil Internet Conference, Gandhigram Rural University, Dindigul, Tamil Nadu. ([http://www.uttamam.org/papers/16\\_02.pdf](http://www.uttamam.org/papers/16_02.pdf)).

Renganathan, Vasu (2014). Computational Phonology and the Development of Text-to-Speech Application for Tamil. In the Proceedings of the International conference on Tamil Internet, 2014, Pondicherry, India. ([http://www.uttamam.org/papers/14\\_35.pdf](http://www.uttamam.org/papers/14_35.pdf)).

Renganathan, Vasu (2001). Development of Morphological Tagger for Tamil, In the Proceedings of the International Conference on Tamil Internet 2001, Kuala Lumpur, Malaysia. ([http://www.uttamam.org/papers/01\\_34.pdf](http://www.uttamam.org/papers/01_34.pdf))

சண்முகம், செ.வை. 2005. மொழி ஆய்வு. மணிவாசகர் நூலகம். சென்னை.



## Corpus Development for Malaysian Tamil

### மலேசியத் தமிழ்த் தரவக மேம்பாடு

C.M. Elanttamil<sup>1</sup>, Saravanan Ramachindran<sup>2</sup>, Neelavathi Samykanu<sup>3</sup>  
University Malaya (UM), Malaysia

#### சுருக்கம்

மொழியியல் துறையில் தரவக மொழியியலும் கணினி மொழியியலும் மிகப்பெரிய வளர்ச்சியைப் பெற்றுள்ளன. தகவல் தொழில்நுட்பத்தின் தொடர் வளர்ச்சியானது, மொழி ஆய்விலும் கற்றல் கற்பித்தலிலும் பெரும் மாற்றத்தைக் கொண்டுவந்துள்ளது.

தரவக மொழியியல் என்பது மொழியியல் ஆராய்ச்சிக்காகச் சேகரிக்கப்பட்ட "நிகழ் நேரத்" (real time) தரவுகள், தரவுத்தளங்களில் கணினிமயமாக்கப்பட்டு, அதன் அடிப்படையில் மேற்கொள்ளப்படும் மொழியியல் ஆய்வு ஆகும். இது மொழிப் பயன்பாட்டு அடிப்படையிலான துல்லியமான ஆய்வுகளுக்குப் பெரும் துணையாக அமைகிறது.

ஒரு மொழிக்கான தரவகத்தை அறிவியல் அடிப்படையில் உருவாக்கவும் உருவாக்கப்பட்ட தரவகத்தை முறையான மொழியியல் ஆய்வு வழிமுறைகளைப் பின்பற்றி ஆய்வுசெய்யவும் தேவையான அறிவைத் தரவக மொழியியல் அளிக்கிறது. ஒரு மொழியில் காணப்படும் நுணுக்கமான கூறுகளையும் இதுவரை கண்டறியப்படாத மொழியியல் கூறுகளையும், மிகப்பெரிய தரவினை ஆய்வதன் வழி கண்டுபிடித்துச் சொல்லக்கூடிய கருவிகளை நாம் பெற்றுள்ளோம். தரவக மொழியியல் என்பது மொழி தரவுகளின் துணைக்கொண்டு மொழியினை ஆய்வதாகும்.

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

#### Keywords:

- A தரவகம்
- B தரவக மொழியியல்
- C கணினி மொழியியல்
- D நிகழ் நேரம்
- E பகுப்பாய்வு

#### Corresponding Author:

C.M.Elanttamil,  
Faculty of Languages and Linguistics,  
University Malaya  
Email: elanttamil@um.edu.my

#### 1. அறிமுகம்

மலேசியத் தமிழ்த் தரவகம்

மலேசியா போன்ற பன்மொழிச் சூழல் நிறைந்த நாட்டில் தாய்மொழியை முதன் மொழியாகவோ (L1) இரண்டாம் மொழியாகவோ (L2) கற்பிப்பதற்கும் கற்பதற்கும் தரவக ஆய்வுகள் பெரிதும் துணைபுரியும். மேலும் ஒரு மொழியின் சொற்களஞ்சியம், இலக்கணம் ஆகியவற்றின் இன்றைய கட்ட வளர்ச்சியைத் தெளிவாகத் தெரிந்து கொள்ளத் தரவக மொழியியல் துறை மிகவும் உதவும். இதற்கு முதலில் ஒரு முழுமையான தரவகம் தேவை. இக்கட்டுரை மலேசியத் தமிழுக்கு உருவாக்கப்பட்ட தரவக மேம்பாட்டினை விளக்குகின்றது.

ஒரு மொழியில் தரவகம் உருவாக்கப்படுவதற்குப் பல நிலையில் சிந்தித்துத் திட்டமிட வேண்டியுள்ளது. தரவகத்தைப் பேச்சுத் தரவகம், எழுத்துத் தரவகம் என இரு நிலையில் உருவாக்கலாம் அதுமட்டுமன்றி இடம் சார்ந்து உருவாக்கலாம்; காலம் சார்ந்தும் உருவாக்கலாம்; ஒப்பீட்டுத் தரவகம், சிறப்புத் தரவகம் என ஆய்வு விரிவடைகிறது. மலேசியாவில் உருவாக்கப்பட்ட தரவகம் தற்கால மலேசிய எழுத்துத் தமிழ் தரவகமாகும்.

தற்காலத் தமிழைப் பொறுத்தவரை, தமிழ்நாட்டில், தரவக மொழியியல் அடிப்படையில் சில தமிழ் ஆய்வுகள் உள்ளன. ஆனால் தரவக மொழியியல் கொள்கைகள், நெறிமுறைகளின் அடிப்படையில் மலேசியாவில் எந்த ஆய்வும் தரவக அடிப்படையில் இல்லை; தரவக அணுகுமுறையில் திட்டமிடப்பட்ட கற்றல் கற்பித்தல் ஏதும் இல்லை. எனவே தரவகத்தை உருவாக்குவது மிக முக்கியமாகும். ஆயினும் தரவகம் உருவாக்க சில கோட்பாடுகளும் தேவை.

இந்தக் கோட்பாடுகளை அடைவதற்கு இதற்கு முன்னர் தரவகங்களை உருவாக்கியவர்கள் அறிவியல் ஆய்வின் வழி சில அடிப்படைகளை உருவாக்கி சென்றுள்ளனர். அந்த அடிப்படைகளைப் பின்பற்றி மலேசியச் சூழலுக்கு ஏற்ப தரவகம் உருவாக்கப்பட்டது. அதற்கான பின்னணியும் கோட்பாடும் தொடர்ந்து விவரிக்கப்படுகின்றது.

## 2. பின்னணி

### தமிழ் - ஓர் இரட்டை வழக்கு மொழி

உள்ளார்ந்த நிலையில், தமிழ் ஓர் இரட்டை வழக்கு மொழி (Ferguson, 1971). பேச்சுத் தமிழுக்கும் எழுத்துத் தமிழுக்கும் அவற்றின் அமைப்பிலும், செயல்பாடுகளிலும், தமிழ்ச் சமுதாயத்தின் மனப்பான்மையிலும் வேறுபாடுகள் உள்ளன. மலேசியாவிலும் தமிழ்மொழி பேசும் மக்களிடையே இந்த இரட்டை வழக்குநிலை நிலவுகிறது. பேச்சுத் தமிழானது, அமைப்பு, செயல்பாடுகள், சமுதாய மனப்பான்மை என்று அனைத்து நிலைகளிலும் எழுத்துத் தமிழுடன் வேறுபாடுகளைக் கொண்டிருக்கின்றது. தமிழ்நாட்டில் மட்டுமே அங்கு நிலவும் இரட்டை வழக்கு நிலைக்கான சில முறையான மொழியியல் ஆய்வுகள் மேற்கொள்ளப்பட்டுள்ளன (Deivasundaram, 1981). மலேசியாவிலோ, சிங்கப்பூரிலோ, இலங்கையிலோ, அத்தகைய ஆய்வுகள் எதுவும் பெரிய அளவில் கொள்ளப்படவில்லை, எனினும், இங்கெல்லாம் நிலவும் இரட்டை வழக்கு நிலைபற்றிச் சில ஆய்வுக் கட்டுரைகள் குறிப்பிட்டுள்ளன (De Silva, 1976).

மேற்கூறிய நான்கு நாடுகளிலும் தமிழ்ப் பேச்சு வழக்குகளில், சில குறிப்பிடத்தக்க வேறுபாடுகள் உள்ளன; இருப்பினும், எழுத்து வழக்குகளில் வேறுபாடுகள் குறைவாக இருக்கின்றன. அவற்றிடையே சொற்களஞ்சிய வேறுபாடுகளே காணப்படுகின்றன. ஆனால், இந்தக் கருதுகோளை உறுதிப்படுத்தும் முறையான ஆய்வுகள் எதுவும் செய்யப்படவில்லை. மலேசிய தரவக உருவாக்கத்திற்கு ஆய்வாளர் எழுத்துத் தமிழினை மட்டுமே ஆய்விற்கு உட்படுத்தியுள்ளார்.

### தற்கால எழுத்துத் தமிழ்

மேற்கூறியவற்றின் அடிப்படையில், தற்கால எழுத்துத் தமிழின் வரலாறு இரண்டு நூற்றாண்டுகளைக் கொண்டது என்று கூறலாம். இருப்பினும், 1930களிலிருந்து மலேசியாவுக்குத் தனியே ஒரு தமிழ் இலக்கிய மரபு உள்ளது என்று ஆய்வாளர்கள் கூறுகின்றனர் (Rajanthiran et al., 2012). எனவே மலேசியத் தமிழின் தன்மைகளை மலேசியத் தமிழின் தரவகத்தைக் கொண்டு ஆய்வு செய்வது அவசியமாகின்றது. தொடர்ந்து மலேசியத் தமிழின் பயன்பாடு அதிகம் உள்ள தளங்களில் இருந்து பேச்சுத் தமிழின் தரவுகள் சேகரிக்கப்பட வேண்டும். மலேசியாவின் முக்கியத் தமிழ்த் தளங்களாகத் தமிழ்ப்பள்ளிகள், தமிழ் மின்னியல்/அச்சு ஊடகங்கள், இணைய ஊடகங்கள், இலக்கிய நிகழ்ச்சிகள் விளங்குகின்றன.

### அ. தமிழ்ப்பள்ளி/ தமிழ்க்கல்வி

1816-ஆம் ஆண்டு பினாங்கில் முதல் தமிழ் வகுப்புத் தொடங்கப்பட்டுக் கிட்டத்தட்ட இருநூறு ஆண்டுகள் ஆகின்றன (Shoniah & Ramasamy, 2015; Wong, 1982). இப்போது மலேசியாவில் 527 தமிழ்த் தொடக்கப்பள்ளிகள் உள்ளன. இந்தப் பள்ளிகளில் எழுத்துத் தமிழ் கற்பிக்கப்படுகிறது. பரமசிவம் & பாரசையன் (2016) கருத்துப்படி, “மொழிக் கொள்கையில் ஏற்பட்ட மாற்றத்தால் (ரசாக் அறிக்கை 1956), தமிழ்த் தொடக்கப் பள்ளிகளில் தமிழ்மொழிப் பயன்பாட்டினைத் தக்க வைத்துக் கொள்ள முடிந்தது. ஆகவே இக்கொள்கையால் தமிழ்மொழி பாதுகாக்கப்பட்டது. எனவே, இன்றுவரை, மலேசியாவில் உள்ள அனைத்து 527 தமிழ்த் தொடக்கப் பள்ளிகளும் தேசிய மொழியான மலாயையும், இரண்டாம் மொழியான ஆங்கிலத்தையும் தவிர அனைத்துப் பாடங்களையும் தமிழ்மொழியில் கற்பிக்கின்றன.

### ஆ. தமிழ் தகவல் ஊடகங்கள்

தமிழ்மொழி பரவலாகப் பயன்படுத்தப்படும் மற்றொரு களம் ஊடகம். பரமசிவம் & பாரசையன் (2016) பின்வருமாறு கூறுகிறார்கள்: “மலேசியாவில் தொலைக்காட்சி, வானொலி, பத்திரிகை போன்ற பல்வேறு ஊடகங்கள் தமிழின் வளர்ச்சிக்குக் குறிப்பிடத்தக்க பங்களிப்பைச் செய்துள்ளன” (பரமசிவம், 2010). மலேசிய வானொலி நிகழ்ச்சிகளில் அதன் இந்திய மொழி ஒலிப்பரப்பில் 24 மணி நேரமும் தமிழ் ஒலிப்பரப்பு இருந்தது; தமிழ் 95% இடம்பிடித்தது. பாடல்களோடும் இசையோடும் கூட, உரையாடல் நிகழ்ச்சிகள், கவிதை ஒப்பித்தல், நாடகங்கள் ஆகியவற்றிற்கும் கணிசமான அளவு நேரம் ஒதுக்கப்படுகிறது. மலேசியாவில் தற்பொழுது அரசாங்க ஊடகமான மின்னல் பண்பலையும் மலேசியத் தொலைக்காட்சி தமிழ்ப்பிரிவும் செயல்பட்டு வருகின்றன. தனியார் தொலைக்காட்சி நிறுவனமான ஆஸ்ட்ரோவின் தமிழ் ஒளிப்பரப்பும் வானொலி பிரிவான தி. எச் ராகாவின் ஒலிப்பரப்பும் செயல்பட்டு வருகின்றன.

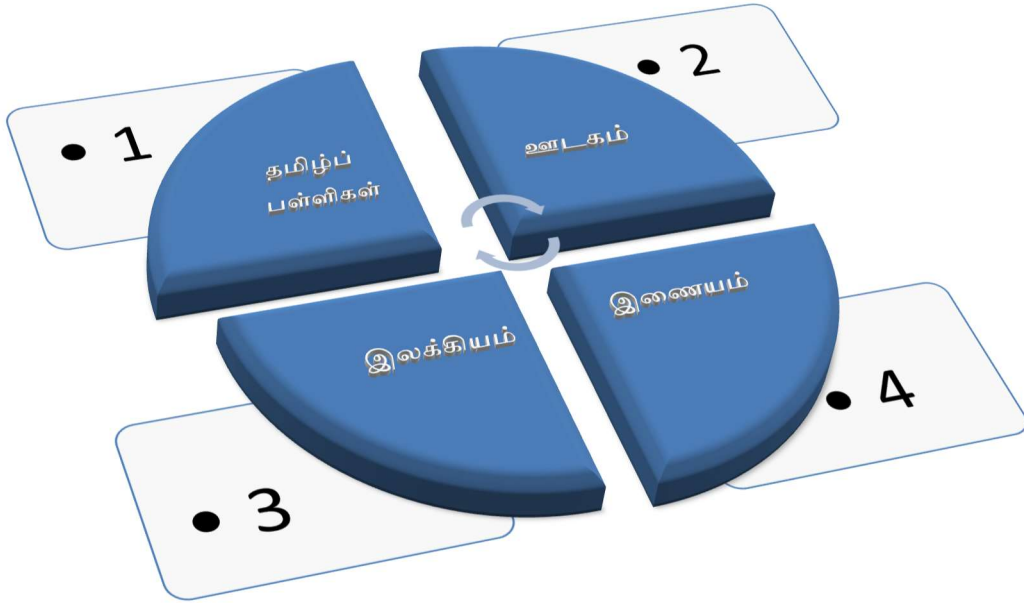
### இ. தமிழ் அச்ச ஊடகங்கள்

மேலும், உள்நாட்டில் அச்சிடப்பட்ட இதழ்களும், நாளிதழ்களும், தமிழ்நாட்டு (இந்தியா) இதழ்களும் மலேசியாவில் கிடைக்கின்றன. இது மலேசியத் தமிழர்களுக்குத் தமிழ்மொழித் திறன்களையும் சொற்களஞ்சியத்தையும் படிக்கவும் வளர்த்துக் கொள்ளவும் தாராளமான வாய்ப்புகளை உருவாக்குகிறது. குறிப்பாக நாளிதழ்களான மலேசிய நண்பன், மக்கள் ஓசை, தமிழ் மலர் தொடர்ந்து வெளிவருகின்றன. மின்னும் நாளிதழான தமிழ் நேசன் நிறுத்தப்பட்டது கவலைக்குரிய செய்தியாகும்.

### ஈ. இணைய ஊடகம்

தமிழின் வளர்ச்சிக்குப் பெரிதும் உதவும் மற்றொரு பயனுள்ள ஊடகம் இணையம். அனைத்து முக்கியத் தமிழ் நாளிதழ்களும் இணையத்தளங்களில் காணப்படுகின்றன. இவை பரந்துபட்ட தமிழ்ச் செய்திகளைக் கொண்டிருக்கின்றன. இது தமிழ்மொழியை இளைய தலைமுறையினர் தக்கவைப்பதற்கும், பயன்படுத்துவதற்குமான மற்றொரு கட்டத்தை உருவாக்குகிறது. மலேசியத் தமிழ் இணையத்தளங்களில் மலேசியா இன்று, வணக்கம் மலேசியா போன்ற தளங்கள் உடனுக்குடன் செய்திகளை இருகின்றன. இது தவிர பொது இயக்கங்கள், தனி நபர்களின் இணையத்தளங்களும், வலைப்பதிவுகளும் தமிழ்மொழியைத் தொடர்ந்து பயன்படுத்தி வருகின்றன. எனவே, பல்வேறு களங்களில் தமிழ் பயன்படுத்தப்பட்டாலும் கூட மலேசியாவில் அதிகமாகத் தமிழ் பயன்படுத்தப்படும் முக்கியக் களங்களாக, கீழ்க்கண்ட நான்கு களங்கள் முதன்மையாகத் இருக்கின்றன.

படம் 1: மலேசியத் தமிழ்த் தளங்கள்.



இத்தளங்களில் எழுத்துத் தமிழ் பயன்படுத்தப்படுகிறது என்பதை மேற்கண்ட தகவல்கள் தெளிவாக நிறுவுகின்றன. மேற்கண்ட களங்களில் பயன்படுத்தப்படும் எழுத்துத் தமிழை மலேசியாவின் தற்கால எழுத்துத் தமிழ் என்று அழைக்கலாம். தற்கால எழுத்துத் தமிழினை முழுமையாக ஆய்வு செய்ய இத்தளங்களில் பயன்படுத்தப்படும் எழுத்துத் தமிழினைத் தொகுப்பது அவசியமாகின்றது. அதன் பின்னர் அதனை மொழியியல் பகுப்பாய்வு செய்வதன் வழி நவீன தொழில்நுட்பக் கருவிகளை உருவாக்க அடிப்படை வேலைகளைச் செய்ய முடியும். இந்த தளங்களில் இருந்து 1980-2020 வரை வெளியிடப்பட்ட தரவுகளை ஆய்வாளர் தரவகம் உருவாக்க தொகுத்துள்ளார்.

தற்கால எழுத்துத் தமிழை விவரிக்க, கடந்த மூன்று நூற்றாண்டுகளாக எழுதப்பட்ட பல கட்டுரைகளும் புத்தகங்களும் உள்ளன, கால்டுவெல், ஜி.யு. போப், வீரமாமுனிவர், ஆர்டன் போன்ற ஐரோப்பிய அறிஞர்கள், 19, 20-ஆம் நூற்றாண்டுகளில் தமிழ்நாட்டில் தங்கியிருந்த காலத்தில் தமிழுக்குச் சில இலக்கணங்களை எழுதினர். 20-ஆம் நூற்றாண்டின் பிற்பகுதியில், தமிழ்நாட்டில் மொழியியல் அறிவியலின் அறிமுகத்தால், அகத்தியலிங்கம், கோதண்டராமன் (Kothandaraman, 1997) (இந்தியா), நுஹ்மான் (Nuhman, 1999) (இலங்கை), சீனி நைனா முகம்மது (மலேசியா), சித்தார்த்தன் (சிங்கப்பூர்) ஆகியோரால் தற்காலத் தமிழுக்குச் சில புதிய இலக்கணங்கள் எழுதப்பட்டன.

இருப்பினும், மேலே குறிப்பிட்டுள்ள பெரும்பாலான இலக்கணங்களும் விவரண மொழியியல் (*descriptive linguistic*) மரபிலிருந்து எழுதப்பட்டுள்ளன. அகத்தியலிங்கமும் கோதண்டராமனும் தங்களது விவரண ஆய்வுகளில் சாம்ஸ்கியின் மாற்றிலக்கண அணுகுமுறையை (*Generative Grammar approach*) அங்கொன்றும் இங்கொன்றுமாக ஏற்றுக்கொண்டனர். ஆனால் இந்த இலக்கணங்கள் தற்காலத் தமிழின் உண்மையான தரவக ஆய்வின் அடிப்படையிலான இலக்கணங்கள் என்று கூற முடியாது. எனவே தரவக மொழியியல் அடிப்படையிலான தற்காலத் தமிழை முழுமையாக ஆய்வு செய்ய வேண்டிய தேவை உள்ளது. இதன் வழி தமிழ்மொழியில் உள்ள நுணுக்கமான, துல்லியமான செய்திகள் வெளிவர அதிகம் வாய்ப்பிருக்கின்றது. மொழிக்கு இலக்கணம் மிக அவசியமான ஒன்று. இவ்விலக்கணத்தை அடிப்படையாக கொண்டே அம்மொழி இயங்குகின்றது. இலக்கணத்தை தரவக அடிப்படையில் ஆய்வு செய்வது இப்பொழுது சாத்தியமாகியுள்ளது.

### 3. மலேசியத் தமிழ்த் தரவக மேம்பாடு

அட்டவணை 1: மலேசியத் தமிழ்க் களங்கள்.

Domain	Medium
Text Book	Primary 1,2,3,4,5,6, Secondary 1,2,3,4,5, Student Writing and Others
Periodical	Tamil News Papers
Popular Magazine	Tamil Magazines
Fiction	Short Story, Novel, Essays, and Others
Internet	Web Portal articles
Academy Journal	Journals
To-Be Spoken	TV/Radio news, and Scripts
Unclassified	Government Notification, NGO Materials, Special edition, and Others

தமிழ்நாட்டில் பயன்படுத்தப்படும் தமிழுக்காக, தமிழ் மெய்நிகர் அகாடமி (சென்னை), அண்ணா பல்கலைக்கழகத்தின் AU-KBC மையம் (சென்னை), மற்றும் இந்திய மொழிகளின் மத்திய நிறுவனம் (மைசூர்) போன்ற சில பல்கலைக்கழகங்கள் மற்றும் ஆராய்ச்சி நிறுவனங்களால் சில தரவக திட்டங்கள் மேற்கொள்ளப்பட்டன. சிங்கப்பூர் எழுத்துத் தமிழுக்கு, சிங்கப்பூர் கல்வி அமைச்சகம் ஒரு தரவகத்தை உருவாக்கியுள்ளது. இலங்கையில், மொரட்டுவா பல்கலைக்கழகம் (கொழும்பு) எழுத்துத் தமிழுக்கான தரவக மேம்பாட்டு பணியில் ஈடுபட்டுள்ளது.

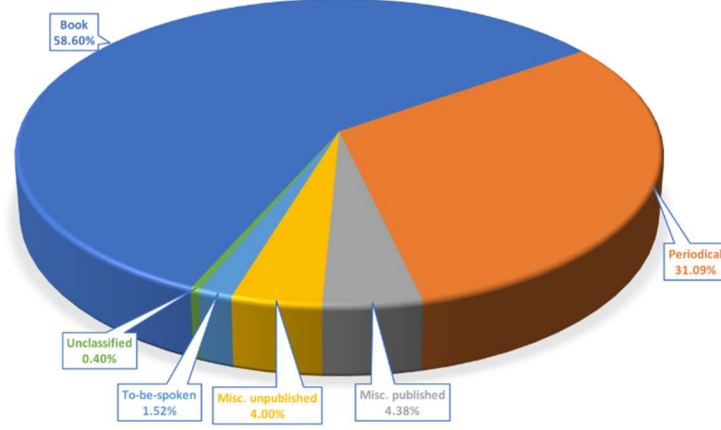
மலேசியத் தமிழுக்கான தரவகம் இன்னும் உருவாக்கப்படவில்லை. எனவே, மலேசியாவின் தற்கால எழுத்துத் தமிழுக்கு, இயற்கையான மொழிப் பயன்பாடு அல்லது மலேசியாவில் உள்ள தமிழ் மொழிப் பயனர்கள் தங்கள் அன்றாட வாழ்க்கைச் சூழல்களில் டௌருவாக்கிய நூல்களை உள்ளடக்கி ஒரு தரவகம் உருவாக்க வேண்டிய அவசியத்தை கருத்தில் கொண்டு, தரவகப் பணியை நிறைவேற்ற இந்த ஆய்வுத் திட்டம் மேற்கொள்ளப்பட்டது.

தரவகம் மேம்பாட்டிற்கு முன் அல்லது தொகுக்கப்படுவதற்கு முன் சில அடிப்படை செய்திகளை உறுதி செய்வது அவசியம் எனப் பார்த்தோம். தரவகத்தின் உட்பிரிவுகள் தேர்வுக்கு மலேசியத் தமிழ்மொழி பயன்பாட்டினை உள்ளடக்கி, BNC COCA எனப்படும் பிரிட்டன் / அமெரிக்கத் தரவகங்கள் வழிகாட்டியாகக் கொள்ளப்பட்டன. தரவகத்தின் எண்ணிக்கைக்கு ப்ராவ்ன் தரவகம் வழிகாட்டியாகக் கொள்ளப்பட்டது.

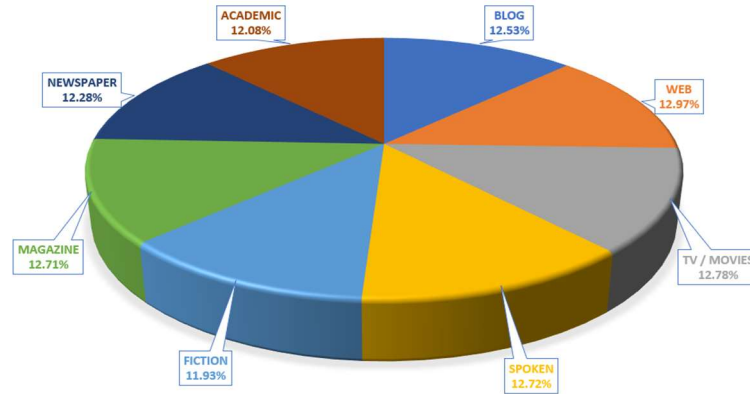
தரவக மொழியியல் துறையில் BNC மற்றும் COCA இரண்டு முக்கியமான தரவகங்களாகும். 20 ஆம் நூற்றாண்டின் (பிற்பகுதி) பிரிட்டிஷ் ஆங்கிலத் தரவகமான BNC, 1991-1995 வரையிலான 100 மில்லியன் பிரிட்டிஷ் ஆங்கில சொற்களை உள்ளடக்கியது. COCA உலகில் பரவலாகப் பயன்படுத்தப்படும் தரவகங்களில் ஒன்றாகும். COCA, 1990-2019 வரை ஒவ்வொரு வருடமும் 20 மில்லியன் சொற்களை உள்ளடக்கி ஒரு பில்லியனுக்கும் அதிகமான சொல் தரவுகளைக் கொண்டுள்ளது.

500 மாதிரிகள் ஒவ்வொன்றும் 2000 சொற்கள் என ஒரு மில்லியன் (10 இலட்சம்) சொற்கள் சேகரிக்கப்பட்டன. இச்சொற்களை எங்கிருந்து எடுக்கப்பட வேண்டும் என்பதில் ஒரு வரையறை உள்ளது. அந்த வரையறையின் அடிப்படையில் மலேசியாவில் தமிழ் அதிகம் புழங்கும் களமாகத் தமிழ்ப்பள்ளிகள், நாளிதழ்கள், சஞ்சிகைகள், இலக்கியப் படைப்புகள், இணையம், ஊடகங்கள் எனப் பட்டியலிடப்பட்டன. தளங்களும் களங்களும் கண்டறியப்பட்டபின், ஒவ்வொரு பிரிவிலும் இருந்து எத்தனை விழுக்காட்டுச் சொற்கள் எடுக்கப்படவேண்டும் என்பதனை பி.என்.சி தரவகத்தை வைத்து வரையறை செய்யப்பட்டது.

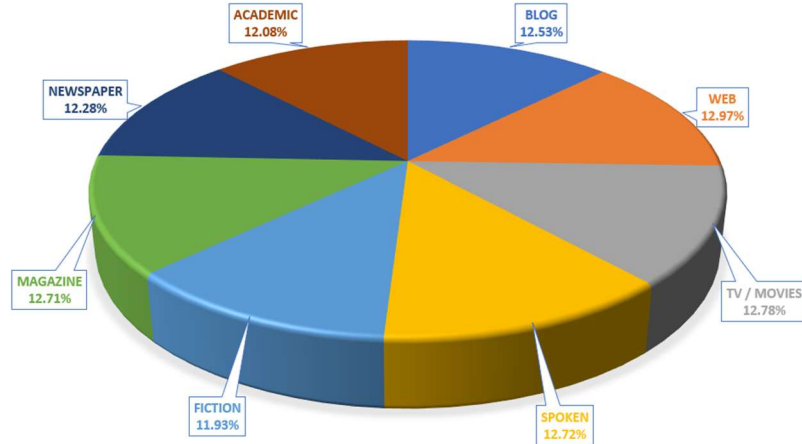
வரைபடம் 2: BNC தரவக உட்பிரிவுகள்



வரைபடம் 3: COCA தரவக உட்பிரிவுகள்



வரைபடம் 4: மலேசியத் தமிழ்த் தரவக உட்பிரிவுகள்.

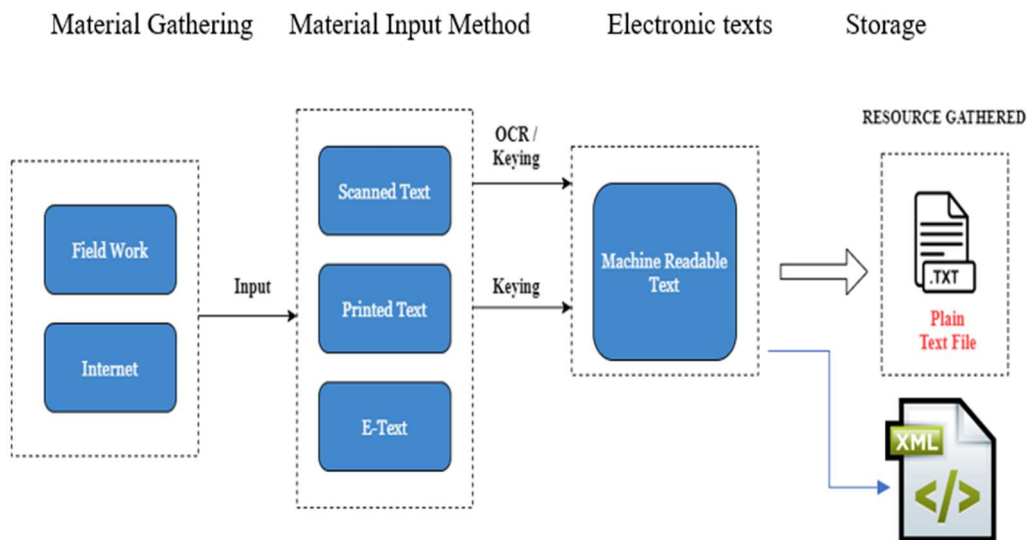


அட்டவணை 2: மலேசியத் தமிழ்த் தரவகமும் பி.என்.சி தரவகமும்.

	British National Corpus	Written Malaysian Tamil Corpus
Domain	Imaginative (19%) Arts (7%) Belief and thought (3%) Commerce/Finance (8%) Leisure (14%) Natural/pure science (4%) Applied Science (8%) Social science (16%) World affairs (20%) Unclassified (1%)	Imaginative (20%) Arts (15%) Belief and Thought (5%) Commerce and Finance (10%) Leisure (10%) Science and Technology (10%) World affairs (10%) Family and Society (20%)
Date	1960-1974 (2.26%) 1975-1993 (89.23%) Unclassified (8.49%)	1980 - 2000 (20%) 2001 - 2010 (30%) 2011 - 2020 (50%)
Medium	Book (58.58%) Periodical (31.08%) Misc. published (4.38%) Misc. unpublished (4.00%) To-be-spoken (1.52%) Unclassified (0.40%)	Textbook (10%) Periodical (20%) Popular magazine (20%) Fiction (10%) Internet (10%) Academy journal (10%) To-be-spoken (10%) Unclassified (10%)

விழுக்காடும் பிரிவுகளும் உறுதி செய்யப்பட்டபின் தரவுகளை உள்ளீடு செய்வதற்கு உள்ளீட்டு படிவ மென்பொருள் உருவாக்கப்பட்டது. இதன் வெளியீடு எக்ஸ்.எம்.எல் கோப்பில் (XML File format) அமையுமாறு வடிவமைக்கப்பட்டது.

வரைபடம் 5: மலேசியத் தமிழ்த் தரவக மேம்பாட்டின் முழுமையான செயல்முறை.



இறுதியாகக் கோப்புகள் எக்ஸ் எம் எல் வடிவத்திலும் உரை கோப்பாகவும் சேமிக்கப்பட்டது.

C.M. Elantamil, University Malaya (UM), Malaysia

வரைபடம் 6: XML கோப்பு வடிவம்.

```
<FileData>
  <FileName>தமிழ்ப்பள்ளிகளை மாற்றான் தாய்ப்பிள்ளையாக
  நடத்துவதை நிறுத்துவீர்.txt</FileName>
  <Source>Internet</Source>
  <SubSource>Malaysia Indru</SubSource>
  <Publisher>Malaysiakini</Publisher>
  <Writer>K.Arumugam</Writer>
  <Year>2012</Year>
  <Date>22.01.2012</Date>
</FileData>
```

#### 4. தரவக அடிப்படையிலான தற்காலத் தமிழ் ஆய்வு

மொழி கற்றல் கற்பித்தலுக்கும், மொழியியல் கருவிகள் மென்பொருள் உருவாக்குவதற்கும் முதலில், நல்லதொரு தரவகம் தேவை என்பதை வலியுறுத்தவே இக்கட்டுரை எழுதப்பட்டுள்ளது. இப்பொழுது உருவாக்கப்பட்டுள்ள தரவகம் 1 மில்லியன் சொற்களைக் கொண்டுள்ளது. தொடர்ந்து தமிழ் பயன்படுத்தும் தளங்களில் இருந்து பெருந்தரவினைப் பெற்று மலேசியத் தமிழுக்கு ஒரு பெருந்தரவகம் உருவாக்க வேண்டிய பணி உள்ளது. அதைத் தொடர்ந்து கணினி நிரல்களை எழுதி அதற்கேற்றவாறு மொழியைப் பயன்படுத்துவது அடுத்தகட்ட மொழி வளர்ச்சிக்கு வித்திடும். தரவக மொழியியல் ஆய்வாளர் மொழியியல் ஆய்வுக்கு Corpus based, corpus driven என்ற இருமுறைகளை முன் வைக்கின்றார். இச்செல்நெறியில் மலேசியத் தமிழுக்கான ஆய்வினை இட்டுச்செல்லாம்.

#### 5. எதிர்பார்ப்பு

அனைத்துக் களங்களிலும் தமிழ்மொழிப் பயன்பாட்டைப் புரிந்துகொள்வதற்கும், அதன் வளர்ச்சிக்கு முறையான மொழித் திட்டமிடலை மேற்கொள்வதற்கும், மொழி கற்பித்தல், அகராதியியல் போன்ற கற்பித்தல் நோக்கங்களுக்காகத் தேவையான மொழியியல் கருவிகளை மேம்படுத்துவதற்கும், மலேசியாவின் தற்கால எழுத்துத் தமிழின் தற்போதைய நிலையையும் வளர்ச்சியையும் தரவகத்தின் துணைகொண்டு ஆய்வு செய்வது அவசியம். இப்பொழுது உருவாக்கப்பட்டுள்ள இத்தரவகம் இம்முயற்சிகளுக்கு வித்திடும் என்பது தெளிவு.

#### REFERENCES

- Aston, G. (1998). The bnc handbook exploring the british national corpus with sara guy aston and lou burnard.
- Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7(1), 1-16.
- Bowker, L., & Pearson, J. (2002). *Working with specialized language: a practical guide to using corpora*. Routledge.
- Cheng, W. (2010). What can a corpus tell us about language teaching. *The Routledge handbook of corpus linguistics*, 319-332.
- Cheng, W. (2011). *Exploring corpus linguistics: Language in action*. Routledge.



- De Silva, M. S. (1976). *Diglossia and literacy* (Vol. 2). Central Institute of Indian Languages.
- Deivasundaram, N. (1981). Tamil diglossia. *Tirunelveli, Nainar Pathippagam*.
- Department of Statistics, M. (2022, 13 May 2022). *Statistics*. Department of Statistics, Malaysia.
- Devi, S. L. (2011). Text Extraction for an Agglutinative Language. *Language in India*, 11(5).
- Ferguson, C. A. (1971). *Language structure and language use: Essays* (Vol. 1). Stanford University Press.
- Francis, M. (2015). A comprehensive survey on parts of speech tagging approaches in dravidian languages. Proceedings of 30th The IIER International Conference, Beijing, China, 26th July,
- Kamil, Z. (1990). *Dravidian linguistics: an introduction*. Pondicherry Institute of Linguistics and Culture.
- Kothandaraman, P. (1997). *A grammar of contemporary literary Tamil*. Int. Inst. of Tamil Studies.
- Krishnamurti, B. (2003). *The dravidian languages*. Cambridge University Press.
- Kulkarni, S., & Sagar, B. (2014). A survey on Named Entity Recognition for South Indian Languages. National Conference on Indian Language Computing,
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- McEnery, T., & Hardie, A. (2012a). *Corpus linguistic: Method, theory and practice*. Cambridge University Press.
- McEnery, T., & Hardie, A. (2012b). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- McEnery, T., & Hardie, A. (2013). The history of corpus linguistics. *The Oxford handbook of the history of linguistics*, 727-745.
- McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh University Press.
- Nuhman, M. (1999). Adippadai Tamil Ilakkanam (Basic Tamil Grammar). *Readers Circle: Kalmunai*.
- Rajantharan, M., Muniapan, B., & Govindaraju, G. M. (2012). Identity and language of Tamil community in Malaysia: Issues and challenges. *International Proceedings of Economics Development and Research at Dubai, UAE*. doi, 10.
- Sarveswaran, K., Dias, G., & Butt, M. (2021). Thamizhimorph: A morphological parser for the Tamil language. *Machine Translation*, 35(1), 37-70.
- Sheshasaayee, A., & Deepa, V. A. (2016). A Conceptual Model for Acquisition of Morphological Features of Highly Agglutinative Tamil Language Using Unsupervised Approach. In *Information Systems Design and Intelligent Applications* (pp. 499-507). Springer.
- Sheshasaayee, A., & VR, A. D. (2015). Morpheme Segmentation for Highly Agglutinative Tamil Language by Means of Unsupervised Learning. *International Journal of Computer Applications*, 975, 8887.
- Shoniah, S., & Ramasamy, K. (2015). 6 Tamil education in Malaysia. *Languages in the Malaysian Education System: Monolingual strands in multilingual settings*, 92.
- Steever, S. B. (2019). *The Dravidian Languages*. Routledge.
- Wong, F. (1982). Education, political development and social equity in Malaysia. *International Journal of Educational Development*, 2(3), 235-248.

## Emotionality in Suicide Notes

N. Nirmeen<sup>1</sup>, N. Vijayan<sup>2</sup>

<sup>1,2</sup>Department of Linguistics, Bharathiar University (BU), India

---

### ABSTRACT

---

#### Keywords:

Forensic Linguistics  
Emotionality  
Genuine Suicide notes  
LIWC  
Automatic Text Analysis

Suicide note is an important forensic text, particularly in the case of equivocal deaths. The major problem for the investigators falls on determining the authenticity of the suicide note discovered. In order to unravel the authenticity of the suicide note, it is essential to understand the characteristics of the genuine suicide notes. Suicidal individuals are more likely to experience depression, stress and anxiety. They are often suffocated by different emotions. Despite the fact that they do not express their emotions, their choice of words, particularly emotive words will reflect the emotions they are going through. This study focuses on analysing the emotionality in a genuine suicide note to detect the emotions expressed by the writer before terminating her life. In recent times, computerised methods are being adopted to analyse and characterise the quantity of linguistic features present in the suicide notes and other types of text. Though, such software are applied to analyse the text, the results are not always exemplary. This study aims to analyse the emotionality in a suicide note using a computerized method of text analysis software; Linguistic Inquiry Word Count and manual analysis. The findings of the study will provide insight into how emotions are conveyed in the suicide note and the issues faced in categorising the emotionality in suicide note using the software.

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

---

#### Corresponding Author:

N. Vijayan,  
Department of Linguistics  
Bharathiar University, India  
Email: [vijayan@buc.edu.in](mailto:vijayan@buc.edu.in)

---

### 1. INTRODUCTION

A written note left by a suicide victim before committing suicide is one of the typical linguistic evidences usually examined by forensic linguists. Suicide note is an important forensic text in equivocal death for the reason that it unravels the truth behind the act of suicide as it contains information about the author's emotional state, messages, last will and motivation. It is vitally pertinent as the beginning of the investigation to expose its content and motive (Richardson & Breyfogle, 1947), to identify the psychological state of the suicide victim (O' Donnel et al, 1993), and to prove the authenticity of the writing (Basim, 2012). Although the language, style, content, and intentions of each suicide note differ, their general form and content are frequently identical (Basim, 2012).

Since 1957, linguistic analysis of suicidal text has been a growing field. Researches like; Clues to Suicide, 1957; Genuine versus simulated suicide notes: An issue revisited through discourse analysis, 1982; An anxiety scale applicable to verbal samples, 1961; Motivation and language behaviour: A content analysis of suicide notes, 1959) etc., pioneered this concept. The goal of these studies were to find characteristics that may be used to distinguish between genuine and fabricated suicide notes from a corpus of 66 suicide notes that Shneidman collected, half of which were genuine and half of which were fabricated. The significant proportion of the early work concentrated on manual analysis and detection of these features, or by concentrating on shallow text characteristics etc. In recent times,

to identify emotions indicated with suicide behaviour there is considerable interest in applying natural language processing (NLP) and machine learning approaches to find suicidal communications.

Analysis of emotions in suicide note is important to understand the psychological behaviour of the writer. Emotion analysis is the technique of detecting and analysing the underlying emotions conveyed in textual data. The text will be evaluated to identify the emotions behind the subjective data. It can be quite hard to ascertain the emotions that are being represented in writing because certain events or situations can elicit underlying feelings. Emotions are not always expressed through an emotive word but an emotion can also be expressed without using specific words that express the emotion. Also, it can be very difficult to distinguish between different emotions solely based on keywords.

Emotions in a suicide note are analysed either manually or by automatic text analyzing software. The most popular software for analysing suicide notes is LIWC software.

LIWC (Linguistic Inquiry and Word Count) is a lexical resource established by a social psychologist James Pennebaker and his team at the University of Texas (Pennebaker et al, 2001). Its lexical data is kept in a dictionary consisting around ninety word categories which classifies English words spanning from linguistic elements like function words to psychologically significant categories, including emotions, cognitive processes, biological processes, and as well as informal language markers. This dictionary can be integrated into a programme that analyses a group of texts and reports the relative frequency of words from each category inside each text. The distribution of these categories within the text can reveal information about the author's psychological state or reflect the author's own circumstances. Each word in a statement is categorised to the appropriate category using a dictionary technique. It uses the terms in its dictionary to count the number of times each word appears in texts. Words are often only listed in the dictionary under the category that corresponds to their most common word sense. It assigns a score for each statement based on the percentage of terms in each category (thereby by default correcting for statement length). The software is extremely basic by computational linguistics standards. However, it is an often utilised research instrument. This study attempts to analyse the emotions in a suicide note by both manual method and LIWC software.

The LIWC dictionary encompasses around 6,500 words. The categorisation of words in are as follows;

Category	Abbrev	Examples	No. of Words in Category	Category	Abbrev	Examples	No. of Words in Category
Word count	WC	-	-	Cognitive processes	Cogproc	Cause, know	797
<b>Summary Language Variables</b>				Insight	Insight	Think, know	259
Analytical thinking	Analytic	-	-	Causation	Cause	Because, so	135
Clout	Clout	-	-	Discrepancy	Discrep	Should, would	83
Authentic	Authentic	-	-	Tentative	tentat	Maybe, if	178
Emotional Tone	Tone	-	-	Certainty	Certain	Always, never	113
Words/sentence	WPS	-	-	Differentiation	Differ	Hasn't, but	81
Words > 6 letters	Sixltr	-	-	Perceptual process	Percept	Look, heard,	436
Dictionary words	Dic	-	-	See	See	View, saw	126
<b>Linguistic Dimensions</b>				Hear	hear	Listen, hearing	93
Total function words	funct	it, to, no, very	491	Feel	Feel	Feels, touch	128
Total pronouns	pronoun	I, them, itself	153	Biological process	Bio	Eat, blood	748
Personal pronouns	ppron	I, them, itself	93	Body	Body	Cheek, hands	215
1 <sup>st</sup> pers singular	I	I, me, mine	24	Health	health	Clinic, flu, pill	294
1 <sup>st</sup> pers plural	we	we, us, our	12	Sexual	Sexual	Horny, love	131
2 <sup>nd</sup> person	You	You, your	30	Ingestion	Ingest	Dish, eat	184
3 <sup>rd</sup> pers singular	She/he	She, her, him	17	Drives	Drives		1103
3 <sup>rd</sup> pers plural	They	They, their	11	Affiliation	Affiliation	Ally, friend	248
Impersonal pronouns	Ipron	It, it's those	59	Achievement	Achieve	Win, success	213
Articles	Article	A, an, the	3	Power	Power	Superior, bully	518
Prepositions	Prep	To, with	74	Reward	Reward	Take, prize	120
Auxiliary verbs	Auxverb	Am, will,	141	Risk	Risk	Danger, doubt	103
Common adverbs	Adverb	Very, really	140	<b>Time orientations</b>			
Conjunctions	Conj	And, but,	43	Past focus	Focuspast	Ago, did	341
Negations	Negate	No, not, never	62	Present focus	Focuspresent	Today, is, now	424
<b>Other Grammar</b>				Future focus	Focusfuture	May, will	97
Common verbs	Verb	Eat, come	1000	Relativity	Relative	Area, bend	974
Common adjectives	Adj	Free, happy	764	Motion	Motion	Arrive, car, go	325
Comparisons	Compare	Greater, best	317	Space	Space	Down, in, thin	360
Interrogatives	Interrog	How, when	48	Time	Time	End, until	310
Numbers	Number	Second, one	36				
Quantifiers	Quant	Few, many	77				

<b>Psychological Processes</b>				<b>Personal Concerns</b>			
Affective processes	Affect	Happy, sad	1393	Work	Work	Job, majors, xerox	444
Positive emotion	Posemo	Love, nice	620	Leisure	Leisure	Cook, chat, movie	296
Negative emotion	Negemo	Hurt, ugly	744	Home	Home	Kitchen, landlord	100
Anxiety	Anx	Worried	116	Money	Money	Audit, cash, owe	226

Anger	Anger	Hate, kill	230	Religion	Relig	Altar, church	174
Sadness	Sad	Crying, grief	136	Death	Death	Bury, coffin, kill	74
Social processes	Social	Mate, talk	756	<b>Informal language</b>	Informal		380
Family	Family	Dad, mom	118	Swear words	Swear	Damn, shit	131
Friends	Friend	Buddy	95	Assent	Assent	Agree, OK, yea	36
Female references	Female	Girl, her	124	Fillers	Filler	I mean, you know	14
Male references	Male	Boy, his, dad	116	Netspeak	Netspeak	Btw, lol, thx	209
				Nonfluencies	Nonflu	Er, hmm, umm	19

Figure 1. LIWC Variable Information

## 2. RESEARCH METHOD

For the study, a handwritten, six-page suicide note has been used. Automatic text analysis using LIWC software and manual method were applied to analyse the emotions. The suicide note was uploaded into the LIWC text analysis software for automatic text analysis, and the findings of words indicating emotions were extracted. The LIWC codes the emotional words into the category of Psychological processes which is further divided as affective processes (affect), positive emotion (posemo), negative emotion (negemo), anxiety (anx), anger (anger) and sadness (sad) as displayed in the figure 2. The suicide note's emotions were painstakingly extracted from the same text manually and categorized the emotions. Then after, the two results were compared, and the challenges in emotion detection in both the methods were explored.

Category	Abbrev	Examples
<b>Psychological Processes</b>		
Affective processes	affect	happy, cried
Positive emotion	posemo	love, nice, sweet
Negative emotion	negemo	hurt, ugly, nasty
Anxiety	anx	worried, fearful
Anger	anger	hate, kill, annoyed
Sadness	sad	crying, grief, sad

Figure 2. Categorization of emotions in LIWC

## 3. RESULTS AND ANALYSIS

The suicide note employed for the study is of 6 pages containing 801 words. The LIWC software extracted the words indicating emotions. Out of 801 words, 46 words indicating moods and emotion were extracted. The 46 words were further categorized into 21 positive emotions, 25 negative emotions, 4 anxiety, 11 anger and 8 sadness as displayed in the following table 1.

Table 1 – LIWC extraction of emotions

Word	Affect	Posemo	Negemo	Anx	Anger	Sad
well	X	X				
lose	X		X			X
lost	X		X			X
loving	X	X				
tortured	X		X		X	
cared	X	X				
love	X	X				
cheating	X		X		X	
lies	X		X		X	
beautiful	X	X				
scared	X		X	X		
pain	X		X			
destroyed	X		X		X	
best	X	X				
rape	X		X		X	
abuse	X		X		X	
torture	X		X		X	
hurt	X		X			X
partying	X	X				
miss	X		X			X
kissing	X	X				
ignore	X		X			
trust	X	X				
embarrassed	X		X	X		
loyal	X	X				
heartbreak	X		X			X
threatens	X		X	X	X	
cheats	X		X		X	
happiness	X	X				
ready	X	X				
appreciated	X	X				
confidence	X	X				
talent	X	X				
bother	X		X		X	
cheated	X		X		X	
special	X	X				
promised	X	X				
engaged	X	X				
selfish	X		X			
cry	X		X			X
loved	X	X				
success	X	X				
empty	X		X			X
promises	X	X				
alone	X		X			X
vulnerable	X		X	X		

The LIWC results were compared to the manual approach, the following issues in emotion detection by LIWC approach were discovered and discussed below.

### 3.1. Homonyms

The word ‘well’ is coded as a positive emotional word. Looking into the context, the phrase is ‘*But I might as well now as I have nothing to lose.*’ The word ‘well’ is not a single word here but an idiom ‘as well’. The LIWC has miscoded the word to the word ‘well’ which means in a good health / state.

‘Kissing’ has been coded as positive emotion. But, in the context, the phrase is ‘*I am kissing my 10 years career and dreams goodbye.*’ Here, ‘kissing’ connotes farewell which is a negative emotion but LIWC has miscoded the word to the word ‘kiss’ which is a form of showing ones love.

### 3.2 Word Sense in Context

The word ‘loving’ is coded as positive emotional word. The phrase in the text is ‘*...I lost myself in loving you.*’ Though, ‘loving’ conveys positive emotion, in the context it conveys a negative emotion. The word ‘loving’ has been combined with the negative word ‘lost’ reflecting a negative emotion.

The words ‘cared, love, beautiful, best, trust and happiness’ are coded as positive emotional words whereas these words in the text appears as; ‘*I have never given so much of myself to someone or cared so much.*’, ‘*You returned my love with cheating and lies*’, ‘*It didn’t matter how beautiful I looked for u.*’, ‘*I thought it would bring best in me*’, ‘*... decided to trust you*’, ‘*All I wanted was you and my happiness*’. These phrases reflect the writer’s expectations, disappointments and her regret for all the love and care she had poured to a non-deserving person.

The same way, words like ‘partying, appreciated, confidence, success etc., are coded to be positive emotional words but it reflects the negative tone of her expectations and disappointment.

### 3.3 Combination of Words

The words that are coded as positive emotional words are mostly preceded by a negative word which reflects a negative emotion viz the word ‘promises’ is coded as positive emotion but it is preceded by a negative word ‘empty’, the word happiness is succeeded by ‘snatched away’ indicating a negative tone.

### 3.4 Unnoticed Words

97.50% of words were present in the LIWC 2015 dictionary. Among 801 words, 20 words were absent in the dictionary as listed in the table 2. 15 of the 20 words belong to affective processes, including 3 positive emotional words and 12 negative emotional words.

Table 2 – Words absent in LIWC Dictionary

Words absent in LIWC2015 Dictionary		
Matter	Karthik	Aborted
Gifts	Lie	Snatched
Shattered	Chase	Esteem
Function	Disrespects	Selflessly
Deserve	Presents	Tone
Mentally	Chose	Goa

The words ‘matter, shattered, function, tore, aborted, snatched, etc.’ impart negative emotions which are not present in the LIWC dictionary.

The words ‘broken, affected, dead, crave, goodbye’ conveys deep negative emotions and the words ‘wish, dream’ conveys positive emotions. These words are present in the dictionary but still LIWC failed to code them as emotions.

### 3.5 Emotive phrases with Non-emotive words

Emotional words alone do not convey emotion; but, the combination of certain non-emotional phrases also does. For instance, the phrases; ‘*I am nothing*’, ‘*I see no light*’, ‘*I wake up not wanting to wake up*’, ‘*I have no reason to breathe anymore*’ etc., holds deep emotions without containing emotional words. These phrases as it contains no emotive words, the LIWC was unable to detect the emotion.

## 4. CONCLUSION

Every method of analysis has its own benefits and drawbacks. The results of LIWC showed that most of the words were accurately categorised whereas few words were miscoded and few were unnoticed which were identified in manual analysis. According to the statistical results of LIWC, 8.86% emotional words were identified; 4.12% positive emotions and 4.74% negative emotions. According to the results of manual analysis; 0.62% positive emotions and 5.61% negative emotions were detected. Though, most of the emotional words were accurately categorized by LIWC, the emotions conveyed were not. The majority of emotions were expressed using non-emotional words, which LIWC couldn't comprehend.

To conclude, LIWC is the most common software used for analysing suicide notes. A suicide note, unlike other texts, functions as a forensic text that has direct applications in forensic linguistics and suicide prevention, thus, it must be keenly analysed in depth. It is a general belief that suicide note is likely to contain words that are related to suicide. For instance, dictionaries of suicide-related keywords have been used successfully to find suicidal blogs (Huang, Goh, & Liew, 2007), despite the fact that there were a lot of false positives. Many researches have shown

that suicide note contains more positive emotions than negative emotions. But, the fact is that the suicide victims take up the decision of suicide due to depression and their chance of expressing negative emotion is higher. It is not always the emotional words that conveys emotion but also the combination of certain words. LIWC does not do word sense disambiguation or take into consideration the context of the words. So, just by extracting the emotional words will not deliver the actual emotions expressed by the writer. Even minor errors in analysis will result in misconception.

Thus, the study suggests that, such text analyzing software and other tools have to be worked on issues discussed in this study. The automatic text analysing tools are mostly for English language, the texts in other language especially for Indian languages remains untouched for such analysis where the only way of analysing relies on manual method. Thus, tools like LIWC has to be developed by the linguists for Indian languages too.

## REFERENCES

- [1] R. Alfiyan, "Meanings in a Suicide Note: An Analysis of Linguistics Pragmatics in Nusadi's Suicide Note." *Diakses dari*, 2018. [Online]. Available: <https://www.researchgate.net/publication/323150837>
- [2] Y. J. Basim, "Author Attribution in Suicide Notes: Evidence from Applied Linguistics," *Comparative Legilinguistics*, 2012. [Online]. Available: <https://www.researchgate.net/publication/323223150>
- [3] Y. Huang., P. Goh, T, and C. L. Liew., "Hunting suicide notes in web 2.0 – preliminary findings," in *IEEE International Symposium on Multimedia Workshops*, ISMW 2007, pp. 517–521, 2007.
- [4] I. O'Donnell, R. Farmer and J. Catalan, "Suicide Notes," *British Journal of Psychiatry*, vol. 1 (163), pp. 45-48, 1993.
- [5] J. W. Pennebaker, et al., "Linguistic Inquiry and Word Count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, pp. 71, 2001.
- [6] J. W. Pennebaker, et al., "The Development of Psychometric Properties of LIWC2015," *Austin,Tx: The University of Texas of Austin*, 2015.
- [7] O. Richardson and H. S. Breyfogle, "Problems of Proof in Distinguishing Suicide from Accident". *The Yale Law Journal*, vol. 56 (3), pp. 482-508, 1947.



## Opinion Mining On Tamil Movie Reviews Using BERT – A Study

Syam Mohan E<sup>1</sup>, R. Sunitha<sup>2</sup>, Amudha T. K<sup>3</sup>  
<sup>1,2,3</sup>Department of Computer Science, Pondicherry University,  
Puducherry, India

---

### ABSTRACT

---

#### Keywords:

Tamil  
Opinion mining  
BERT  
Machine learning  
Deep learning

Tamil is the oldest classical south Indian language that comes under the Dravidian family and is diglossic. The emergence of different social media platforms, increased accessibility to mobile devices, high-speed Internet and ability to express in one's native language have enabled people to express their opinions and emotions on different aspects and instances. Opinion mining helps to identify opinions from textual data, and it can be classified into positive, negative based on their sentiment. Very few works have been carried out in mining the opinions expressed in Tamil, significantly less compared to English. In this paper, the authors performed a study on Tamil opinion mining utilizing a deep learning-based model called BERT. Due to the lack of publicly available datasets in the Tamil language, the authors extracted 200 Tamil movie reviews from various social media platforms and manually labeled them according to their overall sentiment orientation. Furthermore, different machine learning and deep learning approaches are used to compare the results with that of BERT on the Tamil language. It is found that BERT outperformed all other machine learning and deep learning models on Tamil movie reviews. This study assures the suitability of BERT for handling computational problems involving Tamil language.

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

---

#### Corresponding Author:

Syam Mohan E,  
Department of Computer Science,  
Pondicherry University,  
Puducherry, India  
Email: syammohane@gmail.com

---

### 1. INTRODUCTION

Tamil is one of the oldest languages in India, which comes under the Dravidian languages. People started to express themselves on social media platforms in their own native languages like Tamil, since the increased availability of smartphones and high speed Internet. These massive amount of unstructured textual data uploaded to the Internet contain plenty of latent information which can be extracted by analyzing and computational processing for various applications. Opinion mining (OM) is a trending area of research that identifies opinions from textual data by analyzing and processing the data. It can be categorized into mainly three types: document level, sentence level and aspect level based on the level of granularity of processing. It will classify the data into one of the three classes: positive, negative, or neutral, based on the sentiment orientation of the text. This is a trending area of research with significant applications like social media analytics, market prediction, reputation management, etc.

Understanding a natural language by a machine is a difficult process to tackle. Very few works have been carried out in mining the opinions expressed in Indian regional languages, especially Tamil language, and also it is significantly less compared to other foreign languages like English, Chinese, etc. [1][2] In this paper,

the authors performed a study on Tamil opinion mining utilizing a deep learning-based model called BERT. BERT is a transformer-based pre-trained language model developed by Google for language processing problems, which can perform efficiently on language problems, and it supports 104 languages all over the world. [8] This is because it is pre-trained on billions of textual data on Google books and Wikipedia, thereby understanding the language syntactically and semantically better than any other machine learning models. In this work, BERT is used to analyze the opinions expressed in Tamil on Tamil Movies. One of the main reasons for the lack of work on opinion mining in Indian regional languages, including Tamil, is the deficiency of benchmark datasets in these languages. Due to the lack of publicly available datasets in the Tamil language, the authors extracted 200 Tamil movie reviews from various social media platforms like Facebook, YouTube, etc., and manually labeled them according to their overall sentiment orientation. These labeled data are preprocessed and then trained using the multilingual BERT model in the Tamil language, and overall opinion about the movies are classified into positive, negative, or neutral. Furthermore, different machine learning and deep learning approaches, including K-Nearest Neighbors, Decision Tree, Random Forest, Logistic Regression, SVM, Naive Bayes, LSTM, Bi-LSTM, GRU, and Bi-GRU, are used to compare the results with that of BERT on the Tamil language. It is found that BERT outperformed all other machine learning and deep learning models on Tamil movie reviews.

Opinion mining on Indian regional languages, especially Tamil language is one of the less explored area of research in this domain. Vallikannu Ramanathan et al. extracted 50 tweets from Twitter which are movie reviews, and the authors used Tamil SentiWordNet to identify the sentiments of the tweets. They have used the TF-IDF method for sentiment classification, and furthermore, they used domain-specific ontology to improve their results. Sajeetha Thavareesan et al. collected 2691 Tamil comments from social media platforms like Twitter, Facebook, etc., and created a corpus containing Tamil lexicons with sentiments. Tamil OM was done and compared by utilizing the lexicon approach, ML approach, hybrid approach, and K-modes with BoW approach and K-means with Bag of Word (BoW) approach. Due to the unavailability of code-mix Tamil-English annotated datasets, Bharathi Raja Chakravarthi et al. created a corpus containing 15,744 comments extracted from YouTube. This annotated dataset was used to compare with various ML as well as DL models for OM. Bharathi Raja Chakravarthi et al. also created a dataset for multi-modal OM in Tamil and Malayalam, containing 134 videos with manually given transcription for every video. Also, these transcriptions were annotated with respective sentiment values for the task of OM. Sajeetha Thavareesan et al. used k-nearest neighbor and k-means clustering classifiers for carrying out Tamil OM. In this work, the authors have used fastText and BoW for embedding the Tamil movie reviews that they have collected and labeled. Furthermore, it is found that fastText is giving better results when compared to BoW over four experiments they have conducted. Anbukkarasi et al. extracted 1500 Tamil tweets from Twitter for the task of Tamil OM. Unlike other languages, in Tamil, syllables and dialects are different. Hence the authors have used a combined character-based method utilizing the Bi-LSTM model in their work.

## 2. RESEARCH METHOD

The main intention of this study is to evaluate the performance of the BERT model in the Tamil language. Various ML and DL models, such as LR, KNN, DT, NB, SVM, RF, LSTM, Bi-LSTM, GRU, and Bi-GRU, were used to compare with the performance of BERT. These methods are concisely described below.

**Decision Tree (DT):** DT is one of the most popular ML methods which can be applied to both regression and classification problems. A DT employs a tree-like structure to organize decisions and the potential outcomes and repercussions of those actions. Each internal node in this diagram represents a test on an attribute, and each branch the result of the test. A DT's outcome will be more accurate the more nodes it contains. Furthermore, they have the virtue of being simple to use and intuitive, but they need to be more precise. [8][9]

**Random Forest (RF):** A huge number of DT's make up the ensemble learning technique known as RF. Each DT in a RF makes a prediction, and the prediction that receives the most votes is taken as the result. As in DT, a RF model can also be applied to classification and regression issues. The majority of votes are used to determine the RF's result for the categorization challenge. In contrast, the output of a regression task is derived from the mean or average of the predictions made by each tree. [13][14]

**Logistic Regression (LR):** Problems with classification in machine learning are resolved using LR. Similar to linear regression, they are used to predict categorical variables. It can predict the outcome as True, False, 0 or 1, or Yes or No. Instead of providing exact numbers, it rather generates probabilistic values between 0 and 1. [10][11]

**Support Vector Machine (SVM):** The well-known machine learning technique called Support Vector Machine, or SVM, is frequently employed for classification and regression applications. But specifically, it's employed to address classification issues. Finding the best decision boundaries in an N-dimensional space that can divide data points into classes is the basic goal of SVM, and the optimum decision boundary is referred to as a Hyperplane. Support vectors are the extreme vectors that SVM chooses to locate the hyperplane. [12][13]

**Naïve Bayes (NB):** Another well-known classification technique used in machine learning is NB. It is known by this name because it is based on the Bayes theorem and uses the naive (independent) assumption between the features, which is stated as:

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

Every NB classifier assumes that a particular variable's value is independent of all other variables and features. For instance, if a fruit needs to be categorized according to its color, shape, and flavor. Mango will therefore be identified as being yellow, oval, and sweet. Each feature, in this case, operates independently of the others. [15][16]

**K-Nearest Neighbor (KNN):** Assuming that the new and previous cases are comparable, the K-NN technique assigns the new instance to the category most similar to the old ones. The K-NN algorithm is used to classify a unique data point based on similarity after all the prior data has been saved. This indicates that when new data appears in the proper category, the K-NN algorithm can categorize it quickly. Although the K-NN technique is most frequently used for classification problems, it can also be used for regression. [17][18]

**Long Short Term Memory (LSTM):** One major problem with ML models is the exploding gradient, which occurred as a result of the collection of essential and unimportant data. A machine learning model that can determine what information from a paragraph is relevant and remember only relevant information, and discard any extraneous information. Gates are used to resolving this problem. Gates serve as an inbuilt mechanism in the DL models like LSTM (Long-Short-Term Memory) and GRU (Gated Recurrent Unit), controlling which information is retained and which is discarded. LSTM and GRU networks resolve the exploding and disappearing gradient problem in this manner. Every LSTM network essentially has three gates to manage information flow and cells to store information. Information is carried by the Cell States without disappearing from the beginning to later time steps. An LSTM has three gates, viz., input, forget and output gates. Forget gate makes the determination of what information should be carried and what information should be ignored. After the relevant information has been determined, the information is sent to the input gate, which then sends the pertinent information, which updates the cell states. After the input gate has processed the data, the output gate is now activated. The next hidden states are generated by the output gate, and cell states are carried over to the following time step. LSTM operates in a single direction. LSTM systems with bidirectional communication are improvements over unidirectional LSTM. The Bi-LSTM aims to gather data from both directions, from left to right and right to left. Most of the Bi-LSTM concept is identical to that of LSTM, but it increases the models' accuracy. [19][20]

**Gated Recurrent Unit (GRU):** LSTM networks and GRU are similar DL models. A more recent variation of RNN is GRU, where there is no cell state. GRU transmits information using its hidden states. There are just two gates: Reset and Update Gate in GRU. Hence it is more rapid than LSTM. In order to prevent gradient explosion, this gate resets the previous information. Reset Gate decides how much of the past information should be forgotten. Forget Gate and Input Gate are combined to create Update Gate. The forget gate determines which information should be added to memory and which should be ignored. In order to prevent gradient explosion, this gate resets the previous information. Reset Gate decides how much of the past information should be forgotten. [21][22]

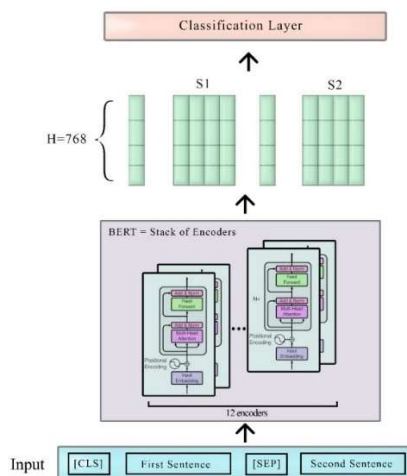


Figure 1 Overall architecture of BERT

**BERT** (Bidirectional Encoder Representations from Transformers): As the name suggests, it is advanced model of Transformer. BERT uses the encoder representations from the Transformer architecture. This model contains multi-head attention mechanism to deeply understand a language. It conditions simultaneously on both the left and the right context to pre-train deep bidirectional representations from the unlabeled text. In order to create cutting-edge models for a range of NLP tasks, the pre-trained BERT model can be improved with just one more output layer. Bidirectional means that BERT learns about a token's context's left and right sides during training. To completely understand the meaning of a language, a model needs to be bidirectional. One of the significant advantages of BERT is its positional, sentence, and token embeddings. These embeddings help the BERT model understand a language better than any other ML models. For many NLP tasks, the BERT model can be enhanced by merely adding a few extra output layers, and hence it is the state-of-the-art method for many NLP tasks. BERT is a pre-trained model, which means that it is trained on billions of textual contents in Google books, Wikipedia, etc. So this helps to understand the context knowledge from all the text data. Another advantage of BERT is that it supports 104 languages, including the south Indian language Tamil. The overall architecture of BERT model is given in figure 1. [23][24][25][26]

The primary intention of this work is to evaluate BERT's performance for Tamil OM. For that, a Tamil movie review dataset is created by extracting 500 Tamil movie reviews from social media platforms like Facebook and Youtube. These reviews are manually annotated with corresponding sentiments. Positive sentiment reviews are annotated with value 1, and negative sentiment reviews as 0. Example for positive and negative Tamil reviews with their corresponding English translations are shown in Table 1 and Table 2 respectively. Now, this dataset is used to do the task of Tamil document-level OM. The overall workflow of Tamil movie review OM is illustrated in figure 2.

Table 1. Positive Reviews with English Translation

Positive Reviews
சிறப்பான தரமான படம் ... இது போன்ற படங்கள் தொடர்ச்சியாக வர நல்வாழ்த்துக்கள் (Excellent quality movie ... Wish you to keep coming with more movies like this.)
அருமையான படம். வெற்றிமாறனுக்கு நன்றி. (Great Movie. Thanks to Vetrimaaran)

Table 2. Negative Reviews with English Translation

Negative Reviews
சட்டம் கடுமையா மாத்தி.. இனி சாதி பற்றி படம் எடுக்க கூடாது. இன்னு வரணும். (If the law is not strict, we should not make films about caste anymore. More to come.)
வலியை அறியாதவனால் இந்த படத்தின் கதையம்சத்தை உணர முடியாது (If you don't know the pain, you can't understand the plot of this film.)

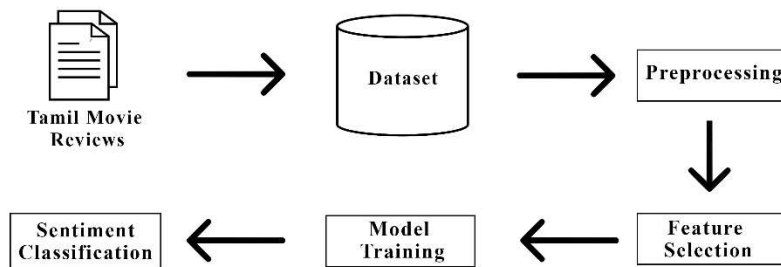


Figure 2 Workflow of Tamil movie review Opinion Mining

Once the Tamil movie review labeled dataset is created, it is passed to preprocessing stage, where all the unnecessary details, like blank spaces, special characters, mentions, etc., are removed from the Tamil movie reviews. Further, the cleaned reviews are input to the feature selection stage for selecting the most significant features. Here TF-IDF feature selection method is used in this work. Unlike ML models, DL models automatically extract the features without explicitly mentioning any feature selection method. Once the Tamil movie review labeled dataset is created, it is passed to preprocessing stage, where all the unnecessary details,

like blank spaces, special characters, mentions, etc., are removed from the Tamil movie reviews. Further, the cleaned reviews are input to the feature selection stage for selecting the most significant features. Here TF-IDF feature selection method is used in this work. Unlike ML models, DL models automatically extract the features without explicitly mentioning any feature selection method. Now various ML and DL models, including the BERT model, are used to train over the Tamil movie review dataset. In the final stage, the corresponding sentiment of the movie reviews is classified into positive or negative.

### 3. RESULTS

In this work, a study on Tamil opinion mining on Tamil movie reviews is done employing the BERT model and also with other ML and DL models. This work used K-NN, DT, RF, LR, SVM, NB, LSTM, Bi-LSTM, GRU, and Bi-GRU models to compare the results with that of BERT. The results show that BERT outscored all other models on the accuracy metrics with an accuracy of 84%. Most least performed model is KNN with an accuracy of 58%. Table 3 shows the results of Tamil OM. The confusion matrix of the BERT model is given in figure 3.

Table 3 Results

Model	Precision	F Score	Recall	Accuracy
DT	0.71	0.70	0.69	0.70
LR	0.84	0.82	0.81	0.69
SVM	0.88	0.85	0.84	0.70
RF	0.83	0.82	0.82	0.74
NB	0.62	0.60	0.59	0.69
KNN	0.41	0.40	0.40	0.58
LSTM	0.77	0.77	0.77	0.76
Bi-LSTM	0.79	0.79	0.79	0.78
GRU	0.79	0.79	0.78	0.78
Bi-GRU	0.80	0.80	0.80	0.79
BERT	0.82	0.82	0.83	0.84

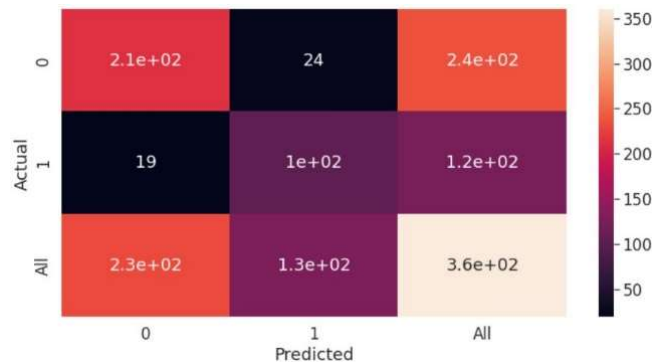


Figure 3 Confusion matrix of BERT model

### 4. CONCLUSION

Tamil OM is one of the less explored areas of research where significantly fewer works happened compared to other languages. One of the main reasons for this is the lack of benchmark datasets on Tamil language to work on. In this work, 500 Tamil movie reviews are collected and labeled according to their sentiments. A document-level OM was conducted on this dataset using BERT and other ML and DL models. It is found that BERT outperforms all other models with a marginal accuracy value. BERT scored 84% accuracy, while the next best-performing Bi-GRU got 79%. KNN attained only 62% accuracy, being the least-performing model. Since the dataset size is comparably small, our models suffer from overfitting even though dropout regularizations are added. The results clearly show that BERT performs well in the Tamil language. In the future, this work will be extended with the expanded dataset for better results and remarks.

### REFERENCES

- [1] V. Ramanathan, T. Meyyappan, and S. M. Thamarai, "Sentiment Analysis: An Approach for Analysing Tamil Movie Reviews Using Tamil Tweets," Recent Advances in Mathematical Research and Computer Science Vol. 3. Book Publisher International (a part of SCIENCEDOMAIN International), pp. 28–39, Oct. 27, 2021. doi: 10.9734/bpi/ramrcs/v3/4845f.
- [2] S. Thavareesan and S. Mahesan, "Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation," 2019 14th Conference on Industrial and Information Systems (ICIIS). IEEE, Dec. 2019. doi: 10.1109/iciis47346.2019.9063341.

- [3] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, 'Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text', Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), 2020, 202–210.
- [4] Chakravarthi, Bharathi Raja, K. P. Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, and John P. McCrae. "Dravidianmultimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam." arXiv preprint arXiv:2106.04853, 2021.
- [5] S. Thavareesan, S. Mahesan, "Sentiment Analysis in Tamil Texts using k-means and k-nearest neighbour," 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS), 2021, pp. 48-53, doi: 10.1109/ICIAfS52090.2021.9605839.
- [6] S. Anbukkarasi and S. Varadhaganapathy, "Analyzing Sentiment in Tamil Tweets using Deep Neural Network," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). IEEE, Mar. 2020. doi: 10.1109/iccmc48092.2020.iccmc-00084.
- [7] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR. 2018;abs/1810.04805. <http://arxiv.org/abs/1810.04805>
- [8] A. Bayhaqy, S. Sfenrianto, K. Nainggolan and E. R. Kaburuan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes," 2018 International Conference on Orange Technologies (ICOT), 2018, pp. 1-6, doi: 10.1109/ICOT.2018.8705796.
- [9] H. T. Phan, V. C. Tran, N. T. Nguyen, and D. Hwang, "Decision-Making Support Method Based on Sentiment Analysis of Objects and Binary Decision Tree Mining," Lecture Notes in Computer Science. Springer International Publishing, pp. 753–767, 2019. doi: 10.1007/978-3-030-22999-3\_64.
- [10] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, pp. 416-419, doi: 10.1109/ICATCCT.2016.7912034.
- [11] W. P. Ramadhan, S. T. M. T. Astri Novianty and S. T. M. T. Casi Setianingsih, "Sentiment analysis using multinomial logistic regression," 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC), 2017, pp. 46-49, doi: 10.1109/ICCEREC.2017.8226700.
- [12] M. Rathi, A. Malik, D. Varshney, R. Sharma and S. Mendiratta, "Sentiment Analysis of Tweets Using Machine Learning Approach," 2018 Eleventh International Conference on Contemporary Computing (IC3), 2018, pp. 1-3, doi: 10.1109/IC3.2018.8530517.
- [13] M. Aufar, R. Andreswari and D. Pramesti, "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study," 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020, pp. 1-7, doi: 10.1109/ICoDSA50139.2020.9213078.
- [14] Huang, Xin, Wenbin Zhang, Xuejiao Tang, Mingli Zhang, Jayachander Surbiryala, Vasileios Iosifidis, Zhen Liu, and Ji Zhang. "Lstm based sentiment analysis for cryptocurrency prediction." In International Conference on Database Systems for Advanced Applications, pp. 617-621. Springer, Cham, 2021. doi: <https://doi.org/10.48550/arXiv.2103.14804>
- [15] Shamrat, F. M. J. M., Sovon Chakraborty, M. M. Imran, Jannatun Naeem Muna, Md Masum Billah, Protiva Das, and O. M. Rahman. "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm." Indonesian Journal of Electrical Engineering and Computer Science, vol. 23, no. 1. Institute of Advanced Engineering and Science, p. 463, Jul. 01, 2021. doi: 10.11591/ijeecs.v23.i1.pp463-470.
- [16] P. Gamallo and M. Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets," Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for Computational Linguistics, 2014. doi: 10.3115/v1/s14-2026.
- [17] Singh, Jyotsna, and Pradeep Tripathi. "Sentiment analysis of Twitter data by making use of SVM, Random Forest and Decision Tree algorithm." In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), pp. 193-198. IEEE, 2021.
- [18] A. Alshamsi, R. Bayari, and S. Salloum, "Sentiment Analysis in English Texts," Advances in Science, Technology and Engineering Systems Journal, vol. 5, no. 6. ASTES Journal, pp. 1683–1689, Dec. 2020. doi: 10.25046/aj0506200.
- [19] Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain, "Sentic LSTM: a Hybrid Network for Targeted Aspect-Based Sentiment Analysis," Cognitive Computation, vol. 10, no. 4. Springer Science and Business Media LLC, pp. 639–650, Mar. 14, 2018. doi: 10.1007/s12559-018-9549-x.
- [20] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," Neural Computing and Applications, vol. 32, no. 13. Springer Science and Business Media LLC, pp. 9713–9729, Sep. 19, 2019. doi: 10.1007/s00521-019-04504-2.
- [21] M. M. Abdelgwad, T. H. A. Soliman, A. I. Taloba, and M. F. Farghaly, "Arabic aspect based sentiment analysis using bidirectional GRU based models," Journal of King Saud University - Computer and Information Sciences. Elsevier BV, Sep. 2021. doi: 10.1016/j.jksuci.2021.08.030.
- [22] F. Liu, J. Zheng, L. Zheng, and C. Chen, "Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification," Neurocomputing, vol. 371. Elsevier BV, pp. 39–50, Jan. 2020. doi: 10.1016/j.neucom.2019.09.012.
- [23] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis." arXiv, 2019. doi: 10.48550/ARXIV.1904.02232.
- [24] Hoang, Mickel, Oskar Alija Bihorac, and Jacobo Rouces. "Aspect-based sentiment analysis using BERT." In Proceedings of the 22nd nordic conference on computational linguistics, pp. 187-196. 2019.

- 
- [25] Z. Gao, A. Feng, X. Song and X. Wu, "Target-Dependent Sentiment Classification With BERT," in IEEE Access, vol. 7, pp. 154290-154299, 2019, doi: 10.1109/ACCESS.2019.2946594.
- [26] H. EL Moubtahij, H. Abdelali, and E. B. Tazi, "AraBERT transformer model for Arabic comments and reviews analysis," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 11, no. 1. Institute of Advanced Engineering and Science, p. 379, Mar. 01, 2022. doi: 10.11591/ijai.v11.i1.pp379-387.



## Memory Based Learning of Tamil Morphology

**K.Rajan**

Department of Computer Engineering,  
Muthiah Polytechnic College, Annamalai University  
Annamalainagar

---

### ABSTRACT

A Morphology of a natural language is a more fundamental and important for any language processing task. Morphological analysis of a word, in agglutinative and morphologically rich languages, is the prerequisite for processing text in higher levels. Models that are able to learn the morphology of natural languages are of great practical interest as tools for descriptive linguistic analysis and for minimizing the expert resources needed to develop morphological analyzers and stemmers. Memory-based learning (MBL) is a machine-learning method based on the idea that examples can be used directly in processing natural language problems. Training examples are stored. During the classification process, the most similar examples from the training data are located, and their class is used to classify the new example [1]. In this paper the MBL technique has been used for the analysis of Tamil morphology. The word forms collected from a tagged corpus have been used to select instances of segmentations. The application of memory-based learning to morphological analysis is reformulated as a classification task in which letter sequences are classified as morphemes and their categories and morpheme boundaries are identified. The generalization performance of memory-based learning algorithms to Tamil morphological analysis task is encouraging.

*Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.*

---

### Keywords:

Machine learning  
Memory based learning  
Instances based learning  
Tamil morphology

---

### Corresponding Author:

Dr.K.Rajan,  
Department of Computer Engineering,  
Muthiah Polytechnic College  
Email: [rajank.mptc@gmail.com](mailto:rajank.mptc@gmail.com)

---

## 1. INTRODUCTION

In morphologically rich languages like Tamil, the word-based language models are not so helpful because of its free word-order nature and out-of-vocabulary problem. The meaning and the category of a word depends mainly on its constituent morphemes. So, identifying the constituents and their meaning within the context is trivial for the analysis of Tamil words. It makes morphological analysis (MA) an important component of a number of NLP systems for morphologically rich languages. The two important issues to be considered for MA are: the segmentation of word into morphemes and the computation of the grammatical characteristics of the overall word. The segmentation of word into morphemes is itself an ambiguous task due to the presence of allomorphs and morphophonemic changes. There are different morphological analysers based on grammar rules, finite state automata and machine learning algorithms.

Memory based learning has been applied to various natural language processing tasks like morphological analysis and parts-of-speech tagging in English, Arabic and Dutch. [2][3]. The number of inflections in any language is finite, all the word-forms can be listed for a given word, assuming the category of a word is known like verb or noun.

In this type of learning, the classifiers keep the examples in memory, without creating abstraction in the form of rules or decision trees. Generalisation is achieved for a new input pattern by retrieving the most similar item

in the memory. A word's part-of-speech (POS) indicates its syntactic function within a sentence, e.g. noun, verb, etc. POS tagging is an important pre-processing step for a variety of NLP tasks. POS information can help to predict which words are likely to occur next, from knowledge of syntactic structure, (through language modelling) in speech recognition system. It is also used in word sense disambiguation (WSD), machine translation (MT) system and as an important part of general corpus-based linguistic research.

Machine learning based techniques have been applied to language learning and acquisition problems among the NLP community. The major problems addressed by machine learning techniques are those of natural language disambiguation appearing at all levels of language understanding process. These problems can be recast as classification problems, a generic type of problem in the field of Artificial Intelligence.

Memory based techniques, based on the principles of non-parametric density estimation, are powerful form of machine learning well-suited for natural language processing. Memory based learning (MBL) algorithms have appeared in many areas of AI with different names: **Similarity-based, example-based, case-based, instance-based, exemplar-based, analogical and lazy-learning**. MBL is a supervised inductive learning algorithm for learning classification tasks. In this paper, we proposed and implemented memory-based learning (MBL) approach for the segmentation of word, which identify the morpheme boundaries, and parts of speech tagging which assign morph level grammatical categories.

## 2. TAMIL MORPHOLOGY

Tamil is an agglutinative and concatenative language, where morphemes are strung together to form long words. Tamil is a verb final language with SOV (Subject Object Verb) pattern. The verbs can have a long sequence of morphemes that express tense, mood and aspects as well as sense of assertion, negation, interrogation, reflection, emphasis, etc. Nouns take plural marker to make plural forms, and they decline to different inflected forms by taking case markers. Concatenative morphology in Tamil involves always suffixation. This is represented as

$$\text{Stem} + (\text{affix})^n$$

where the superscript  $n$  means one or more occurrences of a suffix. Morphs are concatenated as suffixes at the right of a word stem, to produce inflected or derived forms of words.

## 3. RESEARCH METHOD

### Data set and selection of instances

For the memory-based learning approach, the CIIL Tamil corpus of 3 million words has been used as a base for the selection of patterns. For training, the conjugated forms of verbs and the inflected forms of nouns are collected. The Tamil morphological analyser [5] is used to predict the morph level categories (tags) and the word level categories. Un-inflected word forms are omitted. The following table 1 shows the part of the tagged wordlist from the corpus. From these word-forms, stems and suffixes are separated and the distinct suffix patterns are stored as instances. The training data has 6450 unique patterns for various stem and suffix combinations.

The suffix has one or more morphemes. Each word is marked with a minimum of one morph level tag and maximum of four tags. These morph level tags contribute to the meaning of the entire word. The stem category, such as *vb* (verb), *abn* (abstract noun) etc. are marked. For this model, 7 stem types, 15 characters for suffix, and 70 morph-level tags are considered. These morph-level tags are useful to interpret the actual meaning of words according to the stem category.

Table 1. Morphologically analysed word forms with morpheme level tags and word level tags.

Filename	Word	Tagword	Length	Root	Suffix	Num
5AP	நான்னுத்தம்	நான்னு vb த்த pst_e vp <NVvp>	7	நான்னு	த்தம்	1
5AP	வனுத்திரைக்காவும்	வனுத்திரை vb க்கஅ inf வும் part <NVi>	13	வனுத்திரை	க்கஅவும்	1
5AP	அரனுத்தம்	அரனு vb த்த pst_e vp <NVvp>	6	அரனு	த்தம்	1
5AP	வனுட்டம்	வனுட் vb ட் pst_e vp <NVvp>	5	வனுட்	ட்டம்	1
5AP	புட்டம்	புட்ட vb ட் pst_e vp <NVvp>	5	புட்ட	ட்டம்	1
5AP	வனுக்காவும்	வனு vb க்கஅ inf வும் part <NVi>	8	வனு	க்கஅவும்	1
5AP	ஊறினா	ஊற vb இன் pst அ rp <NVrp>	5	ஊற	இனா	1
5AP	அரனுத்தம்	அரனு vb த்த pst_e vp <NVvp>	6	அரனு	த்தம்	1
5AP	வனுட்டம்	வனுட் vb ட் pst_e vp <NVvp>	7	வனுட்	ட்டம்	1
5AP	வனுக்காவும்	வனு vb க்கஅ inf வும் part <NVi>	8	வனு	க்கஅவும்	1
5AP	ஆறினா	ஆற vb இன் pst அ rp <NVrp>	5	ஆற	இனா	1
5AP	உருட்டம்	உருட் vb ட் pst_e vp <NVvp>	6	உருட்	ட்டம்	1
5AP	புட்டம்	புட்ட vb ட் pst_e vp <NVvp>	5	புட்ட	ட்டம்	1
5AP	புனுத்தம்	புனு vb த்த pst_e vp <NVvp>	7	புனு	த்தம்	1
5AP	வனுக்காவும்	வனு vb க்கஅ inf வும் part <NVi>	8	வனு	க்கஅவும்	1
5AP	ஊறினா	ஊற vb இன் pst அ rp <NVrp>	5	ஊற	இனா	1
5AP	வனுத்தம்	வனு vb த்த pst_e vp <NVvp>	5	வனு	த்தம்	1
5AP	வனுட்டம்	வனுட் vb ட் pst_e vp <NVvp>	6	வனுட்	ட்டம்	1
5AP	வனுத்தம்	வனு vb த்த pst_e vp <NVvp>	5	வனு	த்தம்	1
5AP	அரனுத்தம்	அரனு vb த்த pst_e vp <NVvp>	6	அரனு	த்தம்	1

The text file of the CIIL corpus is tokenized and the words are morphologically analysed. In table 1, only the morphologically analysed verbal forms with morpheme level tags and word level tags are shown.

Table 2. Morphemes at different levels with tags

M1	Cat1	Suffix	M2	Cat2	M3	Cat3	Cat5
இந்திய்யு	ppn	வியு	வியு	loc			NNloc
புராராய்ச்	abn	ஐ	ஐ	acc			NNacc
உருவ்யுக்க	vb	இய்யு	இ	pst	ய்யு	rp	NVrp
புராராய்ச்	vb	த்தய்யு	த்தய்யு	pst	ய்யு	pnhup	NNplu
அராய்ச்	hn	கய்யு	கய்யு	plu			NNplu
புராராய்ச்	ppn	த்தய்யு	த்தய்யு	acc			NNacc
இராய்ச்சிய்யு	ian	த்தய்யு	த்தய்யு	acc			NNacc
சுய்யு	abn	கய்யு	கய்யு	plu	இய்யு	loc	NNloc
இராய்ச்சிய்யு	vb	நத்தய்யு	நத்தய்யு	pst	ய்யு	rp	NVrp
கய்யு	abn	த்தய்யு	த்தய்யு	abl	ய்யு	emp	NNemp
அய்யு	hn	இய்யு	இய்யு	gen			NNgen
இராய்ச்சிய்யு	vb	நத்தய்யு	நத்தய்யு	pst	ய்யு	3sn	FV
தய்யு	vb	இய்யு	இ	pst	ய்யு	rp	NVrp
நய்யு	abn	கய்யு	கய்யு	plu			NNplu
நய்யு	abn	த்தய்யு	த்தய்யு	loc	உய்யு	inc	NNinc
கய்யு	abn	ய்யு	ய்யு	loc	உய்யு	inc	NNinc
சுய்யு	abn	த்தய்யு	த்தய்யு	loc	உய்யு	inc	NNinc

In the above table 2, the morpheme M1 is the stem and Cat1 is the stem category. Cat5 is the word category. The distinct suffix forms are selected as instances and stored.

In Tamil similar suffix characters occur with different stem types. A particular stem may belong to a noun or a verb, but the suffix differs based on the type. Instead of using only the suffixes as instances for identifying the root and suffix boundary, the concatenation of three characters from stem ending and upto five characters from beginning of the suffix are used. This character sequence is given in the 'Boundary' column, in table 3.

Table 3. List of roots, suffixes and boundary with left and right context.

