



INTERNATIONAL CONFERENCE ON TAMIL COMPUTING AND INFORMATION TECHNOLOGY

14-16 June 2024

University of Texas, Dallas. USA



Conference Proceedings
மாநாட்டுக் கட்டுரைகள்

**International Conference on Tamil Computing
And Information Technology (ICTCIT 2024)**

Jointly organized by



International Forum for Information Technology in Tamil

and



23rd Annual Conference

CONFERENCE PROCEEDINGS

மாநாட்டுக் கட்டுரைகள்

Edited by

Dr.K. Kalyanasundaram

Published by

International Forum for

Information Technology in Tamil (INFITT)

ISSN : 2313 – 4887

Table of Contents

Conference Schedule	6
Greeting Messages	7-17
 I. Invited Talks	
Chatbot Applications in Agriculture Prof. Sobha L. (Anna Univ KBC Research Center, MIT, Chennai, India)	18
Building Tamil Treebanks Dr. K. Sarveswaran (Univ of Jaffna, Sri Lanka & Univ of Konstanz, Germany)	22
Making of 'soul' and 'body' - A technological pursuit in the world of Artificial Intelligence: Interacting to Multilingual Robots - Scopes and Future Dr. Vasu Renganathan (Univ of Pennsylvania, Philadelphia, USA)	33
AI and Cybersecurity Prof. P. Jayashree, (Anna Univ KBC Research Center, MIT Campus, Chennai)	38
LLMs should be aligned yes, to which and what values? Prof. Monojit Choudhury (Md. Bin Zayed Univ.of Artificial Intelligence, Abu Dhabhi, UAE)	42
Making Large Language Models that can speak Tamil Mr. Raju Kandasamy (Thoughtworks, Coimbatore, India)	45
Text to Sound: Train your Large Language Models Dr. Ruby Annette (Cognizant, Dallas, Texas, USA)	48
Generative AI and Large Language Model Applications in Medicine (not included!) Dr. Malaikannan, S (Samaa Technologies, California, USA)	
 II. Tutorial Workshops	
Speech Recognition and speech synthesis fundamentals, challenges and case studies Prof. B. Bharathi (SSN College of Engineering, Chennai)	51
Deep learning and Machine Learning of Tamil Corpora Dr. Dhivya Chinnappa (J.P. Morgan Chase, Dallas, Texas, USA)	58
Linguistic annotation Essentials for NLP in Tamil Dr. K. Parameswari (International Inst. Of Information Technology, Hyderabad)	60
Machine Translation using Transformers (NMT) Dr. T. Patabhi RK Rao & Dr. Vijay Sundar Ram (Anna Univ KBC Research Center, MIT, Chennai)	64

Utilizing Tech Tools For “Seal of Biliteracy” Tamil Certification Exam Topics Mr. Soundar Jayabal (Avvai Tamil Center, Dallas, Texas, USA)	71
Developing AI applications in Tamil - A Tutorial Dr. Muthu Annamalai (Ezhil Language Foundation, California, USA)	74
A Transformer-Based Sandhi Splitter for Tamil Parameswari Krishnamuthy, Nagaraju Vuppala and Nisha Irene	76
Tools for Smart Wordprocessing in Tamil (not included!) Mr. Rajaraman @Neechalkaran (InfoSys, Chennai)	
Gamification in Tamil Learning (not included!) Mr. Mohan Dhandapani (FedEx, Dallas, Texas, USA)	
Text to Speech in Tamil using Artificial Intelligence (not included!) Mr. Suresh Kumar Harikrishnan (TCS, Dallas, Texas, USA)	

III. Contributed papers

இயல் அமை : கைபேசி மூலம் தமிழ் இணையதளம் தங்கவேலு சின்னசாமி	86
SIGNET: Superior Intelligence for Gesture-based Neural Exploration in Tamil Kanimozhi Suguna S*, Prema S, Vasanthakumari M	91
Facilitating Efficient Web Search Engine for Language Community Prema S*, Kanimozhi Suguna S, Vasanthakumari M	96
கணினி மொழிபெயர்ப்பில் முன்னேற்றங்கள்: தமிழ் - ஆங்கில இயந்திர மொழிபெயர்ப்பு நோக்கு மா. வசந்தகுமாரி*, செ. பிரேமா, செ. கனிமொழி சுகுணா	103
Aspect-Based Sentiment Analysis of Movie Reviews in Tamil- A Study On The Effectiveness of the BERT with MADTRAS Dataset M. Arunmozhi, E. Syam Mohan. R. Sunitha, V. Dhanalakshmi & K. Pajanivelou	108
கூகுள் லென்ஸ்: காட்சித்தேடலும் படச் சொற்களை மொழிபெயர்ப்பதில் ஏற்படும் சிக்கல்களும் முனைவர் சு. பிரபாவதி,	112
Improving Tamil-Telugu Neural Machine Translation using Morpheme-based Tokenization Parameswari Krishnamurthy, Sushvin Marimuthu & Nagaraju Vuppala	118

Tokenization, training and fine-tuning strategies for large language models for Tamil to English text translation task Siddharth Krishna Kumar & Madhavaraj, A	119
பழந்தமிழ் இலக்கியத்தில் மெய்ம்மயக்கத்தின் பாங்கு இராம்பிரசாந்த் வெங்கடக்கிருஷ்ணன் & பாலசுந்தரராமன் இலக்குவன்	120
இயந்திர மொழிபெயர்ப்பு: சிக்கல்கள் - தமிழ்ச் செவ்விலக்கியப் பரவலாக்கத்தை முன்வைத்து) Suja Suyambu	128
ChatGpt 4 : படைப்பாக்க சிந்தனை Kasthuri, M	130
செயற்கை நுண்ணறிவு மூலம் தமிழ்ப் படைப்பாக்கம் நித்திஷ் செந்தூர்	131
தமிழ்நாட்டுப் பல்கலைக்கழகங்களால் தமிழ் மொழியில் ஷோத்கங்காவிற்கு மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளின் பங்களிப்புகள்: ஓர் ஆய்வு சு. கோதைநாயகி	139
இணையத்தில் தமிழ்மொழியின் வளர்ச்சி பி.ஆர். இலட்சுமி	149
காணி இன மக்களின் பேச்சு மொழியும் கணினித் தமிழ் வழி பெற்ற வளர்ச்சி நிலையும் (கன்னியாகுமரி மாவட்டப் பழங்குடியின மக்களை முன் வைத்து) 151 முனைவர் மு.ஜோதிலட்சுமி	

International Conference on Tamil Computing and Information Technology (ICTCIT) 2024

Conference Schedule

Time	Time	Friday 14 June 2024	Saturday 15 June 2024	Sunday 16 June 2024
Dallas US	India IST	Technical Sessions		
08.00 – 08.30	18.30-19.00	Inaugural (30 min)		
08.30 - 09.30	19.00 - 20.00	Invited I (virtual) M. Choudhury	Invited IV (virtual) Raju Kandasamy	Invited VII (virtual) K. Sarveswaran
09.30 – 10.00	20.00 – 20.30	Coffee Break	Coffee Break	Coffee Break
10.00 – 11.00	20.30 – 21.30	Invited II Jayashree P	Invited V Sobha L	Invited VIII S Malaikannan
11.00 – 12.30	21.30 - 23.00	CP Session I 4 papers P1-P4 (each 20 min)	CP Session III 4 papers P5-P8 (each 20 min)	CP Session IV 4 papers P9-P12 (each 20 min)
12.30 – 13.30	23.00 - 24.00	Lunch Break	Lunch Break	Lunch Break
13.30 - 14.30		Invited III Vasu Renganathan	Invited VI Ruby Annette	BOF session
14.30 – 16.30		CP Session II 4 papers P13-P16	Tamil IT workshop for youngsters I Nithish Senthur	
16.30 – 17.00		Tea Break	Tea Break	Valedictory
		Tutorial Sessions		
08.30 – 10.30	19.00-22.00	Tutorial I (virtual): B Bharathi	Tutorial II (virtual) Pattabhi & Vijay	Tutorial VI M. Dhandapani
10.30 - 12.30		Tutorial II Neechalkaran	Tutorial V Dhivya Chinnappa	Tutorial VIII Sunder Jeyabal
13.30 – 15.30		Tutorial III Muthu Annamalai	Tutorial VI: Parameswari K	Tutorial IX S Harikrishnan
16.30 - 17.30				Valedictory
18.00 – 21.00		Cultural Program & Conference Dinner	Cultural Program & Conference Dinner	

** Those highlighted in blue are virtual sessions via Zoom



भारत का प्रधान कौंसल
ह्यूस्टन
**CONSUL GENERAL OF INDIA
HOUSTON**

June 05, 2024


MESSAGE

I would like to extend my best wishes for the International Conference on Tamil Computing and Information Technology, organized by International Forum for Information Technology in Tamil (INFITT) at the University of Texas at Dallas from June 14-16, 2024.

This event stands as a testament to the profound impact of regional languages in the ever-evolving domain of Information Technology, including Artificial Intelligence. Regional languages hold a unique and vital place in global communication and cultural heritage. They not only preserve the rich history and traditions of communities but also play an increasingly vital role in the digital age. By incorporating regional languages into cutting-edge technologies, we ensure that cultural diversity is maintained and celebrated, while also making technology more accessible to a broader audience.

Tamil, one of the world's oldest languages, exemplifies this beautifully. As a classical language with a profound literary history, Tamil's integration into modern computing and AI signifies the preservation and advancement of our linguistic heritage. The role of Tamil in Natural Language Processing and real-time translation technologies is particularly noteworthy. These innovations promise to bridge communication gaps and bring people closer.

I commend the organizers and participants for their dedication and efforts in promoting Tamil computing. May this conference pave the way for significant advancements and inspire future innovations. I am confident that the knowledge shared, and the networks formed here will contribute immensely to the global IT community, enriching both academic and practical applications.


(D.C. Manjunath)



Dr. Jey Veerasamy

Director, Center for CS Education & Outreach &
Professor of Instruction, CS
THE UNIVERSITY OF TEXAS AT DALLAS
800 W. Campbell Road MS EC31
Richardson, Texas 75080-3021



June 1, 2024

On behalf of The University of Texas at Dallas, I am excited to welcome the local families & guests from all over the world to this Tamil Technology conference in our beautiful campus. This is yet another way for us to engage more with the local community, especially with all the families who speak Tamil. I believe ~20% of our international students are from Tamilnadu as well. I hope they will also find this conference enjoyable and valuable to connect to their roots. I also hope that the attendees will make a lot of connections and long-lasting relationships. I am sure the conference will be a great success!

A handwritten signature in black ink, appearing to read 'Jey Veerasamy'.

Dr. Jey Veerasamy

**பன்னாட்டுக் கணித்தமிழ்த் தகவல் தொழில்நுட்ப மாநாடு 2024
வாழ்த்து**

மு. பொன்னவைக்கோ

இயக்குநர், தமிழ் இணையப் பல்கலைக்கழகம் (2000-2003)

துணைவேந்தர்: பாரதிதாசன் பல்கலைக்கழகம் (2007-2010), எஸ்.ஆர்.எம் பல்கலைக்கழகம் (2010- 2013)



முதல் தமிழ்இணைய மாநாடு, 1997-ஆம் ஆண்டு, மே மாதம் சிங்கப்பூரில் 'தமிழ் இணையம் 97' என்ற பெயரில் நடைபெற்றது. Internet என்னும் ஆங்கில சொல்லுக்கு 'இணையம்' என்னும் தமிழ்ச்சொல்லை கணித்தமிழ் உலகிற்கு வழங்கிய பெருமை சிங்கப்பூரைச்சாரும். இரண்டாவது தமிழ்இணைய மாநாடு, 'தமிழ் இணையம் 99' என்ற பெயரில் சென்னையில் 1999-ஆம் ஆண்டு பிப்ரவரி மாதம் நடைபெற்றது. இம்மாநாட்டை அடுத்து 'தமிழ் இணையம் 2000' மாநாடு இலங்கையில் நடத்தத் திட்டமிடப்பட்டு இருந்தது. அந்த மாநாட்டின் முன்னேற்பாட்டுக் கூட்டம் 2000-ஆம் ஆண்டு நவம்பர் மாதம் இலங்கையில் நடைபெற்றது. அந்தக் கூட்டத்தில் சுவிட்சர்லாந்து நாட்டிலிலுள்ள முனைவர் கல்யாணசுந்தரம் அவர்கள், உலகஅளவில் ஒரு இணையத்தமிழ் ஆய்வுக்குழு அமைக்க வேண்டுமென வரைந்து அனுப்பியிருந்த திட்டத்தை திரு.அருண்மகிழ்நன் முன்மொழிந்தார். அந்தக் கலந்துரையாடலில் பிறந்ததுதான் 'உத்தமம்' என்னும் உலகத்தமிழ்த் தகவல் தொழில் நுட்பமன்றம். இம்மன்றத்திற்குப் பெயரிடும் பெருமை எனக்குக் கிட்டியது நான் பெற்ற பேறு. அந்த மாநாட்டை அடுத்து சிங்கப்பூரில் 2000-ஆம் ஆண்டு ஜூலை மாதம் 22-24-ஆம் நாட்களில் நடைபெற்ற தமிழ் இணைய மாநாட்டில் 'உத்தமம்' (INFITT) தொடங்கி வைக்கப்பட்டது. இணையத்தமிழின் ஆய்விற்காக உத்தமத்தில் தமிழ்க் கலைச்சொல் ஆக்கல், யூனிகோடு தமிழ் (UNICODE Tamil) வளர்ச்சி, இணையதள தமிழ் முகவரி வடிவமைத்தல், தமிழ் வரிவடிவக் குறியீட்டுத் தரப்பாடு, ஆங்கில வரிவடிவத் தமிழ்த் தரப்பாடு, தமிழ் எழுத்துரு அறிதல் (Tamil OCR), லினக்ஸில் தமிழ் (Tamil in Linux) ஆகிய பணிகளுக்காக ஏழு ஆய்வுப் பணிக்குழுக்கள் (Working Groups) நிறுவப்பட்டன. பணிக்-குழுக்களின் செயற்பாடுகள் பற்றியும் கணித்தமிழ் மற்றும் இணையத்தமிழ் வளர்ச்சி பற்றியும் ஒவ்வொரு ஆண்டும் உத்தமம் நடத்திய தமிழ் இணைய மாநாடுகளில் கலந்தாய்வு செய்யப்பட்டுள்ளன. இதுவரை (1997-லிருந்து 2014-வரை) சிங்கப்பூர், தமிழ்நாடு,, மலேசியா, அமெரிக்கா, ஜெர்மனி, புதுச்சேரி ஆகிய நாடுகளில் 17 தமிழ் இணைய மாநாடுகள் நடைபெற்றுள்ளன. இப்பொழுது TIC 2020 பதினெட்டாவது மாநாடு இணையவழி நிகழவுள்ளது அறிய மிகுந்த மகிழ்ச்சி அளிக்கின்றது.

இந்த 25 ஆண்டுகளில் நாம் சாதித்தவை என்ன என்று எண்ணிப் பார்க்க வேண்டிய நேரம் இது. ஏராளமான தமிழ் இணையதளங்கள் பிறந்துள்ளன. பல்வேறு எழுத்துரு தரப்பாடுகளிலிருந்து இரண்டு எழுத்துருத்தரப்பாடுகள் - ஒருங்குறி தமிழ் (Unicode Tamil), அனைத்து எழுத்துருத்-தரப்பாடு (TACE-16), ஆகிய இரண்டு தரப்பாடுகள் தமிழக அரசால் அரசின் தரப்பாடுகளாக ஏற்கப் பெற்றுள்ளோம். சொற்செயலிகள், தமிழ் எழுத்துரு அறி மென்மம் (Tamil OCR), பேச்சுத்தமிழை எழுத்துத்தமிழாக்கும் மென்மம், எழுத்துத்தமிழை பேச்சுத்தமிழாக்கும் மென்மம், தமிழ் இயல் மொழிச்செயலாக்கம்

போன்ற பல்வேறு மென்பொருள்கள் உருவாக்கப் பெற்றுள்ளோம். உலகு தழுவிய வாழும் தமிழ் மக்களும், தமிழில் ஈடுபாடு உள்ள மற்றையோரும், தமிழ் மொழியைக் கற்கவும், தமிழர் வரவாறு, கலை, இலக்கியம், பண்பாடு பற்றி அறிந்து கொள்ளவும், மழலைக்கல்வி முதல் பட்டப் படிப்பிற்கான பாடப் பொருள்களையும், மிகப்பெரிய தமிழ் மின்நூலகத்தையும் தன்னுட்கொண்டு தமிழ்ப்பணி ஆற்றி வரும் சிறந்ததொரு தமிழ் இணையப் பல்கலைக்கழகம் கிடைக்கப் பெற்றுள்ளோம்.

ஆனால் இப்படைப்புகளெல்லாம் மக்களைப் போய்ச் சேர்ந்துள்ளனவா? தமிழை ஆட்சி மொழியாகவும் வழக்கு மொழியாகவும் கொண்டுள்ள நாடுகளின் அரசுக்கள் ஏற்று செயல்படுத்துகின்றனவா? இல்லையெனில் அதற்கு உத்தமம் என்ன செய்யவேண்டும்? எப்படிச் செயல்படவேண்டும்? என்பவற்றை இம்மாநாட்டில் கலந்தாய்ந்து முடிவு செய்யவேண்டிய நிலையில் உள்ளோம். . மேலும் இணையதள தமிழ் முகவரி வடிவமைக்கும் பணி இன்னும் நிறைவு பெறாமல் உள்ளது. இப்பணி நிறைவுபெற செய்ய வேண்டியவை பற்றிய முடிவெடுக்க வேண்டிய நிலையில் உள்ளோம்.

இணையப் பயன்பாட்டிற்கு ஒருங்குறி தமிழ் எழுத்துரு (Unicode Tamil) தரப்பாடு என்றும், பிற பயன்பாடுகளுக்கு அனைத்து எழுத்துரு (TACE-16) தரப்பாடு என்றும் தமிழக அரசு அறிவித்துள்ளது. எல்லாப் பயன்பாடுகளுக்கும் அனைத்து எழுத்துரு (TACE-16) தரப்பாடே சிறந்தது என்பதை பல்வேறு ஆய்வுகள் வெளிப்படுத்தியுள்ளன. இந்த அனைத்து எழுத்துரு (TACE-16) தரப்பாட்டின் பயன்பாடு கூடினால், ஒருங்குறி சேர்த்தியம் (Unicode consortium) ஒருங்குறி தளத்தில் 32-பிட்டு அமைப்பில் இந்த அனைத்து எழுத்துரு (TACE-16) தரப்பாட்டினை சேர்க்க இசைவளித்துள்ளது. எனவே, அனைத்து எழுத்துரு (TACE-16) தரப்பாட்டினை பல்வேறு பயன்பாடுகளில் செயல்படுத்தி அதன் பயன்பாட்டினை பெருக்குமாறு கணித்தமிழ் அன்பர்களுையெல்லாம் அன்புடன் கேட்டுக் கொள்கின்றேன்.

இன்று ஜப்பான், கொரியா போன்ற நாடுகளில் செயல்படும் கணிப்பொறிகளுக்கு ஆங்கிலம் தெரியாது. ஜப்பான், கொரிய மொழிகளில் கொடுக்கப்படும் கட்டளைகளை மட்டுமே புரிந்துகொண்டு செயல்படுகின்றன. உலக மொழிகளின் தாய்மொழியாகிய தமிழ் மொழிக் கட்டளைகளால் இயங்கும் கணிப்பொறியை வடிவமைக்க இயலாதா என்ன? அனைத்து எழுத்துரு (TACE-16) தரப்பாடு செய்யப்படும்வரை இயலா நிலை இருந்தது. இப்பொழுது அத்தடை இல்லை. இனி அப்படியொரு கணிப்பொறியைக் காண்பது எப்போழுது? Assembler போன்ற அமைப்புச் செயல்மொழி (System software) windows போன்ற இயக்க மென்பொருள் (Operating System) ஆகியவற்றை அனைத்து எழுத்துரு (TACE-16) தரப்பாட்டில் வடிவமைத்து ஒரு முழுமையான தமிழ்க்கணினியை படைக்க முடியும். இப்படியொரு முழுமையான தமிழ்க்கணினியை படைத்து வழங்குபவருக்கு உருபா ஒரு இலக்கம் பரிசு வழங்கப்படும் என்று மலேசியாவில் நடைபெற்ற 12-வது தமிழ் இணைய மாநாட்டின் போழ்தே அறிவித்திருந்தேன். ஆனால் இன்றுவரை அது நிகழவில்லை. விரைவில் அப்படியொரு முழுமையான தமிழ்க்கணினி பயன்பாட்டிற்கு வரவேண்டும் என்பது எனது பேரவா? என் கனவு நிறைவேறுமா?

பன்னாட்டுக் கணித்தமிழ்த் தகவல் தொழில்நுட்ப மாநாடு 2024 சிறப்பாக நடைபெற
எனது உளங்கனிந்த வாழ்த்துக்களை தெரிவித்துக் கொள்கின்றேன். வாழ்க தமிழ்!
வளர்க கணித்தமிழ்!

அன்புடன்,



(மு.பொன்னவைக்கோ)

1.6.2024



Dr. Nagamanickam Ganesan
NASA Johnson Space Center
Houston, USA
Chair, INFITT



02 June 2024

On behalf of the Executive Council of INFITT, I am excited to welcome the local Texas Tamil families and delegates from all over the world to this Tamil Information Technology conference in the beautiful campus of University of Texas, Dallas. This is yet another way for us to engage more with the local community networking, especially with all the families who speak Tamil. About 20% of the international students at UTD are from Tamilnadu.

There are hundreds of useful technical papers in the INFITT website presented during the Annual conferences. I was the Chairman for INFITT Unicode Committee to enhance Tamil encoding in the ISO 10646 standard, and we did it using the famous Madras Tamil Lexicon as the basis.

Subject Matter Experts will be talking about many aspects of Tamil Computing under development such as Artificial Intelligence and real-time translation between India's languages between speakers on mobiles. Our National Poet, Subramania Bharatiyar said this succinctly:

**காசி நகர்ப் புலவர் பேசும் உரைதான்
காஞ்சியில் கேட்பதற்கோர் கருவி செய்வோம்!**

With Natural Language Processing, AI, and crowd-sourcing from Philological Databases such as Project Madurai, the auto-translation of any two Indian languages between the ends of mobile phones can be made perfect with the involvement of INFITT. This academic conference will be its First step.

Welcome to Texas!

Dr. Nagamanickam Ganesan



Dr Mylswamy Annadurai
Distinguished Scientist
Director(R), ISRO Satellite Centre
Bengaluru, INDIA



வாழ்த்துச் செய்தி


உலகத் தமிழ் தொழில் நுட்பமன்றம் "தமிழ் கணினியம் மற்றும் தகவல்தொழில் நுட்பம்" என்ற தலைப்பில் முன்னெடுத்து நடத்தும் சர்வதேசக் கருத்தரங்கு பற்றி "தெருவெல்லாம் தமிழ் முழக்கம் முழங்கச் செய்வீர்" என்று பாடிய பாரதி எங்கும் கணினி எதிலும் கணினி என்று எல்லாம் கணினி மயம் என்றிருக்கும் இன்று நம்மிடையே இருந்திருந்தால், தனது கணீர்த் தமிழில், எப்படி கணினித் தமிழின் விசாலப் பார்வைக்கு ஒரு விலாசம் தந்திருப்பார்?

இயல், இசை நாடகம் என முத்தமிழை வளர்த்த தமிழ்ச் சங்கம் இன்று இருந்திருந்தால், கணினித் தமிழ் வளர்க்கும், தமிழ்த் தொழில் விற்பன்னர்களுக்கு சம இருக்கை கொடுத்து நாங்கு கால் பாய்ச்சலில் தமிழ் வளர வழி செய்திருக்குமோ?

ஆண்டி முதல் அரசு வரை அனைவருக்குமாக, அனைத்துக்குமாக அறம், பொருள் அறிவு சொன்ன வள்ளுவர் இன்றிருந்தால், செயற்கை நுண்ணறிவின் வழிபிறழா வளர்ச்சிக்கு ஈரடிகள் பலதையும் சீரடிகளாய்ச் செய்திருப்பாரோ?

இளங்கோவின் தமிழ்க் கணினிய காப்பியமும், கம்பனின் கணினிய காதையும் என்று தமிழ்ப் புலமைகளின் என்றைய நீட்சியாய் நம்முன் நிழலாடுகின்றன.

வள்ளுவன், இளங்கோ, பாரதி, கம்பன் என்ற தமிழ்ப் புலமைகள் இன்றைய தமிழ்க் கணினி முயற்சிகள் பற்றி அவரவர் பார்வைகளில் என்ன சொல்லுவார்கள் என்பது ஒருபார்வை. தமிழ், உலக அறிவியல் தொழில் நுட்ப வளர்ச்சியில் எப்படிப் பங்களிக்கிறது, இனி எப்படிப் பங்களிக்க வேண்டும் என்பது அவர்கள் மொழிகளாக இருந்திருக்கும் என்பது நமது கற்பனை. ஆனால் அறிவியல் தொழில் நுட்ப வளர்ச்சி கற்பனையில் உதித்து, செயலாக்கத்தில் உருப்பெறும் ஒரு நிலை. பல நூறு தமிழர்களின் பங்களிப்பு வேளாண் அறிவியல் தொழில் நுட்பத்திலிருந்து, கணிதம், இயல்பியல், உயிரியல், வேதியல், மருத்துவம், அணுவியல், வானியல் எனப் பலப்பல அங்கங்களில் இருப்பதை உலகம் அறியும். அந்தத் துறைகள் அனைத்திலும் கணினியிலின் பங்களிப்பு மிகுதியாகி வருகிறது. கணினியில் பதியப்படும் அனைத்து அறிவியல் கட்டுரைகளும் உடனே தமிழில் வருவது முதல், எந்த மொழியில் எங்கிருந்து பேசினாலும் அதை நம் தாய்மொழி தமிழ் வழி கேட்கும்படி செய்வது


(மயில்சாமி அண்ணாதுரை)

வாழ்த்துரை



உலகின் பல்வேறு துறைகளில் தொழில்நுட்பத்தின் பங்களிப்பு மிகுந்த கவனம் பெறுகிறது. தமிழ் மொழியில் தொழில்நுட்பத்தைப் புகுத்தி வளர்ச்சி அடையச் செய்யும் பணியில் தொடர்ந்து ஈடுபட்டு வரும் உலகத் தமிழ் தகவல் தொழில்நுட்ப மன்றத்தின் (INFITT) பணி பாராட்டுத்தக்கது. 2024 ஆம் ஆண்டு ஜூன் 14 முதல் 16 ஆம் தேதி வரை அமெரிக்காவில் உள்ள டெக்ஸாஸ் பல்கலைக்கழகத்தில் தமிழ் கணினி மற்றும் தகவல் தொழில்நுட்பம் பற்றிய சர்வதேச மாநாட்டினை (International Conference on Tamil Computing & Information Technology) நடத்தும் உத்தமம் மற்றும் அதன் உறுப்பினர்கள் அனைவருக்கும் வாழ்த்துகள்.

சக்தி குழுமங்களின் நிறுவனரான அருட்செல்வர் நா.மகாலிங்கம் அவர்கள், மொழி ஆராய்ச்சி, மொழிபெயர்ப்பு பதிப்புப் பணி, நூல் ஆக்கம், இதழியல் பணி என தமிழ் மொழியின் மீது தான் கொண்ட ஈடுபாட்டை பல்வேறு வடிவங்களில் வெளிப்படுத்தி சிறந்த தமிழ்ப் புரவலராகத் திகழ்ந்தார். இளைஞர்களின் முன்னேற்றத்தின் வாயிலாக தேசத்தை கட்டியெழுப்பும் வகையில் அருட்செல்வர் அவர்களால் தொடங்கப்பட்ட குமரகுரு கல்வி நிறுவனங்கள் 36,000 முன்னாள் மாணவர்களையும், 8000+ மாணவர்களையும் கொண்டு தொழில்நுட்பம், வேளாண்மை, பன்முகக் கலை அறிவியல் மற்றும் மேலாண்மை ஆகிய துறைகளில் மேலான கல்வியை வழங்கிவருகிறது. குமரகுரு வளாகத்தில் நா. மகாலிங்கம் தமிழாய்வு மையம் (N. Mahalingam Tamil Research Centre) தமிழ் மொழி ஆராய்ச்சிக்காக ஊக்கத்துடன் செயல்பட்டு வருகிறது.

இவ்வாய்வு மையத்தில் நிலவியல், தொல்லியல், மானுடவியல், இலக்கியம், கலையியல், நவீனவியல், ஆங்கில நூல்கள், இதழியல், ஆய்வேடுகள், அரிய நூல்கள் எனும் பிரிவுகளில் எண்பதாயிரத்திற்கும் (80,000) மேற்பட்ட நூல்களைக் கொண்ட ஆய்வு நூலகம் இயங்கி வருவதுடன், தமிழ் மொழி சார்ந்த பல்வேறு ஆய்வுகளுக்கு ஆண்டுதோறும் நிதி வழங்கப்படுகிறது. மேலும் பல்வேறு கருத்தரங்குகள் மற்றும் பயிலரங்குகளும் நடத்தப்பட்டு வருகின்றன.

தமிழ் மொழியின் மீது ஆர்வம் கொண்ட மாணவர்களுக்காக குமரகுரு பன்முகக் கலை அறிவியல் கல்லூரியில் இளங்கலைத் தமிழ் படைப்பாக்கம் (8.A. TAMIL CREATIVE WRITING) எனும் புதிய பாடத்திட்டம் உருவாக்கப்பட்டு நவீன யுக்கிகளுடன் பயிற்றுவிக்கப்படுகிறது. எங்கள் நிறுவனர் காட்டிய பாதையில் தமிழ் மொழியின் வளர்ச்சிக்காக பயணிக்கும் எங்கள் செயல்பாடுகளுக்கு மேலும் வலுசேர்க்கும் வகையில் 2023 ஆம் ஆண்டில், செயற்கை நுண்ணறிவுத் தொழில்நுட்பத்தை தமிழ் மொழியுடன் செயல்படுத்துவதற்கான சர்வதேச கருத்தரங்கை குமரகுரு கல்வி நிறுவன வளாகத்தில் உத்தமம் (INFITT) அமைப்பு, எங்களோடு இணைந்து நடத்தியதை தற்போது நான் நினைவு கூர்கிறேன்.

அக்கருத்தரங்கம் ஆய்வாளர்கள் மற்றும் தமிழ் ஆர்வலர்களிடையே நல்ல தாக்கத்தை ஏற்படுத்தியது. குமரகுரு வளாகத்தில் செயற்கை நுண்ணறிவுத் தமிழ் ஆய்வு மன்றம் தொடங்குவதற்கான அடித்தளமாகவும் அமைந்து, அதற்கான பணிகளில் மாணவர்கள் ஆர்வத்துடன் ஈடுபட்டு வருகின்றனர்.

அதே போல தற்போது நடைபெறும் இந்த சர்வதேச மாநாடு தமிழ் கணினி மற்றும் தகவல் தொழில்நுட்ப வளர்ச்சியில் குறிப்பிடத்தக்க தாக்கத்தை ஏற்படுத்தும் என நம்புகிறேன். இதன் ஒருங்கிணைப்பாளர்கள், கட்டுரையாளர்கள் மற்றும் பங்கேற்பாளர்கள் அனைவருக்கும் எனது வாழ்த்துகளைத் தெரிவித்துக் கொள்கிறேன்.

இக்கருத்தரங்கில் விவசாயம், மருத்துவம், பன்மொழி ரோபோக்களுடனான தொடர்புகள், வேர்ச்சொல்லாக்கம், இணையப் பாதுகாப்பு, மொழிமாதிரிகள் உருவாக்கம் மற்றும் பயிற்றுவித்தல் ஆகிய பொருண்மையில் வல்லுநர்களின் உரை அமைகிறது.

இவை கணினித் தமிழில் உள்ள பல்வேறு துறைகளில் வரவேற்கத்தக்க முயற்சிகள் மேற்கொள்ளப்படுவதற்கு வழிவகுக்கும். இக்கருத்தரங்கம் உலகின் மிகப் பழமையான மொழிகளில் ஒன்றான தமிழ் மொழியில் தற்காலத்திற்கேற்ப, கணினி மற்றும் தொழில் நுட்ப ஆய்வுகளைப் புகுத்தி, புதுமையான தாக்கங்களை உருவாக்கும் என நம்புகிறேன். கணினித் தமிழ் வளர்ச்சிக்கான இக்கருத்தரங்கம் மற்றும் சர்வதேச மாநாடு வெற்றி பெற வாழ்த்துகள்.



சங்கர் வாணவராயர்,
தலைவர் - குமரகுரு கல்வி நிறுவனங்கள்



கார்த்திகேய சிவசேனாபதி

தலைவர்

அயலக தமிழர் நல வாரியம்

குட்டப்பாளையம் சாமிநாதன் இல்லம்,
குட்டப்பாளையம், காங்கயம் வட்டம்,
திருப்பூர் - 638 108, தமிழ்நாடு.

Karthikeya Sivasenapathy

Chairman

Non Resident Tamils Welfare Board

Kuttapalayam Saminathan Illam,
Kuttapalayam, Kangayam Taluk,
Tiruppur - 638 108, Tamil Nadu.

☎ +91 74186 74257

☎ +91 96559 12456

✉ nrtwb.chairman@gmail.com

31-05-2024

To

The Organizers, ICTCIT 2024

International Forum for Information

Technology in Tamil (INFITT),

University of Texas at Dallas,

Richardson, Texas, USA. !



SUB: Wishing Success for the ICTIT 2024 Conference.

Dear Organizers,

On behalf of the Non-Resident Tamils Welfare Board, I extend my heartfelt wishes for the grand success of the upcoming International Conference on Tamil Computing and Information Technology (ICTCIT 2024), scheduled to take place from June 14-16, 2024, at the University of Texas at Dallas.

The INFITT's efforts to bring together researchers in Tamil Computing, Artificial Intelligence, and Information Technology are commendable. This conference, now in its 23rd year, serves as, a crucial platform for sharing advancements and fostering innovation in these fields.

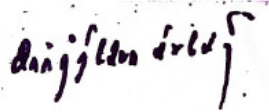
Our Honourable Chief Minister of Tamil Nadu, M.K. Stalin, envisions making Tamil Nadu a one trillion-dollar economy. This vision encompasses leveraging technology and innovation to drive economic growth and improve the quality of life for all citizens. The collaboration and knowledge exchange facilitated by events like ICTIT 2024 are vital to achieving this ambitious goal.

It is also fitting to remember and honour the legacy of Kalaignar Karunanidhi, who introduced India's first Information Technology policy. His pioneering vision laid the foundation for Tamil Nadu's leadership in the IT sector, a legacy that continues to inspire and drive us forward. The focus of ICTIT 2024 on Tamil Computing and Artificial Intelligence is particularly relevant today, as these technologies offer immense potential for bridging language barriers and enhancing communication. The innovative solutions being developed, such as real-time translation between Tamil and other languages, reflect Bharathiyar's dream of seamless communication across linguistic boundaries.

I also take this opportunity to appeal for support and participation from the global Tamil community.

Once again, I extend my best wishes for a successful and impactful conference.

Yours sincerely,



Karthikeya Sivasenapathy
Chairman,
Non-Resident Tamils Welfare Board

Chatbot Applications in Agriculture

Sobha L

AU-KBC Research Centre for Emerging Technologies
MIT Campus of Anna University, Chennai-600044

The need of conversational agents has become acute with the widespread use of personal machines with the wish to communicate and the desire of their makers to provide natural language interfaces (Wilks, 1999).

In 1966, Joseph Weizenbaum at MIT created the first chatbot that came close to imitating a human and it was called ELIZA. Given an input sentence, ELIZA would identify keywords and pattern match those keywords against a set of pre-programmed rules to generate appropriate responses. Since the development of ELIZA, there has been considerable research in the field of intelligent chatbots. The technology advancement in the field of Natural Language Processing (NLP) and in other fields of AI gave intelligent chatbots which allow humans to communicate in their language with computers. The architectures and retrieval processes of chatbots take advantage of advances in machine learning and deep learning to provide faster and reliable information retrieval processes, where responses are generated from the data given for learning. There are chatbots which have adopted generative models to respond; they use machine translation (MT) techniques to translate input phrases into output responses.

A chatbot is a software program that interacts with users using natural language. Different terms have been used for a chatbot such as: conversational agent, machine conversation system, virtual agent, dialogue system, and chatterbot. The purpose of a chatbot system is to simulate a human conversation. The architecture of a chatbot integrates a language model and computational algorithms to emulate informal conversation between a human user and a computer in natural language. Chatbots has wide range of applications in various domains such as education, e-commerce, hospitals and digital libraries.

After ELIZA (Weizenbaum, 1966) there has been several chatbot systems developed in seventies and eighties with wide range of new architectures such as: MegaHAL (Hutchens, 1996), CONVERSE (Batacharia et al., 1999), ELIZABETH (Abu Shawar and Atwell, 2002), HEXBOT (2004) and ALICE (2007). With the new algorithms and techniques of data-mining and machine-learning, decision-making capabilities, availability of corpora, robust linguistic annotations/processing tools chatbots have become more practical, with many commercial applications .

In this talk, I will present a practical chatbot application, showing that chatbots are found in daily life, such as in the domain of agriculture where farmers can get their produce prices, availability of fertilizers in the nearby stores, bank loans available and many more in their native language Tamil.

As said earlier, chatbot is a software designed and developed to converse with human in natural language. Chatbots are very useful in scenarios where the users adequately look for information from knowledge/information rich repositories such as webpages, database etc. It finds its application in B2B and B2C applications. Most of the chatbots are flow-oriented, where they try to spot the keywords and give answers which follow a predefined path. These are rules driven chatbots. The knowledge is limited and presented as different scenarios and paths are defined with rules. Sophisticated chatbots are built using artificial intelligence techniques such as Natural language Understanding (NLU) and Natural Language Generation (NLG). Chatbots can be categorized into following types.

Flow-Oriented Chatbot: These are popular and easy to develop. It has a predefined knowledge source in the form of rules. These are suitable for much focused tasks.

Artificial Intelligence Driven Chatbot: These are considered as most sophisticated chatbots. There are only a few AI driven chatbots. 'Mitsuku' is one of the present day popular AI driven chatbot. These chatbots are driven with NLU and NLG techniques.

Hybrid Bots: As the name suggests these chatbots are built using flow oriented architecture and a limited AI driven mechanism is used whenever the focus of the conversation is deviated for the specified domain.

Human Supported Bots: These are flow oriented chatbots, where the conversation is monitored by human. When the conversation becomes irrelevant the human intervenes and starts giving the answers and these answers are recorded in the knowledge base.

Chatbots are also used for other tasks such as gathering information in call centres and Customer Relation Management (CRM) systems. There are other applications such as automatic telephone answering systems, tools to aid in education, business and e-commerce. Chatbots are also employed in technical call centres, where it understands the requirement of the customers with a set of questions in the conversation and redirects the call to the appropriate executive. The chatbot platform which can handle Tamil conversations pertaining to agriculture analyses the intents of the user in Tamil and relevant responses to the intents are provided to the user in Tamil.

The application mentioned above requires natural language understanding by the system and hence requires deep analysis of language beyond the sentence level. A conversation is a dialogue that happens between two or more persons and it has one to one binding. The minimum unit of conversation is a pair of intent and response $\langle I, R \rangle$ and it can be defined as a cluster of sentences which are related to each other or bound to each other. Hence analysis beyond a sentence is necessary to understand a conversation and this comes under discourse analysis or text analysis.

Hence for any NLU application, discourse needs to be analysed. Recent research in textual information science has increasingly turned to text processing beyond sentence level. Text has

a rich structure in which sentences are grouped and related to one another in a variety of ways. To understand a text, one must understand the relations between its parts and determine how the various propositions fit together.

Halliday and Hasan (1976) suggested that a text is coherent as a result of cohesion which is defined as “a semantic relation between an element in the text and some other element that is crucial to the interpretation of it”. Central to cohesion is the notion of texture created by linguistic features such as reference, substitution, ellipsis, and conjunctions that create thematic relations between two or more clauses or within independent elements in the text. To understand the natural language these linguistic features needs to be analysed. Conversation which is central to a chatbot needs to be analysed at all the above mentioned levels. The chatbot system developed uses the following Methodology and Architecture.

The chatbot converse with human in free running Tamil language. The development of the chatbot are divided into two steps. First is the customization of the chatbot platform according to our need. Second in the core AI driven engines, which receives the user input, understands it, extract the answer from the knowledge base and present in natural language text. The AI engines comprises of three modules; a, Natural Language Understanding Engine Development, b, Natural Language Generation Engine and c, Knowledge base creation Engine. These modules are explained in brief.

Natural Language Understanding Engine: This engine understands the free running Tamil sentences, given as input by the user. These sentences are processed with syntactic processing modules, namely, morphological analyser, POS tagger, Chunker, Named Entity Recogniser and Clause boundary analyser. It is further processed with anaphora resolution engine to resolve the referential entities in the sentence. To resolve the anaphoric expressions, we use two preceding conversations are considered. After these processing of the sentence it is presented in a machine understanding form.

Natural Language Generation Engine: NLG engine has two parts. First part has to find the required answers for the input sentence which has been converted into machine understandable form. Second part is to present the information found from the knowledge base in near natural Tamil sentence construction. We require sentence construction engine and morphological generator engine for these generation task.

Knowledge Based Creation Engine: Knowledge base is the repository of the domain dependent information which is made available in machine understandable form. Here the source documents are converted into knowledge graphs such as conceptual graphs and made suitable for extracting information required for the input sentence from the user. Knowledge base development engine is a generic engine. This allows the customization of chatbot to any required domain.

The chatbot has other AI modules which are used for calculating the prices vegetables, fruits , grains etc. It has separate modules which generated graph for the commodity prices and collecting data from various sources. The agriculture chatbot in Tamil is a complete chatbot

where information related to agriculture needed by farmers, traders and administrators are provided in native Tamil language

References

Abu Shawar, B. and Atwell, E. (2002). A comparison between Alice and Elizabeth chatbot systems. Research Report 2002.19, University of Leeds – School of Computing, Leeds.

Artificial Intelligence Foundation (2007). The A. L. I. C. E. Artificial Intelligence Foundation. Published online: <http://www.alicebot.org> oder <http://alicebot.franz.com/>.

Batacharia, B., Levy, D., A., R. C., Krotov, and Wilks, Y. (1999). CONVERSE: a conversational companion. In Wilks, Y., editor, Machine conversations, pages 205–215. Kluwer, Boston/Dordrecht/ London.

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. English Language Series, London: Longman.

HEXBOT (2004). Hexbot chatbot website. Published online: <http://www.hexbot.com/>.

Hutchens, T. and Alder, M. (1998). Introducing MegaHAL. Published online: <http://cnts.uia.ac.be/conll98/pdf/271274hu.pdf>

Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. Communications of the ACM, 10(8):36–45

Wilks, Y. (1999). Preface. In Wilks, Y., editor, Machine Conversations, pages vii–x. Kluwer, Boston/- Dordrecht/London.

Building Tamil Treebanks

Kengatharaiyer Sarveswaran

Department of Computer Science, University of Jaffna, Sri Lanka.

Department of Linguistics, University of Konstanz, Germany.

sarves@univ.jfn.ac.lk

Treebanks are important linguistic resources, which are structured and annotated corpora with rich linguistic annotations. These resources are used in Natural Language Processing (NLP) applications, supporting linguistic analyses, and are essential for training and evaluating various computational models. This paper discusses the creation of Tamil treebanks using three distinct approaches: manual annotation, computational grammars, and machine learning techniques. Manual annotation, though time-consuming and requiring linguistic expertise, ensures high-quality and rich syntactic and semantic information. Computational deep grammars, such as Lexical Functional Grammar (LFG), offer deep linguistic analyses but necessitate significant knowledge of the formalism. Machine learning approaches, utilising off-the-shelf frameworks and tools like Stanza, UDPipe, and UUParser, facilitate the automated annotation of large datasets but depend on the availability of quality annotated data, cross-linguistic training resources, and computational power. The paper discusses the challenges encountered in building Tamil treebanks, including issues with Internet data, the need for comprehensive linguistic analysis, and the difficulty of finding skilled annotators. Despite these challenges, the development of Tamil treebanks is essential for advancing linguistic research and improving NLP tools for Tamil.