1	Root	Suffix	Boundary
2	இந்தியஆ	வஇவ்	இயஆ+வஇவ்
3	பொருஅசு	ஐ	ர்அசு+ஐ
4	உர்உவஆக்க	இயஅ	ஆக்க+இயஅ
5	புஅட்டி	த்தஅவஅர்கஅள்	அட்டி+த்தஅவஅ
6	அர்அசுஅர்	கஅள்	சுஅர்+கஅள்
7	புஅட்டிபுத்தஇர்அ	த்தஐ	இர்அ+த்தஐ
8	இர்ஆச்சுஇயஅ	த்தஐ	இயஅ+த்தஐ
9	சுஅம்அவஎள்ளி	கஅள்ளிவ்	எள்ளி+கஅள்ளிவ்
10	இர்உ	ந்தஅ	இர்உ+ந்தஅ
11	கஆவ்அ	த்தஇவ்அன்	ஆவ்அ+த்தஇவ்அ
12	அவஎக்கஆண்டஅர்	இன்	ட்டஅர்+இன்
13	இர்உ	ந்தஅத்உ	இர்உ+ந்தஅத்உ
14	தொன்ற	இயஅ	ஒன்ற+இயஅ
15	நண்ணாண்ட	கஅள்	ண்ட+கஅள்
16	நிர்வஆகஅ	த்தஇவ்உம்	ஆகஅ+த்தஇவ்உம்
17	கஅவ்ஐ	யஇவ்உம்	அவ்ஐ+யஇவ்உம்
18	சுஅம்அய்யஅ	த்தஇவ்உம்	அய்யஅ+த்தஇவ்உம்

#### 4. TRAINING

Memory based learning is a form of supervised learning, where examples are represented as a vector of feature values along with an associated category label. Features define a pattern space in which similar examples are associated with similar category. Several features can be associated with the same category which enables polymorphous concepts to be learned. During training, a set of examples is presented to the classifier and added to the memory. In non-parametric estimation, it is assumed that similar inputs have similar outputs. Finding the similar past instances from the training set using a suitable distance measure and interpolating them will find the right output. Because of the need for large memory and computation, this approach was not popular earlier. With advances in computing and memory, such methods have become more widely used recently.

In memory-based learning, all the training instances are stored in a lookup table. Given an input, similar ones should be found from the table using a suitable distance measure and interpolating from them to find the correct output. This technique keeps typical instances for each class. An instance consists of a fixed-length vector of  $n$  feature-value pairs, and an information field containing the classification of that particular feature-value vector. After the instance base is stored, new (test) instances are classified by matching them to all instances in the instance base, and by calculating with each match the distance, given by a distance function

The memory-based learning reproduces the classification of training data flawlessly, as long as there are no identical training instances in memory with different class labels. In morphological analysis, the NLP system produces segmentation for which it has multiple interpretations, and it must decide which interpretation is appropriate in the current context. In order to resolve this ambiguity, it is necessary to consider two or more semantically, syntactically or structurally distinct forms based on the properties of the context. In morphological segmentation application, the ambiguous suffixes are assigned correct category based on their stem type. The training data set has 6450 unique patterns for various stem and suffix combinations. They have a boundary information with left and right context. So they can predict the boundary in given input character sequence.

Memory based learning (MBL) is based on the idea that learning and processing are two steps, where learning is the storage of examples in memory, and processing is similarity-based reasoning with these stored examples. The similarity between a new instance  $X$  and all examples  $Y$  in memory is computed using a similarity metric (that actually measures distance)  $\Delta(X,Y)$ . Classification works by assigning the most frequent class

within the k most similar example(s) as the class of a new test instance

## 5. MORPHEME SEGMENTATION

For morpheme segmentation task, the training data with sequence of 6 characters are prepared, in which 3 characters on the left of the boundary and 3 characters on right. The presence of a boundary is marked. In this task, the character sequence is added with word beginning and word ending information as additional characters. The word is analysed by taking a sliding window of 6 characters and they are matched with the instances stored in the memory. The matching sequence indicates the boundary information.

## 6. RESULTS AND ANALYSIS

The basic metric that works for instances with symbolic features is the *overlap metric* which is used to estimate the distance between instances. The distance between two patterns is simply the sum of the differences between features. The classification (morphological analysis) of the test instance is taken from the closest match from the memory instances. The performance is measured from the precision and recall estimated from morpheme categories predicted correctly. This system maps the letter sequences in context with the categories and boundaries. When multiple matching instances occur, the output will be a sequence of morpheme boundaries. The application of this method produces the boundary between stems and suffixes, boundary between morphemes. The accuracy of segmentation for longer wordforms is greater when compared to short words. These experiments are conducted with varying width of context. The results are compared with the actual segmentations performed with rule based morphological analysers. The output of memory-based learning technique in segmentation is 96% accurate and 92% accurate in morph level tagging.

## 7. CONCLUSION

The morphological feature of Tamil, which is agglutinative can be easily analysed using memory-based learning techniques. As the segmented and tagged data is available for training this algorithm, the grammatical features of the morphemes are also identified. Both the stemming and morpheme segmentaion tasks can be performed with this approach. The longer contextual information produces more accuracy than shorter context. The short suffixes have multiple interpretations.

## REFERENCES

- [1] W. Daelemans, "Memory-based lexical acquisition and processing", In P. Steffens, editor, Machine Translation and the Lexicon, Lecture Notes in Artificial Intelligence, pages 85-98. Springer-Verlag, Berlin. 1995.
- [2] Antal van den Bosch, et al., "Memory-based morphological analysis", Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 285-292, 1999
- [3] Daelemans, et al., "Evaluation of machine learning methods for natural language processing tasks. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, 2002.
- [4] Antal van den Bosch, "Using induced rules as complex features in memory-based language learning", Proceedings of CoNLL 2000, pages 73-78.
- [5] M.Ganesan, "Morph and POS Tagger for Tamil" (Software), Annamalai University, Annamalai Nagar, 2009.
- [6] Alexander Clark, " Partially supervised learning of morphology with stochastic transducers", In Proc. of Natural Language Processing Pacific Rim Symposium, NLPRS 2001, pages 341-348, Tokyo, Japan, November. 2001b.

# Computer-Assisted Learning System for the Tamil Grammar Punarial

Senthamizh Selvi S<sup>1</sup>, Anitha R<sup>2</sup>

<sup>1</sup>Research Scholar, <sup>2</sup>Professor and Head

<sup>1,2</sup>Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Tamil Nadu, INDIA

## ABSTRACT

In Tamil grammar, Conjugation (Punarial) is the spelling of the change in pronunciation that occurs when two words are joined. Of the two words, the first word is called the position term (நிலைமொழி) and the second word is called the derivative term (வருமொழி). When these two words are combined, the state in which no variations appeared in the generated word is called natural formation (இயல்பு புணர்ச்சி), and when variations (appearance, destruction, and distortion) appear in the generated word, it is called strain formation (விகாரப் புணர்ச்சி). To learn this grammar, we propose a rule-based computer-assisted language learning system - a kind of interactive learning method that makes the students improve the efficiency and effectiveness of learning, which can improve the quality of grammar comprehension. As Bhavananthi Munivar mentioned the Punarchi rules in the book Nannool, we developed the system with a few such rules that are frequently taught in schools. We have used Unicode character encoding to display Tamil characters. We have done English to Tamil transliteration to make Tamil typing easier. Since the developed system is a rule-based system, the accuracy is at its highest potential.

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

## Keywords:

Punarchi vidhi,  
Transliteration,  
Tamil Grammar

## Corresponding Author:

Senthamizh Selvi S,  
Department of Computer Science,  
Sri Venkateswara College of Engineering, Tamil Nadu, India  
Email: senthamizhselvi@svce.ac.in

## 1. INTRODUCTION

Tamil is one of the longest-surviving classical languages in the world natively spoken by the Tamil people of the Indian subcontinent. It is a morphologically rich language [1] and has well-structured grammar. But at the same time, it is a low-resource language. Many works are carried out in the process of Tamil computational linguistics. As a part of it, we propose an approach to work in the context of the development of a computational Punarial (புணர்ச்சி) grammar for Tamil words, which will make the students improve their efficiency and effectiveness in learning this grammar. In Tamil grammar, Conjugation (புணர்ச்சி) is a change in pronunciation when two words join. The first word is said to be a positional word and the next word is called a derived word. When two words are joined, variations such as the appearance of a letter or the final letter of the word changing to another letter or the disappearance of a letter occur. This state of variations in the conjugation is called strain formation (விகாரப் புணர்ச்சி). The state in which words are formed without variations is called natural formation (இயல்பு புணர்ச்சி). There is certain grammar to be followed when two words are joined together. Bhavananthi Munivar has mentioned more than 100 Punarchi rules in his book Nannool[2, 3]. We have developed a system by implementing eight such grammar rules.

## 2. CHALLENGES IN DEVELOPING PUNARIAL SYSTEM

Some of the following challenges are faced while developing a Punarial system.

1. Difficulty in understanding the Punarial Grammatical Rules.



2. Nowadays the computational approaches of Natural Language Processing is moved from rule-based model to data-driven models - statistical, machine learning, and deep learning models. For this, a large volume of datasets is required. But for this Punarial, it is yet to be created.
3. Most of the computers people using has English Keyboards. It enables us to type English characters. But to get Tamil characters, the transliteration of English to Tamil is required.

### 3. THE PROPOSED ARCHITECTURE

The proposed system architecture is shown in Fig.1. The user interface is designed to interact with the user. It takes two Tamil words typed in English as input. It displays the transliterated words of input in Tamil, the predicted output of the word after applying the Punarial rules and the list of rules that have been applied for the given words. The transliteration system is to facilitate Tamil typing. It converts English letters to Tamil letters. The fuzzy-based Punarial system takes two Tamil words, understands the words with the help of pre-classified Tamil letters, applies the rules of Punarial to produce the resultant word.

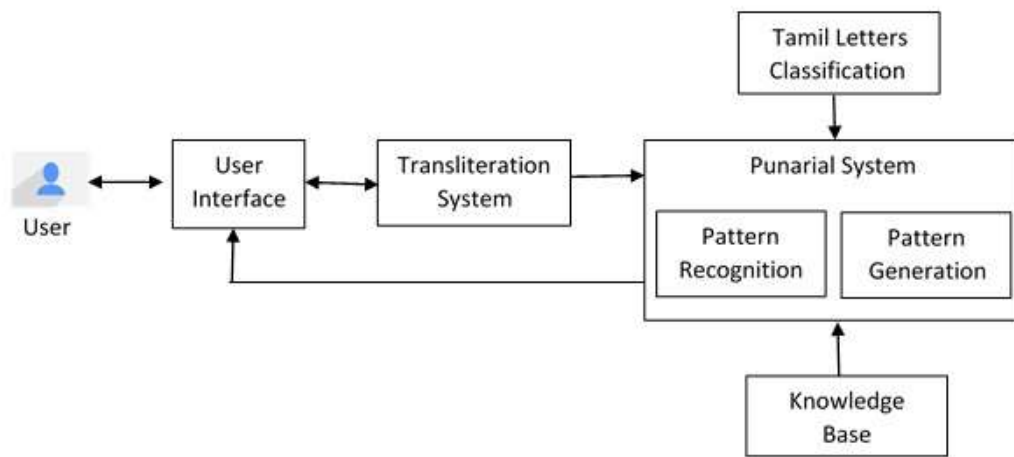


Figure 1. Proposed Architecture

#### 3.1 Transliteration System

Transliteration[4] is a method of mapping text from one script to another that involves swapping letters. In order to make Tamil typing easy, the letters typed in English are transliterated into letters in Tamil. Table 1 shows the equivalent English letters of Tamil vowels. When the letter “a” is typed, the transliteration system converts this to “அ”. Likewise, for the Tamil Consonants, the equivalent English letters are shown in Table 2. The compound form of a Tamil letter is combination of Consonants with Vowels. For the consonant “க”, the compound form is shown in Table 3.

Table 1. Vowels

Tamil Vowels	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஔ	ஓ	ஔ
Equivalent English Letters	a	aa	i	ee	u	oo	e	ae	ai	o	oa	au	h-

Table 2. Consonants

Tamil Consonants	க	ங	ச	ஞ	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள்	ற்	ண்
Equivalent English Letters	k	ng	s	nj	th	n-	p	m	y	r	l	v	zh	L	R	N

Table 3. Compound Letter for the Consonant “க”

k			->	க
k	+	a	->	கா
k	+	aa	->	காா
k	+	i	->	கி



k	+	ee	->	கீ
k	+	u	->	கு
k	+	oo	->	கூ
k	+	e	->	கெ
k	+	ae	->	கே
k	+	ai	->	கை
k	+	o	->	கொ
k	+	oa	->	கோ
k	+	au	->	கௌ

### 3.2 Punarial System

The Punarial system is a fuzzy-based system that works on the levels of possibilities of input words to achieve the definite output word. The Tamil letters are pre-classified into உயிர் எழுத்து, மெய் எழுத்து, உயிர்மெய் எழுத்து, குறில், நெடில், திசை, வல்லினம், மெல்லினம், இடையினம், etc., The classification of Tamil letters enables the system to recognize the pattern that exists in the position word (first word) and derivation word (second word). The knowledge base is constructed with a collection of fuzzy if-then rules that formally represent the knowledge of Punarial rules to be processed during approximate reasoning. The eight Punarial rules that are implemented in the proposed work are shown in Fig. 2. The Pattern generation applies the rules one after another and generates the conjugate word. The order of placing the rules are also placing a major role in the system. For example,

Rule No	Rules
1	உடல் மேல் உயிர் வந்து ஒற்றுவது இயல்பே
2	தனிக்குறில் முன் ஒற்று உயிர்வரின் இரட்டும்
3	இ, ஈ, ஐ வழி யவ்வும் ஏனை உயிர்வழிவவ்வும், ஏமுன் இவ்விருமையும் உயிர்வரின் உடம்படு மெய்யென்றாகும்
4	உயிர்வரின் உக்குறள் மெய்விட்டோடும் யவ்வரின் இய்யாம் முற்றுமற் றொரோவழி
5	பூ பெயர் முன் இன மென்மையும் தோன்றும்
6	நெடிலோடு டுயிர்த் தொடர்க் குற்றுக ரங்களுள் டறவொற் றிரட்டும் வேற்றுமை மிகவே
7	இயல்பினும் விதியினும் நின்ற உயிர்முன் கசதப மிகும்வித வாதன மன்னே
8	திசையோடு திசையும் பிறவும் சேரின் நிலையீற் றுயிர்மெய் கவ்வொற்று நீங்கலும் றகரம் னலவாத் திரிதலும் ஆம்பிற

Figure 2. Punarial Rules

for the input words “கண் + இமை”, if rule 1 “உடல் மேல் உயிர் வந்து ஒற்றுவது இயல்பே” is given before rule 2 “தனிக்குறில் முன் ஒற்று உயிர்வரின் இரட்டும்”, then the output would be “கண்ணிமை”. But the correct order is after applying rule 2, rule 1 should be applied. Rule 2 changes the input words to “கண்ண் + இமை”, which makes rule 1 generate the output to “கண்ணிமை”. On adding a new rule, the system must be tested once again to ensure the correctness of the output.

#### 4. RESULTS AND DISCUSSIONS

The system is implemented in Java. The test data are collected from school textbooks and websites. The system is tested with 72 test samples. The predicted output is the same as the expected output for 69 samples and brought the accuracy of the proposed system close to 96%. The sample output of the transliteration system is shown in Fig. 3. The predicted output that is the same as the expected output with its applied Punarial rules of the proposed system is shown in Fig. 4. For a few test samples shown in Fig. 5, the results are not expected. It is because the system applies more than the required rules. The screenshots of various test samples tested in the proposed system are shown in Fig. 6.

First Word	Second Word	Transliterated First Word	Transliterated Second Word
appaa	udan	அப்பா	உடன்
aangku	avaRRuL	ஆங்கு	அவற்றுள்
aaRu	paalam	ஆறு	பாலம்
uyarvu	adain-thaar	உயர்வு	அடைந்தார்
uyir	ezhuththu	உயிர்	எழுத்து
uL	onRu	உள்	ஒன்று
uRavu	azhaku	உறவ	அழகு
unnai	allaal	உன்னை	அல்லால்
eddi	kaay	எட்டி	காய்
orumai	udan	ஒருமை	உடன்
oadi	senRaam	ஓடி	சென்றான்
kaN	azhaku	கண்	அழகு
kaN	aayiram	கண்	இமை
kaN	aayiram	கண்	ஆயிரம்
kalai	aRivu	கலை	அறிவு
kaLiRu	yaanai	களிறு	யானை
kanaa	kaalam	கனா	காலம்
kaadu	maram	காடு	மரம்
kaadsi	azhaku	காட்சி	அழகு
kiNaRu	thavaLai	கிணறு	தவளை

Figure 3. Output of Transliteration System

Transliterated First Word	Transliterated Second Word	Expected Output	Predicted Output	Applied Rules
அப்பா	உடன்	அப்பாவுடன்	அப்பாவுடன்	3
ஆங்கு	அவற்றுள்	ஆங்கவற்றுள்	ஆங்கவற்றுள்	1,4
ஆறு	பாலம்	ஆற்றுப்பாலம்	ஆற்றுப்பாலம்	6,7
உயர்வு	அடைந்தார்	உயர்வடைந்தார்	உயர்வடைந்தார்	1,4
உயிர்	எழுத்து	உயிரெழுத்து	உயிரெழுத்து	1
உள்	ஒன்று	உள்ளொன்று	உள்ளொன்று	1,2
உறவ	அழகு	உறவழகு	உறவழகு	4
உன்னை	அல்லால்	உன்னையல்லால்	உன்னையல்லால்	3
எட்டி	காய்	எட்டிக்காய்	எட்டிக்காய்	7
ஒருமை	உடன்	ஒருமையுடன்	ஒருமையுடன்	3
ஒடி	சென்றான்	ஒடிச்சென்றான்	ஒடிச்சென்றான்	7
கண்	அழகு	கண்ணழகு	கண்ணழகு	1,2
கண்	இமை	கண்ணிமை	கண்ணிமை	1,2
கண்	ஆயிரம்	கண்ணாயிரம்	கண்ணாயிரம்	1,2
கலை	அறிவு	கலையறிவு	கலையறிவு	3
களிறு	யானை	களிற்றியானை	களிற்றியானை	4,6
கனா	காலம்	கனாக்காலம்	கனாக்காலம்	7
காடு	மரம்	காட்டுமரம்	காட்டுமரம்	6
காட்சி	அழகு	காட்சியழகு	காட்சியழகு	3
கிணறு	தவளை	கிணற்றுத்தவளை	கிணற்றுத்தவளை	6,7
கிழக்கு	நாடு	கிழ்நாடு	கிழ்நாடு	8
கிழக்கு	கடல்	கிழ்க்கடல்	கிழ்க்கடல்	8
குரங்க	யாது	குரங்கியாது	குரங்கியாது	4
கோடு	உயிர்	கோட்டுயிர்	கோட்டுயிர்	1,4,6
சே	இழை	சேலிழை / சேயிழை	சேலிழை / சேயிழை	3
சே	அடி	சேவடி / சேயடி	சேவடி / சேயடி	3
சோறு	சுவை	சோற்றுச்சுவை	சோற்றுச்சுவை	6,7
தனி	ஆழி	தனியாழி	தனியாழி	3
தீ	அணைப்பான்	தீயணைப்பான்	தீயணைப்பான்	3
தெற்கு	மேற்கு	தெல்மேற்கு / தென்மேற்கு	தெல்மேற்கு / தென்மேற்கு	8
தெற்கு	கிழக்கு	தெல்கிழக்கு / தென்கிழக்கு	தெல்கிழக்கு / தென்கிழக்கு	8
தெற்கு	பாண்டி	தெல்பாண்டி / தென்பாண்டி	தெல்பாண்டி / தென்பாண்டி	8
தெற்கு	குமரி	தெல்குமரி / தென்கு	தெல்குமரி / தென்கு	8
பலா	சுவை	பலாச்சுவை	பலாச்சுவை	7
பலா	பழம்	பலாப்பழம்	பலாப்பழம்	7
பாடி	தேடினான்	பாடித்தேடினான்	பாடித்தேடினான்	7
புது	பெயல்	புதுப்பெயல்	புதுப்பெயல்	7
பூ	அழகு	பூவழகு	பூவழகு	3
பூ	கொடி	பூங்கொடி / பூக்கொடி	பூங்கொடி / பூக்கொடி	7, 5
பூ	பாவாய்	பூம்பாவாய் / பூப்பாவாய்	பூம்பாவாய் / பூப்பாவாய்	7, 5
பூ	தளிர்	பூந்தளிர் / பூத்தளிர்	பூந்தளிர் / பூத்தளிர்	7, 5
பூ	தோட்டம்	பூந்தோட்டம் / பூத்தோட்டம்	பூந்தோட்டம் / பூத்தோட்டம்	7, 5
பூ	கூடை	பூங்கூடை / பூக்கூடை	பூங்கூடை / பூக்கூடை	7, 5
பூ	கொத்து	பூங்கொத்து / பூக்கொத்து	பூங்கொத்து / பூக்கொத்து	7, 5
பூ	செடி	பூஞ்செடி / பூச்செடி	பூஞ்செடி / பூச்செடி	7, 5
பூ	சோலை	பூஞ்சோலை / பூச்சோலை	பூஞ்சோலை / பூச்சோலை	7, 5
பூ	உலகில்	பூவுலகில்	பூவுலகில்	3
பூ	உலகம்	பூவுலகம்	பூவுலகம்	3
பூ	ஆல்	பூவால்	பூவால்	3
பொருள்	இயல்	பொருளியல்	பொருளியல்	1
போன	இடம்	போனவிடம்	போனவிடம்	3
மகடுளை	பெரியவள்	மகடுளைப்பெரியவள்	மகடுளைப்பெரியவள்	7
மணி	அழகு	மணியழகு	மணியழகு	3
மணி	அடித்தது	மணியடித்தது	மணியடித்தது	3
மலை	பழம்	மலைப்பழம்	மலைப்பழம்	7
மா	இலை	மாவிலை	மாவிலை	3
மாசு	அற்றார்	மாசற்றார்	மாசற்றார்	1,4
முரடு	காளை	முரட்டுக்காளை	முரட்டுக்காளை	6,7
மேற்கு	காற்று	மேல்காற்று / மேன்காற்று	மேல்காற்று / மேன்காற்று	8
மேற்கு	ஊர்	மேலூர்	மேலூர்	1,8
வடக்கு	கிழக்கு	வடகிழக்கு	வடகிழக்கு	8
வடக்கு	மேற்கு	வடமேற்கு	வடமேற்கு	8
வடக்கு	வேங்கடம்	வடவேங்கடம்	வடவேங்கடம்	8
வயிறு	பசி	வயிற்றுப்பசி	வயிற்றுப்பசி	6,7
வழி	தடம்	வழித்தடம்	வழித்தடம்	7
வாழை	தோப்பு	வாழைத்தோப்பு	வாழைத்தோப்பு	7
வானம்	எல்லாம்	வானமெல்லாம்	வானமெல்லாம்	1
வீடு	தோட்டம்	வீட்டுத்தோட்டம்	வீட்டுத்தோட்டம்	6,7
வேல்	எறிந்தான்	வேலெறிந்தான்	வேலெறிந்தான்	1



## 5. CONCLUSION

A rule-based computer-assisted language learning system for grammar Punarial is proposed. In the book Nannool, more than 100 Punarchi rules are mentioned. The challenge is the requirement of linguistics knowledge of grammar Punarial rules. So, the proposed system is implemented with only eight such grammar rules. The developed system is tested with test samples. It is found that the accuracy is at its highest potential.

## REFERENCES

- [1] Rajendran Sankaravelayuthan, "Computational Linguistics and Technological Development of Tamil", Language in India, Vol. 19 Issue 11, Nov 2019
- [2] Nannool from the website Project Madurai : [https://www.projectmadurai.org/pm\\_etexts/utf8/pmuni0147.html](https://www.projectmadurai.org/pm_etexts/utf8/pmuni0147.html)
- [3] The book “இலக்கணச் சுருக்கம்” by Thiru. ஆறுமுக நாவலர்
- [4] Jayan V, Bhadrar V, “Transliteration from English to Indian Languages based on AnglaMT Machine Translation Perspective”, International Journal of Advanced Research Trends in Engineering and Technology, Vol. 4, Special Issue 6, April 2017



## An Exotic Natural Language Processing Technique for Ancient Tamil Inscriptions: A Linguistic Approach

Ezhilarasi S<sup>1</sup>, UmaMaheswari P<sup>2</sup>

<sup>1,2</sup> Department of Computer Science and Engineering, CEG Campus, Anna University, Chennai, TamilNadu, India

---

### ABSTRACT

---

#### Keywords:

Natural Language Processing  
Morphology  
Ancient inscriptions  
Tamil language  
Grammatical rules  
Linguistic categorizer

Natural Language Processing interlinks a computer to the language. NLP facilitates the language for learning and employing the related applications. The NLP tools improve grammar instruction and understanding of the language. Tamil language of Ancient period differs from the modern era language words. As Tamil language is Agglutinative in nature, ancient morphology is complex. An ancient Tamil phrase may also have a base word with the additional affixes that are connected together. The problems include framing rules for adding and removing suffixes and morphological replacement, complex verb and noun patterns, word beginnings and endings, and poetic scripts are more complicated. In this work, we discuss the strategies for grammar analysis, character analysis, analysing morphology, word level analysing, and verbs, adjectives, adverbs, position suffixes for the analysis at the sentence level that were created using technology based on NLP for Ancient Tamil language texts extracted from inscriptions. The disposition of lexemes with grammatical rules also discussed. The proposed NLP technique of Linguistic categorizer classifies the words into grammatical term categories when analysing text and rules for dissociating of linguistic words in a script. The taxonomy structure of ancient words along with syntactic and lexical analysis is formed. These are quite beneficial for the inscription learners and researchers to comprehend the framework of ancient Tamil language's character, word, and phrase structure in an unconventional manner.

---

#### Corresponding Author:

Ezhilarasi.S,  
Department of Computer Science and Engineering,  
CEG Campus, Anna University, Chennai, TamilNadu, India  
Email: ezhilarasise@gmail.com

---

### 1. INTRODUCTION

The state of the art Natural Language Processing is exited to discover how to use NLP methods for the low-resource language. Naturally, it is rewarding to preserve the oldest language from classical antiquity in the digital sphere. Even for English, scaling the approaches is a challenging task. It is considerably more important to use conventional techniques for morphological analysis given Tamil's complexity. Tamil language are a long way from becoming useful because of the lack of data and need for far greater language resources than English to handle this complexity. In relation to language, morphology is the study of syntactic organization and grammar. We confine ourselves to the examination of the effects of suffixes for the sake of this discussion. We focus solely on the investigation of how suffixes affect the word structure.

That two words are not simply concatenated is the initial idea. The majority of Indian languages have this characteristic, known as sandhi, where the concatenation is indicated by a minor change in writing.

The morphology of classical Tamil is extremely rich and agglutinative. Numerous natural language processing systems have been developed with success using the protocol-based methodology. It can be difficult to deconstruct these combined words in the field of natural language processing. When analysis is performed on a more oldaged and morphologically rich material, such as Tamil inscriptions, the complexity significantly rises. The various difficulties we encountered while attempting to examine the syntactic and semantic features is listed in this work. We also point out the various adjustments that were made to overcome some of the challenges and potential adjustments that might be made to increase the accuracy in the target domain. The task of automatically capturing the agglutinative structure of old Tamil words is difficult. The current work focuses on the design and creation of natural language processing methods for historic Tamil and presents its findings in the conclusion.

### 1.1. Related Works

Selvaraj et al [2] proposed a spellchecker using NLP for Tamil language. The work used spelling correction algorithm using minimum edit distance and suggests a probabilistic and most likely words. Preprocessing is done. Matrix formation algorithm and N gram algorithm is used and for some characters LSTM is tried out.. The dataset used is news corpus Tamil text-7M.txt with comprehensive dictionary. Kengatharaiyer sarveswaran [1] proposed a morphological parser known as Thamizhimorph parser for Tamil. It is developed by FST models with Foma C library and compiler for developing FST. The data set used is Tamil text book and Tamil Universal Dependency tree bank version 2.5 which has 3300 lemmata of Tamil verbs, 80000 Tamil nouns, and 3023 words were analysed from UD Tamil tree bank. This work developed a meta morphological rules, verbs, nouns, orthographical rules with 84.7% accuracy. No benchmark resources exist for evaluation of NLP Tamil and POS tagger is not merged for verbal constructions and complex verbal inflections.

M.Rajasekar et.al [4] proposed a factored method as smart morphological analyser. The work discussed about Tamil language machine translation, provided ML linguistics and considerable ideas about Tamil syntax, POS tagging, splitting affixes, finding noun and verb, identification of morphology, linguistic syntax. Some of the grammatical forms of the word and morphological noun term classification were done. It is tested with 1382 noun words and 1098 verb words. The work is just an outline and not deeply discussed.

J.Benita Antony et al [5] proposed a work on morphological analysis of Tamil Siddha biomedical texts. This deals with orthographical features of language. In this work there is an issue with ambiguity for correct tense and here dictionary is not used.

R.Akilan et al [6] proposed a work on morphological analysis for classical Tamil. The work states a rule based approach. The dataset used is classical Tamil tests of 3257 words with an XML file. This involves preparing a tagsets, rules, used dictionary issues with same syntactic structure and form. The work shows 83% accuracy. Lack of sentence sequence analysis and not able to analyse the texts based on semantic rules. S. Lushanthan et al as in [7] proposed a morphological analysis for Tamil language. Here morphological analyser generates word forms at a particular context and surface forms. It describes only nouns and verbs and uses orthographic rules of Tamil. Totally 2000 nouns and 96 verbs, 80000 noun surface forms and 17166 verb surface forms were processed. The data set used is parliamentary notes that handle two-level morphology with Xerox tool kit. It is FST based and transliteration scheme is defined. This work has not used negative forms, does not consider spoken Tamil and does not focus on sandhi rules and that is not modelled.

Rajendran S et al [8] have proposed morphological analysis for Tamil. The work involves training kernel methods and sequence labelling. It works with morphological features and non linear relationships. The flow comprises of preprocessing, morpheme segmentation, finding morphemes and grammatical elements identification. It used 130000 verbs and 70000 nouns and tested with Amrita POS tagged corpus. Here morpheme dictionary is not used. Vasu Renganathan [9] proposed a work on English-Tamil MT system. It provides syntactic structures from syntactic parser, basic transfer rules from English, MT framework with morphological tagger. It identifies suffixes for literary words in Tamil where it is an interactive approach.



Dhanalakshmi et al.[10] proposed a grammar learning and teaching NLP tools for Tamil. The work proposed learning letters, words and sentences. The development of character analyser, morphological analyser, morphological generator, verb conjugator, POS tagger and dependency parser are done. The dataset used is Dinamani news paper, Yahoo Tamil news, and Tamil short stories with 350000 words. They have developed ML techniques with NLP tools and corpora. B Kumarashanmugam [11] proposed an outline of morphological analyser, tagger, semantic and syntactic rules, transfer rules, sense disambiguation for NLP in Tamil.

## 2. PROPOSED METHODOLOGY

From the literary works it is clear that morphological analysis using NLP techniques for ancient inscription words are not attempted and the research gap is identified. Though the works have proposed a morphological NLP tools for the classical Tamil it is not applicable to ancient Tamil inscriptions. The syntax formation of noun and verb words was not deeply analyzed. But making morphological analysis in ancient Tamil is really difficult. As previously indicated, inflections, derivations, and compounding can all be used to produce unrecognizably long words. Inflections can also vary greatly based on factors like gender, tense, case, and others. There has been consistent development on Tamil morphological analyzers for researching literature. putting the lexicon's terms into a variety of categories (such as nouns, person names, location names, postpositions, pronouns, quantifiers, adjectives, verbs, adverb, conjunctions, grantha borrowed words, brahmi texts, Sanskrit rooted nouns, interjection, language names and so on). The development of morphological process rules is based on these manual annotations.

The proposed exotic NLP technique is a diagnostic module that can be the solution to the investigations and classification of ancient Tamil inscription words. The proposed method is a novel Linguistic Categorizer that gives the categorical terms, sub categories, case suffixes, noun suffixes, verbal suffixes, sandhis and the corresponding descriptions. The algorithm 1 describes the flow which is discussed in 3.2 in detail.

<b>Algorithm 1: Linguistic Categorizer</b>		
<b>input</b>	: Ancient script words $Sw$	//
<b>output:</b>	Linguistic category $C_i$ and base words $Bw$ //	
1	<b>foreach</b> $entry \in Sw$ <b>do</b>	//
2	If Noun phrase:	//
	If OBL and PL exists,	
	SUF and SAN exists,	
	PP and CL exists	
	Analyze dictionary and split	
	Return Noun base word $Bwn$	
3	End if	// End if
4	Else If Verb phrase:	//
	If PR SAN exists,	
	AUX1.....AUX4 exists,	
	TEN GEN CASE exists	
	Analyze dictionary and split	
	Return Verb base word $Bwv$	
5	End if	//
6	<b>End for</b>	//

## 3. RESEARCH METHOD

### 3.1. Characteristics of Tamil

There are several ways that Tamil is different from morphology rich languages. Tamil sentences are redundant in their order. Any combination of the subject, verb, and object is grammatically acceptable. Tamil is an agglutinative language, meaning that suffixes are significant. Suffixes can be included into words in a variety of ways to help them develop naturally and become more meaningful. As a result, there are practically endless numbers of grammatically sound words that cannot be saved. Minor variations in letter composition can significantly change meaning. From the perspective of statistics, this is

more important. When compared to Indo-Aryan languages, Tamil is incredibly context-sensitive. For instance, it could be necessary to use various suffixes when rearranging the sentences. We list the many morphological processing types. Suffix addition with category preservation in inflection (such as noun, verb etc), while keeping the noun's categorization, by adding a suffix (plural noun to singular noun). Derivation is suffix addition that changes the categorization. Compounding is the process of combining two words to create a new term.

### 3.2. Design and Procedure of Proposed NLP technique

The ancient inscription script words are initially pre-processed by performing natural language processing techniques of word tokenization. As any script sentence before morphological analysis undergoes tokenization followed by normalization, punctuations removal. There are few punctuations even in inscription scripts such as ‘-’, ‘,’ and also few symbols and designs which are removed in this module. The linguistic categorizer does the morphological analysis by checking the phrase in noun or verb handling module. According to the inflectional suffixes lemmatization is done by verifying with the case dictionaries. The classification is done where the final noun or verb base is formed and will also be updated in monolingual corpus which can be further used for processing. The base words are clearly split from the rest of the case morphemes. The flow of morphological analysis of Linguistic Categorizer can be shown in figure 1. Case and suffix morphology of Tamil is given in Table 1.

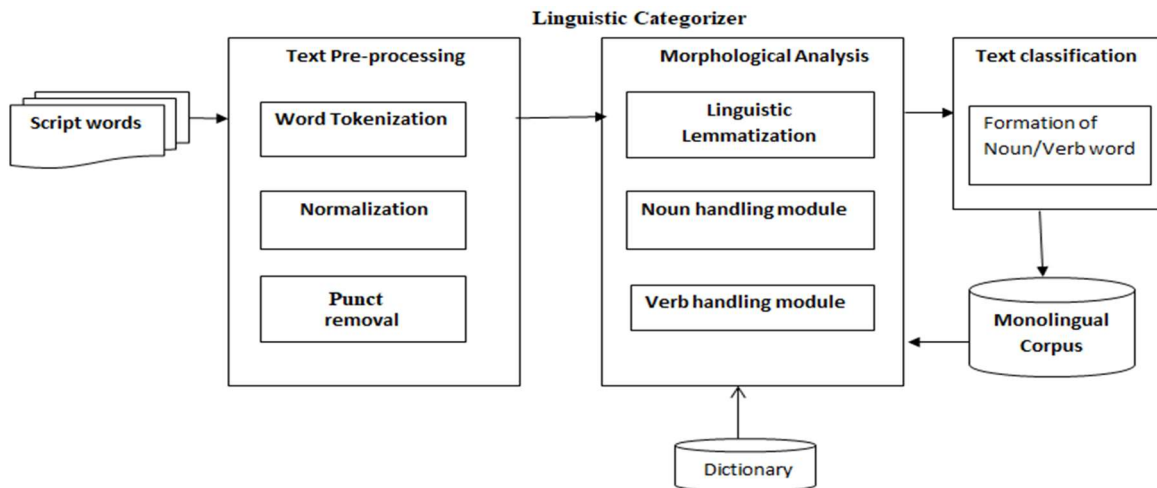


Figure 1. Flow of Morphological analysis of Linguistic categorizer

Table 1. Tamil Suffix morphology

Cases	Suffixes
Accusative	ஏ, ஐ
Instrumental	ஆல்
Sociative	ஓடு, உடன்
Dative	க்கு,ற்கு
Ablative	இருந்து, விட்டு, கொண்டு
Genitive	இன், அது, உடைய
Locative	இல்,உள், இடம்
Vocative	ஆலே
Benefactive	க்காக,ற்காக
Clitics	உம், ஓ, தான்
Selective	ஆவது
Interrogative	ஆ
Optative	வைக்க
Echo	பட,கட
Complement	ஆக
Plural	ஆர், கள், ஈர், ஓம், ஈர்கள், ஆர்கள்,ங்கள்
Singular	ஏன், ஆய், ஆன், ஆள், அது
Causative	இ
Negative	ஆது, ஆதே
Present	கிரு,கின்ற, ஆநின்று
Past	த்,ந்,இன்,ற்
Future	ப்,வ்
Sandhi	க்,ம்,ச்,த்,ந்,ய்,ப்,வ்,ன்

### 3.2.1 Noun handling module

The words are parsed by Noun handling module according to the syntactic structure which is integrated module with oblique (OBL) analysis, plural (PL) analysis, case analysis, sandhi (SAN) analysis, clitics (CL) analysis as described in algorithm 1. The words undergo each phase and are split into separate morphemes. Each analysis involves the corresponding oblique, plural, case, sandhi, clitics dictionaries. This is given in the equation 1.

$$\text{Noun phrase} = [\text{Noun}] + [\text{oblique}] + [\text{plural suffix}] + [\text{case suffix}] + [\text{PP}] + [\text{clitics}] \quad (1)$$

### 3.2.2 Verb handling module

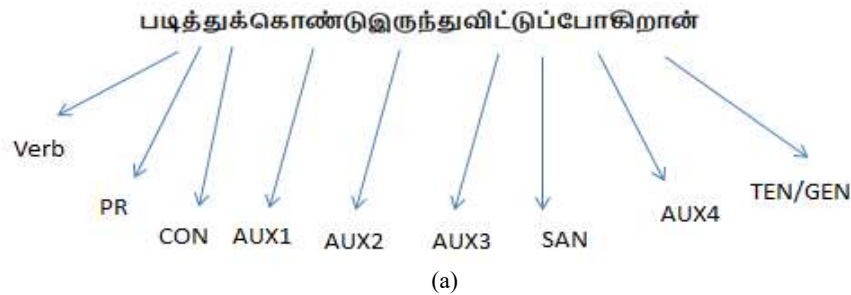
The words are parsed by verb handling module according to the syntactic structure that is integrated with participle (PR) analysis, sandhi (SAN) analysis, auxiliary verb (AUX) analysis, tense (TEN) analysis and gender (GEN) analysis as described in algorithm 1. The word goes through a pipeline of each step to get split. Auxiliary verb is limited to maximum of four in one phrase. Tense defines present, past and future. Each category involves the corresponding participle, sandhi, auxiliary verb, tense and gender dictionaries associated with it. This is given in the equation 2.

$$\text{Verb phrase} = [\text{Verb}] + [\text{verb participle}] + [\text{conjunction}] + [\text{aux1} + \dots + \text{aux4}] + [\text{tense suffix}] + [\text{png}] \quad (2)$$

## 4. RESULTS AND ANALYSIS

### 4.1. Analysis of Test Case Words with Proposed Work

A comprehensive discussion about the analysis of morphology of contemporary Tamil and ancient inscription words from stones with the proposed Linguistic categorizer is experimented with the certain example test cases. For modern Tamil, a verb phrase படித்துக்கொண்டு இருந்துவிட்டுப்போகிறான் is morphologically analyzed as shown in figure 2 (a). Cases such as particle, conjunction, sandhi, tense, gender, and auxiliary verbs are disassociated. A sample stone inscription image is given in figure 2 (b) and few ancient words from the inscription such as ஏழுமஞ்சாடியங்குன்றியாக, ஆடவல்லானென்னுங், ஸ்ரீராஜராஜேஸ்வரமுடையார்க்கு, பாண்டியர்களையும், நாநூற்றுத்தொண்ணூற்றொன்பதின் is analyzed morphologically as shown in figure 2 (c), that involves various case suffixes like sandhi, clitics, plural, benefactive case, genitive case, dative case, ablative case, obliques etc are disassociated. The grammatical categories with the terms of morphology are analyzed using the proposed NLP technique.



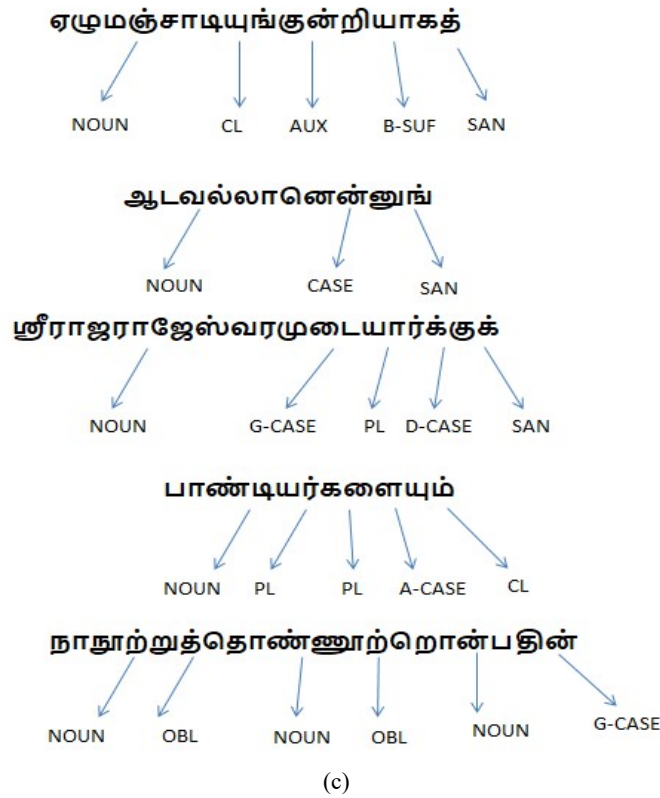


Figure 2. Analysis of Linguistic categorizer (a) Morphology of Verb phrase (contemporary Tamil) (b) Sample stone inscription image (c) Morphology of Noun phrases from inscription words

#### 4.2. Analysis of Existing works

The existing techniques are analyzed with the proposed NLP technique along findings and dataset in Table 2.

Table 2. Analysis of Existing and proposed methods

Work	Technique/Findings	Dataset	Analysis
Smart Morphological analyzer [4]	Morphological factored method for MT	1382 Noun & 1098 verbs	Outline of Morphological forms
Morphological analysis for Tamil [8]	Training kernel methods and sequence labelling	Amrita POS tagged corpus of 130000 verbs 70000 nouns	Morphological features, non linear relationships, doesn't use dictionaries
Challenges in Morphological Analysis of Tamil [5]	Morphological and orthographic features	Tamil Siddha Biomedical texts	Issues in ambiguity in tense and dictionary is not used
English-Tamil Machine Translation system [9]	Morphological tagger MT framework	vbdb.dat, engvb.dat, and lexicon.dat	Parser for literary words of Tamil
Morphological analyzer for classical Tamil [6]	Tagsets and rules	XML file of 3257 words of classical Tamil texts	Lack of sentence sequence analysis, issues with syntax
Morphological analyzer & generator [7]	FST based orthographic rules and transliteration scheme	Parliamentary notes 80000 noun 17166 verb surf forms	Does not cover negative forms
NLP tools for Tamil grammar [10]	Morphological analyzer and generator, NLP tools and corpora, ML techniques	Dinamani news paper, Yahoo Tamil news, Tamil short stories	Grammar teaching tools for learning letters, words and sentence
TamizhiMorph parser [1]	Morphological FST model for Tamil, Orthographical, metamorphological rules	Tamil text book, 3300 verbs, 80000 nouns	POS tagger not merged, not for complex verbal constructions
Spell checker for Tamil using NLP [2]	Spelling correction algorithm, Matrix formation, n gram algorithm, LSTM	News corpus Tamil text 7M.txt	Suggest probabilistic most likely words
Meta-Morph Rules to develop Morphological Analysers [3]	Finite State Morphological Analyzer (FSM)	10000 nouns and 3300 verbs of Tamil	Grammar rules, meta morphological rules, lexical

NLP Technique for Inscription words proposed	Morphological Tamil Linguistic Categorizer	Ancient Tamil Stone inscription words	Morphological grammar categories and case suffixes
---	---	--	---

### 4.3. Evaluation Metrics

For the total vocabulary V or corpus C word level precision, recall and F-measure can be computed for evaluation. Precision is given as the number of correctly predicted morphemes to total number of morphemes predicted  $W_m$ . Recall is also related to measure the accuracy and states that number of correctly predicted morphemes to the total number of actual morphemes to be predicted in Vocabulary V or Corpus C. From this the F-score is computed. The metrics are given as in (3), (4) and (5) as follows

$$\text{Precision} = \frac{\text{No.of correctly predicted morphemes}}{\text{Total no of morphemes predicted } W_m} \quad (3)$$

$$\text{Recall} = \frac{\text{No of correctly predicted morphemes}}{\text{Total no of actual morphemes to be predicted in Vocabulary V}} \quad (4)$$

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

## 5. CONCLUSION AND FUTURE ENHANCEMENT

In this work, a Tamil morphological analyzer based on exotic Natural Language Processing technique is described. This research provided evidence of a new approach used for the morphological analyzer for ancient Tamil inscription words. The Morphological Analyzer is the most significant function of any Natural Language Applications, according to the rule-based approach such as syntactic and grammatical categories. The Morphological Analyzer for ancient Tamil grammatical Linguistic categorizer approach has been described in this work; Due to the successful development of these systems, linguistic tools were created that use a grammatically sound approach for case suffixes. These handles noun and verbs splitting perfectly and returning the base words with the use of dictionaries and corpus. These case suffixes approaches based on rules give more accurate results. This work is further under progress on development on POS Tagger and NER tagger for more understanding of inscription words. This work will be applicable for other language inscription words also.

## ACKNOWLEDGEMENTS

The authors are thankful to the Department of Science and Technology (DST), India for funding research work under Science and Heritage Research Initiative scheme and assistance from TamilNadu state department of archeology under which this work is an extension.

## REFERENCES

- [1] Kengatharaiyer Sarveswaran, Gihan Dias, Miriam Butt, "ThamizhiMorph: A morphological parser for the Tamil language", Machine Translation, 35:37–70 <https://doi.org/10.1007/s10590-021-09261-5>, 2021
- [2] Dr.P.A. Selvaraj, Dr.M. Jagadeesan, Dr.M. Harikrishnan, Dr.R. Vijayapriya, Dr.K. Jayasudha, "Survey on Spell Checker for Tamil Language Using Natural Language Processing, Journal Of Pharmaceutical Negative Results", Vol. 13 Special Issue 07, 2022
- [3] Kengatharaiyer Sarveswaran, Gihan Dias, Miriam Butt, "Using Meta-Morph Rules to develop Morphological Analysers: A case study concerning Tamil", International Conference on Finite-State Methods and Natural Language Processing, pages 76–86, 2019.
- [4] M. Rajasekar, N. Rajasekharan Nair, A. Udhayakumar, "Smart Morphological Analyzer for Tamil", BEST: Journal of Recent Trends in Economics, Ancient history & Linguistic Research (BEST: JRTEAHL) Vol. 3, Issue 1, 15-22, 2017.
- [5] J. Betina Antony and G. S. Mahalakshmi, "Challenges in Morphological Analysis of Tamil Biomedical Texts", Indian Journal of Science and Technology, Vol 8(23), DOI: 10.17485/ijst/2015/v8i23/79350, 2015.
- [6] R. Akilan and E. R.Naganathan, "Morphological Analyzer for Classical Tamil Text: A Rule-Based Approach", ARPN Journal of Engineering and Applied Sciences, Vol. 10, No 20, 2015.
- [7] S. Lushanthan, A. R. Weerasinghe, D. L. Herath, "Morphological analyzer and generator for Tamil Language", International Conference on Advances in ICT for Emerging Regions (ICTer) IEEE Xplore: 2015.
- [8] Anand kumar M, Dhanalakshmi V, Rajendran S, Soman K P, "A Novel Approach to Morphological Analysis for Tamil Language", Proceedings of the Second international conference on Data Engineering and Management, DOI:10.1007/978-3-642-27872-3\_37, 2010

- [9] Dr. Vasu Renganathan, “An Interactive Approach to Development of English-Tamil Machine Translation System on the Web”, University of Pennsylvania, Linguistics
- [10] Dhanalakshmi V, Rajendran S, “Natural Language Processing Tools for Tamil Grammar Learning and Teaching”, International Journal of Computer Applications (0975 – 8887) Volume 8– No.14, 2010
- [11] B Kumara Shanmugam, “Natural Language Processing in Tamil”, AU-KBC Research Centre, MIT, Anna University, Chennai, India
- [12] Loganathan Ramasamy Ondřej Bojar Zdeněk Žabokrtský, “Morphological Processing for English-Tamil Statistical Machine Translation”, Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 113–122, 2012
- [13] R.Akilan, E.R.Naganathan, G.Palanirajan, “Morphological Analyzer for Classical Tamil Texts: A Rule-based approach for Case Marker”, IJSET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 5, 2014
- [14] Anand Kumar M, Dhanalakshmi V, Soman K.P, Rajendran S,” A Sequence Labeling Approach to Morphological Analyzer for Tamil Language”, International Journal on Computer Science and Engineering Vol. 02, No. 06, 1944-1951, 2010
- [15] S.Ezhilarasi, P.UmaMaheswari, “Depicting a Neural Model for Lemmatization and POS Tagging of Words from Palaeographic Stone Inscriptions”, 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 1879-1884, IEEE, 2021
- [16] Sebastian Spiegler, Christian Monson, “EMMA: A Novel Evaluation Metric for Morphological Analysis”, Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1029–1037, Beijing, August 2010
- [17] A. G. Ramakrishnan, Lakshmish N Kaushik, Laxmi Narayana. M, “Natural Language Processing for Tamil TTS”, DOI: 10.13140/RG.2.1.4233.0726, 2007
- [18] Agnieszka Dardzinska, “Natural language processing: Word recognition without segmentation”, Journal of the American Society for Information Science and Technology, 2001
- [19] Itunuoluwa Isewon, Jelili Oyelade, Olufunke Oladipupo “Design and Implementation of Text To Speech Conversion for Visually Impaired People”, International Journal of Applied Information Systems (IJ AIS) – ISSN : 2249-0868, Volume 7– No. 2, 2014

## The Efficacy of the Jamboard Virtual Learning Strategy in Overcoming Grammatical Errors in Tamil

### இலக்கணப் பிழைகளின்றி தமிழ் எழுதிட ஜேம்போர்ட் (JAMBOARD) வழி மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறை

புஷ்பராணி சுப்ரமணி செல்வன்  
மலாயாப் பல்கலைக்கழகம், கோலாலம்பூர்

#### ஆய்வுச் சுருக்கம்

#### Keywords:

- A ஜேம்போர்ட்
- B இலக்கணப் பிழைகள்
- C மெய்நிகர் கற்றல் கற்பித்தல்
- D மொழி
- E இலக்கணம்

மொழியின் அடிப்படைக் கூறுகளில் இலக்கணக் கற்றல் முதன்மையாகத் திகழ்கின்றது. முறையான இலக்கணமே ஒரு மொழியைச் சரியாகப் பேசிட, எழுதிட வழிவகுக்கும். தமிழ் மொழி கற்றல் கற்பித்தல்வழி, தொடக்கப்பள்ளி இறுதியில் மாணவர்கள் அடைய வேண்டிய முக்கியக் கூறுகளில் 'இலக்கண அறிவைப் பெற்றுச் சரியாகப் பயன்படுத்துதல்' என்பது மலேசிய தமிழ்ப்பிரிவு கலைத்திட்டத்தின் முதன்மை நோக்கங்களுள் ஒன்றாகும். ஆனால், தமிழ் மொழியினைத் தாய்மொழியாகக் கொண்டு தமிழ்க் கல்வி பயிலும் மாணவர்களிடையே இலக்கணப் பிழைகள் இன்றி தமிழ் எழுதிடும் ஆற்றல் மிகக் குறைவாகவே உள்ளது. இச்சிக்கலைக் களைய மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறை கையாளப்பட்டுத் தீர்வுக்கான இவ்வாய்வு மேற்கொள்ளப்பட்டது. இவ்வாய்வுக்கு ஜேம்போர்ட் வழி மெய்நிகர் கற்றல் கற்பித்தல் முறை அணுகப்பட்டுள்ளது. இவ்வாய்வில் Online Collaborative Learning Theory (OCL) கோட்பாடு பயன்படுத்தப்பட்டுக் கற்றல் கற்பித்தல் முறை மேற்கொள்ளப்பட்டது. இவ்வாய்விற்காக மலேசியக் கல்வி அமைச்சின் ஆரம்பக் கல்விக்கான பாடநூல்கள் படிநிலை 1 மற்றும் படிநிலை 2 பயன்படுத்தப்பட்டு அதிலுள்ள இலக்கணப் பாடக் கல்வி மட்டுமே ஆய்வுக்குப் பயன்படுத்தப்பட்டுள்ளன. ஆய்விற்கு உட்படுத்தப்பட்ட மாணவர்களுக்கு மெய்நிகர் கற்றல் கற்பித்தல் 2 வாரங்கள் தொடர்ந்து ஜேம்போர்ட் வழி இருவழி தொடர்பில் போதிக்கப்பட்டது. பின்னர், மாணவர்களின் கட்டுரைகளில் காணப்படும் எழுத்துப்பிழைகள் மெய்நிகர் கற்றல் கற்பித்தல் ஆய்வுக்கு முன்னும் பின்னும் ஆராயப்பட்டுக் கலந்துரையாடப்பட்டுள்ளது. நான்காம் ஆண்டு பயிலும் 10 மாணவர்களிடம் ஆய்வுக்கு முன் 5 கட்டுரைகளும் மெய்நிகர் கற்றல் கற்பித்தலில் 10 கட்டுரைகளும் வழங்கப்பட்டுத் திருத்தப்பட்டன. அக்கட்டுரைகளில் காணப்படும் எழுத்துப்பிழைகளின் சராசரி கண்டறியப்பட்டு ஆராய்விற்கு உட்படுத்தப்பட்டன. மெய்நிகர் கற்றலுக்குப் பின் எழுத்துப்பிழைகளின் சராசரி குறைந்து, இலக்கணப்பிழைகளின்றி தமிழ் எழுதிடும் ஆற்றல் மாணவர்களிடையே காணப்பட்டது..

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

#### 1.0 அறிமுகம்

மலேசியக் கல்வி அமைச்சின் தமிழ்ப்பிரிவு கலைத்திட்டக் குவிவு ஆரம்பப் பள்ளி மாணவர்கள் 'இலக்கணப் பிழையின்றி நல்ல மொழியைப் பயன்படுத்தித் தரமான எழுத்துப் படிவங்களைப் படைக்கும்' நோக்கில் வரையறுக்கப்பட்டுள்ளது. இக்கொள்கையின் அடிப்படையில் இலக்கணப் பிழைகளின்றி தமிழ் எழுதும் ஆற்றல் மாணவர்களுக்குச் சிறு வயது முதற்கொண்டே கற்றுத் தரப்பட வேண்டிய ஒரு முக்கியக் கூறாக அமைகின்றது. இலக்கணப் பிழைகளின்றி எழுதும் ஆர்வத்தைத் தூண்ட மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறை சிறந்த களமாகப் பயன்படுகின்றது. மெய்நிகர் கற்றல் கற்பித்தல் மாணவர்களிடையே சுயக் கற்றலை மேலோங்கச் செய்வதோடு, தன்னார்வத்தைத் தூண்டும் வகையில் பல உத்திகள் கையாளப்பட்டு இலக்கணப் பிழைகள் இன்றி தமிழ் எழுத வழிவகுக்கும் நோக்கில் இவ்வாய்வு மேற்கொள்ளப்பட்டது. மெய்நிகர் கற்றல் கற்பித்தல் முறை கல்வித் துறையில் பரவலாகக் காணப்படும் தற்கால போதனா முறை ஆகும். பரவலாகப் பயன்படுத்தப்பட்டு வரும் இணைய உலகில் பல்வேறு மாற்றங்களை உள்ளடக்கிய Web 2.0 தொழில்நுட்பம் கல்வித் துறைக்குப் பெரிதும் பங்காற்றி வருகின்றது. (Al-Kathiri, 2014) இத்தொழில்நுட்பத்தில் ஒன்றான ஊடாடும் வெண்பலகை ஜேம்போர்ட் எனப்படுவது இலவசமாகவும் பாதுகாப்பாகவும் பயன்படுத்தும் தளமாகும். இத்தளத்தின் வாயிலாக ஆசிரியர்கள் மாணவர்கள் என இரு தரப்பினரும் கற்றல் கற்பித்தல் நடவடிக்கையை ஒரே சமயத்தில் மேற்கொள்ளலாம். வகுப்பில் குழு

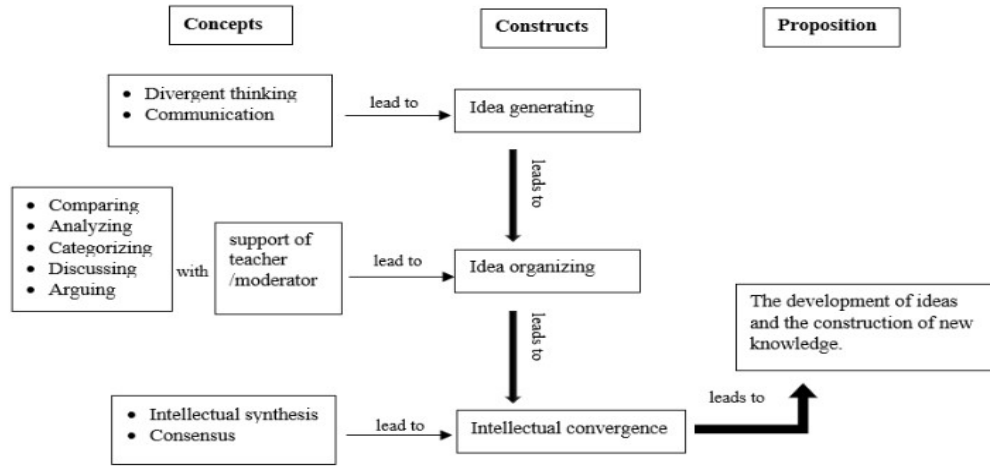


கலந்துரையாடலில் கருத்துகளைப் பகிருகையில் நேரப்பற்றாக்குறை, கருத்துகளைத் தெரிவிப்பதில் அச்சம், கருத்துகளை ஒரே நேரத்தில் தெரிவிக்கையில் சில நிராகரிப்புகள், அனைத்துக் கருத்துகளையும் எடுத்துக் கொள்வதில் சிக்கல், கருத்து பதிவுகளில் அதிருப்தி எனப் பல்வேறு சிக்கல்களை மாணவர்கள் எதிர்கொள்கின்றனர். ஆனால், மெய்நிகர் கற்றல் இவ்வனைத்துச் சிக்கல்களுக்கும் தீர்வாக அமைக்கின்றது. மாணவர்கள் தங்கள் கருத்துகளைத் தெரியமாகவும், ஒரே சமயத்திலும் தெரிவிக்க இயலுக்கின்றது. அதோடு, பதிவிடும் அனைத்துக் கருத்துகளும் ஊடாடும் வெண்பலகையில் இடம்பெறுவதால், மாணவர்களின் அனைத்து கருத்துகளும் சக நண்பர்களால் வாசித்து அங்கீகரிக்கப்படுகிறது. இது மாணவர்களிடையே தங்கள் கருத்துகளைத் தன்னம்பிக்கையுடன் பதிவிட உந்துகோளாக அமைகின்றது. அதோடு, சமூக வலைத்தளங்களின் பங்கு அனைத்துத் துறைகளிலும் பெரும் அளவில் காணப்படுவதனால் அதனைக் கல்வியோடு தொடர்புக் கொண்டு மாணவர்கள் பயன்பெறும் வகையில் மாணவர்களின் எதிர்பார்ப்புகளைப் பூர்த்தி செய்வது கற்றல் கற்பித்தல் முறையில் அவசியமான ஒன்றாகும். (Foster & Neal, 2012) இன்றைய கல்வியாளர்கள் சமூக வலைத்தளங்களாகிய முகநூல், டிவிட்டர், யூடியூப், வலைப்பூக்கள், தெலிகிராம் போன்ற சமூக வலைத்தளங்களைக் கல்வியோடு இணைத்துவிட்டனர். (Balakrishnan & Gan, 2016) ஆகவே, இன்றைய காலத்தில் உடனணைந்து பணியாற்றுவதல், நுண்ணாய்வுச் சிந்தனை, படைப்பாக்கம், கருத்துப்பரிமாற்றம் ஆகியவனவற்றிற்கு எதுவாக ஜேம்போர்ட் தளம் பயன்வழங்கி மாணவர்கள் இலக்கணப்பிழைகளை இன்றி தமிழ் எழுதிட உறுதுணையாகப் பயன் வழங்கியுள்ளது.

## 2.0 செயல்முறை

### 2.1 Online Collaborative Learning Theory (OCL)

கற்றல் கற்பித்தலில் அணுகுமுறைகளைக் கல்வியல் உலகம் நான்கு பெரும் குழுமத்தில் இணைத்துள்ளன. அவை Behaviorism, Cognitivism, Humanism மற்றும் Constructivism ஆகும். அவற்றுள் புதிய அணுகுமுறையான OCL-பரவலாகக் கல்வித் துறையில் பயன்படுத்தப்பட்டு வருகிறது. இந்தக் கோட்பாடு மாணவர்கள் கற்றல் சவாலை உடனணைந்து களைவதற்கு முக்கியத்துவம் வழங்குகின்றது. இந்த அணுகுமுறை 2012 ஆம் ஆண்டு லிண்டா ஹரசிம் என்பவரால் அறிமுகப்படுத்தப்பட்டது. இந்த அணுகுமுறை முற்றிலும் கணினி வழி தொடர்பு மற்றும் சமூக தொடர்பு ஒன்றிணைந்த கற்றல் கற்பித்தலை வழியுறுத்தும் கோட்பாடாகும். இக்கோட்பாடு உடனணைந்து பணியாற்றுகையில் மூன்று வகையாக அறிவு மேம்படுகிறது என்பதனை விவரிக்கின்றது. அவையானவை, தகவலை உருவாக்குதல், தகவலை ஒருங்கிணைத்தல் மற்றும் மூன்றாவது கட்டமாக முன்னறிவைத் திரட்டிப் புதிய தகவலைப் படைத்தல் ஆகியன ஆகும். இக்கோட்பாட்டின் நன்மையானது கூட்டிணைக் கற்றல் நடவடிக்கையோடு, நுண்ணாய்வுச் சிந்தனை ஆற்றல், பகுப்பாய்வு, மதிப்பிடல் ஆகிய உயர்நிலை சிந்தனையாற்றலும் வளர உதவுகின்றது.



படம் 1 : Online Collaborative Learning Theory (OCL) கட்டமைப்பு

மூலம் : லிண்டா ஹரசிம் (2012)

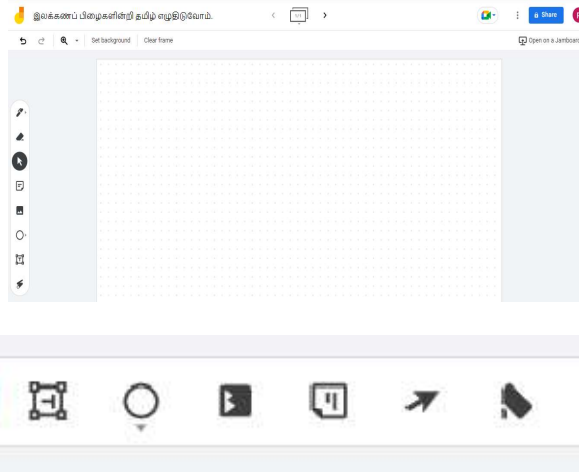
இக்கோட்பாட்டின் அடிப்படையில், ஆய்வின் முதற்கட்டமாக ஆரம்பப்பள்ளிகளுக்கென வரையறுக்கப்பட்ட இலக்கணக் கல்வி மாணவர்களுக்கு ஊடாடும் வெண்பலகை ஜேம்போர்ட் வழியாகக் கற்றுத் தரப்பட்டது. தகவல்களைப் பெற்றபின், மாணவர்கள் வழங்கப்பட்ட தலைப்பிற்கேற்ப கருத்துகளைத் திரட்டினர். பிறகு, திரட்டிய தகவல்களை முதன்மை, துணைக் கருத்து, விளக்கம் ஆகியவற்றிற்கு ஏற்ப ஒருங்கிணைத்தனர். ஒருங்கிணைத்த தகவல்களைக் கட்டுரை வடிவில் ஊடாடும் வெண்பலகை ஜேம்போர்ட்டில் படைத்தனர். பதிவேற்றம் செய்யப்பட்ட கட்டுரைகளைக் குழு நண்பர்கள் வாசித்துக் கட்டுரைகளில் காணப்படும் எழுத்துப்பிழைகளைக் களைய உதவினர். வண்ணக்குறிப்பு அட்டைகளில் தாங்கள் அடையாளங்கண்ட எழுத்துப்பிழைகளைக் குறிப்பிட்டனர். இதன் மூலம் குழு நண்பர்கள் செய்யும் திருத்தங்களைக் களைந்து அடுத்த கட்டுரையில் அமல்படுத்தி உடனுழைப்பு/கூட்டிணை வகையில் கட்டுரைகளில் காணப்படும் எழுத்துப்பிழைகளைக் களைந்தனர். இவ்வாறாகக் குழு நண்பர்களின் பகிர்வுகளால் கட்டுரைகளில் காணப்படும் எழுத்துப் பிழைகளைக் களைந்து குறைவாகவோ/ எழுத்துப்பிழைகள் இன்றியோ மாணவர்களால் கட்டுரைகளைப் படைக்க இயன்றது. OCL கட்டமைப்பு மாணவர்களிடையே கருத்துப்பரிமாற்றம் வாயிலாக அறிவுசார் ஒன்றிணைதலுக்கு உதவியாக அமைந்தது.

## 2.2 வடிவமைப்பு

இவ்வாய்விற்ரு நான்காம் ஆண்டு மாணவர்களில் தமிழ் மொழி எழுத்து தர மதிப்பீட்டில் 4 மற்றும் 5 தர அடைவு நிலையை அடைந்த 10 மாணவர்கள் தேர்ந்தெடுக்கப்பட்டனர். இவர்கள் தர மதிப்பீட்டின் அடிப்படையில் இரு குழுக்களாகப் பிரிக்கப்பட்டனர். முதல் குழுவின் 4 தர அடைவு நிலையை அடைந்த மாணவர்கள். இவர்கள் தர திட்ட மதிப்பீட்டின் அடிப்படையில் நல்ல மொழியறிவும் மொழியாற்றலும் கொண்டவர்கள். இம்மாணவர்களுக்கு ஆய்வுச் சிந்தனையோடு எண்ணங்களையும் கருத்துகளையும் ஏரணச் சிந்தனையோடு வெளிபடுத்தும் திறனும் உள்ளது. இம்மாணவர்கள் ஓரளவு சுயமாகக் கற்கும் திறன் கொண்டவர்கள். இரண்டாவது குழு தர அடைவு 5 நிலையில் உள்ள மாணவர்களாவர். இம்மாணவர்கள் சிறந்த மொழியறிவும் மொழியாற்றலும் கொண்டவர்கள். இம்மாணவர்களுக்கு ஆய்வு, ஆக்கச் சிந்தனையோடு ஏரணமான எண்ணங்களையும் கருத்துகளையும் தெளிவாகவும் விளக்கமாகவும் விளையன்மிகக் வகையில் வெளிபடுத்தும் திறனும் உள்ளது. இந்த மாணவர்கள் சுயமாகக் கற்கும் திறன் கொண்டவர்கள். இம்மாணவர்களுக்கு 2 வாரங்களுக்கு மெய்நிகர் கற்றல் கற்பித்தல் முறை போதிக்கப்பட்டது. இம்மாணவர்களுக்கு இணையம் வழி ஜேம்போர்டைப் பயன்படுத்தி கல்வி கற்கும் முறை முதற்கட்டமாகப் போதிக்கப்பட்டது. அடுத்ததாக, நான்காம் ஆண்டு இறுதியில் மாணவர்கள் அடைய வேண்டிய இலக்கணக் கூறுகள் போதிக்கப்பட்டன. பின்னர், குழுவில் இடம்பெறும் மாணவர்களுக்கான தனித்தனி ஊடாடும் வெண்பலகையில் வெவ்வேறு வண்ணங்களில் பெயர் இணைக்கப்பட்டு அதற்கான இணைப்புப் புலனம் வழி பகிரப்பட்டது. மாணவர்களுக்குத் தினம் ஒரு கட்டுரை தலைப்பு என மொத்தம் 10 நாட்களுக்குப் பத்து கட்டுரைகள் வழங்கப்பட்டன. மாணவர்கள் கட்டுரையை எழுதி அதனை ஜேம்போர்டில் அவர்களுக்கான திரையில் இணைத்தனர். அக்கட்டுரைகளைக் குழு உறுப்பினர்கள் வாசித்து அதில் காணப்படும் இலக்கணப் பிழைகளை அடையாளங்கண்டு அவர்களுக்கென வழங்கப்பட்ட வர்ணத்தின் அடிப்படையில் குறிப்பு அட்டையில் தட்டச்சுச் செய்து பதிவேற்றம் செய்தனர். நான்காம் ஆண்டில் அடைய வேண்டிய இலக்கணத் திறன்களில் சொல்லியல், எழுத்தியல், ஒற்றுப்பிழைகள் அதிகமாக மாணவர்கள் கட்டுரைகளில் காணப்பட்டதால், இம்முன்றினைக் குறைப்பது மட்டுமே நோக்கமாகக் கொள்ளப்பட்டு, நிறுத்தற்குறிகள் பார்க்கப்படவில்லை.

## 2.3 ஜேம்போர்ட்

ஜேம்போர்ட் எனப்படுவது இலவசமாகப் பயன்படுத்தக்கூடிய ஒரு தளமாகும். இத்தளத்தில் ஒரே சமயத்தில் பலர் தங்கள் கருத்துகளை இருவழித் தொடர்பில் உடனுக்குடன் பகிர்ந்து கொள்ளலாம். இது கூட்டிணைக் கற்றலுக்குச் சிறந்த தளமாக இயங்குகின்றது. கீழ்க்காணும் படம் 1 ஜேம்போர்ட்டில் கூட்டிணைக் கற்றலுக்குத் தேவையான சிறப்புக் கூறுகளைக் காட்டுகின்றது.



படம் 1 : ஜேம்போர்ட் கூறுகள்

## ஊடாடும் வெண்பலகை திரை

ஊடாடும் வெண்பலகையில் திரைகளை மாணவர்களின் எண்ணிக்கைக்கு ஏற்ப அதிகரித்துக் கொள்ளக்கூடிய வாய்ப்பினை இத்தளம் வழங்கியுள்ளது. ஒவ்வொரு மாணவருக்கும் தனித்தனி திரை எனப் பத்து மாணவர்களுக்குத் தேவையான பத்து தனித்தனி திரைகளை ஒரே சமயத்தில் ஏற்படுத்தி தரும் வசதியைக் கொண்டுள்ளது. இதனால், வெவ்வேறான இணைப்புகளை உருவாக்குவதற்கான நேரத்தை மிச்சப்படுத்துவதோடு, ஒரே பாடத்தை ஒரே இணைப்பில் பலர் கலந்து கொள்ளும்படி செய்யவும் உதவியாக அமைகின்றது.



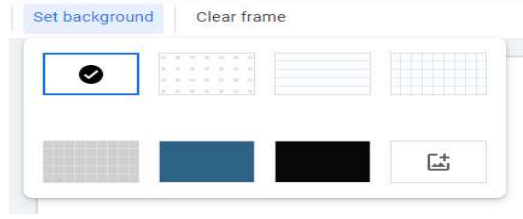
படம் 2: ஊடாடும் வெண்பலகை திரை

## பின்னணி

இத்தளத்தில் மாணவர்கள் தங்களுக்கு ஏற்ற பின்னணியைத் தெரிவு செய்யும் சேவை உண்டு. இப்பின்னணியில் மாணவர்கள் தங்கள் எழுதிய கட்டுரைகளை இணைப்பதன் மூலம் பிறர் எளிதில் கருத்துகளைப் பகிர்ந்து கொள்ளவும் எழுத்துப் பிழைகளைக் கண்டறியவும் உதவியாக அமைந்தது. முழுக்கட்டுரையையும் ஊடாடும் வெண்பலகையில் தட்டச்சுச்

புஷ்பராணி சுப்ரமணி செல்வன், மலாயாப் பல்கலைக்கழகம், கோலாலம்பூர்

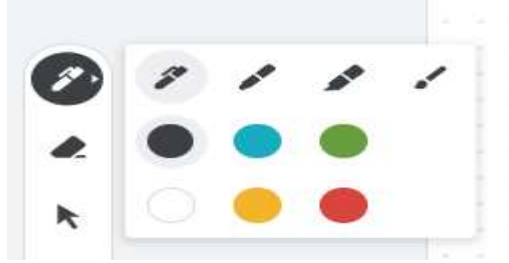
செய்யாமல், தாளில் கட்டுரையை எழுதி அதனைப் புகைப்படம்/ வருடி ஊடாடும் வெண்பலகையில் பின்னணியாக இணைப்பதன் மூலம், மாணவர்களின் நேரத்தை மிச்சப்படுத்தவும் விரைவாக எழுதும் ஆற்றலையும் அதிகரித்தது.



படம் 3: பின்னணி

#### தூரிகைகள்

முன்னிலைப்படுத்துதல், வட்டமிடுதல், குறிப்பிடுதல், அடையாளமிடுதல் போன்ற சேவைகளுக்குத் தூரிகைகள் இணைக்கப்பட்டுள்ளன. இதன்மூலம் மாணவர்கள் சக நண்பர்களின் கட்டுரைகளில் காணப்படும் இலக்கணப் பிழைகளை அடையாளமிட உதவியாக அமைந்தது.



படம் 4: தூரிகைகள்

#### குறிப்பு அட்டை

குறிப்பு அட்டை சேவை பல வண்ணங்களில் இணைக்கப்பட்டுள்ளது. இச்சேவையின் வாயிலாக மாணவர்கள் எளிதில் தங்கள் கருத்துகளை அல்லது குறிப்புகளைத் தட்டச்சு செய்து ஊடாடும் வெண்பலகையில் பொருத்த முடிகின்றது. பல்வேறு வண்ணங்களில் இருப்பதனால் மாணவர்கள் தங்களுக்கென வழங்கப்பட்ட நிறத்தில் மட்டுமே தட்டச்சு செய்வதன் மூலம், குழு உறுப்பினர்களால் குறிப்புகளைப் பதிவிட்டவர் யார் என எளிதில் தெரிந்து கொள்ள முடிந்தது. அதோடு, பல வண்ணங்களில் குறிப்புகள் இடம்பெறும்போது மாணவர்களின் கவனத்தை ஈர்ப்பதோடு, கூட்டிணைக் கற்றலில் கற்கும் ஆர்வத்தையும் மோலோங்கச் செய்தது.



படம் 5: குறிப்பு அட்டை

### 3.0 ஆய்வு முடிவு

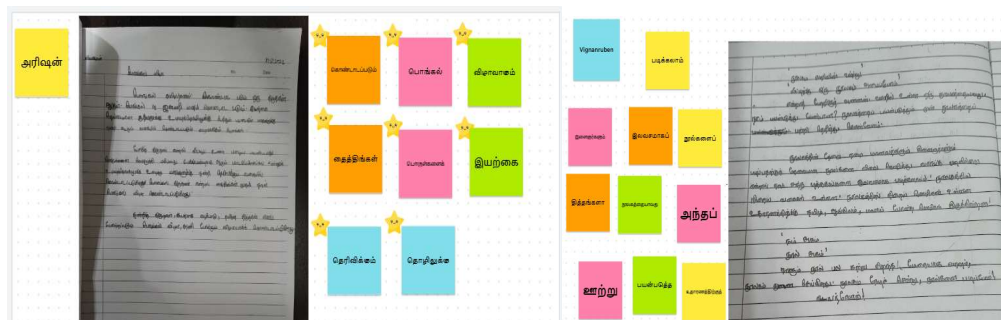
ஜேம்போர்ட் ஊடாடும் வெண்பலகை மாணவர்களின் இலக்கணப் பிழைகளைக் குறைப்பதற்குப் பெரிதும் பங்காற்றியுள்ளது. கீழ்க்காணும் அட்டவணை ! ஆய்விற்கு முன்னும் ஆய்விற்குப் பிறகும் மாணவர்களின் சராசரி இலக்கணப் பிழைகளைக் காட்டுகின்றது.

அட்டவணை 1 : ஆய்வுக்கு முன்னும் ஆய்வுக்குப் பிறகும் மாணவர்களின் சராசரி இலக்கணப் பிழைகள்

மாணவர்கள்	ஆய்வுக்கு முன்			ஆய்வுக்குப் பின்		
	எ.பி	சொ.பி	ஒ.பி	எ.பி	சொ.பி	ஒ.பி
மாணவர் 1	9.2	11.2	8.6	3	2	2.6
மாணவர் 2	7.4	5	12	2	1	0
மாணவர் 3	6	7.2	7.6	2.8	2.2	1.6
மாணவர் 4	6	6.8	9.6	1.2	1	0
மாணவர் 5	7	9.4	11.2	3	1	1.6
மாணவர் 6	10	12.8	9.8	2	1	0
மாணவர் 7	7.2	8.8	7.4	1.6	2.2	1.4
மாணவர் 8	5.2	3.2	6.8	1.4	0.8	3
மாணவர் 9	7.5	3.8	9.6	1.8	0.6	1.2
மாணவர் 10	5.2	6.8	7.5	2.2	0.4	2.6

அட்டவணை I

மெய்நிகர் கற்றல் கற்பித்தலுக்கு முன்னதாக மாணவர்களின் எழுத்துப்பிழைகளின் சராசரி 5< க்கும் மேற்பட்டதாகவும் மெய்நிகர் கற்றலுக்குப் பிறகு சராசரியானது 1-3>க்குள் குறைந்தும் காணப்படுகிறது. சொற்பிழைகளின் சராசரி எண்ணிக்கை மெய்நிகர் கற்றலுக்கு முன்னதாக 3<ற்கும் அதிகமாகவும் மெய்நிகர் கற்றலுக்குப் பின்னர் 0-2.2>க்குள் குறைந்து காணப்படுகிறது. ஒற்றுப்பிழைகளானது 6.8 முதல் 12 வரையிலான சராசரியான மெய்நிகர் கற்றலுக்கு முன்னதாகவும், மெய்நிகர் கற்றலுக்குப் பின்னதாக 0 முதல் 2.6 சராசரிக்குள் குறைந்தும் காணப்படுகிறது. வகுப்பறை குழல் போதனா முறையினைக் காட்டிலும் மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறை மாணவர்களுக்குச் சிறந்ததொரு கற்றல் அனுபவத்தைத் தந்ததோடு, மாணவர்களின் கற்றல் ஆர்வத்தை மேலும் அதிகரித்துள்ளது. தனிப்பர் முறையில் மாணவர்களின் கற்றல் ஐயங்களைக் களைய மேற்கொண்ட அணுகுமுறையும் சக நண்பர்களின் கலந்துரையாடல் மற்றும் பரிந்துரைகள் ஆகியன மாணவர்களின் இலக்கணப் பிழைகளை நிவர்த்தி செய்ய வழிவகுத்தன. இரண்டு வாரங்களுக்குப் பிறகு மாணவர்களின் இலக்கணப் பிழைகளின் எண்ணிக்கை குறைந்த நிலையில் அமைந்ததைக் காண முடிந்தது.



படம் 6 : ஜேம்போர்ட்டில் மாணவர்கள் திருத்தி தட்டச்சு செய்த குறிப்புகள்

## 5.0 முடிவு

மெய்நிகர் கற்றல் கற்பித்தல் மாணவர்களின் இலக்கணப் பிழைகளைக் களைவதில் பெரும் பங்காற்றியுள்ளது. இலக்கணப் பிழைகள் இன்றி எழுத மேற்கொண்ட ஆய்வின் நோக்கம் நிறைவேறியது. இவ்வாய்வின் இறுதியில் ஜேம்போர்ட் வழி கற்றல் கல்வி பெற்ற மாணவர்களின் கட்டுரைகளில் எழுத்துப்பிழைகள் இன்றி/குறைந்தும் எழுதும் ஆற்றல் காணப்பட்டது. அதோடு, மாணவர்களிடையே மேலும் மெய்நிகர் கற்றல் கற்பித்தல் அணுகுமுறையைப் பின்பற்றிக் கல்வி கற்கும் ஆர்வம், தமிழ் மொழியினைப் பிழைகளின்றி எழுதும் ஆற்றலை வளர்த்துள்ளது. மெய்நிகரில் இணைந்து கற்றல் முறை மாணவர்கள் கலந்துரையாடலுக்கும் உடனுக்குடன் பிழைகளைக் களைவதற்கும் பெரும் பங்காற்றியுள்ளது. உடனுக்குடன் எழுத்துப்பிழைகள் திருத்தப்படும்பொழுது மாணவர்களிடையே மெய்நிகர் வாயிலான கலந்துரையாடலும் சிக்கலைக் களையும் நுண்ணறிய ஆற்றலும் மேலாங்கிக் காணப்பட்டது. குழு உறுப்பினர்களின் எழுத்துப்பிழைகளைச் சுட்டிக் காட்டி அப்பிழைகளைத் தாங்கள் தவிர்க்கவும் இந்த ஊடாடும் வெண்பலகை பயன்வழங்கிற்று. தமிழ் மொழி பாடத்திட்டத்தில் வரையறுக்கப்பட்ட அடிப்படை மொழித்திறன்களைக் கொண்டு தமிழ்மொழியைப் பயன்படுத்தி பிறருடன் தொடர்புச் கொள்ளவும் வாசிப்பின்வழி அறிவைப் பெறவும் இலக்கணப்பிழைகளின்றி தமிழ் மொழியைப் பயன்படுத்தித் தமமான படைப்புகளைப் படைக்கும் ஆற்றலை மாணவர்கள் கைவரப் பெறவும் இவ்வாய்வு துணைபுரிந்துள்ளது. 21ஆம் நூற்றாண்டிற்கான திறனும் பண்பும் மாணவர்களிடையே வளர்க்கவும் மெய்நிகர் கற்றல் பயனாக அமைந்தது. ஊடாடும் வெண்பலகை வழியிலான கற்றல் மாணவர்களிடையே தொடர்புக் கொள்ளும் திறனை வளர்த்துள்ளது. பல்வகை ஊடகங்களையும் தொழில்நுட்பத்தையும் பயன்படுத்தி தங்களின் சிந்தனைத்திறன், ஆக்கத்திறன், தகவல் பரிமாற்றம் ஆகியவற்றைக் கொண்டு தன்னம்பிக்கையுடன் தங்கள் கருத்துகளை எழுத்து வடிவில் படைக்க உதவியது. குழுவாகச் செயல்படுகையில் வினையயன்மிகக் வகையில் பிறருடன் ஒத்துழைத்து மற்றவர் கருத்துகளுக்கு மதிப்பளித்து, குழு நண்பர்களின் பங்களிப்பை மதித்துக் குழுவோடு சேர்ந்து பொறுப்பைப் பகிர்ந்து கொள்ளும் பாங்கினை வளர்த்துள்ளது. அதோடு, மெய்நிகர் கற்றல்கற்பித்தல் அணுகுமுறையானது மாணவர்களின் ஆர்வத்தைத் தூண்டும் வகையில் அன்றாட வாழ்க்கைக்கு தொடர்புடையதாகவும், மகிழ்ச்சியூட்டுவதாகவும் அமைந்து மாணவர்களின் முழுமையான ஈடுபாட்டினைக் கொண்டுள்ளது.

புஷ்பராணி சுப்ரமணி செல்வன், மலாயாப் பல்கலைக்கழகம், கோலாலம்பூர்

## நன்றி

இவ்வாய்வினைச் செயல்படுத்த அனுமதி நல்கிய தேசிய வகை மாசாய்க் குழுவகத் தமிழ்ப்பள்ளி (மலேசியா) தலைமையாசிரியர் திரு விஜயன் மாதவன் நாயர் அவர்களுக்கும் ஆய்வில் கலந்து கொண்ட மாணவர்களாகிய, அரிஷன் மகேஷ்வரன், சாருமித்ரா பிரேம்குமார், டேவினா கோபிநாதன், கிரன்ராஜ் வெங்கடேஸ், கீர்த்தி முருகன், குகேஷன் விக்னேஷ்வரன், மனிஷா விக்னேஷ்வரன், ரித்தேஷ் சரவணகுமார், திரனேஷ் கோகுலபாலன், விக்னேஷ் நவீன்குமார் ஆகியோருக்கும் மனமார்ந்த நன்றிகள்.

## REFERENCES

- Al-Kathiri, F. (2014). Beyond the Classroom Walls: Edmodo in Saudi Secondary School EFL Instruction, Attitudes and Challenges.
- Balakrishnan, V., & Gan, C. L. (2016). Students' learning styles and their effects on the use of social media technology for learning. *Telematics and Informatics*, 33(3), 808-821.
- Balakrishnan, V., Liew, T. K., & Pourgholaminejad, S. (2015). Fun learning with Edooware – A social media enabled tool. *Computers & Education*, 80, 39-47.
- Balasubramanian, K., Jaykumar, V., & Fukey, L. N. (2014). A Study on "Student Preference towards the Use of Edmodo as a Learning Platform to Create Responsible Learning Environment". *Procedia - Social and Behavioral Sciences*, 144, 416-422.
- Berns, A., Gonzalez-Pardo, A., & Camacho, D. (2013). Game-like language learning in 3-D virtual environments. *Computers & Education*, 60(1), 210-220.
- Falloon, G. (2015). What's the difference? Learning collaboratively using iPads in conventional classrooms. *Computers & Education*, 84, 62-77.
- Foster, R., & Neal, D. R. (2012). 12 - Learning social media: student and instructor perspectives *Social Media for Academics* (pp. 211-226): Chandos Publishing.
- Gan, B., Menkhoff, T., & Smith, R. (2015). Enhancing students' learning process through interactive digital media: New opportunities for collaborative learning. *Computers in Human Behavior*, 51, Part B, 652-663.
- Grosseck, G., & Holotescu, C. (2010). Microblogging multimedia-based teaching methods best practices with Cirip.eu. *Procedia - Social and Behavioral Sciences*, 2(2), 2151-2155.
- Holotescu, C., Grosseck, G., & Danciu, E. (2014). Educational Digital Stories in 140 Characters: Towards a Typology of Micro-blog Storytelling in Academic Courses. *Procedia - Social and Behavioral Sciences*, 116, 4301-4305.
- Ng, W. (2012). Can we teach digital natives digital literacy? *Computers & Education*, 59(3), 1065-1078.
- Phungsuk, R., Viriyavejakul, C., & Ratanaolarn, T. Development of a problem-based learning model via a virtual learning environment. *Kasetsart Journal of Social Sciences*.
- Samburskiy, D. (2014). Corpus-informed Pedagogical Grammar of English: Pros and Cons. *Procedia - Social and Behavioral Sciences*, 154, 263-267.

## பைத்தான் நிரல்மொழி மூலம் விக்கிமூலம் இயங்கும் முறைகள் (Ways to run a Tamil wikisource through the Python programming language)

சத்தியராஜ் தங்கச்சாமி<sup>1</sup>, க. சண்முகம்<sup>2</sup>, தகவலுழவன்<sup>3</sup>

<sup>1</sup> தமிழ் உதவிப்பேராசிரியர், ஸ்ரீகிருஷ்ணா ஆதித்யா கலை அறிவியல் கல்லூரி, கோயம்புத்தூர் - 641042,  
9600370671, [sathiyaraj@skacas.ac.in](mailto:sathiyaraj@skacas.ac.in)

<sup>2</sup> உதவிப் பேராசிரியர், கணினி அறிவியல் துறை, எஸ்.ஆர்.எம். வள்ளியம்மை பொறியியல் கல்லூரி, சென்னை-603203,  
[shanmugamk.cse@valliammai.co.in](mailto:shanmugamk.cse@valliammai.co.in)

<sup>3</sup> விக்கிமீடியர், சேலம் - 620001, +919095343342, [tha.uzhavan@gmail.com](mailto:tha.uzhavan@gmail.com)

### ABSTRACT

விக்கிப்பீடியா என்ற சொல் இன்று அனைவரின் கவனத்தையும் ஈர்த்துள்ளது. ஆனால் அது விக்கித்திட்டங்களில் ஒரு பகுதித் திட்டமாகும். இந்த விக்கித் திட்டங்களைப் பொதுவாகக் குறிப்பிட விக்கிமீடியா என்ற சொல் பயன்படுத்தப்பட்டு வருகின்றது. இத்திட்டங்களில் 'விக்கிப்பீடியா - கலைக்களஞ்சியம், விக்கிநூல் - அகராதியும் சொற்களஞ்சியம், விக்கிநூல்கள் - பாடநூல்கள், கையேடுகள், வழிகாட்டிகள், விக்கிசெய்திகள் - செய்திகள், விக்கிமேற்கோள் - மேற்கோள்களின் தொகுப்புகள், விக்கிமூலம் - நூலகம், விக்கிப் பல்கலைக்கழகம் - கற்றல் கருவிகளும் செயல்பாடுகளும், விக்கிப் பயணம் - பயண வழிகாட்டிகள், விக்கியினங்கள் - இனங்களின் அடைவு, பொதுவகம் - ஊடகங்களின் களஞ்சியம், மேல் - விக்கி - விக்கி ஊடகத்திட்ட ஒருங்கிணைப்பு, விக்கி ஊடகம் - மென்பொருள் ஆவணப்படுத்தும்' ஆகியன உள்ளன. இதன் பகுப்புமுறை தேவைக்கு ஏற்பப் பல்கும். இத்திட்டங்களில் விக்கிமூலத்தை மேம்படுத்துவதற்கு அல்லது இயங்கும் முறைக்குப் பைத்தான் நிரலின் தேவை உள்ளது என்பதையும் பைத்தான் நிரலில் தமிழ் மொழியை உள்ளீடு செய்து நிரலாக்க மொழியைக் கொண்டு எழுத முடியும் என்பதையும் இவ்வாய்வுரை வெளிப்படுத்தும்.