1.0 Introduction

Treebanks are important because they provide structured, annotated corpora that serve as crucial resources for training and evaluating Natural Language Processing (NLP) models. They offer detailed syntactic and sometimes semantic information, which helps in understanding the grammatical structure of languages.

These treebanks are then used to build tools called parsers, which are employed to parse sentences and obtain their syntactic analyses. Parsing is considered a core task in NLP and is crucial for enabling computational models to understand the syntax of a language.

Treebanks support the development of parsers, which are essential for applications like machine translation, sentiment analysis, and information extraction. For instance, the Penn Treebank provides detailed part-of-speech tagging, hierarchical structures capturing syntactic relationships, and a clear demarcation of phrases (e.g., noun phrases, verb phrases, prepositional phrases). A sentence such as "The teacher explained the complex topic to the students" can be annotated in a Penn Treebank-inspired way as follows:

```

(S
  (NP (DT The) (NN teacher))
  (VP (VBD explained)
    (NP (DT the) (JJ complex) (NN topic))
    (PP (TO to)
      (NP (DT the) (NNS students))))))

```

In this structure, NN, DT, JJ, etc., are part-of-speech tags. The bracket structure shows the hierarchical structure of the sentence. NP, VP, and PP are phrases. Although the Penn Treebank provides a framework for English syntactic annotation, it can be adapted to Tamil or other languages with specific adjustments to account for linguistic differences. For instance, the phrases sometimes need to be reordered to fit the structure. The Tamil translation of the sentence given above is shown in (1), and the syntactic tree for that would be (2).

(1) ஆசிரியர் சிக்கலான தலைப்பை மாணவர்களுக்கு விளக்கினார்
ācīriyar cikkalāṇa talaippai māṇavarkaḷukku viḷakkiṇār
 teacher complex.Adj topic.Acc student.Pl.Dat explain.Past.3SgEpi
 "The teacher explained the complex topic to the students"

(2) (S
 (NP (NN ஆசிரியர்))
 (VP (VB விளக்கினார்)
 (NP (JJ சிக்கலான) (NN தலைப்பை))
 (PP (TO மாணவர்களுக்கு))))

These annotated data for English and Tamil are useful for performing phrase structure alignment, and therefore, useful for developing machine translation applications, for instance.

Treebanks also facilitate linguistic research by providing data that can be used to test linguistic theories and hypotheses. Additionally, treebanks enable cross-linguistic studies and the comparison of syntactic phenomena across different languages. For instance, Futrell (2015) shows that Tamil has the highest degree of free subject and object order using a cross-linguistic analysis from treebanks available in the Universal Dependencies repository.

Although large language models (LLMs) are mostly trained using raw text, treebanks are crucial for evaluating whether LLMs capture syntactic nuances of languages. For instance, Tenney et al. (2019) evaluate the Bidirectional Encoder Representations from Transformers (BERT) model using a treebank to check its linguistic capabilities.

In summary, treebanks are essential resources for training and evaluating computational models and conducting linguistic studies.

In this paper, I discuss how we built Tamil treebanks using various approaches and some of the challenges encountered. The paper consists of three main sections: Building Treebanks, Discussion and Conclusion.

2.0 Treebank annotation formats

Several annotation schemes are used in different treebanks, including the Penn Treebank annotation (Marcus, Marcinkiewicz, & Santorini, 1993), the Prague Dependency Treebank (PDT) annotation (Hajič et al., 2001), the Paninian Dependency framework (Bharati & Sangal, 1993; Begum et al., 2008), and the Universal Dependencies annotation (Nivre et al., 2016). Each of these schemes captures various levels of syntactic and sometimes semantic information and is backed by various linguistic theories.

The primary classification of these schemes can be divided into phrase structure, as followed in resources like the Penn Treebank, and dependency schemes, as used in Universal Dependencies. Analysing the pros and cons of these schemes is beyond the scope of this article.

Among the available formalisms, the dependency grammar formalism is particularly useful for languages such as Tamil, which are morphologically rich and have relatively variable and less rigid word order (Bharati et al., 2009). Since sentences can be constructed with various word orders, phrase structure rules can easily break down. To accurately capture even a simple sentence, multiple phrase structure rules are often required.

Let's take the example sentence in Tamil as shown in (03).

(03)	[வெளியுறவுத்துறை அமைச்சர்]	[மரியாதை நிமித்தமாக]	[அன்வரை]	[சந்தித்தார்]
	<i>veliyuravutturai amaicar mariyātai</i>	<i>nimittamāka</i>	<i>aṇvarai</i>	<i>cantittār</i>
	foreign-depart.Nom minister.Nom	courtesy out-of.Adv	anwar.Acc	meet.Past.3SgEpi
	Foreign Minister paid a courtesy call on anwar			

The sentence in (03) is in the form of NP_{subj} ADVP NP_{obj} V, and this phrase order can be changed in Tamil, for instance, to NP_{obj} NP_{subj} ADVP V or ADVP NP_{subj} NP_{obj} V. To effectively capture this simple sentence, you would need to write at least three different rules in the phrase structure formalism.

On the other hand, a single dependency structure can capture this free-word order nature, as shown in the figure below. Even if the word-order changes, the same dependency rules apply.

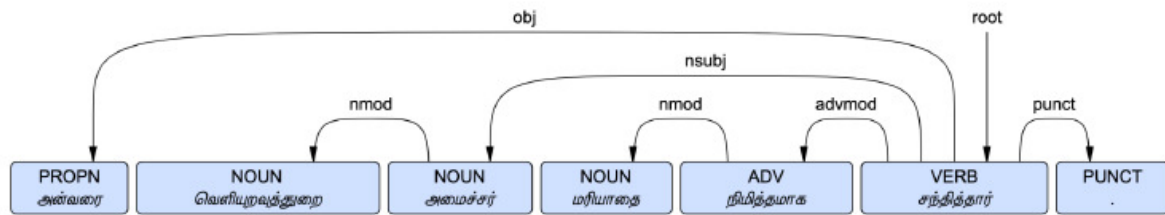


Figure 01: The Universal Dependencies based dependency structure for (03)

There are other linguistically rich and deep grammar formalisms also used to create dependency treebanks. For instance, Lexical-Functional Grammar (LFG) (Kaplan & Bresnan, 1981) provides both a constituency (c-structure) and a dependency (functional structure or f-structure) representation. Figure 02 shows an analysis of a simple construction using the LFG formalism. As illustrated in Figure 02, the constituency structure (c-structure) is represented in the form of a tree, while the dependency structure is shown as an attribute-value matrix in the functional

structure (f-structure). Even if the word order changes, the f-structure remains the same, whereas the c-structure will change.

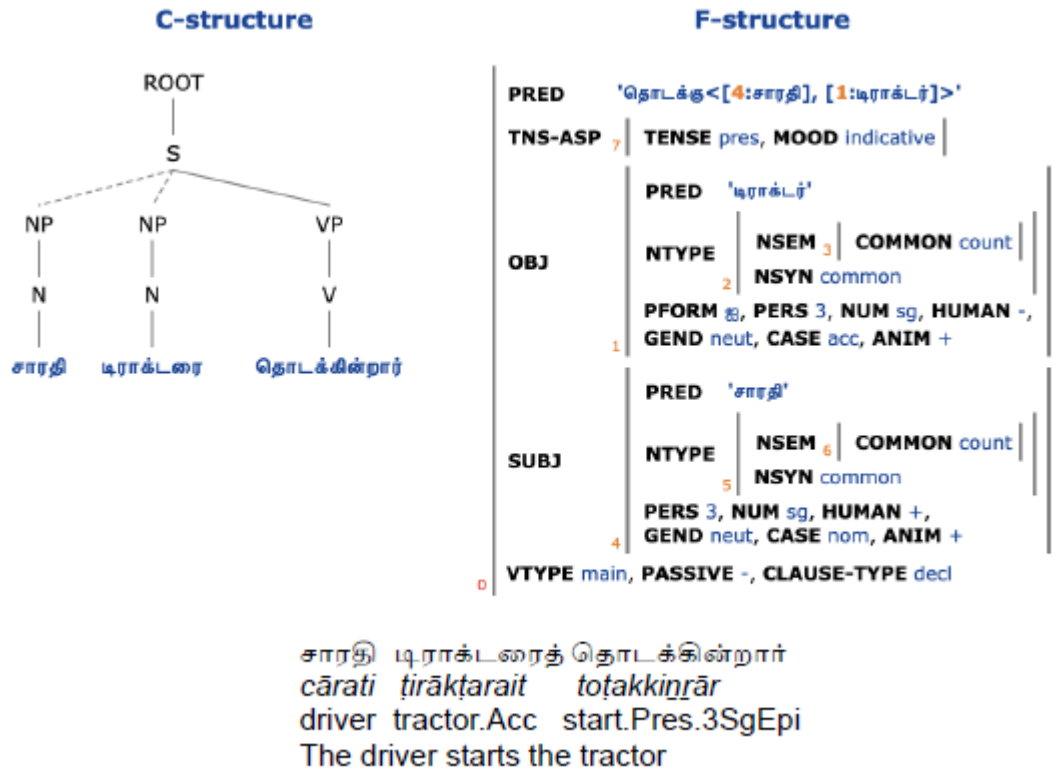


Figure 02: An example for a Lexical Functional Grammar analysis

2.3 Building treebanks

This section outlines the creation of Tamil treebanks using three distinct approaches: manual annotation using the Universal Dependencies, computational grammars, and machine learning techniques. There are several annotation schemes available, as mentioned earlier. In this section, I will discuss the schemes I used to annotate treebanks, as examples.

2.3.1 Building treebanks using manual annotation

Treebanks can be created manually by trained annotators and linguists. For instance, the Modern Written Tamil Treebank and the Aalamaram Tamil Treebank were created manually. This is a time-consuming and tedious process, requiring annotators to have extensive training and an in-depth understanding of Tamil linguistics.

The typical manual annotation process involves selecting and cleaning the data. Then, annotators are trained to annotate the data using an annotation guideline. For languages like Tamil, such guidelines may not exist initially and need to be bootstrapped. Typically, an initial version of the guidelines is taken from another language and then iteratively adapted and refined for Tamil during the annotation process.

Currently, the Universal Dependencies (UD) framework is widely used to build treebanks for various languages. In the latest version of Universal Dependencies, 283 treebanks have been

created using 161 languages worldwide. Some of these treebanks are created based on specific themes. For instance, there is a Vedic treebank created by an institute in Switzerland.

At least four efforts have been made to create a Universal Dependencies treebank for Tamil. One such treebank was initially created using another dependency scheme called the Prague Dependency Treebank (Ramasamy, 2011), and then it was converted automatically using a script. Therefore, there are some flaws in the treebank. The Modern Written Tamil Treebank (MWTT) by Krishnamurthy and Sarveswaran (2021) was created using examples extracted from grammar books. This MWTT consists of 600 sentences. Recently, two other large-scale Universal Dependencies treebanks, each with roughly 100,000 tokens, have been created by two groups of researchers (Abirami et al., 2024).

Table 01 shows the basic information captured in Universal Dependencies. However, all this information can be extended to capture language-specific features. The initial set of features was proposed by researchers considering cross-lingual and multilingual processing. For instance, Abirami et al. (2024) extended this specification to capture Named Entities in the Tamil treebank. The authors used the Misc field to include NER annotation. This field was previously used to include the transliteration of forms and lemmas of the respective sentence.

Table 01: An example of the Universal Dependency annotation (CoNLL-U format)

ID	Form	Lemma	POS	XPOS	Morph-features	Rel	Deprel	Misc
1	பையன்	பையன்	NOUN	_	Case=Nom Gender=Masc Number=Sing Person=3	4	nsubj	-
2	சாவியால்	சாவி	NOUN	_	Case=Ins Gender=Neut Number=Sing Person=3	4	obl:inst	-
3	கதவைத்	கதவு	NOUN	_	Case=Acc Gender=Neut Number=Sing Person=3	4	obj	-
4	திறந்தான்	திற	VERB	_	Gender=Masc Mood=Ind Number=Sing Person=3 Polarity=Pos Tense=Past VerbForm=Fin Voice=Act	0	root	-
5	.	.	PUNCT	_	PunctType=Peri	4	punct	-

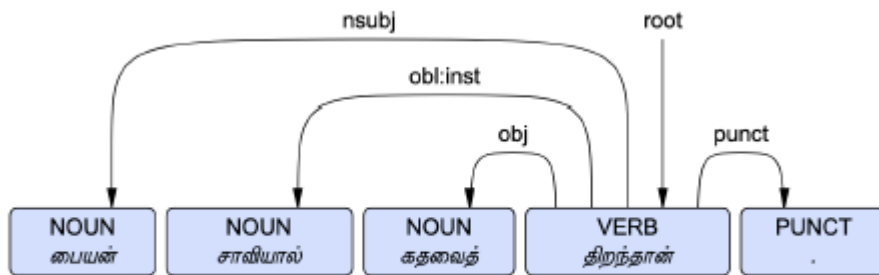


Figure 03: A dependency graph for the annotation given in Table 01.

2.3.2 Building treebanks using Computational Grammars

Grammar-based treebank development is not very popular because it requires significant linguistic knowledge, familiarity with grammar formalisms, and the ability to model these formalisms using computers to generate treebanks. These grammar-based annotations are primarily used for linguistic analyses and comparative studies. There are a few modern linguistic formalisms, such as Head-Driven Phrase Structure Grammar (HPSG) and Lexical

Functional Grammar (LFG), that are actively used to develop grammars for various languages. These two formalisms are also called deep grammar formalisms because they can model deep syntactic structures that are invariant among languages and provide ways to handle various syntactic transformations using syntactic rules, such as passive and dative shifts.

An effort has been made to build a Tamil Lexical Functional Grammar, which is still a work in progress. Currently, the grammar has been implemented to parse very simple sentences taken from elementary Tamil books. Additionally, sentences were obtained from the Parallel Grammar project (ParGram)(Butt et al, 2002), which aims to develop parallel LFG grammars for several languages worldwide to support cross-lingual analyses.

Lexical Functional Grammar is a useful formalism and is now being widely used for various levels of linguistic analysis beyond morphosyntax, including prosody and semantics. However, there is still a long way to go in this respect for Tamil grammar. The environment used to write LFG grammars is called the Xerox Linguistic Environment (XLE). In this environment, phrase structure rules, lexical rules, and lexicon entries are used to build the complete grammar. It is important to note that the XLE environment also supports the integration of morphological analysers developed using Finite-State Transducers (FST), which increases the robustness of the grammar.

Once the grammar is in place, it can be converted into a parse bank (Sulger et al, 2013), which consists of all the annotations along with the sentences. Platforms like the Infrastructure for the Exploration of Syntax and Semantics (INESS) host such parse banks and treebanks, providing access to parallel analyses (Rosén, 2012). There have also been attempts to convert parse banks generated using Lexical Functional Grammar into other treebank formats, such as the Universal Dependencies.

2.3.3 Building Treebanks Using Machine Learning Approaches

Treebanks can also be built using machine learning or deep learning approaches. Several off-the-shelf tools are available for building treebanks using Universal Dependencies parsing, including Stanza, UDpipe, and UUParser. These tools can annotate given sentences using the Universal Dependencies framework. However, to train these parsers, annotated Tamil data is necessary. One approach, called multilingual parsing, helps train a parser for a language using data from similar languages. The effectiveness of this method depends on the accuracy of the data from the other languages.

We built a deep learning-based parser called *ThamizhiUDp* (Sarveswaran & Dias, 2020) using deep learning and a widely used low-resource language processing technology called multilingual processing. Instead of using a machine learning or deep learning approach to annotate the treebank end-to-end, we employed various tools to create the annotations through a multi-stage approach.

First, existing POS-tagged corpora were collected, and a POS tagger was trained using available POS-tagged data to perform part-of-speech tagging with Stanza (Peng. et al, 2020). Then, based

on the POS information, ThamizhiMorph (Sarveswaran et al., 2021), a finite-state morphological analyzer, was used to include UD morphological features in the data. Subsequently, a UUParser-based multilingual parser (Smith, 2018) trained using Hindi, Arabic, Telugu, and Turkish languages was used to annotate the dependency information. We also manually annotated a dataset to validate the parser for syntactic coverage and the parsing accuracy.

Data quality and the amount plays a crucial role in multilingual parsing. For example, although Telugu is in the same language family as Tamil, the parser gave better results when trained with Hindi data, as the Hindi treebank contains a substantial amount of manually annotated data. Some of these experiments were reported in Sarveswaran & Dias (2020).

Several off-the-shelf Universal Dependency parsers are also available online for Tamil. However, since these are trained using the data available online, the quality of the analysis is often very poor.

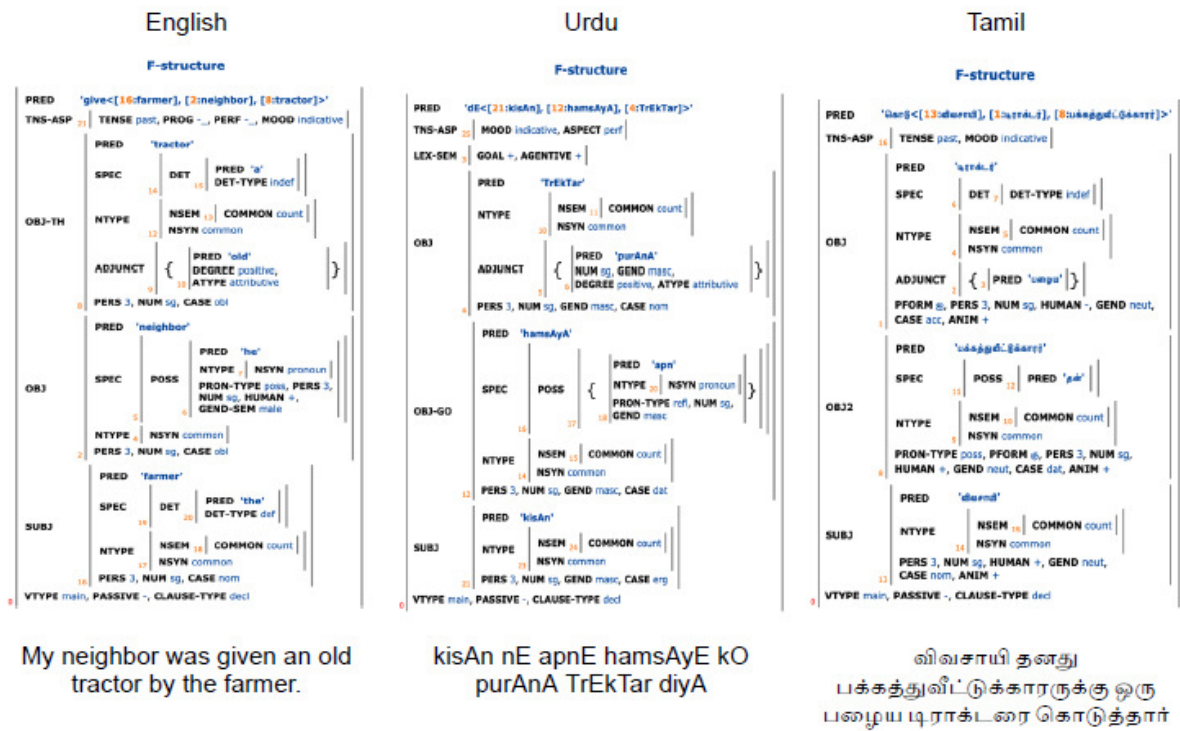


Figure 04: Parallel analyses for “My neighbor was given an old tractor by the farmer” in English, Urdu and Tamil using the Lexical Functional Grammar formalism

4.0 Discussion

We faced numerous challenges while building Tamil treebanks. This section highlights two such challenges related to data and the annotation process.

4.1 The Nature of Internet Data

The sentences taken from grammar books were well-structured and clean. However, the data found on the Internet presented several issues, including the following:

- **Code-mixed Data:** Many sentences include a mix of Tamil and other languages, complicating the parsing and annotation processes. In addition to the language code-mix, there was also junk data found in the dataset, such as HTML tags.
- **Spelling and Grammar Mistakes:** Internet data often contains many spelling and grammatical errors that can complicate the annotation process. Annotators always struggle to get the context in such cases and decide whether to annotate incorrect sentences or not. However, sometimes it is also important to annotate incorrect sentences to train the model effectively.
- **Fragments:** Incomplete sentences or sentence fragments are common, making it difficult to provide accurate syntactic and/or semantic annotations.
- **Dialects:** Tamil has various dialects, some of which are not easily understandable or standardised, posing additional challenges for consistent annotation. It also depends on the treebank design whether to annotate dialects or not.

While these problems are common for any application development using Internet data, they are particularly challenging for treebank creation, where we seek linguistic insights. For instance, syntactic ambiguities can lead to multiple parse trees. Handling code-mixed data is also challenging.

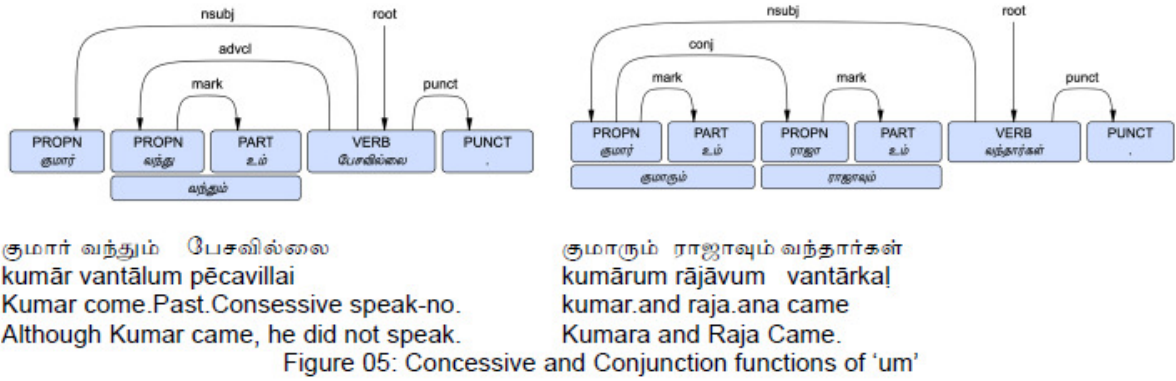
4.2 Linguistic Analysis

This is the most significant challenge we face when building treebanks. Annotators need an in-depth linguistic understanding to accurately annotate sentences. However, since there is no modern comprehensive grammar available for Tamil, identifying analyses can be difficult. For instance, light verb constructions are challenging because identifying and annotating their parts in Tamil can be complex for annotators. Additionally, mixed categories present difficulties; certain constructions in Tamil, such as *vinaiyaalannaiyum peyar*, exhibit both nominal and verbal features, complicating their categorisation and annotation (Butt et al., 2020). Handling and annotating these requires deep linguistic understanding.