Wikipedia projects' Wikipedia - Encyclopedia, Wiktionary - Dictionary and Glossary, Wikisource - Textbooks, Notes, Guides, Wikisnews - News, Wikiquote - Quotes, Wikisource - Library, Wikiuniversity - Learning Tools and Works, wikitour - Media repository, top - wiki - wikimedia project integration, wikimedia - software documentation'. Its analysis is adaptable. This review reveals that there is a need for a Python program to improve or run Wikisource on these projects, and that Python can be written in Tamil.

### Keywords:

- A விக்கிமூலம்
- B தமிழ் இயற்கை மொழி ஆய்வு
- C பைத்தான் நிரல்
- D Wikisource
- E Tamil NLP

### Corresponding Author:

சத்தியராஜ் தங்கச்சாமி,  
தமிழ் உதவிப்பேராசிரியர்,  
ஸ்ரீகிருஷ்ணா ஆதித்யா கலை அறிவியல் கல்லூரி,  
கோயம்புத்தூர் - 641042,  
தமிழ்நாடு, இந்தியா,



## 1. அறிமுகம்

இன்றைய காலத்தில் தொழில்நுட்பம் அனைத்துத் துறைகளிலும் கால் பதித்துள்ளது. அந்தத் தொழில்நுட்பம் பல்கிப் பெருகுவதற்குக் கணினி மொழி அடிப்படையாக அமைந்துள்ளது. இந்தக் கணினிமொழி சிறந்து விளங்கப் பல்வேறு நிரலாக்க மொழி இக்காலத்திலும் சரி எதிர்காலத்திலும் சரி அனைத்துத் துறைகளிலும் சிறந்து விளங்கும் மொழியாகப் பைத்தான் அமைகின்றது.

இப்பைத்தான் மொழியைக் கற்கும்பொழுது, தொழில்நுட்பம் சார்ந்த பல்வேறு செயல்பாடுகளைச் செய்யலாம். தமிழில் நிரல் எழுதுவதற்கு எழில் என்ற நிரலாக்க மொழியும் உள்ளமை குறிப்பிடத்தக்கது. இருப்பினும் பைத்தான் மொழியைப் பெரும்பான்மையான நிரலாளர்களால் பயன்படுத்தப்பெற்று வரும் சிறந்த கணினி நிரலாக்க மொழியாகவும் விளங்கி வருகின்றது. எனவே, தமிழிலேயே பைத்தான் நிரலாக்கங்களை எழுதமுடியும் என்பதையும், அந்நிரலாக்கங்கள் விக்கிமூல மேம்பாட்டிற்குப் பயன் நல்கும் முறைகளையும் அறியத் தருவதாய் இக்கட்டுரை அமைகின்றது.

## 2. ஆய்வணுகுமுறை

பைத்தான் நிரல்மொழியானது விக்கிமூலத்தில் உள்ள நூல்களைத் திருத்தம் செய்வதற்கு எவ்வகையில் பயன் தருகின்றது என்பதையும் பைத்தான் மொழியைத் தமிழிலும் எழுதி, இயக்க வைக்க முடியும் என்பதையும் விளக்கி நிற்பதால், விளக்கவியல் அணுகுமுறை இங்குப் பின்பற்றப்பெற்றுள்ளது.

## 3. ஆய்வுமுடிவும் பகுப்பாய்வும்

இந்த ஆய்வில் ஆங்கிலம் தெரிந்தால்தான் பைத்தான் கற்க முடியும் என்ற சூழலை மாற்றும் முடிவதைத் தருவதைப்போலவே, விக்கிமூலத்தைப் பைத்தான் மொழியைக் கொண்டு மேம்படுத்த இயலும் என்ற சூழலையும் உணரக்கூடிய முடிவைத் தருகின்றது.

பைத்தான், விக்கிமூலம் குறித்த அறிமுகங்களையும் அதனைக் கற்பதற்கான வழிமுறைகளையும் பகுப்பியல் நோக்கில் வகைப்படுத்தி விளக்கப்பெற்றுள்ளது.

### 3.1. விக்கியில் பைத்தான்

பொதுவாகப் பைத்தான் நிரல்களை இயல்பிருப்பாக லினக்ஸ் வகைக் கணினிகளில் இயக்கலாம். ஆனால் விண்டோசு போன்ற இயக்குத்தளங்களில் பைத்தான் பொதிகளை நிறுவ வேண்டும். ஆனால் விக்கி வழங்கியிலேயே இருக்கும் வசதியை, உலாவியிலேயே பயன்படுத்தி நிரல்களை இயக்குதல் என்பது எளிமை.

எண்ணும் எழுத்தும் ஒரு மொழிக்கு அடிப்படை. ஒரு மொழியைப் பேசுகின்ற மனிதனுக்கு அது உயிர் போன்றது. அதனைப் பின்வரும்,

**எண்ணென்ப ஏனை எழுத்தென்ப இவ்விரண்டும்**

**கண்ணென்ப வாழும் உயிர்க்கு. (குறள்.392)**

எனும் திருக்குறள் வெளிப்படுகின்றது. இவ்விரண்டினையும் (எண், எழுத்து), விரைவாகக் கற்பதற்கு உதவுவது, தொழில்நுட்பம் ஆகும். அதில் தலையானது, கணினி மொழியாகும். கணிய மொழிகளில், பைத்தான் (Python) நிரலாக்கம் கற்பதற்கு எளிமையானது. உலகின் பல நாடுகளில், பைத்தான் மொழியானது, இலட்சக்கணக்கானவர்களால் கற்கவும், கற்பிக்கவும் பயன்படுகிறது. இம்மொழியை, கூகுள், யாகூ போன்ற பல வெற்றி அடைந்த வணிக நிறுவனங்களும் பயன்படுத்துகின்றன. விக்கிமீடியத்திட்டங்களின் இணையப்பக்கங்களை உருவாக்குவதற்கும், மேம்படுத்துவதற்கும், தூய்மைப் படுத்துவதற்கும், பல கணிய மொழிகள் பயன்பட்டாலும், இம்மொழியே பெருமளவில் பயன்படுகிறது.

இது ஒரு பொதுக்கள / திறநிலை உரிமத்தில் (open source) இருக்கும் மென்பொருள் ஆகும். இதனால் இது எப்படிச் செயற்படுகிறது என்பதை வெளிப்படையாகத் தெரிவிக்கின்றனர். நமக்குத் தேவையான உதவிக் குறிப்புகளை, இணையத்திலேயே காணலாம். இணையத்தில் இல்லாதவற்றையும், ஐயங்களையும் கேட்டுப் பெறமுடியும்.

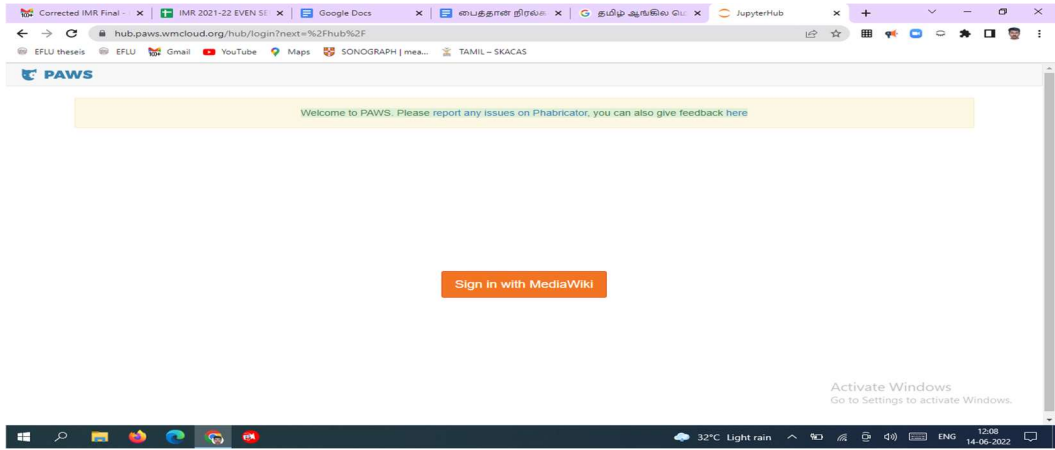
இம்மொழி, மனித மொழியை எளிதாகக் கையாளக்கூடியது. குறிப்பாக, உரைகளை (பனுவல் = text) எளிதாகக் கையாளக்கூடிய திறன் மிக்கது ஆகும். மேலும், இம்மொழியைக் கொண்டு, ஒரு இணையதளத்தை, இணையத்தளச் செயலிகளை (app-s), தேடுபொறியை (search engine like google,yahoo...), கணிய விளையாட்டுக்களை வடிவமைக்க இயலும். இத்திறன்மிகு மொழியை விக்கியில் இயக்கிப் பார்க்கும் வழிமுறைகளை இனிக் காண்போம்.

### 3.2. மின்னியக்கப் பைத்தான் (PAWS)

பாவ்சு (PAWS) என்பதற்கு,

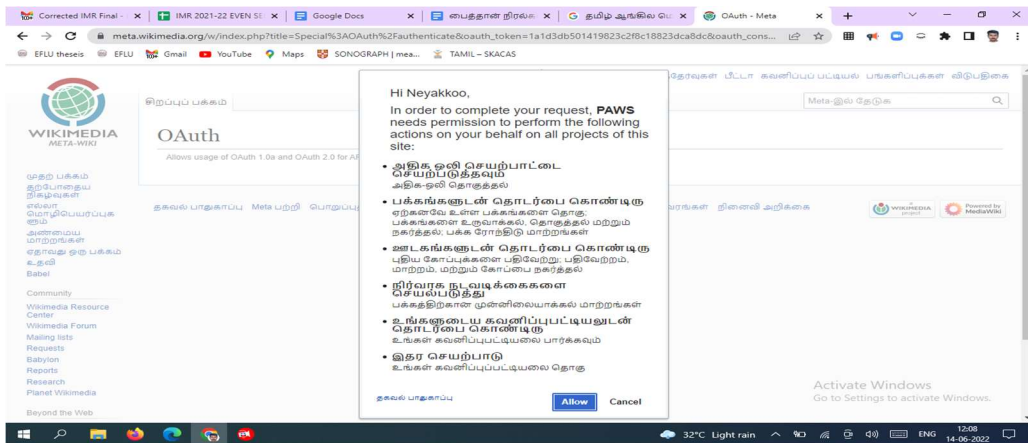
- It is a online browser based programming IDE provided by wikimedia foundation
- It is based on jupyter notebook software

என்பதாகச் சீனிவாசன் விளக்கம் கூறியுள்ளார் (shrini, கங-௦௬-௨௦௨௨ உக:௨௭). இருப்பினும் இதனைக் குளோபல் விக்கியில் உள்ளவருள் முக்கியமான சான் கெராட்டு சாய் என்பவர் Python As a Web Service என விளக்கம் அளித்துள்ளார். மேலே கூறியது போல விக்கித்திட்டத்தில் பைத்தான் நிரலாக்க மொழியை இயக்கிப் பார்ப்பதற்கான இந்தப் பாவ்சு வழிமுறையை இனிக் காண்போம்.



படம்-1. பாவ்சு பக்கம் சென்று புகுபதிகை செய்ய

இப்படம் பாவ்சு பக்கத்திற்குள் நுழைவதற்கான வழிமுறையாகும். இப்படத்தில் காட்டியுள்ளபடி, sign in with mediawiki பொத்தானை அழுத்தப் பின்வரும் படம் தோன்றும்.

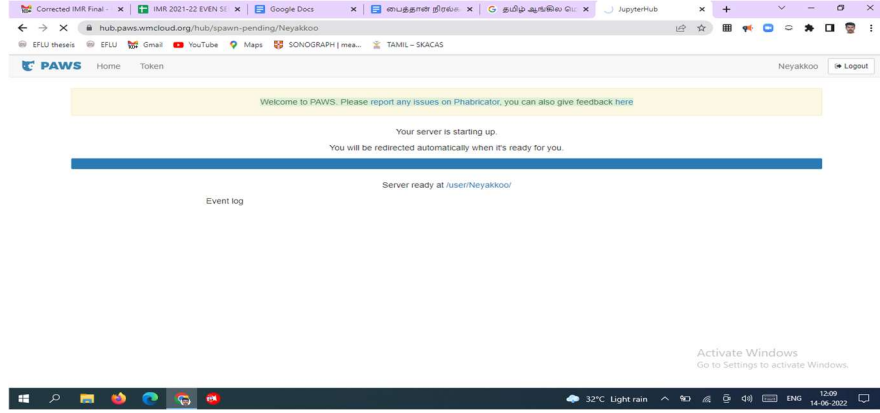


சத்தியராஜ் தங்கச்சாமி, தமிழ் உதவிப்பேராசிரியர், ஸ்ரீகிருஷ்ணா ஆதித்யா கலை அறிவியல் கல்லூரி, கோயம்புத்தூர்



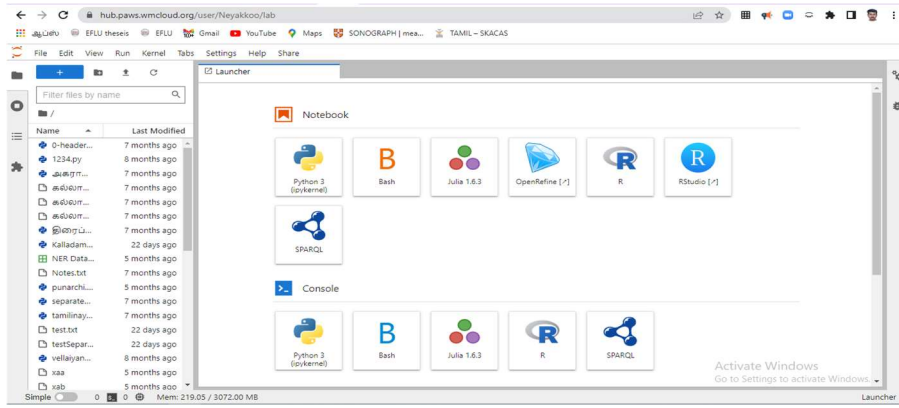
### படம்-2. ஏற்கும் குறிப்பை அழுத்துவதற்கான பெட்டி

இப்படத்தில் காட்டியுள்ள குறிப்புகளை உள்வாங்கிய பின்பு தாங்கள் ஏற்கும் நிலையில் Allow பொத்தானை அழுத்தவும். அது பின்வருமாறு விரியும்.

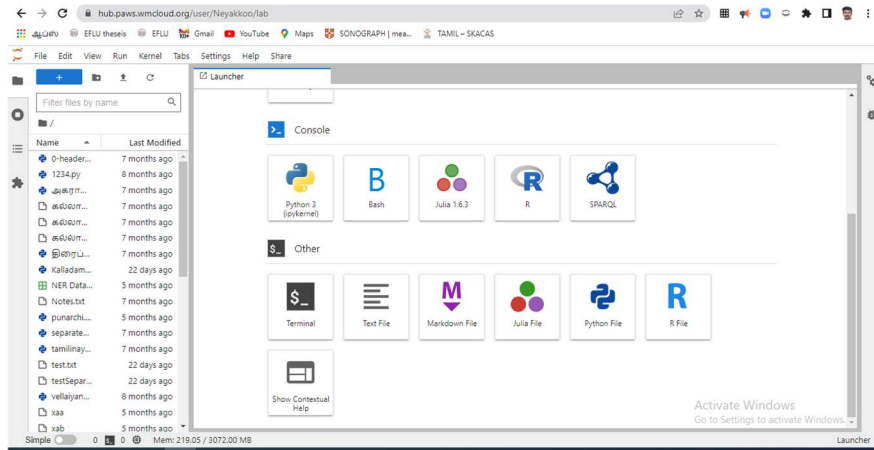


### படம்-3. பாவுக பக்கம் சூபிடருக்குள் செல்லும்

இந்தப் படத்தில் காட்டியுள்ளபடி, சூபிடர் குறிப்பேட்டுப் பக்கத்திற்கு நுழையப் பின்வரும் படம் தோன்றும்.



### படம்-4. பாவுக சூபிடர் பக்கத்தில் குறிப்பேடு, பணியகம், பிற செயல்பாடுகளைக் காணலாம்



### படம்-5. பாவுக சூபிடர் பக்கத்தில் குறிப்பேடு, பணியகம், பிற செயல்பாடுகளைக் காணலாம்

இவ்விரு படங்களில் காட்டியுள்ளபடி குறிப்பேடு (Notebook), பணியகம் (Console), பிற (Other) எனும் மூன்று பகுப்புகள் காணப்படுகின்றன. இங்கு பைத்தான்3 (Python 3 (ipykernel)), பாசு (Bash), சூலியா (Julia 1.6.3), ஓப்பன் ரீபைன் (OpenRefine [7]), ஆர் (R), ஆர் சுடியோ (RStudio [8]), இசுப்பார்குல் (SPARQL) ஆகிய நிரலாக்க மொழிகள் உள்ளன. இவற்றின் மூலம் நமக்கு நாமே நிரல் எழுதலாம்; இயக்கிப் பார்க்கலாம். அதற்கு முன்பு பைத்தான் நிரல் மொழியை எழுதச் சில அடிப்படை விதிமுறைகள் அறிந்து வைப்பது சிறப்பு. குறிப்பாக, ஒரு மொழியைத் திறம்படப் பேசு, எழுத நமக்குத் துணை நிற்பது எழுத்து, சொல், தொடர், பொருண்மை ஆகியனவாகும். அதுபோல் கணினியில் பைத்தான் நிரலாக்க மொழியைத் திறம்படக் கையாளக் குறிச் சொறைகளை அறிந்து வைத்தல் வேண்டும்.

பொதுவகத்தில் இருந்து நூற்குறிப்புகளை எடுத்து, அவற்றை இங்குள்ள அட்டவணைகளில் நிரப்பிய பைத்தான் நிரல் பின்வருமாறு தொடக்கத்தில் எழுதப்பட்டு வந்தது.

```
import csv
import pywikibot
import time
import re
WAIT_TIME = 15
with open('2015-tva-commons-pdf-books-all-info.csv', 'r') as csvfile:
    reader = csv.reader(csvfile, delimiter="~")
    for row in reader:
        #if len(row) == 8:
        # if not 'booktitle' in row:
        #if you use other PAWS, remove the hash to decode well
        wikiPage1 = row[0].replace('File', 'Index')#.decode('utf-8')
        bookAuthor = row[1]
        bookSize = row[3].replace('MB', '')
        indexPages = row[5].replace('pages', '')
        print (wikiPage1)
        print (bookAuthor)
        print (indexPages)
        print (bookSize)

        site = pywikibot.Site('ta', 'wikisource')
        page1 = pywikibot.Page(site, wikiPage1)

        res1 = re.compile("\\| Number of pages=*(\\d+)").search(page1.text)
        if res1:
            print("number of pages is already assign to %s" % res1.group(1))
        else:
            page1.text = page1.text.replace('| Number of pages=', '| Number of
pages='+indexPages)
            page1.save(summary='+ கோப்பளவு = ' + bookSize + ', நூற்பக்கங்கள் =
'+indexPages)

        res2 = re.compile("\\| File size=*(\\d+)").search(page1.text)
        if res2:
            print("File size is already assign to %s" % res2.group(1))
        else:
            page1.text = page1.text.replace('| File size=', '| File size='+bookSize)
```

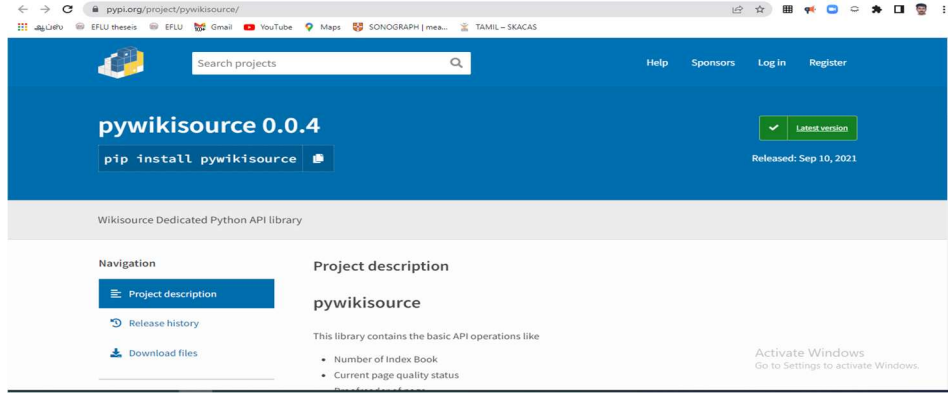
சத்தியராஜ் தங்கச்சாமி, தமிழ் உதவிப்பேராசிரியர், ஸ்ரீகிருஷ்ணா ஆதித்யா கலை அறிவியல் கல்லூரி, கோயம்புத்தூர்

page1.save(summary='+ கோப்பளவு '+' + bookSize+ ', நூற்பக்கங்கள் = '+indexPages)

time.sleep(WAIT\_TIME)

அதன்பின்பு தமிழில் பைத்தான் நிரல் எழுதி, இயக்கிப் பார்க்கப்பெற்றது.

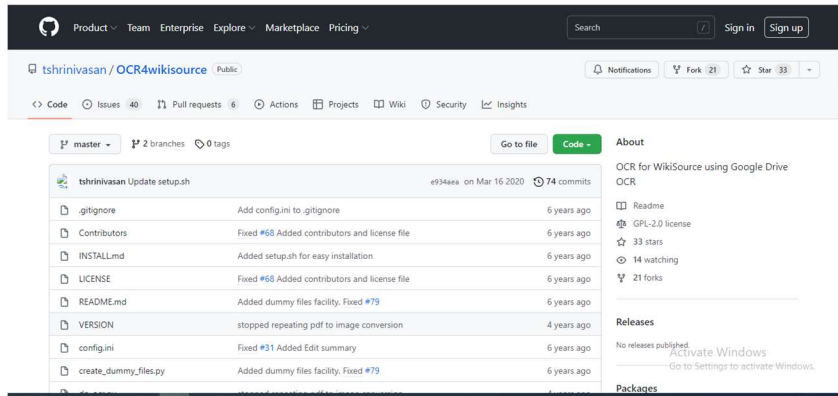
விக்கிப்பீடியா: பைவிக்கிதானியங்கி என்பது பல விக்கித்திட்டங்களுக்கு உதவும் பெரும் நிரல்தொகுப்பாகும். இது பல ஆண்டுகளாகப் பலரால் பைத்தான்2-வில் எழுதப்பட்டு வந்தது. அதில் சிலவற்றைத் தற்பொழுது பைத்தான்3-க்கு மாற்றியுள்ளனர். இந்தப் பைத்தான் நூற்கட்டகத்தினைப் பின்புலத்தில் வைத்து, தமிழ் விக்கிமூலத்திற்கு எனத் தமிழ்வழி இயக்கும் பைத்தான் நூற்கட்டகத்தினை (Library & wrappers) உருவாக்க இயலும். அத்துடன் இந்தப் பைவிக்கிமூல நூற்கட்டகமும் உதவும். அதனைக்காட்டும் படம் வருமாறு;-



படம்-6. பைவிக்கி மூலத்தின் மூலநிரல்

எந்த மொழியினையும், முனையத்தில் (Terminal) தெளிவாகப் படிக்க நீங்கள் பயன்படுத்தும் இயக்குத்தளம், முனையம் மிகமுக்கியம். அதற்கு, டெபியன்10, கன்சோல் சிறப்பு. அதில் பதிப்புகள் மாறினாலும், அனைத்து மொழிகளும் தெளிவாகத் தெரியும்.

சீனிவாசன் எனும் நிரலாளர் அளித்த 'ஓசியார்4விக்கிசோர்சு (OCR4wikisource)' விக்கிமூலத்திற்கு ஒரு வரப்பிரசாதம் என்றே கூறலாம். அதற்குரிய மூலநிரல் படம் வருமாறு;-



படம்-7. ஓசியார்4விக்கிசோர்சு - மூலநிரல்

இது ஆவணத்தப் படித்துத் தட்டச்சுப் படியாக மாற்றும் திறன் கொண்டது. அச்சுப் பக்கத்தில் உள்ளவற்றைத் தட்டச்சுச் செய்து, பின்பு பிழைநீக்கம் செய்வது என்பது எளிதில் முடியக் கூடிய பணி கிடையாது. எனவே, ஓசியார் வரவு வரவேற்கத்தக்கது எனப்பட்டது. இது 90 விழுக்காட்டுப் பணியைத் தானே செய்து முடித்து விடுகின்றது. மீதமுள்ள 10 விழுக்காட்டுப் பணியைத்தான் மனிதவளம் செய்து வருகின்றது.

உலாவி வழியே செயற்படுத்தும் போது, 'பைவிக்கிபாட்' என்ற பைத்தான் நூற்கட்டகத்தின் ஒருங்குறிய வழக்கள் பெருமளவு தவிர்க்கப்படுகின்றன. அது ஒப்பீட்டளவில் அனைவருக்கும் எளிமையானது. எனவே இது யூவி பான்டே என்ற புனைப்பெயரை உடைய சென்னைத் தமிழரால் தோற்றுவிக்கப்பட்டது என்பது குறிப்பிடத்தக்கது.

பைத்தான்2-வின் மூலம் விக்கிமூலத்தை மேம்படுத்த சில நிரலாக்கங்கள் எழுதப்பெற்றன. சான்றாக,

```
#!/usr/bin/python2
```

```
import pywikibot
```

```
aPage = page:அங்கும் இங்கும்.pdf/9
```

```
site = pywikibot.Site('ta', 'wikisource')
```

```
page = pywikibot.Page(site, aPage)
```

```
print u"page"
```

எனும் நிரல்மொழியைக் கூறலாம். இவ்வாறு தொடக்கக் காலத்தில் எழுதி; இயக்கி விக்கிமூலம் மேம்பாடு செய்யப்பெற்று வந்தது. இவ்வாறு எழுதிய நிரலாக்கமொழியின் விளைவுப் பின்வருமாறு அமைகின்றது.

**Output:**ta:page:அங்கும் இங்கும்.pdf/9

இதன்பின்பு பைத்தான்3 வருகை நிரலாக்க மொழியைத் தமிழில் எழுதலாம் என்ற சூழல் உருவானது எனலாம். அதனை வெளிப்படுத்தும் நிரலாக்கம் வருமாறு.

```
#!/usr/bin/python3
```

```
import பைவிக்கிமூலம்
```

```
பைவிக்கிமூலம் = பைவிமூ
```

```
எடுக்கும்பக்கம் = பக்கம்:அங்கும் இங்கும்.pdf/9
```

```
உரலி = பைவிமூ.உரலியிடு(எடுக்கும்பக்கம்)
```

```
விளைவிடு(உரலி)
```

இதன் விளைவு வருமாறு:-

**விளைவு :** ta:பக்கம்:அங்கும் இங்கும்.pdf/9

இதன்மூலம் நிரலாக்கம் எழுதுவது எளிது என்பதை நாம் உணர்ந்துகொள்ளலாம்.

#### 4. முடிவு

இதுவரை விளக்கப்பெற்றதின் அடிப்படையில் பைத்தான்3இல் தமிழில் நிரல்மொழி எழுதி விக்கிமூலத்தினை மேம்படுத்த முடியும் என்பதை அறிந்துகொண்டோம். அது ஒரு தொடக்கப்புள்ளியே. இதுபோன்ற பல நிரலாக்க மொழிகளை இன்னும் அறிந்துகொள்ள [பகுப்பு:பைத்தான் நிரல்கள்](#) எனும் பக்கத்திற்குச் சென்று, கற்கலாம்; விக்கிமூலத்தை மேம்படுத்துவதற்கான மூலநிரல்களை உருவாக்கலாம்.

சத்தியராஜ் தங்கச்சாமி, தமிழ் உதவிப்பேராசிரியர், ஸ்ரீகிருஷ்ணா ஆதித்யா கலை அறிவியல் கல்லூரி, கோயம்புத்தூர்

## உறுதிமொழி (ACKNOWLEDGEMENTS)

இவ்வாய்வை இக்கட்டுரையின் ஆசிரியர்களாகிய நாங்கள்தான் எழுதினோம். இக்கட்டுரை வேறு எங்கும் வெளியிடப்பெறவில்லை என்பதையும் பிறரது கருத்தை எங்களது கருத்து எனக் காண்பிக்கவும் இல்லை என்பதை உறுதிபடக் கூறுகின்றோம்.

## துணைநின்றவை (REFERENCES)

- பைத்தான் - <https://ta.wikipedia.org/s/112>
- [hub.paws.wmcloud.org/user/Neyakkoo/lab](https://hub.paws.wmcloud.org/user/Neyakkoo/lab)
- Hegde, S. U., Hande, A., Priyadharshini, R., Thavareesan, S., Sakuntharaj, R., Thangasamy, S., ... & Chakravarthi, B. R. (2021). Do Images really do the Talking? Analysing the significance of Images in Tamil Troll meme classification. *arXiv preprint arXiv:2108.03886*.
- சத்தியராஜ் தங்கச்சாமி. (2021). மொழி, ஓர் அமைப்பொழுங்கு அணுகுமுறையில் தொல்காப்பிய எழுத்த்திகார நூன்மரபு: Ilakkaṇaviyal aṇukumuraiyil tolkāppiya eḷuttatikāra nūṇmarapu. *இனம் பன்னாட்டு இணையத் தமிழாய்விதழ் (Inam International E-Journal of Tamil Studies)*, 7(28), 58-71.
- <https://kaapiyam.com/tiruvalluvar-tirukkural-meaning-definition-tamil-english-daily-kural/enneppa-enai-ezhuththenpa-ivvirantum-392/>
- விக்கிமூலம் - பைவிக்கிமூலம் - <https://ta.wikipedia.org/s/5ety>
- <https://pypi.org/project/pywikisource/>
- <https://github.com/tshrinivasan/Python-Beginners-Guide>
- <https://github.com/tshrinivasan/tools-for-wiki>
- <https://beginnersbook.com/2019/06/python-user-defined-functions/>
- எழில் நிரலாக்க மொழி - <https://ta.wikipedia.org/s/27xm>
- ஸ்வரம் நிரலாக்க மொழி - <https://ta.wikipedia.org/s/27v2>
- நிரலாக்கம் தலைப்புப் பட்டியல் - <https://ta.wikipedia.org/s/6x3>
- கணினியில் தமிழ் - <https://ta.wikipedia.org/s/5v8>
- வலைவாசல் - தமிழ்க்கணிமை - <https://ta.wikipedia.org/s/3v8>
- விக்கிமூலம்:பைத்தான் நிரல்கள்
- விக்கிமூலம்:பைத்தான் நிரல்கள்/பைவிக்கிமூலம்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/அகரமுதலிச்சொற்கள்தடிமனாக்கம்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/அட்டவணை நிரப்பி தொகுதி
- விக்கிமூலம்:பைத்தான்நிரல்கள்/அட்டவணையின் பகுப்புகள்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/இருபக்கங்களில் பிரிந்த சொல்லிணைப்பு
- விக்கிமூலம்:பைத்தான்நிரல்கள்/கீழடி நடுவில் எண் மட்டும்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/தலைப்பைநகர்த்தல்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/தலைப்பைநகர்த்தலின் பக்கங்கள்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/நூற்றுணைப்பக்க உருவாக்கல்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/பக்கயெண்ணிக்கைப் பகுப்பிடல்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/பகுதிக்குறியீடுகள்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/பகுப்புப்பக்கங்களைஎடுத்தல்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/பொருளடக்கத்துப்புரவு
- விக்கிமூலம்:பைத்தான்நிரல்கள்/பொருளடக்கமில்லா நூலின் துணைப்பக்கங்களை உருவாக்குதல்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/முயற்சி/பிரிந்த சொற்களை இணைத்தல்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/முழுப்பக்கத்துப்புரவு
- விக்கிமூலம்:பைத்தான்நிரல்கள்/மேலடி நடுவில் எண் மட்டும்
- விக்கிமூலம்:பைத்தான்நிரல்கள்/API/பகுப்புப் பக்கங்களை எடுத்தல்
- விக்கிமூலம்:பைத்தானில் அணித்தரவுக்கோப்பு

## Automatic Question Generation using Centrality-based Keyword Extraction Approach for Tamil Text

Senthilkumar P<sup>1</sup>, Nandhini K<sup>2</sup>

<sup>1,2</sup>Department of Computer Science, Central University of Tamil Nadu, Thiruvavur, India.

---

### ABSTRACT

The main objective of an assessment is to measure student's learning abilities and increase such abilities by correcting them in line with their knowledge. Question generation plays a vital role in assessment. The creation of the questions for the assessment is a challenging task, and manually creating a question is a complicated operation that requires expertise, knowledge, and, most importantly, it is a time-consuming process. Automatic question generation (AQG) is a savior to overcome these issues. AQG comprises of six critical stages, with keyword selection being the most significant stage and a crucial component of question quality. Therefore, we focus on the keyword selection method and propose a novel method for automatically extracting keywords from Tamil text to generate questions in Tamil. This keyword selection process involves two important concepts; the first focuses on how significant a keyword is depending on web-based search results and the notable information it contains, the second focuses on giving a keyword weight depending on how frequently it appears and how it is distributed in the corpus. The keyword selection method performs significantly better for Tamil-based question generation when compared to certain other keyword selection techniques in terms of accuracy and question quality.

---

### Keywords:

A Question Generation  
B Natural Language Processing  
C Keyword Selection  
D Tamil Question  
E Tamil Keyword

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

---

### Corresponding Author:

Senthilkumar P,  
Department of Computer Science,  
Central University of Tamil Nadu, India.  
Email: way2sen@gmail.com

---

## 1. INTRODUCTION

Automatic question generation is a concept that takes text as input and generates a question as the output based on the text it received. The created questions can be used to assess skills or check whether the text's information was understood. Automatic question generation is broadly classified into extractive and abstractive is shown in Figure.1.

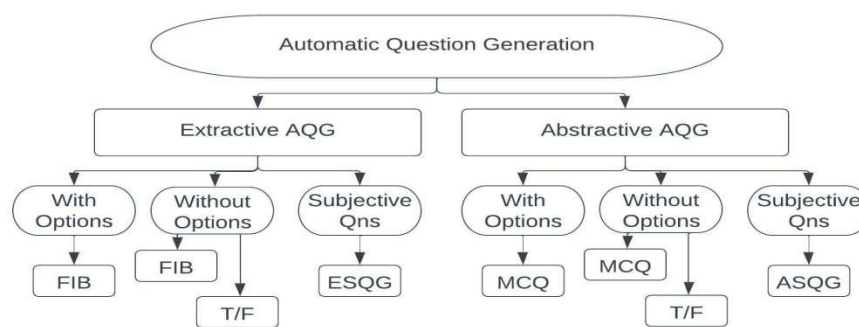


Figure 1. Automatic Question Generation Classification

Table 2. List of Acronyms

Acronym	Meaning of Acronym
FIB	Fill in the blanks Questions.
FIB-with Options	Fill in the blanks Questions with Options.
T/F	True or False Questions.
ESQG	Extractive Subjective Question Generation.
MCQ	Multiple Choice Questions.
ASQG	Abstractive Subjective Question Generation.
ECQG	Extractive Cloze Question Generation.
EOQG	Extractive Open Question Generation.

Extractive questions are extracted from text as such whereas abstractive questions are generated from the text. Our focus is on extractive question generation to evaluate the learning skills of the primary school students from the text. There exists a variety of English language-based automatic question generation technique([1],[2],[3]) but local language-based automatic question generation like the Tamil language-based are very few in number. Automatic question generation has six stages[4] of processing is shown in Figure.2 and every stage have some tasks to be performed. The first simple stage of AQG is Text preprocessing and it includes Text Normalization, Lexical Analysis, Statistical Analysis, and Syntactic Analysis. It simply gives supported text format for processing, and it is the filtering out the computable text.

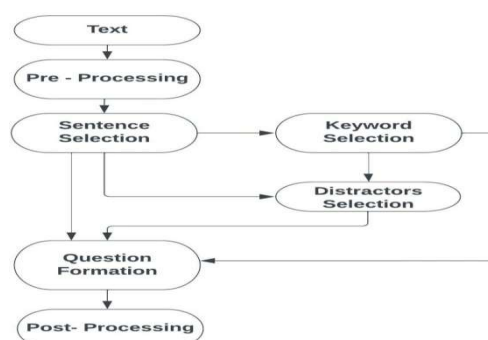


Figure 2. Phases of Automatic Question Generation



















Sentence selection is the second task of AQG, and a sentence that contain questionable information, can act as a candidate for generating questions. Sentence is selected based on some keywords on it. Keyword selection is an essential step that determines the question generation worth, and it is done based on some methods like, Word count[5], POS tagging based[6], Pattern matching[7], and some trained ML based[8].

Question Formation is the next stage of the question generation, it checks the format of question framed correctly. Post processing is the final stage of question generation, it does the job of filtering question[9], ranking question and changing distractors in the questions if needed.

The goal of machine learning is to develop models that operate automatically and learn on their own. NLP, on the other hand, makes it possible for computers to understand and interpret written text. To automatically produce the various question kinds, multiple NLP models have been constructed and work on these models has also been done in many languages, including Tamil. Numerous researchers have presented their work in the domain of AQG, which is still being studied to increase accuracy. According to our proposed methodology, the more established rule-based approach is more reliable and is updated in accordance with trends.

Table 2. The Stages of Question Generation

S.No	Stages	FIB without Options	FIB with Options	True/False QG	Subjective QG
1	Text Cleaning				
2	Sentence Selection				
3	Keyword Selection				
4	Selection of Distractors /Hyponyms				
5	Question Pattern Verification				
6	Post Processing				

Several reviews that have been written have focused on the literature on AQG. The reviews that came later covered the works that had been published after late 2014, while the initial reviews only included works that had been published until that time. Beginning with analyses of question creation for educational purposes, MCQ-style questions derived from ontologies[10], and comparisons of all question generating methods, these reviews then went on to discuss ontologies.

For reading comprehension tests, the Extractive True/False Question Generation method (ETFQ)[3] automatically creates true/false questions from a given passage. There aren't many papers on the True/False Question Generation Approach, which anticipates True or False as the response. The Educational Purpose based Automatic True/False Question consists of a template-based framework that uses various NLP techniques to test the passage's specific knowledge and a generative framework that uses a novel masking-and-infilling strategy to generate more complex and flexible questions.

Fill-In-The-Blank with option questions, or ECQG[10], fall under the genre of electronic assessment-based generating, where a sentence is supplied with one blank and four possibilities to complete it. These questions have recently attracted study interest. The EOQG or Fill-In-The-Blank question with out options differs from the previous Fill-In-The-Blank questions by having a hint instead of alternatives for getting the answer. ECQG are effectively used in active learning, information and communication technology-based education, and intelligent tutoring system for the assessment of learner's content knowledge with slight modifications[11].

Aarish et al, [12] in his work included classification categories and short subjective questions without any manual intervention provides a thorough understanding of this. It effectively supports both the Natural Language Processing community and educational institutions so they can manage the full process of conducting online exams. As a result, the evaluation[13] of subjective questions is a crucial component of the overall system that strives to achieve complete automation in the administration of exams using objective question banks.

## 2. PROPOSED METHODOLOGY

We propose a novel method for keyword extraction using Tamil Wikipedia for automatic question generation. This keyword selection process involves two key ideas: the first is the importance of the keyword based on the content it has in a web-based search for the keyword that is similar to a web-based search and the second is emphasis on keywords that cover the majority of the text corpus because evenly distributed keywords are given more weight.

The proposed method for keyword selection is shown in Figure.3, as the first step, text is cleaned up by removing unwanted text from the corpus and for text processing. The text corpus is then transformed into tokens. By creating a dictionary with a tamil token as the key and an english word as the value—a mapping dictionary—which produce a token that has an equivalent English meaning is shown in Figure.4. In order to use pre-trained machine or deep learning model, there is a need for mapping technique, and which is shown in Figure.4, as the solution to this problem.

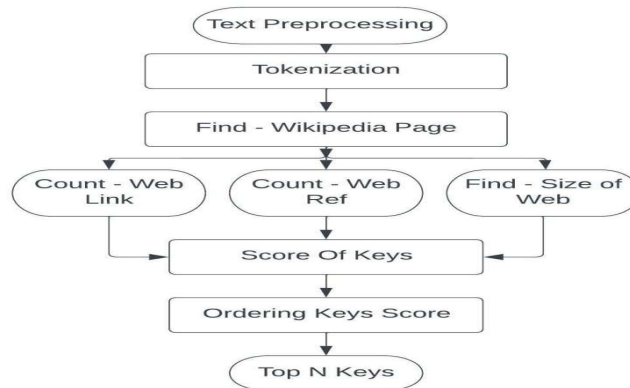


Figure 3. Proposed Method Workflow

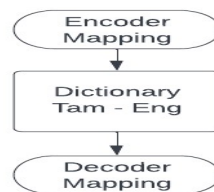


Figure 4. Keyword Mapping Dictionary for Tamil to English

Following tokenization, the occurrence of the tokens searched in web-based database, specifically looking for a wikipedia page for the token word. If a wiki entry is found, then that token is considered as the keyword. The number of links, references, and total content that are available on a keyword's wikipedia page is considered further to analyze. Following the collection of this information, a formula is used to calculate a score for the terms shown in Figure.5. The keywords are then ranked according to this score, and the top n keywords are finally selected based on the keyword score.

$$KwS = (WNL + WNR + WCS) / (n * 100)$$

$KwS$  = Keyword Score  
 $WNL$  = No. of Web links available.  
 $WNR$  = No. of References available.  
 $WCS$  = Size of the available contents in Mb.  
 $n$  = No. of factors deciding the score.

Figure 5. Proposed Keyword Score Formula

Depending on the situation or the domain, the parameters can also be added in the formula to obtain the best possible keywords. Better keywords will undoubtedly guarantee higher-quality question generation after they are chosen, as this is the most crucial step in the automatic question generation process.

### 3. RESULTS AND ANALYSIS

The proposed method for choosing keywords from the entire corpus as well as from each sentence is to use keyword score from the keyword selection formula. Every textual part is considered equally for selecting keywords by the suggested method; for instance, keywords are chosen from each sentence and given equal weight in the corpus. Our method performs better in all types of question generation concepts, and it produce good results.

#### 3.1. Selected Keywords from the Proposed Method

1: Context: தமிழ் தமிழர்களினதும் தமிழ் பேசும் பலரின் தாய்மொழி ஆகும்.

Keywords: ['தமிழ்', 'தாய்மொழி', 'ஆகும்', 'பேசும்']

2: Context: பண்டைத்தமிழில் எழுதப்பட்ட குறிப்பிடத்தக்க காப்பியம், கி.மு 200 முதல் கி.பி 200 காலப்பகுதியைச் சேர்ந்த சிலப்பதிகாரம் ஆகும்.

Keywords: ['தமிழில்', '200', 'சேர்ந்த', 'எழுதப்பட்ட', 'சிலப்பதிகாரம்', 'முதல்', 'ஆகும்']

Table 3. F1-Score Calculation Keywords

S.No	Variable Name	No.of Keys Available	Meaning of the Process
1	TokenKeys	160	Available Tokens in the text.
2	ManualKeys	45	Manually chosen keywords.
3	PredKeys	125	Keywords from the Proposed method as the keyword prediction.
4	PredKeysR	34	Keywords from the Proposed method as the Right keyword prediction.
5	PredKeysW	88	Keywords from the Proposed method as the Wrong keyword prediction.
6	MissedKeys	23	Keywords Missed from the Proposed method.
7	MissdKeysR	21	Keywords Missed rightly from the Proposed method.
8	MissdKeysW	2	Keywords Missed wrongly from the Proposed method.

When creating questions from a text corpus based on Tamil, the proposed keyword selection strategy works well and produces decent results as given in Table 3, as the proposed method for the prediction of keywords results. As an example, In Table 4, it shows the performance of the keyword selection methods Yake, KeyBERT, PKE, and Rake in terms of accuracy, precision, recall, and F1 score[14]. The performance of the proposed method greatly outperforms the alternatives shown in Figure.6.

Table 4. The Performance of Keyword Selection

Method	Accuracy	Precision	Recall	F1-Score
Yake	0.136000	0.188889	0.377778	0.251852
Keybert	0.177570	0.177570	0.513514	0.263889
PKE	0.171717	0.171717	0.459459	0.250000
Rake	0.182540	0.182540	0.621622	0.265477
Proposed	0.272000	0.272000	0.755556	0.400000

How often the machine learning model properly predicted is indicated by accuracy. Precision measures how well a model predicts a given category. How frequently the model was able to identify a particular category is indicated by recall. F1 is the result of combining recall and precision scores. Therefore, accuracy and F1 score have been taken into consideration as two key aspects in determining the quality of the question.

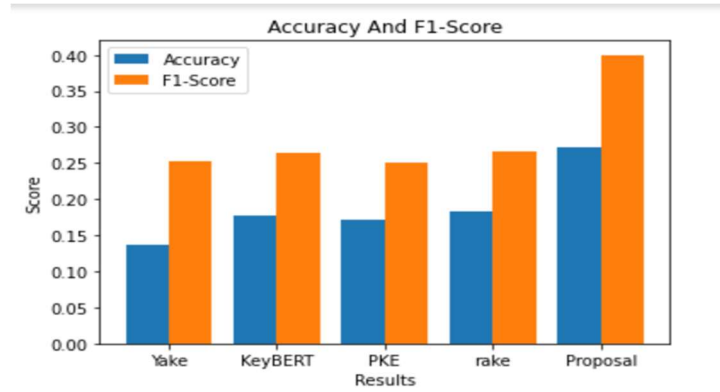


Figure 6. Keyword Accuracy and F1-Score Graph

### 3.2. Proposed Method for Various Question Generation Types

A. QG - True/False: திருக்குறள் ஏறத்தாழ 2,000 ஆண்டுகளுக்கு முன் இயற்றப்படவில்லை.

The TF-QG is a template-based framework with a primary level of question generation and an answer assessment level that intends to test the specific knowledge in the passage by utilising various NLP approaches. When a term is used negatively, we can either alter it or leave it alone depending on whether the sentence is true or false.

B. QG – Without Options: திருக்குறள் ஏறத்தாழ ----- ஆண்டுகளுக்கு முன் இயற்றப்பட்டது.

A Fill in the Blank question[15] has a missing term, sentence, or paragraph, and a blank space in its place. Questions with several blanks can also be built, and the difficulty level for answering these questions varies depending on the type of question. The first sort of fill-in-the-blanks has no alternatives[16], whereas the second has a hint[1], from which, the right one can be selected as the answer. In the given example above, 2000 is the answer and a hint can be given as any form like prior to Jesus' birth.

C. QG - With Options: ----- ஏறத்தாழ 2,000 ஆண்டுகளுக்கு முன் இயற்றப்பட்டது.

- a. திருவிவிலியம் b. சிலப்பதிகாரம் c. குயில் பாட்டு d. திருக்குறள்

The distractors([17],[18],[19]) play a significant part in this form of question production and greatly influence the standard of FIB questions. Fill in the blanks with options will contain some distractors. In the example given below is expected திருக்குறள் as the answer and we may add some distractors like திருவிவிலியம், பொன்னியின் செல்வன், குயில் பாட்டு, சிலப்பதிகாரம் and so on to make the distractions. If the examinee is not sufficiently perplexed by the distractions, he chooses the right response with ease.

D. Subjective Question Generation Using Pre-trained T5 – Model:

The T5 model [20] receives context and keyword as the input and produce questions as the outputs in the automatic question generation concept.

Here we have taken some samples from T5 model outputs as the representation of the results.

Eg.1 Context: தமிழ், உலகில் உள்ள முதன்மையான மொழிகளில் ஒன்றும் செம்மொழியும் ஆகும்.

Keyword: தமிழ்

Subjective Question: உலகின் முதல் மொழிகளில் ஒன்று எது?

Eg.2 Context: இந்தியாவில் கிடைத்துள்ள ஏறத்தாழ 1,00,000 கல்வெட்டுகளில் தொல்லெழுத்துப்பதிவுகளில் 60,000 இதில் ஏறத்தாழ 95 விழுக்காடு தமிழில் உள்ளன.

Keyword: தமிழ்

Subjective Question: எந்த மொழியில் அதிக கல்வெட்டுகள் உள்ளன?

Eg.3 Context: 2005-இல் அகழ்ந்தெடுக்கப்பட்ட சான்றுகள், தமிழ் எழுத்து மொழியைக் கி.மு.600-ஆம் ஆண்டிற்கும் முன்னள்ளியுள்ள.

Keyword: தமிழ்

Subjective Question: அகழ்வாராய்ச்சியில் என்ன மொழி பயன்படுத்தப்பட்டது?

According to the author (Wang Yulong, 2003)[21], the following properties should be present in a sentence to be considered perfectly obvious[22]: accuracy, unity, clarity, coherence, and emphasis. The Proposed model generates questions, which are then submitted to an expert team for quality control.

It performed better on the Likert scale is shown in Figure.7, scoring an average of 6.18 out of 10 according to the report.

If the score falls within the categories of (0 - 2.5), (2.5 - 5.0), (5.0 - 7.5), and (7.5 - 10) accordingly, it is referred to as Poor, Average, Good, and Very Good.

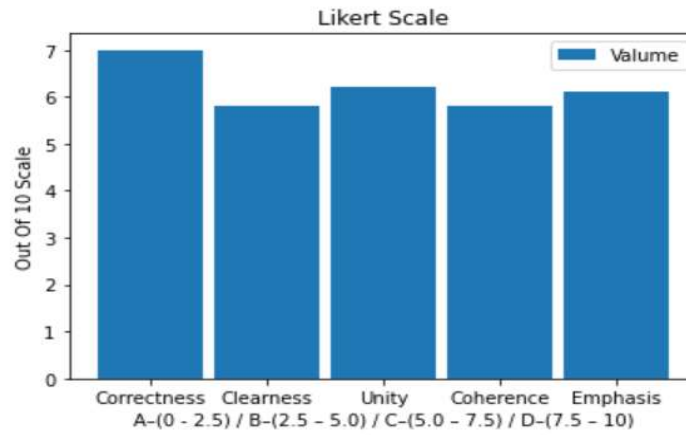


Figure 7. Likert Score of Generated Question

#### 4. CONCLUSION

The questions in the results representation are generated by a trained T5 model, which was initially trained on question and answer pairs in the English language. The extracted keyword and the context are inputs to the T5 transformer, which generates the questions. To create questions in Tamil, we employ a dictionary based mapping technique, which associates each Tamil word with its English equivalent. By the way, Tamil text can be used as the input for the T5 model. Finally we get the question in Tamil as the results after performing the same mapping of the T5 model resultant text. Keywords are crucial in the creation of extracting questions and the suggested centrality-based keyword extraction method outperforms existing methods noticeably in terms of Likert Scale. The construction of a corpus of Tamil questions that includes context, a keyword, and the questions that are framed based on the keyword. The Transformer model's question generation would be of better quality if training and testing are done through Tamil corpus.

#### REFERENCES

- [1] B. Das, "Factual open cloze question generation for assessment of learner's knowledge," 2017, doi: 10.1186/s41239-017-0060-3.
- [2] H. Ali, Y. Chali, and S. a. Hasan, "Automatic Question Generation from Sentences," *Proceedings of TALN 2010*, pp. 19-23, 2010.
- [3] B. Zou, P. Li, L. Pan, and A. T. Aw, "Automatic True/False Question Generation for Educational Purpose," *BEA 2022 - 17th Workshop on Innovative Use of NLP for Building Educational Applications, Proceedings*, no. Bea, pp. 61-70, 2022, doi: 10.18653/v1/2022.bea-1.10.

- [4] D. R. Ch, "From Text : A Survey," vol. 13, no. 1, pp. 14–25, 2020.
- [5] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," *Text Mining: Applications and Theory*, pp. 1–20, 2010, doi: 10.1002/9780470689646.ch1.
- [6] S. Siddiqi and A. Sharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review," *Int J Comput Appl*, vol. 109, no. 2, pp. 18–23, Jan. 2015, doi: 10.5120/19161-0607.
- [7] U. Singh, "A Comparison of Single Keyword Pattern Matching Algorithms," *International Journal of Engineering and Techniques*, vol. 3, [Online]. Available: <http://www.ijetjournal.org>
- [8] N. Firoozeh, A. Nazarenko, F. Alizon, and B. Daille, "Keyword extraction: Issues and methods," *Nat Lang Eng*, vol. 26, no. 3, pp. 259–291, 2020, doi: 10.1017/S1351324919000457.
- [9] M. Last and G. Danon, "Automatic question generation," *Wiley Interdiscip Rev Data Min KnowlDiscov*, vol. 10, no. 6, pp. 1–11, 2020, doi: 10.1002/widm.1382.
- [10] M. Agarwal and L. Technology, "Cloze and Open Cloze Question Generation Systems and their Evaluation Guidelines," no. July, 2012.
- [11] M. Divate and A. Salgaonkar, "Automatic question generation approaches and evaluation techniques," *Curr Sci*, vol. 113, no. 9, pp. 1683–1691, 2017, doi: 10.18520/cs/v113/i09/1683-1691.
- [12] A. Chhabra and S. Mohania, "Model for Short Subjective Questions " 2021.
- [13] M. Blšták and V. Rozinajová, "Automatic question generation based on sentence structure analysis using machine learning approach," *Nat Lang Eng*, vol. 28, no. 4, pp. 487–517, 2022, doi: 10.1017/S1351324921000139.
- [14] J. Li, "A comparative study of keyword extraction algorithms for English texts," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 808–815, 2021, doi: 10.1515/jisys-2021-0040.
- [15] P. Sirithumgul, P. Prasertsilp, and L. Olman, "An Algorithm for Generating Gap-Fill Multiple Choice Questions of an Expert System," *Proceedings of the 55th Hawaii International Conference on System Sciences*, 2022, doi: 10.24251/hicss.2022.901.
- [16] R. Correia, J. Baptista, M. Eskenazi, and N. Mamede, "Automatic generation of Cloze question stems," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7243 LNAI, pp. 168–178, 2012, doi: 10.1007/978-3-642-28885-2\_19.
- [17] J. Shin, Q. Guo, and M. J. Gierl, "Multiple-choice item distractor development using topic modeling approaches," *Front Psychol*, vol. 10, no. APR, pp. 1–14, 2019, doi: 10.3389/fpsyg.2019.00825.
- [18] R. Patra and S. Kumar, "A hybrid approach for automatic generation of named entity distractors for multiple choice questions," 2018.
- [19] A. S. Bhatia, M. Kirti, and S. K. Saha, "Automatic generation of multiple choice questions using wikipedia," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8251 LNCS, no. 2008, pp. 733–738, 2013, doi: 10.1007/978-3-642-45062-4\_104.
- [20] Ramsri Goutham Golla - T5 Transformer Model for Automatic question Generation as Result Presentation.
- [21] Y. Wang Yulong. (2003). College Guide to Writing, Qingdao: Ocean University of China.
- [22] X. Yu, "A Brief Study on the Qualities of an Effective Sentence," *Journal of Language Teaching and Research*, vol. 8, no. 4, p. 801, 2017, doi: 10.17507/jltr.0804.21.

## Local Binary Pattern based Feature extraction and Recognition of Ancient Tamil Palm Leaf Manuscripts Characters using Neural Networks

S Uma Maheswari<sup>[1]</sup> and P Uma Maheswari<sup>[2]</sup>

<sup>1,2</sup> College of Engineering, Guindy, Anna University, Chennai 600 025, INDIA

---

### ABSTRACT (10 PT)

Character Recognition from Palm leaf Tamil Inscriptions is a challenging task that involves a high degree of image preprocessing, segmentation, and recognition. Tamil palm leaf manuscripts are one of the greatest cultural heritage of India that has traditional medicinal practices. The work concentrates on steps taken in digitizing and preserving the PLMs that are more fragile and have a high degree of deterioration on aging. Though there are few high efficiencies online Handwritten Tamil OCRs proved 90% accuracy for modern Tamil alphabet, but they perform with less than 20% accuracy while dealing palm leaf characters. Since the background and foreground colors are similar in PLM scripts and the majority of character strokes are in ancient form. Around 500 palm leaf manuscript images were obtained from State Department of Archeology to accomplish this task. All the PLM images are converted into grey space and binarized to obtain a clear binary image. The images are filtered to remove impurities and normalized to increase the contrast. Then the text lines are segmented first into individual character images and then features are extracted using Local Binary Pattern operator. The feature map generated from LBP will be the input for Artificial Neural Networks based character recognition in which a reasonable accuracy was obtained.

---

### Keywords:

- A Character Recognition
- B Tamil Inscriptions
- C Cultural Heritage
- D Archeology
- E Neural Networks

---

### Corresponding Author:

S Uma Maheswari

College of Engineering, Guindy, Anna University, Chennai 600 025, INDIA

---

## 1. INTRODUCTION

### 1.1 Palm Leaf manuscripts

Tamil is one of the ancient languages that has been well developed and has various mines of heritage knowledge that are inscribed in different mediums like stones, mud plates, copper plates, and palm leaf. In earlier centuries, the traditional Medicinal System (TMS) was practiced by all till the midst of the 21st century and is in practice (due to the emergence of Allopathic medicine) at a lower rate in India, specifically in Tamil Nadu to prevent disease occurrence for the hale and hearty living. As Tamil is an ancient language and its alphabet has evolved over centuries since 3<sup>rd</sup> BC, the traditional medicine information that is found in palm leaves and the like are of ancient form. Since the native people are practicing modern Tamil scripts, reading and understanding the content of ancient Tamil scripts would be a more difficult task which require expert epigraphist. In present days, availability of such epigraphist is very less and the demand will prevail more in future. Around 20% manuscripts that are preserved under state department have been digitized and published but still there is huge volume of palm leaf manuscripts are to be digitized and conserved in modern Tamil. State and federal government has taken many initiatives to digitize and translate the heritage contents, but still technological solutions have to be paid towards to this task for faster, efficient and robust conservation of all such documents in lesser time in such a way to publicly accessible and functional repository. Hence it is a



mandate to devise a novel system which can recognize the ancient Tamil characters present at palm leaf manuscripts. This work focuses on palm leaf manuscript volumes written by the sage Agastya who lived during 7<sup>th</sup> century. There are 105 volumes of each containing 100 to 1500 leaves pertaining to Agastya medicinal manuscripts available at State Library of Tamil Nādu, India. The indigenous siddha medicine and medical procedure content of those manuscripts has high potential to cure various diseases. The current work has involved the work of Agastya.

## 1.2 About Tamil Language

There are 12 vowels and 18 consonants in the Tamil language and there are about 216 compound characters formulated by the combination of vowels and consonants as given in Figure 1. Grantha characters such as (ஐ, ஓ, ஷ, ஹ, ஸ், ஸ்ரீ) are also present in the Tamil palm leaf manuscripts. The numerals and metrics are represented by haracters as given in Figure 1:

	a	ā	i	ī	u	ū	e	ē	ai	o	ō	au
	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஓ	ஔ	
k	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ
ṅ	ங	ஙா	ஙி	ஙீ	ஙு	ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ
c	ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
ñ	ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
t	ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ
n	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
t	த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
n	ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ
p	ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ
m	ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ
y	ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ
r	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ
l	ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	லௌ
v	வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ
ḷ	ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ
ḻ	ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ
r	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
n	ன	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ

க	உ	ந	ச	ரு	கா	எ	அ	கூ	ஓ
1	2	3	4	5	6	7	8	9	0

உ - day   மீ - month   னு - year   பூ - debit   ள - time,   ளை - credit   ஷ - as above   ரூ - rupee  
 ஸ் - numeral   வ - quantity

Figure 1.2: Tamil Characters and numerals found in PLMs

The main challenge in recognition of character in Tamil palm leaf manuscripts is that the palm leaves are brittle the dot over the consonants are not written as they may break the palm leaves which gives a good chance of confusion to determine whether the character is a consonant or a (vowel+consonant) character. Thus, if the dot on the letter is not kept, there are chances of confusion between the consonants and the (vowel+consonant) characters. In the Tamil language, there are 18 consonants that can be misinterpreted with 18 other characters which can also be determined by either writing rules by the spelling of the word but they are tedious. To make the script even more complex there are no full stops(no punctuations) in the script to determine the words is also challenging.

## 1.3 Tamil Palm leaf Manuscripts

It is prevalent that most of the existing character segmentation and recognition methods are designed for specific languages either printed or handwritten. It is observed that the recognition rate is much influenced by feature extraction, and to recognize characters like inscriptions from challenging writing medium with varying style and size is novel and yet to be developed. Hence, the focus of the proposed research work is to segment and recognize the characters from palm leaf manuscript images. PLM consonants do not have a dot modifier on top. Refer to Figure 2, there is no dot on top of any constants.



Figure 1.3 : Sample Palm script test lines

Original Palm leaf manuscripts that are preserved under controlled way is scanned in a high resolution scanner and the scanned image will be stored in png format at PLM dataset. In image enhancement phase, the low quality palm leaf image is processed in various stages to obtain well discriminated text in uniform stroke width and size in order to effectively segment the characters present for feature extraction process. Once the segmented characters are obtained, the structural features like curves, loops, line and the like are extracted and feature vector is generated. Each and every attribute of feature vector will be the input variable for ANN for character recognition.

The greatest challenge in character recognition is that there are several composite characters in tamil language. As far as Tamil language is considered, it has various combinations which gives connected characters when combined. The Tamil dialect characters are cursive with some of them having additional curves and loops. The other challenges like binding holes in different positions and deteriorated leaves need special attention. The limitations and challenges faced during this work will be addressed in future works.



Figure 1.4 : A Full Length PLMs structure

## 2. RELATED WORK

Explaining research chronological, including research design, research procedure (in the form of algorithms, Pseudocode or other), how to test and data acquisition [1]-[3]. The description of the course of research should be supported references, so the explanation can be accepted scientifically [2], [4].

Tables and Figures are presented center, as shown below and cited in the manuscript.

Table 1. The Performance of ...		
Variable	Speed (rpm)	Power (kW)
x	10	8.6
y	15	12.4
z	20	15.3

Deterioration of the Palm Leaf Manuscripts poses a threat in the degradation of the treasured information and thus, rendering Digitization[2,3], an important task for preserving that information. Image Restoration works have been prosecuted in various languages such as Tamil, Malayalam, Telugu and Arabic. The manuscripts were subjected to a number of pre-processing techniques as the primal step for its restoration. Some of the pre-processing techniques [1] are Edge Detection, Contrast Enhancement, Segmentation, etc. Palm Leaf Manuscripts are converted to Grey Scale through the process of Binarization[13]. Through existing Image

Pre-processing techniques such as Noise Removal, Thresholding, Smoothing, etc., the initial phase of Data Processing with images containing High Contrast-Background Separable-Palm Leaf Manuscripts will be obtained. Insights obtained from existing noise removal algorithms are as follows: Binarizing and estimating the optimum parameter value [5] from a subset parameter range, Filtering the noise [6] based on its standard deviation, Development of Automated Document Image Binarization [7] using RoBO (Robust Bayesian Optimization). Apart from the afore-mentioned, Threshold plays a pivotal role in denoising images and thus selection of Threshold becomes vital. This Threshold Selection mechanism can be achieved through Gray Level Histogram[8]. Using Local Otsu Thresholding and Sobel operator-based Image Gradient approximations [4], the proposed algorithm gives the best results for noise removal from palm leaf manuscripts both visually and in terms of PSNR (Peak Signal to Noise Ratio) and MSE (Mean Squared Error) values with less complexity and overhead. Performance of the said approach depends vastly on the Threshold value used for filtering the given Palm Leaf Manuscripts.

The next task of our Research would be to segment each Character, which is a tedious task, as there are a number of issues that may arise while segmenting Characters [14] in Palm Leaf Manuscripts which includes: Difference in the Skew angle between the lines, Overlapping Characters present in the same line and also overlapping that arises between adjacent lines. In General, Character Segmentation through the etched lines can be classified into five categories [15, 16] namely Projection-based method, Smearing method, Grouping method, Hough-based method, and Repulsive-Attractive network (RA-network) method. Using existing methods, we will strive to devise an optimum mechanism that generalizes well for our dataset. – In this particular research setting, the next task in hand will be to extract the features that represent the handwriting style. This process of extracting feature vectors from handwriting style constitutes Feature Extraction Method. Two common groups of feature extraction methods (textural and grapheme-based) are explored [17]. The Feature Extraction Methodology will be based on the idea that the handwriting style of the general population evolves over time. Character Recognition Methodology has taken a drastic turn as the Handwritten Characters inherently paved way for OCR Systems, in which, there exists automated recognition of characters present in digitalized format [18] like paper documents, PDF files and character images. In our Research, it is, therefore, mandatory to create a High-Performance OCR Engine using Machine Learning Technology to analyse digital images and subdividing them into lines and further into characters. Challenges in Character Recognition [9]: Existence of Marginally low contrast between foreground (Etched Letters) and background (Palm Leaf) of the given image and One of the major challenges faced by any existing gray scale conversion algorithms, is that, there exists an unevenness in the rendering of Background color. There exists an ambiguity amongst character symbols if they touch or overlap each other and thus becomes quite a challenge during the task of Character Recognition. Other works on Character Recognition (Other Languages excluding Tamil): Investigation of Character Ligatures of ancient Greek Manuscripts [10], Using SOMs (Self-Organizing Maps), Character Recognition of Lanna Script (an obsolete script of Thailand) [11], Recognition of isolated handwritten Balinese Characters from Palm Leaf Manuscripts [12], The penultimate step of Word Recognition consumes a lot of time and is dependent on the judgement of a Human Entity. Reducing the demand for Experts, we strive to devise a Word Recognition methodology that would group characters to meaningful words using Deep Learning Technologies such as CNN, RNN and LSTMs, given a Lexicon Provider. This annotation of raw Manuscript Text with the most informative transcription will then be transliterated to Contemporary Tamil rendering as the Deliverable for this particular Research Problem.

க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ
ங	ஙா	ஙி	ஙீ	ஙு	ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ
ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ
ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ
ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ
ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ
ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ
ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ
ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	லௌ
வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ
ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ
ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ
ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
ன்	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ

Figure 1. Effects of selecting different switching under dynamic condition

### 3. PROPOSED METHODOLOGY

Around 500 Sage Agastiyar's medicinal palm leaf manuscript images were obtained from State Department of Archeology to this work. Each leaf contains around 7 to 10 lines of text. Approximately there are 100 to 150 characters present in each leaf. Hence the data set is adequate for machine learning. The overall methodology is depicted in Figure 4.

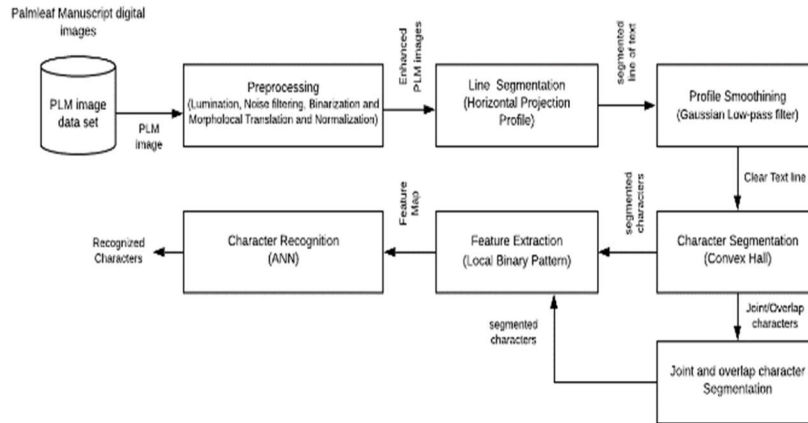


Figure 3 : Proposed System Architecture

#### 3.1 Image Acquisition:

Identified palm leaf manuscripts will be scanned in high resolution scanner with 600 dpi and scanned image is stored in .png format. Such a way the PLM image dataset is created. A sample image is given in Figure 2.



Figure 3.1 : Original PLM image

#### 3.2 Image Enhancement:

The scanned PLM image is first converted into grayscale, undergo filters for noise removal, and then binarized using Adaptive thresholding technique in order to discriminate background foreground subtraction. In adaptive thresholding, threshold is calculated for each pixel in the image  $I_{i,j}$ . If the pixel value is below the threshold constant  $T$ ,  $I_{i,j} < T$ , then that pixel is set to the background value; otherwise, it set to foreground value. Equation (1) depicts adaptive thresholding.

$$T(x,y) = WA(x,y) - C \dots \dots \dots (1)$$

$T(x,y)$  - threshold at pixel location  $(x,y)$

$WA$  - weighted average value calculated for each pixel

After thresholding, the foreground characters present in the PLM is discriminated from its background. This process is called binarization. Further, morphological operations of erosion and dilation is performed to fill any gap in the stroke and to eliminate extra projections in the stroke line.

### Algorithm PLM Image Enhance (PLM Image Dataset)

Input : PLM Images

Output : Enhanced PLM Images

Begin

```

for all PLM Images PLMi
    crop PLMi to eliminate unwanted boundary part
    convert PLMi into gray scale image gPLMi
    gPLMi = luminous (PLMi)
    binarize gPLMi
    bPLMi = Adaptive Thresholding (gPLMi)
    remove noise from Binarized image using Median Filter
    fPLMi = Median filter (bPLMi)
    mPLMi = morphologicaltranslation (fPLMi)
    Store mPLMi in preprocessed dataset
end

```

End Algorithm PLM Image Enhance.

### 3.2 Image Normalization:

Min-Max normalization depicted in eqn (2) is done to ensure uniform distribution of image pixels that makes convergence faster while training the network.

$$N_{mPLMi} = \frac{(MPPi - Min)_{newMax} - newMin + newMin + newMax}{Max - M} \quad (3)$$

### 3.3 Segmentation:

Firstly, the whole image is divided into N parts, and every part retain the same height, Width is 1/N, the text image is then processed by a sequence of operations, and then the upper and lower bounds of the region are determined by the projection scanning in the horizontal direction, from the top to the bottom line by line scanning, and instantaneously obtain the black pixels of each scan line, the number of Pixel black spot accounted. Partition structure Horizontal projection profile of the preprocessed PLM image mPLMi is calculated as given in equation (4). Once after lines are segmented, then convex hull is applied to segment each and every character from the segmented line.

$$S_h = \frac{\sum_{j=ystart}^{yend} \sum_{i=xstart}^{xend-1} (pixel(i,j) + pixel(i+1,j))}{subwidth * subheight} \quad (4)$$

This projection function is smoothed using Gaussian low-pass filtering to reduce noise and remove false projection lines.

Algorithm Segmentation

Input : Normalized image

Output : Segmented characters

```

1. Begin
2. for all mPLMi images
3.     create horizontal projection profile Hp
4.     hp =
5.     do profile smoothing by Gaussian low-pass filter
6.     S(hp) =
7.     Create vertical projection profile Vp
       Vp = contour convex hull (---)
       do profile smoothing by Gaussian low-pass filter
       S(Vp) = Gaussian low pass filter (---)

```

8. end
9. End.

### 3.4 Local Binary Pattern based Feature Extraction:

The performance of the recognition system relies much on the quality of feature extracted. Feature extraction is done using Local Binary Pattern (LBP) algorithm which results in 8-bit binary array in which the character image  $SC_i$  of size  $N \times N$  is divided into fixed size small cells of  $m \times m$  pixels each  $\{p_1, p_2, \dots, p_m\}$  and each pixel  $p_i$  is compared with its 8 neighbours along a circle in counter-clockwise.

$$LBPPR = \sum p - 1S(Gp - Gc)$$

$$P=0$$

Where

$$S(x) = \{1 \text{ if } x \geq 0, 0, X < 0\}$$

The 8-digit binary number obtained is decimal converted. The histogram is generated for each cell and concatenated for entire window to give feature vector of 256 bits [0 ... 255] of the given character image  $SC_i$ . The feature vector for all the character images are generated in such a way and taken for neural network input nodes to classify character images for character recognition.

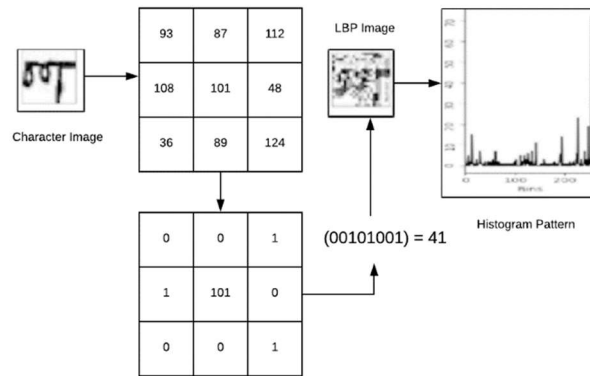


Figure 3.4 LBP pattern generation process

### Algorithm Character Feature Extraction

Input : Segmented Characters

Output: Feature for each image  $SC_i \in Sc$

1. Begin
2. for each image  $SC_i \in Sc$  do
3. Compute Local-Binary Pattern LBP
  - if  $p_i > \text{neighbour}(p_i)$ , then
  - LBP( $p_i$ ) = 0 ; otherwise,
  - LBP( $p_i$ ) = 1.
4. Generate feature vector  $f_i$
5. end for
6. End

### 3.5 Character Recognition

The character images feature vector  $\{f_i | i = 1 \text{ to total number of character images } SC_i\}$  were fed to an Artificial Neural Network for recognition of character they represent. The ANN model has input layer of  $N$  nodes such that  $\{x_1, x_2, \dots, x_N\}$  where  $N$  is the size of  $(f_i)$  and 21 fully connected hidden layers of varying size activated by Relu function at each layer. The output softmax layer is having number of nodes equal to the number of unique characters that are to be recognized. Gradient Decent based Backpropagation algorithm is used to train the network. In this case, frequently appeared characters in the PLM samples were trained initially to observe the model performance. The model is optimized with 'adam' optimizer to reduce the generalization error to the extend possible and to overcome overfitting problem. To further optimize and avoid vanishing gradient, L2 and dropout normalization is used with 25% dropout at all hidden layers. Nearly about 3000 character images, 2500 were used for training purpose and 500 were used for testing. The training accuracy and testing accuracy were measured along with loss at both the stages.



#### 4. RESULTS AND DISCUSSION

The original palm leaf Manuscripts are first undergone image processing. The resultant image after each stage of image processing is given in Figure 4. Three various filters were applied for removal of noise and median filter performs better as depicted in Figure 4e in which their performance is compared with the metric PSNR (Peak Noise to Signal Ratio).



Figure 4 a. PLM raw image

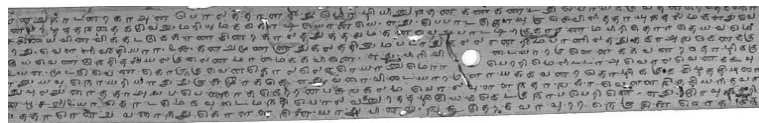


Figure 4 b. Grey scale converted PLM image

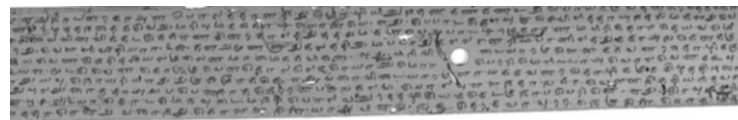


Figure 4 c. Blurred PLM image

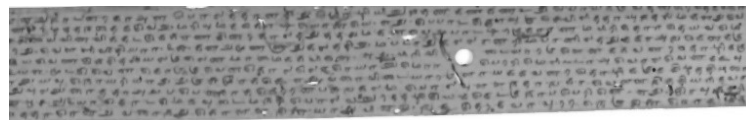


Figure 4 d. PLM image after Bilateral Filtering

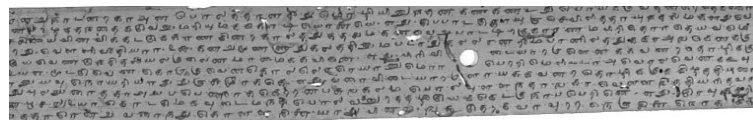


Figure 4 e. PLM image after Median Filtering

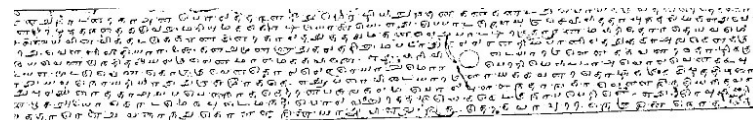


Figure 4 f. After Thresholding – Background discrimination

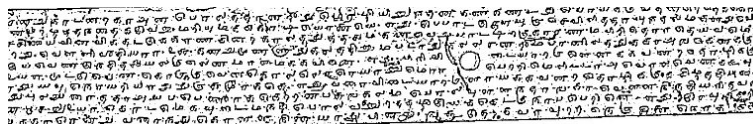


Figure 4 g. Normalized image

Once the noise removal is done, then the background discrimination was done to eliminate the background color and to highlight the characters present in the script using adaptive thresholding with block size = 35 and constant value = 40. Maximum threshold limit set is 255 and the resultant image is given in Figure 4f. Then the binarized image got normalized at 0 to 255 range. The efficiency of different filters are measured by Peak Signal to Noise Ratio value and the median filter that that gives higher PSNR value is taken for noise removal of all PLMs.



The segmentation process is applied to the normalized image using horizontal and vertical projection profile and the rate of segmentation is evaluated with the ratio between the total number of characters present in the manuscript and the number of characters successfully segmented. The segmentation results are given in Figure 5.

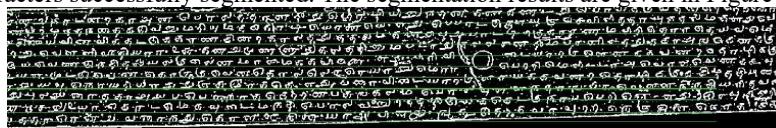


Figure 5a. : Horizontal Segmentation



Figure 5 b : Vertical Segmentation



The rate of Segmentation is measured in terms of ratio between the no.of Characters correctly segmented and the total number of characters present in PLMs.

Segmentation Rate = No.of True segment / Total no. of characters

The segmentation rate observed is 88% in this case. This rate is higher for clear character images and lower for broken characters present. The segmented character images after thinning were applied to LBP for feature extraction. The feature vector and its corresponding histogram is shown in Figure 5 for some character images as sample.

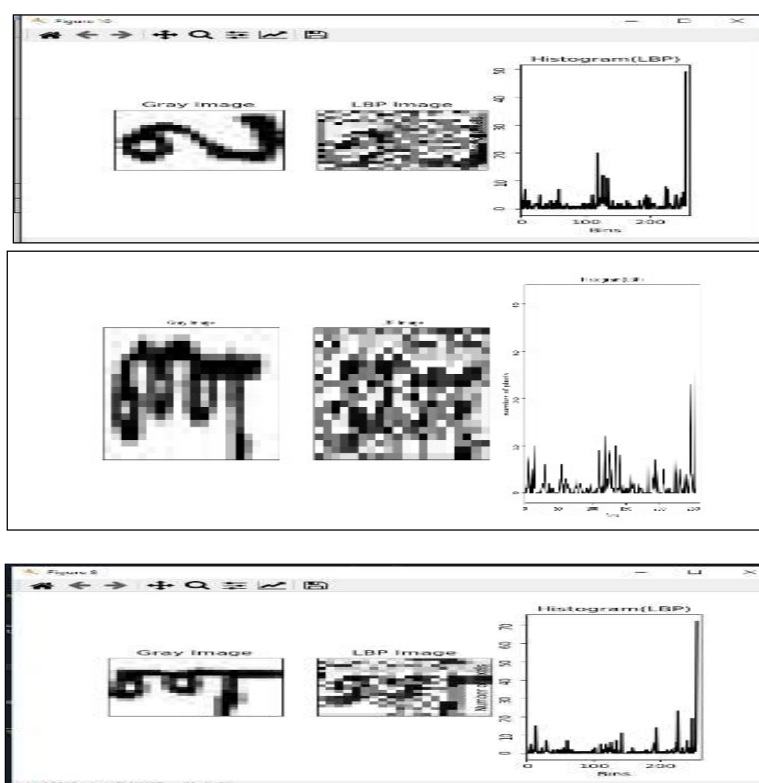




Figure 5 : Feature Extraction – LBP histogram

The ANN performance is evaluated in terms of accuracy. The model is executed for 4000 epochs and the training and testing accuracy and loss percentage observed in each epoch is plotted and shown in Figure 6.

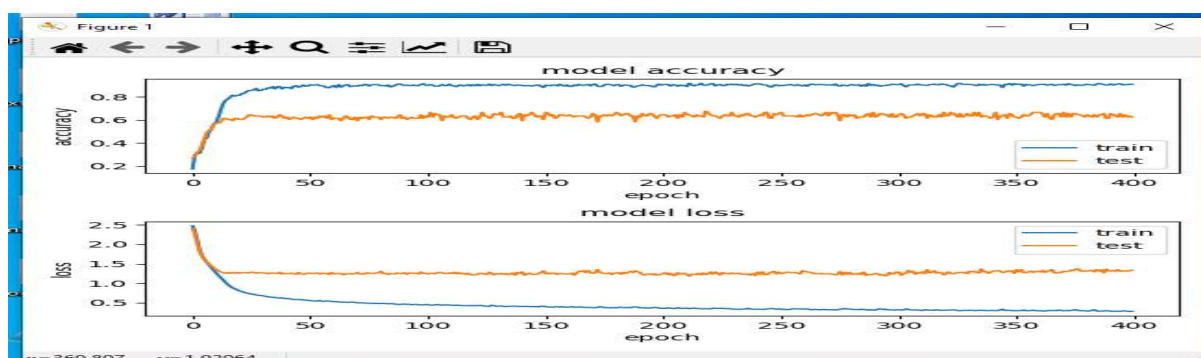


Figure 6 : Accuracy and Loss Curves

## 5. TESTING AND VALIDATION

The proposed model is tested with various test cases which covers the character images of clear, partially deteriorated, broken, overlapped, and other untrained alphabets and numerals. The other character test case also gives results as expected. Unexpected result comes for overlapping character test cases. The training accuracy of 78% is obtained whereas the validation accuracy is 68%. The generalization gap of 10% exists due to the test case contains unclear characters too.

## 6. CONCLUSION

On examining various filter to remove noise present in the PLM image, it is found that median filter has high PSNR value and concluded to be more suitable for PLM images. The proposed Local Binary Pattern (LBP) technique of feature extraction is state-of-art and have innovative idea to use Natural Language script images. It is a texture based approach that extracts character feature in an effective way and occupies less memory and time complexity while training. The ANN model build is giving more than 76% accuracy even early stage of epochs. So that on reduces time taken for training the model. The proposed model is trained and tested for vowels and Consonants and for other single characters. But does not deal the compound character sequences (கா, கெ, கொ and the like) which is really a challenging issue and our extended work is being carried out to resolve this challenge.

## 7. ACKNOWLEDGMENT

This work is non-monetarily supported by The Tamil Nadu state Department of Archaeology and Tamil Virtual Academy for permission to access PLM scripts achieved at state library and expert consultation to read letters that are in ancient forms.

## REFERENCES (10 PT)

- [1] [T.Jerry Alexander, S.Suresh Kumar, "Performance Evaluation of pre-processing techniques for historical Palm Leaf Manuscript image restoration", International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2022
- [2] G. Navirathan, "An Exploratory Review On Digitizing Palm Leaf Manuscripts," IJAR, vol. 7, no. 10, pp. 431–438, Oct. 2019, doi: 10.21474/IJAR01/9846.
- [3] Narenthiran R, Saravanan G, and Ramanujam K, "The Digitization of Palmleaf Manuscripts," in SALIS 2012, Tamil Nadu, 2012, pp. 457–462, doi: 10.13140/2.1.2016.6084.

- [4] Dhanya Sudarsan, Deepa Sankar, "A Novel approach for Denoising palm leaf manuscripts using Image Gradient approximations", International Conference on Electronics Communication and Aerospace Technology, 2019.
- [5] Shijian Lu, Bolan Su, and Chew Lim Tan. 2010. Document image binarization using background estimation and stroke edges. International journal on document analysis and recognition 13, 4 (2010), 303–314.
- [6] A. Buades, B. Coll, J. Morel, A non-local algorithm for image denoising, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, Citeseer, 2005, p. 60.
- [7] Ekta Vats, Anders Hast and Prashant Singh ,Automatic Document Image Binarization using Bayesian Optimization, HIP2017, November 10–11, 2017, Kyoto, Japan
- [8] N. Otsu, "A Threshold Selection Method from Gray-level Histogram," IEEE Trans. Systems Man Cybernet, vol. 9, pp. 62- 66, 1979
- [9] Paramasivam Muthan Eswaran, Dinesh Manib and Sabeenian Royappan Savarimuthu, "Recognizing Tamil Palm-Leaf Manuscript Characters Using Hybridized Human Perception Based Features", ICTACT Journal On Image And Video Processing, 2021.
- [10] K. Pratikakis, I. Petridis, S. Konidakis and S.J. Perantonis, "An Efficient Segmentation-Free Approach to Assist Old Greek Handwritten Manuscript OCR", Pattern Analysis and Applications, Vol. 8, No. 4, pp. 305-320, 2006.
- [11] Arit Thammano and Sakkayaphop Pravesjit, "Recognition of Archaic Lanna Handwritten Manuscripts using a Hybrid Bio-Inspired Algorithm", Memetic Computing, Vol. 7, No. 1, pp. 1-17, 2015.
- [12] Dewa Made Sri Arsa, Gusti Agung Ayu Putri, Remmy Zen and Stephane Bressan, "Isolated Handwritten Balinese Character Recognition from Palm Leaf Manuscripts with Residual Convolutional Neural Networks", International Conference on Knowledge and Systems Engineering, 2020.
- [13] Kaustubh Sadekar, Prajwal Singh, Shanmuganathan Raman , "HDIB1M - Handwritten Document Image Binarization 1 Million Dataset" , IEEE, 2021.
- [14] G. Louloudis, B. Gatosb, I. Pratikakisb, C. Halatsis , "Text line and word segmentation of handwritten documents" , Pattern Recognition, Elsevier, 2008.
- [15] Longlong Ma 1 , Congjun Long2 , Lijuan Duan 3,4,5, Xiqun Zhang3,4,5 , Yanxing Li3,4,5, And Quanchao Zhao "Segmentation And Recognition For Historical Tibetan Document Images", IEEE 2020
- [16] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: A survey," Int. J. Document Anal. Recognit., vol. 9, pp. 123–138, 2007.
- [17] Maruf A. Dhali , Camilo Nathan Jansen, Jan Willem de Wit, Lambert Schomaker, "Feature-extraction methods for historical manuscript dating based on writing style development", Pattern Recognition, Elsevier, 2020.
- [18] R. Babitha Lincy & R. Gayathri, "Optimally configured convolutional neural network for Tamil Handwritten Character Recognition", Multimedia Tools and Applications, Springer, 2020

# Named Entity Recognition for Gynecological domain in Tamil using Machine Learning Algorithms

M. Rajasekar<sup>1</sup>, Dr. Angelina Geetha<sup>2</sup>

Department of Computer Applications, Hindustan Institute of Technology and Science, Chennai, India  
Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Chennai, India

## ABSTRACT

Information Extraction has a number of sub domains to extract such useful information as related to the applications. The sub domains are Relation extraction, Named Entity Recognition, Audio Extraction and Terminology Extraction. NER Tagging is the process of classifying the POS tagged entities into specific groups such as, Person, Place, Organization, Time, and Date. This group entities will increase based on the application of NER Tagging. Named Entity Recognition is also called extraction of objects and identification of objects in pre-defined classes. A lot of applications are created using these NER Tagged entities such as Marketing analysis, Fame analysis of particular product, Sentiment analysis, Movie or media analysis. NER Tagging in regional level languages is complex and is a growing area. In this paper the novel Named Entity Recognition (NER) model is discussed for Tamil gynecological text data which are tagged with its specific POS Tags. The Morphological analysis and Part of speech tagging are already done as preprocessing tasks. The preprocessed and POS tagged datasets are used to extract the named entities in gynecological domain. There are numerous named entity recognition approaches already proposed by NLP researchers. In this research, we have concentrated to extract the gynecological named entities from Tamil text. In this work, the named entities are recognized with the help of machine learning methods Finite state automation method (FSA), Naive bayes classifier and Unidirectional Long-short term memory method (UD-LSTM).

The performance of this three machine learning approaches are evaluated by Precision, Recall and F1-Score method. Based on the performance of these methods, the Finite state automation model performs well and yields high accuracy (89%) for the given domain datasets when compared with the other models.

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

## Keywords:

- A Named Entity Recognition
- B Gynecological Domain
- C Machine Learning Algorithms
- D Tamil Gynecology

## Corresponding Author:

M. Rajasekar,  
Department of Computer Applications,  
Hindustan Institute of Technology and Science, Padur, Chennai, India  
Email: sekarca07@gmail.com

## 1 INTRODUCTION

Information Extraction has a number of sub domains to extract such useful information as related to the applications. The sub domains are Relation extraction, Named Entity Recognition, Audio Extraction and Terminology Extraction. In this chapter, the implementation of Named Entity Recognition (NER) is discussed for Tamil gynecological text data which are tagged with its specific POS Tags. NER Tagging is the process of classifying the POS tagged entities into specific groups such as, Person, Place, Organization, Time, and Date. This group entities will increase based on the application of NER Tagging. Named Entity Recognition is also called extraction of objects and identification of objects in pre-defined classes. A lot of applications are created using these NER Tagged entities such as Marketing

analysis, Fame analysis of particular product, Sentiment analysis, Movie or media analysis. NER Tagging in regional level languages is complex and is a growing area.

## **2. REVIEW OF LITERATURE**

Information was extracted from tourism domain text using a machine learning-based information extraction model<sup>[1]</sup>. Information was extracted from tourist databases using text classification and named entity recognition (NER) approaches. The data for the selected domain was extracted using machine learning technologies like SPACY and BERT. For Named Entity Recognition, it was discovered that the BERT model had the highest accuracy level (99%). For the Text categorization challenge, the accuracy of the BERT and SPACY models was around 95 % to 98 %.

A framework for named entity recognition in Malayalam been proposed<sup>[2]</sup>. It was a deep learning strategy in the NER system for Malayalam. In the case of NER, it was discovered that DL-based techniques greatly outperformed classic shallow-learning-based approaches. In terms of precision, recall, and F-measure, the deep learning strategy for NER system in Malayalam surpassed the others, with an F-score of 8.92 %.

The method was to use a rule-based Kannada named entity recognition system<sup>[3]</sup>. Pre-processing and entity recognition were the two processes. The sentences were obtained from various sources and divided into words during the pre-processing step. The Support Vector Machine model was used to recognise named entities in Kannada words. The regulation was created in order to identify the person's name, location, and classification. This method produced a good accuracy result of 89.32%.

Abinaya et al<sup>[4]</sup> proposed a unique solution for Named Entity Recognition (NER) in Tamil based on the Random Kitchen Sink algorithm. The process of identifying entities in relation to Names (NEs) from text is known as named entity recognition. It entails categorising and identifying preset categories such as people, places, and organisations. NER has conducted a variety of studies utilising various machine learning approaches for English and Indian languages. This study used the Random Kitchen Sink method, a supervised and statistical approach, to develop the NER system for Tamil text. They compared Support Vector Machine (SVM) and Conditional Random Field (CRF) performance. The NER system performed well, with RKS scoring 86.61 percent, CRF scoring 87.21 percent, and SVM scoring 81.62%. The performances in SVM and CRF were 86.06% and 87.20%, respectively, after increasing the size of the corpus.

Srinivasan et al.<sup>[5]</sup> built a new Automated Named Entity Recognition from Tamil Documents. This study used Supervised Learning to provide a novel approach for extracting Named Entities from Tamil language text. The NEs were extracted using a hybrid technique. The features that could be derived based on the Tamil language NEs were used in the Hybrid framework. The approach was assessed using 1028 documents from the standard FIRE corpus, yielding an F-Score of 83.54%.

It was invented the Automated Named Entity Recognition model<sup>[6]</sup>. The algorithm assigned the NER tags to the tokens in Tamil documents using the supervised learning method. The feature extraction from Tamil documents was designed using the Naive Bayes technique. There were 1028 Tamil documents in the corpus. Regex Feature extraction, Morphological Feature extraction, and Context Feature extraction were the three feature extraction models they employed. The Precision and Recall approach was used to assess the model. The last F-measure resulted in a score of 83.54%.

## **3. NAMED ENTITY EXTRACTION IN TAMIL LANGUAGE**

Due to semantic ambiguity identification and extraction of named entities from domain specific Tamil text documents become a difficult task. In Tamil language, sentence is in free word order and a single word can give different meanings. For example,

The Tamil sentence,

எங்கள் கல்லூரி கூடைப்பந்தாட்டத்தில் முதலிடம் பபற்றது -

1(Our College won the first prize in Basket ball Tournament)

எங்கள் கல்லூரிக்கு அடைச்சர் வந்தார் -2

(The minister came to our College)

In sentence 1, an entity கல்லூரி (college) is named as organization

Again in sentence 1, an entity கல்லூரி (college) is named as place

This shows a single entity 'college' gives two meanings in two different contexts.

Tamil language has free word order. Even though we change the order of words in a sentence, it gives the same meaning (Sangakaravelayuthan, R. *et al.*, 2019). Simple word order example is given below.

Table 1. Word Order categories in Tamil

S. No	Sentence	Word Order
1.	பசல்வன் பள்ளிக்கு பசல்கிறான் (Selan to School Going)	- S - O - V -
2.	பசல்வன் பசல்கிறான் பள்ளிக்கு (Selvan going to School)	- S - V - O -
3.	பசல்கிறான் பசல்வன் பள்ளிக்கு (Going selvan to school)	- V - S - O -
4.	பசல்கிறான் பள்ளிக்கு பசல்வன் (Going to school Selvan)	- V - O - S -
5.	பள்ளிக்கு பசல்கிறான் பசல்வன் (To school going Selvan)	- O - V - S -
6.	பள்ளிக்கு பசல்வன் பசல்கிறான் (To school selvan going)	- O - S - V -

When we change the place of each word, all the six sentences give the same meaning. This meaning can be identified by human being but a machine cannot identify the same meaning. If we give these sentences as input to a well trained machine, it cannot give the same meaning. To solve this complex problem by machine learning approach, the proposed model trained with pre named entity tagged corpus. General corpus is tagged as general pre-defined named entities. The following Table 4.2 provides the brief information about Named Entities.

Table 2. General Named Entity Tags

Tag. ID.	Level 1	Level 3	Instances
1	Person	Title	“திரு”
		Individual	“கண்ணன்”
		Group	“ஊணவர்கள்”
		Family Name	“நாயர்”, “ததவர்”
		Public Sector	“தபாக்குவரத்து கழகம்”
		Private Sector	“ஊணா”
		Government	இண்ணெட்டிரீஸ் “தபால்”
		Religious	அலுவலகம்”
		Charitable Trust	“பள்ளிவாசல்”
		Non-profit Organization	“விதவகானந்தா பதாண்டு
2	Organization	Association	நிறுவனம்”
		Media	“அருங்காட்சியகம்”
			“பகண்ணடி கிரிப்பகட் கிளப்”
			“டைம்ஸ் ஆப் இந்தியா”
			“கள்ளிப்பட்டி”
		Village	“தேதுடர்”
		City	“தல்லாகுளம்”
		Panjayat	பஞ்சாயத்து” “பபரியகுளம்”
		Taluk	தாலுகா” “சிவகங்கை”
		District	“ஊவட்டம்”
3	Location	Nation	“இந்தியா”
		Continent	“வை அபெரிக்கா”
		Door.No.	“பந.4/56-2”
		Street Name	“வைக்கு ஊசித்
		Area	பதரு”
		Pincode	“சிம்மைக்கல்”
		Water Bodies	“625540”
			“ஊஞ்சளாறு”

4	Materials	Fruit	"வாடழப்பழம்"
		Vegetable	"பவங்காயம்"
		Tree	"ஆலைரம்"
		Plant	"கத்தரிச்செடி"
		Furniture	"நாற்காலி"
		Electrical	"பதாடலக்காட்சி"
		Motor Vehicle	"தொட்டைர் டசக்கிள்"
		Metal	"இரும்பு"
		Medicine	"பாராசிட்டைல்"
		Chemical	"சல்பர் டை
5	Entertainment	Cloth	ஆக்டைடு"
			"கால்ச்சட்டை"
		Drama	"வள்ளிக்கந்தன் திருணை"
		Sports	"கபாடி"
		Cinema	"ஆயிரத்தில் ஒருவன்"
		Events	"திருணை"
6	Living Things	Conference	"பதற்காசிய
			ைாநாடு"
		Birds	"காகம்"
		Animals	"சிங்கம்"
7	Money Quantity Count	Reptiles	"முதல"
			"ரூபாய்"
			"கிதலாகிராம்"
			"5000"
8	Time Date Year Month Seconds Periods		"ைாதம்"
			"நாட்காட்டி"
			"திங்கள்"
			"வருமை"
			"வினாடி"
			"நிமிஷம்"
			"பநாடி"
			"பபாழுது"
			"நாழிடக"

### 3. NAMED ENTITIES RELATED TO MEDICAL FIELD

The proposed Information extraction system partially identifies the relation of the gynecological domain text. This issue leads to the initiative to develop named entities specifically for health issues

Table 5.3 Medical related Named Entities

Tag. ID.	Level 1	Level 3	Instances
----------	---------	---------	-----------



1	Organs	External	"கண்", "மூக்கு"
		Skeletal	"கால் எலும்பு"
		Integumentary	"தடல் முடி"
		Muscular	"தடசு"
		Respiratory	"நுடரயீரல்"
		Circulatory	"இதயம்"
		Urinary	"சிறுநீரகம்"
		Digestive	"சிறுகுடல்"
		Immune	"இரத்த பவளடளயணுக்கள்"
		Nervous	"மூடள நரம்பு"
2	Diseases	Endocrine	"டதராய்டு சுரப்பி"
		Reproductive	"கருப்ப", "சிடனமுட்டை"
		External	"தடச வலி"
		Internal	"உள் வலி"
		Cancerous	"இரத்தப்பற்றுதநாய்"
		Rheumatism	"வாதம்"
		Disorder in Blood Circulation	"இரத்த அழுவத்தம்", நீரிழிவா "அரிப்பு"
		Skin Disease	"படை"
3	Prevention	Heart Disease	"இரத்த அடைப்பு"
		Lung Disease	"ஆஸ்துதா"
		Reproductive Disease	"குழந்தயின்டை"
4	Disease Identification	Primary	"தடுப்புமுடறகள்"
		Secondary	"நிழற்பைம்"
		Tertiary	"எடுத்தல்"
4	Disease Identification	Screening	"X தர"
		Testing	"பாப்ஸ்மியர்"

#### 4. NE EXTRACTION MODEL

From the pre-processing, Morphological analysis and POS Tagging, the Tagged corpus comes with parts of speech tags for each word. To extract the useful information from the tagged corpus, named entity extraction from the tagged corpus is the important task after POS tagging. The named entities are tagged for the gynecological domain text by using the above Medical related NEs and machine learning methodologies. From the general text NEs, the Medical related NEs are derived for the proposed research work. Initially, general named entities are tagged by simple and powerful machine learning classification method, that is Naive Bayes classification. Before implementing the Naive Bayes classification method the TF-IDF metrics are to be calculated to classify the tagged corpus.

##### 4.1 Metrics for Evaluation

The TF-IDF is a statistical measure of a word's uniqueness that compares the number of times a word appears in a document to the number of documents in which it appears. Term frequency is a metric that determines the proportionate number of terms in a document to the total number of words in that document.

$$TF(t, d) = \frac{\text{count of } t \text{ in doc } d}{\text{total number of words in doc } d} \quad (1)$$

Document frequency is the measure of number of documents in which the term is present.

$$DF(t) = \text{occurrence of } t \text{ in doc } d \quad (2)$$

Inverse-document frequency is the calculation of informativeness of the term t in documents.

$$IDF(t) = \log \frac{1+n}{1+DF(d,t)} + 1 \quad (3)$$

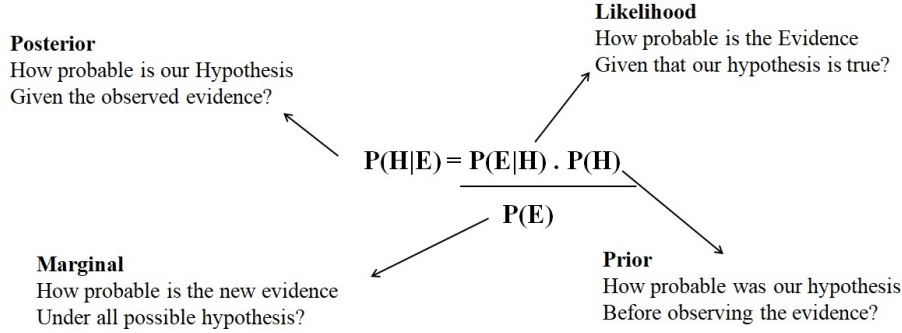
Finally the TF-IDF is calculated as

$$TF - IDF = TF(t, d) \times IDF(t) \quad (4)$$

With the help of TF-IDF measures and the pre-tagged corpus, the named entity identification for the documents is done by the powerful classification algorithm, the Naive Bayes classifier.

The supervised machine learning method Naive Bayes is based on the Bayes theorem. It is a probabilistic machine learning-based classifier that considers that each feature in the input text is independent during classifying. The posterior probability is used to identify the class of a NE classification I in a given set of characteristics.

#### Bayes Theorem



Bayes classifier works based on the trained data as evidence and testing data as outcomes.

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \Rightarrow P(Y|X) = \frac{P(X \cap Y)}{P(X)} \quad (5)$$

$P(\text{Evidence} | \text{Outcome})$        $P(\text{Outcome} | \text{Evidence})$

The Naive Bayes classification calculates the individual probability using the formula given below.

$$P(C_i|W) = P(C_i) \times P(W|C_i) \quad (6)$$

$P(C_i)$  is calculated as,

$$P(C_i) = \frac{\text{total number of features belongs to } C_i \text{ in the training sets}}{\text{total number of features in the training sets}} \quad (7)$$

$P(W|C_i)$  The conditional probability is calculated as

$$P(C_i) = \frac{\text{total number of words (w) belongs to } C_i \text{ in the training sets}}{\text{total number of features belongs to } C_i + |U|} \quad (8)$$

where  $|U|$  represents the total number of unique features in the training documents.

Similarly, each entity has its own collection of features, which are mentioned in the method above. Totally 1635 gynecological documents were tagged as named entities. By implementing the TF-IDF calculations for the selected corpus data, entities of each sentence were named in association with its NER Tags.

Using Naive Bayes classification ten named entities were identified in the Gynecological Tamil text chosen.

We identified this collection of named entity tags from our data.

Table 4. Extracted NER List

Named Entity Tags	NEs in Tamil	Example
Person	<நபர்>	டீனாஜ் பபண்கள், திருண்ணை பபண்கள், கற்பிணிப்பபண்கள், கற்பிணிகள், பிரசவித்த பபண்கள், வயதுக்கு வந்த பபண்கள்
Place	<இடம்>	கருப்ப, சிடனப்படையில், பிறப்புறுப்பில், கருப்ப வாயில்
Problem	<தநாய்>	நீர்க்கட்டிகள், பவளடளப்படுதல், கர்ப்பப்ப புற்றுதநாய், கருப்பவாய் புற்றுதநாய், ஸார்பகப்புற்றுதநாய்
Medicine	<ருந்து>	தடுப்பூசி, ஸாத்திடரகள்
Time	<தநரம்>	கருத்தரிப்பதற்கு முன், குழந்தைப்பற்றபின், குழந்தைகளுக்கு பாலூட்டும் பபாழுது
Number	<எண்>	18, 45, 35, 60, 50
Symptoms	<அறிகுறிகள்>	ஸாதவிஸாய் தள்ளிப்பதாதல், ஸார்பகங்களில் கட்டிகள்
Prevention	<தடுப்புமுடற>	தடுப்பூசி பசுலத்திக்காள்ளுதல், டைல் டைல் அதிகரிக்காதிருத்தல், கர்ப்பத்தடை ஸாத்திடரகடள் அதிக காலம் பயன்படுத்துதல்
Test	<தசாதடன>	எச்.பி.வி. டி.ஐ.வி.ஏ., பாப்ஸ்மியர்
Procedure	<ருத்துவமுடற>	அறுவ சிகிச்சை, ஸுத்துவ ஆதலாசடன

The above Table 4. gives the list of named entities extracted from the Gynecological text data in Tamil. From the extracted named entities, the problem entity has the list of major health problems gathered from the gynecological text data. NE tagged entities related with disease category are given below in Table 5.

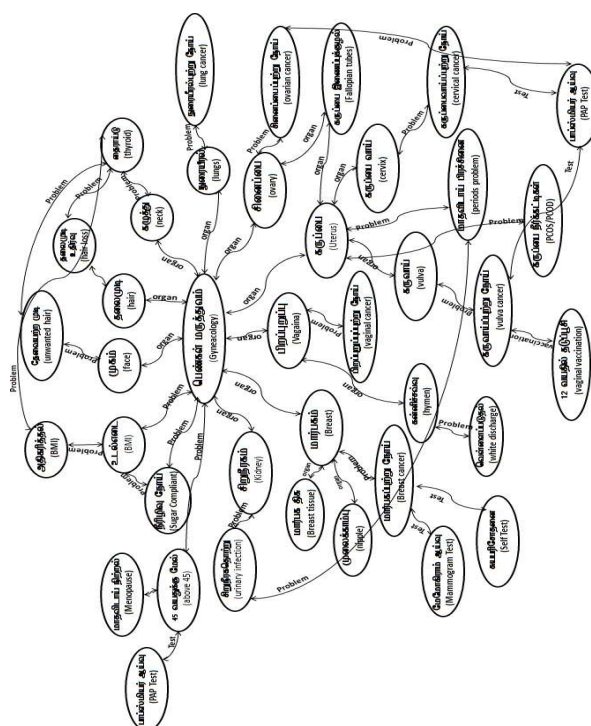
**Table 5. Disease Category NER**

S. No	Disease Categories	
1	ஆதராக்கியபாண்ம்	(Women Health Education)
2	பவள்ட்ளப்ப்டுதல்	(White discharge)
3	சிடன்ப்டபபுற்றுதநாய்	(Ovarian Cancer)
4	கருப்டபநீர்க்கட்டிகள்	(PCOS/PCOD)
5	பிறப்புலுப்புபுற்றுதநாய்	(Vaginal Cancer)
6	கருப்டபவாய்புற்றுதநாய்	(Cervical Cancer)
7	ண்ார்பகப்புற்றுதநாய்	(Breast Cancer)
8	டதராய்டு	(Thyroid)
9	ண்ாதவிண்ாய்பிரச்சிடன	(Menstrual Problems)

Above Table describes the Disease category named entities collected from tagged corpus. There are 9 categories of diseases found in the selected Tamil gynecological text corpus. With this NE tagged entities relation can be done within the same group of NE categories.

## 5. RELATION EXTRACTION

The task of obtaining semantic relationships from a text is known as relationship extraction. Extracted relationships are those that exist between two or more entities of the same type (for example, people, organisations, and places) and fit into one of several semantic categories (e.g. married to, employed by, and lives in). In the proposed tagged corpus data, relation can be made from the disease category named entity list. The relation implemented in the proposed named entity relation is many-to-many relations. The proposed relation created for the gynecological text document is given below.



**Figure 1. Relation extraction for Gynecological Domain Text**

The extracted relation by using named entity attributes full gynecological Named Entities are related with the other entities. Each entity is related with the others by the relation index id. From the NER tagged datasets the relation is created with the related entities tag manually. The noun tokens are assigned as the following NER tags. <நபர்> Person, <இடம்> Place, <தநாய்> Problem, <ஐருந்து> Medicine, <தநரம்> Time, <எண்> Number, <அறிகுறிகள்> Symptoms, <பசய்முடற> Procedure. Each entity in the corpus is stored with relations, client entities. For example, the entity karuppai vaaypputrunoai (கருப்பப வாய்ப்புற்றுதநாய்), has the following relations with the other

```
<P158>
<S1><நோய்>கருப்பவாய்ப்புற்றுநோய்</நோய்></S1>
<S2><நோய்>புற்றுநோய்களில்</நோய்>
<மருந்து>தடுப்பூசி</மருந்து> இதற்கு மட்டுமே இருக்கிறது</S2>
<S3><நேரம்>பாலியல் வாழ்க்கையை தொடங்குவதற்கு முன்பே</நேரம்>
இந்த <மருந்து>ஊசியை</மருந்து> செலுத்திக்கொள்வதே அதிக பயனுள்ளது</S3>
<S4><நபர்>டாக்டரின்</நபர்> <செய்முறை>ஆலோசனைப்படி</செய்முறை>
<எண்>பத்து</எண்> முதல் <எண்>26</எண்> <காலம்>
வயதுக்குள்</காலம்> இதனை <செய்முறை>செலுத்திக்கொள்ளலாம்</செய்முறை></S4>
<S5><எண்>45</எண்> <காலம்>வயது</காலம்> வரை இது
<செய்முறை>பலனளிக்கும்</செய்முறை></S5>
<S6><நபர்>கர்ப்பிணிகள்</நபர்> இந்த <மருந்து>ஊசியை</மருந்து>
<செய்முறை>செலுத்திக்கொள்ளக் கூடாது</செய்முறை></S6>
</P158>
```

entities:

Figure 2. Named Entity Tagging Structure

Table 6. NER Tags Relation Structure

Sentence ID	NER Tags	Tagged Words
S1	<தநாய்>	கருப்பவாய்ப்புற்றுதநாய்
S2	<தநாய்> <ஐருந்து>	புற்றுதநாய் தடுப்பூசி
S3	<தநரம்> <ஐருந்து> <நபர்> <பசய்முடற>	பாலியல் வாழ்க்கைய பதாண்ங்குவதற்கு முன்தப ஊசி ஐாக்டர் ஆதலாசடன
S4	<எண்> <எண்> <காலம்> <பசய்முடற>	பத்து 26 வயது பசலுத்திக்பகாள்ளலாம்
S5	<எண்> <காலம்> <பசய்முடற>	45 வயது பலனளிக்கும்
S6	<நபர்> <ஐருந்து> <பசய்முடற>	கர்ப்பிணிகள் ஊசி பசலுத்திக்பகாள்ளக் கூடாது

Based on the Named entity tags found from the document, the relation has created to extract the useful information. Each named entity has its own attributes to specify from the documents.

A sample relation for Breast cancer entity is shown in Table 7.

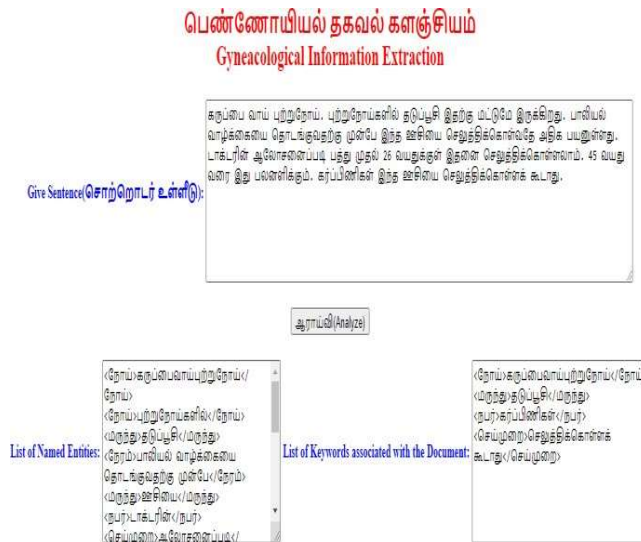
Table 7. Extracted relation for the Breast cancer

Keywords for Breast Cancer	
Name (தநாயின் பபயர்)	Breast Cancer (ஐார்பகப்புற்று தநாய்)
Person (யாருக்கு வரலாம்)	Teenage girls, Married women (பதினஐ வயது பபண்கள், திருணைணான பபண்கள்)

Reason (காரணம்)	Harmon problems and Genes (ஹார்தைன் குடற்பாடு, ஜீன்கள்)
Symptoms (அறிகுறிகள்)	Cyst in breast or armpit, Swelling in breast (ஓர்பகங்களின் கட்டிகள், அக்குளில் கட்டிகள், ஓர்பகங்களில் வீக்கம்)
Test (பரிதசாதன)	Mammogram (சுயபரிதசாதன, தைததாகிராம்)
Prevention (தடுப்பு முடறகள்)	Limit alcohols, Healthy weight (குடிப்பழக்கம், டைல் டைட்)
Medical treatment (ஓருத்துவ முடற)	Surgery, Chemotherapy (அறுடவ சிகிச்ச, கீததாபதரபி)

Finally the related entities and their attributes are related with their appropriate named entities. By using this relational structure Information Extraction framework is developed.

Based on the extracted Named entities, related keywords are found by the model, and the list of keywords is to be used in classification task. These keywords are listed from the named entities by the evaluation of TF-IDF method. The frequency of occurrence of each named entities is calculated. The most optimal occurred entities are listed as keywords. This process is explained in the following figure 2.



**Figure 3. Named Entities and Keywords Extraction**

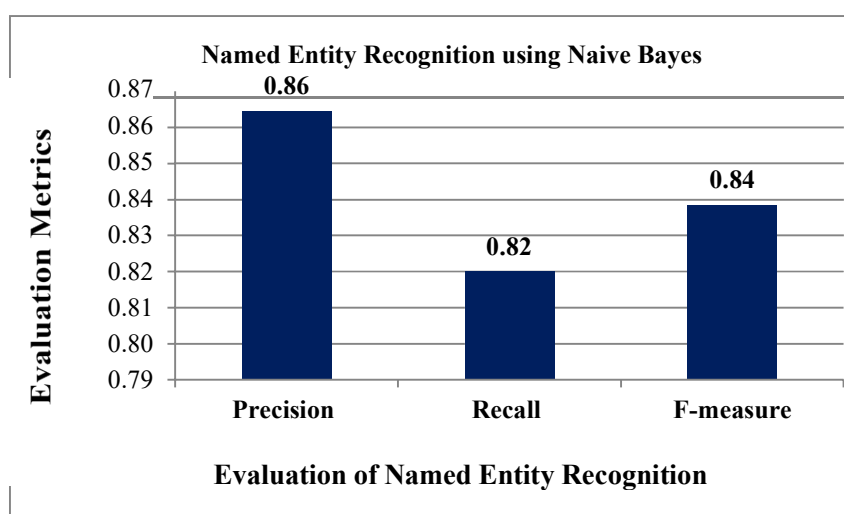
In figure 5.2 Tamil gynecological data are given as input to extract the needed information. From the NER corpus, the list of matching entities is found by Naive Bayes classification method. Extracted named entities are listed. Using TF-IDF calculation method, the list of keywords is selected from the Named entities. The list of named entities and keywords is forwarded to the classification process to extract the useful information as entities which are matched with the Entity corpus.

## 6. RESULTS OF NAMED ENTITY RECOGNITION EXTRACTION

For the named entity extraction, Naive Bayes classifier is used to classify the features with its relations. To evaluate this task general machine learning model evaluation method is used Precision, Recall, and F1-Score are calculated. For the gynecological domain text the entities are different from general domain text. Chosen Entities are Person, Place, Problem, Treatment, Test, Prevention, and Reason. Total tagged words by POS tagging is 62,439. Evaluation results are given the Table 8. and Figure 4. graphically.

**Table 8. Evaluation results for NE Extraction**

Named Entities	Precision	Recall	F-measure
Person	0.90	0.83	0.86
Place	0.91	0.84	0.87
Problem	0.86	0.81	0.84
Treatment	0.82	0.73	0.77
Test	0.85	0.81	0.86
Prevention	0.89	0.88	0.89
Reason	0.81	0.83	0.78

**Figure 4. Naive Bayes based Named Entity Extraction**

Naive Bayes classification method is used to extract the named entities from gynecological domain text in Tamil language. This Naive Bayes model performed well in the selected domain text.

## 7. REFERENCES

- [1] Chantana Chantrapornchai and Aphisit Tunsakul, "Information Extraction based on Named Entity for Tourism Corpus", 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2019, pp. 187-192
- [2] R. Rajimol1, V. S. Anoop2, "A Framework for Named Entity Recognition for Malayalam— A Comparison of Different Deep Learning Architectures", Atlantis Press - Natural Language Processing Research, Vol. 1(1-2), 2020, pp. 14–22
- [3] M. Pushpalatha1, Dr. Antony Selvadoss Thanamani2, "RULE BASED KANNADA NAMED ENTITY RECOGNITION", Journal of Critical Reviews, Vol 7, Issue 4, 2020, pp.255-258
- [4] N. Abinaya\*, M. Anand Kumar and K. P. Soman, "Randomized Kernel Approach for Named Entity Recognition in Tamil", Indian Journal of Science and Technology, Vol 8(24), 2015
- [5] R. Srinivasan and C.N. Subalalitha, "Automated Named Entity Recognition from Tamil Documents", IEEE International Conference on Energy, Systems and Information Processing (ICESIP),2019
- [6] C.N. Subalalitha and R. Srinivasan, "Automated Named Entity Recognition from Tamil Documents", IEEE International Conference on Energy, Systems and Information Processing (ICESIP),2019

# Diffusion of Meaningful Information in Tamil on OSN by Forming Communities with the Aid of HetNet Formation in 5G

G.Ramasubramanian<sup>1</sup>, Dr. S. Rajaprakash<sup>2</sup>

Research Scholar, Vinayaka Mission's Research Foundation, Salem

Associate Professor & Head, Department of CSE,AVIT, Vinayaka Mission's Research Foundation, Chennai

<sup>1</sup> ramasubramanian.csa@avit.ac.in

<sup>2</sup> srajaprakash\_04@yahoo.com

## ABSTRACT (10 PT)

Online Social Networks (OSN) are playing a vital role today for the diffusion of information across the world within a fraction of second. Today OSN support all sorts of regional languages for their users to propagate their messages. Such kind of actions carried out by the OSN users are not only share the messages but also helps the business agencies and individuals to promote and identify their needs. By considering these on mind, a novel framework has been formed out for analyzing and finding the meaning of Tamil posts produced by the users and diffuses the meaningful information with the aid of HetNet formation in 5G networks. The framework being constructed is deploying the sentimental analysis and recommender system for the benefit of faster solution. The proposed framework is useful for the society to group the same users with similar internal traits and topics as well it help to identify the needy in a faster way.

## Keywords:

A Online Social  
B Networks, HetNet  
C Diffusion  
D LDA  
E influencers  
F 5G

## Corresponding Author:

Name Corresponding Author,  
2Department of Computer Science,  
Tamil University India  
Email: correspondingauthor@email.com

## 1. INTRODUCTION

Online Social Networks(OSN) are becoming the popular platforms and are the sources for the campaigns creating, larger crowdsourcing within a fraction of seconds. When a topic is discussed on OSN, it is diffused to all people involved in the networking environment. People who form such virtual by means of posts are categorized to as influencers and those who follow the created campaigns are called as followers. Larger number of people, and business companies are involved in the information on OSN to form opinions and selecting choices on lifestyles, politics, health and product purchases, etc. [1].

There is widespread subjective evidence of formation of electrical campaigns in which the political operatives insert memes into famous SNS like Twitter and Facebook, etc. Besides, there are huge number of campaigns of coordinated spam messages in OSN along with promoted and advertising concepts. A campaign in OSN is a collection of users and posts as like on blogs, comments on OSN and forum posting, etc [2]. The invention of digital technologies diffuses all kinds of messages and the evaluation of the Internet referred to as OSN blur the boundaries between private and public activities [3]. Besides, the recent social network sites like Facebook, Twitter and YouTube, etc., are not only connecting the larger number of users, but also extracts exabytes of information from their daily interactions. OSN also exhibit the three basic characteristics of big data called "three VS"- that is Volume, Velocity and Variety and forms a new meaning for the field of big data [4].

In general, drives are used for marketing purposes, but in another context the concept of campaign management is considered as disturbances at certain extends. Campaign formation and elevation towards users is considered as spam and is meaningfully increasing through emails and social networks [5]. Initial stages of online campaign management are carried out through emails and then spread with the help of blogs, forums



and recently through OSN. Spammers are using events to direct the users of OSN with the help of event related campaigns and now the social networks contain botnets, worms and viruses [6]. The management of is carried out different sorts of people including individuals, firms and business companies to promote their thoughts and outcomes.

Research on social networks to handle effective campaign management shown that consumers are selecting social media marketing to promote their products. For example, one of the popular cool drink companies Pepsi and Coca-Cola applied online customer loyalty events to encourage customers by providing special promotions like free mp3 downloads and CD's, etc. From the client point of view, social networks are the service channels used for engaging real-time bases with businesses [7]. Different types of techniques like content-based classification, clustering, similarity measurements, recommender systems and fuzzy-logic, etc., are used for detecting the campaigns on OSN.

Formation of effective campaign is not only the main task under OSN, but a meaningful information is also to be diffused among all users without any disturbance on network communication. Maximum up to 4G like technologies and broadband, OFC technologies are deployed for the formation of making Internet communication with high speed. These technologies work well under normal conditions, but there are struggles under the disturbance on natural disasters. Recent technologies like 5G with Heterogeneous Networks(HetNet) are helping to tackle such situations, so that the communication in between the devices will never fail and may propagate from smaller HetNet group to broader OSN like sites on the Internet

This paper discusses the support of 5G HetNet architecture for the diffusion of information by forming virtual campaign based on the Tamil keywords and a framework for finding the influencers tweets from Twitter with the help of classifier giving high accuracy results. Bunch of supervised classifiers are utilized over the selected tweets of Twitter and a classifier giving high accuracy is applied for automatic detection and grouping of different kinds of with selected prime words. Organization of the paper is given below

Section II discusses the architecture of 5G HetNet in connection with communication. Section III explores the used classification techniques; Section IV expands the related works and Section V describes the proposed methodology of the present paper. Section VI explores outcomes obtained from the selected samples. Last section of the paper gives conclusion and future development .

## 2. ARCHITECTURE OF 5G HetNet

In the present scenario, new generations of 5G runs applications needing high data rate by densification of networks by deploying smaller cells. The densification of smaller cells yields higher spectral efficiency, low latency and also reduces the power consumption of devices due to the communication with nearby pico-cell. Such arrangement also improves the network coverage.

The overall arrangement is collectively known as Heterogeneous Networks(HetNets) and contains arrangements like Macro, Micro, Pico and femto cells for establishing connection from the minimum level (i.e. from device to device).All these components are operating concurrently to establish communication through the 5G coverage. The architecture of 5G is shown in Figure 1.

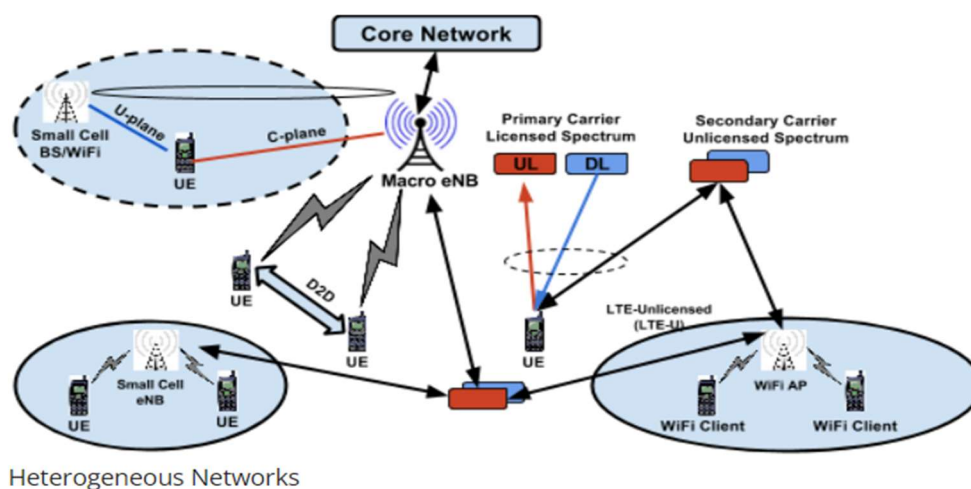


Figure.1 Heterogeneous Network being the part of 5G communication

The small Cell eNB shown in Figure.1 might be a Macro, Micro, Pico or femto cell and the UE is the User Equipment being connected with the network. Entire HetNet arrangement is being the part of a core 5G network. Using of HetNets introduces newer communication methods happening within indoor places.

By taking the advantages of 5G with HetNets arrangements, the entire field of the Internet is using sustained connection at the low signal areas. Especially, this could be very useful for the information diffusion at the time of crisis management, because there are power failures, signal prohibition due to obstacles. A single device would enough to make a peer to peer connection with the nearest eNB cell, so that the entire communication is to be retained.

### 3. CLASSIFICATION TECHNIQUES

Different classification techniques used for effective mining processes to extract meaningful data. In classification, a classifier takes two categories of inputs namely, features and labels. The selected input are further be divided into two sets, train data and test data before given into the classifier. By using the train and test data, the selected classifier learns the way of classifying and after learning, the classifier predicts the correct label (known) for the given new feature (unknown) value. Following are the used classifiers in the present paper.

#### A. Support Vector Machine (SVM)

Support Vector Machine (SVM) seems to be a binary classifier discriminatively harvests new predictions from the given set of training data. The model draws an optimal hyperplane with a new prediction existing on both sides of created hyperplane. The main advantage of SVM is that the model works fitted for both linear and non-linear points and produces optimal outputs. SVM organizes various kernels like linear, non-linear, polynomial and Radial Basic Function (RBF) for classification [8].

#### B. Naïve Bayes (NB)

Naïve Bayes (NB) classifiers are probabilistic classifiers and works on the Bayes theorem. The Naïve Bayes classifier undertakes that all the variables are mutually exclusive and considering the value of the class variable. NB classifier calculates a set of probability by counting the frequency and combination of values in the input. The conditional independence is infrequently valid in many real-world applications [9].

#### C. Logistic Regression (LR)

Logistic Regression is a statistic classifier and identifies the class of new data from the set of available categorical data with the help of training data. The classifier also finds the relationship between the categorical dependent variables and one or more dependent variables [10].

#### D. Decision Tree (DT)

Decision Tree (DT) classifier changes all the facts into decision trees, presenting rules that can be easily understand by the natural language processing. DT process starts from the root of a decision tree until the leaf node is searched recursively in which each branch states the condition and the reached leaf states the decision [11].

#### E. k-Nearest Neighbour (k-NN)

k-Nearest Neighbour is the widely used classifier and examines all the objects in the reference dataset for each unknown query object. The classifier calculates k-nearest neighbours by using the training data and after the calculation, the similarity is tested by taking a single calculated sample [12].

#### F. Random Forest (RF)

Random Forest (RF) is the popular ensemble classification technique and builds an ensemble of decision trees called decision forest. The method uses training bootstraps to build binary-sub trees (decision trees) by using the training boot strap samples and randomly select each node a subset. The constructed decision forest selects classification with high vote as the result from all the constructed decision trees [13].

The present paper utilized all the discussed classifiers and select the classifier with high accuracy for the automatic detection of campaigns in SNS.



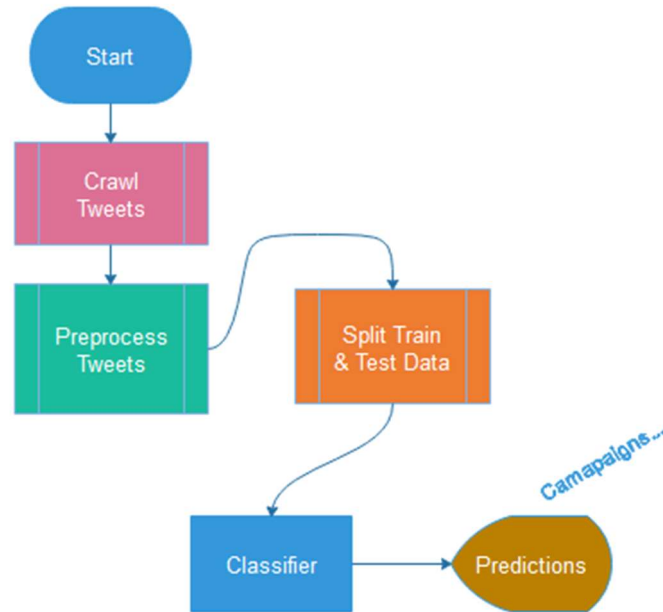


Fig 3. Overall Methodology

The automatic detection of campaign in OSN with the methodology shown is presented as the following algorithm in Table 1.

TABLE 1 AUTOMATIC CAMPAIGN DETECTION (ACD) ALGORITHM

Algorithm: ACD (input: tweets)

Step 1. Apply the steps in Methodology

- a) Crawl tweets
- b) Pre-process tweets
- c) Convert text to vector by using TF-IDF vectorizer
- d) Split Train and Test Data
- e) Apply classification
- f) Select classifier with high accuracy

Step 2. Predict the performance of the classifier with high accuracy

Step 3. Store the detected details for further information diffusion process

Step 4. Stop.

Three main steps are involved in the algorithm ACD presented in Table 1. Step 1 applies all the stages of proposed methodology and Step 2 Predicts the performance by using test and new tweets from users. The predictions are analyzed and results are stored for further information diffusion process like indicating the campaigns to users, group the campaigns for the ease of detecting the offers by the users. The top most campaign is also to be analyzed with the help of ACD algorithm shown in Table 1.

## 6. METHODOLOGY

The classification accuracy of all classifiers for the selected 20,000 samples is plotted in Fig 4. with selected classifier name on x-axis and the percentage accuracy on y-axis.

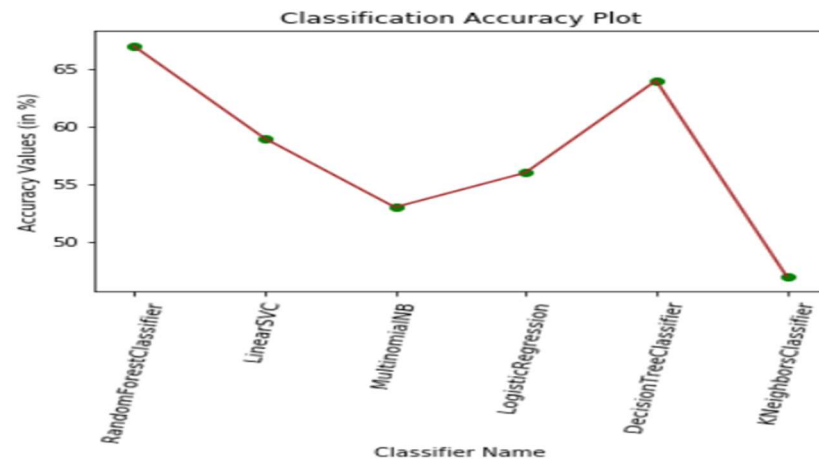


Fig .4 Accuracy of classifiers

The RandomForestClassifier produced better results with high accuracy than the other classifiers as shown in Fig 4. The next high-level classifier is the Decision Tree classifier produced next high accuracy value. Other classifiers produced lower accuracy values and the k-NN is the lowest accuracy classifier produced lower results. The test is repeated with different number of samples like 25,000 and 30,000, etc., and in all cases RandomForest did well classification with the selected data.

## 7. CONCLUSION

The present paper proposed a novel approach for finding Campaign words in Tamil on Twitter with selected number of samples taken from the crawled dataset. The main scope of present paper is to automatically detect the Campaigns containing Tamil words by using the influencing words from the influencers' tweets and identified the Campaign words by adopting a stream of supervised classifiers. The Random Forest classifier produced better results than other classifiers with high accuracy.

The prescribed work on present paper only detected the campaigns by using the classification and the selected data is of the type text. Applying of detected campaign word for further processing like, grouping the campaign originators by using the campaign words, getting users' feedback for the found campaign, identifying the similar users who are interested in found campaigns and using other kinds of social media like images, and so on are the beyond the scope of present research work.

## REFERENCES (10 PT)

- [1] Varol, O., Ferrara, E., Menczer, F., & Flammini, A. (2017). Early detection of promoted campaigns on social media. *EPJ Data Science*, 6(1), 13.
- [2] Lee, K., Caverlee, J., Cheng, Z., & Sui, D. Z. (2011, October). Content-driven detection of campaigns in social media. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 551-556).
- [3] Gonzalez-Bailon, S., & Wang, N. (2013). The bridges and brokers of global campaigns in the context of social media. *SSRN Work. Pap.*
- [4] Tan, W., Blake, M. B., Saleh, I., & Dustdar, S. (2013). Social-network-sourced big data analytics. *IEEE Internet Computing*, 17(5), 62-69.
- [5] Zhang, X., Zhu, S., & Liang, W. (2012, December). Detecting spam and promoting campaigns in the twitter social network. In *2012 IEEE 12th international conference on data mining* (pp. 1194-1199). IEEE.
- [6] Ahmed, F., & Abulaish, M. (2013). A generic statistical approach for spam detection in Online Social Networks. *Computer Communications*, 36(10-11), 1120-1129. doi: 10.1016/j.comcom.2013.04.004.
- [7] Erdoğan, İ. E., & Cicek, M. (2012). The impact of social media marketing on brand loyalty. *Procedia-Social and Behavioral Sciences*, 58, 1353-1360.
- [8] Mohankumar, K., & Srinivasan, B. (2019). Point-of-Interest Based Classification of Similar Users by Using Support Vector Machine and Status Homophily. *International Journal of Machine Learning and Computing*, 9(5), 615-620.
- [9] Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, 7(2), 88-91.
- [10] Deepa B & Jeen Marseline K S (2019). "Social Media Data using Various Classification Algorithms in Datamining." In *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, pp 588-589.
- [11] Dimas Purnomo A, Muhammad Iqbal D & Nova Teguh Sv (2020), "Data Mining for box compression test results classification using decision tree methods, IJSTR, volume 9, issue 03.
- [12] Shah, K., Chauhan, P., & Potdar, M. B. (2015). Design and evaluation of classification algorithm on GPU. *International Journal of Scientific Engineering Research*, 6(5), 639.

- 
- [13] Azar, A. T., Elshazly, H. I., Hassanien, A. E., &Elkorany, A. M. (2014). A random forest classifier for lymph diseases. Computer methods and programs in biomedicine, 113(2), 465-473.
- [14] Weng L,FlamminiA,Vespignani A &Menczer F(2012) Competition among memes in a world with limited attention. SciRep2:354

## விசேட தேவையுடைய மாணவர்களின் கற்றலை மேம்படுத்துவதில் கணினி வழிக் கற்பித்தலை மேற்கொள்வதில் ஆசிரியர்களின் பங்களிப்பு

Mr. N.Koventhan<sup>1</sup>, Mrs. R.Thakshaayini<sup>2</sup>

2Senior Lecturer in Education,  
1,2Department of Education and childcare,  
Faculty of Arts and Culture,  
Eastern university, Sri Lanka

1 [koventhannadarasa@gmail.com](mailto:koventhannadarasa@gmail.com)

### ABSTRACT

#### Keywords:

- A விசேட தேவையுடையோர்
- B கற்றல்
- C கற்பித்தல்
- D கணினி

மட்டக்களப்பு மேற்கு கல்வி வலயத்தை ஆய்வுப் பிரதேசமாகக் கொண்டு அளவீட்டு ஆய்வாக இவ்வாய்வு மேற்கொள்ளப்படுகிறது. விசேட தேவையுடைய மாணவர்கள் தமது கற்றலில் எதிர்நோக்கும் பிரச்சினைகளை இனங்காண்பதுடன் அவர்களது கற்றலை மேம்படுத்துவதில் கணினி வழிக் கற்பித்தலை மேற்கொள்வதில் ஆசிரியர்களின் பங்களிப்பை ஆராய்தல் எனும் நோக்கத்தின் அடிப்படையில் இவ்வாய்வு இடம்பெற்றுள்ளது. ஆய்வுப் பிரதேசமாக மட்டக்களப்பு மேற்கு கல்வி வலயம் தெரிவுசெய்யப்பட்டது. இக்கல்வி வலயத்தில் விசேட தேவையுடைய மாணவர்களைக் கொண்ட 10 பாடசாலைகள் மாதிரிகளாகத் தெரிவு செய்யப்பட்டன. இப் பத்து பாடசாலைகளிலிருந்து 10 அதிபர்கள் நோக்க மாதிரியின் அடிப்படையிலும், விசேட தேவையுடைய மாணவர்களுக்கு கற்பிக்கும் 20 ஆசிரியர்களும், விசேட தேவையுடைய 50 மாணவர்களும், இம்மாணவர்களின் பெற்றோர்கள் 50 பேரும் எளிய எழுமாற்று மாதிரியின் அடிப்படையிலும் ஆய்வு மாதிரிகளாகத் தெரிவு செய்யப்பட்டு வினாக்கொத்து, அவதானம், நேர்காணல் எனும் ஆய்வுக் கருவிகள் பயன்படுத்தப்பட்டுத் தரவுகள் சேகரிக்கப்பட்டன. பெறப்பட்ட தரவுகளை அடிப்படையாகக்கொண்டு ஆய்வு நோக்கங்களின் அடிப்படையில் பண்புரீதியானதும், அளவுரீதியானதுமான கலப்பு முறையில் பகுப்பாய்வு செய்யப்பட்டு ஆய்வின் முடிவுகளாக இம்மாணவர்களின் கற்றலில் ஆசிரியர்களின் கணினி வழிக் கற்பித்தலின் பங்களிப்பு குறைவாக இருப்பதற்கான காரணங்களாகப் ஆசிரியர்களுக்கு விசேட தேவையுடைய பிள்ளைகளைக் கையாள்வது பற்றிய தெளிவின்மை, விசேட தேவையுடைய மாணவர்களுக்கு கற்பிப்பதற்கு போதிய கணினி வசதிகள் இல்லாமை, இம் மாணவர்களுக்கு கற்பிக்கும் ஆசிரியர்கள் கணினி தொடர்பான கற்கை நெறிகளை தொடரமை, பெற்றோரின் ஆதரவு குறைவாக உள்ளமை, பாடசாலையில் பல்வேறு வகையான விசேட தேவையுடைய பிள்ளைகளும் காணப்படுதல் போன்றன இனங்காணப்பட்டன, இம்மாணவர்களின் கற்றலை மேம்படுத்துவதற்குத் தீர்வுகளாக அதிபர்கள் கணினி வசதிகளை ஏற்படுத்திக் கொடுத்தல், ஆசிரியர்களுக்கு ஆசிரிய ஆலோசகர்கள் தகுந்த ஆலோசனைகள் வழங்குதல், பாடசாலையில் கணினி அறிவை வளர்க்கும் வகையில் செயற்றிட்டங்களை செய்தல், அதிபர் ஆசிரியர்களை மாணவர்களின் வீட்டிற்குச் சென்று பெற்றோருடன் மாணவர்களின் கற்றல் தொடர்பில் கலந்துரையாட வைத்தல். போன்றன விதந்துரைக்கப்பட்டுள்ளன.

#### Corresponding Author:

Mr. N.Koventhan  
Senior Lecturer in Education,  
Department of Education and childcare,  
Faculty of Arts and Culture,  
Eastern university, Sri Lanka



## 1. ஆய்வு அறிமுகம்

### 1.1 அறிமுகம்

பொதுவாக சாதாரண மாணவர்களுக்கு தேவைப்படாத விசேட கவனம் மற்றும் குறிப்பிட்ட தேவைகள் அவசியமாக இருக்கும் மாணவர்களை விசேட தேவையுடைய மாணவர்கள் என்பர் (நுடிழலெ ஈழறயசனஇ 2021). பாடசாலையில் விசேட தேவையுடைய மாணவர்களின் கற்றலில் ஆசிரியர்களின் பணி மிகவும் பொறுப்பு வாய்ந்ததாகவும், மகத்துவம் மிக்கதாகவும் காணப்படுகின்றது. பெற்றோர்கள் தம் பிள்ளைகளை புத்திஜீவிகளாக்கி சமுதாயத்திற்கு ஏற்ற சிறந்த பிரஜைகளாக்கும் பொறுப்பை ஆசிரியர்களிடம் விடுகின்றனர்.

விசேட தேவையுடைய பிள்ளைகளிற்கு கல்வி வழங்குவதற்கென்று மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் பாடசாலைகள் இருப்பினும் பெற்றோர்கள் பிள்ளைகளை பாடசாலைக்கு அனுப்பும் வீதம் மிகக் குறைவாகவே உள்ளது ( பிரதேச செயலகத்தின் விசேட தேவையுடைய பிள்ளைகள் பற்றிய அறிக்கை 2018 ).

இன்று கணினிக் கல்வியியல் மூன்று பாத்திரங்களை வகிக்கின்றது. ஒரு போதனையாளன், ஒரு கருவி, ஒரு போதனை பயனாளி இது இன்னும் விரிவடையும் என கல்வியியலாளர்கள் எதிர்வுகூறுகின்றனர் மேல் நாடுகளில் பாடப்புத்தகங்கள் இரண்டாம் நிலைக்கும் கற்றலில் கணினியில் இணையப்பயன்பாடு முதலாவதாகவும் மாறி வருகிறது. இவ் விடயம் விசேட தேவையுடைய மாணவர்களின் கற்றலில் எவ்வாறு உள்ளது மற்றும் இதில் ஆசிரியர்களின் கணினி வழிக் கற்பித்தலின் பங்களிப்பை அறியும் பொருட்டு இவ்வாய்வு மேற்கொள்ளப்படுகிறது.

### 1.2 ஆய்விற்கான பின்னணி : நியாயத்துவம்

பிள்ளைகள் ஒரே தன்மையான ஆற்றல், அனுபவமுடையவர்களாக இருப்பதில்லை. ஒவ்வொரு மாணவரும் பல்வேறுபட்ட கற்றல் தேவையுடையவர்களாக காணப்படுவார்கள். இவர்கள் அனைவரையும் சீரான வழியில் நெறிப்படுத்துவது ஆசிரியரின் பொறுப்பாகும். வகுப்பறையில் மீத்திறன் மாணவர்கள், மெல்லக்கற்கும் மாணவர்கள், கற்றல் இடர்பாடுடைய மாணவர்கள் என பல வகையான மாணவர்கள் காணப்படுவார்கள். இவர்களின் கற்றல் பாங்கு, ஆற்றல், திறன், போன்றன ஒவ்வொருவருக்கும் வித்தியாசமானதாகும். இவர்கள் கற்றுக் கொள்வதில் வித்தியாசமான கற்றல் தேவையுடையவர்களாக காணப்படுவார்கள் (யுயெள யனெ யேறயளவாநநஇ 2019).

மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் கூடுதலான விசேட தேவையுடைய குடும்பங்கள் காணப்படுகின்றது (பிரதேச செயலகத்தின் விசேட தேவையுடைய குடும்பங்களின் கணக்கெடுப்பு அறிக்கை, 2018). இதில் விசேட தேவையுடைய பிள்ளைகளின் தொகையும் அதிகமாகவே காணப்படுகிறது.

உலகெங்கும் தொடர்ச்சியாக பரவிவரும் தொழிநுட்ப மாற்றம் கல்வியிலும் விரைவான மாற்றத்தை ஏற்படுத்தியுள்ளது. அதனால் இன்று விசேட தேவையுடைய மாணவர்களினதும், ஆசிரியர்களினதும் கல்வியை மேம்படுத்தும் பிரதான கருவியாக கணினி, இணைய நூலகங்கள் என்பன பங்குகொள்கின்றது. இதன் தோற்றம் எவ்வாறு என்பதை பார்த்தால் நூல்கள் பல நூற்றாண்டுகளாக மனிதனின் கற்றல் செயற்பாட்டில் உதவி வந்தன. 17 ஆம் நூற்றாண்டில் கொமேனியஸ் எனும் ஐரோப்பிய அறிஞர் பாடநூலை கண்டுபிடித்தார். இதனால் அறிவை பரப்புவதில் ஆசிரியர்களிடம் இருந்த ஏகபோக உரிமை இல்லாதொழிந்தது. இதனால் எந்நேரமும் அறிவைப் பெறலாம் எனும் வாய்ப்புண்டாகியது.

எனினும் நூலை அச்சிடுவது, வெளியிடுவது, பாதகாப்பது என்பது சிரமமானது. என உணரப்பட்டதனால் ஒரு பயனுள்ள மாற்று ஏற்பாட்டை நவீன தொழிநுட்பம் தந்தது. அதில் ஒன்றே கணினி, இணையம் இவ்விவ்விதத்தினால் ஊடகமானது அச்சிடப்பட்ட சொற்களுக்கும் அப்பால் ஒலி, நிறம், இடைத்தொர்பு வசதிகளையும் தருவதோடு விடயத்தை இணையத்தளத்தின் ஊடாக உலகெங்கும் விநியோகிக்கும் ஆற்றல் படைத்ததாகவும் மாறி வந்தது.

குறித்த ஆய்வுப் பிரதேசத்தில் அதிகபடியான விசேட தேவையுடைய மாணவர்கள் எழுத, வாசிக்கத் தெரியாதவர்களாகவும், தங்கள் வேலைகளை சுயமாகச் செய்ய முடியாதவர்களாகவும் காணப்படுகின்றனர். இம்மாணவர்களுக்கான வளங்களும் குறித்த ஆய்வுப்பிரதேசத்தில் அரிதாகவேயுள்ளது. இம் மாணவர்களுக்கு கற்பிக்கும் ஒரு சில ஆசிரியர்கள் விசேட தேவைகள் தொடர்பான பட்டப்படிப்பினை மேற்கொள்ளாதவர்களாகவும் கணினி மூலம் இம்மாணவர்களுக்கான கற்றல், கற்பித்தல் செயற்பாட்டை முன்னெடுக்க முடியாமல் உள்ளனர்.

விசேட தேவையுடைய பிள்ளைகளுக்கு கல்வியை வழங்குகின்ற பாடசாலைகள், கல்வியை உயர்த்தத்தில் வழங்குகின்றனவா? எதிர்காலத்தில் வாழ்க்கைத் தரத்தினை உயர்த்துவதற்கான வழிகள் எத்தகையன? இதில் ஆசிரியர்களின் பங்களிப்பு சிறப்பாக இருக்கிறதா? என்பதை ஆய்வு செய்வதாக உள்ளது. இவ்வாறான பிரச்சினைகளை பின்னணியாகக் கொண்டு இவ் ஆய்வு மேற்கொள்ளப்படுகிறது.

### 1.3 ஆய்வுப் பிரச்சினை :-

மட்டக்களப்பு மேற்குக் கல்வி வலயத்தில் விசேட தேவையுடைய மாணவர்கள் கற்பதற்கான விசேட அலகுகள் நான்கு உருவாக்கப்பட்டுள்ள போதிலும் இம் மாணவர்களுக்கான விசேட தேர்ச்சி பெற்ற ஆசிரியர்களும் மிகக் குறைவாகவே காணப்படுகின்றனர். விசேட தேவையுடைய மாணவர்கள், குறித்த ஆய்வுப் பிரதேசத்தில் இருந்து

பல்கலைக்கழகத்திற்கோ, கல்வியற் கல்லூரிகளுக்கோ செல்வது மிகக் குறைவாகவே உள்ளது. இது ஒரு பாரிய பிரச்சினையாகும்.

மேலும் பெற்றோர், ஆசிரியர்கள், அதிபர் இணைந்து செயற்படும் போக்கு குறித்த ஆய்வுப் பிரதேசத்தில் குறைவாகவே உள்ளது. ஆசிரியர்கள் முன்வந்தால் பெற்றோர்கள் முன்வருவது அரிது அத்துடன் பாடசாலையில் இருந்து இம் மாணவர்களின் வீடுகள் அதிக தூரத்தில் இருப்பதனால் ஆசிரியர்கள் இம்மாணவர்களின் வீடு தரிசித்தல் செயற்பாட்டை முன்னெடுப்பதில் சிரமத்தை எதிர்கொள்கின்றனர். என்பதை ஆய்வாளன் நேரடியாக இம் மாணவர்களின் வீடுகளுக்கு சென்ற போது அவதானிக்க முடிந்தது.

ஆசிரியர்கள் குறைவாக உள்ளமையினால் குறித்த பாடசாலைகளில் தனியாள் வேறுபாடுகளுக்கேற்ப கற்றல், கற்பித்தல் செயற்பாட்டை முன்னெடுப்பதில் சிக்கல் நிலை உள்ளது. இதில் கணனி வழிக்கற்பித்தலை மேற்கொள்வது ஒரு பாரிய பிரச்சினையாக உள்ளது. இன்றைய நிலையில் விசேட தேவையுடைய மாணவர்களின் சிறந்த வகுப்பறைக் கற்றல், கற்பித்தல் செயற்பாடுகளுக்கு வகுப்பறை முகாமைத்துவம் சரியானதாக இல்லை. இம் மாணவர்களுக்கு கற்பிக்கும் ஆசிரியர்கள் ஒரு சிலர் விசேட தேவைக் கல்வி தொடர்பான கற்கை நெறிகளை கற்காமலே கற்பிக்கின்றனர். இதனால் எந்த விசேட தேவையுடைய மாணவர்களை எவ்வாறு கையாள்வது என்பது பற்றிய விளக்கம் இன்மையால் சாதாரண மாணவர்களுக்கு கற்பிப்பது போன்றே கற்பிக்கின்றனர். இதனால் இம் மாணவர்கள் கற்றலில் முன்னேறுவது குறைவாக உள்ளது. இவ்வாறான நிலை மிக முக்கியமான பிரச்சினையாக குறித்த ஆய்வுப்பிரதேசத்தில் காணப்படுகிறது.

#### **ஆய்வுப் பிரச்சினைக் கூற்று :-**

மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் விசேட தேவையுடைய மாணவர்களின் கற்றலை மேம்படுத்துவதில் ஆசிரியர்கள் மேற்கொள்ளும் கணனி வழிக் கற்பித்தல் குறைவாக காணப்படுகின்றமையால் இம் மாணவர்கள் பொருத்தமான கல்வியை பெறுவதில் சிரமப்படுகின்றனர். (மட்டக்களப்பு மேற்கு கல்வி வலய விசேட தேவையுடைய மாணவர்கள் தொடர்பான ஆய்வறிக்கை, 2019).

#### **1.4 ஆய்வினது முக்கியத்துவம் :-**

மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் விசேட தேவையுடைய பிள்ளைகளிற்கான கல்வியை வழங்குவதில் ஆசிரியர்கள் மேற்கொள்ளும் கணனி வழிக் கற்பித்தலானது குறைவாகவே காணப்படுகிறது. இவர்களின் மத்தியில் விழிப்புணர்வை ஏற்படுத்துவதற்காகவும், எதிர்காலத்தில் இந்த ஆய்வுப்பரப்போடு தொடர்புபட்ட ஆய்வுகளை மேற்கொள்பவர்களுக்கு ஆய்வின் நோக்கத்தை பூரணப்படுத்துவதற்கும், தரவுப் பகுப்பாய்வுகளை மேற்கொள்வதற்கும் இவ்வாய்வு முக்கியம் பெறுகின்றது.

#### **2.0 சார்பிலக்கிய மீளாய்வு :-**

பொதுவாக சாதாரண மாணவர்களுக்கு தேவைப்படாத விசேட கவனம் மற்றும் குறிப்பிட்ட தேவைகள் அவசியமாக இருக்கும் மாணவர்களை விசேட தேவையுடைய மாணவர்கள் என்பர் (நுடிழலெ ஈழறயசனஇ 2021).

விசேட தேவைகளை கொண்ட மாணவர்கள் வகுப்பறை கற்றல் நடவடிக்கைகளில் பின்தங்கியிருப்பது கண்டறியப்பட்டது. எனவே ஆசிரியர்கள் கற்றலில் சிரமங்களை எதிர் கொள்கின்ற மாணவர்களைப் பற்றி அதிக கவனம் எடுத்துக்கொள்ளவும், கற்றல் நடவடிக்கைகளில் ஈடுபாட்டினை அதிகரிக்கவும் வேண்டும் (ஸ்ராயணலெஇ 2021).

கணினி, மாணவர்களை நவீன தொழில் நுட்பத்திற்கு பழக்கப்படுத்துவதோடு மட்டுமல்லாமல் புதுமையான, அறிவியல்பூர்வமான கற்றலுக்கும் மாணவர்களைத் தயார் செய்கிறது. ஆசிரியர் விரிவுரை மூலம் கற்பித்தலில் ஈடுபடாமல் கணினியைப் பயன்படுத்தி மாணவர்களையே கற்றுக் கொள்ள வைப்பது "கணினி வழிக் கற்பித்தல்" எனப்படும் (நுடை - 1982).

விசேட தேவையுடைய மாணவர்களின் கற்றல் செயல்பாட்டில் உதவ கணனி தொழில்நுட்பம் முக்கிய பங்கு வகிக்கிறது. இந்த நபர்களுக்கு சுயாதீனமாக செயல்பாடுகளைச் செய்வதில் சிக்கல் உள்ளது, மேலும் இதுபோன்ற மாணவர்களுக்கு உதவ ஆசிரியர்கள் தங்கள் கற்றல் செயல்முறைகளில் கணினியை இணைத்துக்கொள்வார்கள். இத்தகைய தொழில்நுட்பம் குறைபாடுகள் உள்ளவர்களுக்கு அவர்களின் பணிகளைச் செய்வதற்கான திறனை மேம்படுத்த அதிக வாய்ப்புகளை வழங்குகிறது, மேலும் ஊனமுற்றோரின் கல்வியை ஊடாடுவதன் மூலம் மேம்படுத்துகிறது, இதனால் கருத்துகளை மிகவும் திறம்பட கற்பிக்க முடியும். கல்வி செயல்பாட்டில் கணனி இருப்பதால், மாற்றுத்திறனாளி மாணவர்கள் தங்கள் சமூக திறன்களை மேம்படுத்தவும், மேம்பட்ட வாழ்க்கை முறைகள் மூலம் சமூக தொடர்புகளில் ஈடுபடவும் உதவி பெறலாம். (வாக்கர், 2018).

#### **3.0 ஆய்வு முறையியல் :-**

ஆய்வு முறையியலானது, ஆய்வின் பொது நோக்கத்தையும், சிறப்பு நோக்கங்களையும் அடிப்படையாகக் கொண்டு, ஆய்வு வினாக்கள் தயாரிக்கப்பட்டு அவற்றுக்கு விடை காண்பதன் அடிப்படையில் ஓர் அளவு, பண்பு ரீதியான கலப்பு முறையிலான ஆய்வாக வடிவமைக்கப்பட்டுள்ளது. மேலும் ஆய்வுப் பிரதேசம், மாதிரித் தெரிவு, தரவுப் பகுப்பாய்வு முறை போன்ற பல முக்கிய விடயங்களைக் கொண்டதாக அமைகிறது. ஆய்வுப் பிரதேச குடித்தொகையின் அடிப்படையில் எளிய எழுமாற்று மாதிரி, நோக்கமாதிரித் தெரிவுகள் இதில் இடம்பெறுகின்றன.

*N.Koventhan, Senior Lecturer in Education, Department of Education and childcare, Faculty of Arts and Culture, Eastern university, Sri Lanka*

### 3.1 பொது நோக்கம்

விசேட தேவையுடைய மாணவர்கள் கற்றலில் எதிர்நோக்கும் பிரச்சினைகளை இனங்காண்பதுடன் அவர்களது கற்றலை மேம்படுத்துவதில் கணனி வழிக்கற்பித்தலில் ஆசிரியர்களின் வகிபங்கினை ஆராய்வதாகும்.

#### ஆய்வின் விசேட குறிக்கோள்கள்

1. விசேட தேவையுள்ள பிள்ளைகளை இனங்காணல்
2. விசேட தேவையுடைய மாணவர்களுக்கு வழங்கப்படும் கணனி வழிக்கற்றல் செயன்முறைகளைக் கண்டறிதல்.
3. மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் விசேட தேவையுடைய மாணவர்கள் எதிர்நோக்கும் கணனி வழிக்கற்றல் பிரச்சினைகளைக் பகுத்தறிதல்.
4. மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் விசேட தேவையுடைய மாணவர்களின் கணனி வழிக்கற்றல், கற்பித்தலில் ஆசிரியர்களின் ஈடுபாட்டை இனங்காணல்.
5. மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் விசேட தேவையுடைய மாணவர்களின் கணனி வழிக்கற்றலில் ஆசிரியர்களின் ஈடுபாட்டை அதிகரிப்பதற்கு மேற்கொள்ளக் கூடிய நடவடிக்கைகளையும், ஆலோசனைகளையும் பரிந்துரைத்தல்.

### 3.2 ஆய்வின் மாதிரித் தெரிவு

மட்டக்களப்பு மேற்குக் கல்வி வலயத்தில் தெரிவு செய்யப்பட்ட பாடசாலைகளை அடிப்படையாகக் கொண்டு விசேட தேவையுடைய மாணவர்களும், அவர்களது பெற்றோர்களும், இம்மாணவர்களுக்குக் கற்பிக்கும் ஆசிரியர்களும், அதிபர்களும் மாதிரிகளாக தெரிவு செய்யப்பட்டுள்ளனர்.

#### அட்டவணை 01

##### ஆய்வுக்காக தெரிவு செய்யப்பட்ட பாடசாலைகள்

வகை	அதிபர்கள்	ஆசிரியர்கள்	வி.தே.மாணவர்கள்	பெற்றோர்கள்
1யுட - 02	02	10	11	11
1ஊ - 05	05	19	30	30
ஐஐ - 02	02	07	12	12
ஐஐஐ - 01	01	04	07	07
<b>மொத்தம்</b>	<b>10</b>	<b>40</b>	<b>60</b>	<b>60</b>

ஆய்வுக்காக தெரிவு செய்யப்பட்ட 10 பாடசாலைகளில் உள்ள 10 அதிபர்களும் நோக்க மாதிரியின் அடிப்படையில் தெரிவு செய்யப்பட்டனர். இப் 10 பாடசாலைகளில் உள்ள விசேட தேவையுடைய மாணவர்களுக்குக் கற்பிக்கும் 40 ஆசிரியர்களுள் 20 ஆசிரியர்கள் 2:1 என்பதன் படி எளிய எழுமாற்று மாதிரி அடிப்படையில் தெரிவு செய்யப்பட்டனர். இப் 10 பாடசாலைகளில் இனங்காணப்பட்ட விசேட தேவையுடைய மாணவர்கள் ஒரு பாடசாலையில் 5 பேர் எனும் வீதத்தில் 10 பாடசாலைகளிலும் 50 மாணவர்கள் எளிய எழுமாற்று மாதிரி அடிப்படையில் தெரிவு செய்யப்பட்டனர். இப் 10 பாடசாலைகளில் இனங்காணப்பட்ட விசேட தேவையுடைய மாணவர்களது பெற்றோர் ஒரு பாடசாலையில் 5 பேர் எனும் வீதத்தில் 10 பாடசாலைகளிலும் 50 பெற்றோர்கள் எளிய எழுமாற்று மாதிரி அடிப்படையில் தெரிவு செய்யப்பட்டனர்.

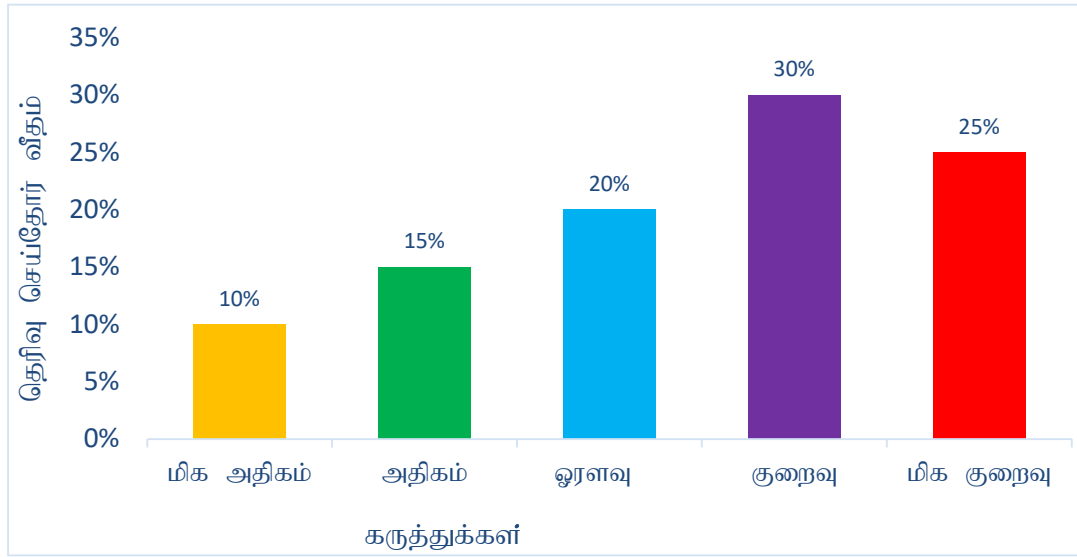
#### 4.0 தரவுப்பகுப்பாய்வும், வியாக்கியானமும், கலந்துரையாடலும் :-

விசேட தேவையுடைய மாணவர்களின் கற்றலில் ஆசிரியர்களின் பங்களிப்பை அறியும் பொருட்டு மேற்கொள்ளப்படுகின்ற இவ் ஆய்வுக்கான தகவல்கள் வினாக்கொத்துக்கள், அவதானம், நேர்காணல் மூலம் திரட்டப்பட்டன. அந்த வகையில் ஆசிரியர்கள், அதிபர்கள், பெற்றோர்கள் வழங்கிய தகவல்களை அடிப்படையாகக் கொண்டு பகுப்பாய்வு செய்து அதனை வியாக்கியானமும், கலந்துரையாடலும் செய்வதாக இவ்வாய்வு அமைகின்றது.

#### 4.1 விசேட தேவையுள்ள பிள்ளைகளை இனங்காணல்

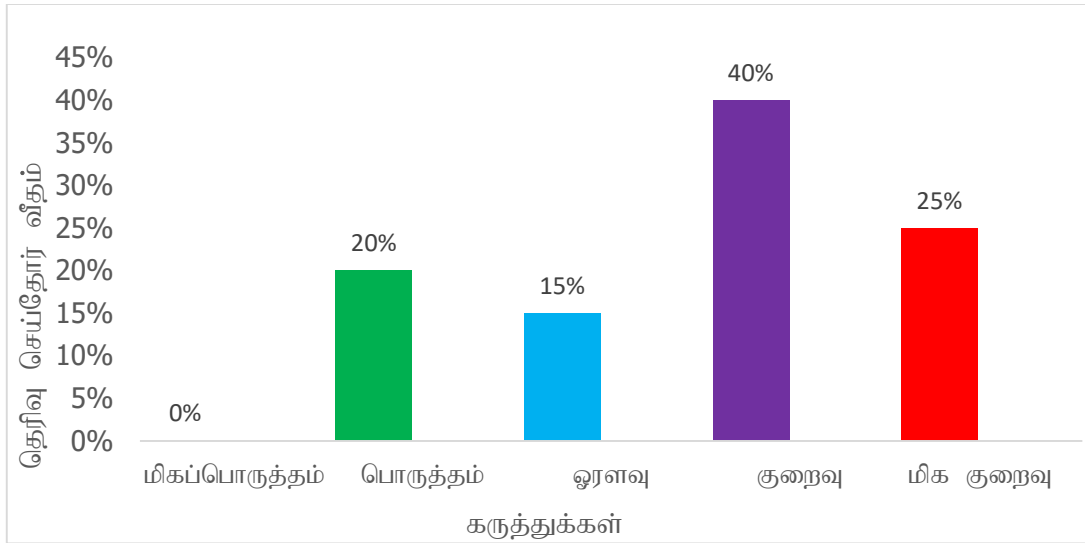
பாடசாலையில் விசேட தேவையுடைய மாணவர்களின் எண்ணிக்கை, தொடர்பில் கேட்கப்பட்ட வினாவிற்கு மிக அதிகம் என 10மு மான ஆசிரியர்களும், அதிகம் என 15மு மான ஆசிரியர்களும், ஓரளவு என 20மு மான ஆசிரியர்களும், குறைவு என 30மு மான ஆசிரியர்களும், மிகக்குறைவு என 25மு மான ஆசிரியர்களும் பதில்களை வழங்கினர். பகுப்பாய்வின்படி சாதாரண மாணவர்களை விட இம்மாணவர்களின் எண்ணிக்கை குறைவாகவே காணப்படுகின்றது. மேற்கொண்ட கலந்துரையாடலில் இதற்கான காரணங்களாக பெற்றோருக்கு இப் பிள்ளைகளின் கற்றல் தொடர்பான விழிப்புணர்வு குறைவாக உள்ளமை, பிள்ளைகளை பாடசாலைக்கு அனுப்பினால் சமூகம் தங்ககளையும், இப் பிள்ளைகளையும் தப்பான கண்ணோட்டத்தில் பார்க்கும் என்பதற்காக பாடசாலைக்கு அனுப்பாமல் மறைத்து வைத்துள்ளனர் எனக்கூறினர். விசேட தேவையுடைய பிள்ளைகளின் எண்ணிக்கை 6247 ஆக அதிகரித்துள்ளபோதிலும் 292 விசேட தேவையுடைய பிள்ளைகளே பாடசாலைக் கல்வியை தொடர்கின்றனர் (முநவாநநளையசயஇ 2014).

மேற்கூறிய ஆய்விலும் இம்மாணவர்கள் பாடசாலைக் கல்வியை தொடர்வது குறைவாகவே உள்ளது. எனவே ஆய்வாளனின் இத் தரவுப்பகுப்பாய்வை உறுதிப்படுத்துகின்றது.



உரு 1 பாடசாலையில் விசேட தேவையுள்ள பிள்ளைகள்

4.2 விசேட தேவையுடைய மாணவர்களுக்கு வழங்கப்படும் கணனி வழிக் கற்றல் செயன்முறைகள்.

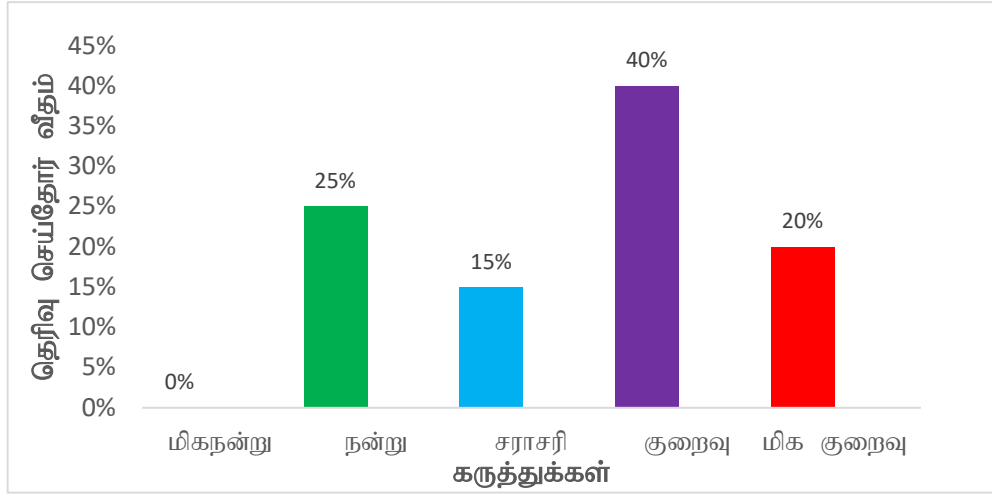


உரு 2 பாடசாலையின் வகுப்பறைச் சூழல்.

விசேட தேவையுடைய மாணவர்களின் கணனி வழிக்கற்றலுக்கான வகுப்பறைச் சூழல் தொடர்பாகப் பெறப்பட்ட தரவுகளின் பகுப்பாய்வு அடிப்படையில், பொருத்தம் என 20% மாண ஆசிரியர்களும், ஓரளவு என்று 15% மாண ஆசிரியர்களும், குறைவு என 40% மாண ஆசிரியர்களும், மிகக் குறைவு என 25% மாண ஆசிரியர்களும் பதில்களை வழங்கியுள்ளனர். இதிலிருந்து இம்மாணவர்களின் கற்றலுக்கு பொருத்தமான வகுப்பறைச் சூழல் குறைவாகவே உள்ளது. அவதானம், கலந்துரையாடல் மூலம் வகுப்பறையில் அதிக மாணவர்கள் காணப்படுகின்றமை, கணனி வசதிகள் இல்லாமை, இடவசதி குறைவு, இம்மாணவர்களின் விசேட தன்மைக்கு ஏற்ப தளபாட வசதிகள் இல்லாமை போன்ற காரணங்களால் இம்மாணவர்களது கற்றலுக்கு பொருத்தமான வகுப்பறை குறைவாக உள்ளது.

விசேட தேவையுடைய மாணவர்களுக்கான வகுப்பறைச் சூழல் சிறந்த முறையில் இல்லை இதனால் இம்மாணவர்கள் கற்றலை சிறப்பாக மேற்கொள்ள முடியாதவர்களாக உள்ளனர். வகுப்பறைச் சூழல் சிறப்பாக அமையும் போதுதான் விசேட தேவையுடைய மாணவர்கள் சிறப்பாக கற்றலை மேற்கொள்வார்கள் (பக்கீர் ஜி.பர், 2004). மேற்கூறிய ஆய்வு முடிவு ஆய்வாளனின் இந் ஆய்வை முன்னகர்த்திச்செல்வதற்கு உதவுகிறது.

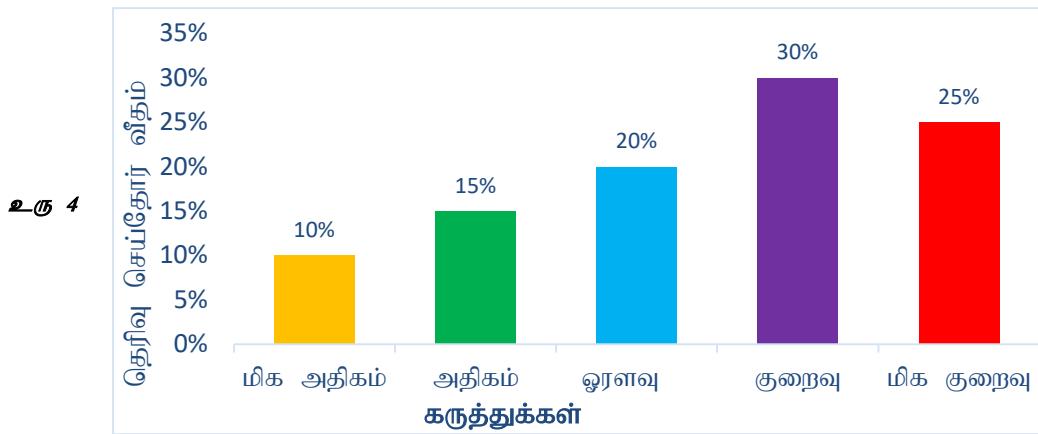
#### 4.3 மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் விசேட தேவையுடைய மாணவர்களின் கணனி வழிக்கற்றலில் உள்ள பிரச்சினைகள்



உரு 3 மாணவர்களின் சுய கற்றல்.

விசேட தேவையுடைய மாணவர்கள் சுய கணனி வழிக்கற்றலில் ஈடுபடுவதுண்டா? என ஆசிரியர்களிடம் கேட்கப்பட்ட வினாவுக்கு 25% மாணவர்கள் நன்று எனவும், சுமார் 15% மாணவர்கள் சராசரி எனவும், 40% மாணவர்கள் குறைவு எனவும், 20% மாணவர்கள் மிகக் குறைவு எனவும் பதிலளித்தனர். இதனால் சுய கற்றலில் ஈடுபடுவது குறைவு. என்ற முடிவிற்கு வரமுடியும். கலந்துரையாடலில் சுய கற்றல் குறைவிற்கான காரணங்களாக ஆசிரியரின் உதவி தேவை, அடிப்படை எழுத்தறிவு குறைவு, அவர்களுக்கான அன்பு, காப்புத் தேவைகள் குறைவாக உள்ளது எனக்கூறினர். விசேட தேவையுடைய மாணவர்களிடையே காணப்படும் தாழ்வு மனப்பாங்குகள், பிறரில் தங்கி நிற்கும் மனநிலை, முன்வரத் தயக்கம், கூச்சகவம், ஞாபகசக்தி குறைவு, பாடசாலைக்கு தொடர்ச்சியாக சமூகமளிக்கமை போன்ற விடயங்கள் இம்மாணவர்களின் சுய கற்றலை குறைக்கின்றது (யுபெநயற யனெ ருரட ஐனெமையடி 2011). எனவே மேற்கூறிய ஆய்வு முடிவு ஆய்வாளரின் இந் ஆய்வை முன்கொண்டு செல்வதற்கு உதவுகின்றது.

#### 4.4 மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் விசேட தேவையுடைய மாணவர்களின் கணனி வழிக்கற்றலில் ஆசிரியர்களின் ஈடுபாடு.



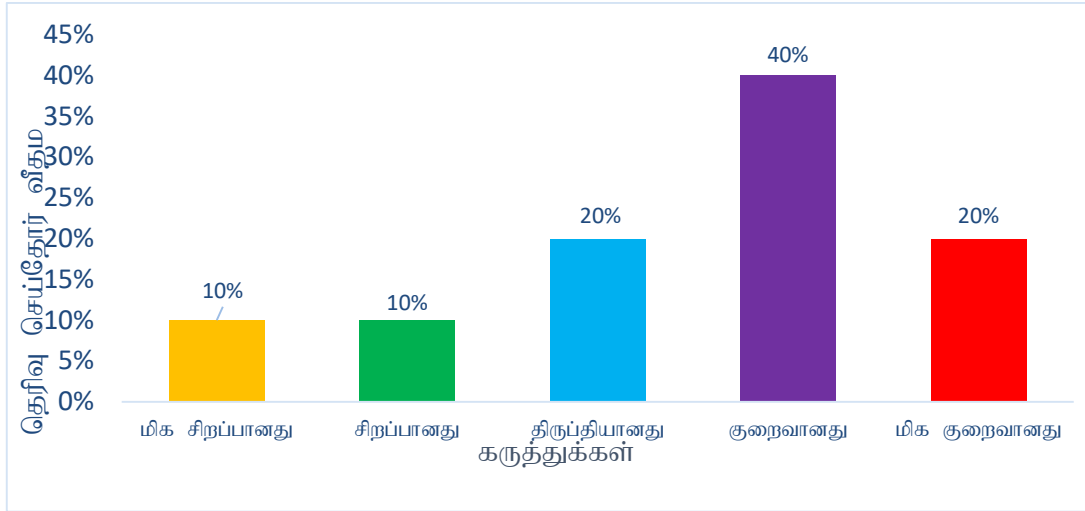
உரு 4

ஆசிரியர்களின் ஒத்துழைப்பு.

விசேட தேவையுடைய மாணவர்கள், ஆசிரியரின் கற்பித்தல் செயற்பாட்டிற்கு ஒத்துழைப்பு வழங்குகின்றனரா? என கேட்கப்பட்ட வினாவிற்கு மிக அதிகம் என 10% மாணவர்களும், அதிகம் என 15% மாணவர்களும், ஓரளவு என 20% மாணவர்களும், குறைவு என 30% மாணவர்களும், மிகக்குறைவு என 25% மாணவர்களும் பதில்களை வழங்கினர். பகுப்பாய்வின்படி மாணவர்களது ஒத்துழைப்பு குறைவாகவே காணப்படுகின்றது. கலந்துரையாடலின் படி ஆ ஒத்துழைப்பு குறைவதற்கான காரணங்களாக பெற்றோர்களின் கவனம் இன்மை, ஆசிரியருடன் தொடர்பாடல் குறைவு, விசேட தேவையுடைய பிள்ளைகளை கவனிப்பது குறைவு, பெற்றோர் வறுமை காரணமாக தூர இடங்களுக்கு தொழிலுக்கு செல்கின்றமை போன்றவற்றை முன்வைத்தனர்.

பெற்றோர் பிள்ளைகள் மீது செலுத்தும் அக்கறை குறைவு, பெற்றோர் பிரிந்து வாழ்தல், பெற்றோரின் கல்வி மட்டம் குறைந்த தன்மை, பொருளாதாரத்தில் பின்தங்கி உள்ளமை போன்ற காரணங்களால் விசேட தேவையுடைய மாணவர்களுக்கான கற்றல் செயற்பாட்டிற்கு பெற்றோரால் ஒத்துழைப்புகள் கூடுதலாக வழங்க முடியாமல் உள்ளனர் (சோபா, 2013). மேற்கூறிய ஆய்வு முடிவுடன் விசேட தேவையுடைய மாணவர்களின் பெற்றோரின் ஒத்துழைப்பு என்னும் விடயத்தில் காணப்பட்ட முடிவுகள் ஆய்வாளனின் இந்த ஆய்வை மேலும் மெருகூட்டுவதாக அமைகிறது.

**4.5 மட்டக்களப்பு மேற்கு கல்வி வலயத்தில் விசேட தேவையுடைய மாணவர்களின் கற்றலில் ஆசிரியர்களின் ஈடுபாட்டை அதிகரிப்பதற்கு மேற்கொள்ளப்படக்கூடிய நடவடிக்கையும், ஆலோசனையும்**



**உரு 5 ஆசிரியர்களின் ஈடுபாடு**

விசேட தேவையுடைய மாணவர்களின் கற்றலில் ஆசிரியர்களின் ஈடுபாடு உள்ளதா? எனக் பெற்றோரிடம் கேட்கப்பட்ட வினாவிற்கு மிகச் சிறப்பாக என 10% மான பெற்றோரும், சிறப்பானது என 10% மான பெற்றோரும், திருப்தியாக என 20% மான பெற்றோரும், குறைவாக என 40% மான பெற்றோரும், மிகக் குறைவாக என 20% மான பெற்றோரும் பதில் வழங்கினர். ஆகவே ஆசிரியர்களின் ஈடுபாடு குறைவாக உள்ளது. மேற்கொண்ட கலந்துரையாடலில் இதற்கான காரணமாக ஆசிரியர்கள் அதிகாலையில் வேலைக்கு செல்கின்றமை, கணனி வசதிகள் இல்லாமை பாடசாலையிலிருந்து வீடு அதிக தூரத்தில் உள்ளமை போன்ற காரணங்களால் இம் மாணவர்களின் கற்றலில் ஆசிரியர்களின் ஈடுபாடு குறைகின்றது. ஆசிரியர்கள், வலயக்கல்வி அதிகாரிகள், சமூக அமைப்புக்கள் இணைந்து பெற்றோருக்கு தகுந்த ஆலோசனைகள் வழங்குவதன் மூலம் இம்மாணவர்களின் கற்றலில் பெற்றோரின் ஈடுபாட்டை அதிகரிக்க முடியும் (டிசைவழி உவமை 2018).

#### 5.0 முடிவுகளும், விதப்புகளும்

- ஆய்வுக்குட்படுத்திய பாடசாலைகளில் சாதாரண மாணவர்களை விட விசேட தேவையுடைய மாணவர்களின் எண்ணிக்கை குறைவாவே உள்ளது.
- ஆய்வுக்குட்படுத்திய பாடசாலைகளில் குறிப்பாக பார்வைக் குறைபாடுடை பிள்ளைகள், கேட்டல் குறைபாடுடைய பிள்ளைகள், மெல்லக் கற்கும் பிள்ளைகள், உடல் குறைபாடுடைய பிள்ளைகள் காணப்படுகின்றனர்.
- இம் மாணவர்களிடையே கணனி பாவனை மிக குறைந்தே வருகின்றது.
- ஆய்வுக்குட்படுத்தப்பட்ட பாடசாலைகளில் வகுப்பறை இம்மாணவர்களின் கணனி வழிக்கற்றல், கற்பித்தலுக்கு பொருத்தமானதாக இல்லை.
- ஆய்வுக்குட்படுத்திய பாடசாலைகளில் இம்மாணவர்கள் கிரகித்தல், மனனம் செய்தல், பிரச்சினை தீர்த்தல், பகுத்தல், தொகுத்தல், முதலிய செயற்பாடுகளில் குறைந்த மட்டத்தில் உள்ளனர்.
- கேள்வித் தாள்களைத் தயாரிக்க, விடைத் தாள்களை மதிப்பிட மற்றும் தேர்வு முடிவுகளைப் பகுத்தறிய கணனி பயன்படுத்தப்படுவது குறைவாக உள்ளது.
- ஆய்வுக்குட்படுத்திய பாடசாலைகளில் இம்மாணவர்களின் கற்றல் நிலை குறைவாக உள்ளது. இம்மாணவர்களை சாதாரண மாணவர்களுடன் ஒப்பிடும் பொழுது குறுகிய நேரத்தில் களைப்படைந்து விடுவர். போசாக்கு இல்லாத காரணத்தினால் நீண்ட நேர கற்றல் செயற்பாட்டில் இவர்களால் ஈடுபடமுடிவதில்லை.