Very long sentences found in formal writing are also problematic. For instance, we encountered sentences with more than 40 tokens. When the number of tokens increases, marking dependencies becomes very challenging for humans, even though most of these are central embeddings or modifiers.

Tamil words often contain packed linguistic information. To annotate them, we need to break them into individual pieces that can be marked for syntactic information. For example, we tokenize clitics from the forms to mark their syntactic roles. For instance, Figure 05 shows how *-um* is tokenised to mark concessiveness and the conjunction.

Furthermore, people tend to stack more and more tokens to form compounds, complicating language processing. In such cases, we need to break them into multiple tokens to capture their syntactic information. However, breaking such tokens is not straightforward due to the nature of Unicode encoding and the abugida writing system. Additionally, deep linguistic knowledge is required to break them accurately.



Identifying good annotators passionate about linguistic annotation is very challenging. While we can find native Tamil speakers, not many have an understanding of linguistic phenomena or it was hard to train them. Moreover, many linguists today prefer other branches of linguistics, and few are interested in studying language structure.

Despite these challenges, the development of Tamil treebanks is crucial for advancing linguistic research and improving NLP tools for Tamil. The efforts to overcome these challenges contribute to more robust linguistic resources.

5.0 Conclusion

In this paper, I have briefly outlined three approaches for building treebanks: manual annotation, grammar-based parsing, and machine learning approaches. Each of these methods captures different levels of information. However, the amount of information they can capture can be expanded using language-specific features. A significant advantage of formalisms and annotation schemes discussed is their expandability. Among these, Lexical Functional Grammar (LFG)-based parsed treebanks are highly deterministic and built on a solid linguistic foundation. Therefore, they are particularly useful for linguistic analyses.

Furthermore, these treebanks are crucial in the current era of Large Language Models (LLMs) as they can be used to fine-tune and evaluate these models, making the models more effective for a wide variety of tasks. Although there have been some efforts reported in creating treebanks, there is still a long way to go. More studies related to Tamil linguistics are needed, especially to capture contemporary Tamil and dialectal variations.

Acknowledgement

Some of these reported works were carried out in collaboration with many scholars around the world, including Miriam Butt, Gihan Dias, Parameswari Krishnamurthy, Keerthana Balasubramani, A. M. Abirami, Wei Qi Leong, Hamsawardhini Rengarajan, D. Anitha, R. Suganya, and many others.

References:

- Abirami, A. M., Leong, W. Q., Rengarajan, H., Anitha, D., Suganya, R., Singh, H., ... Shah, R. R. (2024, May). Aalamaram: A Large-Scale Linguistically Annotated Treebank for the Tamil Language. In G. N. Jha, S. L., K. Bali, & A. K. Ojha (Eds.), *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation* (pp. 73–83).
- Begum, R., Husain, S., Dhvaj, A., Sharma, D. M., Bai, L., & Sangal, R. (2008). Dependency annotation scheme for Indian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.
- Bharati, A., & Sangal, R. (1993, June). Parsing free word order languages in the Paninian framework. In *31st Annual Meeting of the Association for Computational Linguistics* (pp. 105–111)
- Bharati, A., Gupta, M., Yadav, V., Gali, K., & Sharma, D. M. (2009, August). Simple parser for Indian languages in a dependency framework. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)* (pp. 162–165).
- Butt, M., Rajamathangi, S., & Sarveswaran, K. (2020). Mixed Categories in Tamil via Complex Categories. In M. Butt & I. Toivonen (Eds.), *Proceedings of the LFG'20 Conference, On-Line* (pp. 68–88).
- Butt, M., Dyvik, H., King, T., Masuichi, H., & Rohrer, C. (2002). The Parallel Grammar Project. In *COLING-02: Grammar Engineering and Evaluation*.
- Futrell, R., Mahowald, K., & Gibson, E. (2015, August). Quantifying word order freedom in dependency corpora. In *Proceedings of the third international conference on dependency linguistics (Depling 2015)* (pp. 91–100).
- Kaplan, R. M., & Bresnan, J. (1981). *Lexical-functional grammar: A formal system for grammatical representation*. Massachusetts Institute Of Technology, Center For Cognitive Science.
- Krishnamurthy, P., & Sarveswaran, K. (2021, December). Towards building a modern written tamil treebank. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)* (pp. 61–68).
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020a. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.
- Ramasamy, L., & Žabokrtský, Z. (2012, May). Prague dependency style treebank for Tamil. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1888–1894).
- Sarveswaran, K., & Dias, G. (2020, December). *ThamizhiUDp: A Dependency Parser for Tamil*. In P. Bhattacharyya, D. M. Sharma, & R. Sangal (Eds.), *Proceedings of the 17th International Conference on Natural Language Processing (ICON)* (pp. 200–207).

Sarveswaran, K., Dias, G., & Butt, M. (2021). *ThamizhiMorph*: A morphological parser for the Tamil language. *Machine Translation*, 35(1), 37-70.

Smith, A., Bohnet, B., Nivre, J., de Lhoneux, M., Stymne, S., & Shao, Y. (2018). 82 treebanks, 34 models: Universal dependency parsing with cross-treebank models. Retrieved from <https://research.google/pubs/82-treebanks-34-models-universal-dependency-parsing-with-cross-treebank-models/>

Sulger, S., Butt, M., King, T., Meurer, P., Laczko, T., Rákosi, G., Dione, C., Dyvik, H., Rosén, V., De Smedt, K., Patejuk, A., Çetinoğlu, ., Arka, I., & Mistica, M. (2013). ParGramBank: The ParGram Parallel Treebank. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 550–560). Association for Computational Linguistics.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. An open infrastructure for advanced treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco (eds.) *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey, May 2012.

Making of ‘soul’ and ‘body’ - A technological pursuit in the world of Artificial Intelligence: Interacting to Multilingual Robots - Scopes and Future

Vasu Renganathan
University of Pennsylvania

vasur@sas.upenn.edu

(http://robot.tamilnlp.com/multilingual_robot.html)

Introduction:

In my efforts to build a robot that can speak and understand Tamil along with accessing information in Tamil from databases and express in both written and spoken form, I do make a constant progress. Thanks to the evolving AI technology with the efforts of many companies and institutions, including Google, Microsoft, Open AI, Facebook and so on. My earlier papers on this topic were published in the proceedings of the international conference on Tamil Computing conducted by the International Forum for Information Technology in Tamil (INFITT). (cf.

http://www.uttamam.org/papers/21_32.pdf, http://www.uttamam.org/papers/13_11.pdf and <http://uttamam.org/papers/tic2023.pdf>). It has been a very challenging pursuit, and these attempts made some progress over the period of time in the way how a machine can be used to interact with in human languages. I present here the recent development of this research showing enormous progress over my previous attempts. What is presented here is different from the earlier attempts in that this research makes use of the advances of Google’s speech recognition and text to speech technology (cf. <https://cloud.google.com/text-to-speech>) in a much more efficient manner than before. As a result, the robot that is described here is enabled to deal with multiple languages, not restricted to Tamil. All of what is discussed in this paper can be extended to other languages also without any restrictions whatsoever. In a way, the robot that is attempted here is multilingual in nature with the ability to switch between languages through the spoken medium.



Single board computer and Linux Operating system:

With the advent of single board computers such as Raspberry PI and use of an efficient operating system such as Linux along with powerful programming languages like Python, PHP etc., it is now possible to make use of all the technological advancements in a portable platform so building robots can be possible containing all of these resources. What is introduced in this work is a small box capable of moving around with four wheels and can be commanded fully by voice. With the possibility to connect to internet and making a local network, it becomes possible to connect to internet to retrieve relevant resources and make them available to the users over the voice technology as well as textual medium in the local webserver. With the use of MySQL database and Apache webserver, it becomes possible to build a local network where the robot can interact with constantly and retrieve commands as stored in the database.

Salient features of the multilingual robot:

1) This Robot can interact in any language from one's choice, provided corresponding table for commands are already supplemented using the portal from local webserver. So far, it is ready to be used for Tamil, English, French, Hindi, Telugu and Malayalam. The input voice is passed over to Google's speech technology application through WIFI and transforms into text so the robot can make use of it in a meaningful manner based on the input text. The application of this robot is exponential provided the scope is well thought out with sufficient changes in this application.

2) Movements: Once a language is chosen, we can start interacting with it in the respective language. We can give commands like go forward, backward, circle around, go to kitchen and order coffee, take this pencil to the living room and so on. Corresponding Tamil commands are: 'முன்னால போங்க', 'பின்னால போங்க', 'போய்க்கிட்டே இருங்க', 'சமையலறைக்குப் போய் காப்பி சொல்லுங்க', and so on. We can also record a set of movements and ask it to go to kitchen and come back; get coffee from kitchen, distribute these snacks to guests and so on. These can be done in all of the languages that Google supports English, Telugu, Malayalam, Japanese etc. (<https://cloud.google.com/translate/docs/languages>):

3) Translation: We can ask it to translate to other languages. Once we tell the target language, it will translate all what you say in the target language using the Google's Translation technology. The relevant commands in Tamil for this are: 'ஹிந்திலெ சொல்லுங்க', 'குஜராத்திலெ சொல்லுங்க' and so on. In English the corresponding commands are: 'translate to Hindi', 'translate to French' and so on. The command 'translate to English' brings back to the monolingual state.

4) Making quizzes: It can pronounce difficult expressions in the language we interact with and ask you to repeat. So, lines from poems, complex spoken expressions, vocabularies etc., can be included in the form of a quiz. It switches to quiz mode and quizzes you. New quizzes can be added interactively from a localhost webpage. Some of the tests that are stored for Tamil are: 'திருக்குறள் தேர்வு', 'புறநானூறு தேர்வு'. Corresponding commands in English are: 'difficult expressions', 'English test 1' and so on.

5) Audio files stored locally or on the internet can be accessed and played through any command we add in local server's portal. What is available in the dialogue form in the Tamil learning website <http://learn.tamilnlp.com> can be accessed with corresponding commands such as: 'தொட்டுக்க என்ன இருக்கு', 'ரொம்ப வெலெ சொல்றீங்க' and so on. In this respect, any audio files including music, dialogues etc., that are freely available on the internet can easily be accessed using this resource.

6) One can save their schedules and things to do, and it will remind you whenever you start the conversation with your name in the language of your choice. One can also add more reminders for it to remind you at a later time. Relevant commands in Tamil for this purpose are: நான் என்ன செய்யணும், நீக்கிவிடுங்கள், பதிவு செய்யுங்கள் and so on. In English, the corresponding commands are 'add to my schedule', 'what is my schedule', 'remove my schedule' and so on.

7) Wikipedia resources can be accessed as needed: It is possible to get information from Wikipedia and report it in the language of our choice. The command relevant to it in Tamil is using the phrase 'ஐப் பற்றி சொல்லுங்க' as in மயிலாடுதுறை பத்தி சொல்லுங்க, மதுரைத்

திட்டம்ப் பற்றிச் சொல்லுங்க and so on. In English, the relevant command can be built with phrases like ‘tell me about...’, ‘what is...’, ‘who is ..’ and so on.

8) Tell a joke: It says jokes getting randomly from a text file/json file. Tamil command for it is ‘ஜோக் சொல்லுங்க’ and the English command for it is ‘tell me a joke’.

9) It is a speech companion, and it will interact with anyone in any language of one’s choice. For example commands in English can be listed as follows: What is your name? How are you? Go forward, go backward, circle around, what are the examinations you have, difficult expressions, my name is vasu, translate to Hindi/Tamil/Telugu etc., did you eat?, who is the first President of the United States, Who is the first Prime minister of India, switch language and so on.

10) Using JSON files to retrieve text: Literary texts and other relevant texts in JSON format can easily be accessed with this application. So far, the Tamil literary texts such as ‘Tirukkural’, ‘Kuruntokai’, ‘Purananuru’ as stored in JSON format and can be accessed with corresponding ids. For example, ‘திருக்குறள் 25’, ‘குறுந்தொகை 45’, ‘புறநானூறு 235’ etc., would access corresponding poems and play in spoken format, correspondingly display them in text format in the local webserver.

This robot can do all these functions in almost all the languages that Google supports, but each language resource needs to be customized using the user-friendly portal as needed. It is possible to customize this tool for other languages by making suitable changes from the table for English. See <https://cloud.google.com/translate/docs/languages> for all the languages that Google supports for translation, text to speech and speech to text. This Robot uses all these resources and offers an experience of interacting with a machine in the form of "man-machine interaction partner".

The following images show how the local webserver is built to store, edit and control the robot.

Fig. 1 Local webserver’s customizable page showing English commands:

Fig. 2 Local webserver’s customizable page showing Tamil related commands:

The leftmost column lists all of the available resources such as customizing each language with relevant introductory commands, commands for quizzes and question/answering systems, adding new language, displaying the users’ commands and the responses of this robot and so on. The links to the commands of each language can be customized in their own scripts, so this application can use Google’s text to speech technology to convert them to speech form. The middle column is meant for listing the commands the users would use and the third column is meant to include the corresponding Python commands along with adding texts for quizzes and question answering. For example, to make a quiz, the text should start with the word ‘teach:’ and words to be pronounced must be listed with the delimiter comma. The word ‘text:’ in the front of the text would indicate the application is to read out aloud the text that follows using text to speech library. Thus, the application is written in such a way that the text that is entered in this page are customizable accordingly.

be listed as below: a) a pedagogical companion to teach languages to improve spoken skills along with literary knowledge, b) a home companion to do simple tasks to move around the house and perform some errands such as making announcements, dealing with one's daily schedule, using internet audio resources and so on, c) with sufficient improvement on interacting with database, it is possible to make this application a machine learning device to continuously build its knowledge base by interacting with human.

References:

Google's Text technology: <https://cloud.google.com/translate/docs/languages>.

Renganathan, Vasu (2023). Large Language Models (LLM) and the Role of linguists in the World of AI. In the Proceedings of the International Conference on Tamil Computing, Coimbatore: Kumaraguru Technical Institute of Technology.

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).

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).

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)

AI and cyber security

Dr. P Jayashree,

Anna University, KBC Research Center, MIT, Chennai

Internet, social networks and digital technologies, leads to a data driven era where data plays a vital role in our lives for major decision making and deriving insights for data centric business. On the other hand computationally intensive applications keep emerging that necessitates GPU and parallel architectures. With the availability of large data and computational power in recent years, there is a remarkable growth in the fields of Artificial Intelligence (AI), Machine learning (ML) and Natural language processing (NLP). Similarly, another growing domain that receives critical attention is Cyber security. In this digital age, with smart devices and sophisticated technologies, cyber security is no more a choice but a mandate to be prioritized. Cyber landscape is constantly evolving with a wider access to tools and technologies by adversaries. Thanks to the advancements in artificial intelligence and machine learning technologies, AI tools like chatbots, virtual assistants, recommendation systems and navigation applications make our lives easy. At the same time, they also present security challenges. The generative AI systems and LLMs though have tremendous potential in automating and assisting intelligent tasks, pose security risks that need to be taken care of. The recent release of AI risk management framework by NIST emphasizes the need for rapid research into this new direction involving a synergy between AI and security.

Cyber landscape

Cyber threats span across a large spectrum ranging from data breaches to sophisticated cyber-attacks, happening globally every day. The first form of attack is an email spam that happened in ARPANET in 1900s followed by review spams in commercial websites and phishing with the growth of social networks in 2000s. With proliferation of AI, AI based spams keep growing..

All cyber attacks happen by exploiting the vulnerabilities in hardware and software. Major vulnerabilities include weakness in input validation, memory overflow, memory corruption, SQL injection, Cross site scripting and directory traversals leading to various web based attacks. The impact varies from simple code bypassing to information leak to Dos attacks. Common vulnerabilities and exposures are listed in many public vulnerability databases make awareness of known information security vulnerabilities. Malware attacks that are malicious programs to gain access to system and network resources to cause harm or access useful information from the resources. They range from simple worms, virus, trojan horses to adware, spyware, ransomware to sophisticated botnets that consist of many compromised computing resources to launch attacks. Denial of service attacks(DoS) are attacks from a single source or multiple sources to flood the network with requests aiming at halting e-mail, web and other services. Phishing and spoofing attacks target users to make them reveal sensitive information through disguised text or e-mail or voice messages. IBM Security Threat Intelligence report, 2023,

phishing is the most common type of malware in 41% of security breaches and data breaches. Code injection attacks include SQL injection and XSS attacks that leverages system vulnerabilities to inject malicious codes. Currently, social engineering attacks keep increasing, exploiting human emotions and characters like fear, greed or sympathy to collect information through phishing, tailgating or pretexting. With the growth of 5G and IoT, compromising IoT devices to form bot nets is simplified. AI powered attacks like adversarial AI, AI generated social engineering, dark AI and Deep Fake attacks are easily launched leveraging AI tools and techniques.

Role of ML and NLP in defensive security

In recent years, adoption of AI techniques for addressing cyber security needs in various applications is increasing. ML and NLP models are built for various security tasks such as threat detection, malware analysis, and anomaly detection. Attack detection solutions broadly fall under two category namely signature matching or anomaly based methods. Basically both methods look for known or unusual data patterns to conclude on attack and type of attack. ML models can be trained to look for these patterns for classification of the data.

The performance of any ML model heavily depends on the training data. Data preprocessing and feature engineering are the crucial steps for building good models. Most of the threat and attack identification are attributed to known or learned. ML models can be trained to analyze vast amounts of data to identify patterns and anomalies that may indicate malware, intrusions, or unauthorized access attempts. In the literatures quite a number of classification algorithms like Random Forests, Support Vector Machines (SVM), and Logistic regression are used for cyber threat detection. These supervised learning algorithms are trained on labeled data like malware samples to classify unknown samples as benign or malicious data. Similarly clustering algorithms like K-Means and Hierarchical Clustering are used for anomaly detection, by grouping unusual patterns or outliers in data that may indicate cyber threats. With data volume no longer a constraint, deep learning (DL) networks are increasingly utilized for intrusion detection and malware and attack classification. As DL models can learn complex patterns and automatically extract features from raw data, they have better efficiency for detecting sophisticated threats. Most of the researchers in this domain used CNNs for low level and high level feature extraction followed by different DL networks for further classification tasks. With generative AI networks like GANs, there is a visible superiority in the performance on feature extraction and classification tasks. Ensemble models are predominantly employed in research for better attack detection and classification. Reinforcement learning algorithms are used in adaptive cyber defense, where the system learns to make trial-and-error actions for optimizing the intended results.

NLP also plays a role in cybersecurity defense, for detecting and responding to threats more effectively. NLP techniques including named entity recognition, sentiment analysis, and topic modeling are used to analyze unstructured data like emails and social media to identify potential threats. Malware reports, security logs, incident reports and other textual data are analyzed to identify malware patterns, and incidents. Similarly emails and URLs can be

analyzed for finding phishing attacks. One of the key applications of NLP is text classification wherein NLP models are trained for detecting spam and fake reviews, that increasingly exists in e-commerce and online platforms. Fake reviews can also be identified applying sentiment analysis methods based on review sentiments or anomaly detection using anomalous contents or review behavior patterns. With the emergence of large language models, such as GPT and BERT, identifying deceptive language patterns, suspicious links, and other indicators of malicious intent, extracting relevant information, summarizing key insights from large unstructured data becomes promising and effective. AI enabled security systems can be devised to automatically respond to detected threats and vulnerabilities isolating compromised systems, patching vulnerabilities, or reconfiguring settings to save time and damages.

Security threats to NLP/ML models

Though ML and NLP models provide potential security defense solutions, they pose risks and are subjected to cyber threats and attacks. Some of the common threats and possible ways to address them are discussed. Attacks can be classified based on attack timing, attack goals and attacker information. The information security attacks happen during model development and model deployment as well. Most common attacks during model development phase are data poisoning attack, model extraction attack and backdoor attack. Attackers inject malicious samples into the training data, causing the model to learn undesirable behaviors. Data sanitization is an important task to be carried out during preprocessing that can avoid data poisoning. During reverse engineering attacks, attackers can attempt to steal or reconstruct the trained model by querying the system and analyzing its outputs. Building robust models with data augmentation and adversarial training helps in evading these attacks. Data anonymity and differential privacy can be applied to training data and noise can be added to the model's outputs to protect sensitive information.

Machine learning and deep learning models are vulnerable to adversarial attacks. Attacks that are targeted during deployment includes adversarial attacks and evasion attacks in addition to out of distribution generalization. Adversaries can attempt perturbations in input text data to mislead the ML/NLP model, causing it to make incorrect predictions or expose sensitive information. Noise attacks and patch attacks can be deployed on image data by adversaries. Adversarial training, gradient masking and data normalization can help resolving these attacks. Another important attack to be addressed is privacy attack where sensitive information, such as personal data or proprietary information are extracted from the training data or model parameters by adversaries. If data is not properly anonymized or protected, it could lead to privacy breaches. Differential privacy and data anonymization are the techniques used to evade privacy attack. Further, NLP models can inherit and amplify biases present in their training data, leading to unfair outputs and decisions.

Building language models involve multiple stages of processing namely lexical analysis, syntax analysis, semantic analysis, discourse analysis and ..., and each stage is susceptible to different types of attacks. To mitigate these security risks, secure model training, robust model architectures, input validation, privacy-preserving techniques are the recommended solutions.

While AI can enhance cybersecurity defenses, it can also be used by attackers to develop more sophisticated threats.. Adversarial machine learning techniques are used to study and defend against AI-based cyber attacks, such as adversarial examples or model evasion attacks. However, it is also important to address the potential risks and challenges associated with the use of AI in cybersecurity, such as data quality, model bias, and adversarial attacks.