- ஆய்வுக்குட்படுத்திய பாடசாலைகளில் விசேட தேவையுடைய மாணவர்களின் கற்றலுக்கு ஆசிரியர்கள் பங்களிப்பு வழங்குவது குறைவாகவே காணப்படுகிறது. இது வரைக்கும் எந்தவொரு செயற்றிட்டங்களும் இம் மாணவர்களின் கற்றலை மேம்படுத்துவதற்கு குறித்த பாடசாலைகளில் நடைபெறவில்லை.
- இம் மாணவர்களின் பெற்றோரின் கல்வி மட்டம் குறைவாக உள்ளதனால் இம் மாணவர்களுக்கு வீட்டில் கல்வி கற்றுக்கொடுப்பது சவாலாக உள்ளது.
- குறித்த ஆய்வுப்பிரதேசத்தில் இம்மாணவர்கள் பல்கலைக்கழகத்திற்கு செல்வது குறைவு.
- ஆய்வுக்குட்படுத்திய பாடசாலைகளில் விசேட தேவையுடைய மாணவர்களின் பெற்றோர்கள் இப் பிள்ளைகளின் எதிர்காலம் தொடர்பில் விழிப்புணர்வு அற்றவர்களாகவும், இப்பிள்ளைகள் விசேட தேவையுடையவர்களாக காணப்படுவதால் இவர்கள் பாடசாலைக்கு சென்று எதையும் சாதிக்கமாட்டார்கள் என்துடன் தங்களுக்கு பணச்செலவுதான் அதிகரிக்கும் என்பதால் தன்னம்பிக்கை குறைந்தவர்களாகவும் காணப்படுகின்றனர்.
- ஆய்வுக்குட்படுத்திய பாடசாலைகளில் பெற்றோர்கள் இம்மாணவர்களுக்கு பொருத்தமான இணைப்பாடவிதான செயற்பாடுகளில் ஈடுபடுவது குறைவாகவே காணப்படுகிறது.

#### விதப்புரைகள்

- இம் மாணவர்களின் விசேட தன்மைக்கு ஏற்ப அவர்களை இனங்காணல் வேண்டும். ஆசிரியர்கள் விசேட தேவையுடைய மாணவர்களின் கற்றலுக்கு உதவ வேண்டும், குறிப்பாக பெற்றோர்களும் இப் பிள்ளைகளை சரியான முறையில் இனங்கண்டு அவர்களை கையாள வேண்டும். பிரதேசத்திற்கு பொறுப்பான சமூக சேவை அதிகாரி விசேட தேவையுடைய பிள்ளைகளை இனங்கண்டு பாடசாலையில் சேர்ப்பதற்கான வழிவகைகளை பெற்றோருக்கு செய்து கொடுத்தல் வேண்டும்.
- குறிப்பிட்ட கால அளவில் இப்பிரச்சனையைத் தீர்க்க ஆசிரியரை அதிகளவில் கற்பித்தலில் ஈடுபடச் செய்விப்பதே கணினி வழிக் கற்பித்தலின் அடிப்படை நோக்கமாகும். ஆசிரியர் மாணவர் மற்றும் கணினி ஆகிய மூன்றும் இணைந்து செயல்படும்போதே கணினி வழிக் கல்வி சிறப்பாக அமையும்.
- தொழிநுட்ப முன்னேற்றத்திற்கு ஆசிரியர்களின் இடத்தினை பெற்றுக்கொள்ள முடியாவிட்டாலும், கணினி மென்பொருள் மற்றும் இணையப்பாவனை என்பன இன்று பிள்ளைகளின் கற்றல் உபகரணங்களாக மாறியுள்ளன (புசநநகெநடைன ஞரணரமடை 1998).
- கணினி வழிக் கற்றலுக்கு மாணவரை அதிகமாகக் ஊக்கப்படுத்துதல் வேண்டும்.
- மாணவர்கள் அவர்களின் நோக்கங்களை அடைவதற்கு ஏற்ற அனுபவங்களைக் கொடுக்கிறது. மீளக் கற்றலுக்கு கணினியின் பயன்பாடு மிக முக்கியம்.
- கணினி வழிக் கற்பித்தலில், கற்றல் - கற்பித்தலை மேம்படுத்த பல வகையான மென்பொருள்கள் (ஞழகவறயசந) உள்ளன. மாணவர்கள் இந்த மென் பொருள்களில் பதிந்துள்ள தகவல்களைத் திரும்பத் திரும்ப பயிற்சி செய்து (னுசடை யனெ ிசயஉவடைந) மனதில் இருத்திக் கொள்ள வேண்டும். இல்லையெனில், கற்றது குறுகிய காலத்திற்கே நினைவில் இருக்கும். எனவே சரியாக நடைமுறைப்படுத்தல்.
- விசேட தேவையுடைய மாணவர்களுக்கேற்ப வகுப்பறைச் சூழல் அமைக்கப்பட வேண்டும். அதற்குப் பாடசாலையின் உதவியுடன் பெற்றோர், கல்வி அதிகாரிகளின் உதவிகளைப் பெற வழிவகுத்தல் அவசியம். இம்மாணவர்களின் விசேட நிலையை ஆசிரியர் உணர்ந்து பொருத்தமான கற்றல், கற்பித்தல் செயற்பாடுகளை முன்னெடுக்கவேண்டும்.
- இம் மாணவர்களின் விசேட தன்மைக்கு ஏற்ப அவர்களுக்கான கற்பித்தல் செயற்பாடுகளை மேற்கொள்ள வேண்டும். அடிப்படை தேர்ச்சிகளை விசேட தேவை உடையவர்களுக்கு ஏற்ற விதத்தில் நெகிழ்ச்சி தன்மை உடையதாக பிரத்தியேகமாக தயாரித்தல்.
- ஆசிரியர் பாட, இணைப்பாடவிதான செயற்பாடுகளில் இம்மாணவர்களுக்கு முன்னுரிமை வழங்கி ஊக்குவிக்க வேண்டும்
- மேலும் பயிற்றப்பட்ட விசேட தேவையுடைய மாணவர்களுக்கு பொருத்தமான ஆசிரியர்களை நியமித்தல். விசேட செயற்றிட்டங்களை மேற்கொண்டு இம்மாணவர்களின் கற்றலை உறுதிப்படுத்தல். அதிபர் சிறந்த முறையில் மேற்பார்வை செய்தல் வேண்டும்.
- செயல்நிலை ஆய்வுகளில் ஆசிரியர் ஈடுபட்டு இம்மாணவர்கள் கற்றலில் எதிர்நோக்கும் பிரச்சினைகளை இனங்கண்டு பரிகார கற்பித்தல் செயற்பாடுகளை மேற்கொள்ள வேண்டும்.
- இம்மாணவர்களுக்கான பயிற்சிப் பாசறைகள், ஆலோசனை நிகழ்ச்சித்திட்டங்களை மேற்கொள்ள வேண்டும்.





## Deep Analyses of the Evolution of Tamil characters from Stone Inscriptions: Digital Conservation perspective

Karishma V R<sup>1</sup>, P Uma Maheswari<sup>2</sup>

<sup>1,2</sup> Department of Computer Science and Engineering, College of Guindy (CEG), Anna University (AU), India

---

### ABSTRACT

---

---

#### Keywords:

Stone Inscriptions  
Evolution of Tamil  
Classification  
Recognition

Tamil is the classical language of India. It has literary and epigraphical evidence, that its origin is ancient and has an independent tradition. Stone inscriptions are a reliable source of information about ancient India. The challenges in Extracting the characters from the stone inscription are differentiating the foreground pixel from the background stone images, perspective distortion, different light illumination, the same kind of background/foreground, eroded stones, lack of shape and size of the text, and poor writing skill of the inscriber. As ancient Tamil characters have a similar pattern to different characters, it is very difficult to classify the characters. So, it is very important for the researchers to understand the character pattern and to classify them accordingly to train the model. For better recognition, proper classification of characters is required. In this paper, deep analyses of the evolution of Tamil characters from the stone inscriptions are made with the proper ground truth for the ancient characters. This study will give an insight into the language evolved over the period and it provides a strong base for the model to recognize the characters optimally.

---

#### Corresponding Author:

Karishma V R,  
Department of Computer Science and Engineering, College of Guindy (CEG), Anna University (AU), India  
Email: karishmaraghu5@gmail.com

---

### 1. INTRODUCTION

The study of ancient inscriptions is significant for understanding history. The scripts used in these inscriptions may date from different eras and are classified according to the ruling dynasty at the time. Tamil-Brahmi, commonly known as Tamizhi or Damili, was a southern Indian version of the Brahmi script. It is the first developed script during Ashoka's period. The Tamil-Brahmi script has been paleographically and stratigraphically dated between the 3<sup>rd</sup> century BCE and the 1<sup>st</sup> century CE, and it is the oldest known writing system found in several regions of Tamil Nadu, Kerala, Andhra Pradesh, and Sri Lanka. Vattezhuthu probably started developing from Tamil-Brahmi around the 4<sup>th</sup> or 5<sup>th</sup> century CE. As this script was written with more

ursive lines so, it was known as Vattezhuthu and Vattam. Most of the Vattezhuthu scripts are from Kerala and Sri Lanka regions. Parallelly, the Pallavas popularised the practice of writing Sanskrit letters in Tamil Nadu, commonly known as the Grantha script. This continued for nearly two centuries i.e., from the 4<sup>th</sup> – the 6<sup>th</sup> century. The Tamil script evolved from the Grantha script around the 7<sup>th</sup> century CE. The challenges in Extracting the characters from the stone inscription are differentiating the foreground pixel from the background stone images, perspective distortion, different light illumination, the same kind of background/foreground, eroded stones, lack of shape and size of the text, and poor writing skill of the inscriber. As ancient Tamil characters have a similar pattern to different characters, it is very difficult to classify the characters. Robust feature extraction is very important to improve the performance of the Ancient Tamil character recognition system. Many Deep Learning models work efficiently for the recognition of characters. But not all the characters are recognized. Most of the characters are not addressed because of the complexity of understanding the ancient Tamil characters. This paper focuses on the understanding evolution of Tamil characters and their variations, to train the deep learning model adequately for better recognition.

## 2. EVOLUTION OF TAMIL SCRIPTS

Tamil is the classical language of India. Tamil is more than a language; it is an intrinsic aspect of Tamil culture. It is difficult to fix the age of the evolution of the language because of its rich vocabulary. One of the pieces of evidence of the existence of the Tamil language is stone inscriptions. The scripts featured in these inscriptions are from various eras and are grouped according to the governing dynasty at the time. Figure 1. Depicts the evolution of Tamil scripts from the olden era to the modern era.

HISTORY OF TAMIL SCRIPT																													
நூற்றாண்டு	a ā i ī u ū ē ē ai o ō										நூற்றாண்டு	Kñcñṭṇṭnṭnṃyṛlv!lṛm																	
Century	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஓ	Century	க்	ங்	ச்	ஞ்	ட்	ண்	த்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ம்	
BC 3 <sup>rd</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	BC 3 <sup>rd</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒
AD 2 <sup>nd</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 2 <sup>nd</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 3 <sup>rd</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 3 <sup>rd</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 4 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 4 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 5 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 5 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 6 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 6 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 7 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 7 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 8 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 8 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 9 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 9 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 10 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 10 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 11 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 11 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 12 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 12 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 13 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 13 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 14 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 14 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 15 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 15 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 16 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 16 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 17 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 17 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 18 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 18 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	
AD 19 <sup>th</sup> C	𑌀	𑌁	𑌂	𑌃	𑌄	𑌅	𑌆	𑌇	𑌈	𑌉	AD 19 <sup>th</sup> C	𑌕	𑌔	𑌐	𑌕	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	𑌒	

Figure 1. Evolution of Tamil Script.  
(Source courtesy “Dakshina Chitra”, Chennai)

## 2.1 Tamil-Brahmi

Brahmi is the earliest Indian alphabetical script. As per its regional variations, it is identified as Tamil-Brahmi, Asokan-Brahmi, Northern-Brahmi, Southern-Brahmi, and Sinhala-Brahmi[1], [10]. All modern Indian scripts are the evolved forms of Brahmi. Tamil-Brahmi inscriptions about 93 in number are found on natural caverns and rock beds in 31 places in Tamil Nadu[37]. Further potteries from excavations,

coins, seals, and rings collected from river beds also bear the Tamil-Brahmi script. The distribution of Tamil-Brahmi inscriptions covers roughly all parts of Tamil Nadu. Due to overseas trade, a few pottery pieces and a touchstone bearing personal names in this script are recently noticed in Egypt and Thailand also. Tamil-Brahmi inscriptions are classified as Early Tamil-Brahmi and Late Tamil Brahmi and are dated between the 3<sup>rd</sup> century BCE and the 3<sup>rd</sup> century CE. Figure 2. Shows the Tamil Brahmi inscription of Jambaimalai. Figure 3. Represents 2<sup>nd</sup> century BCE Tamil Brahmi inscription from Arittapatti, Madurai India.

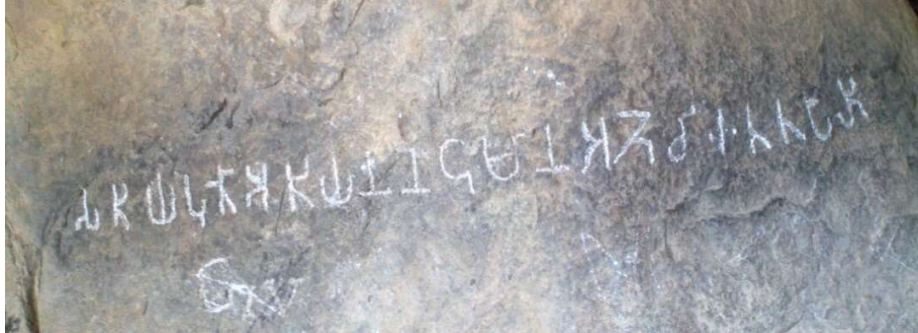


Figure 2. Jambai Tamil Brahmi Inscription.



Figure 3. Tamil Brahmi inscription from Arittapatti, Madurai India.

## 2.2 Vattezhuthu

Vattezhuthu probably started developing from Tamil-Brahmi around the 4<sup>th</sup> or 5<sup>th</sup> century CE[10], [11]. As this script was written with more cursive lines. so, it was known as Vattezhuthu and Vattam. It is also known as Tekkan Malayalam and *Nanamona*. Earlier inscriptions in this script were mostly noticed in southern districts, and occasionally in other areas. So far, it is not noticed in Thanjavur and adjoining regions of the Kaveri River delta known earlier as Chola mandalam. Due to special courses and workshops on Tamil Epigraphy conducted by the State Department of Archaeology to students, teachers, and others interested in epigraphy, in recent decades many important inscriptions in this script are noticed. They are the Pulankurichi inscription in the Sivaganga district, many memorial stone inscriptions in northern districts of Tamil Nadu, and several in Villupuram and adjoining districts. These discoveries now help palaeographers to form a complete picture of this script's evolution[37]. Sendan Maran's irrigation inscription in vattezhuthu 7<sup>th</sup> century CE, Vaigai river bed, Madurai is shown in figure 4. Donative inscription in vattezhuthu, Pandya Maranjadaian 8<sup>th</sup> century CE, Tiruttangal, Virudhunagar district is shown in figure 5.

## 2.3 Grantha

The development of the Grantha script in Tamil Nadu may be divided into four periods. The archaic and ornamental, the transitional, the medieval, and the modern. Archaic and ornamental variety



commonly known as Pallava Grantha. Mahendravarman's Tiruchirappalli rock-cut cave and other cave temple inscriptions, Narasimhan's Mamallapuram, Kanchi Kailasanatha and Saluvankuppam temple inscriptions, Mutharaiyar's Senthalai inscriptions are examples of this variety. The transitional variety of Grantha inscriptions roughly belongs to three centuries between the 6<sup>th</sup> century CE and the 9<sup>th</sup> century CE. Later Pallava's (Nandivarman's Kasakudi, Udayendram plates, etc.) and Pandyan Nedunjadaiyan's Anaimalai inscriptions are samples of this. The medieval variety is from about the 9<sup>th</sup> century CE to the 12<sup>th</sup> century CE. Inscriptions of imperial Cholas of Thanjavur are examples of this. The modern variety belongs to the later Pandya and Vijayanagara periods. After the introduction of printing machines, many Sanskrit books transcribed from palm leaves were printed in Grantha script[12], [37]. Figure 6. Shows inscription in grantha script, Mutharaiyar Chiefs, 9<sup>th</sup> century CE, Sendalai, Thanjavur district. Figure 7. depicts Mahendravarman, Pallava king, in the stone inscription which is in grantha script, dated about 7<sup>th</sup> century CE, Tiruchirappalli.

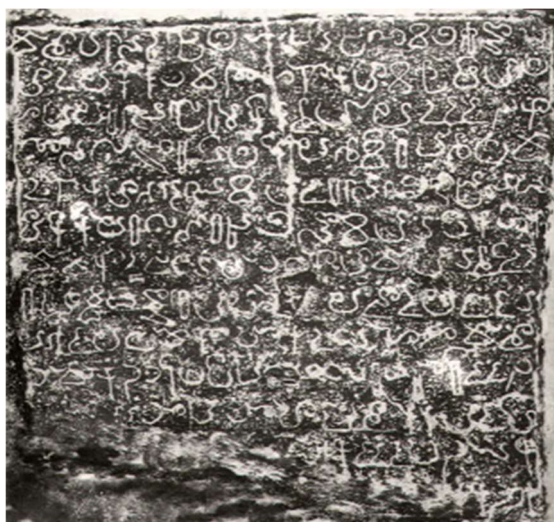


Figure 4. Sendan Maran's Irrigation  
Inscription in Vattezhuthu.  
(Source courtesy Department of Archaeology)



Figure 5. Donative Inscription in  
Vattezhuthu.  
(Source courtesy Department of Archaeology)

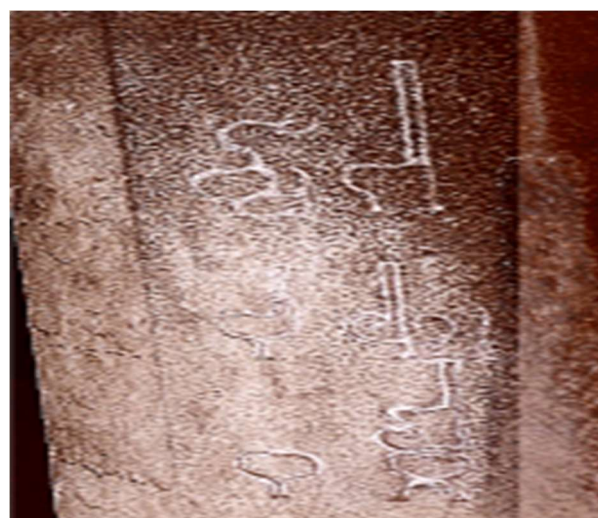
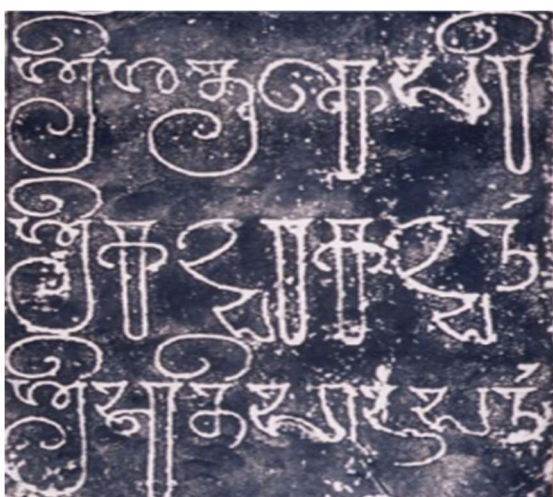


Figure 6. Inscription in Grantha Script,  
Thanjavur.  
(Source courtesy Department of Archaeology)

Figure 7. Inscription in Grantha Script,  
Tiruchirappalli.  
(Source courtesy Department of Archaeology)

### 3. CHALLENGES IN EXTRACTING CHARACTERS FROM STONE INSCRIPTION IMAGES

The character extraction from the camera-captured stone inscriptions is difficult due to various factors like light illumination, similar background, and foreground, eroded stones, and lack of text shape, size, and noise. Other than the technical issues in handling images there are more challenges in extracting characters from the images because the image contains characters from ancient Tamil scripts. The variations in the character formation over the evolution of Tamil Script are a major impact on recognition. Text line boundaries are muddled as the interline spacing becomes narrow due to the crowded writing style. Characters overlap making them difficult to separate. From the 11<sup>th</sup> century CE (urban Palaeography) the writing of the Tamil script had a standard pattern so, it was easy to classify the characters with reference to modern Tamil script and there will be a good recognition rate. But the 6<sup>th</sup> century CE to 10<sup>th</sup> century CE (rural Palaeography) has different writing styles and for most of the stone inscriptions, the ground truth is not found, so it's difficult to understand the characters without epigraphist guidance.

### 4. FRAMEWORK FOR CHARACTER RECOGNITION

The essential processes in processing any image are image acquisition, image enhancement, segmentation, and feature extraction. The classification is based on the extracted features, and the character is recognized. The accuracy of the recognition depends on training the model with the proper classification by understanding the unique patterns of the characters. In this paper, we mainly concentrate on various recognition models. So, a detailed discussion of feature extraction and classification models is done.

#### 4.1. Feature extraction models

Feature extraction is used in the approach for dimensionality reduction. It is possible to handle groupings of raw data. By selecting and grouping variables into features, it is possible to extract the best feature from data sets. It is helpful whenever it's needed to remove resources without losing important data and to reduce the amount of duplicated data in the data set. It provides the pertinent shape information found in a pattern. It can facilitate effective classification methods. It has been transformed into a representation of a feature vector map if its data sets are very dimensional. This stage can aid in increasing the recognition rate[26].

Table 1. Feature extraction methods

S.no	Title of Research	Feature Extraction methods
1	A hybrid group search optimization: firefly algorithm-based big data framework for ancient script recognition- Soft Computing – Springer[31] 2020.	Hough Transformation is used for detecting straight lines, circles, and ellipses. Group Search Optimization- Firefly- is for feature selection
2	Soft computing approaches for character credential and word prophecy analysis with stone encryptions- Soft Computing, Springer [33] 2020.	Zoning feature- to extract the geometric features of the images, HOG feature- is to determine the object, Zernike feature- Zernike feature distinctively describes functions on the entity disk and analyzes the shape of the object,

3	Pattern Matching Model for Recognition of Stone Inscription Characters- The Computer Journal. [11] 2021.	Speeded-Up Robust Feature-it has three main stages Feature Point Detection, Confined Region Description, and Feature Matching.
4	Repossession and recognition system: transliteration of antique Tamil Brahmi typescript- Current Science. [25] 2021.	Zernike moment- distinctively describes functions on the entity disk and analyzes the shape of the object and zoning feature- to extract the geometric features of the images
5	A novel nearest interest point classifier for offline Tamil handwritten character recognition- Pattern Analysis and Applications, Springer. [24] 2020.	Taking the Nearest Interest point from Speeded-Up Robust Feature-it has three main stages Feature Point Detection, Confined Region Description, and Feature Matching.
6	An NN-based analytic approach to symbol level recognition for degraded Bengali printed documents, springer. [8] 2020.	Image attributes include things like circles, lines, textures, and contour shapes. Before designing a feature extraction network, the features that need to be extracted should be determined. The training phase of CNN offers automated feature extraction. There are several convolutional layers and pooling layer pairings in the feature extraction network.

#### 4.2. Recognition Models

The analysis process ends with this step, which involves predicting the target classes. A target, label, or category is the name given to the group. Supervised learning is used when the recognition is based on the labeled class. In doing so, the algorithm practices each image individually and picks up certain techniques. It begins doing the recognition by the learned technique. Unsupervised learning is used when it is unknown what the labels are. It uses a grouping method to learn on its own. This will allow it to identify the character. Reinforcement learning operates by evaluating prior experience and generating independent judgments in light of collected facts. Additionally, it's a trial-and-error procedure.

Table 2. Recognition Models

S.no	Title of Research	Recognition models
1	A hybrid group search optimization: firefly algorithm-based big data framework for ancient script recognition- Soft Computing – Springer[31] 2020.	Artificial Neural Network (ANN), in classification tasks, has the capability of handling documents with high-dimensional features and documents with noisy and contradictory data.
2	Soft computing approaches for character credential and word prophecy analysis with stone encryptions- Soft Computing, Springer [33] 2020.	SVM classification algorithm to maximize class margin and minimize error. convolutional neural networks (CNNs) and ImageNet architecture



		to recognize the character after image classification. It is a powerful deep-learning algorithm that operates explicitly on images and is also used to decrease the error rate and accurately identify the data.
3	Pattern Matching Model for Recognition of Stone Inscription Characters- The Computer Journal. [11] 2021.	BoG- code word model for each character to reduce the complexity, and using similarity matching recognition is done.
4	Repossession and recognition system: transliteration of antique Tamil Brahmi typescript- Current Science. [25] 2021.	Neural Network is a machine learning tool with a gradient descent method for recognizing ancient characters.
5	A novel nearest interest point classifier for offline Tamil handwritten character recognition- Pattern Analysis and Applications, Springer. [24] 2020.	The nearest Interest Point Classifier is working based on similarity matching code.
6	An NN-based analytic approach to symbol level recognition for degraded Bengali printed documents, springer. [8] 2020.	LSTM neural network is a special kind of RNN (Recurrent Neural Network) that is capable of remembering information for a long period and eliminating long-term dependence problems.

## 5. CONCLUSION

A trustworthy source of knowledge about ancient India is stone inscriptions. Separating the foreground pixels from the background stone images, perspective distortion, different light illumination, the same type of background/foreground, eroded stones, lack of shape and size of the text, and the inscriber's poor writing ability are challenges in extracting the characters from the stone inscription. It is quite challenging to categorize the ancient Tamil characters because they share a pattern with other characters. Therefore, researchers must comprehend character patterns and classify them appropriately in order to train the model. Correct character classification is necessary for improved recognition. Various challenges in the character extraction of character from the stone inscription images. These analyses will provide insight into the way the language changed through time and a solid foundation for the model to best recognize the characters.

## ACKNOWLEDGEMENTS

The Tamil Virtual Academy and the Tamil Nadu state department of archaeology provide non-financial assistance for this initiative for Reading Letters in Ancient Form.

## REFERENCES

- [1] A Vidhyavani, T manoranjitham, "A Survey on Recognition of Ancient Tamil Brahmi character from epigraphy", Journal of Pharmaceutical Negative Results, vol. 13, no. 3, 2022.
- [2] Ankan Kumar Bhunia, Aishik Knower, Ayan Kumar Bhunia, "Script identification in natural scene image and video frames using an attention-based Convolutional-LSTM network", Pattern Recognition, Elsevier, vol. 85, pp. 172-184, 2019.
- [3] Bapu Chendage , Rajivkumar Mente , Vikas Magar, " A Survey on Ancient Marathi Script Recognition from Stone Inscriptions", Compliance Engineering Journal, vol. 11, no. 8, pp. 142-157, 2020.

- [4] Bapu Chendage, Rajivkumar Mente, "Study On Ancient Marathi Script Improvement Using Digital Image Processing Techniques", Journal of the Maharaja Sayajirao University of Baroda, vol. 55, no.3, pp. 26-38, 2021.
- [5] Chunxia Zhang, Longxue Li, Xudong Li, "A Survey of Chinese Character Recognition Research Based on Deep Learning", International Conference on Network and Information Systems for Computers, pp. 926-931, 2021.
- [6] Dhanya Sudarsan, Deepa Sankar, "A Novel Complete Denoising Solution for Old Malayalam Palm Leaf Manuscripts", Pattern Recognition and Image Analysis, vol. 32, no. 1, pp. 187-204, 2021.
- [7] H. T. Chandrakalaa, G. Thippeswamyb, Roshan Joy Martis, "Impact of Total Variation Regularization on Character Segmentation from Historical Stone Inscriptions", Pattern Recognition and Image Analysis, vol. 31, no. 1, pp. 35-48, 2021.
- [8] Jayati Mukherjee, Swapan K Parui, Uptal Roy, "NN-based analytic approach to symbol level recognition for degraded Bengali printed documents", Indian Academy of Sciences, sadhana, vol. 45, no. 263, pp. 1-22, 2020.
- [9] Jinhu Sun, Peng Li, Xiaojun Wu, "Handwritten Ancient Chinese Character Recognition Algorithm Based on Improved Inception-ResNet and Attention Mechanism", IEEE International Conference on Software Engineering and Artificial Intelligence, pp. 31-35, 2022.
- [10] K. Durga Devi, P. Uma Maheswari, "Digital acquisition and character extraction from stone inscription images using modified fuzzy entropy-based adaptive thresholding", Soft Computing, Springer, vol. 23, pp. 2611-2626, 2019.
- [11] K. Durga Devi, P. Uma Maheswari, Phani Kumar Polasi, R. Preetha, M. Vidhyalakshmi, "Pattern Matching Model for Recognition of Stone Inscription Characters", Computational Intelligence, Machine Learning, and Data Analytics, The Computer Journal, pp. 1-11, 2021.
- [12] Kaladevi Ra, Revathi Ab, Manju Ac, "Analyzing the Evolution of Modern Tamil Script for Natural Language Processing", ECS Transactions, vol. 107, no. 1, pp. 5219-5226, 2022.
- [13] Kavitha Subramani, Murugavalli Subramaniam, "Creation of original Tamil character dataset through segregation of ancient palm leaf manuscripts in medicine", Experts System, Wiley, vol. 38, pp. 1-13, 2020.
- [14] Liu Guoqing, Hao Changning, Yan Jingbo, Dong Jing, Zhao Zuolong, Hao Lujia, "Stroke Extraction Algorithm of Clerical Script in Han Dynasty Based on Contour: Take "Stele of Cao Quan" as an Example", Mobile Information systems, Hindawi, pp. 1-10, 2022.
- [15] M. Merline Magrina, "Convolution Neural Network based Ancient Tamil Character Recognition from Epigraphical Inscriptions", International Research Journal of Engineering and Technology, vol. 7, no. 4, 2020.
- [16] P Preethi, HR Mamatha, Hrishikesh Viswanath, "Study of using hybrid deep neural networks in character extraction from images containing text", Computer Science and Information Technology, pp. 45-52, 2021.
- [17] P Ravi, C Naveena, Y H Sharathkumar, "OCR for historical Kannada documents using clustering methods", Indian Journal of Science and Technology, pp. 3652-3663, 2020.
- [18] P. Dharani Devi, V. Sathiyapriya, "Brahmi Script Recognition System using Deep Learning Techniques", Third International Conference on Inventive Research in Computing Applications, IEEE Explore, pp. 1-4, 2021.
- [19] Padmaprabha Preethi, Hosahalli Ramappa Mamatha, "Region-Based Convolutional Neural Network for Segmenting Text in Epigraphical Images", Artificial Intelligence and Applications, pp. 1-9, 2022.
- [20] Pravin Savaridass M, Haritha J, Balamurugan V T, "CNN Based Character Recognition and Classification in Tamil Palm Leaf Manuscripts", International Conference on Communication, Computing, and Internet of Things, IEEE Xplore, pp. 1-6, 2022.
- [21] Preeti P, Anish Kasi, Manish Shetty, Mamatha H R, "Denoising and Segmentation of Epigraphical Estampages by Multi-Scale Template Matching and Connected Component Analysis", Procedia Computer Science, Elsevier, vol. 171, pp. 1877-1886, 2020.
- [22] R Prabavathi, J Shiny Duella, V Brindha Devi, "Prehistoric Stone Image Tamil Character Recognition using Optimized Deep Neural Network using Zernike Moments and Simplex Method", Turkish Journal of Computer and Mathematics Education, vol. 12, no. 11, pp. 5983-5591, 2021.
- [23] R. Jayakanthan, A. Hiran Kumar, N. Sankarram, B. S. Charulatha, Ashwin Ramesh, "Handwritten Tamil Character Recognition Using ResNet", International Journal of Research in Engineering, Science and Management, vol. 3, no. 3, pp. 133-137, 2020.
- [24] R. N. Ashlin Deepa, R. Rajeswara Rao, "A novel nearest interest point classifier for offline Tamil handwritten character recognition", Pattern Analysis and Applications, Springer, vol. 23, pp. 199-212, 2020.
- [25] S. Brindha, S. Bhuvaneshwari, "Repossession and recognition system: transliteration of antique Tamil Brahmi typescript", Current Science, vol. 120, no. 4, pp. 654-665, 2021.
- [26] S. Dhivya, J. Rene Beulah, "Ancient Tamil Character Recognition from Stone Inscriptions – A Theoretical Analysis", Asian Conference on Innovation in Technology, IEEE, pp. 1-8, 2022.

- [27] Sakkayaphop Pravesjit, Krittika Kantawong, Vitou That, “Segmentation of Broken Khmer Characters”, International Conference on Big Data Analytics and Practices, IEEE Explore, pp. 64-68, 2022.
- [28] Shashaank M, Aswatha, Ananth Nath Talla, Jayanta Mukhopadhyay, Partha Bhowmick, “A Method for Extracting Text from Stone Inscriptions using Character spotting”, pp. 1-14, 2015.
- [29] Shikha Magotra, Baijnath Kaushik, Ajay Kaul, “A Comparative analysis for identification and classification of text segmentation challenges in Takri Script”, Sadhana, Indian Academy of Sciences, vol. 45, no. 146, pp. 1-20, 2020.
- [30] Suganya Athisayamania, Dr. A. Robert Singhb, Dr. T. Athithan, “ Recognition of Ancient Tamil Palm Leaf Vowel Characters in Historical Documents using B-spline Curve Recognition”, Procedia Computer Science, Science Direct, vol. 171, pp. 2302-2309, 2020.
- [31] T. S. Suganya, S. Murugavalli, “A hybrid group search optimization: firefly algorithm-based big data framework for ancient script recognition”, Soft Computing, Methodologies and Applications, Springer, vol. 24, pp. 10933-10941, 2020.
- [32] T.Jerry Alexander, S.Suresh Kumar, B.Sowmya, “Performance Analysis of Fuzzy based Restoration Technique for Ink Bleed-through Degraded Documents”, Fourth International Conference on Electronics, Communication and Aerospace Technology, IEEE Xplore, pp.1429-1434, 2020.
- [33] V. Vani, S. R. Ananthalakshmi, “Soft computing approaches for character credential and word prophecy analysis with stone encryptions”, Soft Computing, Methodologies and Applications, Springer, vol. 24, pp. 12013-12026, 2020.
- [34] Vijayalakshmi. R, Dr. J M Gnanasekar, “A Review on Character Recognition and Information Retrieval from Ancient Inscriptions”, International Conference on Smart Structures and Systems, IEEE, pp. 1-7, 2022.
- [35] V T Chellam and Naanchil C Natarajan, “Linguistic History of Ancient India: A study on Archaeological Evidences”, Academia, pp. 1-26, 2020.
- [36] <https://dl.acm.org/doi/fullHtml/10.1145/3402891>.
- [37] <https://www.tnarch.gov.in/epigraphy/inscriptions>.
- [38] <https://tamilnadu-favtourism.blogspot.com/2016/01/jambai-malai-jambai-villupuram.html>.

## மகிழ்வூட்டும் கற்றலுக்கான மின்னிலக்கப் புதிர் அறை Using Digital Escape Rooms to Make Learning Fun

முனைவர் ராமன் விமலன் Dr. Raman Vimalan

Lecturer, Tamil Language & Culture Division

Asian Languages and Cultures AG I National Institute of Education (NIE)

Nanyang Technological University (NTU), 1 Nanyang Walk, Singapore 637616

### ABSTRACT

இருபத்தொன்றாம் நூற்றாண்டு மின்னிலக்க உலகாக உருமாற்றம் அடைந்துள்ளது. மருத்துவம், பொறியியல், விஞ்ஞானம் போன்ற துறைகளைப் போன்றே கல்வித்துறையிலும் தொழில்நுட்பத்தின் பயன்பாடு மிகுந்துள்ளது. கொவிட்-19 தொற்றுநோய்ப் பரவல் சூழலில் வகுப்பறைகள் மின்னியல் வகுப்பறைகளாக மாறின. கற்றல் கற்பித்தலில் புதுப்புது அணுகுமுறைகளும் தொழில்நுட்பங்களும் புகத்தொடங்கின. அதனடிப்படையில் வகுப்பறைக்குள் புகுந்த தொழில்நுட்பந்தான் மின்னிலக்கப் புதிர் அறை (Digital Escape Room).

#### Keywords:

A மகிழ்வூட்டும் கற்றல்  
B மின்னிலக்கப் புதிர் அறை  
C புதிர் அறை  
D மின்னிலக்கப் புதிர்  
E மின்வழிக் கற்றல்

“பசித்தவனுக்கு மீனை உண்ணக் கொடுப்பதைவிட, மீன் பிடிக்கக் கற்றுக்கொடுப்பதே சிறந்தது” என்னும் பழமொழியை மெய்ப்பிக்கும் வகையில் ‘மின்னிலக்கப் புதிர் அறை’ நடவடிக்கை கற்றல் கற்பித்தலில் பயன்படுகிறது. மாணவர்களை எப்படி? என்ற வினாவிற்கு விடைகாணத் தூண்டுவதோடு அவர்களே சுயமாக முன்னேறிச் சென்று சிக்கல்களுக்கான தீர்வினைக் காண ஊக்குவிக்கிறது. சுயகற்றல், அனுபவவழிக் கற்றல், கூடிக்கற்றல் போன்றவை நிகழ இம்மின்னிலக்கப் புதிர் அறை வழிகோலுகிறது. மேலும், ஈடுபாடுமிக்க மகிழ்வூட்டும் கற்றலுக்கும் அடித்தளம் அமைக்கிறது.

மின்னிலக்கப் புதிர் அறை மாணவர்களிடம் படைப்பாற்றல், பிரச்சனைக்குத் தீர்வு காணுதல், தொடர்பாற்றல் திறன் போன்றவற்றுக்கும் உதவி செய்கிறது. மாணவர்களை விளையாட்டுகளில் ஆர்வத்துடன் ஈடுபடச் செய்வதுடன் சிந்தித்துச் செயலாற்றவும் உதவி செய்கிறது. ஆசிரியர்கள் எந்தப் பாடத்தையும் இதன் மூலம் ஒருங்கிணைத்துக் கற்பிக்க இயலும். மாணவர்கள் எங்கு செல்ல வேண்டும்? எதைப் படிக்க வேண்டும்? என்பதை ஆசிரியர் வழிகாட்டாமலும் அறிவுறுத்தாமலும் புதிர் நிலையில் கண்டறிந்து விமர்சன நோக்கிலும் தர்க்க ரீதியாகவும் சிந்தித்துத் தீர்வுகான இது வழியமைக்கிறது. விளையாட்டு வழிக் கற்றல் இதன் சிறப்பம்சமாகும். மேலும் ஆசிரியர்கள் தங்கள் வகுப்புகளின் தரநிலைக்கேற்றவாறு இதனைச் சொந்தமாகவும் வடிவமைக்க முடியும்.

மாணவர்கள் ஒவ்வொரு புதிரையும் பூட்டைத் திறந்து விடைகாண்பர். இதனைத் தனிநிலையிலோ, குழுவிலையிலோ செயற்படுத்த இயலும். கூகுள் படவில்லைகள் (Google slides), கூகுள் படிவம் (Google form) போன்ற இலவச இணைய வளங்களைக்கொண்டு இதனைச் செயற்படுத்தலாம். மேலும், ஒலி ஒளிப்பகுதிகளை இணைத்தும் விடைகாணவும் மதிப்பீடு செய்யவும் ‘மின்னிலக்கப் புதிர் அறை’ நடவடிக்கை உதவுகிறது.

மாணவர்கள் புதிர் அறைக்குள் நுழைந்தவுடன் ஒரு புதிருக்கு விடை கண்ட பின்னரே பூட்டுத் திறக்கிறது. பின்னர், அடுத்த பூட்டைத் திறக்க முயற்சிக்கிறார்கள். சரியான விடையைக் கண்டறிந்த பின்னர், அப்பூட்டும் திறக்கிறது. இவ்வாறு அவர்கள் ஒவ்வொரு பூட்டையும் திறந்து இறுதியில் புதிர் அறையிலிருந்து தப்பித்து வெற்றி இலக்கை எட்டுகின்றனர். இந்நடவடிக்கைகள் யாவும் கேட்டல், பேசுதல், வாசித்தல், எழுதுதல் ஆகிய அடிப்படைத் திறன்களோடு ஒருங்கிணைத்துச் செயற்படுத்தப்படுகின்றன. இப்புதிர் அறை நடவடிக்கையில் மாணவர்களுக்கு உதவிக்குறிப்புகளும் ஆசிரியர்களின் வழிகாட்டுதலும் வழங்கப்படுவதால் மாணவர்கள் வெற்றி இலக்கை எளிதில் எட்டுகின்றனர். மொத்தத்தில் மின்னிலக்கப் புதிர் அறை (Digital Escape Room) மகிழ்வூட்டும் கற்றலுக்கு வலுகோலாகிறது என்பது உண்மை.

#### Corresponding Author:

முனைவர் ராமன் விமலன்,  
ஆசிரியமொழிகள் மற்றும் பண்பாடுகள்,  
தேசியக் கல்விக் கழகம்  
நன்யாங் தொழில்நுட்பப் பல்கலைக் கழகம், சிங்கப்பூர்.  
Email: raman.vimalan@nie.edu.sg

## 1. முன்னுரை

இருபத்தொன்றாம் நூற்றாண்டு மின்னிலக்க உலகாக உருமாற்றம் அடைந்துள்ளது. மருத்துவம், பொறியியல், விஞ்ஞானம் போன்ற துறைகளைப் போன்றே கல்வித்துறையிலும் தொழில்நுட்பத்தின் பயன்பாடு மிகுந்துள்ளது. உலகைப் புரட்டிப்போட்ட கொவிட்-19 தொற்றுநோய் பரவல் சூழலில் வகுப்பறைகள் மின்னியல் வகுப்பறைகளாக மாற்றம் கண்டன. கற்றல் கற்பித்தலில் புதுப்புது அணுகுமுறைகளும் தொழில்நுட்பங்களும் புகத்தொடங்கின. ஜாம், கூகுள் மீட் போன்ற மெய்நிகர் தளங்கள் வகுப்பறைக் கற்றல் தடைபெறா வண்ணம் கல்வியைச் சாத்தியமாக்கின. மாணவர் கற்றல் தளங்களும் மின்னியல் வழியிலான கற்றல் வளங்களும் மாணவர்களைக் கற்றலில் ஈடுபடுத்தின. ‘மின்னிலக்கப் புதிர் அறை’ (Digital Escape Room) தொழில்நுட்பமும் வகுப்பறைக்குள் புகுந்து மாணவர்களை வெகுவாகக் கவர்ந்தது. இம்மின்னிலக்கப் புதிர் அறைத் தொழில்நுட்பத்தைப் பற்றியும் அது மகிழ்வூட்டும் கற்றலுக்கு வித்திடும் வழிவகைகளைப் பற்றியும் இக்கட்டுரை விளக்குகிறது.

## 2. மின்னிலக்கப் புதிர் அறை

மின்னிலக்கப் புதிர் அறை என்பது தொடக்கத்தில், ஒரு பொழுதுபோக்கு வடிவமாகத் தொடங்கப்பட்டது. இது, “ஒரு விளையாட்டை விளையாடுவதற்காகப் பூட்டப்பட்டிருக்கும் ஒரு மின்அறை. இவ்வறையில் ஓர் இலக்கை அடைவதற்குக் குறிப்பிட்ட நேரத்திற்குள் தொடர்ச்சியான புதிர்களைத் தீர்க்க வேண்டும். அதாவது, அறையைத் திறப்பதற்கான சாவியைக் கண்டறிய வேண்டும். மேலும், புதிர் அறையிலிருந்து வெறியேறும் வழியைக் கண்டுபிடிக்க நேரத்தையும் புதிர்களுக்கான துப்புகளையும் சரிவரப் புரிந்து கொள்ள வேண்டும்” (Oxford Languages and Google). இதனடிப்படையில், விளையாட்டு, வேடிக்கை போன்ற ஆர்வமூட்டும் அம்சங்கள் நிறைந்ததாக மின்னிலக்கப் புதிர் அறை அமையும். விளையாட்டை மையமிட்டு உருவாக்கப்பட்ட பொழுதுபோக்கு அம்சம் நிறைந்த புதிர் அறை விளையாட்டுகள் பின்னர், குழுப்பண்பை வளர்க்கும் நிலையில் பயிற்சிப் பட்டறைகளாக நடத்தும் நிலைக்கு வளர்முகம் கண்டன. இன்றைய கல்விச்சூழலில் வளமூட்டும் நடவடிக்கைகளாகவும் கற்றல் கற்பித்தலுக்கு உதவும் நடவடிக்கையாகவும் இம்மின்னிலக்கப் புதிர் அறைத் தொழில்நுட்பம் பயன்படுத்தப்பட்டு வருகின்றது.

“பசித்தவனுக்கு மீனை உண்ணக் கொடுப்பதைவிட, மீன் பிடிக்கக் கற்றுக்கொடுப்பதே சிறந்தது” என்னும் பழமொழியை மெய்ப்பிக்கும் வகையில் ‘மின்னிலக்கப் புதிர் அறை’ நடவடிக்கை கற்றல் கற்பித்தலில் இடம்பெறுகிறது. மின்னிலக்கப் புதிர் அறைகளை “இணையவழித் தப்பிக்கும் அறைகள்” அல்லது “மெய்நிகர் தப்பிக்கும் அறைகள்” என்றும் அழைக்கின்றனர். இப்புதிர் அறைகள் ‘புதையல் வேட்டை’, ‘புதிர்ங்கம்’ போன்று விளையாட்டு அடிப்படையில் அமைவன. இவை மாணவர்களுக்குப் பாரம்பரிய வகுப்பறைக் கற்பித்தல் முறைகளைக் காட்டிலும் சற்று மாறுபட்டு அமைகின்றன. (Lesley Speller, 2022) ஆசிரியர் கற்பிக்கும் அல்லது மாணவர் கற்கும் பாடப்பொருளை ஆர்வமும் ஈடுபாடும் மிக்க வகையில் மின்னியல் வழியில் வழங்குகின்றன.

மாணவர்களுக்கு ஒரு புதிய பாடத்தைக் கற்பிப்பதற்கோ அல்லது கற்பித்த பாடப்பொருளை மதிப்பீடு செய்வதற்கோ உதவும் ஒரு சிறந்த வழியாக இவை கல்வியாளர்களால் ஏற்றுக்கொள்ளப்பட்டுள்ளன. மெய்நிகர் வழியில் செயற்படுத்தப்படும் இம்மின்னிலக்கப் புதிர் அறை ஜாம் மற்றும் பிற கற்றல் தளங்களின் வழியாக நடத்தப்படும் ஒரு கற்றல் செயல்பாடாகவும் அமைகிறது.

## 3. நோக்கமும் செயல்பாடும்

மின்னிலக்கப் புதிர் அறை நடவடிக்கையின் முதன்மை நோக்கம் ‘புதிர் அறையிலிருந்து தப்பித்தல்’ என்ற முதன்மைக் குறிக்கோளைக் கொண்டுள்ளது. (Aaron Hallaway, 2015) விளையாட்டாளர்கள் தங்களுக்கு நிர்ணயிக்கப்பட்ட கால எல்லைக்குள் புதிர் அறைக்குள் இடம்பெற்றுள்ள புதிர்களையும் சவால்களையும் எதிர்கொண்டு, சிந்தித்துச் செயலாற்றி வெற்றி இலக்கை எட்டுவதை இதில் இடம்பெறும் நடவடிக்கைகள் தீர்மானிக்கின்றன. மேலும் சுயகற்றல், உடனிலைந்து கற்றல் போன்ற கற்றல் செயல்பாடுகளை நிகழ்த்துவதற்கும் வாய்ப்பளிக்கிறது. விளையாட்டுவழிக் கற்றல் அணுகுமுறையின் வாயிலாகக் கற்றல் இலக்கை எய்துவதையும் இம்மின்னிலக்கப் புதிர் அறை நோக்கமாகக் கொண்டுள்ளது.

## 4. கற்றல் கற்பித்தல் வளம்

மின்னிலக்கப் புதிர் அறை கற்றல் கற்பித்தலுக்கு உதவும் மெய்நிகர் கற்றல் வளமாகும். மாணவர்களை எப்படி? என்ற வினாவிற்கு விடைகாணத் தூண்டுவதோடு அவர்களே சுயமாக முன்னேறிச் சென்று சிக்கல்களுக்கான தீர்வினைக் காணவும் ஊக்குவிக்கிறது. மாணவர்களை வட்டத்திற்கு வெளியே (Thinking outside the box) சிந்திக்கத் தூண்டி சுயகற்றல், அனுபவவழிக் கற்றல், கூடிக்கற்றல் போன்ற கற்றல் செயல்பாடுகள்

நிகழவும் இம்மின்னிலக்கப் புதிர் அறை வழிகோலுகிறது. மேலும், ஈடுபாடுமிக்க கற்றலுக்கும் அனுபவவழிக் கற்றலுக்கும் வழியமைத்து மகிழ்வூட்டும் கற்றலை உறுதிப்படுத்திக் கற்றலை இனிதாக்குகிறது.

மின்னிலக்கப் புதிர் அறையில் இடம்பெறும் நடவடிக்கைகள், மாணவர்களிடம் படைப்பாற்றல், பிரச்சனைக்குத் தீர்வு காணுதல், தொடர்பாற்றல் திறன் போன்ற திறன் வளர்ச்சிகளுக்கு உதவிபுரிகின்றன. மாணவர்களை விளையாட்டுகளில் ஆர்வத்துடன் ஈடுபடச் செய்வதுடன் சிந்தித்துச் செயலாற்றவும் தூண்டுகிறது. பாலர் பள்ளி மாணவர்கள் முதல் உயர்வகுப்புகளில் பயிலும் மாணவர்கள் வரை இப்புதிர் அறை நடவடிக்கைகளில் ஈடுபட்டுக் கற்றல் இலக்கை அடைகிறார்கள். மகிழ்வூட்டும் நிலையில் அமையும் புதிர் நடவடிக்கைகள் மூலம் ஈடுபாடுமிக்க கற்றலை மாணவர்களிடம் வளர்க்க முடிகிறது. இந்நடவடிக்கை விளையாட்டு நிலையில் அமைவதால் துடிப்புடன் கற்கும் சூழல் உருவாகிறது. ஆசிரியர்கள் மாணவர்களுக்கு வழிகாட்டாமல், அவர்களுக்கு ஊக்கம் கொடுத்து முயற்சி செய்யுமாறு கூறும் வழிகாட்டுநர்களாகச் செயல்படுவதால் முயன்று தவறிக் கற்கும் அணுகுமுறையில் மாணவர்களின் கற்றல் மேம்படுகிறது.

ஆசிரியர்கள், மாணவர்களுக்குக் கற்பிக்க விரும்பும் எந்தப் பாடத்தையும் இம்மின்னிலக்கப் புதிர் அறை மூலம் ஒருங்கிணைத்துக் கற்பிக்க முடிகிறது. தாய்மொழிப் பாடத்தில் இடம்பெறும் கட்டுரை, கருத்தறிதல், இலக்கணம் சார்ந்த மொழிப்பயிற்சிகள் போன்ற பாடப்பகுதிகளை இப்புதிர் அறைகள் மூலம் சலிப்பின்றிக் கற்க இயலுகிறது. மேலும், மாணவர்கள் எங்குச் செல்ல வேண்டும்? எதைப் படிக்க வேண்டும்? என்பதை ஆசிரியர் வழிகாட்டாமலும் அறிவுறுத்தாமலும் சுயகற்றலில் ஈடுபட்டு, விமர்சன நோக்கிலும் தர்க்க ரீதியாகவும் சிந்தித்துத் தீர்வுகாண வழியமைக்கிறது. விளையாட்டு வழிக் கற்றலைச் சிறப்பம்சமாகக்கொண்ட இந்நடவடிக்கையை ஆசிரியர்கள் தங்கள் வகுப்புகளின் தரநிலைக்கேற்றவாறு சொந்தமாக வடிவமைத்துக் கற்றல் கற்பித்தலை ஆழமுடையதாகவும் விருப்பமுடையதாகவும் ஆக்க உதவுகிறது.

## 5. புதிர் அறை உருவாக்கம்

ஆசிரியர்கள், மின்னிலக்கப் புதிர் அறையை உருவாக்குவதற்கு ஒருசில படிநிலைகளைப் பின்பற்ற வேண்டும். அவ்வகையில் ஒருசில முன்னேற்பாடுகளை ஆசிரியர்கள் மேற்கொள்ள வேண்டும். வகுப்பின் தரநிலையை அறிந்து அதற்கேற்ற வண்ணம் மின்னிலக்கப் புதிர் அறை நடவடிக்கைகளை உருவாக்கத் திட்டமிடுதல் அவசியம். பின்னர், கற்பிக்கப் புகும் பாடத்தையும் அதனையொட்டிய கருப்பொருளையும் (Theme) தெரிவுசெய்துகொள்ள வேண்டும். பின், புதிர் அறைக்கான தலைப்பையும் அதன் உள்ளடக்கத்தையும் தீர்மானம் செய்துகொள்ள வேண்டும். அதன் பிறகு, மாணவர்களின் ஆர்வத்தைத் தூண்டும் நடவடிக்கைகளையும் அவற்றைத் தயாரிக்க உதவும் புதிர் அறை அமைப்பையும் திட்டமிட்டு உருவாக்கிக்கொள்ள வேண்டும். (Lesley Speller, 2022).

மின்னிலக்கப் புதிர் அறையை அமைக்கும்போது, மாணவர்கள் விரும்பக்கூடியதாகவும் அவர்களை ஈர்க்கக்கூடியதாகவும் இருப்பது அவசியம். பின்னர், மாணவர்கள் புதிர் அறையில் இடம்பெறவுள்ள சவால்கள் அல்லது புதிர்களுக்கான தடயங்களைக் கண்டறியும் வழிகளைத் தீர்மானித்துக்கொண்டு, புதிர்கள் மற்றும் விளையாட்டுகளை உள்ளடக்கிய புதிர் அறையை வடிவமைக்க வேண்டும். இவ்வாறு வடிவமைக்கும்போது மாணவர்களின் அனுபவ நிலைக்கு ஏற்றவாறு வடிவமைப்பது அவசியம். மேலும், ஒவ்வொரு சவாலும், மாணவரின் கொள்திறன்களைச் சோதிக்கும் வகையில் அமைய வேண்டும். எடுத்துக்காட்டாக, பாலர் பள்ளி மாணவர்களுக்கு, படங்களைப் புதிர் அறைகளில் தரலாம். படங்களையோ பொருட்களையோ வரிசைப்படுத்துதல் போன்ற எளிய பணிகளை வழங்கலாம். வாசிப்பில் பின்தங்கிய நிலையில் இருக்கும் மாணவர்களுக்கு, எழுத்துவழிப் பேச்சுணரித் தொழில்நுட்பத்தைப் பயன்படுத்திப் புதிர்களை வழங்கலாம்.

ஆரம்ப நிலையில் ஒரு புதிர் அறையில் தொடங்கிப் படிப்படியாக இரண்டு முதல் ஐந்து வரையிலான புதிர் அறைகளை உருவாக்கிக் கற்றல் கற்பித்தலை வழிநடத்த வேண்டும். அதே வேளையில் சவாலையும் கடினப்படுத்துவது அவசியம். குழுநிலையில் தொடங்கிய பணியைத் தனிநிலையில் செய்யும் வகையில் மாணவர்களைத் தயார்படுத்துவது கற்றலை ஆழப்படுத்துவதோடு சுயகற்றலுக்கும் வழிவகுக்கும். மாணவர்களின் கவனத்தை ஈர்க்கும் வகையிலமைந்த படங்கள், காணொளிகள், புதிர்கள் போன்றவற்றை இப்புதிர் அறைக்குள் உட்புகுத்தி பாடப்பொருளை ஆர்வமூட்டும் வகையில் அமைத்துக்கொள்வது மாணவர்களின் ஈடுபாட்டை மிகுவிக்கும். இவற்றைக் கூகுள் ஒளிப்படவிலைகளிலோ (Google Slides) அல்லது கூகுள் படிவத்தின் (Google Forms) துணைகொண்டோ உருவாக்க வேண்டும். நிறைவாக, மாணவர்கள் ஒவ்வொரு புதிர் அறையைப் பயன்படுத்துவதற்கான காலஅளவை நிர்ணயம் செய்து புதிர் அறையின் செயல்பாட்டை உறுதி செய்ய வேண்டும். ஒரு புதிருக்கு ஐந்து நிமிடங்கள் முதல் பதினைந்து நிமிடங்கள் வரை கால நிர்ணயம் செய்து கொள்வது அவசியம்.

## 6. செயல்பாடும் வழிமுறைகளும்

ஆசிரியர்கள், மாணவர்களின் கவனத்தை ஈர்க்கும் புதிர் அறைகளை உருவாக்கி அதனைச் செயல்பாட்டு நிலைக்குக் கொண்டுவந்த பின்னர், இம்மின்னிலக்கப் புதிர் அறை நடவடிக்கையைக் கூகுள் வகுப்பறையிலோ (Google Classroom) மாணவர் கற்றல் தளத்திலோ (Student Learning Space) பதிவேற்றம் செய்ய வேண்டும். அதன்பிறகு, மாணவர்களுக்கு பணி அனுமதியை வழங்கிப் புதிர் அறை விளையாட்டில் ஈடுபடுத்த வேண்டும். இதனை விரைவுத் தகவல் குறியீட்டைப் (QR Code) பயன்படுத்தியும் மாணவர்களுடன் பகிர்ந்துகொள்ளலாம்.

மின்னிலக்கப் புதிர் அறைக்குள் உள்நுழையும் மாணவர்கள், அவ்வறையில் இடம்பெற்றுள்ள ஒவ்வொரு புதிருக்கான பூட்டையும் திறந்து விடைகாண முயல்வர். அவர்கள் புதிர் அறைக்குள் நுழைந்தவுடன் ஒரு புதிரைத் தீர்க்க முற்படுவர். அவர்களின் விடை சரியாக இருந்தால், பூட்டுத் திறக்கிறது. அவர்கள் அப்புதிர் அறையிலிருந்து தப்பிக்கிறார்கள். இல்லையெனில், அவர்கள் மீண்டும் புதிருக்குச் சென்று அங்கு இடம்பெற்றுள்ள உதவிக்குறிப்புகளைப் பயன்படுத்தி வெற்றி இலக்கை எய்துகின்றனர். பின்னர், அடுத்த பூட்டைத் திறக்கும் முயற்சியில் இறங்குவர். அங்கு இடம்பெற்றுள்ள புதிருக்கான சரியான விடையைக் கண்டறிந்த பின்னர், அப்பூட்டும் திறக்கிறது. இவ்வாறு அவர்கள் ஒவ்வொரு பூட்டையும் திறந்து இறுதியில் புதிர் அறையிலிருந்து தப்பித்து வெற்றி இலக்கை அடைவர். இந்நடவடிக்கையைத் தனிநிலையிலோ, குழுநிலையிலோ செயற்படுத்த இயலும். கூகுள் படவில்லைகள் (Google slides), கூகுள் படிவம் (Google form) போன்ற இலவச இணைய வளங்களைக்கொண்டு இதனைச் செயற்படுத்த முடிகிறது. மேலும், ஒலி ஒளிப்பகுதிகளை இணைத்தும் மாறுபட்ட தளங்களில் கற்கும் வாய்ப்பினை வழங்க முடிகிறது. மாணவர்கள் சிந்தித்துச் செயலாற்றவும் சிக்கலுக்குத் தீர்வுகாணவும் கற்பித்தலை ஆசிரியர்கள் மதிப்பீடு செய்யவும் இம்மின்னிலக்கப் புதிர் அறை நடவடிக்கை பெருந்துணை புரிகிறது.

இந்நடவடிக்கைகள் யாவும் கேட்டல், பேசுதல், வாசித்தல், எழுதுதல் ஆகிய அடிப்படைத் திறன்களோடு ஒருங்கிணைத்துச் செயற்படுத்தப்படுகின்றன. இப்புதிர் அறை நடவடிக்கையில் மாணவர்களுக்கு உதவிக்குறிப்புகளும் ஆசிரியர்களின் வழிகாட்டுதலும் வழங்கப்படுவதால் அவற்றின் துணைகொண்டு மாணவர்கள் வெற்றி இலக்கை எளிதில் எட்டுகின்றனர். மொத்தத்தில் மின்னிலக்கப் புதிர் அறை (Digital Escape Room) மகிழ்வூட்டும் கற்றலுக்கு வழிகோலுகிறது என்பது உண்மை.

## 7. மீள்நோக்குதல்

மின்னிலக்கப் புதிர் அறை நடவடிக்கைக்கான நேரம் முடிந்ததும், நடவடிக்கை பற்றிய கலந்துரையாடலை அமைத்துக்கொள்வது அவசியம். மாணவர்கள் புதிர் அறைகளைத் திறந்து வெளியேறி வெற்றி பெற்றாலும், அடுத்தமுறை மின்னிலக்கப் புதிர் அறை நடவடிக்கையை மேற்கொள்ளும்போது மீண்டும் மாணவர்களிடம் அதன் செயல்முறையை விளக்குவது இன்றியமையாதது. (Christine Danhoff, 2022) அதுமட்டுமல்லாமல், ஆசிரியர் எதை மேம்படுத்த வேண்டும்? எப்பகுதியில் கடினத்தன்மை இருந்தது? எது எளிமையாக இருந்தது? என்பதையும் மீள்நோக்கம் செய்து அடுத்த முறை மின்னிலக்கப் புதிர் அறை உருவாக்கத்தின்போது, புதிர் விளையாட்டை மேலும் ஆர்வமூட்டுவதாகவும் சிறப்புவாய்வுததாகவும் ஆக்கிக்கொள்ள வேண்டும்.

## 8. விளைவும் பயனும்

மின்னிலக்கப் புதிர் அறை நடவடிக்கையில் ஈடுபடும் மாணவர்கள் மெய்நிகர் வழியில் விளையாட்டுகளில் மூழ்கி ஆர்வத்துடன் நடவடிக்கைகளில் பங்கெடுக்கின்றனர். பையப்பயிலும் மாணவர்களும் இதில் ஆர்வத்துடன் பங்கேற்று வெற்றி இலக்கை அடைய முடிகிறது. மேலும், மாணவர்களிடம் குழுவுணர்வு, ஒத்துழைப்பு, சிக்கல்களுக்குத் தீர்வு, படைப்பாக்கம், புத்தாக்கம் போன்ற பல்வேறு திறன்களும் வளரும் வாய்ப்புள்ளது. தொழில்நுட்பப் பயனாளர்களாக வலம் வரும் இக்கால மாணவர்கள், கற்றலில் தகவல் தொழில்நுட்பத்தைப் பயன்படுத்தித் தங்கள் தகவல் தொழில்நுட்ப அறிவை மேம்படுத்திக் கொள்ளவும் முடிகிறது.

மின்னிலக்கப் புதிர் அறையை உருவாக்குவது மாணவர்களின் ஈடுபாடுமிக்க கற்றலை ஊக்குவிக்கும் ஒரு சிறந்த வழியாகும். புதிய தலைப்பை அறிமுகப்படுத்துவதற்கோ அல்லது மாணவர்கள் ஏற்கனவே கற்றுக்கொண்ட பாடப்பொருளை மதிப்பீடு செய்வதற்கோ இந்த மின்னிலக்கப் புதிர் அறை பயன்படுகிறது. முதலில் இதனை வடிவமைக்கச் சற்று நேரமெடுக்கலாம். ஆனால், நாம் ஒரு புதிர் அறையை வடிவமைத்துச் செயற்படுத்தியவுடன் அதே நடைமுறைகளையும் உபகரணங்களையும் மற்ற புதிர் அறை உருவாக்கத்திற்கும் பயன்படுத்திக்கொள்ள முடிகிறது. மேலும், புதிய தலைப்புக்கு ஏற்ற வண்ணம் புதிர்களையும் சவால்களையும்



அமைப்புமுறையையும் மாற்றியமைத்து மாணவர்களுக்குப் புத்தாக்கத்துடன் கற்றல் கற்பித்தலை நிகழ்த்த முடிகிறது.

இம்மின்னிலக்கப் புதிர் அறை விளையாட்டில் ஈடுபட ஒரு தனி நபர் பல கோணங்களில் சிந்திக்க வேண்டும். அவர் மற்றவர்களுடன் கலந்துரையாடி விடைகளைக் கண்டுபிடிக்க முயல்வதால் அவருடைய சமூகத் தொடர்புத் திறன்கள் மேம்படுகின்றன. புதிர் அறைகளில் இடம்பெறும் சவால்களைச் சமாளிக்க முயலும்போது உணர்ச்சிகள் மேம்பட்டு ஆர்வத்துடன் நடவடிக்கையில் ஈடுபடும் வாய்ப்புக்கிட்டுகிறது. படிப்படியாகவும் கதையோட்டத்துடனும் நடவடிக்கைகள் இடம்பெறுவதால், விளையாடும் நபரின் நினைவாற்றல் திறன் மேம்படுகிறது. கேட்டல், பேசுதல், வாசித்தல், எழுதுதல் என்னும் அடிப்படைத் திறன்களை மையமிட்டு விளையாடப்படுவதால் மாணவர்கள் முழுமையான கற்றலில் ஆர்வத்துடன் ஈடுபடுகின்றனர். மேலும், மின்னிலக்கப் புதிர் விளையாட்டை நிறைவு செய்ய வேண்டும் என்னும் குறிக்கோள் இருப்பதால், மாணவர்களின் ஆர்வநிலையும் முனைப்பும் கூடுகிறது.

மெய்நிகர் வழியில் செயல்படுத்தப்படும் இம்மின்னிலக்கப் புதிர் அறை நடவடிக்கையில் நன்மைகள் பல இருப்பினும் ஒருசில தீமைகளும் இருக்கவே செய்கின்றன. மெய்நிகர் வழியில் செயற்படுத்தும்போது மாணவர்களின் விருப்ப நிலைக்கேற்ப வகுப்பறைக்கு வெளியில் இருந்தும் செய்யும் வாய்ப்பிருப்பதால் ஆசிரியரின் நேரடிப்பார்வை இல்லாமல் போவதும் மாணவர்களின் கவனக்குறைவிற்கும் இடமளிக்கிறது. புதிர் அறையில் இடம்பெறும் பாடவளங்களையும் புதிர்களையும் ஒருசில மாணவர்கள் புரிந்துகொள்வதில் இடர்ப்பாடு நேரிடுகிறது. இது மாணவர்களிடம் சலிப்பை ஏற்படுத்துகிறது. எனவே, மாணவர்கள் அனைவரும் ஈடுபாட்டுடன் கற்பதற்கு உதவும் பாடவளத்தை உருவாக்குவதில் ஆசிரியர்களின் முயற்சி அதிகம் இருக்க வேண்டும்.

## 9. முடிவுரை

‘மின்னிலக்கப் புதிர் அறை’ (Digital Escape Room) தொழில்நுட்பம் கற்றல் கற்பித்தலுக்கு உதவும் மெய்நிகர் வழியிலான கற்றல் வளமாகும். பாரம்பரிய வகுப்பறைக் கற்பித்தல் முறைகளிலிருந்து சற்று மாறுபட்டு, பாடப்பொருளை ஆர்வமும் வேடிக்கையும் விளையாட்டும் நிறைந்ததாகவும் மாற்றுகிறது. இதனால் மாணவர்களின் ஈடுபாடு மிகுந்து கற்றலை ஆழப்படுத்துகிறது. மாணவர்களுக்கு ஒரு புதிய பாடத்தைக் கற்பிப்பதற்கும் கற்பித்த பாடப்பொருளை மதிப்பீடு செய்வதற்கும் இது ஒரு சிறந்த வழியாக அமைகிறது. இலவச இணைய வளங்களைக்கொண்டு உருவாக்கப்படும் இம்மின்னிலக்கப் புதிர் அறை கூகுள் வகுப்பறை, ஜூம், மாணவர் கற்றல்தளம் போன்ற பிற கற்றல் தளங்களின் வழியாகவும் நடத்தப்படும் ஒரு கற்றல் செயல்பாடாக அமைகிறது. சுயகற்றல், உடனிலைந்து கற்றல், விளையாட்டுவழிக் கற்றல், அனுபவவழிக் கற்றல், முயன்றுதவறிக் கற்றல் போன்ற நிலைகளில் இதன் செயல்பாடு அமைகிறது. மேலும், மாணவர்களிடம் படைப்பாற்றல், பிரச்சனைக்குத் தீர்வு காணுதல், தொடர்பாற்றல் திறன் போன்ற திறன் வளர்ச்சிகளுக்கும் உதவிபுரிகின்றது. கேட்டல், பேசுதல், வாசித்தல், எழுதுதல் ஆகிய அடிப்படைத் திறன்களை ஒருங்கிணைத்துச் செயற்படுத்தப்படுவதால் மாணவர்களின் மொழித்திறன் மேம்பாட்டிற்கும் வழியமைக்கிறது. மாணவர்கள் விளையாட்டு அடிப்படையில் சவால்களை எதிர்கொண்டு வெற்றி இலக்கை எளிதில் எட்டுவதால் மின்னிலக்கப் புதிர் அறை மகிழ்வுட்டும் கற்றலுக்கு வழிகோலுகிறது என்பது உண்மையினும் உண்மை.

## REFERENCES

1. Christine Danhoff, (2022). *Using Digital Escape Rooms to Make Learning Fun*, <https://www.edutopia.org>
2. Creative Classroom, Education Resources, (October 5, 2022). *WeAreTeachers Staff*,
3. Lesley Speller, (2022) *Learning with Escape Rooms*, University of Arkansas
4. Aaron Hallaway, (2015), “What is a ‘Room Escape’?”. <https://geekandsundry.com/what-is-a-room-escape/>

## கம்பயில்லா உணரி வலைதளத்தில் பாதுகாப்பான மற்றும் திறமையான தரவு பரிமாற்றத்திற்கான மூன்று நிலை பாதுகாப்புகள்

இரா. நேசமலர்<sup>1</sup>, முனைவர் கா. இரவிக்குமார்<sup>2</sup>  
1 முனைவரப்பட்ட ஆய்வாளர், கணிப்பொறி அறிவியல் துறை, தமிழ்ப் பல்கலைக்கழகம், இந்தியா  
மின்னஞ்சல்: ranesamalar@gmail.com  
2 ஆய்வுநெறியாளர், கணிப்பொறி அறிவியல் துறை, தமிழ்ப் பல்கலைக்கழகம், இந்தியா  
மின்னஞ்சல்: ravikasi2008@gmail.com

### ABSTRACT

#### Keywords:

- A குறியாக்கவியல்
- B, குறியாக்கம்
- C, மறைகுறியாக்கம்
- D, கையொப்பம்
- E அணுகல் கட்டுப்பாடு.

கம்பயில்லா உணரி வலைதளங்கள் (WSNs) வரையறுக்கப்பட்ட ஆற்றலுடன் சிறிய உணரி முனைகளைக் கொண்டிருக்கும். இத்தகைய முனைகள் கம்பி ஊடகத்தின் தேவையின்றி இயற்கை சூழல்களை தெரிவிக்கும் திறனைக் கொண்டுள்ளன. கம்பயில்லா உணரி வலைதளங்கள் தன்னாட்சி மற்றும் திறந்த வெளியில் உபயோகப்படுத்தப்படுகின்றன. கம்பயில்லா உணரி வலைதளத்தில், உணரி முனையின் முக்கியப் பணியானது தரவை உணர்ந்து அதை பல்வேறு வகையிலான சூழலில் அடிப்படை நிலையத்திற்கு அனுப்புவதாகும். சேகரிக்கப்பட்ட தரவின் சரியான தன்மைக்கு உத்தரவாதம் அளிக்க அடிப்படை நிலையத்திற்கும் மற்றும் உணரிகளுக்கு இடையே பாதுகாப்பான ஊடகத்தை அமைப்பது முக்கியம். சேகரிக்கப்பட்ட தரவு சிதைந்தால், தரவு பகுப்பாய்வின் முடிவுகள் நம்பமுடியாத மற்றும் இன்னும் பேரழிவு சேதத்திற்கு வழிவகுக்கும். இந்தச் சிக்கலைச் சமாளிக்க, கம்பயில்லா உணரி வலைதளத்தில் பாதுகாப்பான மற்றும் திறமையான தரவு பரிமாற்றத்திற்கான ஹாஷ் செய்தி அங்கீகார குறியீட்டு கையொப்பம், டோக்கன் அடிப்படையிலான அணுகல் கட்டுப்பாட்டுடன் கூடிய மறைக்குறியீட்டு மாற்றுதல் மற்றும் கலப்பு குறியாக்கவியல் பயன்படுத்தி மூன்று அடுக்கு பாதுகாப்பை இந்த ஆராய்ச்சிப் பணி முன்மொழிகிறது. முதலில், உணரி முனைகள் மூலம் உணரப்படும் தரவை மறைக்குறியீட்டு மாற்றுதல் நுட்பத்துடன் கூடிய சமச்சீர்-விசை குறியாக்கவியல் பயன்படுத்தப்படுகிறது. இரண்டாவதாக டோக்கன் அடிப்படையிலான ஹாஷ் செய்தி அங்கீகார குறியீட்டு கையொப்பம் மாற்றப்பட்ட மறைக்குறியீட்டிற்காக உருவாக்கப்படுகிறது. மூன்றாவதாக, ஒவ்வொரு உணரி முனைகளிலிருந்து அடிப்படை நிலையத்தால் பெறப்படும் மாற்றப்பட்ட மறைக்குறியீட்டு உரையை மீண்டும் ஒரு முறை குறியாக்க மறைக்குறியீட்டு மாற்றுதல் நுட்பத்துடன் கூடிய பொது-விசை குறியாக்கவியல் பயன்படுத்தப்படுகிறது. பின்னர் அடிப்படை நிலையம் இந்த மாற்றப்பட்ட மறைக்குறியீட்டு உரையை நிர்வாகிக்கு அனுப்புகிறது. கூடுதலாக, டோக்கன் அடிப்படையிலான அணுகல் கட்டுப்பாட்டு நுட்பம், அனுப்புநர் உணரி முனையின் டோக்கனைக் கொண்ட நிர்வாகியை மட்டுமே பெறப்பட்ட மறைக்குறியீட்டு உரையை அணுக அனுமதிக்க பயன்படுகிறது. மேலும், பெறப்பட்ட மறைக்குறியீட்டு தகவல் பாதுகாப்பானதா இல்லையா என்பதை

சரிபார்க்க ஹாஷ் செய்தி அங்கீகார குறியீட்டு கையொப்பம் பயன்படுத்தப்படுகிறது. முன்மொழியப்பட்ட குறியாக்கவியல் நுட்பம் ஒரே நேரத்தில் ரகசியத்தன்மை, அங்கீகாரம், அணுகல் கட்டுப்பாடு மற்றும் நன்பகத்தன்மை ஆகியவற்றை சரிபார்கின்றது. முன்மொழியப்பட்ட பாதுகாப்பு நுட்பமானது குறியாக்கம் மற்றும் மறைகுறியாக்கத்திற்கு குறைவான கணக்கீட்டு நேரத்தை எடுத்துக்கொள்கிறது மற்றும் ஏற்கனவே உள்ள மற்ற குறியாக்கம் நுட்பங்களுடன் ஒப்பிடும்போது குறைந்த ஆற்றலைப் பயன்படுத்துகிறது என்பதை சோதனை முடிவு காட்டுகிறது.

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

### Corresponding Author:

இரா.நேசமலர்

முனைவர்ப்பட்ட ஆய்வாளர், கணிப்பொறி அறிவியல் துறை, தமிழ்ப் பல்கலைக்கழகம், இந்தியா

மின்னஞ்சல்: ranesamalar@gmail.com

### 1. அறிமுகம்

கம்பயில்லா உணர் வலைப்பின்னல் (WSNs) என்பது சுய-கட்டமைக்கப்பட்ட, உள்கட்டமைப்பு இல்லாத கம்பயில்லா வலைப்பின்னல் ஆகும், இது சுற்றுச்சூழல் நிலைமைகளைக் கண்காணித்து பதிவு செய்கிறது மற்றும் தரவுகளை ஒரு அடிப்படை நிலையத்தில் சேமிக்கிறது. இதன் குறைந்த விலை, கச்சிதமான அளவு மற்றும் மருத்துவம், ஆயுதப் படைகள் மற்றும் அதிக கண்காணிப்பு போன்ற பல்வேறு களங்களில் பொருத்தக்கூடிய தன்மை காரணமாக, பல்வேறு பயன்பாடுகளுக்கு அதிகமாக பன்படுத்தப்படுகின்றது[1]. கம்பயில்லா உணர் வலைப்பின்னல்களில் உள்ள அபாயகரமான அல்லது அணுக முடியாத பகுதிகளில் உணரிகள் பல்வேறு இடங்களில் பொருத்தப்பாடுகின்றன[2]. உணர் முனைகள், நுழைவாயில் முனைகள், அடிப்படை நிலையங்கள் மற்றும் பயனர்கள் அனைத்தும் இந்த வலைப்பின்னல் பகுதிகளாகும்.

உணர் முனைகள் வரையறுக்கப்பட்ட செயலாக்க திறன்கள் மற்றும் நினைவக அளவைக் கொண்டுள்ளன. இவை தகவல்களை சேகரித்து நுழைவாயில் மூலம் பெறப்பட்ட தரவுகளை பயனர்களுக்கு அனுப்புகிறது. பாதுகாப்பற்ற பாதையில் தரவு வழங்கப்படுவதால், அங்கீகரிக்கப்படாத அணுகல், முறைகேடான கண்காணிப்பு மற்றும் மாற்றியமைத்தல் உள்ளிட்ட ஆபத்துகளுக்கு எதிராக இதை திறம்பட பாதுகாக்கப்பட வேண்டும். குறியாக்கவியல் என்பது உணர் முனைகளின் தரவு மற்றும் இரகசியத்தன்மை, ஒருமைப்பாடு மற்றும் செல்லுபடியாகும் தன்மையை உறுதி செய்வதற்கான ஒரு சக்திவாய்ந்த கருவியாகும்[3]. சுருக்கமாக, குறியாக்கவியல் என்பது பாதுகாப்பற்ற தரவை படிக்க முடியாத பாதுகாப்பான தரவுகளின் தொகுப்பாக மாற்றுவதற்கான நுட்பங்களின் தொகுப்பாகும். தரவு குறியாக்கத்திற்கு, குறியாக்கவியல் பொதுவாக பாதுகாப்பு விசைகளைப் பயன்படுத்துகிறது, இது குறியாக்கம் மற்றும் மறைகுறியாக்க செயல்பாட்டில் நெகிழ்வுத்தன்மையை வழங்குகிறது. குறியாக்கவியல் வழிமுறைகள், வெவ்வேறு கணித வழிமுறைகளை அடிப்படையாகக் கொண்டவை மற்றும் வெவ்வேறு நுட்பங்களைப் பயன்படுத்துகின்றன, பொதுவாக, செயலாக்கம் மற்றும் நினைவக செலவுகள் மற்றும் தாக்குதல் எதிர்ப்பு ஆகியவற்றின் அடிப்படையில் வேறுபட்ட செயல்திறன் இருக்கும், இது மிகவும் பொருத்தமான வழிமுறைகளைத் தேர்ந்தெடுப்பது ஒரு முக்கியமான வடிவமைப்பு முடிவாகும்[4]. வளக் கட்டுப்பாடுகள் காரணமாக, கம்பயில்லா உணர் முனைகளின் கட்டமைப்புகள் பாதுகாப்பு அச்சுறுத்தல்களான சுற்று சிதைவு, ஊடுருவல் மற்றும் தவறான செய்தி போன்றவற்றால் பாதிக்கப்படக்கூடியவை. இதன் விளைவாக, குறிப்பிட்ட கம்பயில்லா உணர் வலைதளங்களின் அம்சங்களுக்கு இணங்கக்கூடிய மிகவும் பயனுள்ள பாதுகாப்பு முறைகள் வலைதளத்திற்கு அனுப்பப்படுகின்றன[5]. சமச்சீர் குறியாக்கவியல் [6], [7] என்பது பாதுகாப்பு அம்சங்களை வழங்கக்கூடிய மிகவும் பரவலான பாதுகாப்பு அணுகுமுறையாகும். இரண்டு முனைகள் ஒன்றுடன் ஒன்று தொடர்பு கொள்ள விரும்பினால், அவை அத்தகைய பாதுகாப்பு முறைகளில் குறியாக்கம் மற்றும் மறைகுறியாக்க செயல்முறைக்கு பொதுவான விசையைப் பயன்படுத்துகின்றன. தகவல்தொடர்பு பாதுகாப்பு மற்றும் அங்கீகாரத்தை வழங்க முனைகள் இந்த சமச்சீர் விசையை முன்பே தேர்ந்தெடுத்து பகிர்ந்துள்ளன. விசை மேலாண்மை [8] என்பது பகிரப்பட்ட சமச்சீர் விசைகளை உருவாக்கும் செயல்முறையாகும் [9].

பாதுகாப்பான தரவு பரிமாற்றத்திற்காக, சமச்சீர் மற்றும் சமச்சீரற்ற விசை (H-SAKAT) வழிமுறைகளை கொண்ட பாதுகாப்பான மற்றும் திறமையான கலப்பின டோக்கனை இந்தத் ஆய்வு கட்டுரை வழங்குகிறது. டோக்கன் உருவாக்கம், ரகசியம், பொது, தனிப்பட்ட விசை உருவாக்கம், தரவு குறியாக்கம் மற்றும் மறைகுறியாக்கம் ஆகியவை முன்மொழியப்பட்ட அணுகுமுறையில் நான்கு செயல்முறைகளாகும். வலைப்பின்னலில் உள்ள ஒவ்வொரு உணர் முனைக்கும், அடிப்படை நிலையம் முதலில் சமச்சீர் மற்றும் சமச்சீரற்ற விசைகளை உருவாக்குகிறது. உணர் முனைகள் தனிப்பட்ட டோக்கன்களை உருவாக்குகிறது, இது உணர் முனை மற்றும் தரவின் ஒருமைப்பாட்டை சரிபார்க்க நுழைவாயில் முனை பயன்படுத்தும். உணரப்பட்ட தரவை குறியாக்க (H-SAKAT) குறியாக்க வழிமுறை பயன்படுத்தப்படுகிறது. மறைகுறியாக்கப்பட்ட செய்தி நுழைவாயில் முனையிலிருந்து அடிப்படை நிலையத்திற்கு அனுப்பப்படுகிறது. மறைகுறியாக்கப்பட்ட செய்தி அடிப்படை நிலையத்தால் பெறப்படுகிறது, இது (H-SAKAT) மறைகுறியாக்க முறையைப் பயன்படுத்தி மறைகுறியாக்குகிறது. இந்த தாளின் மீதமுள்ள பகுதி பின்வருமாறு ஒழுங்கமைக்கப்பட்டுள்ளது: பகுதி 2 பாதுகாப்பான தரவு பரிமாற்றம் தொடர்பான பணிகளை மதிப்பாய்வு செய்கிறது. பகுதி 3 சமச்சீர் மற்றும் சமச்சீரற்ற முக்கிய தரவு பரிமாற்றத்துடன் முன்மொழியப்பட்ட கலப்பின டோக்கனை விளக்குகிறது. இறுதியாக பிரிவு 4 இல் வழங்கப்பட்ட முன்மொழியப்பட்ட அணுகுமுறையின் உருவகப்படுத்துதல் முடிவு கட்டுரையை முடிக்கிறது.

இரா. நேசமலர், முனைவர்ப்பட்ட ஆய்வாளர், கணிப்பொறி அறிவியல் துறை, தமிழ்ப் பல்கலைக்கழகம், இந்தியா

## 2. தொடர்புடைய படைப்புகள்

கம்பயில்லா உணரி வலைப்பின்னல் முக்கியத்துவத்துடன், தரவு பாதுகாப்பு மற்றும் பரிமாற்றம் எப்போதும் ஒரு பிரச்சினை. தற்போது, பெரும்பாலான ஆராய்ச்சி முயற்சிகள் பாதுகாப்பு மற்றும் பரிமாற்றத்தை மேம்படுத்துவதில் கவனம் செலுத்துகின்றன. அனுப்புநருக்கும் பெறுநருக்கும் இடையில் தரவு ஒருமைப்பாட்டைப் பேணுவதற்கு மறைகுறியாக்கப்பட்ட தரவின் பரிமாற்றம் அவசியமாகும். கம்பயில்லா உணரி வலைதளத்தின் வழியாக தரவு நகரும் போது, அதை அங்கீகரிக்கப்படாத பயனர்களிடமிருந்து பாதுகாக்கப்பட வேண்டும். கம்பயில்லா உணரி வலைப்பின்னலில் பாதுகாப்பான தரவு பரிமாற்றம் தொடர்பான சில வேலைகளை இந்தப் பகுதி விளக்குகிறது..

மறைகுறியாக்கப்பட்ட தரவை மாற்றுவதற்கு, மில்ரா மற்றும் சாஹு [10] ஒரு புதிய வழிமுறையை பரிந்துரைத்தனர். அனுப்பப்படும் தரவின் மறைகுறியாக்கப்பட்ட வடிவத்தைப் பெற, தற்போதைய ஆர்எஸ்எ மற்றும் டிரிபிள் டிஇஎஸ் ஆகியவற்றின் கலப்பினமாக்கல் பயன்படுத்தப்படுகிறது. பிரகாஷ் மற்றும் ராஜ்புத் [11] நேரத்தைச் சேமிக்கும் நுட்பத்தை உருவாக்கினர், இது குறைந்தபட்ச கணினி வளங்களை உட்கொள்ளும் போது தரவு ஒருமைப்பாடு மற்றும் இரகசியத்தை உறுதி செய்கிறது. இந்த ஆசிரியர் சமச்சீற்ற விசை வழிமுறையான ECC (நீள்வட்ட வளைவு குறியாக்கவியல்), மற்றும் சமச்சீர் விசை வழிமுறையான AES (மேம்பட்ட குறியாக்க தரநிலை) ஆகியவற்றைப் பயன்படுத்துகிறார். விசை உருவாக்கம் மற்றும் பகிர்வுக்கு, ECC பயன்படுத்தப்படுகிறது, மேலும் AES தரவு குறியாக்கம் மற்றும் மறைகுறியாக்கத்திற்கு பயன்படுத்தப்படுகிறது..

முனை மற்றும் முனை தொடர்புக்கு, காசி மற்றும் பலர், [12] பாதுகாப்பான அங்கீகாரம் மற்றும் தரவு குறியாக்கத்தை புதுமையான முறையில் பரிந்துரைக்கின்றனர். நீள்வட்ட வளைவு டிஜிட்டல் கையொப்பம் குறியாக்கவியல் திட்டத்தின் உதவியுடன், இத்திட்டமானது முனை மற்றும் முனை தொடர்பு வலைதளத்திற்கு பாதுகாப்பை வழங்குவதோடு மட்டுமல்லாமல், விசை உருவாக்க நேரம், ஹலோ செய்திகளின் எண்ணிக்கையை அளவிடுவதற்கான பொருத்தமான வழிமுறையை வழங்குவதன் மூலம் முனைகளின் நினைவக இடத்தையும் சேமிக்கிறது. மேலும், கம்பயில்லா பாதுகாப்பான தொடர்பு முனைகளின் பாதுகாப்பான தகவல்தொடர்புக்கு உதவுகிறது, இது முழு வலைதளங்களின் சிறந்த மற்றும் திறமையான பாதுகாப்பிற்கு உதவுகிறது. ஒரு அங்கீகார நெறிமுறையின் உதவியுடன், வலைதளத்தில் உள்ள ஆபத்து மற்றும் பாதுகாப்பு அபாயங்களின் தன்மையை குறைக்கிறது.

அமிடேபே மற்றும் பலர், [13] விவசாய வயல் பயிர் கண்காணிப்புக்கான இணைய அடிப்படையிலான அமைப்பில் தரவு பரிமாற்றத்திற்கான பாதுகாப்பான சேவை கட்டமைப்பை வழங்குகிறது. வடிவமைப்பு அளவைப் பொறுத்து, இந்த மறைநூல் அமைப்பு சமச்சீற்ற மற்றும் சமச்சீர் குறியாக்கவியல் நுட்பங்களை ஒருங்கிணைத்து ஒரு கலப்பு அணுகுமுறையைப் பயன்படுத்துகிறது. வடிவமைப்பு அளவைப் பொறுத்து, இந்த மறைநூல் அமைப்பு சமச்சீற்ற மற்றும் சமச்சீர் குறியாக்கவியல் நுட்பங்களை ஒருங்கிணைத்து ஒரு கலப்பு அணுகுமுறையைப் பயன்படுத்துகிறது.

பரிமாற்றத்தின் போது தரவு ஒருமைப்பாட்டை பராமரிக்க, பாபேர் மற்றும் அல்-அஹ்மதி [15] உணர்திறன் ஆற்றல் திறமையான உணரி வலைதளத்தின் நெறிமுறை மற்றும் வாட்டர்மார்க்கிங் முறைகளின் அடிப்படையில் இலகுவாக, பாதுகாப்பான தீர்வை வழங்குகின்றனர். இந்த அமைப்பில் பயன்படுத்தப்படும் ஒரே மாதிரியான குறியாக்கம், உணரி முனைகள் வேகமாகவும் திறமையாகவும் அடையாளம் காணும் அதே வேளையில், சிங்க்ஹோல் கண்டறிதல் மற்றும் தடுப்புக்கு குறைந்த ஆற்றலைப் பயன்படுத்துகிறது.

ஜஹான் மற்றும் பலர் [16], ஒரு புதிய பல-செயல்பாட்டு பாதுகாப்பான தரவு மொத்த நுட்பத்தை பரிந்துரைக்கின்றனர், இது அசல் தகவலை, மதிப்பு, ஒழுங்கு மற்றும் சூழல் பாதுகாப்பை செயல்படுத்துவதற்கு நன்கு வரையறுக்கப்பட்ட திசையன்களாக மாற்றுகிறது, எனவே பல-செயல்பாட்டுத் திரட்டலுக்கான கட்டுமானத் தொகுதிகளை வழங்குகிறது. உள்ளே மறைக்குறியீடு ஒருங்கிணைப்பு மற்றும் இறுதி முதல் இறுதி பாதுகாப்பு ஆகியவற்றை அடைய, ஒரே மாதிரியான குறியாக்க நெறிமுறைகள் பயன்படுத்தப்படுகிறது.

## 3. முன்மொழியப்பட்ட முறை

இந்த பகுதி முன்மொழியப்பட்ட பாதுகாப்பான தரவு பரிமாற்றத்தை விளக்குகிறது. முன்மொழியப்பட்ட கம்பயில்லா உணரி வலைப்பின்னல் ஆனது அடிப்படை நிலையம், நுழைவாயில், மற்றும் உணரி முனை ஆகியவற்றைக் கொண்டுள்ளது. அடிப்படை நிலையம் வலைப்பின்னலின் மைய அலகு ஆகும். இது பிணையத்தில் உள்ள ஒவ்வொரு உணரி முனைக்கும் விசைகளை உருவாக்குகிறது. ஒவ்வொரு உணரி முனையும் தனித்தனி டோக்கன்களை உருவாக்குகிறது. உணரப்பட்ட தரவு டோக்கன்களுடன் சமச்சீர் மற்றும் சமச்சீற்ற விசையைப் பயன்படுத்தி குறியாக்கம் செய்யப்பட்டு நுழைவாயில் முனை வழியாக அடிப்படை நிலையத்திற்கு அனுப்பப்படுகிறது. நுழைவாயில் முனை உணரிமுனை டோக்கனின் அடிப்படையில் உணரப்பட்ட தகவலைச் சரிபார்த்து, அதை அடிப்படை நிலையத்திற்கு அனுப்புகிறது.

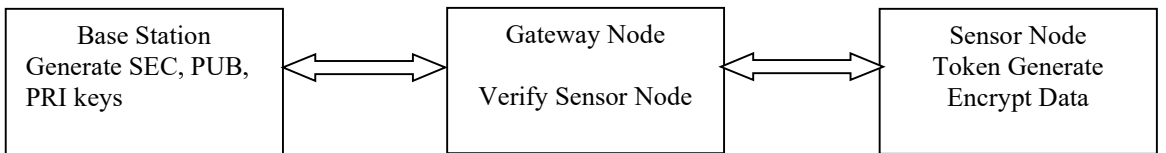


Figure shows the flow of data in WNS

நெறிமுறை முன்மொழியப்பட்ட (H-SAKAT) தரவு பரிமாற்றத்தை விளக்குகிறது. இந்த நெறிமுறையில், அடிப்படை நிலையம் ப்ளோஃபிஷ் மற்றும் ஆர்எஸ்எ நெறிமுறைகளைப் பயன்படுத்தி இரகசிய, பொது மற்றும் தனிப்பட்ட விசையை உருவாக்குகிறது (படி 1-6). டோக்கன் உருவாக்கும் செயல்முறை விளக்கப்பட்டுள்ளது (படி 7-13). படி (13-18) தரவு குறியாக்க செயல்முறையை விளக்குகிறது. டோக்கன் சரிபார்ப்பு (படி 15-20) இல் விவரிக்கப்பட்டுள்ளது. மறைகுறியாக்க செயல்முறை விளக்கப்பட்டுள்ளது (படி 21-23)

நெறிமுறை : *H-SAKAT* தரவு பரிமாற்றம்

**Input:** Base Station (BS), Gateway Node (GN) and Sensor Node  $SN=\{n_1, n_2, n_3, \dots, n_k\}$

**அடிப்படை நிலையம் உருவாக்க விசை**

Step 1: For  $i = 1$  to  $k$   
 Step 2:  $KYG = \text{GenerateKey}()$   
 Step 2:  $SKEY[i] = \text{KeyGen.SecKey}$   
 Step 3:  $PUK[i] = \text{KeyGen.PubKey}$   
 Step 4:  $PRK[i] = \text{KeyGen.PrivKey}$   
 Step 5: Distribute to  $i^{\text{th}}$  sensor node  
 Step 6: End If

**உணரி முனைகள் டோக்கனை உருவாக்குதல்**

Step 7:  $Nid = \text{Get sensor node id}$   
 Step 8:  $a1$  and  $a2 = \text{select two random numbers from } 0 \text{ to } 9$   
 Step 9:  $uc1$  and  $uc2 = \text{select two char from } A \text{ to } Z$   
 Step 10:  $lc1$  and  $lc2 = \text{select two char from } a \text{ to } z$   
 Step 11:  $\text{Token} = \text{append}(nid, a1, a2, uc1, uc2, lc1, lc2)$   
 Step 12: Use SHA-1 to generate hash value  
 Step 13:  $\text{Token1} = \text{SHA1}(\text{Token})$

**உணரி முனை தரவு குறியாக்கம்**

Step 14:  $\text{PubKey} = \text{Get public key}(nid)$   
 Step 15:  $\text{Encry1} = \text{Encrypt data using RSA algorithm with PubKey}$   
 Step 16:  $SK = \text{Get Secret Key}(nid)$   
 Step 17:  $\text{Encry2} = \text{Encrypt data using Blowfish algorithm using SK}$   
 Step 18:  $\text{EncryMg} = \text{Enc2}$

**நுழைவாயில் முனை டோக்கனை சரிபார்க்கிறது**

Step 19:  $\text{Get hTK}[nodeId];$   
 Step 20: If  $(\text{hTK}[nodeId] == \text{hash}(\text{Token}))$   
 Step 21: Send  $\text{encMsg}$  to Base Station  
 Step 22: Else  
 Step 23: Discard  $\text{encMag}$   
 Step 24: EndIf

**அடிப்படை நிலையம் தரவை மறைகுறியாக்கல்**

Step 25:  $SK = \text{Get Secret key}(nid)$   
 Step 26:  $\text{Dcry1} = \text{Decrypt data using Blowfish algorithm with SK}$   
 Step 27:  $\text{PrivKey} = \text{Get Private Key}(nid)$   
 Step 28:  $\text{Dcry2} = \text{Decrypt data using RSA algorithm using PrivKey}$   
 Step 29:  $\text{OrigMg} = \text{Dec2}$

#### 4. உருவகப்படுத்துதல் முடிவுகள்

இந்த பகுதி முன்மொழியப்பட்ட வேலையின் உருவகப்படுத்துதல் முடிவுகளை விளக்குகிறது. ஆரம்பத்தில், நெட்வொர்க் 50 உணரி முனைகளுடன் உருவாக்கப்படுகிறது, அவை குறிப்பிட்ட பகுதியில் தோராயமாக பொருத்தப்படுகின்றன. வலைபின்னலில் ஒரு அடிப்படை நிலையம் மற்றும் சில எண்ணிக்கை (6 - 10) கேட்வே முனைகள் உள்ளன. சுற்றுச்சூழல் நிலை (வெப்பநிலை, காற்றின் வேகம் போன்றவை) தொடர்பான தகவல்களை, மருத்துவம் தொடர்பான தரவு (இதய துடிப்பு, நாடித்துடிப்பு போன்றவை) உணரி முனை உணரும். இந்தத் தரவு நுழைவாயில் முனை வழியாக அடிப்படை நிலையத்திற்கு அனுப்பப்படும். முன்மொழியப்பட்ட *H-SAKAT* ஜாவாவைப் பயன்படுத்தி செயல்படுத்தப்பட்டது மற்றும் முன்மொழியப்பட்ட நெறிமுறையின் செயல்திறன் குறியாக்க நேரம், மறைகுறியாக்க நேரம் மற்றும் மறைகுறியாக்கப்பட்ட செய்தி அளவு ஆகியவற்றின் அடிப்படையில் மதிப்பிடப்பட்டது.

அட்டவணை 1: குறியாக்க நேரத்தின் ஒப்பீட்டை காட்டுகிறது

Message Size (bytes)	THCA	HCA	HET	H-SAKAT
608	992	533	453	161
25514	1033	987	467	193
34091	1062	1029	479	240
62586	3297	2254	484	257
143516	3845	3287	492	296
209856	4786	4279	974	439

அட்டவணை 2 மறைகுறியாக்க நேரத்தின் ஒப்பீட்டைக் காட்டுகிறது.

Message Size (bytes)	THCA	HCA	HET	H-SAKAT
608	588	213	207	118
25514	754	647	379	169
34091	838	787	393	187
62586	889	798	397	234
143516	910	874	406	346
209856	1530	1542	792	525

##### 5. முடிவுரை

கம்பயில்லா உணர் வலைப்பின்னலில் சிறந்த பாதுகாப்பான தரவு பரிமாற்றத்தை வழங்க, இந்த தாள் டோக்கனுடன் ஒரு கலப்பின சமச்சீர் மற்றும் சமச்சீரற்ற விசை நெறிமுறையினை முன்மொழிகிறது. இந்தத் தாள் முதலில் ஆர்எஸ்எ (விசைச்சீர்) நெறிமுறையை பயன்படுத்தி உணரப்பட்ட தகவலை குறியாக்கம் செய்து மீண்டும் எஇஎஸ்(சமச்சீர்) வழிமுறையைப் பயன்படுத்தி குறியாக்கம் செய்யப்படுகிறது. இது உயர் அளவு பாதுகாப்பை வழங்குகிறது. குறியாக்க நேரம், மறைகுறியாக்க நேரம் மற்றும் மறைகுறியாக்கப்பட்ட செய்தியின் அளவு ஆகியவற்றின் அடிப்படையில் உரைநடை வேலையின் செயல்திறன் மதிப்பிடப்படுகிறது. ஏற்கனவே உள்ள மற்ற அணுகுமுறையுடன் ஒப்பிடும்போது முன்மொழியப்பட்ட வழிமுறை அனைத்து அளவீடுகளையும் குறைக்கிறது என்பதை முடிவு காட்டுகிறது.

##### REFERENCES

- [1] Y. Jin, K. S. Kwak, and S.-J. Yoo, "A novel energy supply strategy for stable sensor data delivery in wireless sensor networks", IEEE Syst. J., vol. 14, no. 3, pp. 3418–3429, 2020.
- [2] A. H. Mohajerzadeh, H. Jahedinia, Z. Izadi-Ghodousi, D. Abbasinezhad-Mood, and M. Salehi, "Efficient target tracking in directional sensor networks with selective target area's coverage", Telecommun. Syst., vol. 68, no. 1, pp. 47–65, May 2018.
- [3] K.A. Shim, "A Survey of Public-Key Cryptographic Primitives in Wireless Sensor Networks", IEEE Commun. Surv. Tutor., vol. 18, pp. 577–601, 2015
- [4] D.G. Costa, S. Figueredo, & G. Oliveira, "Cryptography in wireless multimedia sensor networks: A survey and research directions", Cryptography, vol. 1, no. 1, 2017
- [5] M. F. Moghadam, M. Nikooghadam, M. A. B. A. Jabban, M. Alishahi, L. Mortazavi and A. Mohajerzadeh, "An Efficient Authentication and Key Agreement Scheme Based on ECDH for Wireless Sensor Network," in IEEE Access, vol. 8, pp. 73182–73192, 2020
- [6] M. S. Yousefpoor and H. Barati, "Dynamic key management algorithms in wireless sensor networks: A survey", Comput. Commun., vol. 134, pp. 52–69, Jan. 2019.
- [7] F. Zhan, N. Yao, Z. Gao, and G. Tan, "A novel key generation method for wireless sensor networks based on system of equations", J. Netw. Comput. Appl., vol. 82, pp. 114–127, Mar. 2017.
- [8] S. Athmani, A. Bilami, and D. E. Boubiche, "EDAK: An efficient dynamic authentication and key management mechanism for heterogeneous WSNs", Future Gener. Comput. Syst., vol. 92, pp. 789–799, Mar. 2019

- [9] C. Mishra and B. Sahu, "Transmission of Encrypted data in WSN: An Implementation of Hybridized RSA-TDES Algorithm," 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), pp. 1-6, 2020
- [10] S. Prakash, & A. Rajput, "Hybrid cryptography for secure data communication in wireless sensor networks", In Ambient Communications and Computer Systems, Springer, pp. 589-599, 2018
- [11] R. Qazi, K.N. Qureshi, F. Bashir, et al. "Security protocol using elliptic curve cryptography algorithm for wireless sensor networks", J Ambient Intell Human Comput, vol. 12, pp. 547-566, 2021
- [12] A.F. -X. Ametepe, S.A.R.M. Ahouandjinou and E.C. Ezin, "Secure Encryption by Combining Asymmetric and Symmetric Cryptographic Method for Data Collection WSN in smart Agriculture," 2019 IEEE International Smart Cities Conference (ISC2), pp. 93-99, 2019
- [13] A. Wang, J. Shen, P. Vijayakumar, Y. Zhu, L. Tian, "Secure big data communication for energy efficient intra-cluster in WSNs", Information Sciences, vol. 505, pp. 586-599, 2019
- [14] H.A. Babaeer and S.A. Al-Ahmadi, "Efficient and Secure Data Transmission and Sinkhole Detection in a Multi-Clustering Wireless Sensor Network Based on Homomorphic Encryption and Watermarking," in IEEE Access, vol. 8, pp. 92098-92109, 2020
- [15] P. Zhang, J. Wang, K. Guo, F. Wu, G. Min, "Multi-functional secure data aggregation schemes for WSNs", Ad Hoc Networks, vol. 69, pp. 86-99, 2018.



## Mapping Comparative constructions in Tamil and English for Machine Translation

Dhanalakshmi V<sup>1</sup> and Rajendran S<sup>2</sup>

*1School of Tamil, Pondicherry University, Pondicherry  
dhanagiri@pondiuni.ac.in*

*2CEN, Amrita Vishwa Vidyapeetham, Coimbatore  
rajushush@gmail.com*

---

### ABSTRACT

Comparative construction is a linguistic technique for demonstrating two or more items to show their similarities and differences. Typically, a comparative construction consists of a predicate and two noun phrases, one of which serves as the “criterion” of the comparison and the other as the object of comparison (the comparee NP). The sentences that are comparable to sentences like “Raja is taller than Roja”, where the noun phrase after the item “than” is the standard NP, are prototype examples of comparative constructions in the languages of the globe. As suggested by Dixon (2012), a prototypical comparative construction has three elements: participants of comparison (comparee and standard of comparison), the property (parameter of comparison), and the index of comparison. In this paper, we are not concerned with the typological study of comparative constructions. We have focused on mapping comparative constructions in Tamil with that of English and formulated rules for Machine Translation. The translation between Comparative construction sentences in Tamil and English are tested with the existing Machine Translation system and their results are discussed.

---

### Keywords:

A linguistic technique  
B prototypical  
C Machine Translation  
D NP  
E noun phrases

---

### Corresponding Author:

Dhanalakshmi V  
School of Tamil, Pondicherry University, Pondicherry  
dhanagiri@pondiuni.ac.in

---

## 1. INTRODUCTION

A prototypical comparative construction involves a quality or property whose extent is compared, the entity being compared, and the standard of comparison. Comparison is a mental act by which two or more items are examined in order to assess similarities or differences between them. The comparison can be made with regard to a certain gradable, one-dimensional property, and the items are then assigned a position on a predicative scale. This mental act of comparison finds its linguistic encoding in comparison constructions, especially comparative constructions for the expression of comparison of inequality or equative constructions for the expression of comparison of equality. The linguistic literature has especially been concerned with comparison of inequality and comparative constructions as found in the following English sentences (Yvonne Treis. 2018).

- (1) Mary is tall-er than Peter - Comparee - Parameter - Parameter/Degree Marker -Standard Marker - Standard
- (2) Mary is more intelligent than Peter - Comparee - Parameter/Degree Marker - Parameter - Standard Marker - Standard

- Comparee (COM) = what is being compared against some standard of comparison (*Mary*); alternative terms used in the literature: Item compared.
- Standard of comparison (SOC) = what the comparee is being compared against (*Peter*)
- Standard Marker = marker of the grammatical function of the standard (*than*), alternative terms used in the literature: marker, pivot, relator
- Parameter of comparison (POC) = property of comparison (*tall, intelligent*); alternative terms used in the literature: Quality or Quantity, comment, (comparative) predicate.
- Degree Marker or Parameter Marker : It marks the degree of presence or absence of a property in the comparee (*more* or *-er*); alternative terms used in the literature: Index, comparative concept.

Dixon (2005) makes use of the terms comparee, index, parameter, mark and standard. The following example will exemplify the use of these terms.

John is more famous than Bill.  
comparee index parameter mark standard

In traditional grammar of English, four degrees of comparison of the adjective are distinguished. They are positive degree, equative degree, comparative degree, and superlative degree (Yvonne Treis. 2018).

Positive degree: basic form of the adjective; *Susan is tall* - positive construction

Equative degree: parameter is ascribed to the comparee and the standard to the same extent; *Susan is as tall as Peter* - equative construction

Comparative degree: Parameter applies to the comparee to a higher extent than to the standard; *Susan is tall-er than Peter* - comparative construction

Superlative degree: shows the highest degree of the parameter applied to the comparee; *Susan is the tallest of her family* - superlative construction

Tamil does not make use of degree marker or parameter marker; it makes use of only parameter of comparison. In Tamil, there is no morphological distinction between positive degree, comparative degree and superlative degree. Rajendran (1976-77) elaborately studied comparison of inequality and equality in Tamil.

## 2. COMPARISON OF INEQUALITY

In a prototypical comparative construction in Tamil and English, the comparee occupies the subject position, and the standard of comparison occupies the predicate position. In English, the standard of comparison occupies a position at the end of the comparative construction after the parameter of comparison whereas in Tamil, the standard of comparison occupies the predicative position before the parameter of comparison. The standard of comparison is marked for accusative case in Tamil. The parameter of comparison does not make use of a comparative degree marker; it is pronominalized to agree with the subject NP. English makes use of *than* as parameter of comparison and Tamil makes use of *viTa* or *kaaTTilum* as parameter of comparison.

- (3) raaNi raataiy-ai vita/kaaTTilum azahkaana-vaL  
Rani Radha-ACC POC ADJ-PN  
'Rani is more beautiful than Radha'

The mapping between Tamil and English can be given as follows:

### Mapping rule 1:

$NP_{COM} + NP_{-ai} + vita/kaaTTilum + ADJ-PN = NP_{COM} + BE + more + ADJ + than + NP_{SOC}$

In the superlative comparative construction, the parameter of comparison is marked for superlative degree in English. In Tamil, the parameter of comparison is not marked for superlative degree. The standard of comparison has to be an inclusive NP of superlative nature meaning 'of all', 'among all' and so on. In Tamil too, the standard of comparison in superlative comparative construction must be inclusive nature: *avarkaL elloorilum* 'among all'.

- (4) raaNi avarkaL elloor-ai-yum viTa ahakaana-vaL  
Rani they all-ACC-EMP than ADJ-PN  
Rani is most beautiful among all

### Mapping rule 2:

$NP + NP_{-ai} + viTa/kaaTTilum + ADJ-PN = NP_{COM} + BE + most + ADJ + among\ all.$

In English, certain adjectives inflect for comparative degree is marked with -er instead of *more*.

- (5) raaNi raataiy-ai viTa/kaaTTilum uyaramaanavaL  
Rani Radha-ACC POC ADJ-PN  
Rani is taller than Radha

**Mapping rule 3:**

$NP_{COM} + NP_{-ai} + \text{vita/kaaTTilum} + ADJ-PN = NP_{COM} + BE + ADJ\text{-er} + \text{than} + NP_{SOC}$

Some adjectives in English, inflect for comparative marker -est instead of *most*.

- (6) raaNi avarkaL elloor-ai-yum viTa uyaramaana-vaL  
Rani they all-ACC-EMP than tall-PN  
Rani is tallest among all

**Mapping rule 4:**

$NP + NP_{-ai} + \text{viTa/kaaTTilum} + ADJ-PN = NP_{COM} + BE + ADJ\text{-est} + \text{among all}$ .

In the place of ADJ-PN Tamil can make use of noun denoting quality +adverbial marker when followed by the be-verb *iru*. English makes use of be-verb and adjective combination only. The following sentence will exemplify this statement.

- (7) raaNi raataiyai viTa/kaaTTilum azhak-aaka iru-kkiR-aaL  
Rani Radha-ACC than beauty-ADVP be-PRE-3FS  
Rani is more beautiful than Radha

**Mapping rule 5**

$NP_{COM} + NP_{-ai} + \text{vita/kaaTTilum} + N\text{-ADVP iru-TEN-PNG} = NP_{COM} + BE + \text{more} + ADJ + \text{than} + NP_{SOC}$

Similar to adjectives, adverbs too can make comparison of inequality. Consider the following example.

- (8) raaNi raataiy-ai viTa/kaaTTilum veekamaaka ooT-in-aaL  
Rani Radha-ACC than fast run-PAS-PNG  
Rani ran faster than Radha

In Tamil, the standard of comparison is marked for accusative marker; the standard marker *viTa/kaaTTilum* comes next and the parameter of comparison which is an adverb comes after standard marker and before the verb. In English, the verb comes before the parameter of comparison which is an adverb which is inflected for the comparative marker -er; the standard of comparison comes at the end and the standard marker *than* comes before the standard of comparison.

**Mapping rule 6:**

$NP_{COM} + NP_{-ai} + \text{viTa/kaaTTilum} + ADV + V\text{-TEN-PNG} = NP_{COM} + V\text{-TEN} + ADV\text{-er} + \text{than} + NP_{SOC}$

### 3. COMPARISON OF EQUALITY

If two or more items are found to be similar quantitatively or qualitatively they can be subjected to comparison of equality. Consider the following sentence.

- (9) raaNi raataiy-aip poola azhakaana-vaL  
Rani Radha-ACC like ADJ-PN  
'Rani is as beautiful as Radha'

In the above sentence, Rani is the compare, i.e. item compared; Radha is the standard of comparison; *poola* is the standard marker; and *azhakaana* 'beautiful' is the parameter of comparison. In English, the standard of comparison comes at the end; the parameter of comparison comes in-between the standard marker as-----as.

**Mapping rule 7:**

$NP + NP_{-ai} + \text{poola} + ADJ-PN = NP + BE + \text{as-ADJ-as} + NP_{SOC}$

Similar to adjectives, adverbs also undergo comparison of similarity. Consider the following sentence.

- (10) raaNi raataiy-aip poola veekamaaka ndaTa-kkiR-aaL  
Rani Radha-ACC like walk-PRE-3FS  
'Rani walks as fast as Radha'

In Tamil, the subject function as the comparee; the standard of comparison marked for accusative case follows it; the standard marker *poola* follows next; the parameter of comparison which is an adverb follows it and the verb which inflect for tense and person-number-gender (PNG) occupies the final position of the construction. In English, the subject NP function as the compare. The verb which is inflected for tense comes next. The parameter of comparison comes in between the standard marker 'as---as'.

#### Mapping rule 8:

$$NP_{COM} + NP_{-ai} + poola + ADV + V-TEN-PNG = NP_{COM} + V-TEN + as-ADV-as + NP_{SOC}$$

Comparison can be made without the explicit expression of adverb. In that context, the comparative construction become ambiguous. Consider the following example:

- (11) raaNi raataiy-aip poola ooTu-kiR-aaL  
Rani Radha-ACC like run-PRE-3FS  
Rani runs like Radha

The Tamil sentence is ambiguous as it can be interpreted in a number of ways: 'Rani, runs (instead of walking) like Radha', 'Rani runs in the same speed like Radha', Rani runs in the same style or manner like Radha' and so on. Consider the following example,

- (13) raaNi-kku raataiy-aip poola ceelai iru-kkiR-atu  
Rani-DAT Radha-ACC like sari be-PRE-3NS  
'Rani has sari like Radha'

The Tamil sentence is ambiguous inviting different interpretations: Rani has similar sari like Radha, Rani has sari of same colour like Radha, Rani has sari of same texture like Radha, and so on.

Comparison can be made without explicitly expressing the parameter of comparison. Consider the following example:

- (14) raaNi raataiy-aip poola iru-kkiR-aaL  
Rani Radha-ACC like be-PRE-3FS  
'Rani resembles Radha.'

The not-expression of parameter of comparison makes this sentence ambiguous allowing different interpretation from the point of view of quality and quantity.

#### Mapping rule 9:

$$NP_{COM} + NP_{-ai} + poola + iru-TEN-PNG = NP_{COM} + resemble-TEN + NP_{SOC}$$

*iru* can be replaced by *toonRu* 'appear' in the above construction of equality.

- (15) raaNi raatay-aip poola toonRu-kiR-aaL  
Rani Radha-ACC like appear-PRE-3FS  
Rani appears like Radha'

The addition of emphatic -ee can make the resemblance more closer.

- (16) raaNi raataiy-aip poolav-ee iru-kkiR-aaL  
Rani Radha-ACC like-EMP be-PRE-3FS  
'Rani resembles Radha very much.'

Instead of *poola* 'like', *maatiri* 'like', *aLavukku* 'as much', *attanai* 'that many' can be made use of as standard marker.

- (17) raaNikku raataiy-ai maatiri pasi.  
Rani-DAT Radha-ACC like hunger  
'Rani is hungry like Radha'
- (18) raaNikku raataiy-ai aLavukku pasi.  
Rani-DAT Radha-ACC that much hunger  
'Rani is that much hungry like Radha'
- (19) raaNikku raataiy-ai attanai pasi.  
Rani-DAT Radha-ACC that much hunger  
'Rani is as much hungry as Radha'

*aLavu* and *attanai* specifies quantity. Another way of expressing quantity for the sake of comparison is using *ettanai* 'how much' and *attanai* as exemplified in the following comparative construction.

- (20) raaNi-kku ettanai ceelai iru-kkiR-at-oo attanai ceelai raatai-kk-um iru-kkiR-atu  
Rani-DAT how-much be-PRE-3NS-Q that-much saree Radha-DAT-EM be-PRE-3NS  
'Radha has as many as saris Rani'

*ettanai---attanai*, *evvaLavu---avvaLavu* can be equated with *as many as* and *as much as* respectively. Similarly

eppaTi---appaTi and *evvaaRu--avvaaRu* can be equated with English *what manner--that manner* as exemplified in the following example.

- (21) raaNi eppaTi ooT-in-aaL-oo appaTi raat-aiy-um ooT-in-aaL  
Rani how run-PAS-3FS-Q that-manner Radha-ACC-EMP  
'Rani waked in the same manner like Radha'

Exact resemblance can be expressed by making use of the emphatic markers -ee and taan as exemplified by the following sentence.

- (22) raaNi raatai-ee taan  
Rani Radha-EMP EMP  
'Rani is exactly like Radha'

camamaaka 'equally', iNaiyaaka 'equally' can be used to specify the exactness in the resemblance.

- (23) raaNi raatai-kku samamaaka/iNaiyaaka paaT-in-aaL  
Rani Radha-DAT equally sing-PAS-3FS  
'Rani sang as equally as Radha'

One can notice similarity expressed in the following comparative phrases. In these constructions poonRa is used as standard marker

- (24) mati poonRa mukam  
moon like face  
'moon like face'
- (25) taamarai poonRa mukam  
lotus like face  
'lotus like face'

Comparison of equality can be studied elaborately, but such an elaborative method is not adopted here.

#### 4. RESULTS AND DISCUSSION

Comparative construction Sentences in Tamil are translated to English using the existing Translation Systems. The table given below shows the translation rules for the Comparative construction Sentences between Tamil and English and also the results of the existing Translation Systems.

Rule No	Tamil sentence & Rule	English Translation & Rule	Google	Bing - Microsoft	Systran
1.	ராணி ராதாவை விட அழகானவள் (NP <sub>COM</sub> + NP-ai + vita/ kaaTTilum + ADJ-PN)	Rani is <b>more</b> beautiful than Radha (NP <sub>COM</sub> + BE + <b>more</b> + ADJ + than + NP <sub>SOC</sub> )	Rani is <b>more</b> beautiful than Radha Yes	Rani is <b>more</b> beautiful than Radha Yes	<b>Queen</b> is <b>better</b> than Radha No
2.	ராணி அவர்கள் எல்லோரையும் விட அழகானவள் (NP+ NP-ai + viTa/kaaTTilum + ADJ-PN)	Rani is the <b>most</b> beautiful among all (NP <sub>COM</sub> + BE + <b>most</b> + ADJ + among all)	Rani is <b>more</b> beautiful than all of them	The <b>queen</b> is <b>more</b> beautiful than all of them	The <b>Queen</b> is <b>more</b> beautiful than them all
3.	ராணி ராதாவை விட உயரமானவள் (NP <sub>COM</sub> + NP-ai + vita/kaaTTilum + ADJ-PN)	Rani is <b>taller</b> than Radha (NP <sub>COM</sub> + BE + <b>ADJ-er</b> + than + NP <sub>SOC</sub> )	Rani is <b>taller</b> than Radha	Rani is <b>taller</b> than Radha	<b>Queen</b> is <b>taller</b> than Radha

4.	ராணி அவர்கள் எல்லோரையும் விட உயரமானவள் (NP+ NP-ai + viTa/kaaTTilum + ADJ-PN )	Rani is the <b>tallest</b> among all  (NP <sub>COM</sub> + BE + <b>ADJ-est</b> + among all)	Rani is <b>taller</b> than all of them	The <b>Queen</b> is <b>taller</b> than all of them	The <b>Queen</b> is <b>taller</b> than them all
5.	ராணி ராதையைக் காட்டிலும் அழகாக இருக்கிறாள் (NP <sub>COM</sub> + NP-ai + vita/ kaaTTilum + N-ADV iru-TEN- PNG)	Rani is <b>more</b> beautiful than Radha  (NP <sub>COM</sub> + BE + <b>more</b> + ADJ + than+ NP <sub>soc</sub> )	Rani is <b>more</b> beautiful than Radha	Rani is <b>more</b> beautiful than Radha	<b>Queen</b> is <b>more</b> beautiful than queen
6.	ராணி ராதாவை விட வேகமாக ஓடினாள் (NP <sub>COM</sub> + NP-ai +viTa/ kaaTTilum + ADV +V-TEN-PNG )	Rani ran <b>faster</b> than Radha  (NP <sub>COM</sub> + V-TEN+ ADV-er +than + NP <sub>soc</sub> )	Rani ran <b>faster</b> than Radha	Rani ran faster than Radha	<b>Queen</b> ran <b>faster</b> than Radha
7.	ராணி ராதாவைப் போல் அழகானவள் (NP + NP-ai + pool + ADJ-PN)	Rani is <b>as beautiful</b> <b>as</b> Radha  (NP + BE + as- ADJ-as + NP <sub>soc</sub> )	<b>Beautiful like</b> Rani Radha	Rani is <b>as</b> <b>beautiful as</b> Radha	<b>She is as</b> <b>beautiful as</b> queen Radha
8.	ராணி ராதாவைப் போல் வேகமாக நடக்கிறாள் ( NP <sub>COM</sub> + NP-ai + poola + ADV + V- TEN-PNG )	Rani walks <b>as fast</b> <b>as</b> Radha  (NP <sub>COM</sub> + V-TEN + as-ADV-as +NP <sub>soc</sub> )	Rani walks <b>as</b> <b>fast as</b> Radha	Rani walks <b>as</b> <b>fast as</b> Radha	<b>Queen acts</b> <b>swiftly as</b> Radha
9.	ராணி ராதாவைப் போல் இருக்கிறாள் (NP <sub>COM</sub> + NP-ai + poola + iru-TEN- PNG)	Rani looks/ resembles Radha.  (NP <sub>COM</sub> + Look/ resemble-TEN + NP <sub>soc</sub> )	<b>Looks like</b> <b>Rani Radha</b>	Rani <b>looks like</b> Radha	<b>She is like</b> <b>Rani Raddha</b>

10.	ராணி ராதையே தான்	'Rani is exactly like Radha'	Rani is Radha	It's Rani Radha	Rani Raddha
	Correct output percentage		60%	70%	40%

## 5. CONCLUSION

The Comparison of equality and inequality needs elaborate study. Only certain important aspects of comparison are studied here from the point of view of machine translation. We try to map comparative constructions of equality and inequality in Tamil and English by positing mapping rules. The Comparative construction Sentences in Tamil are translated to English using the existing Translation Systems and the results indicate that the existing Translation system needs to concentrate on the specific pattern of comparative construction Sentences between Tamil and English. The Proposed Mapping rules will enhance the results of the Machine Translation systems.

## REFERENCES (10 PT)

- [1] Dixon, R.M.W. 2005. Comparative constructions in English. *Studia Anglica Posnaniensia*, 41.
- [2] Dixon, R.M.W. 2008. Comparative constructions: A cross linguistic typology. *Studies in Languages* 32,4:787-817
- [3] Dixon, R.M.W. 2012. *Basic Linguistic Theory. Volume 3. Further Grammatical Topics*. Oxford: Oxford University Press.7
- [4] Rajendran, S. 1976-77. Comparative constructions in Tamil: *Bulletin of the Deccan College Research Institute*, volume xxxvi nos 1-4.
- [5] Yvonne Treis. 2018. Comparative constructions: An Introduction. *Linguistic Discovery* volume 16, Issue 1



## An Efficient Approach for Computer Aid Teaching and Learning Using DEEDs Lab

Mr.Dharmaraj<sup>1</sup>, Mr.Kirubakaran<sup>2</sup> Dr.Anastraj<sup>3</sup>

<sup>1,2,3</sup>Head & Assistant professor, Department of Computer Application, Loyola College, Vettavalam, Thiruvannamalai.

---

### ABSTRACT (10 PT)

The student performance prediction is an important role of teachers to analysis student need additional assist. To predict the difficulties of the student will use a digital design course namely technology enhanced learning (TEL) called Digital electronic education and design suite (DEEDS). The machine learning algorithm having a random forest, k-nearest neighbor and Support Vector Machine (SVM).the DEED system allows the student to solve the digital exercise different level of difficulty during logging the data. The input variable is current study of the students for each and every individual session. The output variables were the student's grade for the respective session. We trained the machine learning of the data from the five sessions and tested the algorithm remaining last session. To performed 10-fold cross –validation and computed the receiver operating characteristic to evaluate the performance of the model. The results show the SVM given higher accuracy. The SVM can easily combine the TEL system. An expect to instruct to improve the student performance an every session.

---

### Keywords:

A Technology Enhanced Learning (TEL)  
B. Linear Regression  
C. Artificial Neural Network (ANN),  
D. Support Vector Machine (SVM).  
D. Root Mean Square Error  
E ), Digital Electronic Education and Design Suite (DEEDS)

---

### Corresponding Author:

Mr.Dharmaraj  
Head & Assistant professor, Department of Computer Application,  
Loyola College, Vettavalam, Thiruvannamalai.

---

## 1. INTRODUCTION

Forecasting student presentation is necessary for educators to get near the beginning reaction and take direct action or early safety measures if essential to get better the student's presentation. This forecast can be managed by place the basis of the problem. Should it be from additional behavior that the student is contributing in family troubles, or health problems? This entire factor can have a most important effect on student performance. By means of having a dataset for student's performance can help us study such cases. They are K- Nearest Neighbor (KNN) to predict the final grade of the student which falls in the maximum.

In the progress of a nation's economy and literate society development. Education also enhances decision-making process and constructs the competitive present generation. Different data mining tools are used to predict the student performance. The presentation of the student different throughout a year and the number of a student failure enlarge due to the decrement in the student performance with numerous features. The most known reason of the failure of the student is don't have of deep knowledge regarding the course, complicated to adjust the new environment, social media usage, romantic affiliation and teacher approach to the student and course. The prediction of the student performance has the advantage of indicating earlier the factor and provides prevention solution. Data mining is used to generate the meaningful information from massive data set using some patterns.

The utilization of Data Mining in the Educational perspective is mentioned as Educational Data Mining (EDM) and explained By the International Educational Data Mining Society in as “new coming discipline, concentrate with improve Methods for examine the single types of data that appear from Educational surroundings. Data mining techniques applied to foretell the educational achievement of the learners based on their socio-economic condition Using data mining techniques assist for anticipate the student accomplishment, indicate resource and aids to make a decision.

To predict the students' realization is a very important part of a higher education, as the whole enlargement of the Education system is directly relative to develop the success rate of the learner, increase students' product and reduce the dropout rate. Therefore there are many Situations where the performance of the students' needs to be predicted, for example, to differentiate allowed students for joins in placement activities and to find the pathetic students so that different essential achievement took to modify.

A decision tree is a flow –charts like tree structure which is made of nodes and arcs .Each internal nodes represented by rectangles and the leaf node represented By Oval. Decision tree mostly used for the use of decision making. It always initiate from the node to take steps, and from this position, the user drop an each node continuous.

Classification techniques is one of the important part used applications of data mining. The major function of classification is allocating class label to a set of possible class values to an unseen instance constitute a set of variables. the r classification techniques applied to foresee the student achievement. Using data mining techniques support for approximation the student achievement indicate resource and assist to make a result. Prediction the educational achievement is essential Part of a higher education, as the whole enlargement of the education system is directly proportional to develop the success rate of the students increase students' result and reduce the dropout rate. Therefore there are many circumstances.

## 2. RELATED WORK

Artificial neural network (Multilevel perception) ANN, another technique used in the different paper which improves the student and instructors performance, Similar to a human brain, Artificial neuron network are predominantly shown as a system of interconnected neurons that interchange information between each other. Multilevel perception present best between all classifier and more well-organized when there are large data set. It is originate in the weka tools to make investigation by using the name sequential minimal optimization. Accuracy values, which calculate the efficiency of the models, are all at least approximately 90% [5].

Lotfi Najdi et al [19] Predictive modeling system random forest and CART are operated to predict the student retention and graduation. In this system the student at risk of dropout identified. Ensemble method baggings develop in the system to enlarge the performance of the algorithm. The very important attribute in the paper are GPA and secondary school .The accuracy of the algorithm is 88%.

Marbouti et al. (2016) used logistic regression, support vector machines (SVMs), decision trees (DTs), ANNs and a Naïve bayes classifier (NBC) to identify at- risk students in advance of the next course. This study used input features, such as grades, attendance, quizzes, weekly homework, team participation, project milestones, mathematical modeling activity tasks, and exams from an offline course. Analysis of the results found that the NBC algorithm provided satisfactory accuracy (85%)

## 3. PROPOSED WORK

It is a well-known fact that the selection of an optimal set of classifiers is an important part of multiple classifier systems and the independence of classifier outputs is generally considered to be an advantage for obtaining better multiple classifier systems. In terms of classifier combination, the voting methods

demand no prerequisites from the classifiers. When multiple classifiers are combined using voting methodology, we expect to obtain good results based on the belief that the majority of experts are more likely to be correct in their decision when they agree in their opinion.

## Dataset

Data may be obtained from many different and varied data sources. This stage comprises of gathering all available information on students. The set of factors that can affect the students' performance is first identified and collected from various sources of data available. This is then integrated into a single data set. One of the key parts in EDM is prediction. It is necessary to investigate and predict students' academic progress and performance. It is a complex research undertaking to identify and indicate the issues the difficulties students' academic performance. Several unrelated and redundant data can also be found in academic information which affects the outcomes of prediction. To maximize relevancy of features and minimize the redundancy of data. An organized literature evaluation was provided on clustering algorithm and its suitability and serviceability in the environment of EDM.

### Activity Selection

The details about the activities and their meanings are as follows:

For example the keywords of our interest relevant to the course and educational activities of the students and if nothing is found, we assigned "Other".

Abbreviation of activities:

Es: Exercise#: Number

Deeds: Digital Electronics Education and Design Suite

Diagram: Simulation Timing Diagram  
FSM: Finite State Machine Simulator  
Description of activities:

Study\_Es\_# session of exercise#

It indicates that a student is studying / viewing the content of a specific exercise.

Deeds\_Es\_# session of exercise#

It indicates that the student is working on a specific exercise inside the Deeds simulator (Digital Circuit Simulator)

Deeds\_Es

The student is on Deeds simulator but it is not clear what exercise he is working on.

As we consider the 'exercise' feature from the moment the student study the content of an exercise to the moment he changes to another exercise, this can be estimated for assigning the number of exercise to Deeds as well.

TextEditor\_Es\_# session of exercise#

The student is writing the results of his work to submit later to the instructor. The students use a text editor (Word, Office, etc.) to answer to the questions and explain the solution they found through Deeds simulator  
TextEditor\_Es

It indicates that the student is working on an exercise in the text editor but it is not clear which exercise it is. This happens due to change of file names by the student, so we cannot recognize automatically which exercise he works on. Again, the suggestion given above on Deeds\_Es holds.

TextEditor

It shows that the student is using the text editor but not on exercises, this can contain other activities related to the text editor, for instance when they just open it.

When the students use Simulation Timing Diagram' to test the timing simulation of the logic networks, while using the Deeds simulator. It also contains these components: "Input Test Sequence" and "Timing Diagram View Manager ToolBar".

Deeds simulator, Simulation Timing diagram contain the properties window, which allows setting all the required parameters of the component under construction. For instance, the Properties can contain: "Switch Input", "Push-Button", "Clock properties", "Output properties", "textbox properties". Label all as 'Properties'. To understand if 'Properties' refer to Deeds simulator or SimulationTiming diagram.

#### Study\_Materials

The student is viewing some materials relevant to the course (provided by the instructor).

#### FSM\_Es\_# session of exercise#

When the student is working on a specific exercise on 'Finite State Machine Simulator' FSM\_Related

When the student is handling the components of Finite State Machine Simulator.

When the student is not viewing any pages described above, then we assigned 'Other' to the activity. This includes, for majority of cases, the student irrelevant activity

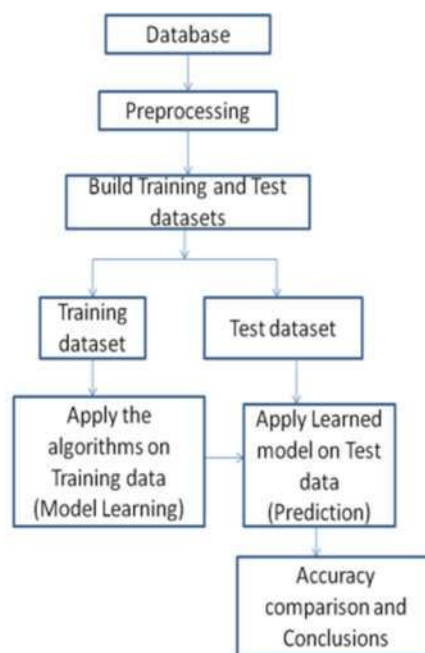


Fig 1. Architecture of prediction

### 3.1 Data Pre-Processing: -

Data preprocessing explain several type of processing achieve on unprocessed data to arrange it for a different processing procedure. Pre-processing of data is considered as a very important task in this work as quality and reliability of available information is needed which directly affects the results attained. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. Preprocessing is an important in machine learning before the classifying the data. It is very important to clean and prepare the DEED log

data using the preprocessing techniques. To extract the input variables those are more related to student difficulty. Because the performance of the models depends on the preprocessing methods.

### 3.2 Random Forest Algorithm

The Random Forest is one of the best machine learning models for predictive analytics, creation an industrial workhorse for machine learning. The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as: where the final model is the sum of simple base models. every derived classifier is a simple decision tree. A random forest is classifiers that consist of several decision trees and outputs the class that is the method of the classes result by individual trees. In essence, a random forest consists of a collection of unrepresentatively simple decision trees, each capable of producing a response when presented with a set of predictor values. A random forest has shown to run very efficiently on large datasets with large number of variables. This broad technique of using multiple models to obtain better predictive performance is called model assembling.

#### Disadvantage

Random forests have been experimental to fit for a number of datasets with noisy classification/regression tasks. For data counting definite variables with many number of levels, random forests are biased in help of individual attributes with additional levels. Therefore, the variable significance scores from random forest are not consistent for this type of data.

### 3.4 K-Nearest Neighbor

The most common extension to KNN is the use of standardized attribute for distance calculation. Since the attribute have greatly varying value ranges, using them directly in distance metrics would effectively give more weight to attribute with larger values. For example, even a dozen of *lesson attempts* would be irrelevant measure up to *gross speed* with value attainment up to hundreds. Z-score standardization is commonly used in data analysis to avoid such problems and to give all features roughly equal weight. KNN is the simplest of all machine learning algorithms. It has got a wide applications in different fields such that, pattern recognition, marketing of internet, analysis of Image databases cluster, etc. Sometimes it is helpful to avoid tied votes by choosing k to be an odd number. A single number k is given to determine the total number of neighbors used for classification. When k=1, then the nearest neighbors for a sample will determine its class. KNN require an integer k, a training data set and a metric to measure closeness

#### Disadvantages:

- It is Lazy learners
- It is susceptible to the limited structure of the data.

### 3.5 Support vector machine:-

Support Vector Machines (SVMs) belong to a family of generalized linear models which achieves a classification or regression decision based on the value of the linear combination of attribute. The mapping function in SVMs can be either a classification function or a regression function. For classification, nonlinear kernel functions are often used to transform the input data to a high dimensional feature space in which the input data becomes more separable compared to the original input space. Then, the maximum-margin hyperplanes are created to best possible divide the classes in the training data. Two parallel hyperplanes are created on every side of the hyperplane that divide the data by most of the distance between the two parallel hyperplanes. An assumption is prepared that the better the edge or distance among these parallel hyperplanes the enhanced the generalization error of the classifier

**Advantage**

- Flexibility in the selection of the structure of the threshold
- Robustness towards small number of data points

**4. EXPERIMENTAL RESULT EVALUATION METHODS**

<u>predictor</u>		<u>true</u>	<u>false</u>
	<u>true</u>	<u>True positive(TP)</u>	<u>False positive(FP)</u>
	<u>False</u>	<u>False negative(FN)</u>	<u>True negative(TN)</u>

**5. PERFORMANCE OF ALGORITHMS IN R STUDIO**

We performed multiple approach on our dataset and analyzed which attribute of the student's performance is more contributing towards each session used to analysis the fast learners and slow learners in multiple regression... We also performed various machine learning algorithms in R studio like Random forest , K Nearest neighbor Classifier , Support vector Machine and tabulated the accuracy in table, here it is obvious that , Support vector Machine give high accuracy compared to other methods. Also performances of algorithms depend on nature of the dataset.

**5.1 Confusion matrix:-**

Each instance is classified into two classes in a binary classification model. The two classes are true and false class. This gives rise to four possible classifications for each instance namely: True Positive (TP): The number of correct predictions that an object is positive. False Positive (FP): The number of incorrect predictions that an object is positive. False Negative (FN): The number of incorrect predictions that an object is negative. True Negative (TN): The number of correct predictions that an object is negative. This circumstances can be correspond to as a confusion matrix also called possibility.

The observed classifications for a phenomenon are compared with the predicted classifications of a model in a confusion matrix. In table 1 the classification that are shown along the major diagonal of the table are the correct classifications refereed as true positives and true negatives. The model errors are signified by the other fields. Only the true positive and true negative fields would be filled out for a perfect model and the other fields would be set to zero. From the confusion matrix, a number of model performance metrics can be derived. The most common metric is accuracy which is defined as the overall success rate of the classifier and is computed as

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

Other performance metrics include Sensitivity/Recall and Specificity/Precision.

Sensitivity is defined as percentage of correctly classified instances. Specificity is defined as percentage of incorrectly classified instances.

$$\text{Sensitivity/Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Specificity/Precision} = \text{TN} / (\text{TN} + \text{FP})$$

## 5.2 Cross-validation set:-

In the first set of experiments, we used the original dataset was composed in all records. Based on the 10-fold cross-validation, the support vector machines produced the best results with an overall prediction rate of 87.23% experiments are conducted to assess the predictive ability of the three ensemble models. Based on the 10-fold cross validation methodology, the information fusion type ensemble model produced the best results with an overall prediction rate of 87%, followed by the bagging type ensembles and boosting type ensembles with overall prediction rates (sensitivity and specificity) of 89.5% and 85% respectively. Even though the prediction results are slightly better than the individual models, ensembles are known to produce more robust prediction systems compared to a single-best prediction model.

<u>Classifier</u>	<u>Accuracy</u>	<u>Specificity</u>	<u>Sensitivity</u>
<u>Random forest</u>	<u>89%</u>	<u>87.5%</u>	<u>89.5%</u>
<u>K-NN</u>	<u>85%</u>	<u>85%</u>	<u>86%</u>
<u>SVM</u>	<u>95.5%</u>	<u>94%</u>	<u>95%</u>

## 6. CONCLUSION

Machine learning concentrate on the growth of computer programs that can observed data and use it study for machine themselves. The achievement of machine learning for students prediction system based on using better machine learning algorithms, choose the right algorithm for the problem is very significant to accomplish the best results. The results get hold of here confirms that the pattern does be present. even though a many number factors manipulate the final result of a student, the high accuracy obtained by the machine learning models .the problem of predicting student performance in an DEEDs lab using TEL system .The reliable predictions of the performances of the students, the teachers can notified their efforts on those students and problem that need concentration the majority. The SVM method for predicting whether students will complete in all session the results demonstrate that SVM can produce such predictions accurately, particularly for the grade. An experimentally calculate a large amount of extension and feature sets, and found out that differences in predictor performances and constraint values. Finally the Support Vector Machine gives the High accuracy 95.5%.

## REFERENCES

- [1] Dursun Delen , "A comparative analysis of machine learning techniques for student retention management" Elsevier doi:10.1016/j.dss.2010.06.003.
- [2] Murphy PM, Aha DW (1995) UCI repository of machine learning databases, (Machine Readable Data Repository). Dept. Inf. Comput. Sci., Univ. California, Irvine, CA
- [3] L. J. Elkan, S. A. (1976) „The distance-weighted k-nearest neighbour rule“, IEEE Transactions on Systems, Man and Cybernetics, vol. 6, no 4, pp. 325–327.
- [4] Cover, T. and Hart, P. (1967) „Nearest neighbour pattern classification“, IEEE Transactions on Information Theory, vol. 13, issue 1, pp 21–27.
- [5] Tuomas Tanner and Hannu Toivonen,, “Predicting and preventing student failure using the k-nearest neighbour method to predict student performance in an online course environment”,.
- [6] Ryan S.J.D baker and Kalina yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, Article 1, Vol 1, No 1, Fall 2009
- [7] K V Krishna Kishore, Venkatramaphanikumar S, Sura Alekhya, "Prediction of student academic progression: A case study on Vignan University," International Conference on Computer Communication and Informatics, 2014, pp. 1 - 6
- [8] K. P. Shaleena, Shaiju Paul, "Data mining techniques for predicting student performance," IEEE International Conference on Engineering and Technology (ICETECH), 2015, pp. 1 – 3



- [9] Mushtaq Hussain, Wenhao Zhu<sup>1</sup> and Wu Zhang<sup>1</sup> ,”Using machine learning to predict student difficulties from learning session data”, School of Computer Engineering and Science, Shanghai University, Shanghai, China, 10th February 2018.
- [10] Ge X, Liu J, Qi Q, Chen Z (2011) A new prediction approach based on linear regression for collaborative filtering. In: 8th International 2011 conference on fuzzy systems and knowledgediscovery(FSKD),pp 2586–2590. <https://doi.org/10.1109/FSKD.2011.6020007>
- [11] Acharya A, Sinha D (2014) Early prediction of students performance using machine learning techniques. *Int J Comput Appl* 107(1):37–43. <https://doi.org/10.5120/18717-9939>.
- [12] Donzellini G, Ponta D (2007) A simulation environment for e-learning in digital design. IEEE *TransIndElectron* 54(6):3078–3085. <https://doi.org/10.1109/TIE.2007.907011>.
- [13] Elbadrawy A, Studham RS, Karypis G (2015) Collaborative multi-regression models for predicting students’ performance in course activities. In: 5th International conference on learning analytics and knowledge (LAK ’15), pp 103– 107. <https://doi.org/10.1145/2723576.2723590>.
- [14] Fernandez-Delgado M, Mucientes M, Vazquez-Barreiros B, Lama M (2014) Learning analytics for the prediction of the educational objectives achievement. In: 44th IEEE Frontiers in Eeducation conference (FIE), pp 2500–2503. <https://doi.org/10.1109/FIE.2014.7044402>..

## TEACHERS' WILLINGNESS TO USE INSTRUCTIONAL TECHNOLOGY FOR TEACHING AND LEARNING

கற்றல் கற்பித்தலில் தொழில்நுட்ப அறிவுறுத்தலைப்  
பயன்படுத்துதலில் ஆசிரியர்களின் முனைப்பு

Tasaratha Rajan Anamalai<sup>1</sup>, Maizatul Hayati Mohamad Yatim<sup>2</sup>

<sup>1</sup> Sultan Idris University of Education, Malaysia; tasara83@gmail.com

<sup>2</sup> Sultan Idris University of Education, Malaysia; maizatul@fskik.upsi.edu.my;

### ABSTRACT

This study examined primary school teachers' willingness to use instructional technology for teaching and learning (T&L). This study adopted the quantitative approach and used a questionnaire as a research instrument. The sample consisted of 180 primary school teachers in district X. Study's variables included perception of teachers' readiness, knowledge of teachers' readiness, skills linked to teachers' readiness, and challenges and issues. The analysis results showed that all variables obtained moderate mean scores, specifically, perceptions of teachers' readiness ( $M = 2.87$ ,  $SD = .455$ ); knowledge of teachers' readiness ( $M = 2.86$ ,  $SD = .600$ ); skills linked to teachers' readiness ( $M = 2.61$ ,  $SD = .554$ ); and challenges and issues ( $M = 2.87$ ,  $SD = .460$ ). In conclusion, the results imply that primary school teachers in district X have moderate readiness to use instructional technology in T&L. Hence, school administrators and the Ministry of Education need proactive efforts to increase teachers' interest.

Copyright © 2022 International Forum for Information Technology in Tamil.  
All rights reserved.

### Keywords:

Teachers' readiness,  
instructional technology,  
perspective,  
knowledge,  
skills,  
challenge

### Corresponding Author:

Tasaratha Rajan Anamalai,  
Faculty of art, computing creative industry,  
Sultan Idris University of Education, Malaysia,  
Email: tasara83@gmail.com

## 1. INTRODUCTION

Today's generation has witnessed significant and rapid information and communication technology development (Fariduddin, Azidah, and Aziah, 2019). This situation has also impacted education development regarding teaching and learning (T&L) approaches, techniques, and methods in Malaysia. Various new methods and techniques have been introduced in the education field to facilitate faster, innovative, and seamless communication and interaction to transcend the limitations of time and place (Ramya and Poongodi, 2022). In line with efforts to improve the quality of education, the Malaysian National Education Policy (NEP) is constantly reviewed to ensure the effectiveness of curriculum implementation to prepare Malaysians to face the educational, economic, political, and social challenges in this highly globalised world (Abdul Halim et al. 2020).

Parallel to the changes implemented by the Malaysian Ministry of Education changes in the NEP, the Malaysian Education Development Plan 2013-2025 or PPPM (2013-2025) has been outlined to restructure the

Malaysian education system, from preschool, primary and secondary levels (Ministry of Education Malaysia, 2018a ). The Malaysian Higher Education Education Development Plan 2015-2025 or PPPM(PT) (2015-2025) was also established to discuss the issues and challenges of 21st-century education in tertiary education (Ministry of Education Malaysia, 2018 b).

Instructional technology has created a significant technological ground for teaching. Using instructional technology in the education system can encourage interaction between students and instructors to improve the efficiency of T&L (Sophonhiranrak, 2021) . According to Lim and Lee (2021) , the younger generation in Malaysia is adept at sharing knowledge with friends and obtaining information using the latest instructional technology devices, such as personal computers, projectors, tablets, netbooks, smartphones, and laptops (Omar and Ismail, 2020) .

Instructional technology is a new concept implemented in teaching and learning due to easy access to technological devices. Implementing instructional technology is similar to electronic learning or e-learning in the classroom. However, instructional technology constitutes a learning process that can occur outside the classroom anywhere and at any time (Umi et al., 2019) . According to Ching (2018) , instructional technology includes machine learning that allows students to access learning materials anywhere and anytime through a computer system or mobile devices with Internet access.

Furthermore, Instructional technology is medium that can improve students' thinking skills and knowledge. According to Halizayanie (2019) , using instructional technology in education can help students enrich their learning experience by stimulating various problem-solving skills, 21st-century skills, and higher-order thinking skills (HOTS). Teachers need to apply effective pedagogy in the classroom. They need to update themselves with the latest knowledge and use teaching techniques suitable for the current generation to ensure students can master the standards the Ministry of Education outlined. As Azizah and Suziyani (2021) mentioned, delivering effective educational knowledge in T&L will result in quality human capital. Teachers also need to diversify creative teaching methods, and interesting strategies must align with current developments to achieve the teaching and facilitation objectives. This study generally focuses on primary school teachers' willingness to use instructional technology in teaching and learning.

## 2. PROBLEM STATEMENT

According to Syaza, Fauzi, and Faridah (2018), wireless mobile devices allow the learning process to occur informally regardless of time and location limitations, even without a teacher's or instructor's presence.

A study by Ching (2018) strongly emphasized the inclusion of entertainment elements in the KSSR Malay language teaching and facilitation process can create a relaxed and fun learning atmosphere that can increase students' interest in learning. Moreover, Noor Desiro and Husnin (2021) found that the knowledge of rural secondary school teachers about Google Classroom is low even though their readiness to use Google Classroom in Teaching and Facilitation (T&L) is high.

Studies found that the concept of instructional technology can increase students' motivation as its teaching mechanisms can increase interactivity and promote understanding (Handani, Suyanto, and Sofyan, 2016) . Thus, the concept of instructional technology attracts interest and motivates students' creativity and innovation (Halizayanie, 2019) .

### *Research objective*

To study the problems discussed, the research objectives are as follows:

- i Identifying teachers' perceptions of instructional technology in teaching and learning.
- ii Identifying teachers' knowledge of instructional technology in teaching and learning.

- iii Identifying teachers' skills in using instructional technology in teaching and learning.
- iv Identifying the challenges teachers face in implementing instructional technology in teaching and learning.

### 3. METHODOLOGY

This study followed the quantitative research design. A questionnaire was used as the main instrument to collect data to answer the research questions in district X. The questionnaire instrument contains two parts, Part A and Part B. Part A contains questions related to the respondent's demographics, specifically gender, age, level of education, work experience, and type of school. Part B contains questions about Primary School teachers' readiness to use instructional technology in teaching and learning. A five-point Likert scale was used for each item in Part B.

The demographic data were reported through frequency, percentage, and mean, while the hypothesis testing was conducted using Statistical Package for Social Science, Version 12.0 (SDSS). The analysis results are presented through tables and diagrams to give the researcher a deeper understanding of the research topic.

In line with Kamarul Azmi Jasmi's study (2009), this study used percentage and mean values to report the findings on the perception, knowledge, and skills levels. The levels ranged from or low (1.01-2.00), moderately low (2.01-3.00), moderately high (3.01-4.00), and high (4.01-5.00) based on the mean values obtained.

The researcher used a simple random sampling method to determine that the selected respondents represent the study population, specifically the primary school teachers. The researcher distributed the questionnaires to all district X, primary schools, and 180 completed responses were returned.

The questionnaire was divided into two parts to capture teachers' perceptions, knowledge, skills, and challenges in using instructional technology in teaching and learning

Part I : Respondents' Demographics

Part II : Teacher's readiness to use instructional technology in teaching and learning

- i. Teachers' perceptions of the use of instructional technology in teaching and learning
- ii. Teachers' knowledge of using instructional technology in teaching and learning
- iii. Teachers' skills in using instructional technology in teaching and learning
- iv. Challenges and issues faced by teachers in using instructional technology in teaching and learning

### 4. STUDY FINDINGS AND DISCUSSION

This section answers the research question, 'What is the level of primary school teachers' willingness to use instructional technology in teaching and learning'? The analysis of the teacher's readiness involves several variables: perception, knowledge, and skills in using information technology in T&L in district X.

#### *Respondents' Demographics*

This section explains the results on the respondents' demographic backgrounds- gender, age, level of education, and work experience of current teachers. The analysis results are presented in Table 1.

**Table 1. Respondents 'Demographics**

No	Demography		Frequency	Percentage(%)
1	Gender	Men	50	28.0
		Female	130	72.0
		<b>Total</b>	<b>180</b>	<b>100.0</b>
2	Age	20 - 30 years	10	5.6
		31 – 40 years old	69	38.3
		41 – 50 years old	62	34.4
		51 - 60 years old	39	21.7
		<b>Total</b>	<b>180</b>	<b>100.0</b>
3	Educational status	Diploma	20	11.1
		Bachelor	138	76.7
		Masters	21	11.7
		Doctor of Philosophy	1	.5
		<b>Total</b>	<b>180</b>	<b>100.0</b>
4	Work experience	Less than 5 years	9	5.0
		5 – 10 years	32	17.8
		11 – 20 years	75	41.7
		21 – 30 years old	45	25.0
		Over 30 years	19	10.5
		<b>Total</b>	<b>180</b>	<b>100.0</b>

Table 1 shows the respondents' demographic backgrounds. 50 (28.0%) respondents are males and 130 respondents (72.0%) are females. In terms of age, a total of 10 respondents (5.6 %) are aged between 20 to 30, 69 (38.3 %) are aged between 31 and 40 years, 62 (34.4 %) are aged between 41 and 50, and 39 (21.7 %) are aged between 51 and 60. As for the level of education, a total of 20 respondents (11.1%) obtained a diploma, 138 respondents (76.7%) graduated with a bachelor's degree, 21 respondents (11.7%) obtained a master's degree, and one (0.5%) graduated with a doctor of philosophy. Concerning their work experience, nine respondents (5.0%) have less than five years of work experience, 32 respondents (17.8%) with five to 10 years, 75 respondents (41.7%) with 11 to 20 years, 45 respondents (25%) have been working for 21 to 30 years and 19 respondents (10.5%) have worked for over 30 years.

### *Mean Analysis of Teachers' Perceptions*

Table 2 shows the mean level indicator as mentioned in Landell (1977).

**Table 2. Landell Mean Level Indicators (1977)**

Mean Value	Level Indicator
1.00 – 2.33	Low
2.34 – 3.67	Simple
3.68 – 5.00	Height

The results of the mean analysis of teachers' perceptions of the use of instructional technology in teaching and learning are shown in Table 3.

**Table 3. Mean Analysis of Perceptions of Teacher Readiness**

No	Item	Mean	Standard deviation
1	The use of instructional technology in T&L is more interesting	3.21	.610
2	The use of instructional technology in T&L can improve student achievement	3.12	.588
3	The use of instructional technology facilitates student understanding	3.14	.617
4	The use of instructional technology in T&L can increase students' interest in learning	3.24	.613
5	Teachers need to be prepared to explore aspects of using instructional technology in T&L	3.15	.642
6	Teachers need to be positive about the use of instructional technology in T&L	3.19	.615
7	The content of the instructional technology needs to be diversified in the T&L process	3.17	.615
8	The use of instructional technology in T&L is not effective for students	2.18	.877
9	The use of instructional technology in T&L is a waste of time	2.13	.877
10	The use of instructional technology in T&L causes students to lose focus	2.18	.838
Overall Mean		2.87	.455

Table 3 shows that teachers' perception of the use of instructional technology in teaching and learning is at a moderate level with an overall mean value of Mean=2.87, SD=.455. The first item on the perception of teachers' readiness, Most items have a moderate mean score; item 1, "The use of instructional technology in T&L is more interesting," (M=3.21, SD=.610), item 2, "The use of instructional technology in T&L can increase student achievement" (M=3.12, SD=.588), item 3 "The use of instructional technology facilitates student understanding" (M=3.14, SD=.617), item 4 "The use of instructional technology in T&L can increase students' interest in learning," (M=3.24, SD=.613), item 5, "Teachers need to be prepared to explore aspects of the use of instructional technology in T&L" is at a moderate level (M=3.15, SD=.642), item 6, "Teachers need to be positive about the use of instructional technology in T&L" is at a moderate level (M=3.19, SD=.615, and item 7 "The content instructional technology needs to be diversified in the T&L process " (M=3.17, SD=.615. Three items have lower mean scores, item 8, "The use of instructional technology in T&L is not effective for students" (M=2.18, SD=.877), and item 9, "Use instructional technology in T&L wastes time (M=2.13, SD=.877), and item 10 "The use of instructional technology in T&L causes students not to focus" (M=2.18, SD=.838).

### **Knowledge Mean Analysis**

The mean analysis results on teachers' knowledge in using instructional technology in teaching and learning are shown in Table 4.

**Table 4. Mean Analysis of Knowledge of Teacher Readiness**

No	Item	Min	Standard deviation
1	I know how to use instructional technology in T&L	2.63	.727
2	I know that instructional technology makes T&L more creative	2.94	.695
3	I know that instructional technology helps in engaging students	3.08	.628
4	I know how to get teaching aids online	2.86	.719
5	I know how to download instructional applications	2.85	.684
6	I know instructional applications are easily attainable/accessible	2.80	.732
Overall Mean		2.86	.600

Table 4 shows that teachers' knowledge of using instructional technology in teaching and learning is at a moderate level, with an overall mean value of  $M=2.86$  and  $SD=.600$ . All items show moderate mean scores, item 1 "I know using instructional technology in T&L" ( $M=2.63$ ,  $SD=.727$ ), item 2 "I know instructional technology makes T&L more creative" ( $M=2.94$ ,  $SD=.695$ ), item 3 "I know that instructional technology helps in attracting students' interest" ( $M=3.08$ ,  $SD=.628$ ), item 4 "I know how to get teaching materials through online" ( $M=2.86$ ,  $SD=.719$ ), item 5 "I know how to download instructional applications" ( $M=2.85$ ,  $SD=.684$ ) and item 6 "I know that instructional applications can be reached/accessed easily" ( $M=2.80$ ,  $SD=.732$ ).

#### *Mean Skill Analysis*

The mean analysis results for teachers' skills in using instructional technology in teaching and learning are shown in Table 5.

**Table 5. Mean Analysis of Skills in Teacher Readiness**

No	Item	Min	Deviation Standard
1	I am proficient in using instructional technology in T&L	2.46	.714
2	I can conduct classes using instructional technology related to certain topics during T&L	2.63	.703
3	I can conduct a class using instructional technology during T&L	2.61	.690
4	I can't use the instructional technology app in T&L	2.20	.812
5	I can find T&L-related information in the instructional technology	2.76	.691
6	I'm good at using the mobile devices	2.80	.698
7	I am good at using a laptop/computer	2.89	.651
8	I am proficient in using instructional applications for T&L	2.51	.723
Overall Mean		2.61	.554

Table 5 shows the teacher's skill level in using instructional technology in teaching and learning is moderate, with an overall mean value of  $M=2.61$  and  $SD=.554$ . All items also obtained moderate mean scores, item 1 "I am good at using instructional technology applications in T&L" ( $M=2.46$ ,  $SD=.714$ ), item 2 "I can handle the class using instructional technology related to certain topics during T&L" ( $M=2.63$ ,  $SD=.703$ ), item 3 "I can handle the class using instructional technology during T&L" ( $M=2.61$ ,  $SD=.690$ ), item

4 “I cannot use instructional technology applications in T&L” (M=2.20, SD=.812), item 5 “I can find information related to T&L in instructional applications” (M=2.76, SD=.691), item 6 “I am good at using the mobile devices” (M=2.80, SD=.698), item 7 “I am good at using a laptop/computer” (M=2.89, SD=.651) and item 8 “I am good at using instructional applications for T&L” (M=2.51, SD=.723).

### ***Mean Analysis of Challenges and Issues***

The mean analysis results on the challenges and issues teachers faced in using instructional technology in teaching and learning are shown in Table 6.

**Table 6. Mean Analysis of Challenges and Issues**

No	Item	Min	Standard deviation
1	Teachers lack technical training and courses in instructional technology in teaching and learning	3.07	.718
2	Teachers lack courses on instructional technology in teaching and learning	3.10	.711
3	I lack awareness of the benefits of T&L in using instructional technology in teaching and learning	2.55	.788
4	I have time constraints to prepare instructional technology material	2.95	.699
5	Teachers do not have internet access facilities at school	3.04	.780
6	I am not ready to use instructional technology in teaching and learning	2.92	.696
7	Instructional technology devices facilities for teaching and learning are insufficient	2.33	.855
8	The facilities provided such as netbooks & computers for T&L provided by the school	2.97	.712
Overall Mean		2.87	.460

Table 6 shows that teachers’ challenges and issues in using instructional technology are moderate, with an overall mean of M=2.87 and SD=.460. All items also show a moderate mean score; item 1 “Teachers lack technical training and courses in instructional technology in teaching and learning” (M=3.0, SD= .718), item 2 “Teachers lack courses on instructional technology in teaching and learning” (M=3.10, SD=.711), item 3 “I lack awareness of the benefits of T&L in using instructional technology in teaching and learning” (M=2.55, SD=.788), item 4 “I have time constraints to provide instructional technology materials” (M=2.95, SD=.699), item 5 “Teachers do not have access to the internet at school” (M=3.04, SD=.780), item 6 “I am not ready to use instructional technology in teaching and learning” (M=2.92, SD=.696), item 7 “Instructional technology devices facilities for teaching and learning are insufficient” (M=2.33, SD=.855) and item 8 “The facilities provided such as netbooks & computers for T&L provided by the school” (M=2.97, SD=.712).

## **5. CONCLUSION**

The data were analysed to determine primary school teachers’ readiness to use instructional technology in the T&L in district X. The data analysis showed that teachers in district X moderately use devices and facilities provided by the school, such as netbooks & computers for T&L . All constructs obtained moderate mean scores- perception of teacher readiness (M=2.87, SD=.455), knowledge of teacher readiness (M=2.86,



SD=.600), skills linked to teachers' readiness ( $M=2.61$ ,  $SD=.554$ ) and challenges and issues ( $M=2.87$ ,  $SD=.460$ ). This finding shows moderate teachers' willingness to use instructional technology in T&L in primary schools in district X. Hence, proactive efforts from administrators and the education ministry are necessary to increase teachers' interest in using this technology in T&L. In conclusion, as the level of readiness of primary school teachers to use instructional technology for T&L in the district X is still moderate, it needs to be strengthened, specifically in the aspect of teachers' knowledge and skills and the convenience of using instructional technology in schools. Teachers need to equip themselves with information technology knowledge and skills and be ready to use instructional technology to increase students' achievements. Therefore, to ensure that instructional technology can be used optimally in all primary schools in the district X, all parties need to play their respective roles in increasing the availability of information technology facilities in schools, in addition to improving the teachers' skills related to the use of simple technology through continuous training.

## REFERENCES

- Abdul Halim Abdullah, Hayati Nazirwan, Kumaresan Pumalai, and Nor' Hidayah Mohd Amin. 2020. "PISA Assessment: Where Does Malaysia Rank for Mathematical Literacy Among Southeast Asian Countries?" *Science Magazine* (June).
- Ching, Melvina Chung Hui. 2018. "Development And Evaluation Of Mobile Educational Software 'M-Education' For Learning Malay In Primary Schools."
- Halizayanie Binti Kimlin, Siti Izani Binti Idris. 2019. "Crea8tif: Digital Art Project Management Smart Application." *Journal on Technical and Vocational Education (JTVE), Vol 4 No 3: Special Edition NASCO (2019)* 4(3).
- Handani, Sitaesmi Wahyu, M. Suyanto, and Amir Fatah Sofyan. 2016. "Application of Gamification Concept to E-Learning for 3-D Animation Learning." *Telematics* 9(1):42–53.
- Lim, Thian Li, and Angela Siew Hoong Lee. 2021. "Extended TAM and TTF Model: A Framework for the 21st Century Teaching and Learning." in *Proceedings - International Conference on Computer and Information Sciences: Sustaining Tomorrow with Digital Innovation, ICCOINS 2021*.
- Malaysia Education Ministry. 2018. "PPPM 2018 Annual Report 2013-2025." *Cereel For* 51(1):51.
- Muhammad Fariduddin Wajdi Anthony, Azidah Abu Ziden, and Aziah Ismail. 2019. "Fuzzy Delphi Technique: Design And Development Of GeSVa Acceptance Model In m-Learning Teacher Education Institute." *Journal of Educational Research & Indigenous Studies* 2(1).
- Noor Azizah Mohd Said, and Suziyani Mohamed. 2021. "Relationship between Teacher Personality and Child Moral Formation (Relationship between Teacher Personality and Child Moral Formation)." *World Journal of Education* 3(1).
- Noor Desiro Saidin, and Hazrati Husnin. 2021. "Google Classroom as an M-Learning Platform : Level of Knowledge and Level of Preparedness of Rural Middle School Teachers." *World Journal of Education* 3(2).
- Omar, Mohd Norakmar, and Siti Noor Ismail. 2020. "Mobile Technology Integration in the 2020s: The Impact of Technology Leadership in the Malaysian Context." *Universal Journal of Educational Research* 8(5).
- Ramya, D., and OT Poongodi. 2022. "A Study on the Usage of Information Communication Technology Tools in the Teaching - Learning Process of Engineering Education." *Journal of Applied Science and Engineering (Taiwan)* 25(2).
- Sarah Alia Mohamed Faisal, and Nor Hafizah Adnan. 2021. "Teachers' Level of Readiness and Acceptance in Practicing the Use of RI4.0 Digital Technology as Teaching Aids in Primary Education." *International Journal of Advanced Research in Islamic Studies and Education* 1(3).
- Sophonhiranrak, Samoeakan. 2021. "Features, Barriers, and Influencing Factors of Mobile Learning in Higher Education: A Systematic Review." *Helion* 7(4).
- Syaza HAZwani, Zaini, Mohd Ayub Ahmad Fauzi, and Yahya Faridah HAnim. 2018. *Elements of Innovation Diffusion Theory in the Acceptance and Use of M-Learning Among Students*.
- Umi Azmah Hasran, Nur Fadhilah Abdul Jalil, Wan Ramli Wan Daud, Rosseni Din, and Siti Fadhilah Mat Noor. 2019. *Multidisciplinary Technology Education: Introducing Fuel Cell with Game-Based Learning (Multidisciplinary Technology Education: Introducing Fuel Cell with Game-Based Learning)*. Vol. 11.

## Efficient Technique for Corpus Creation of Code-Mixed Data on Tamil and English Text from Social Forum

M.Sangeetha<sup>1</sup> Dr.K.Nimala<sup>2</sup>

<sup>1</sup>Research Scholar

<sup>2</sup>Assistant Professor

Computer Science and Engineering  
SRM Institute of Science and Technology,  
Kattankulathur, Tamilnadu, India

[nigeetha2005@gmail.com](mailto:nigeetha2005@gmail.com),

[nimalak@srmist.edu.in](mailto:nimalak@srmist.edu.in)

---

### ABSTRACT (10 PT)

NLP refers to "natural language processing," a subfield of AI that entails teaching computers to interpret, analyze, and glean relevant information from human language. There are a number of scenarios in which using a text's sentiment analysis would be helpful. One such use is analyzing the common attitudes expressed in online discussions. In contrast, many social media comments are written in scripts that aren't the author's native language and don't adhere to standard English grammar. For a language like Tamil, which has minimal resources available, the lack of annotated code-mixed data compounds the difficulty of this issue. To combat this, we developed a code-mixed, sentiment-annotated Tamil-English corpus. In this article, we'll talk about how the corpus was assembled and how the polarities were assigned. The level of agreement between annotators and the outcomes of trained sentiment analyses using this corpus.

---

### Keywords:

A Sentiment Analysis  
B Tamil-English Code-mix  
C Natural Language Processing  
D Corpus  
E Grammar rule

---

### Corresponding Author:

Dr.K.Nimala  
Assistant Professor  
Computer Science and Engineering  
SRM Institute of Science and Technology,  
Kattankulathur, Tamilnadu, India.  
E-mail: [nimalak@srmist.edu.in](mailto:nimalak@srmist.edu.in)

---

## 1. INTRODUCTION

Blog entries, reviews, and comments on social media sites like Facebook, Twitter, YouTube, etc. provide a forum for individuals to share their perspectives on any issue[1] claim that the grammar of any language can be used by anyone sees fit, hence multilingual individuals often communicate their views on a topic in more than one language[2]. Combining linguistic elements from different languages into a new text is known as "code-mixing"[3].

As the Internet has expanded rapidly and become increasingly popular, data volume has grown substantially [4]. Because of the huge network of texts that includes mixed content like text, audio, and video in numerous languages, humans are also compelled to deal with complexity.

In order to get the most out of a dataset, SA should be utilized to identify, examine, and extract the respondents' subjective attitudes and emotions [5][6]. SA is a branch of natural language processing that goes by a variety of other names, including opinion mining, orientation analysis, sentiment classification, and subjective analysis[7]. Typical examples of NLP difficulties encountered in SA tasks include entity recognition, word polarity disambiguation, and aspect extraction. A strong association exists between the difficulty of a SA task and the problems that end users encounter when using their particular applications.

### A. Code-Mixing and its types

Multilingual populations are more likely to engage in code-mixing. English words, phrases, and connectives are frequently used in everyday spoken Tamil due to its prominence as the language of education and higher learning. This is most noticeable in those with less education or training [8], but it also occurs in those with higher degrees of education or training. Many Tamil-English (Tanglish) lines are written in Roman style [9] because of English's widespread use on the internet. Code-mixing examples from Tamil corpora, with English translations, are provided in Table.1.

Table 1. Code mixing examples in a Tamil dataset

Changing Code Type	Example	Translation
Avoid Using Other Languages in Place of Tamil (Written in Tamil Script)	எந்தப் பிரச்சனை என்றாலும், எத்தனை பேர் கருத்து சொல்லி முடித்திருந்தாலும் நம் கருத்தைக் பகடதேற்கும் சிலர் ஆர்வமாக இருப்போர்கள்	No matter the issue, no matter how many people have already commented, some people will be interested in hearing our opinion
Combining English and Tamil in the same sentence (Tamil written only in Tamil script)	தமிழை என்னை பிறக்க வைத்த என தாயிற்கு நன்றி... இப்பேடி ஒரு டோடனலயிற்பிரிய இனம் அனமப்போளருக்கு நன்றி... டேலாசிரியர்க்கு நன்றி...i dont watch these kind of emotional songs...just listened it and im loving it...	Thanks to my mother who gave birth to me as a Tamilian...Thanks to the music director who composed such a song...Thanks to the film writer...I don't watch these kinds of emotional songs...just listened to it and I'm loving it...
No other language than Tamil (Written in Latin script)	Ellarume anbukaga enginal yar than anbu seluthuvathu?...neenga elarmelaiyum anbu seluthunga .....	Does everyone yearn for love Who will give his love ?..You love everyone...
Combination of inter- and intra-sentential discourse (Tamil written in both Tamil and Latin script)	உண்மை உய கா இந்தேடத்திலுந்நடிகத்திருவர ுபமமிகசேரிய உயரத்தன அனடயதகுதி உடையவர்கள் ..சினிமாமட்டும்திவிதிலக்காஎன்னை ..thaguthi udaiyorai thangi pidikkayarumillai	True brother both actors in this movie deserve to reach great heights ..Cinema alone is no exception ..There is no one to hold the meritorious wearer
Language Switching Between English and Tamil Within the Same Sentence (written in Latin script only)	Vaa endru solummunney varungindra nyabagam....kannileyilaye kadha lum nenjamey kadhalin thaayagam"... soulful lyrics with heavenly cinematography and music...	The memory that comes before you say come The leaf in the eye is love and the heart is the home of love"... Lyrics with heavenly cinematography and music in Soulful...
Switching between Morphologically Distinct Codes (Written in both Tamil and Latin script)	ஒ இந்த டோட்டுக்கு இப்பேடித்தான் viewerssekurengla	Oh, this is how you gather more viewers for this song

When examining code-mixed sentences, we made a distinction between intra-sentential switching, tag-switching, and inter-sentential switching. Almost all of the feedback was written in Tamil script but used English grammar or a combination of Tamil grammar and English lexicons. The comment contained both Tamil characters and some English words. The principle is explained by these instances.

நீங்கள் வாழ்க்கையைக் கற்க விரும்பினால், படிப்பில் மட்டுமல்ல, உங்கள் சகாக்களிடமிருந்தும் கற்றுக்கொள்ள வேண்டும். அதற்கு வகுப்பறையே சிறந்த இடம் - ஆங்கில மொழி பெயர்ப்பு விருப்பம் இல்லாமல் தமிழ் எழுத்தில் எழுதப்பட்ட வார்த்தைகள்.

- If you want to learn life, you should learn from your peers and not just in the course. The classroom is the best place for that-
- Words written in Tamil script, without any English translation option..
- Thalapathy dance pathi comment podunga.....
- Leave a comment on the Thalapathy dance .. -Tag switching with English words.
- What a surprise...Nothing can beat, enna oru energy of vaathi coming pattu..
- What a surprise...nothing can beat, what an energy of Vaathi coming song-Inter-sentential switch
- communicate mattum share people any message on Social media uthavukirathu
- Communicate and share people any message with social media. Intra-sentential switch between clauses.

The lack of annotated data for code-mixed situations motivated us to focus on sentiment analysis in Tamil [10] [11]. However, due to the mixing of languages at different stages of linguistic analysis [12], a code-mixed usage is not supported by an annotated corpus designed for monolingual data.

#### B. Sentiment Polarity of Tamil Reviews

Positive:

Tamil: எல்லாரும் தனலகுனிந்து நடப்பதேற்குக்காரணம்,

ஸ்மார்ட் போனும் ப்ளாஷியல் மீடியாக்களும்தான்

English: Smartphones and social media are the reason everyone is walking with their heads down

Tamil: இன்டர்சநட் இன்னைசயன்றால்,

சிலருக்குப் பவறு எந்தப் பவனையும் செய்யப் பவமுடியாது.

English: Without the internet, some people would not be able to do any other work.

Neutral

Tamil: புதியேனடப்புகள எதிர்த்து ஒரு கூட்டம் காத்திருக்கிறது. நண்பர்

English: A crowd is waiting for new creations. Friend.

#### C. General frameworks for Corpus creations

A summary of the procedures used to create our dataset is shown in Figure 1. The YouTube Comment Scraper programme was used to collect the comments. These remarks were used by us to create datasets for sentiment analysis with manual annotations. With enough representation for each feeling, we aimed to gather comments that feature code-mixing throughout the text. It was more challenging to glean the necessary and appropriate text from the comment area due to the presence of comments written in languages other than the target language.

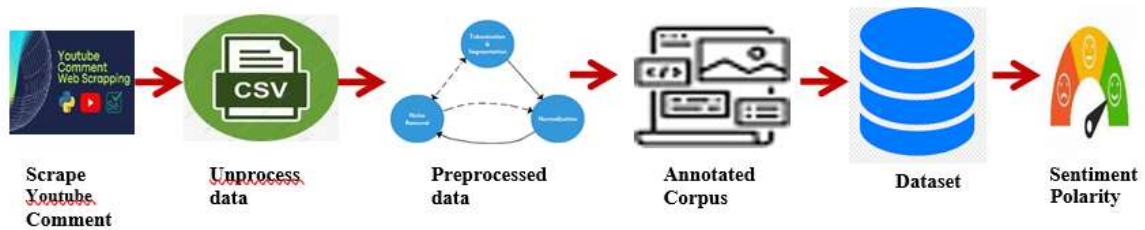


Figure 1. Dataset Preparation

*Title of manuscript is short and clear, implies research results (First Author)*

## 2. RESEARCH METHOD

The problem of sentiment analysis has been tackled in a number of ways, the most well-established being lexicon-based methods and machine-learning techniques [14]. WordNet-Affect [15], SentiNet, and SentiWordNet [16] were the most extensively used new lexicons since the field of sentiment analysis began utilizing them in 1966. Despite their famed ease of use, standard machine learning and lexicon-based algorithms fail to deliver when applied to user-generated data because of the data's inherent volatility. Here, deep learning approaches excel because of their ability to effectively adapt to the ever-changing nature of user-generated data. There are a few different transfer learning methods available, each with its own set of benefits and drawbacks. Three such programs are GloVe [17], Word2Vec[18], and fastText [19].

Many people of Tamil descent in Singapore, India, Sri Lanka, and elsewhere in the diaspora speak Tamil, a Dravidian language[20][21]. Tamil [24] and other Indian languages have large native-speaker populations, making them a promising market for commercial NLP applications. Consequently, there have been several studies on sentiment analysis in these languages [23][24].

Code-mixed YouTube comments in Tamil and English have been annotated [26]. The 15,744 comments in the corpus were annotated using at least three distinct coding systems, and the corpus as a whole was annotated by a total of 59 people. Common ML algorithms were used, such as the Support Vector Machine (SVM), Decision Trees (DT), Multinomial Naive Bayes (MNB), Logistic

Regression (LR), k-Nearest Neighbor (kNN), and Random Forest (RF), with features such as Term-Frequency-Inverse Document- Frequency (TF-IDF) of word n-grams from  $n = 1$  to 5000. (1, 3).

Two Deep Learning (DL) models, the 1D Convolutional Long Short-Term Memory (1DConvLSTM) and the LSTM have also been implemented; the former uses the Keras embedding, while the latter employs the Dynamic Meta Embedding (DME). The authors have developed a transformer-based classifier for SA of Tamil-English code-mixed text. The classifier makes use of multilingual Bidirectional Encoder Representations (mBERT). [22] released Twitter and blog datasets for code-mixed Telugu-English content that is specific to sentiment analysis. The authors utilized both classic ML models (SVM, MNB, DT, LR, KNN, and RF) and deep learning models (Convolutional Neural Network, BiLSTM, and hybrids thereof; BiLSTM+CRF, and BiLSTM+LSTM) to predict feelings in code-mixed Telugu-English text. Character and word n-grams in the range  $n=(1,3)$  are used to train ML models, along with hand-selected attributes such as the number of special characters, capital letters, and numerals. The BiLSTM+LSTM model performed at an accuracy of 0.98 on the Blog dataset, while the BiLSTM+CRF model performed at an accuracy of 0.99.

A 5,536-strong Kannadaguli community has created a collection of scrobbed videos from YouTube that have a combination of Tulu and English. While creating the dataset, the author left comments in either Tulu or a code that combined Tulu and English; he or she then extracted just the remarks written in Latin script. Krippendorff aimed to establish standards for scholars to follow when annotating data. The Tulu-English annotated dataset was used to train a wide variety of NLP (NB, LR, DT, k-NN, RF, SVM, and Principal Component Analysis) and DL (BiLSTM and Contextualized Dynamic Meta Embeddings) models, as well as a transformer-based classifier (BERT). TF-IDF scores and Keras embeddings are two features that can be helpful for ML and DL models. When compared to its rivals, the BiLSTM model performed admirably. The Dravidian languages of Tamil, Kannada, Malayalam, and Telugu have been notably overlooked in the literature on sentiment analysis. Tulu English code-mixed YouTube videos collected by a 5,536-com Kannadaguli. [2]. While building the dataset, the author exclusively extracted Latin-script comments, all of which were in Tulu or an English-Tulu code. To guarantee uniformity in the annotation process, Krippendorff sought to establish a consensus between the annotators. Multiple ML models (NB, LR, DT, k-NN, RF, SVM, and Principal Component Analysis), Deep Learning models (BiLSTM and Contextualized Dynamic Meta Embeddings), and a transformer-based classifier with BERT models were trained using the annotated Tulu-English dataset. ML and DL models can take advantage of features such as TF-IDF scores and Keras embeddings, for instance. The BiLSTM model outperformed the majority of the alternatives. Literature reviews reveal that the Dravidian languages of Tamil, Kannada, Malayalam, and Telugu are rarely considered in sentiment analysis due to a lack of resources.

## 3. METHODOLOGY

### A. Creating and Annotating a Corpus

It was our intention to create a data set suitable for scientific investigation that would incorporate both Tamil and English encoding. The YouTube comments were gathered using the YouTube Comment Scraper Tool. We started by collecting 12100 Tamil sentences from the comments sections of YouTube videos. If you typed "Online class during the covid era" into YouTube in 2021, you would get a slew of

Positive State: There is a explicit or implicit clue in the text suggesting that the speaker is positive state, i.e., Happy, Admiring, Relaxed, Forgiving etc.  
 நேர்மனதில் நின்று பேசுகின்ற நேர்மனதையுடைய நிலை என்று உதாரையில் சொல்லப்பட்டுள்ளது. நேர்மனதையுடைய நிலை என்பது மனப்போக்கைக் குறிக்கிறது. நேர்மனதில், நிதர்னமமாக, மனநிலையில் போன்றவை.

☐ Understand

☐ No

Negative State: There is an explicit or implicit clue in the text suggesting that the speaker is Negative State i.e., Sad, Angry, Anxious, Violent, etc.

☐ Understand

Both Positive and Negative or Mixed Feelings. There is an explicit or implicit clue in the text suggesting that the speaker is both Positive and Negative feeling. Comparing two

① Understand

Neutral State: There is no explicit or implicit indicator of the speaker's emotional state. Example for asking like or subscription or questions about release date or movie dialog etc.

☐ Understand

### a. Annotations

□ Positive words-These words stand for the enthusiastic support, inspiration, admiration, and adulation of any individual, thing, or circumstance. They also symbolize the positive expression of an emotion, feeling, or viewpoint. Values ranging from "0" to "1" are assigned to these words that have an upbeat effect.

□ Negative words—Certain words have negative connotations that can be used to convey disapproval, criticism, failure, melancholy, or a pessimistic mindset. Negative words are those that range from “-1” to “0,” and are referred to as such.

□ Neutral words-Neither a good nor a bad feeling may be found in these words. For the purpose of analyzing any statement, certain terms are not taken into consideration. The value "0" is assigned to these words.

☐ Mixed feelings: There is evidence in the text—either explicit or implicit—that the speaker is feeling both good and bad: Comparing two movies

### b. Annotators

When we were done with the Google form, we made sure that men and women each contributed an equal amount of annotations.

We have collected these comments using the form shown in figure 2.