References:

1. Jha, Akshita, and Chandan K. Reddy. "Codeattack: Code-based adversarial attacks for pre-trained programming language models." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 12. 2023.
2. Xu, Han, et al. "Adversarial attacks and defenses in images, graphs and text: A review." International journal of automation and computing 17 (2020): 151-178.
3. Shayegani, Erfan, et al. "Survey of vulnerabilities in large language models revealed by adversarial attacks." arXiv preprint arXiv:2310.10844 (2023).
4. Cheng, Pengzhou, et al. "Backdoor attacks and countermeasures in natural language processing models: A comprehensive security review." arXiv preprint arXiv:2309.06055 (2023)
5. Huang, Yujin, et al. "Training-free lexical backdoor attacks on language models." Proceedings of the ACM Web Conference 2023. 2023.
6. Sai, Siva, et al. "Generative ai for cyber security: Analyzing the potential of chatgpt, dall-e and other models for enhancing the security space." IEEE Access (2024).
7. M. Mahfuri et al., "Transforming Cybersecurity in the Digital Era: The Power of AI," 2024 2nd International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 2024, pp. 1-8, doi: 10.1109/ICCR61006.2024.10533072
8. Ismail, Walaa Saber. "Threat Detection and Response Using AI and NLP in Cybersecurity.", 2023.
9. Sufi, Fahim. "An innovative GPT-based open-source intelligence using historical cyber incident reports." Natural Language Processing Journal 7 (2024): 100074.
10. X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He and Z. Shen, "Deep Stable Learning for Out-Of-Distribution Generalization," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 5368-5378,
11. <https://www.crowdstrike.com/cybersecurity-101/cyberattacks/most-common-types-of-cyberattacks/>
12. <https://www.cvedetails.com/>

LLMs Should be Aligned, yes, but to Which and Whose Values?

Monojit Choudhury

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi

Email: monojit.choudhury@mbzuai.ac.ae

As Large Language Models and their applications become more ubiquitous across domains (Chui et al., 2023) from marketing and sales to product R&D and software engineering, from healthcare to education, numerous such ethical decisions must be taken every moment. Imagine an LLM deployed to help respond to and moderate conversations on an online forum for HIV+ youths in Africa (Karusala et al., 2021) or one that helps farmers in India to decide whether inorganic or organic pesticides are good for their context (Barik, 2023). In this talk, I will argue that LLMs should not be designed and developed to work with specific moral values because as a generic model, they are expected to be used for a variety of downstream applications, to be deployed across geographies and cultures, and used by a heterogeneous group of end-users.

The moral stance taken during the decision-making process, which could even mean whether to show a specific auto-complete suggestion or not, should be decided by various actors involved during the application development, deployment, and usage phases. LLMs should be capable of generic and sound ethical reasoning, where given a situation and a moral stance, it should be able to resolve the dilemma whenever possible or ask for more specific inputs on the moral stance that are necessary for resolving the dilemma. In other words, I would like to argue against value alignment of LLMs, and instead make a case for generic support in LLMs for value alignment at application development stage or by the end-user.

Due to their lack of transparency, a host of ethical issues related to LLMs and downstream tasks built on top of them have been brought out by researchers (Bender et al., 2021; Basta et al., 2019). There have been efforts towards alignment of LLMs to avoid inappropriate, offensive or unethical use. However, due to value pluralism, as we demonstrated in our recent work (Rao et al., 2023; Khandelwal et al., 2023), extensive alignment is rather detrimental to the ethical reasoning ability of the models. An emerging and more suitable practice is to either build application-specific content filters and post-processing modules (Del Vigna et al., 2017; Ji et al., 2021), or to embed the moral principles and ethical policies in prompts (Schick et al., 2021). While the former is limited in power and its ability to generalize across tasks, the latter depends on the ethical reasoning ability of the underlying LLM. Instead, in our work, we propose a framework to specify ethical policies in prompts and a systematic approach to assess the ethical reasoning capability of an LLM. The framework consists of carefully crafted moral dilemmas reflecting conflicts between interpersonal, professional, social and cultural values, and a set of ethical policies that can help resolve the dilemmas one way or the other. The framework is agnostic to and therefore, can support different approaches to normative ethics, such as deontology, virtue and consequentialism, and policies can be specified at different levels of abstraction.

Our study of 5 SOTA models in the GPTx series including GPT-4 and ChatGPT suggests that: (a) the ethical reasoning ability of the models, in general, improves with their size with GPT-4 having nearly perfect reasoning skills, (b) GPT-3 and ChatGPT have strong internal bias towards certain moral values leading to poor reasoning ability, (c) reasoning ability of the models depend heavily on the language one presents the ethical dilemma and policies with poorer performance for low resource languages; and (d) most models, including GPT-4, exhibit bias towards democratic and self-expression values that are mainly observed in Western and English-speaking societies over traditional and survival values that are characteristic of Global South and Islamic cultures (Inglehart and Welzel, 2010).

There are important repercussions of these findings for designing ethically versatile and consistent future LLMs.

[This abstract heavily quotes and borrows from Section 1 of Rao et al. (2023).]

References

- Soumyarendra Barik. 2023. Chatgpt on whatsapp: Govt's bhashini initiative to use ai for beneficiaries of welfare schemes.
- Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.
- Emily M. Bender, Timnit Gebru, Angelina McMillanMajor, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Michael Chui, Eric Hazan, Robert Rogers, Alex Singla, Kate Smaje, Alex Sukharevsky, Lareina Yee, and Rodney Zemmerl. 2023. The economic potential of generative AI: The next productivity frontier.
- Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.
- Ronald Inglehart and Chris Welzel. 2010. The WVS cultural map of the world. World Values Survey.
- Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2021. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.
- Naveena Karusala, David Odhiambo Seeh, Cyrus Mugo, Brandon Guthrie, Megan A Moreno, Grace JohnStewart, Irene Inwani, Richard Anderson, and Keshet Ronen. 2021. “that courage to encourage”: Participation and aspirations in chat-based peer support for youth living with hiv. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17.
- Aditi Khandelwal, Utkarsh Agarwal, Kumar Tanmay, and Monojit Choudhury. 2024. Do Moral Judgment and Reasoning Capability of LLMs Change with Language? A Study using the Multilingual Defining Issues Test. In *Proceedings of the 18th Conference of the European*

Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2882–2894, St. Julian’s, Malta. Association for Computational Linguistics.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13370-13388.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP.

Making LLMs that Can Speak Tamil

Mr. Raju Kandasamy,
Thoughtworks, Coimbatore, India

Abstract

The Tamil community has effectively preserved its rich literary heritage by digitizing palm leaves, inscriptions, and out-of-print books. Over the past few decades, significant progress has been made in defining digital standards such as TSCII and Unicode for Tamil, enabling native speakers to participate in the digital space. However, there are several challenges in making these digital archives usable for the current AI-dominated generation. This paper discusses the challenges faced in building Tamil LLMs (Large Language Models) and the advancements made in this area. Additionally, it demonstrates a small Llama model built from scratch, capable of writing traditional Tamil venba fluently. Emphasizing the need for a dedicated foundational model for Tamil, the paper explores why building such a model from a Tamil-only dataset is crucial. The primary reasons include pretraining biases from other languages, confined token vocabulary, and the potential loss of Tamil's rich grammatical rules in multilingual models. Existing models often use translated content for finetuning, resulting in grammatical, coherence, and fluency errors, and employ suboptimal tokenization processes. This paper proposes a foundational model that addresses these issues by leveraging a curated, rich Tamil-only dataset and including all forms of Tamil: Iyal, Isai, and Nadagam. Furthermore, it underscores the importance of curating a Tamil culture-specific finetuning dataset with Tamil-specific practices.

Introduction

Tamil literature and culture have been meticulously preserved and brought into the digital age. Despite this progress, the integration of this rich heritage into contemporary AI applications presents several challenges. This paper aims to identify these challenges and propose solutions to develop a foundational Tamil LLM that can effectively serve the Tamil-speaking community.

Challenges in Building Tamil LLMs

Pretraining Bias from Other Languages

Multilingual models are pretrained on datasets containing multiple languages, leading to inherent biases. These biases affect the accuracy and relevance of the model's output when applied to Tamil, as the model may inadvertently incorporate linguistic and cultural elements from other languages.

Confined Token Vocabulary

The token vocabulary in multilingual models is often limited, preventing the model from fully capturing the intricacies of the Tamil language. This limitation affects the model's ability to generate coherent and contextually appropriate text in Tamil.

Cultural and Grammatical Integrity

Internet-scale pretraining can introduce non-Tamil cultural references and practices, diluting the authenticity of the generated content. Additionally, the rich grammatical rules of the Tamil language are often overlooked, resulting in outputs that lack linguistic accuracy and cultural relevance.

Limitations of Existing Models

Use of Translated Content for Finetuning

Many existing models rely on translated content for finetuning, leading to grammatical, coherence, and fluency errors. These errors stem from the inherent differences between languages, which translation processes may not adequately address.

Suboptimal Tokenization

Current models often use character-level tokenization or other suboptimal methods, resulting in fragmented and contextually incorrect outputs. A more effective approach involves hand-built tokenization by Tamil scholars and linguists, ensuring adherence to grammatical rules.

Proposed Foundational Model

Accessibility and Sustainability

To ensure widespread accessibility, the proposed foundational model should be built as a Small Language Model (SLM) that can operate in resource-constrained environments. This approach also promotes sustainability by reducing the computational resources required for model training and deployment.

Curated Tamil-Only Dataset

A key component of the proposed model is a curated, cleaned, and rich Tamil-only dataset. This dataset should include OCR-scanned and archived documents, covering all three forms of Tamil: Iyal (literature), Isai (music), and Nadagam (drama).

Culture-Specific Finetuning Dataset

Incorporating a Tamil culture-specific finetuning dataset is essential. This dataset should include texts that reflect Tamil-specific practices and cultural nuances, ensuring that the model's output is contextually and culturally appropriate.

Conclusion

Developing a dedicated foundational model for Tamil is crucial for preserving and promoting Tamil literature and culture in the digital age. By addressing the challenges of pretraining biases, confined token vocabulary, and maintaining cultural and grammatical integrity, the proposed model aims to provide accurate and culturally relevant outputs. The inclusion of a curated Tamil-only dataset and a culture-specific finetuning dataset will further enhance the model's effectiveness and accessibility.

References

1. Abhinand Balachandran. "Tamil-Llama: A New Tamil Language Model Based on Llama 2" arXiv, 2023, eprint: 2311.05845.
2. Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. "Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model." arXiv, 2024, eprint: 2402.07827.

Text to Sound: Train your own LLM

Dr. Ruby Annette Jayaseela Dhanaraj,

GenAI Principal Architect & Senior Member, IEEE, Texas, USA

It was scientifically established that ragas in Indian music evoke a range of emotional responses, from happiness and calmness to tension and sadness. Indian music operates on the concept of ‘nava rasas,’ which include shringara (romance), hasya (humor), karuna (grief), raudra (anger), veera (heroism), bhayanaka (fear), vibhatsa (disgust), adbhuta (wonder), and shanta (peace). Each raga predominantly expresses one of these emotions. Researchers from the National Brain Research Centre in Haryana, India, and the University of Reading, UK, found that ragas like ‘Desh’ and ‘Tilak Kamod’ made listeners feel calm during the slower, arrhythmic alaap and happy during the faster, rhythmic gat. Similarly, ‘Shree’ and ‘Miyan ki Todi’ evoked feelings ranging from sadness to tension. In Carnatic music, ragas such as Amrithavarshini, Reethigowlai, Nattai, and Abheri are generally considered happy, whereas ‘Shivaranjani’ and ‘Vasantha Bhairavi’ often convey sadness, depending on how they are performed.

This article discusses the approaches and methods applied by an open-source research on ‘Text to Sound’ for generating guitar sounds and how the lessons learnt in the process can be leveraged to succeed in our new initiative in collaboration with Dr. Lalitha Jawahar, Assistant Professor from the “Tamil Isai Kallori”, Chennai. The new collaborative initiative is to develop an AI model capable of creating Carnatic music based on the emotions conveyed through text in video scene description and how it can be used to in pyschotherapeutic interventions. Also explore new AI advancements like ChatGPT and the QLoRA model for improving previous approaches for generative AI.

Addressing AI Challenges faced in Generating Sound from Musician Commands

The core challenge was enabling an AI system to generate specific guitar sounds based on a musician's voice commands. For example, if a musician says, "Give me a bright guitar sound," the AI must understand the context and produce the appropriate timbre. Words like ‘bright’ have different meanings in general contexts but signify a particular quality in music.

Dataset Challenges and Solutions

Challenge 1: Creating a Guitar Music Domain Dataset

A major hurdle was the absence of datasets specifically for guitar music. To tackle this, the team created their own dataset. This dataset included conversations among musicians about guitar sounds, sourced from platforms like Reddit, but expanded further using data augmentation techniques. They used BiLSTM deep learning models to generate contextually enriched datasets.

Challenge 2: Annotating Data and Creating a Labeled Dataset

Another challenge was annotating the data to create a labeled dataset. General datasets that train large language models like ChatGPT needed fine-tuning for specific domains. The word “bright,” for instance, might refer to light or sound quality. Using the annotation tool Doccano, musicians labeled data with terms related to instruments and timbre qualities. An active learning approach was also applied to automate part of the labeling process, leveraging domain expertise.

Challenge 3: Modeling as an ML Task – NER Approach

Deciding on the appropriate modeling approach was also crucial. The team chose Named Entity Recognition (NER) to identify and extract music-related entities. They used spaCy’s NLP pipeline, incorporating transformer models like RoBERTa from HuggingFace. This allowed the AI to understand the context-specific meanings of terms like “bright” and “guitar” in the music domain.

Model Training Challenges and Solutions

Overfitting and Memory Issues

During model training, overfitting was a major concern due to limited data. Overfitting occurs when a model performs well on training data but poorly on unseen data. To mitigate this, the team used data augmentation, creating multiple test sets, including context-based ones, to ensure the model’s robustness. They also faced memory issues with spaCy and addressed these by splitting the training set into parts and training them separately.

Model Performance and Accuracy

Ensuring real-world performance was critical. Despite limited training data, the model consistently achieved over 95% accuracy, thanks to the pre-trained RoBERTa model and spaCy’s capabilities. Tests with various datasets, including context-based and content-based ones, confirmed the model’s efficacy.

Standardizing Named Entity Keywords

Real musicians provided feedback indicating a wide range of vocabulary for sound and music, some of which were non-standard terms like “temple bell.” The team created a solution called standardizing named entity keywords, mapping these terms to standardized keywords using domain experts and clustering methods such as cosine and Manhattan distances. This bridged the gap between the musicians’ language and the AI’s training data.

Future Approaches with ChatGPT and QLoRA Model

ChatGPT for Data Collection and Annotation

ChatGPT can assist in data collection, annotation, and pre-processing. Its text generation capabilities reduce the effort needed for data gathering and annotation, making it a valuable tool in the early stages of model development.

QLoRA Model for Efficient Fine-Tuning

The QLoRA model, with its ability to quantize large language models to 4 bits, offers an efficient way to fine-tune models, reducing memory usage without compromising speed. Fine-tuning with low-rank adapters helps preserve most of the original model's accuracy while adapting it to domain-specific data, providing a cost-effective and faster alternative to traditional fine-tuning methods.

Leveraging Vector Databases

Using vector databases like Milvus or Vespa can help find semantically similar words, enhancing the model's performance by identifying contextually relevant terms beyond simple word-matching algorithms.

Conclusion

The lessons learnt from this open source project through the challenges of dataset preparation, annotation, and model training led to innovative solutions and valuable insights is being leveraged in our new initiative to develop an AI model capable of creating Carnatic music based on the emotions conveyed through text described about the video scene. With advancements like ChatGPT and QLoRA, we have powerful new tools to address these challenges more effectively.

Speech Recognition and speech synthesis fundamentals, challenges and case studies

B. Bharathi

Department of Computer Science and Engineering
Sri Siva Subramaniya Nadar College of Engineering
Kalavkkam, Chennai, India

Introduction

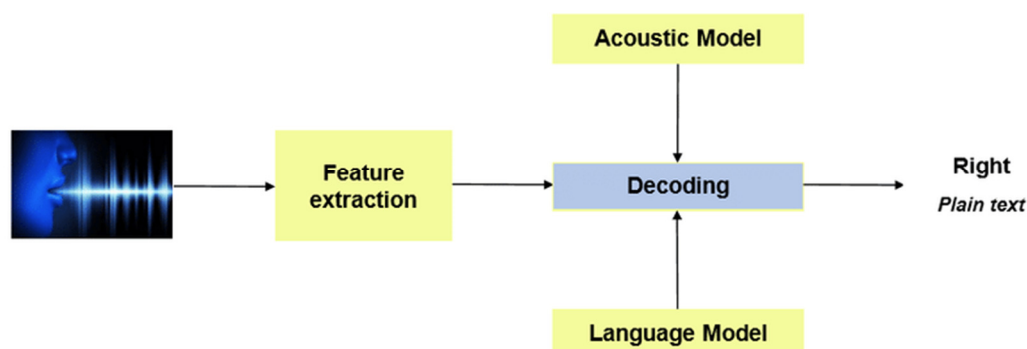
Speech, as a form of communication, is fundamental to human interaction and has several advantages over other forms of communication. Speech recognition and speech synthesis are two fundamental components of natural language processing (NLP) systems that enable computers to understand and generate human speech, respectively. Speech recognition, also known as automatic speech recognition (ASR) or speech-to-text (STT), involves the process of converting spoken language into text. It allows computers to understand and interpret spoken commands, queries, or conversations. Speech recognition systems employ sophisticated algorithms to analyze audio signals, extract relevant features, and map them to textual representations. These systems find applications in various domains, including voice assistants, transcription services, language learning, accessibility tools, and automation. Speech synthesis, also known as text-to-speech (TTS), is the process of generating spoken language from text input. It enables computers to produce human-like speech output, allowing for auditory communication with users. Speech synthesis systems utilize linguistic and acoustic models to convert textual information into speech waveforms, incorporating aspects like pronunciation, intonation, and expressiveness. Applications of speech synthesis range from assistive technology for visually impaired individuals to navigation systems, entertainment media, language learning tools, and personalized audio content creation.

Automatic speech recognition

Speech recognition is a crucial component of natural language processing that has gained widespread attention due to its numerous applications. This tutorial provides an in-depth look at speech recognition technology, covering fundamental concepts, signal processing techniques, feature extraction methods, acoustic modeling, and language modeling.

Fundamentals of ASR

In order to translate spoken language into text, ASR systems usually go through many stages:



1. Signal Processing

Using a microphone, record the speech signal, then transform it into a digital representation as the initial step. Pre-processing is applied to this digital signal to extract features, normalize it, and reduce noise. Key features such as Mel-Frequency Cepstral Coefficients (MFCCs) are extracted, as they effectively represent the phonetic content of the speech.

Different frequencies are blended to form our speech. Specifically, the Fast Fourier Transform (FFT) technique may be used to break down the signal into its component frequencies and perform an effective analysis of the data. Spectrogram creation is one method for visualizing the audio data. The frequency fraction transformer (FFT) is used to separate the audio signal into time frames and then into individual frequency components.

2. Acoustic Modeling:

This stage involves the use of statistical models to depict the relationship between the extracted features from the audio speech signal and the phonetic units of speech (phonemes). For this, Hidden Markov Models, or HMMs, have been extensively utilized. However, nowadays more modern systems use deep learning models such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs).

3. Language Modeling

Getting the corresponding text output from continuous speech input is the primary goal of automatic speech recognition. However, the most likely word combinations cannot be discovered only by the acoustic model. Contextual information is needed for this, and it can be taught to the model.

To estimate the likelihood of a word sequence, ASR systems make use of language models. The language model includes information specific to a given language including terminology, syntax, and sentence structure. N-gram models are widely used. By using probabilities of single words, ordered pairs, triples, etc., N-grams approximate the sequence

probability with the chain rule. Combining these probabilities with the ones from the Acoustic Model helps in eliminating language ambiguities from the sequence options and provides a more accurate text representation of the utterance.

But more recent advanced techniques like Recurrent Neural Network Language Models (RNNLMs) and Transformer-based models (e.g., BERT) are used to capture the syntactic and semantic structure of language.

4. Decoding

The last step decodes the most likely word order from the speech input by combining acoustic and language models. This is figuring out the optimal route across a network of potential word sequences, frequently with the use of algorithms like the Viterbi algorithm.

Challenges in ASR

Even with great progress, ASR systems still have a number of issues that affect their accuracy and usefulness:

1. Diversity in Speech

Accents, dialects, speaking speeds, and intonation are only a few examples of the many variations in human speech. It is challenging for ASR systems to generalize across many speakers and circumstances because of this heterogeneity.

2. Background Noise

Ambient noise and overlapping speech (in busy areas, for example) can severely impair ASR system performance. To lessen these impacts, effective speech augmentation and noise reduction techniques are crucial.

3. Homophones and Ambiguity

One of the challenges faced by ASR systems is the existence of homophones, which are words with similar sounds but distinct meanings, such as "two" and "too". For such statements to be well understood, context is essential.

4. Limited Data for Low-Resource Languages

Large volumes of training data have benefited ASR systems for frequently spoken languages like English, but many low-resource languages lack enough transcribed data, which makes it difficult to construct reliable ASR models for these languages.

5. Computational Resources

ASR systems, especially those built on deep learning, need a lot of processing power to train and run. For businesses with limited access to high-performance computer equipment, this might be a hindrance.

Evolution of ASR

1950s-60s: Isolated word and number recognition.

1970s: 1000 word recognition

1980s: New model of speech recognition - thousands of words of continuous speech.

1990s: 1st commercial recognition system with advent personal computers

2000s: Google voice search- users searches 230 billion words

2010s: ...

- IBM watson
- Apple Siri on iPhones
- Google Assistant on 400+ millions devices
- Microsoft cortana on windows device
- Amazon Alexa - integrated with more hardware and software for smart home use.