#### (a) Sample Google Form-1

Choose the Best Sentiment  
சுல்லோரும் தலை குனிந்து நடப்பதற்குக் காரணம், சுமார்தப்போதும்  
சோவியல்மீடியாக்களும் தான்

☐ Positive  
☐ Negative  
☐ Mixed-Feelings  
☐ Unknown state  
☐ Not Tamil

---

Choose the Best Sentiment  
இன்டர்நெட் இல்வையென்றால், சிலருக்கு வேறு எந்த வேலையும் செய்யவே  
முடியாது.

☐ Positive  
☐ Negative  
☐ Mixed-Feeling  
☐ Unknown State  
☐ Not Tamil

---

Choose the Best Sentiment  
இந்த பட்டிக்கை இப்படித்தான் viewers sekureengla

☐ Positive  
☐ Negative  
☐ Mixed-Feeling  
☐ Unknown State  
☐ Not Tamil

#### (b). Sample Google Form-2

Figure 2: Google Form

There were 15 willing participants. They were all native Tamil speakers, however, their genders, educational backgrounds, and mediums of instruction varied. The annotators are listed in Table 1. The volunteers were given a Google form to fill out, and 100 sentences to read and rate. Each volunteer has the option of annotating as many or as few sentences from the corpus as they would like, and if they volunteer to annotate more, they will receive another Google form with another batch of 100 sentences. Our annotators, both male, and female are evenly split among the forms we send them. Annotator details are provided in Table 2.

**Table 2: Criteria for Annotators**

Gender	Male	9
	Female	6
Education	Undergraduate	4
	Graduate	4
	Post Graduate	7
Medium of Schooling	English	8
	Tamil	7
Total		15

#### C. Inter-Annotator Agreement

This is quantified by calculating the inter-annotator agreement, which shows the degree to which raters generally agree with one another. This is important because it guarantees that the annotation system is reliable and that different raters will all assign the same sentiment label to the same comment. Both of the following questions are related to the issue of annotator agreement. How do the different annotators' annotations compare and contrast with one another? Is there a likelihood that the annotators' level of agreement or disagreement might be explained by chance alone? The percentage of agreement is more straightforward to calculate than the second question.

We utilized Krippendorff's alpha( $\alpha$ ), a popular but computationally costly method, to gauge annotators' level of agreement. In contrast, Krippendorff's alpha ( $\alpha$ ) is more applicable given its flexibility in terms of sample size, the number of raters, domains, and measurement scales (it may be used for the nominal, ordinal, interval, and ratio scales) as well as its resistance to missing data. Since not

Table 3: Sharing of Information

Category	Tamil-English
+Ve sentence	11657
-Ve sentence	2057
Different feelings	1982
Neutral state	920
Other Languages	467
	17083

For instance, a conflict between the annotators' classifications of Positive and Negative feelings is more significant than one between Mixed and Neutral feelings is able to manage these conflicts.  $\alpha$  is outlined as:

The annotators' examined divergence ( $D_e$ ) in sentiment labeling is compared to the projected divergence ( $D_p$ ) when the coding of sentiments may be related to chance rather than a characteristic of the sensation itself.

$$D_e = \frac{1}{n} \sum_c \sum_k o_{ckmetric} \delta_{ck}^2 \quad \text{Equation 1.2}$$

$$D_p = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_{kmetric} \delta_{ck}^2 \quad \text{Equation 1.3}$$

Using nominal and interval metrics, we determined the level of agreement between annotators. The range of the symbol is from 0 to 1, or from 1 to 0. If is 1, then all annotators agree 100%, and if it's zero, then there's no consensus at all and the outcomes are completely at random. To provide a starting point, we applied a number of deep learning methods to identify the sentiments of Tamil-English code- mixed YouTube posts.

#### 4. EXPERIMENTAL SETTINGS

Based on the corpus created 3 models were validated with the corpus generated.

##### A. BERT-Multilingual:

[28] The transform-based bidirectional encoder representation for encoding languages was presented. Its primary purpose is to pretrain on unlabeled text, and it can be trained and refined by adding a final layer. BERT is used for any classification problem that involves text [29]. We investigate methods for classifying code-mixed data according to appropriate sentiments. Figure 3 depicts the overall pre-training and tuning processes of the BERT

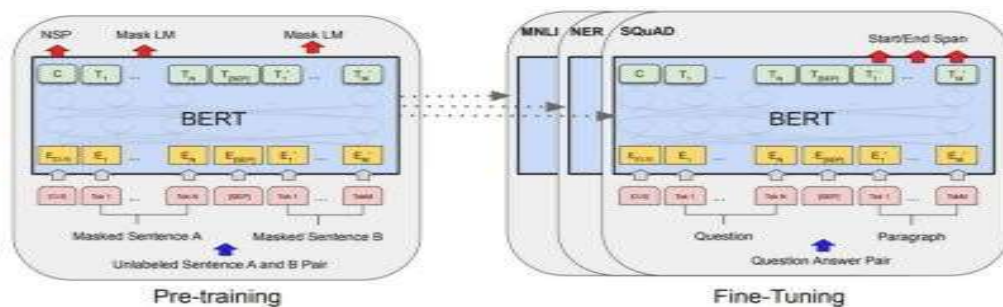


Figure.3. BERT's Pre-Training and Fine-Tuning Procedures



The BERT model is trained to utilize the Masked language model (MLM) and the subsequent sentence prediction task (NSP). To produce predictions for masked tokens, the MLM uses a deep bidirectional model trained on a random sample of input tokens.

	<b>Category</b>	<b>+ve sentence</b>	<b>-ve sentence</b>	<b>Different feeling</b>	<b>Neutral position</b>
	<b>Support</b>	<u>2980</u>	<u>674</u>	<u>560</u>	<u>680</u>
<b>Precision</b>	<b>BERT</b>	<u>0.71</u>	<u>0.41</u>	<u>0.41</u>	<u>0.44</u>
<b>Recall</b>		<u>0.85</u>	<u>0.39</u>	<u>0.13</u>	<u>0.37</u>
<b>F-Score</b>		<u>0.78</u>	<u>0.4</u>	<u>0.2</u>	<u>0.41</u>

Table 4 . Comparison of performance metric for BERT

#### A. RoBERTa

Robert [30] is an alternative to BERT that can predict the next sentence without using a training target. Fine-tuning Roberta is depicted in a single statement (Figure 4). Rather, with the Masked Language Modelling objective in mind, training of the language model makes use of larger mini-batches and learning rates. Roberta's well-considered architecture allows it to outperform the standard BERT baseline on downstream NLP tasks. When conducting experiments, we made use of Roberta's features.

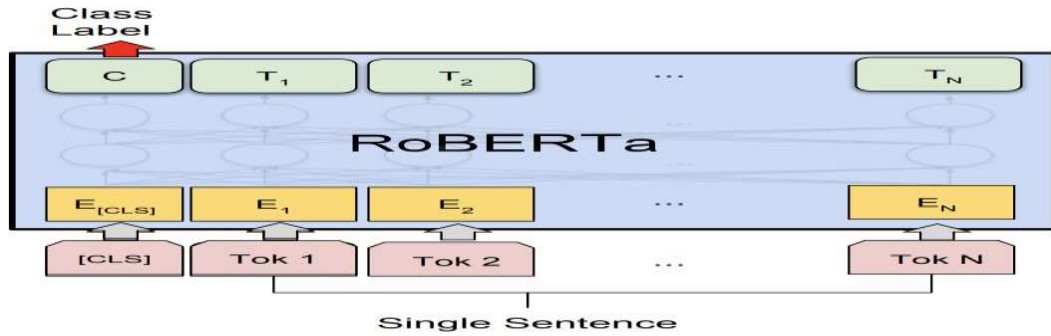


Figure.4. Fine-tuning Roberta in single sentence

Table 5. Comparison of performance metric for Roberta

	<b>Category</b>	<b>+ve sentence</b>	<b>-ve sentence</b>	<b>Different feeling</b>	<b>Neutral position</b>
	<b>Support</b>	<u>2980</u>	<u>674</u>	<u>560</u>	<u>680</u>
<b>Precision</b>	<b>Roberta</b>	<u>0.7</u>	<u>0.39</u>	<u>0.34</u>	<u>0.46</u>
<b>Recall</b>		<u>0.82</u>	<u>0.42</u>	<u>0.12</u>	<u>0.36</u>
<b>F-Score</b>		<u>0.75</u>	<u>0.4</u>	<u>0.18</u>	<u>0.4</u>

#### A. XLM-Roberta

It was shown that XLM-Roberta, an unsupervised cross-lingual representation strategy, outperformed multi-lingual BERT on a variety of cross-lingual benchmarks, hence it was offered as a viable alternative [31]. The XLM-R model was developed for evaluating and inferring on a wide range of downstream tasks after being trained on data from Wikipedia articles written in 100 different languages. The XLM-Roberta model is depicted in Figure.5 and is utilized as an Entity alignment model.

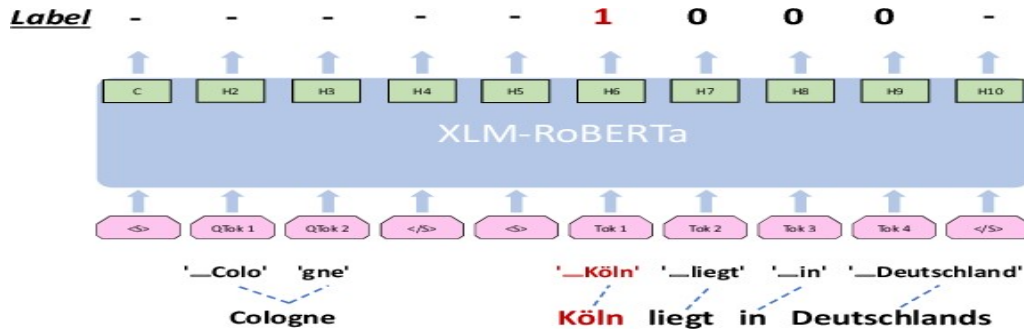


Figure.5.XLM-Roberta-Entity alignment model.

	Category	+ve sentenc e	-ve sentenc e	Differen tfeeling	Neutra l positio n
	Support	2980	674	560	680
Precision	XLM- Roberta	0.73	0.49	0.41	0.62
Recall		0.86	0.54	0.19	0.46
F-Score		0.79	0.51	0.34	0.53

Table 6.Comparison of performance metric for XLM-Roberta

## 5. RESULTS AND DISCUSSION

The data set is not normally distributed. Table 4 shows that out of a total of 17,803 sentences, 67% fall into the Positive category. When compared to the Neutral and Mixed groups, the Positive group performed higher on tests of precision, memory, and the F-measure. These two classes present unique challenges for human annotators, and as noted in Section 3, they are rather uncommon in the dataset. But the Negative and Other Language groups did really well. We think this is because the data has a much higher proportion of negative comments and more overt hints for negative and non-Tamil words. To illustrate how deep learning works, Figure 5 displays an analysis of classifiers that use the BERT model.

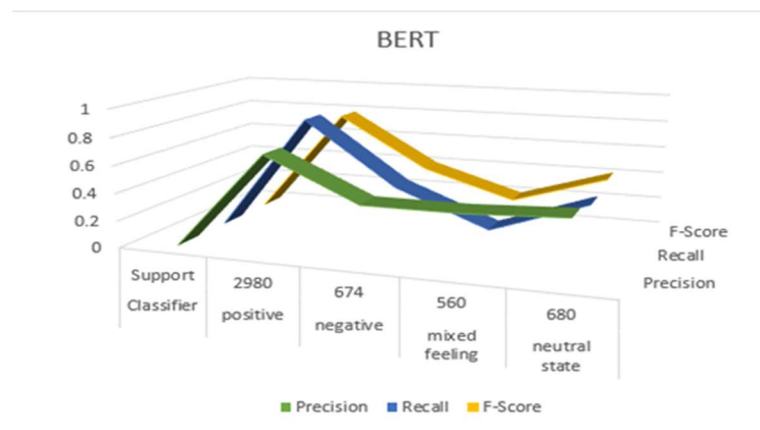


Figure 5: Analysis of classifiers based on the BERT model

Figure. 6 clearly shows the Analysis of classifiers based on the Roberta model –a Deep learning technique techniques.

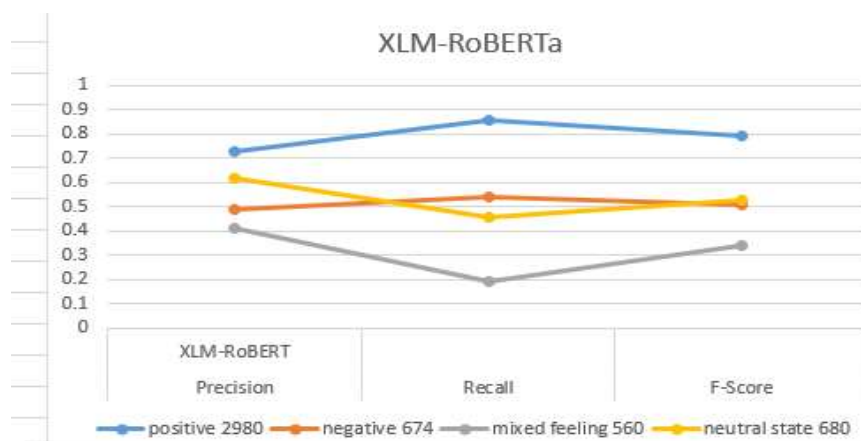
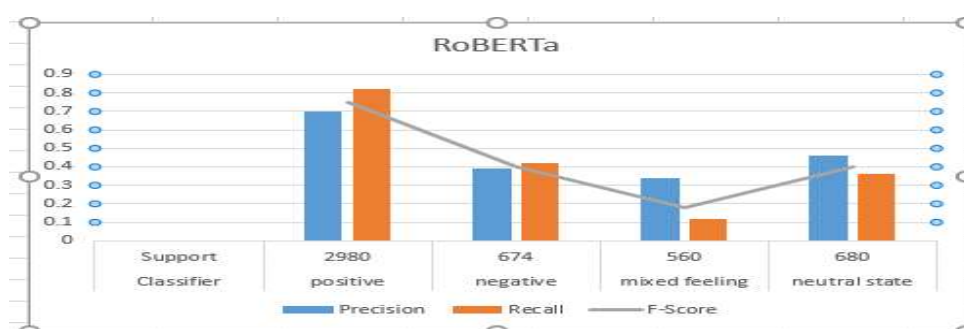


Figure 7. Analysis of classifiers based on the XLM-Roberta model



## 6. CONCLUSION

In this article, We introduce the Tamil-English corpus, a bilingual dataset of annotated YouTube comments suitable for sentiment analysis. Researchers will be able to undertake code-mixed sentiment analysis research and receive useful data thanks to this annotation project. In addition to making the corpus available to researchers, we also give a measure of inter-annotator agreement, calculated using Krippendorff's alpha and a set of baseline values.

## REFERENCES

- [1] Scotton, C. M. (1982). The Possibility of CodeSwitching: Motivation for Maintaining Multilingualism. pages 432–444.
- [2] Sangeetha, M., Mathivanan, R. (2022). Exploration of Sentiment analysis Techniques on a Multilingual Dataset dealing with Tamil-English reviews. In Proceedings of the ACCAI, IEEE. DOI: 10.1109/ACCAI53970.2022.9752612
- [3] Suryawanshi, S., Chakravarthi, B. R., Verma, P., Arcan, M., McCrae, J. P., and Buitelaar, P. (2020). A Dataset for Troll Classification of Tamil Memes. In Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation, pages 7–13.
- [4] Coffman, K. G., & Odlyzko, A. M. (2002). Internet growth: Is there a “Moore’s Law” for data traffic? In Handbook of massive data sets (pp. 47-93). Springer, Boston, MA. <http://dx.doi.org/10.2139/ssrn.236108>
- [5] Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA. [https://doi.org/10.1007/978-1-4614-3223-4\\_13](https://doi.org/10.1007/978-1-4614-3223-4_13)
- [6] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In A practical guide to sentiment analysis (pp. 1-10). Springer, Cham. [https://doi.org/10.1007/978-3-319-55394-8\\_1](https://doi.org/10.1007/978-3-319-55394-8_1)
- [7] Al-Saqqah, S., & Awajan, A. (2019). The use of word2vec model in sentiment analysis: A survey. In Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control (pp. 39-43). AIRC’19, Cairo, Egypt, December 14-19.
- [8] Krishnasamy, K. (2015). Code mixing among TamilEnglish bilingual children. International Journal of Social Science and Humanity, 5(9):788.

- [9] Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. In Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for UnderResourced Languages) (SLTU-CCURL 2020), Marseille, France, May. European Language Resources Association (ELRA).
- [10] Phani, S., Lahiri, S., and Biswas, A. (2016). Sentiment analysis of Tweets in three Indian languages. In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016), pages 93–102, Osaka, Japan, December. The COLING 2016 Organizing Committee
- [11] Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. In Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for UnderResourced Languages) (SLTU-CCURL 2020), Marseille, France, May. European Language Resources Association (ELRA)
- [12] Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2018). Improving wordnets for under-resourced languages using machine translation. In Proceedings of the 9th Global WordNet Conference (GWC 2018), page 78.
- [13] R. Nareshkumar and K. Nimala, “An Exploration of Intelligent Deep Learning Models for Fine-Grained Aspect-Based Opinion Mining,” 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICES), Jul. 2022, doi: 10.1109/ices55317.2022.9914094.
- [14] Habimana, O., Li, Y., Li, R., Gu, X., and Yu, G. (2019). Sentiment analysis using deep learning approaches an overview. Science China Information Sciences, 63(1):111102, Dec.
- [15] Valitutti, R. (2004). WordNet-Affect: an effective extension of wordnet. In Proceedings of the 4th International Conference on Language Resources and Evaluation, pages 1083–1086.
- [16] Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC’06, pages 417–422.
- [17] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In EMNLP, volume 14, pages 1532–1543.
- [18] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- [19] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017a). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.
- [20] Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019a). Comparison of different orthographies for machine translation of under- resourced Dravidian languages. In 2nd Conference on Language, Data and Knowledge (LDK 2019). SchlossDagstuhl-Leibniz- Zentrum fuer Informatik.
- [21] Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019b). WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation, pages 1–7, Dublin, Ireland, 19 August. European Association for Machine Translation.
- [22] Chakravarthi, B. R., Priyadharshini, R., Stearns, B., Jayapal, A., S, S., Arcan, M., Zarrouk, M., and McCrae, J. P. (2019c). Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. In Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, pages 56–63, Dublin, Ireland, 20 August. European Association for Machine Translation.
- [23] Padmamala, R. and Prema, V. (2017). Sentiment analysis of online Tamil contents using recursive neural network models approach for Tamil language. In 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), pages 28–31, Aug
- [24] Das, A. and Bandyopadhyay, S. (2010). SentiWordNet for Indian languages. In Proceedings of the Eighth Workshop on Asian Language Resources, pages 56–63, Beijing, China, August. Coling 2010 Organizing Committee.
- [25] Ranjan, P., Raja, B., Priyadharshini, R., and Balabantaray, R. C. (2016). A comparative study on code-mixed data of Indian social media vs formal text. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), pages 608–611
- [26] Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. In Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for UnderResourced Languages) (SLTU-CCURL 2020), Marseille, France, May. European Language Resources Association (ELRA).
- [27] Kannadaguli, P., 2021, November. A Code-Diverse Kannada-English Dataset For NLP Based Sentiment Analysis Applications. In 2021 Sixth International Conference on Image Information Processing (ICIIP) (Vol. 6, pp. 131-136). IEEE.
- [28] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Proceedings of NAACL-HLT 2019, Minneapolis, Minnesota, June 2 - June 7, 2019. c 2019 Association for Computational Linguistics.
- [29] Tayyar Madabushi, H., Kochkina, E., and Castelle, M. (2019). Cost-sensitive BERT for generalizable sentence classification on imbalanced data. In Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda pages 125–134, Hong Kong, China, November. Association for Computational Linguistics.
- [30] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692

- 
- [31] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, 2020. Computational Linguistics.

## Digital Library and its Features

J.Lingeswaran <sup>1</sup>, P.Mangayarkarasi <sup>2</sup>

<sup>1</sup> Principal, Sri Venkateswara College of Education, Parasur, Tiruvannamalai Dist, Tamilnadu, India

<sup>2</sup> Assistant Professor and Head Department of Linguistics,  
Tamil University, Thanjavur, Tamilnadu, India.

---

### ABSTRACT (10 PT)

This research article deals with the role of digital library particularly focuses the sites used for digital library and advanced features. It gives the vivid idea about the digital library and its uses. The researcher points out the digital library and its kinds of materials, e-book collections, and digital formats are briefly explained in the introductory chapter. The researcher also deals with the types of digital library and some of the examples relevantly and he clearly explains how to access them. This paper lists out the soft ware and hard ware of the digital library. Finally it consists of the importance of digital library.

---

### Keywords:

A Digital Library  
B Software  
C Hardware  
D e-library

---

### Corresponding Author:

J.Lingeswaran  
Principal, Sri Venkateswara College of Education  
Parasur, Tiruvannamalai Dist,  
Tamilnadu, India.

---

## 1. INTRODUCTION

Digital libraries are Internet sites consecrated to the creation and preservation of e-book collections and holdings of other kinds of materials, without the need for end users to purchase the materials they want to consult and read. Creating and preserving these collections involves the participation of a large number of intermediate institutions, which is part of what makes digital libraries so interesting. Among the participants are those institutions that secure from the publishers the rights to transform or distribute their materials in digital formats, and libraries that purchase the rights for the members of their institutions to access these materials while respecting certain conditions. In some case, libraries do not acquire copyright but merely are licensed by publishers and distributors to consult these materials. Digital libraries are mainly stocked with sources of information that are available on the Internet in open access format, and they are remarkable for the ease of access to collections, the networking possibilities they offer, and the universal availability of their collections. These libraries are places where new digital objects are added to conventional documents already housed there.

Among the most noteworthy of these digital libraries are the Project Gutenberg, the World Digital Library and the Europeana Library. The World Digital Library was created by the U.S. Library of Congress and inaugurated on 21 April 2009; the Europeana digital library, inaugurated on 20 November 2008, is an open access library that serves Europe. There are digital library projects sponsored by national libraries, among which the Miguel de Cervantes Digital Library [Biblioteca Virtual Miguel de Cervantes] of the Biblioteca Nacional de España, the Gallica digital library of the Bibliothèque Nationale de France and other such projects stand out. The contents of these libraries are fundamentally works in the public domain. Though national legislation may vary, a work is typically considered to be in the public domain 70 years after the death of the author. As a result of this lack of international consensus, there are countries where works can be in the public

domain only 50 years after the author's death, which explains how in one country an author's works may be freely available while in another they are not.

## 2. RELATED LITERATURES

In the 1980s, libraries card catalogs were being replaced by Online Public Access Catalogs (OPACs). These were usually closed systems that could contain little more than bibliographic data. Most OPACs were done in Machine Readable Cataloging (MARC) format. It generally represents an individually published item or "information product," and describes the physical characteristics of the item itself (Brenner et al, 2006). In the 1990s and beyond, digital libraries changed the way we have thought about how we retrieve information. What exactly is a digital library? According to Donald Waters, digital libraries are "organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities" (Waters, 1998). This definition allows for a great degree of interpretation. The concept of digital library has multiple senses that one might invoke in various contexts. For example, the concept may refer simply to the notion of collection without reference to organization, intellectual accessibility or service attributes. This extended sense seems to be in play, for example, when we hear the World Wide Web described as a digital library. The concept might also refer to the organization underlying the collection, or even more specifically to the computer-based system in which the collection resides (DLF, 1995). Digital libraries represent the meeting point of a large number of disciplines and fields, i.e., data management, information retrieval, library sciences, document management, information systems, the Web, image processing, artificial intelligence, human-computer interaction, and others (Ioannidis, 2005). The Alex Catalogue of Electronic Texts is one example of a digital library.

## 3. TYPES OF DIGITAL LIBRARY AND EXAMPLES

### Stand-alone Digital Library (SDL)

This is the regular classical library implemented in a fully computerized fashion. SDL is simply a library in which the holdings are digital (i.e., electronic scanned or digitized). The SDL is self-contained the material is localized and centralized. In fact, it is a computerized instance of the classical library with the benefits of computerization. Examples of SDLs are the Library of Congress (LC) and its National Digital Library (NDL) (<http://www.loc.gov>), and the Israeli K12 Portal Snunit (<http://www.snunit.k12.il>) .

### Federated Digital Library (FDL)

This is a federation of several independent SDLs in the network, organized around a common theme, and coupled together on the network. A FDL composes several autonomous SDLs that form a networked library with a transparent user interface. The different SDLs are hetero and are connected via communication networks. The major challenge in the construction and maintenance of a FDL is interoperability (since the different repositories use different metadata formats and standards). Examples of FDLs are the Networked Computer Science Technical Reference Library (NCSTRL) (<http://www.ncstrl.org>) and Networked Digital Library of Theses and Dissertations (NDLTD) (<http://www.ndltd.org>).

### Harvested Digital Library (HDL)

This is a virtual library providing summarized access to related material scattered over the network. A HDL holds only metadata with pointers to the holdings that are one click away in Cyberspace. The material held in the libraries is harvested (converted into summaries) according to the definition of an Information Specialist (IS). However, a HDL has regular DL characteristics, it is finely grained and subject focused. It has rich library services, and has high quality control preserved by the IS, who is also responsible for annotating the objects in the library. The HDL harvesting model is further detailed in section 3. Examples of HDLs are the Internet Public Library (IPL) (<http://www.ipl.org/>) and the Virtual Library (<http://www.vlib.org/>).

### EXAMPLES OF SOME DIGITAL LIBRARIES AND HOW TO ACCESS THEM.

LOC – Library of Congress American Memory (<http://memory.loc.gov/ammem/>)

NSDL – National Science DL (<http://nsdl.org>)

IPL – Internet Public Library (<http://www.ipl.org>)

CDL – California DL (<http://www.cdlib.org>)  
 ADL Alexandria DL (<http://www.alexandria.ucsb.edu>)  
 BL – British Library (<http://www.bl.uk/>)  
 NZDL New Zealand DL (<http://www.nzdl.org/>)  
 Einstein Archives Online (<http://www.alberteinstein.info/>)  
 IEEE Digital library ([www.ieeedl.com](http://www.ieeedl.com))  
 ACM Digital library ([www.acmdl.org](http://www.acmdl.org))  
 Networked Digital Library of Theses and Dissertations (NDLTD)-(<http://www.ndltd.org>).  
 ArticleCentral.com! (<http://www.articlecentral.com/>)  
 Networked Computer Science Technical Reference Library (NCSTRL) – (<http://www.ncstrl.org>)  
 Israeli K12 Portal Snunit (<http://www.snunit.k12.il>).

#### 4. SOFTWARES INVOLVED

There are different softwares used in digital library such as:

Alfresco (software):

Alfresco is a free/libre enterprise content management system for Microsoft Windows and Unix-like operating systems. It is used for Enterprise content management for documents, web, records, images, and collaborative content development.

Cambridge imaging system:

It was founded in 1996, is a software company based near Cambridge, UK that specializes in enterprise video platforms. It has one subsidiary company, Screenocean, based in London, UK, an online digital library containing program material and related metadata from the Channel 4 archive.

Digital Commons:

Digital Commons is a hosted open access institutional repository and publishing solution, combining traditional institutional repository functionality with tools for peer-reviewed journal publishing, conference management, and multimedia.

DSpace:

DSpace is an open source repository software package typically used for creating open access repositories for scholarly and/or published digital content. While DSpace shares some feature overlap with content management systems and document management systems, the DSpace repository software serves a specific need as a digital archives system, focused on the long-term storage, access and preservation of digital content.

EXo-Platform:

EXo Platform is an open source, standard-based, Enterprise Social Platform written in Java and distributed under the GNU Lesser General Public License. The platform is sold and distributed by eXo Inc., a global company with U.S. headquarters in San Francisco, California, global headquarters in France, and offices in Tunisia and Vietnam.

Fedora Commons:

Fedora (or Flexible Extensible Digital Object Repository Architecture) is a digital asset management (DAM) architecture upon which institutional repositories, digital archives, and digital library systems might be built. Fedora is the underlying architecture for a digital repository, and is not a complete management, indexing, discovery, and delivery application. It is a modular architecture built on the principle that interoperability and extensibility are best achieved by the integration of data, interfaces, and mechanisms (i.e., executable programs) as clearly defined modules.

Greenstone (Software):

Greenstone is a suite of software tools for building and distributing digital library collections on the Internet or CD-ROM. It is open-source, multilingual software, issued under the terms of the GNU General Public License. Greenstone is produced by the New Zealand Digital Library Project at the University of Waikato, and has been developed and distributed in cooperation with UNESCO and the Human Info NGO in Belgium.

Intra-text:



IntraText is a digital library that offers an interface while meeting formal requirements. Texts are displayed in a hypertextual way, based on a Tablet PC interface. By linking words in the text, it provides Concordances, word lists, statistics and links to cited works. Most content is available under a Creative Commons license it also offers publishing services that enable similar advantages. The IntraText interface applies a cognitive ergonomics model based on lexical hypertext and on the Tablet PC or touch screen interface. It uses a set of tools and methods based on HLT (Human Language Technologies). IntraText is a reading, reference and search tool. It can be used to read a work, to browse a text as hypertext, to search for words and phrases just through a simple click of your pen or mouse.

#### Invenio:

Invenio is an open source software package that provides the tools for management of digital assets in an institutional repository. The software is typically used for open access repositories for scholarly and/or published digital content and as a digital library. Invenio is developed by the CERN Document Server Software Consortium, and is freely available for download. Free and paid support models are available. The service provider TIND Technologies was established in 2013 to accommodate the growing demand for support of Invenio.

#### Islandora:

Islandora is an open source digital repository system based on Fedora Commons, Drupal and a host of additional applications. It is open source software (released under the GNU General Public License) and was developed at the University of Prince Edward Island by the Robertson Library. Islandora may be used to create large, searchable collections of digital assets of any type and is domain-agnostic in terms of the type of content it can steward. It has a highly modular architecture with a number of key features

#### Knowledge Tree:

Knowledge Tree, Inc. provides online software that helps sales and marketing teams discover, manage, and refine the collateral they use in sales engagements. The technology is tuned for sales, sales operations, and marketing teams. Based in Raleigh, North Carolina, the company also has an office in Cape Town, South Africa.

The company's product, also called Knowledge Tree, makes use of the Amazon EC2 cloud computing platform and Salesforce.com's Force.com platform. Knowledge Trees features including content discovery, reporting, and editing are designed to support B2B sales situations that depend on collateral and documents. The service is available on a subscription basis.

#### Pleade:

Pleade is an open source search engine and browser for archival finding aids encoded in EAD (an XML standard for encoding archival finding aids). Based on the SDX platform, it is a very flexible web application.

#### SABDA:

SABDA or SABDA Bible Software is an Indonesian integrated Bible study platform that's based on the Online Bible engine, [1] with multilingual Bibles available in the program (including Indonesian, Malay, English, Greek and Hebrew, and many local languages of Indonesia). The word sabda is the Indonesian word for Logos (via Sanskrit: shabda), and also abbreviation of "Software Alkitab, Biblika Dan Alat-alat" (Bible Software, Biblical Resources, And Tools). It is produced and managed by Yayasan Lembaga SABDA (SABDA Foundation) which translated and made available freely more than 100 Biblical modules in Indonesian since 1994, besides the default OLB modules.

## 5. HARDWARE INVOLVED

The hardware involved are as follows:

- Computer, mobile phone or any device that can access the network.
- Storage device or Database where data and information are kept and stored
- Scanner that will be used to convert traditional object into Digitized objects.
- Printer will be used to print out digitized object.
- Internet Modem which will be used to access the network.
- Traditional Materials Such as books, magazines e.tc.

## 6. IMPORTANCE OF DIGITAL LIBRARY

Scholarly communication, education, research such as E-journals, e-books, and data sets, e-learning. Access to cultural collections such as Cultural heritage, historical & special collections, museums, biodiversity. E-governance such as Improved access to government policies, plans, procedures, rules and regulations, and Archiving and preservation.

## 7. ADVANTAGES OF DIGITAL LIBRARY

No physical boundary. The user of a digital library needs not to go to the library physically; people from all over the world can gain access to the same information, as long as an Internet connection is available. Round the clock availability a major advantage of digital libraries is that people can gain access 24/7 to the information.

Multiple access:

The same resources can be used simultaneously by a number of institutions and patrons. This may not be the case for copyrighted material: a library may have a license for "lending out" only one copy at a time; this is achieved with a system of digital rights management where a resource can become inaccessible after expiration of the lending period or after the lender chooses to make it inaccessible (equivalent to returning the resource).

Information retrieval:

The user is able to use any search term (word, phrase, title, name, and subject) to search the entire collection. Digital libraries can provide very user-friendly interfaces, giving click able access to its resources.

Preservation and conservation:

Digitization is not a long-term preservation solution for physical collections, but does succeed in providing access copies for materials that would otherwise fall to degradation from repeated use.

Space:

Whereas traditional libraries are limited by storage space, digital libraries have the potential to store much more information, simply because digital information requires very little physical space to contain them and media storage technologies are more affordable than ever before.

Added value:

Certain characteristics of objects, primarily the quality of images, may be improved. Digitization can enhance legibility and remove visible flaws such as stains and discoloration.[14]

## 8. CONCLUSION

These digital libraries do completely different things, have completely different interfaces, and use different technology to display information for their users. The Alex Catalog of Electronic Texts has a very simple, clean layout. It also has a simple search function. There is no advanced search available with this digital library. American Memory by the Library of Congress is a little more advanced than the Alex Catalog of Electronic Texts. American Memory has a little more complicated set up because it offers a wider variety in its collections. The Alexandria Digital Library has a much more sophisticated search function. While using their National Geospatial Digital Archive, one can search anywhere throughout the United States and recover satellite images, multiple varieties of air photos, and maps.

There will be continued expansion of digital library activities. Digital libraries will build upon work being done in the information and data management area. Digital libraries provide an effective means to distribute learning resources to students and other users. Planning a digital library requires thoughtful analysis of the organization and its users, and an acknowledgment of the cost and the need for infrastructure and ongoing maintenance (Adams, Jansen, and Smith 1999).

## REFERENCES

- [1] Academic Info. (n.d.). Retrieved from Academic Info: Digital Library: <http://www.academicinfo.net/digital.html>
  - [2] Adams, W. J. (1999) . Planning, building, and using a distributed digital library. Third International Conference on Concepts in Library and Information Science. Dubrovnik, Croatia.
  - [3] AlderMan, J. (1998). Digital Library Project. Retrieved from <http://www.unf.edu/~alderman/DigitalLibraries/DLProjects.html>
  - [4] Besser, H. (n.d.). Historical Background of Digital library. Retrieved from Digital Humanities : <http://www.digitalhumanities.org/companion/>
  - [5] CSDL. (2007). The Center for the Study of Digital Libraries. Retrieved from <http://www.cSDL.tamu.edu/cSDL/center/center.htm>
  - [6] Different types of Digital Library. (2005). Retrieved from Computer Technology File Storage: [www.helpme.com](http://www.helpme.com)
- J.Lingeswaran, *Principal, Sri Venkateswara College of Education, Parasur, Tiruvannamalai Dist, Tamilnadu, India*

## Tools for constructing AI/ML solutions in Tamil

Abdul Majed Raja RS, Muthiah Annamalai  
[littlecoder@gmail.com](mailto:littlecoder@gmail.com)  
[ezhillang@gmail.com](mailto:ezhillang@gmail.com)

---

### ABSTRACT (10 PT)

We contend creation of new AI/ML applications in Tamil is still hard despite relative abundance of Tamil datasets [1]; this is due to scarcity of Tamil tools. However the accessibility of fully-trained models and capability of providing pre-trained models, like huggingface [2], are much harder and still require domain expertise in hardware and software.

While individuals have published [3-4] some small Jupyter notebooks, and articles, but they still remain inadequate to scale the breadth of Tamil computing needs in AI world among:

(1) NLP – Text Classification, Recommendation, (2) Spell Checking, (3) Correction tasks, TTS – speech synthesis tasks, and ASR – speech recognition

While sufficient data exist for 1, the private corpora for speech tasks (அருந்தமிழ் பட்டியல்), the public corpora of a 300hr voice dataset recently published from Mozilla Common Voice (University of Toronto, Scarborough, Canada leading Tamil effort [5a]) have enabled data completion to a large degree for tasks 2 and 3. Private repositories exist for voice data under Penn LDC.

Ultimately the missing tooling can provide capability to quickly compose AI services based on open-source tools and existing compute environment to host services and devices in Tamil space. We propose for community to build a pytorch-lightning [5b] like API for Tamil tasks across NLP, TTS, ASR via AI so that newer AI/ML applications are easily built. Role of central institutions and governments is also explored.

---

### Keywords:

A NLP  
B Tamil  
C Text Classification  
D Recommendation  
E Spell Checking

---

### Corresponding Author:

Muthiah Annamalai,  
INFITT, USA  
Email: [ezhillang@gmail.com](mailto:ezhillang@gmail.com)

---

## 1. INTRODUCTION

The main text Recently, DALL-E images (Generative AI) by Open-AI, and Stable Diffusion models by Emad Mostaque of Stability AI provides promise generative capabilities to average users unleashing creativity (Fig. 1). These tools and technologies provide pathways to adapt some fast AI applications for good (provide TTS in voice of disabled person who has lost voice) and nuisance, or mischief (fake-news) etc. Generative AIs have their unresolved problems we list under biases portion of this paper.

AI technologies allows several applications for Tamil community but we have report that adapting them safely and creatively with positive outcomes require more work from side of tooling, data curation among other metrics.



Fig. 1: Prompt to DALL-E from OpenAI [6] describing (a) street temple-car festival (Thiruvizha) in Madurai at night; (b) Tamil family celebrating festival of lights Deepavali in Madurai; same for bottom row as well. The prompt asked AI to generate in pastel style.



## 2 MODELS

Traditionally, Machine Learning Models were built for specifics - Specific Task, Specific Language - like Text Classification for English, Text Classification for Tamil and so on. Recently, Since the rise of Transformer-based models like BERT, The lines of these specifics have gotten blurred. Thanks to Large Language Models that are trained on huge datasets and Millions and Billions of Parameters, The same model that's used for English Translation can also be used for Tamil Translation.

### 2.1 Zero-Shot Usage

Zero-shot learning is a machine learning technique that allows a model to recognize an object or a concept that it has never seen before. While many, if not most, machine learning models require a large amount of training data, zero-shot learning can recognize an object or concept without any new training data. Large Language Models are trained on a large amount of text data. These models can be used to answer questions about text, such as "what is the most likely next word in a sentence?" Hence these models work fairly good out-of-box making them ideal candidates for a good Zero-shot usage.

An example of a Zero-shot usage is to use a Large Language Model like GPT-3 for Sentiment Classification for Tamil Language. Thus eliminating the need for new training data.

## 2.2 Model Fine-Tuning

Model Fine-tuning is a technique that allows a model to be trained on a new dataset. Large Language Models or Foundational Models can be fine-tuned to get improved performance on a new dataset. This became quite popular since the beginning of Transfer Learning. Transfer Learning is the process of applying knowledge gained in one context to a different context. For example, if you have learned how to use Microsoft Word, you can apply that knowledge to using OpenOffice. In the same way, Foundational Models or Large Language Models trained for Text Generation tasks can be used for other applications like Sentiment Analysis, Entity Extraction, Grammar Correction and so on.

While Zero-shot Learning can work fairly well in a general context and is good for the English language, It can improve the performance of these models very well if they get trained on a relatively smaller dataset. Fine-tuning a Large Language Model to let the fine-tuned model perform NLP for the dataset that is similar to the fine-tuned dataset can be a very effective way to use Foundational LLMs for Tamil NLP tasks.

This does not mean that these models cannot be used in Zero-shot capability but means they do a lot better if they are fine-tuned on relevant dataset which in our case is new Tamil Dataset.

### 2.3 Model Serving

One of the least addressed problems in ML and AI is how to serve the Model to developers and end-users. It is important that we serve both Developers, who would build on top of our toolkit and end-users who would directly use our toolkit to leverage AI/ML for their Tamil NLP requirements. Hence we propose two distinct ways to serve these models as a central toolkit for Tamil AI/ML

1. A Python Library for Developers
2. A Gradio App

The Python Library that can be hosted for free on PyPi can serve the Tamil developers who want to use our Toolkit to build applications and services leveraging Tamil AI/ML while the Gradio App that can be run locally on any computer (preferably GPU) or hosted for free on Hugging Face Spaces can serve the end-users like Tamil Content Creators who want to include our Toolkit as part of their workflow.

## 2.4 Model Selection

While there is a growing number of Large Language Models every single day, It's very important for us to pick the right model that can work well for Tamil Language. One of the easiest ways to select the right model for Tamil is by looking at the training dataset information.

Most open source Large Language Models indicate their training dataset composition. From that information, We can understand which of those existing Large Language Models have got the most Tamil Data during the Model Training. This is primarily applicable for a Zero-shot Learning since Fine-tuned models mostly would have been fine-tuned on Tamil Dataset.

For example, Big Science's BLOOM model was 46 Natural Languages and 13 Programming Languages. Tamil is one of those Natural Languages of the Indic category which is ~4% of all the languages. Even though Tamil is a very small part of the entire language set, The Zero-shot tasks like Text Generation that we experiment for Tamil works fairly fine.



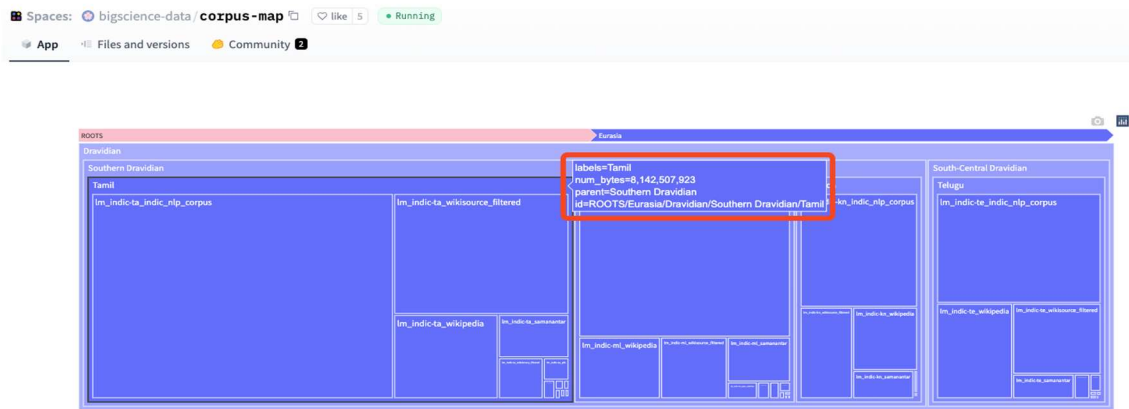


Fig. 2: Corpus map used to train a specific model [7]

BLOOMZ and mT0, a family of models capable of following human instructions in dozens of languages zero-shot. BLOOMZ and mT0 are finetuned from BLOOM and mT5 pretrained multilingual language models on crosslingual task mixture (xP3) and the resulting models are capable of crosslingual generalization to unseen tasks & languages. In the case of BLOOMZ & mT0, Tamil is just 0.5% of the fine-tuned data, Yet the model is capable of performing tasks like Sentiment analysis, Text generation, Keywords creation and so on.

### 3 AI APPLICATIONS FOR TAMIL

RoBERTa and BERT models are customized for Tamil by finetuning the final layers for classification of idioms in work [17]. We report in this section how various NLP, TTS applications can be solved using AI/ML models.

#### 3.1 Spelling Correction with LLM

We may use masked words as '<mask>' when input sentence to check for spelling correction on certain words in sentence that are out-of-dictionary or not correctable by known rules [8];

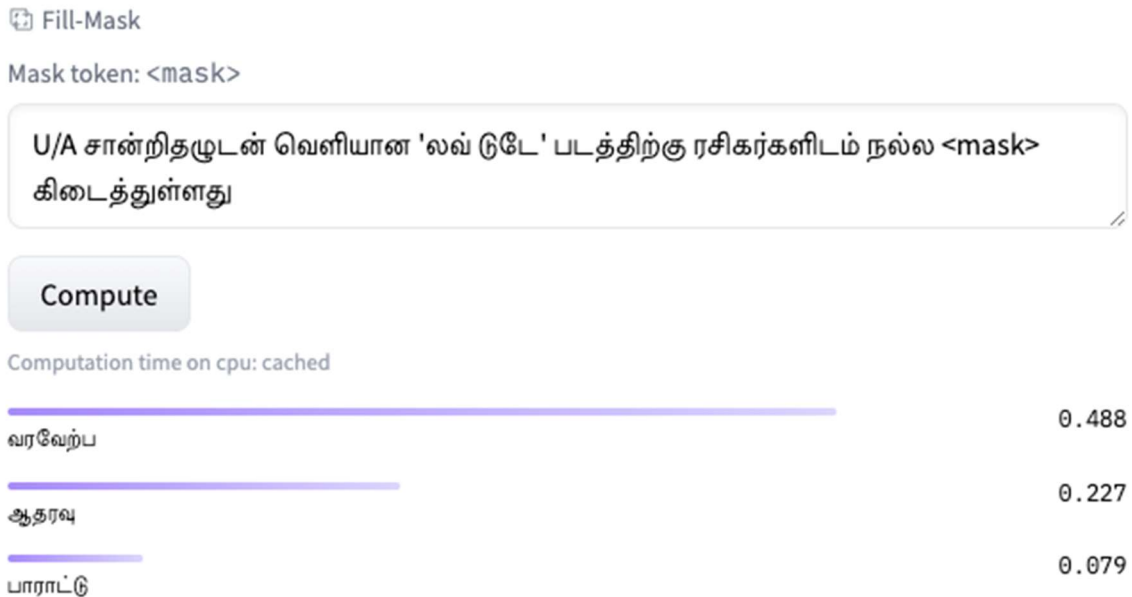


Fig. 3.1: Spelling checker functionality of LLM using masking; missing word is recommended வரவேற்பு.

#### 3.2 Sentiment Recognition with LLM

Sentiment Recognition in NLP is the task of identifying the correct sense of a word in a given context. This is one of the most used tasks in NLP given how much text data is available in the world. It's also largely sought after given the business applications of Sentiment Analysis Models.



Fig. 3.2: sentiment recognition of text by using LLMs.

With the help of LLMs, We can use the existing Foundational models for Sentiment Analysis in Tamil Language without the need for a new training dataset. For example, We used **BLOOMZ** LLM for performing Sentiment Analysis of a Tamil Review in a Zero-shot Context.

### 3.3 Named-Entity Recognition with LLM

Named-Entity Recognition is the task of identifying the names of people, places, organizations, and other entities of interest in text. This is a key component of many natural language processing applications. Using Large Language Models for Named-Entity Recognitions can be a very good application.

Below is an example of using BLOOMZ model for Named-Entity Recognition.

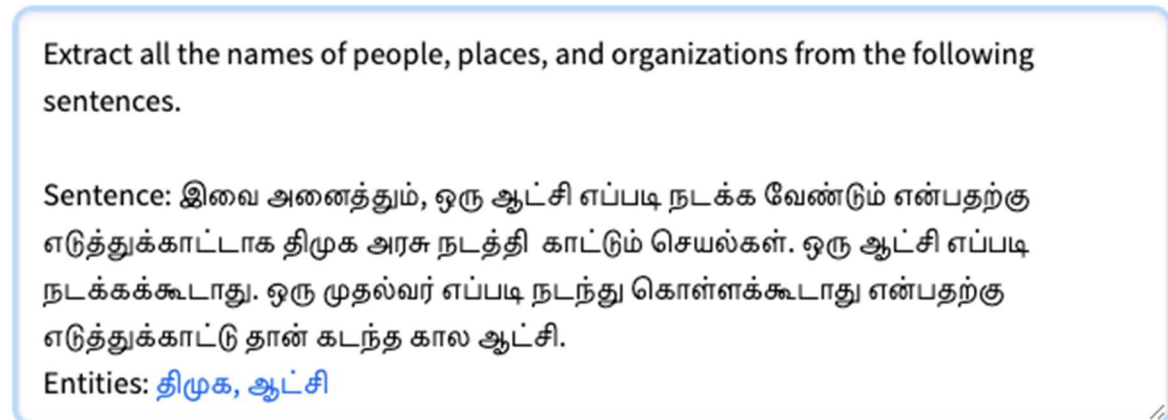


Fig. 3.3: Name-entity recognition using LLMs.

### 3.4 Audio and Voice Applications - ASR, TTS

ASR and TTS models based on sequence-to-sequence transformation pioneered by researchers at Meta (Facebook) have been adopted by authors to present a good demonstrations of TTS applications in Tamil, and other major Indian languages [15]. We note however number to words conversion remains a sore point in this implementation as

compared to work [20].

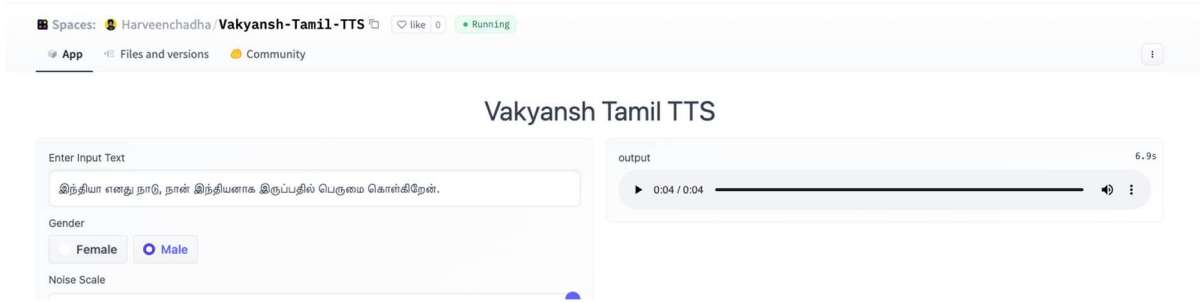


Fig 3.4: Demo Space for work

Clearly we can see the improved quality of AI/ML based TTS over unit-selection synthesis based approaches.

OpenAI's Whisper [16], as reported in [18], is demonstrated to translate high-quality lyrical Tamil audio with transcription and errors highlighted in the following figure.

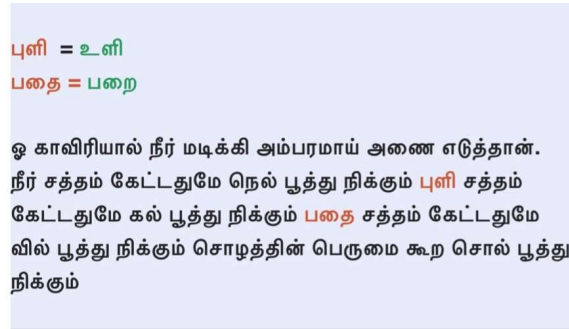


Fig 3.5: Experimental results of Malaikannan [18] using OpenAI Whisper for Tamil ASR based on the popular song “பொன்னி நதி” from the movie “பொன்னியின் செல்வன்.” This is showing low word-error rate 6.4%.

#### 4 TAMIL TOOLING GAPS

Our proposal is the following to address the gaps, and we also understand many of the steps are further problems on their own:

1. Develop a open-source toolbox for pre-training and task training specialization
2. Identify good components to base effort
3. Contribute engineering effort, testing, and validation
  1. R&D – DataScience, Infra, AI framework
  2. Engineering Validation – DataScience, Tamil language expertise
  3. Engineering – packaging, documentation, distribution
  4. Project management
4. Library to be liberally licensed MIT/BSD
5. Open-Source license for developed models
6. Find hardware resources for AI model pre-training etc.
7. Managed by a steering committee / nominated BDFL
8. Scope – decade time frame



## 9. Financial support for such a wide effort

### 4.1 Datasets related Tooling

Currently hosted datasets [1] not consumable in uniform interface for Torch or with TensorFlow in a uniform format; we have only raw data today.

### 4.2 Model Related Tooling

- model attribute, training time, standardized accuracy metrics, training dataset, notions of biases etc. are absent

### 4.3 Compute related challenges

Free compute is limiting on what can be done; Google Cloud CoLaboratory is limited in credits that are freely available; training CNN or LSTM takes lot of time on laptop scale hardware.

There is a chronic need for special purpose AI Accelerators (GPU, RDUs, etc.) for large scale models pre-training; there needs to be efforts in private-public collaboration to subsidize cost and sponsorship these activities.

### 4.4 Problems and Biases

Just a decade ago the auto-complete in Google search query with the words “Tamil “ will always end with “Tigers,” limiting what an uninformed lay-person could learn about Tamil people, language or culture; which such a subjective bias has been removed it remains largely un-tested in various areas. This would be considered as harmful bias against Tamils by virtue of language marker in the discourse of [10].

Large language models (LLMs) are known to have problems with representing minorities along various margins, problems with performing math (calculators), potential to be environmentally harmful, repeat harmful stereotypes on minorities by age, nationality, race or other marking criteria [10], etc.

Language models exhibit a variety of expertise to work as auto-pilots in coding tasks [11], as email marketing assistants [12] etc. however as autonomous agents still much remains to be achieved [13] - current generation of AI models and agents are in rung-1 of 3-step ladder of causation [14] and act based on observation but not in a causal framework of learning which would be the creation of near-human level intelligence.

Specifically for Tamil language, as a largely under-resourced language, we find the nature of AI-systems to largely dependent on public data sets (uncurated) and few private data sets, and goodwill of giant corporations like Google or Meta (Facebook) to develop models for tasks. In such cases the pre-trained models are not qualified for biases. Additionally where data is not available or incorrect data is available the systems will not be able to reason correctly causing problematic consequences for applications of such AI models for Tamil community. Overall sufficient availability of compute, data, correctness and bias measures for Tamil tasks are needed to quantify bias in AI models.

Advent of generative AI models like DALL-E, Stable Diffusion etc. have created a chaotic situation of attribution, fair-use and copyright.

As a Tamil community we would want our real-world language, cultural, audio-visual, written and oral cultural milieu to be within the “in-distribution” of training set of the language/visual/multi-modal models for AI. When such an ecosystem of data driven AI modeling, and harm reducing systems exist perhaps someday we can hope to eliminate biases about individuals, groups, or minorities (by various labels) for creation of a oracular AI agents which can be native to Tamil.

## 5. SUMMARY AND CONCLUSIONS

AI/ML systems rely of good data; we note dominance of Tamil data reflects in metrics like OpenAI's Whisper (ASR model) performing on Tamil audio to have lowest word-error rate (at 20.6%) among Indian languages (even compared to Hindi at 26.9%) perhaps evidence of data prevalence and seeds of digitization and open-content in parallel corpora (audio + transcribed text) available in Tamil [16].

We have presented various aspects of AI/ML systems which can benefit the Tamil community in general and gaps in tooling which can accelerate the delivery of AI based applications in hands of general developer and community members, democratizing AI.

t format consists of a flat left-right columns on A4 paper (quarto). The margin text from the left and top are 2.5cm, right and bottom are 2 cm. The manuscript is written in Microsoft Word, single space, Time New Roman 10pt or Vijaya 9 pt and maximum 10 pages, which can be downloaded at the website: <https://tamilinternetconference.org/>. ஆய்வுக் கட்டுரைகளை ஆங்கிலம் / தமிழ் / இரண்டிலும் சமர்ப்பிக்கலாம். தமிழ் கட்டுரைகள் விஜய எழுத்துருக்களில் இருக்க வேண்டும்.

A title of article should be the fewest possible words that accurately describe the content of the paper. Omit all waste words such as "A study of...", "Investigations of...", "Implementation of...", "Observations on...", "Effect of...", "Analysis of..." etc. Indexing and abstracting services depend on the accuracy of the title, extracting from it keywords useful in cross-referencing and computer searching. An improperly titled paper may never reach the audience for which it was intended, so be specific.

The Introduction should provide a clear background, a clear statement of the problem, the relevant literature on the subject, the proposed approach or solution, and the new value of research which it is innovation. It should be understandable to colleagues from a broad range of scientific disciplines. Organization and citation of the bibliography are made in IEEE style in sign [1], [2] and so on. The terms in foreign languages are written italic (italic). The text should be divided into sections, each with a separate heading and numbered consecutively. The section/subsection headings should be typed on a separate line, e.g., **1. Introduction** [3]. Authors are suggested to present their articles in the section structure: **Introduction - the comprehensive theoretical basis and/or the Proposed Method/Algorithm - Research Method - Results and Discussion – Conclusion.**

Literature review that has been done author used in the chapter "Introduction" to explain the difference of the manuscript with other papers, that it is innovative, it are used in the chapter "Research Method" to describe the step of research and used in the chapter "Results and Discussion" to support the analysis of the results [2]. If the manuscript was written really have high originality, which proposed a new method or algorithm, the additional chapter after the "Introduction" chapter and before the "Research Method" chapter can be added to explain briefly the theory and/or the proposed method/algorithm [4].

## REFERENCES (10 PT)

- [1] <https://huggingface.co/spaces/Harveenchadha/Vakyansh-Tamil-TTS>
- [2] INFITT அருந்தமிழ் - Awesome Tamil resource list, <https://github.com/INFITTOfficial/awesome-tamil> (accessed Nov , 2022).
- [3] T. Wolf, "Huggingface's transformers: State-of-the-art natural language processing," (2019).
- [4] M. Annamalai, "AI and Tamil Computing opportunities", tutorial at Tamil Internet Conference (2021) link
- [5] (a) AbdulMajedRaja Bloomz model for AI, <https://www.youtube.com/1littlecoder> (accessed Nov 14, 2022);
- [6] (b) Niklas Muennighoff, et-al, "Crosslingual Generalization through Multitask Finetuning," arXiv:2211.01786 (2022)
- [7] (a) UTSC Tamil Digital Studies Program Common Voice project <https://tamil.digital.utoronto.ca/en/tamil-common-voice>
- [8] (b) Pytorch Lightning <https://www.pytorchlightning.ai/> 4
- [9] DALL-E - Generative AI images from text by Open-AI, 2022 (accessed Nov 1, 2022)
- [10] Tamil portion of Corpus map of BigScience model, <https://huggingface.co/spaces/bigscience-data/corpus-map> (accessed Nov 28, 2022)
- [11] M. Annamalai, T. Shrinivasan, "Algorithms for certain classess of Tamil spelling correction," Tamil Internet Conference, Chennai, India (2019).
- [12] R. Bommasani et-al, "On the Opportunities and Risks of Foundation Models," Stanford Center for Research on Foundation Models Report, August (2021).
- [13] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" Proc. of ACM Conference FAccT '21, New York, NY, USA, pages 610–623, (2021)
- [14] Github Autopilot, <https://github.com/features/copilot> (accessed 2022)
- [15] Jasper AI (<https://www.jasper.ai>) , (2022)
- [16] Pearl, Judea, and Dana Mackenzie. "AI can't reason why." Wall Street Journal (2018).

- [17] Pearl, Judea, and Dana Mackenzie, "The Book of Why: The New Science of Cause and Effect," Basic Books, (2018).
- [18] Harveen Singh Chadha, et-al, "Vakyansh: ASR Toolkit for Low Resource Indic languages," arXiv:2203.16512 [cs.CL] (2022).
- [19] Alec Radford, et-al "Robust Speech Recognition via Large-Scale Weak Supervision," OpenAI Report (2022).
- [20] Briskilal, J. and Subalalitha, C.N., 2022. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. Information Processing & Management, 59(1), p.102756.
- [21] Malaikannan S, private communication (Nov, 2022).
- [22] (a) Malaikannan S, "Can a machine write a story ?," blog post (2016).
- [23] (b) Anderj Karpathy, "The Unreasonable Effectiveness of Recurrent Neural Networks," (2015).
- [24] Annamalai, Muthiah, and Sathia Mahadevan. "Generation and Parsing of Number to Words in Tamil." Tamil Internet Conference (2020).

## கணினிவழி அகரநிரல் உருவாக்கல்

நீச்சல்காரன் சே. இராஜாராமன்

கணித்தமிழ் ஆர்வலர்

<https://www.neechalkaran.com>

### ABSTRACT (10 PT)

#### Keywords:

- A திரட்டிப் பட்டியலிட்டு
- B அகரநிரல்
- C கலைச்சொற்களை
- D சொற்கட்டி(index)
- E கணினி மொழியியல்

பொதுவாக ஒரு நூலில் உள்ள சொற்களை எல்லாம் திரட்டிப் பட்டியலிட்டு அகரநிரல் அல்லது சொற்கட்டி(index) என்ற பெயரில் கடைசிப் பக்கங்களில் வெளியிடப்படும். இதன் மூலம் நூலிலுள்ள கலைச்சொற்களை எளிதில் அடையாளம் காணமுடியும். இன்னும் சில இலக்கிய நூல்களுக்கு ஒரு சொல் எங்கே வந்தன, அதன் பொருள் என்று ஒரு சொல்லடைவும்(concordance) உருவாக்கப்படும். முன்பு இவற்றை எல்லாம் மனிதர்களே செய்ய வேண்டியிருந்தன. ஆனால் தற்போதைய கணினி மொழியியல் வளர்ச்சியில் இவற்றைக் கணினியே செய்ய முடியும். எப்படிச் செய்யலாம் அதன் பின்னே உள்ள நுட்பங்களை இக்கட்டுரை விவரிக்கிறது.

#### Corresponding Author:

நீச்சல்காரன் சே. இராஜாராமன்

கணித்தமிழ் ஆர்வலர்

<https://www.neechalkaran.com>

## 1. INTRODUCTION

ஓர் எழுத்தாளரின் சொல்லாற்றலைக் கணிக்க அவர் பயன்படுத்தும் சொற்களின் எண்ணிக்கையும், அதன் முடுக்கமும் முக்கிய காரணியாகும். அதாவது எத்தனைப் புதிய சொற்களைக் கையாண்டார் என்றும் புதிய சொற்களின் பிரயோகம் அடுத்தடுத்த பக்கங்களில் எவ்வளவு மாறுகிறது என்றும் கணித்தால் அவரின் சொல்லாற்றலை மதிப்பிட முடியும். அந்த வகையில் கணித்துக் காட்ட சூளகு (<http://apps.neechalkaran.com/sulaku>) என்ற கருவி ஏற்கனவே உள்ளது.

அதில் ஏதேனும் ஒரு படைப்பைத் தந்து அடிச்சொல் என்று தேர்வு செய்தால், அப்படைப்பில் உள்ள சொற்களை உருபனியல் பகுப்பாய்வு செய்து அடிச்சொற்களைப் பட்டியலிடும். இதன்மூலம் அதிகம் பயன்படுத்தப் பட்ட சொற்களை அறியமுடியும். மேலும் எழுத்து, சீர் என்று கூடுதலாகவும் பல தகவல்களைத் தரும். ஆனால் இது முழுமையான சொல்லடைவை உருவாக்காமல் சொற்பட்டியலை மட்டுமே எடுத்துத் தரும். எந்தச் சொல் எந்தப் பக்கம் என்று கணிக்கமுடியாது.

இதற்குமாற்றாக VaaniNLP பொதியின் துணையுடன் சொல்லடைவை உருவாக்க முடியும். எவ்வாறு உருவாக்கலாம் என்று பைத்தான் மொழியினை மையமாகக் கொண்டு கீழே விவரிக்கப்படுகிறது. ஒரு docx அல்லது odt கோப்புகளை அதற்கான நிரல் கொண்டு படித்து கீழே உள்ள வழியில் உரலிக்குப் பதில் பக்க உள்ளடக்கத்தைக் கொடுத்து, அகரநிரலினை உருவாக்கலாம். இந்த உதாரணத்தில் இணையத்தளத்தினை எடுத்துள்ளோம். விரும்பிய உரலிகளை array மாறிலியாக எடுத்துக் கொள்ள வேண்டும்.

**சுளகு**

சொல்லாப்வுக் கருவி இது வாணியின் வெளியீடு உதவிக்கூறிப்பு

**மொத்தம் எழுத்து சீர் இதர அழி**

டிசம்பர் 6-ம் தேதி அதிகாலை 4 மணிக்கு அண்ணாமலையார் கோயில் மூலவர் சன்னதியில் பரணி தீபமும், மாலை 6 மணிக்கு 2,668 அடி உயரமுள்ள அண்ணாமலை உச்சியில் மகா தீபம் ஏற்றப்பட உள்ளன. கார்த்திகைத் தீபத் திருவிழாவுக்கு 30 லட்சம் பக்தர்கள் வருவார்கள் என ஆட்சியர் பா.முருகேஷ் தெரிவித்துள்ளார். பக்தர்களின் வசதிக்காக டிசம்பர் 5-ம் தேதி முதல் 8-ம் தேதி வரை 20 சிறப்பு ரயில்கள் இயக்கப்படும் என தென்னக ரயில்வே அறிவித்துள்ளது. சிறப்பு ரயில்களை தவிர, திருவண்ணாமலை வழியாக தினசரி வழக்கமாக இயக்கப்படும் அனைத்து ரயில்களும் தடையின்றி இயக்கப்பட உள்ளன.

அடிச்சொல்	எண்ணிக்கை
மணி	8
தேதி, ம்	7
திருவண்ணாமலை	6
தீபம்	5
இரவு, டிசம்பர், இயக்கு, ரயில், சிறப்பு, 6	4
திருவிழா, கார்த்திகை, இரு, வழி, 45, 8, 5	3
பக்தர், வா, வரை, முதல், புதுச்சேரி, வந்தடை, அதிகாலை, வேலூர், புறப்படு,	2

```
#Index_generator
import requests
from bs4 import BeautifulSoup
from tamil import vaaninlp
urls=["https://tech.neechalkaran.com/2022/11/venmurasu-concordance.html",
      "https://tech.neechalkaran.com/2022/11/blog-post.html",
      "https://tech.neechalkaran.com/2022/09/blog-post.html",
      "https://tech.neechalkaran.com/2018/02/tamil-sorting.html"
      ]
```

அதன் பின்னர் beautifulsoup அல்லது Scrapy, Selenium, Octoparse போன்ற ஏதேனும் இணையப் பக்கத்தைச் சுரண்டும் நிரலகம் கொண்டு உள்ளடக்கத்தை மட்டும் எடுத்துக் கொள்ளவேண்டும். அந்த உள்ளடக்கத்தை word\_tokenize வழி சொற்களாகப் பகுத்துக் கொள்ள வேண்டும்.

```
masterword={}
for i in range(len(urls)):
    xpage= requests.get(urls[i])
    xpage.encoding = 'utf-8'
    prose=xpage.text
    xsoup = BeautifulSoup( xpage.text)#html5lib
    body =xsoup.find("article")
    sol = vaaninlp.word_tokenize(body.text)
```

பின்னர் lemmatize வழி ஒவ்வொரு சொற்களின் அடிச்சொற்களைப் பிரித்துக் கொள்ள வேண்டும். பொதுவாக அனைத்துச் சொற்களையும் ஒரேநேரத்தில் அடிச்சொல் பிரிப்பதைவிட ஒரு வரையறை வைத்து கொஞ்சம் கொஞ்சமாகப் பிரித்துக் கொள்ளலாம். இந்த vaaninlp பொதியானது ஏறக்குறைய பெரும்பாலான அடிச்சொற்களைப் பிரிக்கும் ஆனால் அதனால் பிரிக்கமுடியாதவற்றை உள்ளது உள்ளபடி கொடுத்துவிடும்.

```

leng=0
limit=1500#size of the array size in lemmatize
stem=[]
nonstem=[]
while leng<len(sol):
    result= vaaninlp.lemmatize(sol[leng:leng+limit])
    leng=leng+limit
    for j in range(len(result)):
        if(result[j]["Flag"]==True):
            stem.append(result[j]["RootWords"][0])
        else:
            nonstem.append(result[j]["Userword"])
    stem=list(set(stem))
    nonstem=list(set(nonstem))

```

இறுதியாக கிடைத்துள்ள தனித்த சொற்களை பக்கக் குறியுடன் இணைத்து, மொத்தமாகச் சேர்த்தால் சொற்கட்டி கிடைத்து விடுகிறது.

இங்கே p1, p2, p3 என்று உதாரணத்திற்கு இடப்பட்டுள்ளது. அதை விரும்பிய குறியீட்டில் குறித்துக் கொள்ளமுடியும்.

முனைவர் p2  
சற்று p2  
கூகிள் p2,p3  
ஆர்வம் p2  
பட்டம் p2  
விதி p2  
சேர் p2  
தமிழி p2  
நீக்கு p2,p3  
உரையாசிரியர் p2  
கிரந்தம் p2,p3  
கட p2  
எழுத்துரு p2  
சார் p2  
தேடு p2  
தற்போது p2  
பெரும்பாலும் p2,p3

#### வரம்புகள்:

- இந்தப் பொது அனைத்து விதமான சொற்களையும் புரிந்து கொள்ளாது. சுமார் 65000 அடிச்சொற்களுடன் சுமார் 13 கோடி சொற்களைப் புரிந்து கொள்ளும். தற்காலத் தமிழுக்கு உருவாக்கப்பட்டதால் சமகால இலக்கியமல்லாத அனைத்து படைப்புகளிலும் முழுமையாகப் பிரிக்காமல் போகலாம்.
- பெயர்ச் சொற்கள், வினைச் சொற்களைத் தவிர மற்றவற்றைப் பிரிக்காது, அடையாளமும் காட்டாது. “வளர்கிறதே” என்பதை “வளர்” என்று பிரித்துக் காட்டும், “நாகப்பட்டினத்தில்” என்பதை “நாகப்பட்டினம்” எனப் பிரித்துக் காட்டும். ஆனால் “இவ்வகை”, “பெரும்பாலும்”, “நாகைப்பட்டினத்தில்” போன்ற இடைச்சொற்கள், தனிப் பெயர்கள் போன்றவற்றைப் பிரிக்காது. அனைத்துச் சொற்களும் கலந்தே இருக்கும் என்பதால் கடைசியாக ஒருமுறை சரிபார்த்துக் கொள்ள வேண்டும்.
- ஒன்றிற்கும் மேற்பட்ட சொற்களைக் கொண்ட பெயர்களை அடையாளம் காணாது. உதாரணமாக “தமிழ் நாடு” என்று இருந்தால் தமிழ் தனியாகவும் நாடு தனியாகவும் கணக்கில் கொள்ளப்படும். பெயர் பொருள் சுட்டி (Named Entity Recognition) திறனையும் எதிர்காலத்தில் பயன்படுத்தி இன்னும் மேம்பட்ட அகரநிரலை உருவாக்க வேண்டும்.

- கணினியால் பிரிக்க முடியாதவை என்றால் புதிய சொல் என்று எடுத்துக் கொள்ள முடியாது ஆனால் நடைமுறையில் அதிகம் புழங்காத சொற்கள் என்று சொல்லமுடியும். நாளை மொழியியல் திறன் கூடும் போது இந்த எண்ணிக்கையில் மாற்றங்கள் நிகழும் என்பதையும் கருத்தில் கொள்ளலாம்.

#### சொல்வங்கி

இந்த முறையில் இரண்டு இலக்கியங்களின் சொற்கள் சேகரிக்கப்பட்டன. இரண்டும் இணையத்தில் உரைவடிவில் கிடைப்பதாலும் இந்த இரண்டு இலக்கியங்களின் உரைநடை தனித்த கவனம் பெற்றதாலும் தேர்ந்தெடுக்கப்பட்டன.

#### பொன்னியின் செல்வன்

அமரர் கல்கியின் வரலாற்றுப் புதினமான பொன்னியின் செல்வன் அதன் நிற்காத மொழிநடைக்குப் புகழ்பெற்றது. <https://book.ponniyinselman.in> என்ற முகவரியில் உள்ளவற்றைச் சுரண்டி ஆய்வுசெய்கையில் கீழே உள்ள புள்ளிவிவரங்கள் கிடைக்கின்றன.

பகுதிகள்	மொத்தச் சொற்கள்	வாணியால் பகுக்க முடியாதவை
பகுதி 1	74179	2164
பகுதி 2	148252	4220
பகுதி 3	214437	6121
பகுதி 4	63492	1578
பகுதி 5	198743	4669
மொத்தம்	699103	18752

முழுத் தரவு <https://www.kaggle.com/datasets/neechalkaran/ponniyinselman>

#### வெண்முரசு

எழுத்தாளர் ஜெயமோகனின் வெண்முரசு புதினத்தின் இணையத்தளத்திலிருந்து <https://venmurasu.in/> சுரண்டி ஆய்வு செய்யப்பட்டது. இப்புதினத்தின் 26 நூல்களில் 1932 அத்தியாயங்களில் சுமார் 1.38 கோடி சொற்களைக் கணினிவழியாக அலசப்பட்டன. அதில் 90% சொற்களின் அடிச்சொல் கணிக்கப்பட்டும் இதர சொற்கள் உள்ளவாறே பட்டியலாகியுள்ளன.

நூல்	மொத்தச் சொற்கள்	வாணியால் பகுக்க முடியாதவை
முதற்கனல்	84543	7718
மழைப்பாடல்	173733	15391
வண்ணக்கடல்	141509	13881
நீலம்	56463	8567
பிரயாகை	153956	9996
வெண்முகில்நகரம்	343689	29066
இந்திரநீலம்	525807	47110



காண்டிபம்	678381	58313
வெய்யோன்	159062	15516
பன்னிரு படைக்களம்	331091	33345
சொல்வளர்காடு	452717	45029
கிராதம்	159620	16135
மாமலர்	347741	32027
நீர்க்கோலம்	195108	18865
எழுதழல்	348760	35363
குருதிச்சாரல்	505304	52012
இமைக்கணம்	598061	61434
செந்நா வேங்கை	753298	77181
திசைதேர் வெள்ளம்	902531	93482
கார்கடல்	1070304	109038
இருட்கனி	1195815	119636
தியின் எடை	1301463	128115
நீர்ச்சுடர்	1416884	137438
களிற்றியானை நிரை	1568449	150526
கல்பொருசிநுரை	157435	11071
முதலாவிண்	183548	13618
மொத்தம்	<b>13805272</b>	<b>1339873</b>

முழுத் தரவு: <https://www.kaggle.com/datasets/neechalkaran/venmurasu>

#### பயன்பாடு

ஒரு மொழி வளர் அம்மொழியில் இலக்கியங்கள் வளர் வேண்டும். அந்தவகையில் வெண்முரசு மற்றும் பொன்னியின் செல்வன் தமிழில் இலக்கியத்தில் முக்கியப் புதினங்களாகும். புதினம் முழுக்க பல புதிய சொற்கள் கையாளப்பட்டுள்ளன. புதியவர்களுக்கு வாசிக்க கடுமையானதாக இருந்தாலும், வாசிக்கக் கூடியவர்களுக்கு மொழியின் இனிமையையும் உணர்த்துகிறது. சொல்லாய்வு செய்யவும் மீளாய்வு செய்யவும் இச்சொற்கள் பயன்படும். அறிவியல்பூர்வமாக எத்தனைப் புதிய சொற்கள் உள்ளன எனக் கண்டுபிடிக்கவும் அகரநிரல் உருவாக்கவும் இந்த முயற்சி பயன்படும்.

கடந்த நூற்றாண்டு முதல் வெளிவந்த அனைத்துத் தமிழ் நூல்களுக்கும் அகரநிரல் உருவாக்க முடியும். ஒவ்வொரு நூலாசிரியரின் சொல் வளத்தை மதிப்பிடமுடியும்.

நூல்கள் மட்டுமல்லாமல் இணையத்தளத்தையும் மதிப்பிடலாம். ஒவ்வொரு வலைப்பதிவு, டிவிட்டர் பதிவு, செய்தித்தளங்கள் என்று அகரநிரல் உருவாக்கினால் மொழிதாண்டி தகவல் திரட்டவும் பேருதவியாக இருக்கும்.



வெளியிணைப்புகள்:

- [1] X. S. Li, *et al.*, "Analysis and Simplification of Three-Dimensional Space Vector PWM for Three-Phase Four-Leg Inverters," *IEEE Transactions on Industrial Electronics*, vol. 58, pp. 450-464, Feb 2011.
- [2] மூல நிரல்: [https://github.com/neechalkaran/VaaniNLP/blob/main/Samplecode/index\\_generator.py](https://github.com/neechalkaran/VaaniNLP/blob/main/Samplecode/index_generator.py)
- [3] VaaniNLP பொது <https://pypi.org/project/VaaniNLP/>
- [4] பொன்னியின் செல்வன் சொல்வங்கி <https://www.kaggle.com/datasets/neechalkaran/ponniyinselvan>
- [5] வெண்முரசு சொல்வங்கி <https://www.kaggle.com/datasets/neechalkaran/venmurasu>

சித்த மருத்துவக் களத்தின் சமூக-மூலப்பொருண்மையியல் அமைப்பும் தகவல் மீட்டும்

எஸ். வீர அழகிரி

எஸ். இராசேந்திரன்

அமிர்தா விஸ்வ வித்யபீடம்

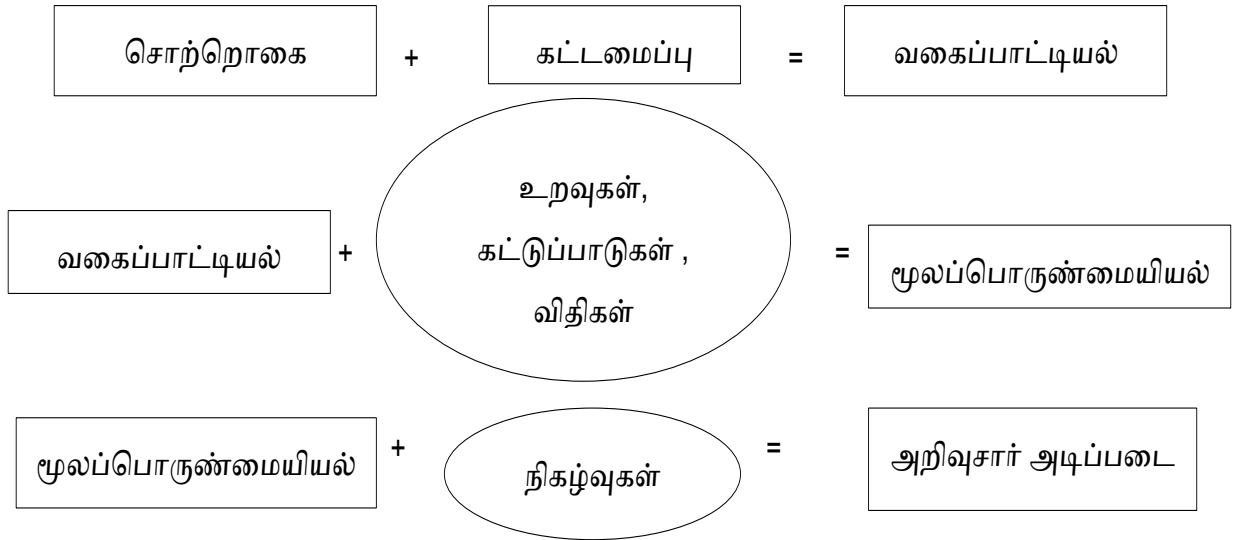
கோயம்புத்தூர்

## 1.முன்னுரை

சித்த மருத்துவம் என்பது தமிழ்நாட்டில் தோன்றிய பாரம்பரிய மற்றும் பயனுள்ள மருத்துவ முறைகளில் ஒன்றாகும். இந்த முறை இந்தியாவில் மிகவும் பழமையான மற்றும் மிகவும் பிரபலமான மருத்துவ முறைகளில் ஒன்றாகக் கருதப்படுகிறது. சித்தர் என்ற சொல் உடல் மற்றும் ஆன்மீக நடைமுறைகளில் சிறந்து விளங்குபவர்களைக் குறிக்கிறது. அவர்கள் சித்த மருத்துவத்தின் ஆரம்ப பயிற்சியாளர்கள். பெரும்பாலான சித்த இலக்கியங்கள் தமிழில் எழுதப்பட்டவை; இதன் சரியான புரிதலுக்கும் பயனுள்ள பயிற்சிக்கும் தமிழ் மொழியில் புலமை தேவை. மனிதகுலத்தின் நல்வாழ்வுக்கு, குறிப்பாக நவீன வாழ்க்கை முறைக்கு சித்த மருத்துவ முறையில் சிறந்த நடைமுறைகள் மிகவும் தேவைப்படுகின்றன. இதற்கு சித்தா அமைப்பின் மூலம் தெரிவிக்கப்படும் அறிவின் சரியான புரிதல் மற்றும் சிறந்த பிரதிநிதித்துவம் தேவைப்படுகிறது, இதை இறுதிப் பயனர், இணையப் பயன்பாடுகள் மற்றும் தானியங்கு அறிவார்ந்த அமைப்புகளால் எளிதாக அணுகும்படி செய்ய இயலும். மேற்கூறிய தேவைகளுடன் இந்த பாரம்பரிய அறிவைப் பிரதிநிதித்துவப்படுத்துவதற்கான மிகச் சரியான வழிகளில் ஒன்று மூலப்பொருண்மைகளைப் பயன்படுத்துவதாகும். மூலப்பொருண்மை முறையான பிரதிநிதித்துவ முறைகளை மாற்றியமைக்கிறது, இதன் மூலம் ஒருவர் சித்தமருத்துவக் களத் தகவல்களை விளங்கக்கூடிய நிலையில் பொருண்மை மயக்கம் இன்றி பகிர்ந்து கொள்ள முடியும். இந்த ஆய்வானது தமிழ் மொழியில் கட்டமைக்கப்பட்ட ஒரு சித்தமருத்துவக் கள மூலப்பொருண்மையை முன்மொழிகிறது. எந்தவொரு பயன்பாட்டிலும் சித்தமருத்துவ அறிவை அணுகுவதற்கும் விளக்குவதற்கும் இது உதவியாக அமையும்.

தனிப்பட்ட மூலப்பொருண்மையியல் அமைப்பு சமுதாயத்தை வல்லுநர்களின் அறிவாண்மையில் இருந்து விலக்கி வைக்கும். எனவே சமூகம் பகிர்ந்து கொள்ளும் அறிவைப் பயன்படுத்தி சுகாதாரக் களத்தின் மூலப்பொருண்மையியல் கட்டமைப்பை உருவாக்குவதற்கு இங்கு முன்மொழியப்பட்டுள்ளது. முன்மொழியப்பட்ட அறிவு அடிப்படையிலான அமைப்பை உருவாக்கும் போது சமூகத்தால் பகிரப்படும் அறிவு கணக்கில் எடுத்துக்கொள்ளப்படும், சமூக மூலப்பொருண்மையியல் அமைப்பு சுகாதார களத்தின் பிரதிநிதித்துவத்தில் மிகவும்

பயனுள்ளதாக இருக்கும். சமூக-மொழியியல் அமைப்பு மற்றும் இயற்கை மொழி செயலாக்கம் (NLP) இதற்கு பங்களிக்க முடியும். சமூகத்தின் கண்ணோட்டத்தையும் ஆர்வத்தையும் உள்ளடக்கிய உடல்நலக் களத்தின் தமிழ்ப் பிரதிநிதித்துவம் இன்றைய தேவை. ஆங்கிலம் பொது மூலப்பொருண்மையியல் கட்டமைப்பையும், சுகாதார களம் உட்பட பிற குறிப்பிட்ட மூலப்பொருண்மைக் கட்டமைப்பையும் உருவாக்கியுள்ளது. ஆனால், சித்தமருத்துவ அறிவை அடிப்படையாகக் கொண்டு சுகாதாரக் களத்தில் சமூக-மூலப்பொருண்மையியல் கட்டமைப்பை உருவாக்கும் தற்போதைய முன்மொழிவு ஒரு புதிய முயற்சியாகும். மூலப்பொருண்மையியல் என்பது ஒரு பொருண்மைக் களத்தில் உள்ள சொல் அலகுகளை சொல்சார் மற்றும் பொருண்மைசார் உறவுகள் மூலம் இணைத்து, பகிர்ந்துகொள்ளும் முறையின் அடிப்படையில் அமையும். மூலப்பொருண்மையியல் அமைப்பை பின்வருமாறு வெளிப்படுத்தலாம்.



ஏன் ஒருவர் மூலப்பொருண்மையியலை உருவாக்க வேண்டும் என்ற கேள்வி எழலாம். பின்வருவன சில காரணங்களாகும்:

- மக்களுக்கும் அல்லது மென்பொருள் செயலிகளுக்கும் இடையில் தகவலின் அமைப்பின் பொதுவான புரிதலை பங்கிட்டுக்கொள்ள
- பொருண்மைக்கள அறிவை மீளப்பயன்படுத்த இயலச்செய்ய
- பொருண்மைக்கள ஊகங்களை வெளிப்படைப்படுத்த
- செயற்படுத்தும் அறிவிலிருந்து பொருண்மைக்கள அறிவைப் பிரிக்க
- பொருண்மைக்கள அறிவை ஆய

மூலப்பொருண்மையியல் ஒரு பொருண்மைக்களத்தில் தகவல்களைப் பங்கிட்டுக்கொள்ளத் தேவைப்படும்/விரும்பும் நபர்களுக்கு ஒரு பொதுவான சொற்றொகையை வரையறை விளக்கம் செய்கிறது.

## 2. முந்தைய ஆய்வுகள்

CYC மற்றும் IEEE SUMO போன்ற மூலப்பொருண்மையியல்கள் மற்றும் மூலப்பொருண்மை அமைப்புகளை உருவாக்குவதில் சில குறிப்பிடத்தக்க முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. Upper CYC திட்டம் பொதுவான அறிவின் 3000 மிக முக்கியமான கருத்துகளைப் படம்பிடித்து, மூலப்பொருண்மையை உருவாக்குகிறது. SUMO (Standard Upper Ontology) திட்டம் உயர் நிலை, மெட்டா, தத்துவம் மற்றும் பொதுவான வகையான கருத்துகளைப் படம்பிடிப்பதில் உள்ள சிக்கலை நிவர்த்தி செய்து அவற்றை இன்னும் விரிவாக முன்வைக்கிறது. சில குறிப்பிட்ட களம் சார்ந்த மூலப்பொருண்மையியல்களும் இந்த வடிவவாதங்களைப் பயன்படுத்தி உருவாக்கப்பட்டுள்ளன. இந்த முந்தைய முயற்சிகளில் இருந்து, பெரிய அளவிலான மூலப்பொருண்மைகளை உருவாக்குவது ஒரு புதிய சிக்கல்களை உருவாக்குகிறது என்பதை நாம் அறிவோம், குறிப்பாக மூலப்பொருண்மையியல்கள் உறைந்த வளங்களைக் காட்டிலும் "நேரடி களஞ்சியங்களாக" பார்க்கப்படும். எடுத்துக்காட்டாக, எந்தவொரு பெரிய அளவிலான மூலப்பொருண்மையியல் வடிவமைப்புக் குழுக்களும் பின்வரும் முக்கிய சிக்கல்களைக் கருத்தில் கொள்ள வேண்டும்: தற்போதுள்ள மூலப்பொருண்மையியல் பின்னலமைப்பில் புதிய களங்களைச் சேர்ப்பது, அறிவு வடிவமைப்பை மாற்றுவது, செயல்திறன் சிக்கல்கள் மற்றும் சேவையின் உத்தரவாதத் தரம், அளவிடுதல் மற்றும் எந்த மாற்றங்களையும் செய்வதில் எளிமை.

நைடா (1975) முன்மொழியப்பட்ட பொருளின் கூறு பகுப்பாய்வுக் கோட்பாட்டின் அடிப்படையில் உருவாக்கப்பட்ட தமிழ் சொற்களஞ்சியத்திற்காக (1982, 2001) ராஜேந்திரன் ஒரு மூலப்பொருண்மையியல் சொற்களஞ்சியத்தை உருவாக்கியுள்ளார், இது இந்திய பாரம்பரியமான நிகண்டு மற்றும் அரிஸ்டாட்டிலிய மரபுகள் மற்றும் இனங்கள் ஆகியவற்றிற்கு ஏற்றவாறு மேம்படுத்தப்பட்டது. ராஜேந்திரன் (2001) சொற்களஞ்சியத்தால் குறிப்பிடப்படும் கருத்துருக்களை நான்கு வகைகளாக வகைப்படுத்தியுள்ளார்: நைடாவைப் (1975) பின்பற்றி இவர் சொற்களை/கருத்துருக்களை இருப்புப்பொருள்கள், அருவங்கள், நிகழ்வுகள் மற்றும் தொடர்புகள் எனப் பகுத்துள்ளார். இருப்புப்பொருள்கள் உறுதியான கருத்துகளின் குறிப்பு அர்த்தங்களைக் கொண்டிருக்கின்றன, நிகழ்வுகள் முக்கியமாக வினைச்சொற்கள் மற்றும் வினைப் பெயர்ச்சொற்களைக் கொண்டுள்ளன மற்றும் அருவங்கள் முக்கியமாக பெயரடைச் சொற்கள் மற்றும் வினையடைச்சொற்கள் ஆகியவற்றைக் கொண்டிருக்கும். தொடர்புபடும் செயற்பாட்டுச் சொற்கள், முன்னொட்டுகள், பின்னொட்டுக்கள், இணைப்பான்கள் போன்றவற்றை உட்படுத்தும்.

## 3. சொல்சார் மற்றும் பொருண்மைசார் உறவுகளால் சித்தமருத்துவச் சொற்றொகையைக் கட்டமைத்தல்

### சொல்சார் உறவுகள்

தமிழ் சொற்களஞ்சியத்தின் மூலப்பொருண்மையியல் கட்டமைப்பில், சொல் அலகுகளை ஒன்றுடன் ஒன்று இணைக்க அல்லது தொடர்புபடுத்தக்கூடிய குறைந்தபட்சம் நான்கு சொல்சார் அல்லது அர்த்த உறவுகள் உள்ளன.

ஒருபொருள்பன்மொழி: மருத்துவர்: வைத்தியர்

உள்ளடங்குமொழி-உள்ளடக்குமொழி: மருத்துவம்: சித்தமருத்துவம்

சினைமொழி-முழுமொழி: கை, கால்: உடல்

இசைவு: கண்ணோய்-கண்சிவப்பு

இசைவின்மை: சித்தமருத்துவம் – ஆயுர்வேதமருத்துவம்

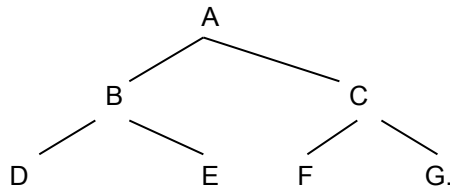
### சொல்சார் மரபுரிமை

உள்ளடக்குமொழி-உள்ளடங்குமொழி மற்றும் சினைமொழி-முழுமொழி ஆகியவை சொல் அலகுகள் பொருண்மைப் பண்புக்கூறுகளைப் பெறுவதை உறுதிப்படுத்துகின்றன.