Promising Research Directions

Many interesting research for future study are being investigated in an effort to advance automatic speech recognition technology:

- **Unsupervised Learning:** Especially for low-resource languages.
- **Transfer Learning:** Using huge datasets with pre-trained models to enhance performance on certain tasks or languages with sparse data.
- **End-to-End Systems:** Streamlining the speech recognition process by creating systems that translate speech to text instantly, eliminating the need for intermediate human processes and perhaps producing more accurate and efficient models.

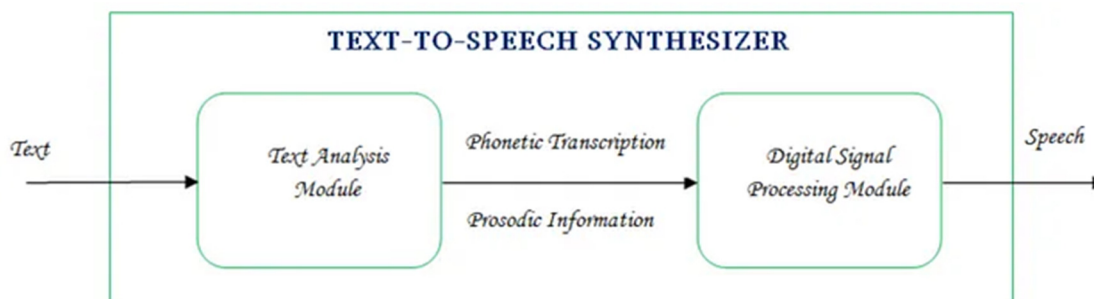
Similar to how convolutional neural networks (CNNs) extract characteristics from image data, deep neural networks-in particular, RNNs can extract pertinent information from spectrograms. For sound unit sequencing, an RNN and Connectionist Temporal Classification (CTC) layer combination can take the role of the acoustic model. Even though end-to-end DNNs could still contain language mistakes, the system's performance can be further enhanced by rescoring the probabilities of spelling and context through the use of N-grams or Neural Network Language Models (NLM) trained on large amounts of text.

Speech Synthesis

The process of producing human speech artificially is called speech synthesis, or text-to-speech (TTS). Automating customer service, virtual assistants, assistive technologies, and other sectors are among the fields in which this technology which speaks written text has applications. Producing expressive, comprehensible, and natural speech from text is the aim of voice synthesis technology.

Fundamentals of Speech Synthesis

Speech synthesis incorporates a number of essential elements:



1. Text Analysis

Text normalization, which involves formatting the input text into a standard format, is the first step in the process. This phase involves processing numbers, clearing out ambiguities, and extending abbreviations. Linguistic processing is also used in text analysis to detect grammar, semantic meaning, and elements of speech.

2. Phonetic Transcription

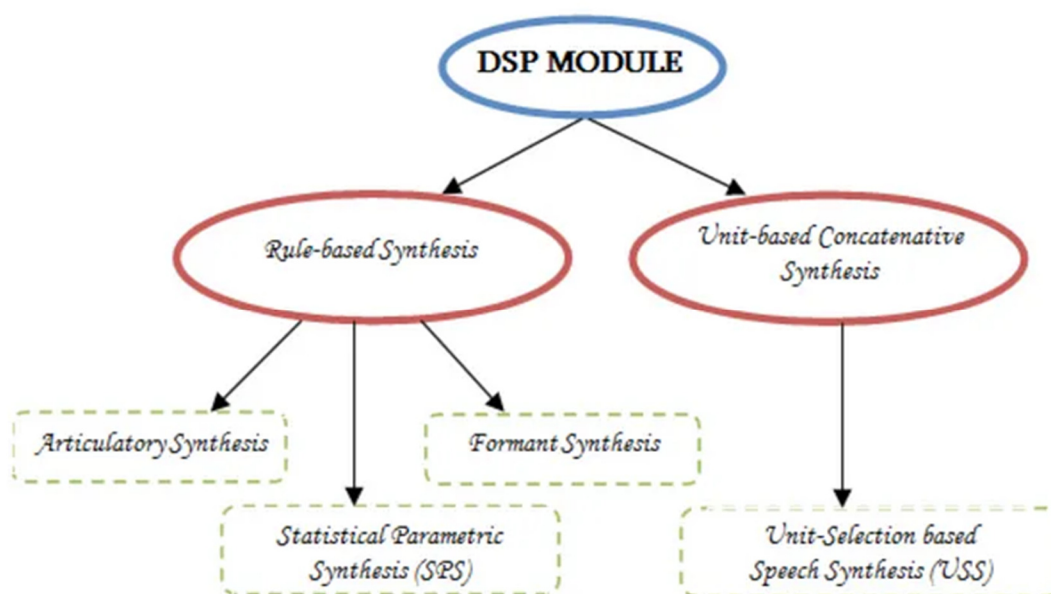
The text has been normalized and then transformed into phonetic representations that indicate the correct pronunciation of each word. This entails mapping individual letters or groups of letters to the associated sounds (phonemes) through a process known as grapheme-to-phoneme conversion.

3. Prosody Generation

The rhythm, emphasis, and intonation of speech are referred to as prosody. Finding the right pitch, length, and strength for each phoneme is what this stage entails in order to produce a natural-sounding synthesized speech. The creation of prosody is essential for accentuating and expressing emotion in speech.

4. Speech Synthesis

To conclude, the speech waveform needs to be generated. Speech synthesis can be done in a number of ways, such as:



- **Concatenative Synthesis:** Speech is produced by concatenative synthesis, which creates whole utterances by assembling pre-recorded speech pieces from a large database. It frequently generates speech that sounds natural and of excellent quality, but it needs a massive amount of data.
- **Formant Synthesis:** Generates speech by using mathematical models to simulate the acoustic properties of the human vocal tract, producing sounds through the manipulation of formant frequencies. It is highly flexible and can produce a wide varieties of sounds but often sounds less natural.
- **Parametric Synthesis:** Generates speech by using statistical models to produce speech parameters, such as pitch and duration, which are then used to synthesize the speech waveform. It balances naturalness and flexibility. Speech synthesis systems (HTS) based on Hidden Markov Models (HMM) are classified as statistical parametric speech synthesis (SPS) techniques.
- **Neural Synthesis:** Modern systems generate speech by using deep learning models like WaveNet and Tacotron to directly produce speech waveforms from text, significantly enhancing naturalness and expressiveness.

Challenges in Speech Synthesis

Naturalness and Intelligibility: One of the biggest challenges in creating synthetic speech is making it seem natural and intelligible to humans. Even while brain synthesis techniques have advanced significantly, it is still challenging to produce speech that is identical to human speech in all situations.

Expressiveness: Emotional depth and expression in humans are difficult to capture and reproduce. Extensive prosody modeling and context-aware synthesis are necessary for expressive TTS.

Resource Intensity: Deploying neural synthesis, which produces high-quality speech on devices with minimal processing capacity is difficult because it demands a significant amount of computing resources for training and real-time synthesis.

Language and Dialect Coverage: It is difficult to create high-quality TTS systems for all languages and dialects because many languages lack comprehensive linguistic and phonetic data.

Handling Ambiguities and Context: It takes sophisticated natural language comprehension and disambiguation skills to accurately synthesize speech that accurately reflects the content and context of the text (e.g., homographs like "lead" vs. "lead").

Future Directions

- **Improved Neural Models**

Neural network topologies and training methodologies would substantially improve the expressiveness and naturalness of speech produced.

- **Multimodal Synthesis**

TTS systems may be made more accurate and expressive by including contextual and visual cues, such as lip movements and facial emotions.

- **Low-Resource Language Support**

Creating high-quality text-to-speech (TTS) systems for low-resource languages by utilizing data augmentation and transfer learning approaches.

- **Personalization**

Constructing TTS systems with the ability to gradually adjust to the preferences and speech patterns of individual users, so offering a more customized user experience.

- **Real-Time Processing**

Improving the effectiveness of TTS systems to allow for high-quality, real-time speech synthesis on edge devices with constrained computing resources.

Conclusion

Both speech recognition and synthesis technologies have undergone significant advancements in recent years, driven by innovations in machine learning, deep learning, and signal processing techniques. These advancements have led to improved accuracy, naturalness, and versatility in speech-based applications, contributing to their widespread adoption across various industries and domains.

Empowering Tamil Natural Language Processing and Machine Learning Research: Resources and Methodologies

Dhivya Chinnappa

JP Morgan Chase & Co, Dallas, Texas, USA

Abstract

This tutorial aims to provide a beginner-level session for students, researchers, practitioners, and enthusiasts, offering a comprehensive understanding of the resources, tools, and methodologies available for conducting research in Tamil Natural Language Processing (NLP) and Machine Learning (ML). Participants will learn about the existing resources for working with Tamil NLP and methodologies to build and evaluate NLP models for the Tamil language. The session will encompass a detailed walkthrough of available NLP resources and data sources, various machine-learning problems applicable to Tamil corpora, and an exploration of academic venues dedicated to Tamil AI research. Additionally, participants will engage in hands-on activities using a Tamil Large Language Model (LLM).

Resources:

Understanding the available resources is a critical first step in Tamil NLP/AI research. The tutorial will introduce participants to key repositories and catalogs, tools, datasets, and pre-trained models tailored for Tamil NLP. These resources are essential for tasks ranging from tokenization and morphological analysis to more complex applications like machine translation and sentiment analysis. Other resources that could be exploited for Tamil NLP include classical and modern Tamil literature, Newspapers, Wikipedia, Blogs, tweets and blogs, movie scripts, and Tamil lyrics.

Gaps

Tamil NLP models, including Large Language Models differ from English NLP models due to the scarcity of resources, lack of strong fundamental research, linguistic complexity, and cultural differences. By understanding these gaps, participants can better tailor their research efforts to the unique aspects of Tamil NLP.

Academic Venues and Collaboration

Advancing Tamil NLP and ML research requires active engagement with the academic community. This tutorial will highlight prominent conferences and workshops dedicated to Tamil AI research, including:

- ICON (International Conference on Natural Language Processing): A key venue for presenting cutting-edge NLP research.
- SPELLL (International Conference on Speech and Language Technologies for Low-Resource Languages): Focuses on linguistic research and engineering aspects.
- DravidianLangTech workshops: Dedicated to the computational processing of Dravidian languages, providing a collaborative platform for researchers.

These venues in addition to accepting individual research also conducts several shared tasks encouraging young ML enthusiasts to get into Tamil AI research. Participants will learn how to engage with these venues, submit papers, and collaborate with other researchers to further their work.

Hands-on experience with Tamil LLM

The session will culminate in a hands-on workshop where participants interact with a Tamil Language Model (LLM). They will learn about model training and fine-tuning on a dataset and see how the results change with respect to different hyper-parameters. Participants will also learn about common best practices when building a ML model

Conclusion

By the end of this tutorial, participants will have a solid foundation in the resources, tools, and methodologies essential for Tamil NLP and ML research. They will be equipped with practical skills to develop and evaluate NLP models tailored to the Tamil language, fostering further advancements and collaborations in the field.

A Tutorial on Linguistic Annotation Essentials for Tamil

Parameswari Krishnamurthy

International Institute of Information Technology, Hyderabad, India

Introduction

Linguistic annotation is a critical aspect of computational linguistics and natural language processing (NLP). It involves marking up a text with labels that provide information about linguistic phenomena such as morphology, syntax, semantics and pragmatics. For Tamil, a Dravidian language spoken predominantly in the Indian state of Tamil Nadu and in Sri Lanka, linguistic annotation requires attention to its unique syntactic, morphological, and phonological properties. This write-up provides an overview of the essentials of linguistic annotation for Tamil.

1. Morphological Annotation

Tamil is an agglutinative language, it forms words by stringing together morphemes (the smallest meaningful units of language) without altering them. Morphological segmentation involves identifying and labeling these morphemes within a word.

For example, in the word "பிள்ளைகளிடமிருந்துதானா" needs segmentation like பிள்ளை+கள் கள் + (இடம் + இருந்து)+தான் +ஏ. It has the following morphological information

1. பிள்ளை (pillai) - Root
2. கள் (kal) - Plural marker
3. இடம்+இருந்து (itamiruntu) - Case marker
4. தான் (tān) - Emphatic particle used for emphasis
5. ஆ (ā) - Interrogative particle indicating a question

Similarly, verbs are also rich in adding conjugation includes tense, aspect, mood, person, gender and number. Words in Tamil are not only formed through the concatenation of two or more morphological elements but also involve multilevel derivation. This morphological complexity demands morphological segmentation for further analysis.

2. Part-of-Speech Tagging

Part-of-speech (POS) tagging annotation in Tamil is a vital process that systematically assigns grammatical categories to each word in a sentence, thus aiding in linguistic analysis and computational processing of the language. Each word in a Tamil sentence is annotated with its part of speech (POS), such as noun, verb, adjective, adverb, etc. Tamil POS tagging also requires distinguishing between various types of verbs (main, auxiliary), nouns (proper, common), and some other categories. The POS tagging guidelines for Tamil are meticulously designed to capture the language's morphological, syntactic, and semantic intricacies, encompassing 16 distinct tags that classify words into various parts of speech such as nouns, verbs, adjectives, adverbs, and others. For instance, adjectives in Tamil are often identifiable by the suffix -ஆ and describe attributes like size or emotion, while adpositions typically follow the nouns they modify, providing relational context. The guidelines also account for

specific language features, such as tagging gerunds as nouns and ordinal numbers as adjectives. By applying these detailed tagging rules, POS tagging in Tamil enables precise word-role identification and facilitates the development of sophisticated language processing tools, reflecting the rich and complex grammatical structure of Tamil.

3. Syntactic Annotation

Syntactic annotation often involves dependency parsing, where the syntactic structure of a sentence is represented by a tree with words as nodes and syntactic relations as edges. For Tamil, this means identifying subject-verb-object (SVO) relationships, as well as other syntactic dependencies such as modifiers and auxiliaries. Our approach adheres to universal dependencies guidelines, incorporating Tamil-specific features into the process. We utilize the CONLLU format for annotation.

Here's an example of a sentence in CONLLU format in Tamil.

#Sent_id = 1

#text= நாம் சாப்பிடும் பழ வகைகளில் அனைத்து தேவையான சத்துக்களும் அடங்கியுள்ளன.

ID	Form	POS	Lemma	Morph	Relation	Head
1	நாம்	PRON	நாம்	Case=Nom Number=Sing Person=1	nsubj	2
2	சாப்பிடும்	VERB	சாப்பிடு	Polarity=Pos Tense=Fut VerbForm=Part	acl	4
3	பழ	NOUN	பழம்	Number=Sing	compound	4
4	வகைகளில்	NOUN	வகை	Case=Loc Number=Plur	obl	9
5	அனைத்து	DET	அனைத்து	—	det	7
6	தேவையான	ADJ	தேவையான	—	amod	7
7-8	சத்துக்களும்	-	-	-	-	-
7	சத்துக்கள்	NOUN	சத்து	Case=Nom Number=Plur	nsubj	9
8	உம்	PART	உம்	—	mark	7
9-10	அடங்கியுள்ளன	-	-	-	-	-
9	அடங்கி	VERB	அடங்கு	Polarity=Pos VerbForm=Conv	root	0
10	உள்ளன	AUX	உள்	Gender=Neut Number=Plur Person=3 Tense=Pres VerbForm=Fin	aux	9
11	.	PUNCT	.	—	punct	9

4. Semantic Annotation

Semantic annotation involves marking up text with information about its meaning, beyond just its syntactic structure. It aims to capture the intended semantics or the underlying concepts expressed in the text. This can include identifying named entities such as people, organizations, locations, and dates, as well as categorizing words or phrases into semantic classes such as events, actions, or entities. Semantic annotation can also involve linking words or phrases to external knowledge bases or ontologies to provide additional context and disambiguation. Overall, semantic annotation helps in understanding the deeper meaning and context of textual

data, which is essential for various natural language processing tasks such as information extraction, question answering, and sentiment analysis.

4.1 Named Entity Recognition (NER)

Semantic annotation includes tasks like Named Entity Recognition, where proper nouns are categorized into predefined categories such as persons, locations, organizations, dates, etc. For example, "சென்னை" (Cennai) would be tagged as a location.

4.2 Semantic Role Labeling (SRL)

Semantic role labeling involves identifying the roles that different words play in a sentence. For instance, in the sentence "ரவி புத்தகம் வாசிக்கிறான்" (Ravi puttakam vācikkirān), "ரவி" (Ravi) is the agent (doer of the action), "புத்தகம்" (puttakam) is the theme (object being acted upon), and "வாசிக்கிறான்" (vācikkirān) is the action (verb).

5. Pragmatic and Discourse Annotation

Pragmatic and discourse annotation involves marking up text with information about how language is used in context to convey meaning and facilitate communication. Pragmatics focuses on the study of language in use, including aspects such as implicature, presupposition, and speech acts, which go beyond the literal meaning of words and sentences. Discourse annotation, on the other hand, deals with the structure and organization of connected stretches of language, such as conversations, narratives, or arguments. It includes identifying discourse relations between sentences or utterances, annotating rhetorical structures, and marking up phenomena like anaphora, coherence, and cohesion. Both pragmatic and discourse annotation are essential for understanding the nuanced meaning and communicative intent embedded in natural language text, and they play a crucial role in various natural language processing tasks such as dialogue systems, text summarization, and sentiment analysis.

5.1 Discourse Markers

Pragmatic annotation deals with elements beyond the sentence level, such as discourse markers and connectors. Tamil uses a range of particles and connectors, such as "ஆனால்" (āṇāl - but) and "எனவே" (eṇavē - therefore), which need to be annotated to understand the flow of discourse.

5.2 Coreference Resolution

This involves identifying when different expressions in a text refer to the same entity. For example, in the sentences "ரவி பள்ளிக்கூடத்திற்கு செல்கிறான். அவன் புத்தகம் எடுத்துக்கொண்டு செல்கிறான்." (Ravi paḷḷikkūṭattirkku celkirān. Avaṇ puttakam eṭuttukkoṇṭu celkirān.), "அவன்" (avaṇ) refers to "ரவி" (Ravi).

Conclusion

Linguistic annotation for Tamil is a multifaceted process that requires careful consideration of its unique linguistic features. Proper orthographic and phonological annotation is foundational, while morphological and syntactic annotations help in understanding the structure of the language. Semantic and pragmatic annotations further enrich the understanding by providing insights into meaning and use in context. As computational tools and resources for Tamil

continue to develop, detailed and accurate linguistic annotation will be essential for advancing research and applications in NLP for this rich and ancient language.

In this tutorial, we've delved into the intricate world of linguistic annotation for Tamil, a language rich in history and complexity. From the foundational aspects of orthographic and phonological annotation to the more nuanced morphological, syntactic, semantic, and pragmatic annotations, we've explored the multifaceted nature of processing Tamil text computationally. Understanding Tamil in the area of natural language processing requires meticulous attention to its unique linguistic features, and thorough annotation is crucial for advancing research and applications in NLP for this ancient and culturally significant language. As computational tools and resources continue to evolve, this tutorial serves as a guide for navigating the complexities of Tamil in the digital landscape, facilitating deeper insights into its structure, meaning, and usage in context.

Machine Translation using Transformers (NMT)

Vijay Sundar Ram and Patabhi RK Rao

AU-KBC Research Centre,

MIT Campus of Anna University, Chennai, India

sundar@au-kbc.org, patabhi@au-kbc.org

1. Introduction

Machine Translation (MT) is one of the most dealt fields of Natural Language Processing (NLP), since World War II, yet we have not achieved near human translation. And specifically in the Indian language scenario there is lot more to be done. MT is the automated process of decoding the meaning of source text and recording the meaning in target language without loss of information. Automated machine translation system understands the source sentence and automatically generates sentence in the target language. This achieved using many different techniques.

2. History of Translation

First Machine Translation in Tamil: Tamil to Russian

The fig 1 below, pictorially presents MT evolution on timeline.

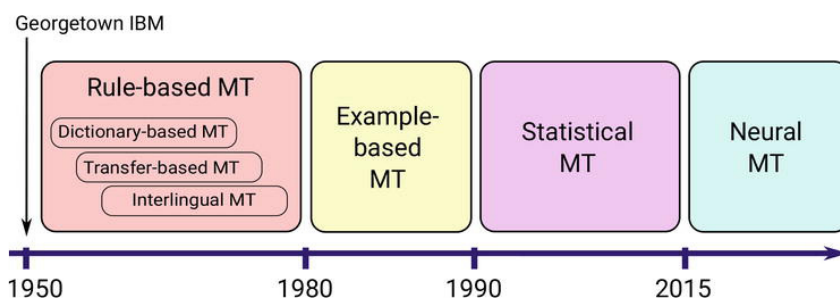


Fig1: MT Evolution - Timeline chart

3. Different Types of Translation Techniques

The early translation engines started with word-to-word translation and evolved to the present Neural machine translation. The various techniques across years are listed as follows.

- Dictionary-based Machine translation
- Rule-Based Machine Translation systems
- Transfer Based Machine translation
- Interlingual Machine translation
- Example-based Machine translation
- Analysis-Transfer-Generation
- Statistical Machine Translation systems
- Neural Machine translation

4. Neural Machine Translation (NMT)

In this section, we start with the basics of Artificial Neural Networks (ANN), followed by the incremental improvements in Neural machine translation techniques from its initial stages in 2015 to present promising state-of-art techniques. We have also discussed the challenges in developing NMT for low-resource language and morphologically rich languages. And the methods to mitigate these challenges.

4.1 Basics of Artificial Neural Networks

Artificial Neural Networks (ANN) is algorithms based on brain function and are used to model complicated patterns and forecast issues. The Artificial Neural Network (ANN) is a deep learning method that arose from the concept of the human brain Biological Neural Networks. They are among the most powerful machine learning algorithms used today. The development of ANN was the result of an attempt to replicate the workings of the human brain.

Basically, there are three layers in the network architecture: the input layer, the hidden layer (more than one), and the output layer. Neural Network is a set of connected, INPUT/OUTPUT UNITS, where each connection has a WEIGHT associated with it.

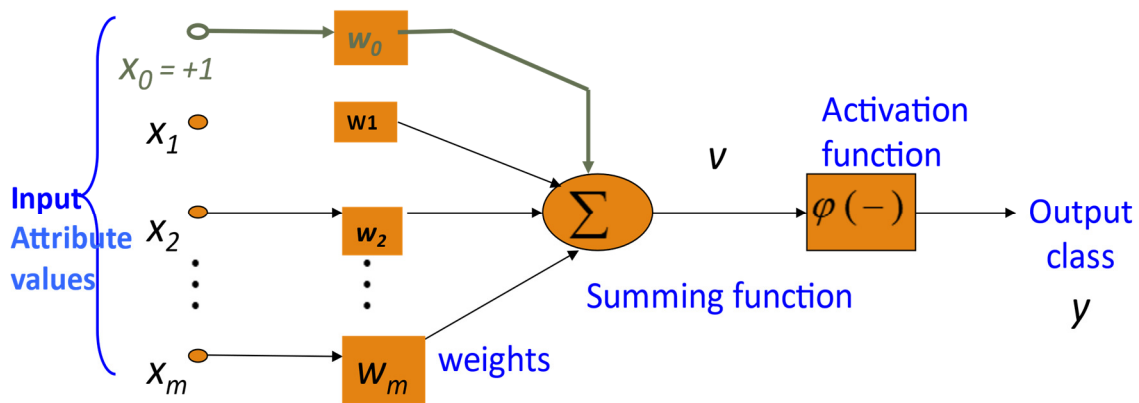


Fig2: Artificial Neural Network

4.2 Journey of NMT

Neural Machine Translation (NMT) started with the successful works by Kalchbrenner and Blunsom [8], Sutskever et al., [14] and Cho et al., [3], where an encoder encoded the source sentence to a fixed-length vector, from which a decoder generated the translation. Sutskever et al., [14] reported a NMT system built based on RNN with long short term memory (LSTM), this surpassed the performance of the previous start-of-art performance. These seq2seq models work well for short sentences, but do not perform well for long sentences due to the vanishing gradient problem. Bahdanau et al., [1] presented an extended encoder-decoder to handle the problem of encoding of source sentence into a fixed-length vector. They used a bidirectional recurrent neural network (RNN) consisting of forward and backward RNN to focus around the word. Attention mechanism was introduced in the decoder to decide the part of the source sentence to pay attention. Luong et al., [12] simplified the attention mechanism by considering the hidden states at the top layer of both encoder and decoder. This attention mechanism attends to the entire input sequence.

With improvements in attention mechanism Vaswani et al., [15] introduced a new architecture called transformer with encoder and decoder that relies solely on attention mechanism. The Transformer model relies on self-attention where all input sequence members are compared with each other, and modifies the corresponding output sequence position. Though these NMT systems (Bahdanau et al., [1], Vaswani et al., [15]) has led to a greater improvement in translation of high resource languages, the translation of morphologically rich and low resource languages is a major challenge.

4.3 Challenges in Low Resource Languages and Ways to Mitigate

Low resource of parallel data is a bottleneck in many language pairs. Different approaches were executed to overcome or reduce the problem. We briefly discuss these techniques in the following section.

4.3.1 Increasing the data using Back Translation

Senrich et al., [13] introduced back translation technique, where the monolingual data of the target language is translated to source language using available MT system and combined with the training data. This helps in improving the translation quality.

4.3.2 Phrase Table Injection

Zhao et al., [23] presented a method to combine the SMT and NMT by utilizing the phrase table generated in the SMT training. It is combined with the data in training the NMT.

4.3.3 Leveraging the pre-trained models

Pre-trained models such as BERT, Glove, RoBERTa are commonly used in NMT to improve the quality of translation. These models are used in fine-tuning the NMT training.

4.3.4 Combining the Corpus

When similar languages are on the target side, in this technique, the knowledge is exploited to translate the mixed language better. Banerjee A et al., [1] has presented a technique, where English-Hindi and English-Marathi corpus are combined to train the NMT and English-Marathi corpora is used to fine-tune the NMT training.

4.3.5 Transfer Learning

Transfer learning is the process of applying an existing training Machine learning model to a new, but related problem. In pivot-based transfer learning, first they pre-train a source-pivot model with a source-pivot parallel corpus and a pivot-target model with a pivot-target parallel corpus. Then initialize the source-target model with the source encoder from the pre-trained source-pivot model and the target decoder from the pre-trained pivot-target model. Now the training with a source-target parallel corpus is continued. Kim et al., [9] has proposed three methods to increase the relation among source, pivot, and target languages in the pre-training: 1) step-wise training of a single model for different language pairs, 2) additional adapter component to smoothly connect pre-trained encoder and decoder, and 3) cross-lingual encoder training namely autoencoding of the pivot language.

Domain term translation in most of languages is a challenging task and in Indian languages it is more challenging due to very less availability of parallel domain terms. Hema Ala et al., [7] has proposed to handle domain terms in NMT using the back translation technique, where Domain specific back translation using monolingual and generates synthetic data. They have conducted experiments on Chemistry and Artificial Intelligence domains for Hindi and Telugu in both directions.

Unsupervised NMT (UNMT) is one of the up-coming techniques to overcome the low-resource problem. It is shown that UNMT works for source and target languages are similar and in same domain. Sai Koneru et al., [11] had presented an experiment on UNMT for Dravidian languages (Kannada, Tamil, Telugu and Malayalam) to English.

4.4 Challenges in Morphological Rich Language

Translation of morphologically rich languages using NMT has the following challenges, a) large number of inflected forms lead to a larger vocabulary and thus causes data sparsity. b) Generating sentences with correct linguistic agreement and expressing exact semantics of the input sentence is a challenge.

These challenges are handled using the following techniques;

- a) Breaking the word forms into sub-word units, so that the overall vocabulary size is reduced.
- b) Training with linguistic features such as lemma-tag strategy.

Sub-word units are generated using

- a) Statistical methods such as Sub-word Text Encoder (STE) and Byte Pair Encoding (BPE)
- b) linguistically motivated word segmentation methods using tools such as Morfessor and Morphological analysers.

Dominik Machacek et al., [5] compared the linguistically motivated method morfessor and derivational dictionaries-based method and statistical methods such as STE and BPE in German to Czech translation. Their experiments showed the non-linguistically motivated method performed better. Goyal et al., [6] has presented Hindi to English NMT, where they generalised the embedding layer of the Transformer model to incorporate linguistic features such as PoS, lemma, and morphological features. There was a significant increase in the BLEU scores. Dewangan et al., [4] has presented an elaborate NMT experiments to understand the poor performance of the Dravidian languages compared to Indo-Aryan languages. They used Byte Pair Encoding (BPE) method to understand the BPE in Indian languages. From their study, they presented that the optimal value for BPE merge for Indian languages is between 0-5000, which is low compared to that observed for European languages.

5 Recent Machine Translation Techniques: MT using LLMs

With the introduction of multi-lingual large language models (LLMs) such as GPT-3 and ChatGPT, new techniques have emerged, machine translation using LLMs. Here the LLMs are fine-tuned with prompts for learning translation.

Evaluation of the translation is another important aspect of MT ecosystem. In the following section, we have presented the details of different types of evaluation metrics.

6 Automatic Evaluation Metrics

Automatic quality metrics are divided into **string-based metrics** and **machine learning-based metrics**.

6.1 String-based metrics

String-based metrics generally measure the word or character distance between the target sentence and the reference translation.

Examples:

- BLEU
- METEOR
- NIST
- chrF
- TER

String-based are used in research papers and competitions because they are explainable and fair, and they can support any language pair.

But string-based metrics can punish translations that convey the correct meaning, and the scores cannot be compared across language pairs.

The scores generally do not correlate well with human evaluation scores when translation quality is high.

6.2 Machine learning-based metrics

Machine learning-based metrics use sentence embeddings to calculate the difference between the generated target sentence and the reference translation, or even between the target sentence and the source sentence.

Examples:

- COMET
- YiSi
- BERTscore

Machine learning-based metrics require a model that was trained on data with the source and target languages.

The score can correlate well with human evaluation scores.

But the scores are not explainable or fair, so they cannot be used in a research competition.

6.3 Human evaluation metrics

Human evaluation is the gold standard.

- MQM
- SQM
- Average score and average z-score
- TrueSkill
- Adequacy and fluency judgement
- Relative ranking
- Constituent ranking
- Yes or no constituent judgement
- Direct assessment

But human evaluation is slow, expensive, and subjective.

7 Summary

NMTs offers many key benefits that make them particularly well-suited:

- NMTs can learn and model non-linear and complicated interactions, which is critical since many of the relationships between inputs and outputs in natural language are non-linear and complex.

- Neural Networks can generalize – After learning from the original inputs and their associations, the model may infer unknown relationships from anonymous data, allowing it to generalize and predict unknown data.

References

- [1] Bahdanau D., Cho K., and Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
- [2] Banerjee, A., Jain A., Mhaskar S., Deoghare S, D. Sehgal A., and Bhattacharya, P. (2021). Neural Machine Translation in Low-Resource Setting: a Case Study in English-Marathi Pair. In Proceedings of the 18th Biennial Machine Translation Summit - Volume 1: Research Track, MTSummit 2021 Virtual, pp 35-47
- [3] Cho, K., van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014).
- [4] Dewangan, S., Alva, S., Joshi, N., Bhattacharyya, P. (2021). Experience of neural machine translation between Indian languages. *Machine Translation* 35, 71–99
- [5] Dominik Macháček, Jonáš Vidra, Ondřej Bojar (2018): Morphological and Language-Agnostic Word Segmentation for NMT. In: Proceedings of the 21st International Conference on Text, Speech and Dialogue—TSD 2018, pp. 277-284, Springer-Verlag, Cham, Switzerland, ISBN 978-3-030-00794-2
- [6] Goyal, Vikrant and Kumar, Sourav and Sharma, Dipti Misra. (2020). Efficient Neural Machine Translation for Low-Resource Languages via Exploiting Related Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp 162-168
- [7] Hema Ala, Vandan Mujadia, Dipti Misra Sharma. (2021). Domain Adaptation for Hindi-Telugu Machine Translation Using Domain Specific Back Translation. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp 26-34
- [8] Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1700–1709. Association for Computational Linguistics.
- [9] Kim, Y., Petrov, P., Petrushkov, P., Khadivi, S., and Ney, H. (2019). Pivot-based transfer learning for neural machine translation between non-English languages. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 866–876, Hong Kong, China. Association for Computational Linguistics
- [10] Klein G., Hernandez F., Nguyen V., and Senellart J. (2020) The opennmt neural machine translation toolkit: 2020 edition. In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020), pages 102–109.
- [11] Koneġ, Sai; Liu, Danni; Niehues, Jan. (2021). Unsupervised Machine Translation On Dravidian Languages, In 16th conference of the European Chapter of the Association for Computational Linguistics (EACL), Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages.

- [12] Luong M., Pham H., and Manning D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.
- [13] Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- [14] Sutskever, I., Vinyals, O., and Le, Q. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS 2014)*
- [15] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, U.; Polosukhin, I. (2017) Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 4–9
- [16] Zhao, Y., Y. Wang, J. Zhang, and C. Zong (2018). Phrase table as recommendation memory for neural machine translation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pp. 4609–4615.

Utilizing Tech Tools For “Seal of Biliteracy” Tamil Certification Exam Topics

**Soundar Jayabal,
Avvai Tamil Center, Dallas, Texas**

- What is the Seal of Biliteracy?
- What is the Global Seal of Biliteracy?
- What is the Difference Between Them?
- Benefits of Seal of Biliteracy Certifications
- Decoding Exam Structures
- Understanding the Syllabus for Exam Preparation
- Navigating the Right Study Materials
- Utilizing Technology for Tamil Language Learning
- Using Tech Tools to Enhance LSRW Skills
- Technical Specifications for Online Test Participation
- Utilizing Practice Tests and Online Resources
- Effective Approaches to Exam Success

Why “Seal of Biliteracy” Certifications for Tamil Language - A background information.

அமெரிக்காவில் சமீபத்தில் நடந்த மக்கள் தொகை கணக்கெடுப்புப்படி கிட்டத்தட்ட 300,000 தமிழ் மக்கள் அமெரிக்கா முழுவதும் வசிக்கிறார்கள். நியூ ஜெர்சி, நியூ யார்க், கலிபோர்னியா, டெக்சாஸ், ஜார்ஜியா, இலியனாய்சு, ஃப்ளோரிடா, வாஷிங்டன் ஆகிய பகுதிகளில் அதிக அளவில் வசிக்கிறார்கள்.

அமெரிக்காவில் தமிழ்க்கல்வி வளர்ச்சியை நாம் மூன்று பகுதிகளாகப் பிரிக்கலாம்.

1. 1995 ஆம் ஆண்டுக்கு முன்னர்.
2. 1995 ஆம் ஆண்டில் இருந்து 2010 வரை.
3. 2010 ஆம் ஆண்டிற்குப் பிறகு.

1995 ஆம் ஆண்டுக்கு முன்பு வந்தவர்கள் பெரும்பாலும் மருத்துவர்கள், பொறியாளர்கள், அறிவியல் ஆராய்ச்சியாளர்களாகச் சிறு எண்ணிக்கையில் இருந்தனர். அவர்கள் அனைவரும் ஆண்டுக்கு சில முறை பண்டிகை நாட்களில் சந்தித்துக் கொண்டனர். தமிழ்ச் சங்கங்கள் பெரிய அளவில் இப்போது நடப்பது போல் இயங்கவில்லை. அவர்களுடைய குழந்தைகளின் தமிழ்க்கல்வி பெரும்பாலும் வீட்டில் பெற்றோர்கள் கற்றுக் கொடுத்த சில அரிச்சுவடி பாடங்களுடன் நின்றுவிட்டது. முறையான கல்வி நிறுவனங்களோ, கட்டமைப்புகளோ, தமிழ் கற்பதற்கான கற்பிப்பதற்கான வசதிகளோ, பாடத்திட்டங்களோ இல்லை.

1995 ஆம் ஆண்டிற்குப் பிறகு கணிப்பொறி தொழில்நுட்ப வல்லுநர்களின் வருகை அதிகரிக்கத் தொடங்கியது. பெரும் எண்ணிக்கையிலான அவர்களுடைய குழந்தைகளுக்குத் தமிழ் கற்பிக்க வேண்டிய அவசியமும் தேவையும் ஏற்பட்டது. 1999 ஆம் ஆண்டு கலிபோர்னியா தமிழ்க்கல்விக்கழகம் தொடங்கப்பட்டது. வேறு சில நகரங்களிலும் தமிழ் கற்றுக் கொடுக்க தமிழ்ப்பள்ளிகள் தொடங்கப்பட்டது. அப்பள்ளிகள் பெரும்பாலும் கோயில்களிலும் தேவாலயங்களிலும் நூலகங்களிலும்

அல்லது ஒரு சிலரின் வீடுகளிலும் நடத்தப்பட்டது. சிறிய அளவில் தன்னார்வ ஆசிரியர்களுக்குத் தெரிந்த அளவிலான பாடத்திட்டங்களைக் கொண்டு அல்லது தமிழ்நாட்டில் இருந்து கொண்டுவரப்பட்ட பாடப் புத்தகங்களில் இருந்து வகுப்புகள் நடத்தப்பட்டன.

சிகாகோ நகரில் 1984ல் தொடங்கப்பட்ட தமிழ்ப் பள்ளிகள் முறையாக ஒருங்கிணைக்கப்பட்டு 2004ஆம் ஆண்டு "அமெரிக்கத் தமிழ்ப் பள்ளிகள்" என்ற அமைப்பாக முறைப்படி இலியனாய்சு மாநிலத்தில் பதிவு செய்யப்பட்டு இன்றளவும் தொடர்ந்து நடத்தப்பட்டு வருகிறது. 2006 ஆம் ஆண்டு தமிழ்ப் பள்ளிகளை இணையத்தில் இணைக்கும் tamilschools.net என்ற முயற்சி எடுக்கப்பட்டது. ஆனால் அந்த முயற்சி வெற்றி பெறவில்லை. இதற்கிடையே 2005ஆம் ஆண்டிலிருந்து பேரவை நடத்தும் தமிழ் விழாக்களில் தமிழ் ஆசிரியர்கள், தமிழ்க் கல்வி ஆர்வலர்கள் தொடர்ந்து சந்தித்து தங்களுக்குள் தகவல்களைப் பரிமாறிக் கொண்டு வந்தனர்.

இப்படி 2010ஆம் ஆண்டு வரை பல்வேறு வழிகளிலும் தமிழ் கற்றுக் கொடுக்க முயற்சி செய்யப்பட்டுத் தொடர்ந்து நடத்தப்பட்டு வந்தது. அப்போது புலம்பெயர்ந்த தமிழ் குழந்தைகளுக்குத் தமிழ் கற்பித்தலில் உள்ள பிரச்சினைகளை உணர முடிந்தது. அதில் முதன்மையானது இங்குள்ள குழந்தைகளுக்கு முதன்மை மொழி ஆங்கிலம் என்பதும் தமிழ் ஒரு இரண்டாம் மொழியாக வெளிநாட்டு மொழியாக மட்டுமே இருக்கிறது என்ற உண்மையும், உணர்வும் ஆகும். ஆங்கில வழியாகத் தமிழ் கற்பதால் குழந்தைகளும் தமிழை ஆங்கிலத்துடன் மொழி ஒப்பீடுகள் செய்தனர். அப்படி செய்ததில் தமிழ் அதிக எழுத்துக்கள் கொண்ட கற்பதற்குக் கடினமான மொழியாகவும், சிக்கலான பயன்பாட்டில் இல்லாததாக இலக்கணம் இருப்பதையும் உணர்ந்தனர். அதோடு எழுத்து மொழிக்கும், பேச்சு மொழிக்கும் அதிக அளவிலான வேறுபாடுகள் இருப்பதை அறிந்தனர். தமிழ் கற்பிக்கும் / கற்கும் முறை வழக்கமாக அவர்கள் பள்ளிகளில் ஆங்கிலம் கற்றுக் கொள்ளும் முறையில் இருந்து முற்றிலும் மாறுபட்டு வேறு வகையாக இருந்ததையும் உணர்ந்தனர். இப்படி பல்வேறு சிக்கல்களுக்கு இடையே பல ஆண்டுகளாக மாணவர்கள் தமிழ் மொழியைக் கற்றுக் கொண்டு வந்தனர். அதோடு பழைய முறையிலான பாடத்திட்டங்கள், பாடப்புத்தகங்களைப் பயன்படுத்த வேண்டியிருந்தது. தமிழை முழுமையாக 8 ஆண்டுகள் கற்றுக் கொண்ட பின்னரும் அவர்களுக்கு எந்த நேரடிப் பயனும் இல்லாமல் இருந்தது. கல்லூரிகளிலோ வேலைவாய்ப்புகளிலோ சிறப்புக் கவனமோ அல்லது சலுகையோ இல்லாமல் இருந்தது.

இந்த பிரச்சனைகளை எல்லாம் சரி செய்ய பல்வேறு முயற்சிகள் எடுக்கப்பட்டது. அதன் ஒரு பகுதியாக 2010ஆம் ஆண்டில் அமெரிக்கத் தமிழ்க் கல்விக் கழகம் தொடங்கப்பட்டது. அமெரிக்கச் சூழ்நிலைகளுக்கு ஏற்றவாறு ஒருங்கிணைந்த பாடத் திட்டங்களும் பாடங்களும் முழுக்க முழுக்க தன்னார்வலர்களைக் கொண்டு உருவாக்கப்பட்டது. ஆசிரியர் பயிற்சிகளும், தேர்வு நடத்துவதற்கும் உதவிகள் செய்யப்பட்டது. அமெரிக்கா முழுவதும் பெரிய நகரங்களிலும் இரண்டாம் கட்ட நகரங்களிலும் முறைப்படி பல்வேறு தன்னார்வ தமிழ்ப்பள்ளிகள் தொடங்கப்பட்டன. கலிபோர்னியா தமிழ்க் கல்விக் கழகமும் உலகத் தமிழ்க்கல்வி கழகமாகப் பெயர் மாற்றப்பட்டு அதன் கீழும் பல்வேறு தமிழ்ப்பள்ளிகள் இணைந்தன. 2018 ஆம் ஆண்டில் அமெரிக்கத் தமிழ் ஆசிரியர்களை ஒருங்கிணைக்கும் "அமெரிக்கத் தமிழாசிரியர் கழகம்" தொடங்கப்பட்டது. 2021ஆம் ஆண்டு ஒரு புதிய வழியில் தமிழ்

கற்றுக்கொடுக்கும் முயற்சியோடு “அமெரிக்கத் தமிழ்மொழிக் கல்வி நிறுவனம்” தொடங்கப்பட்டது.

இப்படியான பல்வேறு தொடர் செயல்பாடுகளால் இப்போது கிட்டத்தட்ட 300க்கும் மேற்பட்ட தன்னார்வத் தமிழ்ப் பள்ளிகள் உள்ளன. 30,000க்கும் மேற்பட்ட மாணாக்கர்கள் ஒருங்கிணைந்த பாடத் திட்டத்தின் கீழ் தமிழ்க்கல்வி பயில்கின்றனர். இருந்தாலும் ஒரு சில தமிழ்ப்பள்ளிகளே உள்ளூர் மாவட்ட கல்வி நிலையங்களால் அங்கீகரிக்கப்பட்டு தமிழ் படிக்கும் மாணாக்கர்களுக்கு அவர்கள் படிக்கும் உயர்நிலைப் பள்ளிகள் மூலம் இரண்டாம் மொழிக்கான மதிப்பீட்டுப் புள்ளிகள் (language credits) வழங்குகிறது. அப்படி உள்ளூர் கல்வி மாவட்ட நிலையங்களால் அங்கீகரிக்கப்படாத பள்ளிகளில் படிக்கும் மாணவர்கள் மாநில அரசின் ‘இருமொழி முத்திரை’ சான்றிதழ், மொழித்திறன்களை சோதிக்கும் தனியார் அமைப்புகளான AAPPL, ALTA, Avant ஆகியவை நடத்தும் தேர்வுகள் மூலம் மதிப்பீட்டுப் புள்ளிகள், “Global Seal of Biliteracy” அமைப்பு வழங்கும் சான்றிதழ்கள் பெறலாம். அதற்கான தேர்வுகள் அனைத்தும் ACTFL வகுத்துள்ள மொழி கற்பித்தலுக்கான தரப்பாடுகளின் (standards) படியே நடத்தப்படுகிறது. அத்தரப்பாடுகள் கேட்டல், பேசுதல், படித்தல், எழுதுதல் என்ற நான்கு திறன்களை அடிப்படையாகக் கொண்டது.