சொற்றொகை இரண்டு வகையான கட்டமைப்பை அனுமதிக்கிறது: கிளை படிநிலை அமைப்பு, கிளை அல்லாத படிநிலை அமைப்பு. கிளைப் படிநிலை அமைப்பு இருவகைப்படும்: வகைப்பாட்டியல் படிநிலை அமைப்பு, முழு-சினைப் படிநிலை அமைப்பு. கிளையல்லாத படிநிலை அமைப்பு இரண்டு வகைப்படும்: தொடர் அமைப்பு, விகித அமைப்பு.

### கிளைப் படிநிலை அமைப்பு

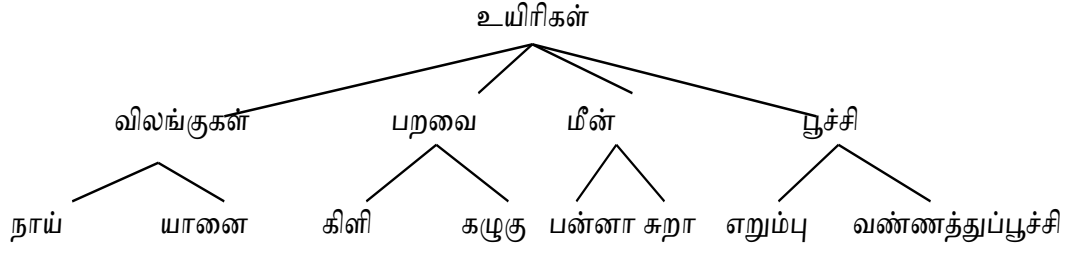
படிநிலையை இரண்டு உறவுகளின் அடிப்படையில் வகைப்படுத்தலாம்: ஆதிக்க உறவு, வேறுபாட்டு உறவு. A மற்றும் B, A மற்றும் C, B மற்றும் D மற்றும் B மற்றும் E, C மற்றும் F மற்றும் C மற்றும் G ஆகியவற்றுக்கு இடையே உள்ள உறவு ஆதிக்கத்தின் உறவு. B மற்றும் C, D மற்றும் E மற்றும் F மற்றும் G ஆகியவற்றுக்கு இடையே உள்ள உறவு வேறுபாட்டு உறவு



### வகைப்பாட்டியல் படிநிலை அமைப்பு

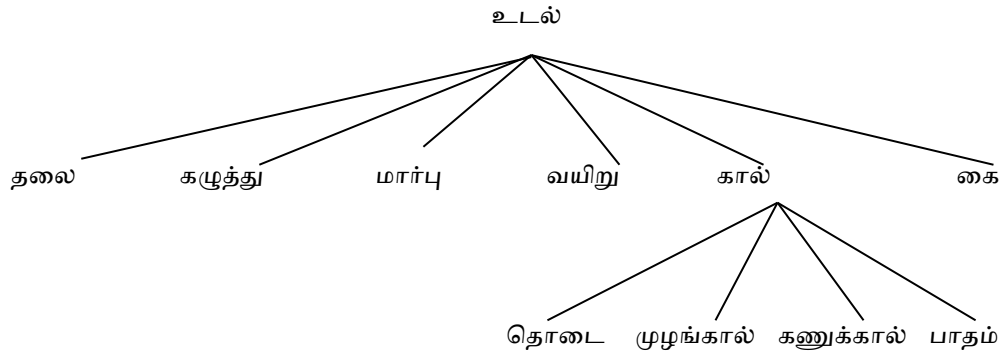
வகைப்பாட்டியல் படிநிலைகள், அடிப்படையில் வகைப்படுத்தும் அமைப்பாகும், மேலும் அவை ஒரு மொழி பேசுபவர்கள் உலக அனுபவத்தை வகைப்படுத்தும் விதத்தை பிரதிபலிக்கின்றன. நன்கு

உருவாக்கப்பட்ட வகைப்பாட்டியல் வெவ்வேறு நிலைகளில் உள்ள வகைகளின் ஒழுங்கான மற்றும் திறமையான தொகுப்பை வழங்குகிறது.



### முழு-சினை படிநிலை அமைப்பு

முழு-சினை அல்லது பாக உறவுகள் என்பது மூலப்பொருண்மையியல் உறவுகள் ஆகும், அவை எங்கும் நிறைந்த, வகைபிரித்தல் துணைத்தொகுப்பு உறவுகளாக அடிப்படையாகக் கருதப்படுகின்றன



### கிளையில்லா அமைப்பு

கிளையில்லா அமைப்பு பலவகைப்படும்.

#### வரிசை அமைப்பு:

கிளைகளின்றி ஒன்றுக்குப்பின் ஒன்றாக வரிசைப்படுத்தும் அமைப்பு வரிசை அமைப்பாகும்;. இது இரண்டுவகைப்படும்: இருதுருவ அமைப்பு, ஒருதுருவ அமைப்பு.

#### இருதுருவங்கள்

வரிசை அமைப்பில் எளிமையானது ஒரு இணை எதிர்மொழிகள் ஆகும். எதிர்மொழிகள் என்ற தலைப்பில் சொல்லப்பட்டதைத் தவிர, இந்தக் கட்டமைப்புகளைப் பற்றிச் சொல்வதற்கு பெரிதாக ஒன்றும் இல்லை.

#### இருதுருவச்சங்கிலிகள்

இருமுனை சங்கிலிகள் அளவின் ஒவ்வொரு முனையிலும் எதிர் துருவமுனைப்பின் மறைமுகமான மிகையான சொற்களைக் கொண்டுள்ளன. பின்வருவது இருதுருவச்சங்கிலிக்கு எடுத்துக்காட்டு ஆகும்:

{மிகநுண்ணிய, நுண்ணிய, சிறிய, பெரிய, மிகப்பெரிய}

#### ஒருதுருவச் சங்கிலிகள்

ஒருதுருவச் சங்கிலிகளில், சங்கிலியின் முனைகளில் உள்ள சொற்கள் எதிர் திசைகளில் நோக்கப்படுகின்றன என்பதில் எந்த அர்த்தமும் இல்லை. பின்வருபவை பல்வேறு வகையான ஒருதுருவச் சங்கிலிகள் ஆகும்:

1. அளபுகள்
2. நிலைகள்
3. அளவீடுகள்
4. தரவரிசைகள்
5. தொடர்வரிசைகள்

#### அளபுகள்

அளவு அல்லது தீவிரம் போன்ற தொடர்ச்சியான அளவிடப்பட்ட சில பண்புகளின் வெவ்வேறு அளவுகளை அவற்றின் அர்த்தத்தின் ஒரு பகுதியாக அளபுகள் இணைக்கின்றன, ஆனால் சேர்ப்பதில் எந்த தொடர்பும் இல்லை. பின்வருபவை எடுத்துக்காட்டுகள் ஆகும்:

{தோல்வி, வெற்றி, தனிமதிப்பு}

{பாறை, குன்று, மலை, மாமலை}

#### நிலைகள்

நிலைகள் என்பது ஏதாவது ஒரு வாழ்க்கைச் சுழற்சியில் புள்ளிகள் மற்றும் பொதுவாக முன்னேற்றம் என்ற கருத்தை உள்ளடக்கியது. பின்வருபவை எடுத்துக்காட்டுகள்:

{பட்ட முன்படிப்பு, இளநிலைப் பட்டம், முதுநிலை பட்டம், முனைவர் பட்டம்}

{முட்டை, லார்வா, பியூபா, வண்ணத்துப்பூச்சி}

#### அளவீடுகள்

அளவீடுகள் ஒரு பகுதி-முழு உறவை அடிப்படையாகக் கொண்டவை, ஒவ்வொன்றும் ஒரே எண்ணிக்கையிலான ஒரே பகுதிகளாகப் பிரிக்கப்படுகின்றன: பொதுவாக அருகிலுள்ள சொற்களால் குறிக்கப்பட்ட அளவிடப்பட்ட பண்புகளின் மதிப்புகளுக்கு இடையே ஒரு வடிவியல் உறவு உள்ளது.

{செகண்டு, மினிட், மணி, நாள், வாரம், மாதம், ஆண்டு}

{மில்லிமீட்டர், சென்டிமீட்டர், மீட்டர்}

{இஞ்சு, அடி, கஜம், பர்லாங்கு, மைல்}

{ஆழாக்கு, உழக்கு, பக்கா, மரக்கால்}

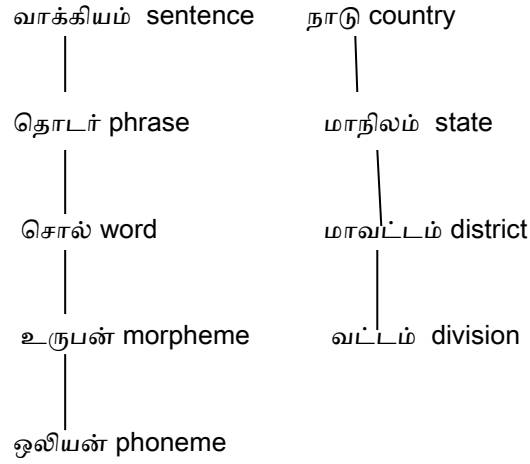
## வரிசைகள்

வரிசைகளும் வரிசைப்படுத்தப்பட்ட சொற்கள் ஆகும், ஆனால் அண்டை சொற்களைப் பொறுத்தவரை 'அதிகமாகவோ அல்லது குறைவாகவோ' எந்த பண்பும் இல்லை.

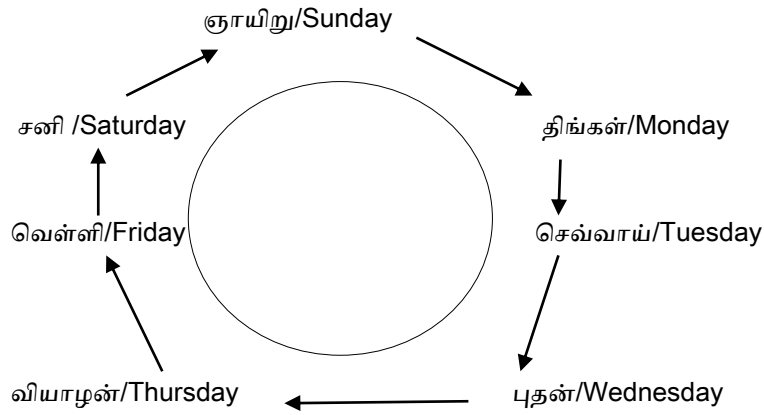
{ஒன்று, இரண்டு, மூன்று, ..., }

{இறந்தகாலம், நிகழ்காலம், எதிர்காலம்}

{கார்காலம், கூதிர்காலம், முன்பணிக்காலம், பின்பணிக்காலம், இளவேனிற்காலம், முதுவேனிற்காலம்}



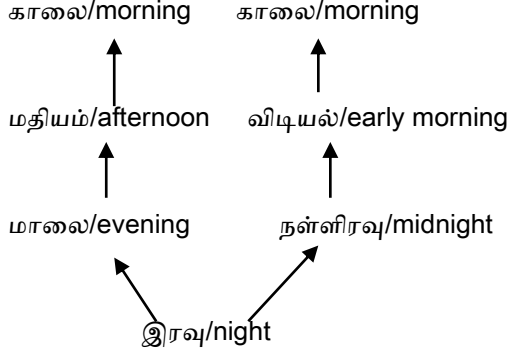
## சுற்றுக்கள்



## சுருளிகள்

வரிசை மற்றும் சுற்றுக்களின் கலவையானது சுருளிகளின் வரையறுக்கும் பண்புகளாக எடுத்துக்கொள்ளப்படலாம். சுருள் இணைப்புகள் பொதுவாக காலங்களைக் குறிக்கின்றன.





#### 4. சித்த மூலப்பொருண்மையியல் அமைப்பு உருவாக்கத்திற்கான முன்னேற்பாடுகள்

தரவுகளை சேகரிக்க வார்த்தை உட்பொதிக்கும் நுட்பங்கள் (Word embedding techniques to collect data)

சித்த மருத்துவம் பற்றிய தரவுகளை சேகரிக்க வார்த்தை உட்பொதிக்கும் நுட்பங்களைப் பயன்படுத்துகிறோம்.

வேர்ட் எம்பெடிங் என்பது சொல் பிரதிநிதித்துவத்தின் ஒரு நுட்பமாகும், இது இயந்திர கற்றல் வழிமுறைகளால் ஒத்த பொருளைக் கொண்ட சொற்களைப் புரிந்துகொள்ள அனுமதிக்கிறது. தொழில்நுட்ப ரீதியாக, இது நரம்பியல் நெட்வொர்க், நிகழ்தகவு மாதிரி அல்லது சொல் இணை நிகழ்வு மேடிக்ஸில் பரிமாணக் குறைப்பு ஆகியவற்றைப் பயன்படுத்தி உண்மையான எண்களின் திசையன்களாக வார்த்தைகளை மேப்பிங் செய்வதாகும்.

பல்வேறு மொழி மாதிரிகளைப் பயன்படுத்தி சொல் உட்பொதிப்பைக் கற்றுக்கொள்ளலாம். எடுத்துக்காட்டாக, 'நாய்' வெக்டரால் [0.75, 0.22, 0.66, 0.97] குறிப்பிடப்படும். அகராதியில் உள்ள அனைத்து சொற்களும் இவ்வாறு குறியாக்கம் செய்யப்பட்டால், வார்த்தைகளின் திசையன்களை ஒன்றோடொன்று ஒப்பிடுவது சாத்தியமாகும், எடுத்துக்காட்டாக, திசையன்களுக்கு இடையே உள்ள கோசைன் தூரம் அல்லது யூக்ளிடியன் தூரத்தை அளவிடுவதன் மூலம்.

வார்த்தைகளின் நல்ல பிரதிநிதித்துவம், 'சாக்கர்' அல்லது 'இயந்திரம்' என்ற சொல்லை விட 'செல்லம்' என்ற வார்த்தை 'நாய்' என்ற சொல்லுக்கு நெருக்கமாக இருப்பதைக் கண்டறிய முடியும். எனவே, உட்பொதிக்கப்பட்ட வெக்டார் ஸ்பேஸில், ராஜா-ஆண்+பெண் = ராணி அல்லது லண்டன்-இங்கிலாந்து+இத்தாலி = ரோம் என்ற சமன்பாடு கூட இருக்கும் என்று நம்புவதற்கு இந்தப் பிரதிநிதித்துவங்கள் நம்மை அனுமதிக்கின்றன.

சித்த மருத்துவ அமைப்பு பின்வரும் பல்வேறுபட்ட கிளைகளைக் கொண்டுள்ளது

குணபாடம் (Pharmacology),

நஞ்சுமருத்துவம் (Toxicology),

நோய்நாடல் (Pathology),

மருத்துவம் (General Medicine),

சூல் மற்றும் மகளிர் மருத்துவம் (Obstetrics and Gynaecology),

குழந்தை மருத்துவம் (Paediatrics),

அறுவை மருத்துவம் (Surgery),  
 தோல் மருத்துவம் (Dermatology),  
 கண், மூக்கு, தொண்டை மருத்துவம் (E. N. T),  
 கண்மருத்துவம் (Ophthalmology),  
 கிரிகை நோய்மருத்துவம் (Psychiatry),  
 வர்மம் (Pressure Manipulation Therapy)  
 புறமருத்துவம் (External Therapy),  
 முதியோர் மருத்துவம் (Geriatrics)  
 காய கற்பம் (Rejuvenation therapy).

சித்த மருத்துவ ஒழுங்கமைப்பு நான்கு முதன்மையான பிரிவுகளைக் கொண்டுள்ளது,

1. வாதம்/இரசவாதம் (Chemistry/Alatrochemistry Alchemy),
2. வைத்தியம் (Treatment),
3. யோகம் (Yogic Practices)
4. ஞானம் (Wisdom)

அடிப்படைக் கோட்பாடுகள் - 96 தத்துவம்

சித்த முறைப்படி மனிதன் தத்துவம் எனப்படும் அடிப்படைக் கோட்பாடுகள் / கருவிகளால் கட்டமைக்கப்படுகிறான், அவை 96 எண்ணிக்கையில் உள்ளன. அவை மனித உடலின் அடிப்படை செயல்பாடுகளைக் கையாளும் அறிவியலாகக் கருதப்படுகின்றன. இந்த 96 கொள்கைகள் உடல், செயல்பாட்டு, உளவியல் மற்றும் அறிவுசார் கூறுகளை உள்ளடக்கியது.

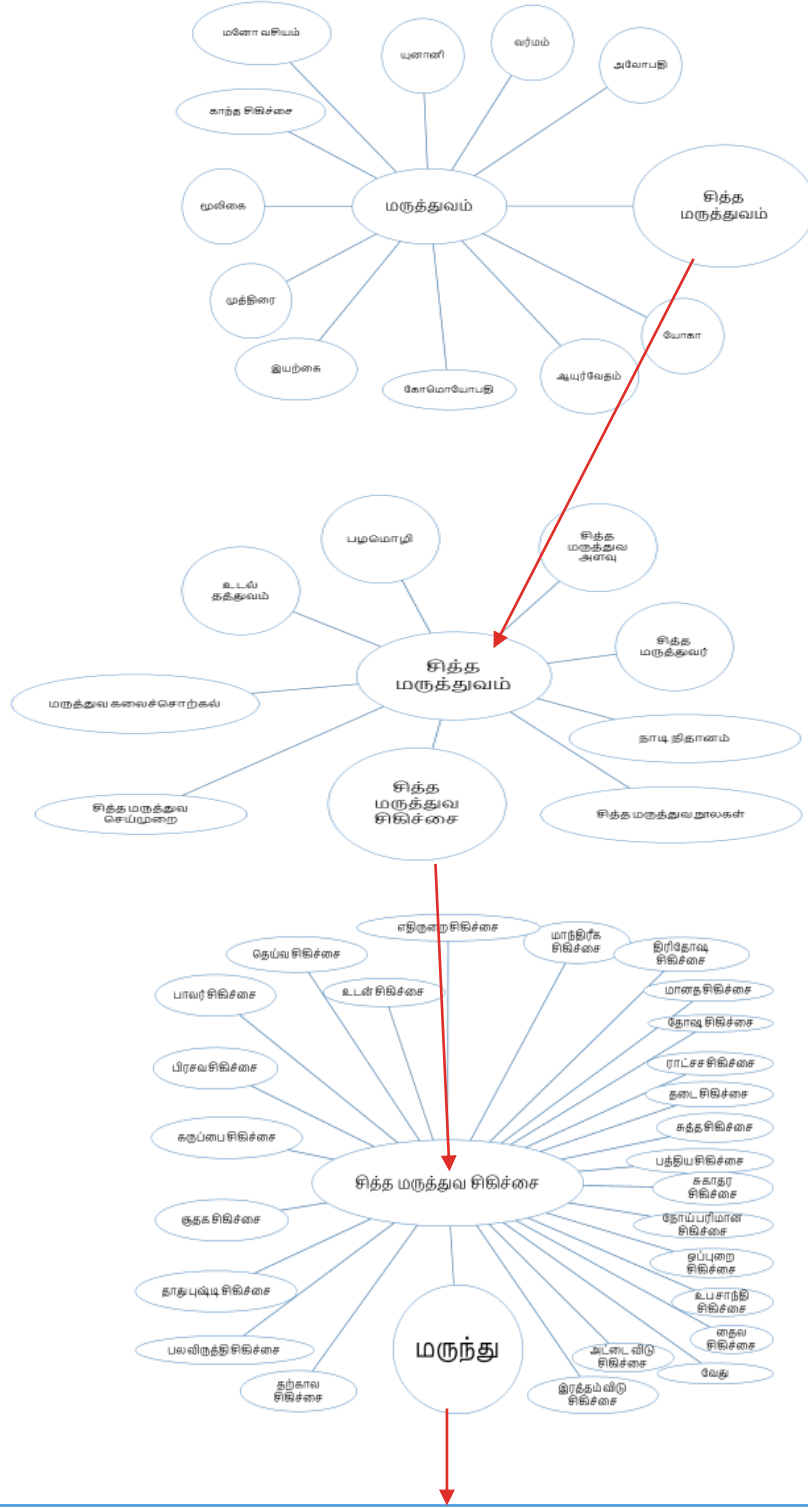
அகக்கருவிகள் 36

புறக்கருவிகள்-60

நோய் அறிதல் (Diagnosis)

1. நாடி பரிசோதனை (Siddha way of Pulse examination)
2. பரிசுப் பரிசோதனை (Touch and palpation)
3. நாப் பரிசோதனை(Tongue examination)
4. நிறப் பரிசோதனை (Colour, complexion, discolouration etc.)
5. மொழிப் பரிசோதனை (Voice examination)
6. விழிப் பரிசோதனை(Eyes examination)
7. மலப் பரிசோதனை (Stools examination)
8. மூத்திரப் பரிசோதனை (Urine examination)
- a. Urine Sign
- b. நெய்க்குறி (Oil on Urine Sign)

பின்வருவது சித்தமருத்துவ மூலப்பொருண்மையின் ஒரு எடுத்துக்காட்டாகும்.



விளக்கம்: சித்த மருத்துவத்தில் தொன்றுதொட்டு உள்ளாட்சி மருந்துகள் 32 வகைகள் எனவும் வெளியாட்சி மருந்து 32 வகைகள் எனவும் கூறப்பட்டுள்ளன. இவையே அகமருந்தென்றும் புறமருந்தென்றும் கூறப்படும்.

## 5. முடிவுரை

சித்த மருத்துவ களத்திற்கான முன்மொழியப்பட்ட மூலப்பொருண்மையியல் கட்டமைப்பு அடிப்படையிலான அறிவு அமைப்பு மருத்துவ நடைமுறையில் ஈடுபடுபவர்களுக்கு பெரும் உதவியாக இருக்கும். சித்த களத்தின் முன்மொழியப்பட்ட சமூக-மூலப்பொருண்மையியல்-அறிவு அடிப்படையிலான கட்டமைப்பின் மூலம் பொது மக்களும் பயனடையலாம்.

#### **நோக்கிட்டு நூல்கள்**

Busa, F., Calzolari, N., and Lenci, A. 2001. "Generative Lexicon and SIMPLE Model: Developing Semantic Resources for NLP." In: Bouillon P. and Busa, F. (ed.) The language of Word Meaning, Cambridge University Press, Cambridge, pp. 333-362.

Busa, F., Calzolari, N., Lenci, A. and J. Pustejovsky. 2001. "Building a Semantic Lexicon: Structuring and Generating Concepts." In: H. Bunt et al. (ed.) Computing meaning, Volume 2, 29-51, Kluwer Academic publishers, Netherlands.

Fellbaum, Christiane (Ed). WordNet: An electronic lexical database, MIT Press, 1999.

Lenat, D. B. and R. V. Guha. Building Large Knowledge Based Systems. Reading, Massachusetts: Addison Wesley, 1990

Nida, E.A. 1975a. Componential Analysis of Meaning: An Introduction to Semantic Structure. Mouton, The Hague.

Nida, E.A. 1975.b. Exploring Semantic Structure. The Hague: Mouton.

Rajendran, S. 2001. taRkaalat tamizc coRkaLanjiciyam [Thesaurus for Modern Tamil]. Tamil University, Thanjavur.

Rajendran, S. 1995. "Towards a Compilation of a Thesaurus for Modern Tamil". South Asian Language Review. 5.1:62-99.

Vasudeva Varma. Building Large Scale Ontology Networks. Downloaded from internet.

## கல்வியில் விளையாட்டுகள்: தமிழ்க் கற்றல் கற்பித்தலில் சொல் விளையாட்டுகளை ஒரு துணைக் கருவியாகப் பயன்படுத்துதல்

முகிலன் முருகன் - Muhelen Murugan  
OMTAMIL DOTCOM, Malaysia

### சுருக்கம்

குறிச்சொற்கள்:

அ கல்வித் தொழில்நுட்பம்  
ஆ விளையாட்டு செயலி  
இ சொல் கிடங்கு  
ஈ திறமூல இயக்குதலம்

செயலி கருவிகளின் பயன்பாடு தற்போதைய கல்வி சூழலில் அதிகளவில் வளர்ந்து வருவது கல்வியாளர்களிடையும் கல்விக் கழகங்களிடையும் மாபெரும் கவனத்தை ஈர்த்துள்ளது. கல்வியில் விளையாட்டு என்பது கற்றலுக்கு இடையூறாக பார்க்கப்படலாம். ஆனால் கல்வியில் அதன் பங்கு மாணவர்களின் ஊக்கத்தையும் ஈடுபாட்டையும் பல மடங்கு அதிகரிப்பதுடன் அவர்களின் காட்சி, ஒருங்கிணைப்பு திறன்களை நண்பர்களுடன் மேம்படுத்த உதவுவதுதான் உண்மை நிலைமை. கற்றல் கற்பித்தலை எளிமைப்படுத்தவும் மேம்படுத்தவும் கல்வித் தொழில்நுட்பம் பயன்படுத்தப்படுகிறது. ஆனால் கல்வியில் தொழில்நுட்பத்தைப் பயன்படுத்துவது மட்டுமே மாணவர்களின் திறன்களை நேரடியாக உயர்த்தாது. ஏனெனில் தொழில்நுட்பக் கருவிகள் பாடத்திட்டத்திற்கு உட்பட்டு இயங்க வேண்டும். தமிழ்க் கற்பித்தலில் சொல் விளையாட்டு செயலிகளைச் செயல்படுத்துவதும் பயன்படுத்துவதும் குறித்து இக்கட்டுரை விவாதிக்கிறது. 'ஓம்தமிழ்' திறமூல இயக்குதலத்திலுள்ள சொல் ஊகித்தல், சொல் தேடல், குறுக்கெழுத்து புதிர் மற்றும் சொல் அடுக்கு போன்ற விளையாட்டுகளின் வடிவமைப்பும் இந்த தாளில் விவாதிக்கப்படுகிறது.

Copyright © 2019 International Forum for Information Technology in  
Tamil.

### Corresponding Author:

MUHELEN MURUGAN,  
A-1-03, Cybersquare,  
Jalan Teknokrat 6, Cyberjaya,  
63000 Selangor Malaysia  
Email: muhelen@omtamil.com

## Corpus Analysis of ditransitive verb “koṭu” in Modern Tamil for Information Retrieval System

A.Murugaiyan<sup>1</sup> and

<sup>1</sup>Ecole Pratique des Hautes Etudes, Paris, a.murugaiyan@gmail.com

<sup>2</sup>Pondicherry University, dhanagiri@pondiuni.ac.in

This paper presents an analysis of “koṭu” (to give) a ditransitive verb in simple predicate constructions in Modern Tamil for building Information Retrieval System. . A “Ditransitive Construction is a construction with a verb denoting transfer of an entity (T) from an agent (A) to a recipient (R), such as *Kim gave Lee a box*” (Haspelmath 2015). Information retrieval is understood as the process of selecting relevant information from a collection of information sources to meet a specific information demand. Tamil is considered as a verb final language which permits a relatively free word order. Typologically a nominative-accusative language Tamil is potentially a differential object-marking language. On the other hand, the dative case encodes different arguments like experiencer, recipient, and beneficiary.

In this paper we examine 1) whether there is any hierarchical ordering of different participants / arguments like agent, theme, beneficiary, recipient and purposive, 2) how these different participants are encoded and 3) how these information would be integrated in the process of information retrieval in Tamil.

1. nāṇ rāmaṇukku puttakattukkukāka paṇam koṭuttēṇ (Agent-Beneficiary-Purposive-Theme-Verb)
2. nāṇ kōyilukku naṅkoṭaiyāka rāmaṇiṭam paṇam koṭuttēṇ. (Agent-Beneficiary-Purposive-Recipient-Verb)

### References:

- Haspelmath Martin. 2015. ‘Ditransitive Constructions’, *Annual Review of Linguistics*, 2015:1(1), 19-41.
- Heine, Bernd & König, Christa. 2007. On the linear order of ditransitive objects. Ms., University of Cologne.

**இயந்திர மொழிபெயர்ப்பிற்கான அகராதி உருவாக்கம்**  
(Lexicon for Machine Translation)

**முனைவர் ப. குமார்**  
உதவிப் பேராசிரியர்  
தமிழ்த்துறை  
தமிழ்நாடு மத்தியப் பல்கலைக்கழகம்  
திருவாரூர் – 610 005  
மின்னஞ்சல்: [pkumar@cutn.ac.in](mailto:pkumar@cutn.ac.in)

மொழிபெயர்ப்பிற்குப் பயன்படக்கூடிய முக்கியமான கருவிகளில் ஒன்று அகராதி ஆகும். அகராதி போன்றதொரு அமைப்பு மனித மூளையில் (*Mental Dictionary*) இருப்பதால்தான் மனித மொழிபெயர்ப்பு எளிமையாக நடைபெறுகிறது. கணினியின் வருகைப்பிறகு மொழியாய்வுக் கருவிகள் பலவும் உருவாக்கப்பட்டன. இக்கருவிகளை ஒன்றிணைத்து கணினி மொழிபெயர்ப்புகள் பல நடந்து வந்தாலும், தமிழில் முழுமையான கணினி மொழிபெயர்ப்பு இன்னும் வரவில்லை. இதுதொடர்பான ஆய்வுகள் பலவும் நடைபெற்று வருகின்றன. இதன் ஒரு பகுதியானது தான் இயந்திர மொழிபெயர்ப்பிற்கான அகராதி உருவாக்கம் என்பது. அந்த வகையில் இன்றைக்குக் கிடைக்கக்கூடிய அச்சு அகராதித் தரவுகளைக்கொண்டு கணினி மொழிபெயர்ப்புகளைச் செய்ய இயலாது என்பதையும், சொற்களின் மொழிப் பொருளைப் புரிந்து கொள்ளும் வகையில் கணினிக்கான அகராதி எவ்வாறு உருவாக்கப்பட வேண்டும் என்பதையும் இக்கட்டுரை விரிவாக ஆராய்கிறது.

# Migrating TamilPesu to Cloud based Deployment

Authors: Surendhar Ravichandran <[surendhar.r@proton.me](mailto:surendhar.r@proton.me)>, T. Shrinivasan <[tshrinivasan@gmail.com](mailto:tshrinivasan@gmail.com)>, Muthiah Annamalai [ezhillang@gmail.com](mailto:ezhillang@gmail.com)

## Abstract:

Open-Tamil project has expanded to provide its API as web-service via tamilpesu.us website since 2018 [1]. In this article we share the process of migrating the deployment of this API server through cloud based app-platform with a service providers thereby providing significant advantages to users of site like: secure https access, quick time from code commit to deployment, and ease of maintenance for the project developers. We propose these identifications as easier tools for maintenance and growth of Tamil web applications and cause for wider adoption in our community.

## References:

1. Syed Abuthahir et-al, "Growth and Evolution of Open-Tamil," Tamil Internet Conference (2018).



## Natural Language Resources in Tamil

### ABSTRACT

Today we are living in the world of communication. The world of communication interlinks everyone through its various media. In this aspect Computers play a major role by bringing the world under the user's fingertip. Grammar is the legal advocacy to the human art of communication. But learners get annoyed with the language rules and the old teaching methodology.

Interlinking the computer to the language through Natural language Processing (NLP) paves a way to solve this problem. The innovative NLP applications are used to generate language learning and teaching tools which enhance the teaching and learning of Grammar. In this paper we present the Grammar teaching tools for analyzing and learning character, word and sentence of Tamil Language. Tools like Character Analyzer for analyzing character, Morphological Analyzer and Generator and Verb Conjugator for the word level analysis and Parts of Speech Tagger, Chuker and Dependency parser for the sentence level analysis were developed using machine learning based technology. These tools are very useful for second language learners to understand the character, word and sentence construction of Tamil language in a non-conceptual way. General Terms Tamil grammar, Agglutinative language, Natural Language processing, Machine Translation.

**Keywords:** Grammar Learning and Teaching, Machine Learning, Character Analyzer, Morphological Analyzer and Generator, Verb Conjugator, Parts of Speech Tagger and Chuker, Dependency parser.

BY; DR, MOHAMMED AFSAL

Afsalkhanmon06@gmail.com

# Symmetries in Number Forms of Tamil and Dravidian Languages

Authors: Muthiah Annamalai [ezhillang@gmail.com](mailto:ezhillang@gmail.com)

## Abstract:

We propose the Tamil number forms are equivalent by isomorphism (single rule over all numbers to corresponding numeral) in Telugu, Kannada and Malayalam. The latter being almost indistinguishable from Tamil except for prosody; this is based on intuition of digit forms [1]. We further contend that algorithm for generating numerals in any of the four languages are structurally identical due to the equivalence of numerals in a abstract way. We propose common algorithm for generating and parsing number forms [2] in these languages to/from text and into audio TTS generation. These can be used in various applications like token-queue systems, spoken calculators, etc.

## References:

1. Wikipedia on Tamil Numeral Influence, [https://en.wikipedia.org/wiki/Tamil\\_numerals#Influence](https://en.wikipedia.org/wiki/Tamil_numerals#Influence) (accessed Nov 14, 2022)
2. M. Annamalai, S. Mahadevan, "Generation and Parsing of Number to Words in Tamil," Tamil Internet Conference, 2020.

## Linguistic issues in machine translation in Tamil

### தமிழ்மொழியில் இயந்திர மொழிபெயர்ப்பில் மொழியியல் சிக்கல்கள்

Selvajothi Ramalingam  
Faculty of Languages and Linguistics  
Universiti Malaya, Kuala Lumpur, Malaysia

#### ஆய்வுச் சுருக்கம்

#### Keywords/ கருச்சொற்கள்:

இயந்திர மொழிபெயர்ப்பு  
கூகுள் மொழிபெயர்ப்பு  
மனித மொழிபெயர்ப்பு  
தமிழ்மொழி

மொழிபெயர்ப்பு என்பது ஒரு மொழியில் வழங்கப்பட்ட தகவல்களை வேறு மொழிபேசுவோர் அறிந்து கொள்ளும் வகையில் அந்தக் குறிப்பிட்ட மொழிக்கு மாற்றி அமைக்கும் நடவடிக்கையாகும். இந்த நவீன யுகத்தில் மொழிபெயர்ப்பு நடவடிக்கையானது தொழில்நுட்பம் துணைகொண்டு மிகவும் எளிதாகச் செய்யக்கூடிய ஒன்றாக அமைந்திருக்கிறது. அவ்வகையில் இயந்திர மொழிபெயர்ப்பானது குறிப்பாக, கூகுள் மொழிபெயர்ப்பு (*google translation*), மொழிபெயர்ப்பு நடவடிக்கைக்கு மிகப்பெரிய உந்துதலாக அமைகின்றது. கூகுள் மொழிபெயர்ப்பில், இதுவரை 134 மொழிகளில் இலவசமாக மொழிபெயர்க்கக்கூடிய வசதி அமைந்துள்ளது. இருப்பினும் இந்தக் கூகுள் மொழிபெயர்ப்பு இன்னும் முழுமையான; தரம்வாய்ந்த ஒன்றாக அமையவில்லை என்பது இயந்திர மொழிபெயர்ப்பின் சிக்கலாக அமைந்திருக்கிறது. ஆனால் இச்சிக்கல் எல்லா மொழிகளுக்கும் ஒரே நிலையில் அமைவதில்லை. சில மொழிகளில் மொழிபெயர்ப்புத் தரம் சிறப்பாகவும், சில மொழிகளில் மொழிபெயர்ப்புத் தரம் மோசமாகவும் அமைந்திருப்பதைக் காணமுடிகிறது. இந்த ஆய்வானது கோவிட்-19 பெருந்தொற்றுக் காலகட்டத்தில் மலேசிய நாட்டில் வெளிவந்த மலாய், ஆங்கில நாளிதழ்களில் கோவிட்-19 பெருந்தொற்றுத் தொடர்பான செய்திகளை மொழிபெயர்க்கும் போது ஏற்பட்டுள்ள மொழியியல் சிக்கல்களை அடையாளம் கண்டுள்ளது. அவ்வகையில் குழப்பமான வாக்கிய அமைப்பு, பொருத்தமற்ற சொற்பயன்பாடு, தவறான பதிலிடுபெயர்கள், இலக்கணப் பிழைகள் போன்றவை இயந்திர மொழிபெயர்ப்பின் சிக்கல்களாக இவ்வாய்வில் வெளிக்கொணரப்பட்டுள்ளன. மேலும் இயந்திர மொழிபெயர்ப்பு வெறும் நேரடி பொருண்மையை மட்டுமே மொழிபெயர்த்துள்ளதே தவிர சூழலியல் பொருண்மை மொழிபெயர்ப்பில் இடம்பெறவில்லை என்பதுவும் கண்டறியப்பட்டுள்ளது. ஆக, இயந்திர மொழிபெயர்ப்பில் மொழிபெயர்க்கப்பட்ட பனுவல் அல்லது தகவல் மொழிபெயர்ப்புக்கான இலக்கை அடையாமல் அதன் தரத்தை பாதிக்கின்ற வகையில் அமைந்துள்ளதை நாம் அறிய முடிகிறது.

Copyright © 2019 International Forum for Information Technology in Tamil.  
All rights reserved.

#### Corresponding Author:

Selvajothi Ramalingam,  
Department of Malaysian Languages and Applied Linguistics,  
Faculty of Languages and Linguistics, Universiti Malaya, Kuala Lumpur, Malaysia  
Email: selvajothi@um.edu.my

#### 1. முன்னுரை

மொழிபெயர்ப்பு என்பது ஒரு மூல மொழியிலிருந்து மற்றொரு மொழிக்கு (இலக்கு மொழிக்கு) கருத்தை மாற்றும் செயலாகும் (*Derrida & Venuti, 2001*). தகவல் தொடர்புச் சாதனங்களான வானொலி, தொலைக்காட்சி, செய்தித்தாள், இணையம், போன்ற அனைத்து ஊடகங்களும் மொழிபெயர்ப்புப் பணிக்குப் பயன்படுத்தப்படுகின்றன. இன்றைய சூழலில் பெரும்பாலான நாடுகள் ஒருமொழி பேசும் மக்கள் கொண்ட நாடாக இருப்பதில்லை. குறைந்தபட்சம் இரண்டு மொழிகளாவது பயன்பாட்டில் இருக்கிறது. அதாவது அந்த நாட்டு மக்களின் பெரும்பான்மையினரின் தாய்மொழி அல்லது தேசிய மொழியும் அனைத்துலக மொழியான ஆங்கிலம் பயன்பாட்டு மொழியாகவும் கல்வி மொழியாகவும் இருக்கிறது (*Mukadam, Sommerlad & Livingston, 2017*). இவ்வாறான பன்மொழிச்சூழலில் மொழிபெயர்ப்பு நடவடிக்கையானது முக்கியம் வாய்ந்த ஒன்றாக அமைகிறது. ஒரு மொழியில் வழங்கப்பட்ட தகவல்கள் இன்னொரு மொழி பேசும் சமூகத்தினர் புரிந்து கொள்ளும் வகையில் இந்த மொழிபெயர்ப்பு நடவடிக்கை மேற்கொள்ளப்படுகிறது. இது தொடர்பாக முந்தைய காலங்களில்

மொழிபெயர்ப்பு நடவடிக்கையின்போது அகராதியும் சொற்களஞ்சியமும் பெருந்துணையாக இருந்தது. ஆனால் தற்போதைய தொழில்நுட்ப வளர்ச்சி காரணமாகக் கணினி, இணையம், செயலிகள் பெருந்துணையாக இருந்து வருகிறது. இவற்றின் வாயிலாக மிக விரைவில் மொழிபெயர்ப்பைச் செய்துவிடும் சூழலும் அமைகிறது. அவ்வகையில் இயந்திர மொழிபெயர்ப்பானது குறிப்பாக கூகுள் மொழிபெயர்ப்பு, மொழிபெயர்ப்பு நடவடிக்கையை இன்னும் துரிதப்படுத்துகின்ற செயலாகப் பார்க்க முடிகிறது. இந்தக் கூகுள் மொழிபெயர்ப்பில் இதுவரை மொத்தம் 134 மொழிகள் இடம்பெற்றுள்ளன (<https://translate.google.com.my>). இவ்வாறான இயந்திர மொழிபெயர்ப்பில் தட்டச்சு செய்து மொழிபெயர்ப்பதும் பேச்சுரை வழி மொழிபெயர்ப்பதும் மொழிபெயர்ப்புப் பணியை மேலும் துரிதப்படுத்துகிறது.

இயந்திர மொழிபெயர்ப்பு (machine translation) என்பது ஒரு வகை தானியங்கி மொழிபெயர்ப்பு ([https://en.oxforddictionaries.com/definition/machine\\_translation](https://en.oxforddictionaries.com/definition/machine_translation)). அதாவது கணினியின் துணைகொண்டு இணையம்வழி மொழிபெயர்ப்புப் பணிகள் மொழிபெயர்ப்பாளர்களால் செய்யப்படுவது. இது அவர்களின் மொழிபெயர்ப்புப் பணிகளைத் துரிதப்படுத்த உதவுகின்றது. இன்றைய தகவல் யுகத்தில் தகவல்களை விரைந்து பெறுவதற்கும் மொழிபெயர்ப்புப் பணியை விரைவுபடுத்துவதற்கும் கூகுள் மொழிபெயர்ப்பு உதவுகிறது (Nur Naimah Akmar & Harun, 2016). இயந்திர மொழிபெயர்ப்பு மொழிபெயர்ப்பாளர்களுக்குப் பல நன்மைகளைக் கொண்டிருந்தாலும், மனித ஆற்றலோடு ஒப்பிட இயலாது என்பது நிதர்சனமான உண்மை. சிக்கலான மொழி அமைப்பின் காரணமாகத் தரமான மொழிபெயர்ப்புகளை உருவாக்குவதில் மனித நிபுணத்துவம், பண்பாட்டு பின்புல அறிவு, சுற்றுச்சூழல், மொழியியல் பண்புகள் உட்பட பல்வேறு காரணிகள் சிறந்த மொழிபெயர்ப்புக்குத் துணைபுரிகின்றன (Wan Rose Eliza, 2013). இயந்திர மொழிபெயர்ப்பில் மொழிபெயர்பெயர்க்கப்பட்ட பனுவல் பொதுவாக நேரடி மொழிபெயர்ப்பாகவே இருக்கிறது. இதனால் மொழிபெயர்ப்பில் குழப்பமான தகவல்களும், பொருள்மயக்கமும், இலக்கணப் பிழைகளும் அதிகமாகவே காணப்படுகிறது (Haroon & Daud, 2017). இச்சிக்கலை அடிப்படையாகக் கொண்டு இந்த ஆய்வு மேற்கொள்ளப்பட்டது. இந்த ஆய்வின் நோக்கமானது தமிழ்மொழியில் இயந்திர மொழிபெயர்ப்பு குறிப்பாகக் கூகுள் மொழிபெயர்ப்பில் காணப்படும் மொழியியல் சிக்கல்களை அடையாளம் காண்பதாகும். இச்சிக்கல்களைக் கண்டறிந்து சீர் செய்யப்படும்போது அதன் மொழிபெயர்ப்புத் தரம் மேலோங்க வாய்ப்புள்ளது. இங்குதான் மொழிபெயர்ப்பாளரின் பங்கும் தேவைப்படுகிறது. ஆக இயந்திர மொழிபெயர்ப்புச் செய்தாலும் மனித நிபுணத்துவமும் இணைந்து மொழிபெயர்க்கும்போது மொழிபெயர்ப்புப் பணி துரிதமாகவும் தரமாகவும் அமைய சூழல் ஏற்படும் (Nur Naimah Akmar & Harun, 2016).

மொழிபெயர்ப்பு என்பது உலகம் முழுவதும் வேகமாக வளரத் தொடங்கிய ஒரு துறையாகும். மொழிபெயர்ப்பு நடவடிக்கையின் வழி ஒரு சமூகத்தின் கருத்துருவாக்கம், பண்பாடு, வாழ்வியல் கூறு ஆகியவற்றை முழுமையாக விளங்கிக் கொள்ள இயலும். ஆயினும், பண்டைய காலங்களில், மொழிபெயர்ப்பின் நடவடிக்கையானது குறிப்பிட்ட ஒரு சில தரப்பினரால் மட்டுமே செய்யக்கூடியவையாக, குறிப்பாக இரண்டு மொழிகளில் புலமைபெற்றவர்களும் கல்வித் தகுதிபெற்றவர்களும் மட்டுமே மொழிபெயர்ப்புப் பணிகளைச் செய்து வந்ததாக இருந்தது. Garcia (2009) அவர்களும் இதே கருத்தினை முன்வைக்கின்றார். அதாவது உயர்க்கல்வி கற்றவர்களும் மொழிபெயர்ப்புக்கென சிறப்புத் தகுதி பெற்றவர்களும் மட்டுமே மொழிபெயர்ப்புக்கும் திறன் கொண்டவர்களாகக் கருதப்பட்டனர். அவர்களே மொழிபெயர்ப்புப் பணிகளைச் செய்து வந்தனர்.

ஆங்கில, மலாய் மொழிகளில் பல மொழிபெயர்ப்புக் கருவிகள் இருந்தாலும் தற்போது தமிழ்மொழியில் கூகுள் மொழிபெயர்ப்பு மட்டுமே பெரும்பாலானோரால் பயன்படுத்தப்படுகிறது. 1990களுக்குப் பிறகு கணினியும் இணையப் பயன்பாடும் அதிகரித்திருந்த வேளை Google, Yahoo, Bing போன்ற தேடுபொறிகள் மொழிபெயர்ப்பின் போது ஒரு தகவலைப் பெறுவதற்கும் பொருண்மையை அறிந்து கொள்வதற்கும் மொழிபெயர்ப்பாளருக்கு மிகவும் உதவியாக இருந்து வந்தது (Haroon & Daud, 2017). பார்க்கப்போனால், இன்றைய காலகட்டத்தில் இணையத்தின் பயன்பாடு இந்த மொழிபெயர்ப்புப் பணியை எளிதாக்க பெருந்துணையாக அமைகிறது.

தொடர்ந்து, அகராதி மொழிபெயர்ப்பதற்கான மிக முக்கியமான கருவிகளில் ஒன்றாகக் கருதப்படுகிறது. ஆனால் தற்போதைய காலகட்டத்தில் ஒரு சொல்லின் பொருளை மிக விரைவில் தெரிந்து கொள்வதற்கு இணையத்தையும் செயலிகளையும் பயன்படுத்துகின்றனர். மொழிபெயர்ப்பாளர்கள் இப்போது இந்த அகராதிகளைக் கையடக்க திறன்பேசி; கணினி வாயிலாகவே அணுகி பெறுகின்ற வாய்ப்புகள் அமைந்துவிட்டன. எங்கிருந்து வேண்டுமானாலும்; எந்நேரத்திலும் தங்கள் மொழிபெயர்ப்புப் பணியை முடிக்க மொழிபெயர்ப்பாளர் அகராதியைக் கொண்டுவர வேண்டிய அவசியமில்லை. அகராதியை அணுகுவதற்கு மொழிபெயர்ப்பாளருக்குப் போதுமான; வேகமான இணையத் தரவு மட்டுமே தேவையாக இருக்கிறது.

கூகுள் மொழிபெயர்ப்பு கூகுள் நிறுவனத்தினரால் அறிமுகப்படுத்தப்பட்டு நிருவகிக்கப்படும் மொழிபெயர்ப்புச் செயல்பாடாகும். மொழிபெயர்ப்பாளர்கள் தங்கள் மொழிபெயர்க்கவிருக்கும் உரைநடைப்பகுதியைக் கூகுள் மொழிபெயர்ப்பில் பதிவு செய்த ஒருசில விநாடிகளிலேயே நேரடியாக இலக்கு மொழிக்கு மொழிபெயர்க்கப்பட்டுவிடும். இது முழுமையாக இயந்திர மொழிபெயர்ப்பாக அமைகிறது. கூகுள் மொழிபெயர்ப்பில் தற்போது உள்ள 134 மொழிகளுள் தமிழ்மொழியும் அடங்கும். Sofer (2009) என்பவர் இயந்திர மொழிபெயர்ப்பானது மொழிபெயர்ப்புப் பணியை விரைவுபடுத்துவது மட்டுமே மாறாக அதன் தரம் அதாவது மொழிபெயர்க்கப்பட்ட தகவல்கள் சரியாகவும் பொருத்தமாகவும் மொழிபெயர்க்கப்பட்டுள்ளதா அல்லது பொருண்மை சரியாகவும் பொருத்தமாகவும் அமைந்துள்ளதா என்பது உறுதியற்றது என்று கூறுகிறார். Zetzsche (2007) என்பவரோ இயந்திர மொழிபெயர்க்கும் பனுவலின் தரத்தையும் மனிதர்கள் மொழிபெயர்க்கும் பனுவலின் தரத்தையும் ஒப்பிட இயலாது என்று கூறுகிறார். அதாவது மனித ஆற்றலோடு செய்யப்படுகின்ற மொழிபெயர்ப்பே தரமானதாக அமையக்கூடிய சாத்தியக் கூறுகள் இருக்கின்றன. காரணம் இயந்திர மொழிபெயர்ப்பானது நேரடி மொழிபெயர்ப்பை மட்டுமே செய்ய இயலும். சூழலியல் பொருண்மையை அறிந்து மொழிபெயர்க்கும் ஆற்றல் அதற்குக் கிடையாது. தொடர்ந்து இயந்திர மொழிபெயர்ப்பானது சொல்லுக்குச் சொல் மொழிபெயர்ப்பு செய்யக்கூடியதே தவிர ஒட்டு மொத்தப் பத்தி அல்லது பனுவலின் நோக்கம் அறிந்து, உள்ளடக்கம் அறிந்து உரைக் கோவை (discourse) அடிப்படையில் மொழிபெயர்ப்பதில்லை. இக்கருத்தை Iman Santoso (2011) என்பவர் மேலும் வலுவூட்டுகிறார். அதாவது கூகுள் மொழிபெயர்ப்பு நிறைவான ஒன்று அல்ல என்றும்

பொருத்தமான மொழிபெயர்ப்பை வழங்குவதில்லை என்றும் கூறுகிறார். அதுமட்டுமல்லாமல் கூகுள் மொழிபெயர்ப்பில் 134 மொழிகள் இருந்தாலும் மொழிபெயர்ப்புத் தரமானது ஒரு மொழியிலிருந்து இன்னொரு மொழிக்கு வேறுபட்டுள்ளது. இதற்குக் காரணம் அந்தந்த மொழியின் தொடரியல் அமைப்பு முறை, இலக்கணம், சொல்லாக்க அணுகுமுறை ஆகியவற்றில் மாற்றம் உள்ளது.

## 2. ஆய்வு அணுகுமுறை

இந்த ஆய்வு விளக்கமுறையிலான பண்புசார் வகையைச் சார்ந்த ஆய்வாக அமைந்திருக்கிறது. உள்ளடக்கப் பகுப்பாய்வு முறையில் தரவுகள் பகுப்பாய்வு செய்யப்பட்டுள்ளன. கோவிட்-19 பெருந்தொற்றுக் காலகட்டத்தில் மலாய்; ஆங்கில நாளிதழ்களில் வெளிவந்துள்ள பெருந்தொற்றுச் செய்திகளை இயந்திர மொழிபெயர்ப்பில் மொழிபெயர்க்கப்பட்டு ஆய்வுத் தரவுகளாகப் பயன்படுத்தப்பட்டது. இந்த மொழிபெயர்ப்பைச் செய்தவர்கள் மலேசியாவில் அமைந்துள்ள பொதுப் பல்கலைக்கழகம் ஒன்றில் மொழியியல் துறையில் பயிலும் 10 மாணவர்கள் ஆவர். இவர்கள் கூகுள் மொழிபெயர்ப்புக் கருவியைப் பயன்படுத்தி மொழிபெயர்ப்புச் செய்தனர். காரணம் தமிழ்மொழியில் கூகுள் மொழிபெயர்ப்பு மட்டுமே பெரும்பாலானோரால் பயன்படுத்தப்பட்டு வருகிறது. அவர்கள் மொத்தம் 10 செய்திகள் (5 மலாய் மொழியும் 5 ஆங்கிலமும்) தெரிவு செய்து அதனை இயந்திர மொழிபெயர்ப்பின் வாயிலாக மொழிபெயர்த்த பின் மொழிபெயர்த்ததைச் சீர்செய்தனர். இவையே ஆய்வுத் தரவுவாகப் பயன்படுத்தப்பட்டது. மலாய்; ஆங்கிலம் இரண்டு மொழிகளிலும் ஒருசில பத்திகள் மட்டுமே தோராயமாகத் தேர்ந்தெடுக்கப்பட்டு ஆய்வுத் தரவாகப் பயன்படுத்தப்பட்டது.

## 3. ஆய்வு முடிவுகள்

தேர்ந்தெடுக்கப்பட்ட செய்திகளின் மொழிபெயர்ப்பை உற்று நோக்கியதில் மொழியியல் கூறுகளான வாக்கிய அமைப்பு, சொல்லாக்கம், இலக்கணம் ஆகிய கூறுகள் மொழிபெயர்ப்பில் (இலக்கு மொழியில்) சீர்செய்யப்பட வேண்டும் என்பது அறிய முடிகிறது. தேர்ந்தெடுக்கப்பட்ட நாளிதழ்ச் செய்திப் பகுதியில் சில வரிகள் (மூல மொழி), இயந்திர மொழிபெயர்ப்பின் உரை வடிவம், மாணவர்களின் மொழிபெயர்ப்பு, இயந்திர மொழிபெயர்ப்பில் ஏற்பட்ட சிக்கல்கள் ஆகியவை அடங்கியுள்ளன. இந்த ஆய்வு முடிவுகள் அட்டவணை 1 முதல் 6 வரையில் வழங்கப்பட்டுள்ளன.

அட்டவணை 1. மலாய்-தமிழ் இயந்திர மொழிபெயர்ப்பும் மனித மொழிபெயர்ப்பும்				
எண்	மலாய்மொழி	இயந்திர மொழிபெயர்ப்பு (Machine Translation)	மனித மொழிபெயர்ப்பு (Human Translation)	மொழிபெயர்ப்புச் சிக்கல்கள்
1	Assalamualaikum dan Salam Sejahtera, saudara-saudari rakyat Malaysia yang saya kasih.	என் அன்பான மலேசிய சகோதர சகோதரிகளான அஸ்ஸலாமுவாலிகம்-ம் மற்றும் சலாம் செஜ்தெரா.	என் அன்புக்குரிய மலேசிய மக்கள் அனைவருக்கும் வணக்கம்.	<b>பொருத்தமான நிகரன்</b> அஸ்ஸலாமுவாலிகம் மற்றும் சலாம் செஜ்தெரா என்ற சொற்றொடருக்குப் பதிலாக வணக்கம் என்ற ஒரு சொல் பயன்படுத்தியிருக்கலாம். <b>தவறான வேற்றுமை உருபு</b> [சகோதர சகோதரிகளான – சகோதர சகோதரிகளுக்கு]
2	Malam ini, sekali lagi saya berada di hadapan saudara-saudari sekalian untuk memaklumkan situasi semasa ancaman wabak Covid-19 yang telah menimbulkan kebimbangan bukan sahaja di Malaysia, malahan di seluruh dunia.	இன்றிரவு, கோவிட் -19 தொற்றுநோயின் அச்சுறுத்தலின் போது நிலைமையை தெரிவிக்க மீண்டும் நான் சகோதர சகோதரிகளுக்கு முன்னால் இருக்கிறேன், இது மலேசியாவில் மட்டுமல்ல, உலகம் முழுவதும் கவலையை ஏற்படுத்தியுள்ளது.	மலேசிய மக்கள் மட்டுமல்லாமல், உலகளவில் அனைவரையும் அச்சுறுத்தி வருகின்ற கோவிட் 19 பெருந்தொற்றுநோயின் தற்போதைய நிலவரத்தைத் தெரிவிக்க இன்றிரவு உங்கள் முன் மீண்டும் நிற்கிறேன்.	<b>தவறான சொல்லாட்சி</b> [கவலை – அச்சுறுத்தல்] [அச்சுறுத்தலின் போது – அச்சுறுத்தி வருகின்ற] <b>சொல்லியைப்</b> [முன்னால் இருக்கிறேன் – முன் நிற்கிறேன்] <b>நீண்ட வாக்கியம் –</b> [குழப்பத்தை ஏற்படுத்தும் இலக்கு மொழி] <b>தவறான பதிலிடு பெயர்</b> [சகோதர சகோதரிகளுக்கு – உங்கள்]
3	Sehingga kini, wabak Covid-19 telah merebak di 135 buah negara.	இன்றுவரை, கோவிட் -19 தொற்றுநோய் 135 நாடுகளுக்கு பரவியுள்ளது.	கோவிட் -19 பெருந்தொற்று நோய் இதுவரையிலும் 135 நாடுகளுக்குப் பரவியுள்ளது.	<b>பொருத்தமான நிகரன்</b> [இதுவரை – இன்றுவரை]
4	Seramai 162,711 orang telah disahkan positif Covid-19 di seluruh dunia.	உலகளவில் மொத்தம் 162,711 பேர் கோவிட் -19க்கு சாதகமாக இருப்பது உறுதி செய்யப்பட்டுள்ளது.	உலகளாவிய நிலையில் மொத்தம் 162,711 பேருக்கு இப்பெருந்தொற்றுநோய் தொற்றியுள்ளதாக உறுதி செய்யப்பட்டுள்ளது.	<b>தவறான சொல்லாட்சி</b> [சாதகமாக – தொற்றியுள்ளது]
5	Daripada jumlah ini, seramai 6,443 orang telah meninggal dunia.	இவர்களில் 6,443 பேர் இறந்துள்ளனர்.	இந்த மொத்த எண்ணிக்கையில் சுமார் 6443 பேர் இப்பெருந்தொற்று நோயினால் பலியாகியுள்ளனர்.	<b>சில சொற்கள் விடுபட்டுள்ளது அல்லது பயன்படுத்தப்படவில்லை</b> [Jumlah - மொத்த எண்ணிக்கையில்] [Seramai = சுமார்]

6	<i>Di Malaysia, telah berlaku peningkatan kes Covid-19 secara mendadak, iaitu 190 kes semalam, disusuli 125 kes baharu hari ini, menjadikan jumlah keseluruhan individu yang dijangkiti wabak ini adalah sebanyak 553 orang.</i>	மலேசியாவில், கோவிட் -19 வழக்குகளில் கூர்மையான அதிகரிப்பு ஏற்பட்டுள்ளது, இது நேற்று 190 வழக்குகள், இன்று 125 புதிய வழக்குகள், தொற்றுநோயால் பாதிக்கப்பட்ட மொத்த நபர்களின் எண்ணிக்கையை 553 பேருக்கு கொண்டு வந்துள்ளன.	மலேசியாவில் நேற்று 190 சம்பவங்களும் தொடர்ச்சியாக இன்று 125 சம்பவங்களும் பதிவாகியுள்ளதால் கோவிட்-19 சம்பவங்கள் திடீரென அதிகரித்துள்ளது. ஆக, இது கோவிட்-19 பெருந்தொற்றுநோயினால் பாதிக்கப்பட்டவர்களின் மொத்த எண்ணிக்கை 553ஆகக் குறிப்பிடப்படுகிறது.	<b>தவறான சொல்லாட்சி</b> [வழக்குகளில் – சம்பவங்கள்] [கூர்மையான அதிகரிப்பு - சம்பவங்கள் திடீர் உயர்வு] <b>குழப்பத்தைத் தரும் நீண்ட வாக்கியம்</b> இ.மொ - ஒரே வாக்கியத்தில் 23 சொற்கள் ம.மொ - இரண்டு வாக்கியம் 1- 15 வாக்கியம் 2- 10
7	<i>Daripada jumlah tersebut, seramai 511 orang sedang dirawat di hospital, manakala 42 orang telah pun pulih.</i>	அந்த எண்ணிக்கையில், மொத்தம் 511 பேர் மருத்துவமனையில் சிகிச்சை பெற்று வருகின்றனர். அதே நேரத்தில் 42 பேர் ஏற்கனவே குணமடைந்துள்ளனர்.	அந்த மொத்த எண்ணிக்கையில், சுமார் 511 பேர் மருத்துவமனையில் சிகிச்சை பெற்றுக் கொண்டிருக்கின்ற நிலையில் 42 பேர் இத்தொற்றிலிருந்து குணமாகியுள்ளனர்.	<b>தவறான நிகரன்</b> [ஏற்கனவே குணமடைதனர் – குணமடைதனர்]  ‘ஏற்கனவே’ என்ற சொல் தேவையில்லை - பொருள் மயக்கம்
8	<i>Keutamaan kerajaan pada masa ini adalah untuk mencegah penularan baharu wabak ini yang dibimbangi akan menjangkiti lebih ramai rakyat.</i>	அரசாங்கத்தின் தற்போதைய முன்னுரிமை இந்த புதிய தொற்றுநோய் பரவாமல் தடுப்பதாகும்.	இன்னும் மக்கள் பலர் பாதிப்படைவர் என்ற அச்சுறுத்தலை ஏற்படுத்தும் இத்தொற்றுநோய் இனி பரவாமல் தடுப்பதே அரசாங்கத்தின் தற்போதைய முதன்மை நோக்கமாகும்.	<b>சில சொற்கள் பயன்படுத்தப்படவில்லை</b> dibimbangi akan menjangkiti lebih ramai rakyat - மக்கள் பலர் பாதிப்படைவர் என்ற அச்சுறுத்தலை ஏற்படுத்தும்]
9	<i>Situasi semasa wabak ini memerlukan tindakan drastik diambil bagi memulihkan keadaan secepat mungkin.</i>	இந்த தொற்றுநோயின் தற்போதைய நிலைமை விரைவில் நிலைமையை மீட்டெடுக்க கடுமையான நடவடிக்கை எடுக்க வேண்டும்.	இத்தொற்று நோயின் தற்போதைய நிலைமையிலிருந்து விரைவில் மீண்டு வர தீவிரமான நடவடிக்கைகளை மேற்கொள்வது அவசியமாகும்.	<b>தவறான வேற்றுமை உருபு</b> [நிலைமை விரைவில் நிலைமையை- நிலைமையிலிருந்து விரைவில்] [நடவடிக்கை - நடவடிக்கைகளை]
10	<i>Untuk itu, kerajaan memutuskan untuk melaksanakan Perintah Kawalan Pergerakan, mulai 18 Mac 2020, iaitu lusa hingga 31 Mac 2020, di seluruh negara.</i>	எனவே, இயக்கம் கட்டுப்பாட்டு ஆணையை 2020 மார்ச் 18 முதல் நாளை மறுநாள் 2020 மார்ச் 31 வரை நாடு தழுவிய அளவில் செயல்படுத்த அரசு முடிவு செய்துள்ளது.	எனவே, அரசாங்கம் நாளை மறுநாளான 18 மார்ச் முதல் 31 மார்ச் வரை நாடு தழுவிய அளவில் நடமாட்டக் கட்டுப்பாட்டு ஆணையைச் செயல்படுத்த முடிவெடுத்துள்ளது.	<b>தவறான சொல்லாட்சி</b> [இயக்கம் – அரசாங்கம்]

எண் 1இல் இயந்திர மொழிபெயர்ப்பில் ஏற்பட்டுள்ள சிக்கல் யாதெனில் “அஸ்ஸலாமுவாலிகம் மற்றும் சலாம் செஜ்தொரா” என்ற சொற்றொடர் இயந்திர மொழிபெயர்ப்பில் மொழிபெயர்க்க இயலவில்லை. அப்படியே “சலாம் செஜ்தொரா” என்று பயன்படுத்தப்பட்டுள்ளது. மேலும் “அஸ்ஸலாமுவாலிகம்” என்பது இஸ்லாமியர்கள், இஸ்லாமிய முறைப்படி வணக்கம் சொல்வது ஆகும். ஆனால் இந்தப் பண்பாட்டு வழக்கத்திற்கு முக்கியத்துவம் இல்லாத நிலையில் இயந்திர மொழிபெயர்ப்பில் அச்சொற்றொடரை அப்படியே “அஸ்ஸலாமுவாலிகம்” என்றே பயன்படுத்தப்பட்டுள்ளது. இவ்வாறான சூழலில் பண்பாட்டுப் பின்னணியை அறிந்து மொழிபெயர்க்க வேண்டும். அதாவது தமிழ்மொழி பயனர்களை இலக்காகக் கொண்டிருப்பதனால் “வணக்கம்” என்ற சொல்லைப் பயன்படுத்துவதே பொருத்தமாக அமைந்திருக்கும்.

அடுத்து இயந்திர மொழிபெயர்ப்பில் தவறான வேற்றுமை உருபு பயன்படுத்தப்பட்டிருக்கிறது. “சகோதர சகோதரிகளுக்கு” என்று பயன்படுத்தியிருக்க வேண்டிய இடத்தில் “சகோதர சகோதரிகளான” என்ற தவறான வேற்றுமை உருபு பயன்படுத்தப்பட்டிருக்கிறது. அடுத்து “நிலைமை விரைவில் நிலைமையை” என்ற சொற்றொடரில் வேற்றுமை உருபு தவறாகப் பயன்படுத்தப்பட்டுள்ளது. பார்க்கப்போனால் “நிலைமையிலிருந்து விரைவில் மீண்டு வர” என்ற சொல்லே பொருத்தமாக அமைகிறது. அடுத்த எடுத்துக்காட்டு “நடவடிக்கைகளை” என்று இருக்க வேண்டிய இடத்தில் “நடவடிக்கை” என்று மொழிபெயர்க்கப்பட்டுள்ளது. ஆக தமிழ்மொழியின் இலக்கண அமைப்பை இயந்திர மொழிபெயர்ப்பு அறிந்திருக்காத நிலையில் இவ்வாறான இலக்கணப் பிழைகள் ஏற்பட்டுள்ளதை இதன் வழி நாம் அறிந்துகொள்ள முடிகிறது.

மேலும் இயந்திர மொழிபெயர்ப்பில் தவறான சொல்லாட்சி பயன்படுத்தப்பட்டுள்ளன. “உலகம் முழுவதும் கவலையை ஏற்படுத்தியுள்ளது” என்ற சொற்றொடரில் “கவலை” என்ற சொல்லின் பொருள் “அச்சுறுத்தல்” என்றே இடம்பெற்றிருக்க வேண்டும். இங்குப் பயன்படுத்திய கவலை என்ற சொல் பொருத்தமற்ற சொல்லாகவே அமைந்திருக்கிறது. அடுத்து “சாதகமாக இருப்பது” என்பது சற்றும் பொருத்தமற்ற ஒன்றாக அமைகிறது.

உண்மையில் “தொற்றியுள்ளது” என்பதே மிகவும் பொருத்தமானதாகும். “சம்பவங்களை” வழக்குகள் என்று கூறப்பட்டதும் “தொடர்ச்சியாக அதிகரித்தது” என்று இடம்பெற்றிருக்க வேண்டிய இடத்தில் “கூர்மையான அதிகரிப்பு” என்று இருப்பதும் தவறான சொல்லாட்சியாகவே கருதப்படுகின்றன. அடுத்து “அரசாங்கம்” என்று இருக்க வேண்டிய இடத்தில் “இயக்கம்” என்று இருப்பது வாசகர்களுக்குத் தவறான புரிதலைக் கொடுத்துள்ளது. இவ்வாறான தவறான சொற்பயன்பாடு வாசகர் புரிதலில் குழப்பத்தை ஏற்படுத்தவல்லது. மேலும் மொழிபெயர்ப்பின் நோக்கத்தை வெற்றிபெறச் செய்வதில் தோல்வி அடைந்துள்ளது என்றே கூறலாம்.

அடுத்து, வாக்கியங்களைப் பிரித்து மொழிபெயர்க்கும் திறன், இயந்திர மொழிபெயர்ப்புக்கு இல்லை. எண் 6இல் ஒரே வாக்கியத்தில் 23 சொற்கள் பயன்படுத்தப்பட்டுள்ளது. இவ்வாறான நீண்ட வாக்கியங்கள் தகவல் பரிமாற்றத்தில் குழப்பத்தை ஏற்படுத்தும் வகையில் அமைந்துள்ளது. அதுவே மனித மொழிபெயர்ப்பு நடவடிக்கையின்போது அது இரண்டு வாக்கியங்களாகப் பிரித்து, பொருத்தமான இடைச்சொல் (ஆக) கொண்டு மொழிபெயர்க்கப்பட்டுள்ளதால் கூறவரும் கருத்து தெளிவாக விளங்கும் வகையில் அமைந்துள்ளது.

எண் 2இல் “நான் சகோதர சகோதரிகளுக்கு முன்னால் இருக்கிறேன்” எனும் சொற்றொடரில் “சகோதர சகோதரிகளுக்கு முன்னால்” என்று கூறுவதைவிட “உங்கள் முன்னால்” என்று மொழிபெயர்க்கப்பட்டிருக்க வேண்டும். இங்குத் தவறான பதிலிடு பெயர் பயன்படுத்தியிருப்பது தமிழ்மொழிக்குச் சற்றும் பொருத்தமற்ற சூழலில் அமைந்துள்ளது. இதை அடுத்து, இயந்திர மொழிபெயர்ப்பில் சொல் இயையும் தவறாகப் பயன்படுத்தப்பட்டுள்ளது. “முன்னால் இருக்கிறேன்” என்ற சொற்றொடருக்கு, “முன்னால் நிற்கிறேன்” என்ற சொல் இயைபு பொருத்தமாக இருக்கிறது.

தொடர்ந்து, ஆங்கில மொழிபெயர்ப்பு மலாய்மொழி இயந்திர மொழிபெயர்ப்புப் போன்றே சில சிக்கல்களைக் கொண்டுள்ளன. அச்சிக்கல்களைத் தொடர்ந்து காண்போம்.

#### அட்டவணை 2. ஆங்கிலம் தமிழ் இயந்திர மொழிபெயர்ப்பில் பொருத்தமற்ற கலைச்சொல் பயன்பாடு

ஆங்கிலமொழி	தமிழ்மொழி (Machine Translation)	தமிழ்மொழி (Human Translation)
KUALA LUMPUR, Jan 4 -- The Ministry of Science, Technology and Innovation (MOSTI) has stressed that the COVID-19 vaccine will not cause long-term harmful side effects on people's health.	கோலாலம்பூர், ஜனவரி 4 - கோவிட் -19 தடுப்பூசி மக்களின் ஆரோக்கியத்தில் நீண்டகால தீங்கு விளைவிக்கும் பக்க விளைவுகளை ஏற்படுத்தாது என்று அறிவியல், தொழில்நுட்பம் மற்றும் கண்டுபிடிப்பு அமைச்சகம் (மோஸ்டி) வலியுறுத்தியுள்ளது.	கோலாலம்பூர், ஜனவரி 4 - கோதணி நச்சு தடுப்பூசிகள் மக்களின் ஆரோக்கியத்திற்கு நீண்டகாலப் பாதிப்பை ஏற்படுத்தக்கூடிய கொடிய பக்க விளைவுகளை வழங்காது என்று அறிவியல், தொழில்நுட்பம் மற்றும் புத்தாக்க அமைச்சு (MOSTI) வலியுறுத்தியுள்ளது.

மேற்காணும் எடுத்துக்காட்டில் “அறிவியல், தொழில்நுட்பம் மற்றும் கண்டுபிடிப்பு அமைச்சகம்”, இயந்திர மொழிபெயர்ப்பு ஆகும். இதில் “கண்டுபிடிப்பு எனும் சொல் ஆங்கிலத்தில் – innovation என்று பொருள்படுகிறது. இதற்குத் தமிழில் மிகப் பொருத்தமான சொல் “புத்தாக்கம்” என்று பயன்படுத்தப்பட்டு வருகிறது. மேலும் “அமைச்சு” என்று பயன்படுத்தியிருக்க வேண்டிய இடத்தில் அமைச்சகம் என்று பயன்படுத்தியிருப்பதும் அச்சொல்லின் பொருண்மையில் குழப்பத்தை ஏற்படுத்தியிருக்கிறது.

#### அட்டவணை 3 ஆங்கிலம் தமிழ் இயந்திர மொழிபெயர்ப்பில் இயைபுத்தன்மையற்ற சொல் பயன்பாடு

ஆங்கிலமொழி	தமிழ்மொழி (Machine Translation)	தமிழ்மொழி (Human Translation)
Its deputy minister Ahmad Amzad Hashim said the vaccine had been clinically tested and would be regulated by the National Pharmaceutical Regulatory Agency (NPR) of the Ministry of Health.	அதன் துணை மந்திரி அஹ்மத் அம்சாத் ஹாஷிம், தடுப்பூசி மருத்துவ ரீதியாக பரிசோதிக்கப்பட்டதாகவும், சுகாதார அமைச்சின் தேசிய மருந்து ஒழுங்குமுறை நிறுவனம் (என்.பி.ஆர்.ஏ) கட்டுப்படுத்தும் என்றும் கூறினார்.	(MOSTI) துணை அமைச்சர் அஹ்மத் அம்சாத் ஹாஷிம், அந்தத் தடுப்பூசி மருத்துவ ரீதியாக பரிசோதிக்கப்பட்டதாகவும், சுகாதார அமைச்சின் தேசிய மருந்துகள் கட்டுப்பாட்டுப் பிரிவு நிறுவனம் (NPR) அதை ஒழுங்குமுறைப்படுத்தும் என்றும் கூறினார்.

அடுத்ததாக இயைபுத் தன்மையற்ற சொற்பயன்பாடும் இயந்திர மொழிப்பெயர்ப்பில் கண்டறியப்பட்டுள்ளது. அதாவது மலேசிய நாட்டுச் சூழலில் “துணை மந்திரி” என்னும் சொற்பயன்பாடு வழக்கில் அல்லது பயன்பாட்டில் இல்லை. மலேசியாவில் “துணை அமைச்சர்” என்றே பயன்படுத்தப்பட்டு வருகிறது. இந்நிலையில் “துணை மந்திரி” என்று கூறும்போது இது இயைபு தன்மைக்குப் புறம்பாக இருக்கிறது. ஆக இவ்வாறான சொற்களை நாம் பண்படுத்தும் போது வாசகர்களுக்கு அச்செய்தி அந்நியமாக இருக்கும் என்பதையும் படம்பிடித்துக் காட்டுகிறது. ஆக இவ்வாறான கூறுகளை அறிந்துகொண்டுதான் மொழிபெயர்ப்பாளர் மொழிபெயர்ப்புப் பணியைச் செய்ய வேண்டும்.

#### அட்டவணை 4. ஆங்கிலம் தமிழ் இயந்திர மொழிபெயர்ப்பில் தவறான கருத்து கொண்ட வாக்கிய அமைப்பு

ஆங்கிலமொழி	தமிழ்மொழி (Machine Translation)	தமிழ்மொழி (Human Translation)
------------	---------------------------------	-------------------------------



"All vaccines have side effects, what is important is that the side effects are not serious or harmful, so far based on the data we received the vaccine does not have serious side effects."	"அனைத்து தடுப்பூசிகளும் பக்க விளைவுகளைக் கொண்டிருக்கின்றன, முக்கியமானது என்னவென்றால், பக்க விளைவுகள் தீவிரமானவை அல்லது தீங்கு விளைவிப்பவை அல்ல, இதுவரை நாங்கள் தடுப்பூசி பெற்ற தரவுகளின் அடிப்படையில் கடுமையான பக்க விளைவுகளை ஏற்படுத்தவில்லை."	"அனைத்து தடுப்பூசிகளும் அதன் பக்க விளைவுகளைக் கொண்டிருக்கின்றன. இதில் முக்கியமானது, கொடிய மற்றும் ஆபத்தை விளைவிக்கக்கூடிய பக்க விளைவுகளை கொண்டிருக்கக்கூடாது. எனவே, இதுவரை எங்களுக்கு கிடைக்கப்பெற்ற தடுப்பூசி பற்றிய தரவுகளின் அடிப்படையில் அவை கடுமையான பக்க விளைவுகளை ஏற்படுத்தக்கூடியவையல்ல."
---	---	---

அடுத்ததாக இயந்திர மொழிபெயர்ப்பில் வாக்கிய அமைப்பு முறையிலும் குழப்பம் உள்ளது. "Based on the data we received" என்ற வாக்கியம் "நாங்கள் தடுப்பூசி பெற்ற தரவுகள்" என்று மொழிபெயர்க்கப்பட்டுள்ளது. இது தவறான தகவலைப் பிரதிபலிக்கிறது. பார்க்கப்போனால் "எங்களுக்குக் கிடைக்கப்பெற்ற தடுப்பூசி பற்றிய தரவுகளின்" என்று மொழிபெயர்க்கப்பட்டிருக்க வேண்டும். இந்த இரண்டு வாக்கியங்களில் வெளிப்படுத்தப்பட்டுள்ள கருத்துகள் முற்றிலும் மாறுபட்டு இருக்கிறது. ஆக இதுவே இயந்திர மொழிபெயர்ப்பின் குறைபாடாகப் பார்க்கலாம். பொருத்தமற்ற மொழிபெயர்ப்பு ஒருபுறம் இருக்க தவறான கருத்தை வாசகர்களுக்குக் கொண்டு சேர்க்கும் நடவடிக்கையானது மொழிபெயர்ப்பின் நோக்கத்திலிருந்து விலகிச் சென்றுவிட்டதை அறிய முடிகிறது.

#### அட்டவணை 5. ஆங்கிலம் தமிழ் இயந்திர மொழிபெயர்ப்பில் மொழியணிகளின் நேரடி மொழிபெயர்ப்பு

ஆங்கிலமொழி	தமிழ்மொழி (Machine Translation)	தமிழ்மொழி (Human Translation)
"Most side effects are temporary such as injection pain, some may have fever or nausea," he said when met after appearing as a guest on Bernama Radio Jendela Fikir programme today.	"பெரும்பாலான பக்க விளைவுகள் ஊசி வலி போன்ற தற்காலிகமானவை, சிலருக்குக் காய்ச்சல் அல்லது குமட்டல் இருக்கலாம்" என்று பெர்னாமா ரேடியோ ஜென்டெலா ஃபிகிர் நிகழ்ச்சியில் விருந்தினராக தோன்றிய பின்னர் சந்தித்தபோது அவர் கூறினார்.	"பெரும்பாலான பக்க விளைவுகள் தற்காலிகமானது மட்டுமே. எடுத்துக்காட்டாக, தடுப்பூசியைச் செலுத்தியவுடன் வலி ஏற்படுவது, சிலருக்கு காய்ச்சல் அல்லது குமட்டலும் ஏற்படலாம்," என்று பெர்னாமா ரேடியோ ஜென்டெலா ஃபிகிர் நிகழ்ச்சியில் விருந்தினராகப் பங்கேற்ற பின்னர் சந்தித்தபோது அவர் கூறினார்.

அடுத்ததாக, இயந்திர மொழிபெயர்ப்பில், மொழியணிகள் குறிப்பாக உருவகம் அல்லது உவமையை மொழிபெயர்க்கும்போது அவற்றை நேரடியாக மொழிபெயர்த்துவிடுகிறது. அப்படி மொழிபெயர்த்தால் அதன் பொருள் விளங்காது அல்லது தவறான பொருளைப் கொடுக்கும் நிலையை ஏற்படுத்திவிடுகிறது. எடுத்துக்காட்டாக, "such as injection pain" என்ற சொற்றொடர் தற்காலிகமானவையே என்று பொருள்கொள்ளப்பட வேண்டும். ஆனால் மொழிபெயர்ப்பில் "பெரும்பாலான பக்க விளைவுகள் ஊசி வலி போன்று தற்காலிகமானவை" என்று பொருள்படுகிறது. "ஊசி வலி" என்ற இந்தச் சொற்றொடர் நேரடியாகவே மொழிபெயர்க்கப்பட்டுள்ளது. தமிழில் "ஊசி வலி" என்ற சொற்றொடர் தற்காலிகம் என்ற பொருண்மையைக் கொடுக்கவில்லை; பொருளும் தெளிவாக விளங்கவில்லை. ஆக இச்செய்தியைப் படிக்கும் வாசகர்களுக்குக் குழப்பம் ஏற்படுத்தும் வகையில் அமைந்துள்ளது.

#### அட்டவணை 6. ஆங்கிலம் தமிழ் இயந்திர மொழிபெயர்ப்பில் ஒலிபெயர்ப்பு

ஆங்கிலமொழி	தமிழ்மொழி (Machine Translation)	தமிழ்மொழி (Human Translation)
Malaysia has signed an agreement with Pfizer-BioNTech to procure 12.8 million vaccine doses and another 6.4 million doses will be provided by AstraZeneca as well as several other suppliers.	ஃபைசர்-பியோஎன்டெக் நிறுவனத்துடன் மலேசியா 12.8 மில்லியன் தடுப்பூசி அளவுகளை வாங்குவதற்கான ஒப்பந்தத்தில் கையெழுத்திட்டுள்ளது. மேலும் 6.4 மில்லியன் டோஸ் அஸ்ட்ராஜெனெகா மற்றும் பல சப்ளையர்களால் வழங்கப்படும்.	மலேசியா, Pfizer-BioNTech நிறுவனத்துடன் 12.8 மில்லியன் தடுப்பூசி அளவுகளை வாங்குவதற்கான ஒப்பந்தத்தில் கையெழுத்திட்டுள்ளது. மேலும், 6.4 மில்லியன் அளவுகள் AstraZeneca மற்றும் பிற வழங்குனர்கள் மூலமாக வழங்கப்படும்.

இயந்திர மொழிபெயர்ப்பில் சில சொற்களுக்குப் பொருத்தமான நிகரன்களைப் பரிந்துரைக்காமல் அதாவது சொல்லாக்கம் செய்யாமல் அப்படியே ஒலிபெயர்த்து கூறுகின்ற குழுவும் அமைந்திருக்கின்றது. "suppliers" என்று பயன்படுத்தப்பட்ட மொழிபெயர்ப்புச் சொல் "வழங்குநர்கள்" அல்லது "விநியோகர்கள்" என்ற சொல்லைப் பயன்படுத்தாமல் "சப்ளையர்களால்" என்று ஒலிபெயர்ப்புச் செய்யப்பட்டு வழங்கப்பட்டுள்ளது. இவ்வாறான குழுவில் இயந்திர மொழிபெயர்ப்புச் சில சொற்களை மொழிபெயர்க்காமல் அப்படியே ஒலிபெயர்த்து வழங்கப்பட்டிருப்பது, மொழிபெயர்ப்பு நோக்கத்தை அடைவதில் பொருத்தமான கலைச்சொல் பயன்படுத்துவதிலிருந்து விலகி நிற்பதை அறிய முடிகிறது.



#### 4. முடிப்பு

இந்த ஆய்வு இயந்திர மொழிபெயர்ப்பில் ஏற்படும் சிக்கல்களை அடையாளம் காட்டியுள்ளது. அதாவது மொழிபெயர்ப்பு நடவடிக்கையின்போது குழப்பமான வாக்கியத் தொடர், பொருத்தமற்ற கலைச்சொல் பயன்பாடு, இலக்கணப் பிழைகள், பொருத்தமற்ற பதிலிடுபெயர்கள் போன்றவை எவ்வாறு இயந்திர மொழிபெயர்ப்பில் இடம்பெற்றிருக்கிறது என்பதை இவ்வாய்வின்வழி அறிய முடிகிறது.

பொருண்மை அடிப்படையில் பார்க்கும் பொழுது இயந்திர மொழிபெயர்ப்புச் சொல்லுக்குச் சொல் மொழிபெயர்த்து வழங்குகிறது. அதனால் நேரடி பொருண்மை மட்டுமே மொழிபெயர்ப்பில் இடம்பெற்றுள்ளது. மாறாக சூழலியல் பொருண்மை சரியாகக் கையாளப்படவில்லை. ஒரு சொல் பல பொருண்மைகளைக் கொண்டிருக்கலாம். ஒவ்வொரு சூழலுக்கும் ஏற்ற வகைகள் அந்தப் பொருண்மை மாற்றம் அடையும். ஆகவே முழு வாக்கியத்தின் கருத்தை இலக்கு மொழிக்குச் சரியாக மொழிபெயர்க்கும் பணியைச் செய்வதில் இயந்திர மொழிபெயர்ப்பு தவறி விடுகின்றது. இது ஒட்டு மொத்த மொழிபெயர்ப்பின் தரத்தைப் பாதிக்கின்றது; மொழிபெயர்ப்பின் நம்பகத்தன்மையைக் குறைத்தும் விடுகிறது. இவ்வாறு சொல்லின் பொருண்மையைப் பொருத்தமான சொல் கொண்டு மொழிபெயர்க்கும்போது அது வாக்கிய அளவோடு நின்றுவிடாமல் ஒட்டுமொத்த உரைநடைப்பகுதியையும் உள்வாங்கிக் கொண்டு மொழிபெயர்க்க வேண்டும். இங்குதான் மொழியியல் திறனான உரைக்கோவை முக்கியத் துறையாகத் துணைநிற்கின்றது.

மொத்தத்தில் இந்த ஆய்வின்வழி இயந்திர மொழிபெயர்ப்பு பல்வேறு குறைபாடுகளைக் கொண்டுள்ளது என்பதை நாம் அறிய விழைகிறோம். இருப்பினும் இயந்திர மொழிபெயர்ப்பு நமக்கு எவ்வாறு உதவுகின்றது என்றால் மொழிபெயர்ப்புப் பணியைத் துரிதப்படுத்துவதற்கு ஆகும். மொழிபெயர்ப்பு உரைநடைப் பகுதியில் ஒரு சில வாக்கியங்கள் மிகவும் பொருத்தமாக மொழிபெயர்க்கப்பட்டிருக்கும். அதே வேளை ஒரு சில சொற்களுக்கும் மிகப்பொருத்தமான நிகரங்களும் பயன்படுத்தப்பட்டிருக்கும். இவ்வாறான சொற்களையும் வாக்கியங்களையும் அப்படியே நாம் பயன்படுத்திக் கொள்ளலாம். ஆனால் குறைபாடுடைய வாக்கியங்களையும் சொற்களையும் மனித ஆற்றல் துணைகொண்டு சீர்செய்ய வேண்டும் மனித மொழிபெயர்ப்பின் பங்கு அளப்பரியதாக அமைகிறது. அதாவது இயந்திர மொழிபெயர்ப்புத் தனித்து தரமான மொழிபெயர்ப்பை வழங்கிவிட இயலாது. மனித ஆற்றல் என்று பார்க்கும்போது ஒரு மொழிபெயர்ப்பாளர் மூல மொழி; இலக்கு மொழி ஆகியவற்றில் நன்கு புலமைப் பெற்றவர்களாக இருத்தல் வேண்டும், இலக்கு மொழியின் பண்பாட்டு நடைமுறைகளை அறிந்துகொண்டு வைத்திருப்பவர்களாகவும் இருக்க வேண்டும். மேலும் மொழிபெயர்ப்பாளர்களுக்குத் துறைசார்ந்த முன்னறிவு இருப்பதும் சாலச் சிறந்ததாகும். அவ்வாறு இல்லாத பட்சத்தில் மொழிபெயர்ப்பாளர்கள் இணைய வசதியைத் பயன்படுத்தி தேவையான தகவல்களை அறிந்துகொண்டு மொழிபெயர்ப்பதும் தரமான மொழிபெயர்ப்பை வழங்கிட உதவும். ஆக இயந்திர மொழிபெயர்ப்பும் மனித மொழிபெயர்ப்பும் இணைந்து ஒருங்கே மொழிபெயர்ப்புப் பணி செய்யும்போது விரைந்தும் தரமான மொழிபெயர்ப்பையும் வழங்கிவிட முடிகிறது. இதற்கும் மேலாக தரமான இயந்திர மொழிபெயர்ப்பு அமைய வேண்டுமெனில் சமூகத்தின் பங்கும் அளப்பரியது. அதாவது கூகுள் மொழிபெயர்ப்பில் மொழிபெயர்க்கப்பட்ட உரைநடைப்பகுதியைப் பயனர்களே சீர் செய்து பரிந்துரைக்கலாம். அதிகமாகச் சீர்மைப் பணிகள் இடம்பெறும்போது தமிழ்மொழியில் மொழிபெயர்ப்பின் தரமும் உயரும் என்பதைக் கருத்தில் கொள்ள வேண்டும். அதாவது நாம் பயனர்களாக மட்டுமல்லாது பங்களிப்பாளர்களாகவும் இருக்க வேண்டும், தமிழ்மொழி வளர்ச்சிக்கு நமது பங்காக இது அமையும்.

#### மேற்கோள் நூல்கள்

- [1] Derrida, J., & Venuti, L. (2001). What is a "relevant" translation?. *Critical inquiry*, 27(2), 174-200.
- [2] Garcia, I. (2009). Beyond Translation Memory: Computers and the professional translator. *The Journal of Specialised Translation*, 12, 199-214.
- [3] Haroon, H., & Daud, N. S. (2017). The translation of foreign words in an English novel into Malay. *GEMA Online Journal of Language Studies*, 17(1).
- [4] Mukadam, N., Sommerlad, A., & Livingston, G. (2017). The relationship of bilingualism compared to monolingualism to the risk of cognitive decline or dementia: a systematic review and meta-analysis. *Journal of Alzheimer's Disease*, 58(1), 45-54.
- [5] Nur Naimah Akmar, K & Harun, B. (2016). Penggunaan Google Translate Dalam Aktiviti Terjemahan Kata Arab. Prosiding Persidangan Kebangsaan Isu-Isu Pendidikan Islam 2016 (ISPEN-i 2016), 21 – 22 Mei 2016, Kuala Lumpur Malaysia ISBN 978-967-14229-0-8, hlm 274-279.
- [6] Sofer, M (2009). The Morry Sofer's translator's handbook, edisi ke-7. USA: Schreiber Publishing Inc
- [7] Wan Rose Eliza, A. R. (2013). Penggunaan peralatan dan teknologi dalam penterjemahan. *Asas Terjemahan Dan Interpretasi. Pulau Pinang: Penerbit Universiti Sains Malaysia*.
- [8] Zetzsche, J. (2007). Machine Translation Revisited. *Translation Journal*, Volume (11). (<http://accurapid.com/journal/39MT.htm>)