இப்படிப்பட்ட சான்றிதழ்கள் உயர்நிலைப் பள்ளி மாணவர்களுக்கான மதிப்பீட்டுப் புள்ளிகளைப் பெறுவதோடு அல்லாமல் கல்லூரிச் சேர்க்கைகளுக்கும் ஒரு சில வேலை வாய்ப்புகளுக்கும் உதவக்கூடும். ACTFL நிறுவனத்தின் மொழிகளுக்கான தரப்பாடு அமெரிக்காவில் மட்டுமல்லாது உலகின் பல்வேறு நாடுகளில் அங்கீகரிக்கப்பட்டுப் பயன்படுத்தப்பட்டு வருகிறது. இதன்மூலம் உலக அளவிலான ஒரு அங்கீகாரமும் தமிழ் மொழி கற்றலுக்குக் கிடைக்கும்.

சௌந்தர் ஜெயபால்

இணை நிறுவனர், அவ்வை தமிழ் மையம் (ATC)

இணை நிறுவனர், அமெரிக்கத் தமிழ்க் கல்விக்கழகம் (ATA)

இணை நிறுவனர், அமெரிக்கத் தமிழ் ஆசிரியர் கழகம் (ATTA)

Generative AI and LLM for Tamil

Muthiah Annamalai

Founder PanMo Cloud, San Francisco, CA

<ezhillang@gmail.com>

ABSTRACT

We discuss applications of the LLM for Tamil NLP with demonstration and discuss some of the guardrails of LLMs in Tamil.

1. INTRODUCTION

Generative AI and Large Language Models (LLMs) have provided novel automation for tasks previously unable to be handled at a acceptable quality. Particularly introduction of ChatGPT (Nov, 2022) has gained a wide acceptance among the general public outside of practitioners of AI; subsequent rise of open-source LLMs like Llama, Llama2, Llama3 and Mistral and their fine-tuned variants like Tamil-Llama have provided a lot of horse-power for capable engineering teams to integrate these models and capabilities into their applications. Questions remain on how effective these models can be in tackling NLP tasks, their accuracy and bias and hardware limitations to deployment.

2. Application in NLP

2.1 Spelling Correction

By using a masked word in a sentence, we can make the LLM complete this masked word and compare against user input for spelling correction.

e.g. if the user input is “அன்பே சுவம் என்று சொல்வார் சிலர்” then the masked word technique can be applied for the suspicious word and sent to LLM as input, “அன்பே [MASK] என்று சொல்வார் சிலர்,” and the LLM would complete is as “அன்பே சுவம் என்று சொல்வார் சிலர்” thereby performing a spelling correction.

Since LLMs prediction is of order of tokens/second we will have to run the masked word on a pre-processed sentence where only a few words are tagged by a prediction model for potential error.

3. Quality and Bias in LLMs

LLMs are not oracles even if they appear so; popular works like “stochastic parrots” etc. indicate the LLMs are mainly auto-regression tools which are good at learning the distribution of the training data and generating tokens in that distribution; they are not sentient even if they may give some persuasive outputs to that effect.

Particularly LLMs suffer from limitations in arithmetic capability; for example, Tamil-LLAMA paper reports a sentence completion of factorial function (தொடர்பெருக்கு) incorrectly; whereas such limitations are found over by the function-calling LLAMA or Tool-LLAMA variants the fundamental nature of the LLMs remain.

Further LLMs need to be qualified for various levels of bias and metrics established - do they represent Tamil, people, ethnography, culture in an unbiased manner? What are their biases

regarding one or more of these categories? To our knowledge this is not done to any quantifiable way.

4. Conclusions

Tamil LLAMA model despite its limitations presents a significant advancement in generative AI open-source language model; we hope to see more community participation from Tamil department academics, engineers and humanities to quantify the bias and guardrails in the data set. Future for AI applications remains bright but limited by access to training, fine-tuning and running such models.

A Transformer-Based Sandhi Splitter for Tamil

Parameswari Krishnamuthy, Nagaraju Vuppala and Nisha Irene

<param.krishna@iiit.ac.in>, <nagaraju.vuppala@research.iiit.ac.in>, nishairene@gmail.com

International Institute of Information Technology - Hyderabad

Abstract

Sandhi splitting is a crucial step in natural language processing (NLP), especially for languages with agglutinative morphology. Morphological analysis involves identifying the morphemes constituting a word, and Sandhi splitting is integral to this process. This study focuses on identifying instances of breaking words with external sandhi in the context of composite non-compound words. Tamil, an agglutinative language, constructs words by attaching various morphemes (meaningful word units) as suffixes to a root or base word. These morphemes convey grammatical information such as gender, number, case markers, tense, aspect, mood, etc. Nouns and verbs in Tamil are highly inflected, with extensive conjugation patterns. Sandhi, derived from Sanskrit, denotes the process wherein two or more morphemes or word forms unite to form a complex word, facilitating smooth word transitions during word formation. We approach Sandhi splitting as a Seq2Seq task for machine learning, utilizing different transformer models. Our experiments aim to evaluate whether transformer models are good at sandhi splitting. We train our models using data containing sentences with sandhi splitting annotations. The training dataset encompasses both splitting and non-splitting words to enable the model to discern when to split words. Through this study, we seek to enhance sandhi splitting accuracy in Tamil text processing, ultimately contributing to the advancement of NLP applications in morphologically complex languages.

1 Introduction

Sandhi, in linguistics, is a process in which two or more morphemes or word forms unite to form a complex word. It involves the alteration of sounds at word boundaries when two words are combined to form a new word. Sandhi, as derived from Sanskrit, means ‘join together’. It refers to the natural phonetic transformations that occur when two or more words are juxtaposed.

These transformations are often guided by a particular language’s phonological rules. These phenomena can include merging phonemes (units of sound), omitting phonemes, and adding phonemes to facilitate smooth word transitions. Sandhi can be understood in two ways. Internal sandhi refers to different types of sound changes that occur within words (internally) in the word-formation processes, such as inflection and derivation, whereas external sandhi refers to sound changes that happen between two or more fully-formed word boundaries (externally). In our study compound words are considered as single tokens and can not be split because they derive new lexical sense on combining. The composite words which are non-compound words in External sandhi do not derive new meanings on combining, these word forms are considered

as two different tokens and we consider these to be the split points in our model. The detailed explanation of types of sandhi is explained in section 5.

This explicit study on splitting composite words is essential as it necessitates a deep understanding of morphological distinctions, syntactic variations, and semantic clarity. Composite words are usually formed with multiple word forms with different roots including suffixes and their grammatical features. Along with morphological distinctions they exhibit various syntactic relationships that affect overall sentence structure and meaning. They also may carry multiple lexical senses, and splitting them into constituent parts clarifies their individual contributions to the overall context. These linguistic aspects are interconnected and essential for understanding the complexity of language.

2 Sandhi in Tamil

Tamil is an agglutinative language as it forms words by attaching various morphemes (meaningful word units) as suffixes to a root or base word. These morphemes can convey grammatical information such as gender, number, person, case markers, tense, aspect, mood, and more. Tamil nouns can be highly inflected to indicate case and number. Tamil verbs are also richly inflected, with extensive conjugation patterns for tense, aspect, mood, gender, person, and number. In Tamil, there are contexts where two/more grammatically formed word forms can be joined together. We refer to them as composite words or non-compounds as they must be split before processing them. These split words are considered tokens.

The rest of the paper is organized as follows: Section 3 discusses the review and comparison of existing approaches. Sections 4 and 5 explain the sandhi splitting and types of sandhi. Section 7 explains the rules for sandhi splitting. Section 8 discusses an overview of the experiment and evaluation of approaches ending with section 10 with conclusion and future work.

3 Related Work

Some research has been carried out on sandhi splitting in Tamil and other Indian languages. However, most of this research has centered on internal sandhi phenomena. Our study, on the other hand, addresses the splitting of external sandhi words in cases where two or more fully formed composite words are combined.

A CRF based sandhi splitter is built as part of a morphological analyser. We have tested it with our test set and found that only 26% words are splitted..

In the paper Kuncham (2015) the authors have employed a statistical method to perform Sandhi splitting in the Tamil and Malayalam languages. The testing results indicate an accuracy of 89.07% for Tamil and 90.50% for Malayalam, demonstrating the effectiveness of the approach. This methodology comprises two main components: segmentation and word generation, both of which utilize Conditional Random Fields (CRFs) as a key tool. Devadath (2014) in their work, they devised a hybrid method that leverages the phonological changes occurring when words are joined together in the context of external sandhi. This approach

combines the statistical identification of split points with the application of predefined character-level linguistic rules. As a result, their system currently achieves an accuracy rate of 91.1%. In the study by Devadath (2016) authors have addressed the issues related to Tamil treebanking, in the context of external sandhi. Explained the challenges posed by external sandhi in the syntactic annotation of Tamil sentences. Their experiment using the statistical parser to empirically validate the improvements made to the treebank stating that even the separation of a single type of external sandhi significantly enhances overall parsing accuracy.

The study by vempaty et al. (2011) introduced a method that employs finite state automata to identify possible words within compound words in Telugu. It is built using the syllables of base words, enabling the recognition of candidate words within compound structures. In a study conducted by Nair et al. (2011) they designed an algorithm aimed at breaking down Malayalam compound words into a series of morphemes by employing multiple levels of finite state automata. In the paper M.R. Shree (2016) they have adopted an approach to the internal Sandhi splitting technique in the Kannada language. In the work by Gupta (2009) authors conducted Sandhi-Vichched in Hindi words and assessed their software using a dataset of over 200 words through their rule-based algorithm.

4 Sandhi Splitting

Sandhi splitting is the process of splitting a given composite word into its constituent word forms\footnote{the first word form is termed as W1, the second word form as W2, and subsequent words are termed W3... Wn.}. The task of sandhi splitting becomes complex in agglutinative languages like Tamil as tokens obtained through tokenization can contain more than one morphological word within them. So, these tokens require segmentation at a different level. Sandhi splitting must be done as part of the tokenization step, as composite words cause the fusion of tokens. Tokenization serves as the initial step before engaging in sandhi splitting. This sequential approach ensures that the text is appropriately structured and analyzed, thus facilitating a deeper understanding of the intricate morphological processes.

5 Types of Sandhi

Sandhi in Tamil is realized in two ways: Internal Sandhi and External Sandhi. Detailed explanations of Internal sandhi and External sandhi are given here:

5.1 Internal Sandhi

Internal Sandhi, also known as *antar* sandhi, refers to phonological changes that occur within a single word, typically due to morphological processes like inflection and derivation.

5.1.1 Inflections:

Inflected words are base words that undergo grammatical changes to convey different meanings, such as verb conjugations, noun plurals, prepositions or postpositions, and case

markers are considered as inflectional markers. Table1 explains the inflectional suffixes in Tamil.

S.No	Word form	Word+inflectional suffix	Gloss
1.	பேனாவைக்கொண்டு	பேனாவைக்+கொண்டு pen + INS	‘with pen’
2.	வலதுபுறத்தில்	வலதுபுறம்+இல் right side + LOC	‘towards right side’
3.	பற்கள்	பல்+கள் tooth +PL	‘teeth’
4.	ஆடுவாள்	ஆடு+வாள் dance +will (FUT)+SG-3-F	‘she will dance’

Table1 : Inflectional Suffixes in Tamil

5.1.2 Derivations

Derivation involves creating new words or modifying the meaning of existing words by adding prefixes, suffixes, or infixes. These affixes alter the root word's meaning or grammatical category. Table2 explains the derivational suffixes in Tamil.

S.No	Word Form	Word+derivational suffix	Gloss
1.	அறிவான	அறிவு+ஆன Intelligent +ADJ	‘intelligent ‘
2.	நேராக	நேர்+ஆக Straight + ADV	‘straightly’
3.	பார்க்கக்கூடாது	பார்க்கக்+கூடாது see+not-AUX	‘do not see’

Table2: Derivational suffixes in Tamil

5.2 External Sandhi

External sandhi, also known as *bAhya* sandhi, involves phonological changes that occur at the boundaries of words when they come into contact, either due to word combination or sentence formation as a result of stylistic variation. External sandhi can be divided into two types.

5.2.1 Compound Words:

Compound words are formed by combining two or more complete words to create a new word with a distinct meaning.

S.No	Word form	W1+W2
1.	மதியவேளை	மதிய+வேளை 'afternoon'
2.	துவரம்பருப்பு	துவரம்+பருப்பு 'toor dal'

Table3: Compound words in Tamil

5.2.2 Composite words or Non-compound Words:

Composite words, also known as Non-compound words, contain two or more complete words within them but do not derive new or distinct meanings from this combination. Since these composite words do not derive new meanings on combining, these word forms are considered as two different tokens and hence to be split.

S.No	Word form	W1+W2
1.	சமைத்துக்கொடுத்தான்	சமைத்து+கொடுத்தான் cook +give-PST.SG.3.M
2.	கொடுமைசெய்வான்	கொடுமை+செய்வான் torture +do-FUT.SG.3.F
3.	விரைவாகப்போ	விரைவாக+போ fast -ADV +go

Table4: Composite words or non-compound words

6. Need for Sandhi Splitting

This explicit focus on the sandhi split of composite words is essential for several key reasons:

1. **Morphological Distinctions:** Composite words often consist of two or more fully formed word forms, each carrying different roots and corresponding grammatical features. These morphological differences are crucial to understanding the structure and meaning of the word. The language maintains clarity in its morphological structure by splitting these word forms.
2. **Syntactic Variations:** In composite words, the constituent word forms can display various syntactic relationships. These relationships help in understanding the overall sentence structure and meaning. Splitting these word forms helps disambiguate syntactic relationships and ensure the sentence retains its intended syntax.

3. **Semantic Clarity:** Composite words may convey multiple, distinct lexical senses when examined as separate components. Splitting them into their constituent parts enables a clearer understanding of the individual lexical meanings they contribute to the overall context. Combining multiple roots and grammatical features often leads to nuanced and context-specific meanings. Sandhi rules help in preserving and revealing these semantic nuances, ensuring that the intended message is conveyed accurately.

7 Rules for Sandhi Splitting:

Sandhi rules allow us to unravel the complexities of composite words, leading to a deeper comprehension of their structure, syntax, and meaning in language analysis and interpretation. These rules for sandhi are made based on the distinction between the Major and Minor categories. Major categories in Tamil, such as nouns (N), verbs (V), pronouns (PR), adjectives (JJ), and number words, have the capacity to take inflectional and derivational suffixes and often carry the core meaning in a sentence. Minor categories in Tamil, including quantifiers (QT), particles (RP), quotatives (UT), intensifiers (INTF), negations (NEG), etc., are often considered as closed-class words, and they are more stable in their form and do not readily take inflections or any derivational suffixes. We identified four major contexts in which word forms are conjoined that need to be split. The contexts are:

1. Major Categories + Major Categories

This combination involves sandhi between two major categories of words. Examples are given in table5.

S.No	W1+W2	W1	W2
1.	அளவிலிருக்கும்	அளவில் (N) 'in amount'	'இருக்கும் (N) 'be-FUT.3.NEU'
2.	அதிகமானவை	அதிகம் (N) 'More'	ஆனவை (V) 'become-PST.PL.3.NEU'
3.	அதிசயமாகும்	அதிசயம் (N) 'miracle'	ஆகும் (v) 'is'
4.	அதற்குள்ளாகவே	அதற்கு (N) 'for that'	உள்ளாகவே(ADV) 'within'
5.	ஆரம்பகாலங்கள்	ஆரம்பம் (ADJ) 'begin'	காலங்கள் (N) 'seasons'
6.	ஆர்வமுள்ளவர்கள்	ஆர்வம் (N) 'interest'	உள்ளவர்கள் (N) 'Who have-PRE.PL'

Table5: Major Categories + Major Categories

2. Minor Categories + Minor Categories

This combination involves sandhi between two minor categories of words. Examples are given in table6.

S.No	W1+W2	W1	W2
1.	அவ்வாறில்லை	அவ்வாறு 'like that'	இல்லை 'not'
2.	பிறகங்கே	பிறகு 'then'	அங்கே 'there'
3.	வேறெல்லாம்	வேறு 'something else'	எல்லாம் 'all'
4.	எப்படியென்றால்	எப்படி 'how'	என்றால் 'means'

Table6: Minor Categories + Minor Categories

3. Major Categories + Minor Categories

This combination is the interaction between major categories and minor categories. Major categories provide the core meaning, while minor categories modify the sense of the sentence being in the W2 position. Examples are given in table7.

S.No		W1+W2	W1	W2
1.	Concessive	தடங்கலிருப்பினும்	தடங்கல் (N) 'obstacles'	இருப்பினும் 'inspite of'
2.	Conditional	அவனிருந்தால்	அவன் (PRON) 'he'	இருந்தால் 'if'
3.	Quantifiers	கவலையனைத்தும்	கவலை (N) 'worries'	அனைத்தும் 'all'
4.	Interrogatives	பாடுவதெங்கே?	பாடுவது (N) 'to sing'	எங்கே? 'where'
5.	Distal	அவனங்கு	அவன் (PR) 'he'	அங்கு 'there'
6.	Proximal	இவனிங்கு	இவன் (PR) 'he'	இங்கு 'here'

Table 7: Major Categories + Minor Categories

4. Minor Categories + Major Categories

In this combination, minor categories are supplementary to major categories, often modifying or specifying the meaning of the major category word in the W2 position. Examples are given in table8.

S.No		W1+W2	W1	W2
1.	Quantifier	பலபேர்	பல 'many'	பேர் 'members'
2.	Interrogatives	எக்காரியம்?	எந்த 'which'	காரியம்? 'matter'
3.	Distal	அங்குள்ளோர்	அங்கு 'there'	உள்ளோர் 'people-be'
4.	Proximal	இங்குள்ளோர்	இங்கு 'here'	உள்ளோர் 'people-be'

Table8: Minor Categories + Major Categories

8. Implementation

8.1 Data Collection and Preprocessing:

We have collected a total of 1 Lakh sentences for this experiment. Out of which 7000 are manually verified. The remaining sentences are augmented using a sandhi bag of word equivalents. If a sandhi word is found in the corpus, then the corresponding target side is split using this database. If there is no match found, then the target side sentence remains as it is. This database has 4000+ entries which are collected manually.

All the sentences are tokenized for further processing. It is important that the model not only learns when it should split but also when not to split. So, to overcome this issue we are also including sentences which necessarily do not have sandhi words.

8.2 Workflow:

We are using the Transformer model architecture, a state-of-the-art deep learning model, which has shown a high success rate in various natural language processing tasks. Transformers can capture contextual information making them well-suited for sandhi split tasks.

Like every transformer model, our model comprises an encoder-decoder architecture, where the encoder processes the input text, and the decoder predicts where the sandhi splits occur. We implemented the transformer model using the OpenNMT Pytorch based toolkit.

8.3 Training:

The dataset is split into train and validation sets. The training set contains 39000 sentences and the validation set comprises 10000 sentences. An additional set of 1000 sentences are used as the test set. We have conducted 3 experiments on the dataset to figure out the best performing model as shown in table7.

8.3.1 Experiments

In the first experiment we trained the model using a basic transformer architecture model. The

model is trained up to 10000 steps. Since the transformer model required huge data this model results are moderate.

In the second experiment, we have arranged the data in such a way that each line consists of only one word on the source side and the target side consists either one word or its equivalent sandhi words separated by a '+' symbol. We call it the "Word Model". This data when trained using transformer architecture showed great improvement on the first experiment. The model is trained up to 10000 steps. The results can be seen in table9.

In the third experiment, we have arranged the word level data such that each single character recognized by Unicode standard is separated by a space. We call it the "Character Model". This data when trained using transformer architecture showed significant improvement over the first and second experiments. The model is trained up to 10000 steps. The results can be seen in the table9.

9 Evaluation:

For evaluation we have used Bleu score and ChrF2++ metrics. As mentioned in related work when tested with the test set CRF based splitter could work with only 26% accuracy which we consider as baseline to compare our models. From the results it is clear that Word level and Character level models perform better. Since bleu score works on sequence of words we are providing ChrF2++ score for this model.

Results:

Experiment Type	Bleu score	ChrF2++
Sentence level Transformer model	17.7	44.7
Word level Transformer Model	41.1	66.6
Character level Transformer Model	NA	75.0

Table 9: Results

10 Conclusion and Future Work:

Tamil being an agglutinating language needs external sandhi splitting for the downstream applications in NLP. From the results, it can be inferred that the performance improves significantly when trained at Word level and Character Level with transformer models. Future experiments can be conducted using more data with pre-trained transformer models and fine tuning with Large Language Models (LLM) to get more improved results.

References

Devadath V V, Litton J Kurisinkel, Dipti Misra Sharma, and Vasudeva Varma. 2014. A Sandhi Splitter for Malayalam. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 156–161, Goa, India. NLP Association of India.

Devadath, V. V., and Dipti Misra Sharma. 2016. "Significance of an accurate sandhi-splitter in shallow parsing of dravidian languages." *Proceedings of the ACL 2016 Student Research Workshop*.

Gupta, Priyanka, and Vishal Goyal. 2009. "Implementation of rule based algorithm for Sandhi-Vicheda of compound hindi words." *arXiv preprint arXiv:0909.2379*.

Kuncham, Prathyusha, et al. 2015. "Statistical sandhi splitter for agglutinative languages." *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I* 16. Springer International Publishing.

Nair, Latha Ravindran and S. David Peter. 2011. "Development of a rule based learning system for splitting compound words in Malayalam language." *2011 IEEE Recent Advances in Intelligent Computational Systems* (2011): 751-755.

Shree, M. Rajani, Sowmya Lakshmi, and B. R. Shambhavi. 2016. "A novel approach to Sandhi splitting at Character level for Kannada Language." *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*. IEEE.

Klein, Guillaume, et al. 2020. "The OpenNMT neural machine translation toolkit: 2020 edition." *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*.

Vaswani, Ashish, et al. 2017. "Attention is all you need." *Advances in neural information processing systems* 30.

Deshmukh, R., Bhojane, V., & PIIT, N. P. 2014. Sandhi splitting techniques for different Indian languages. *International Journal of Engineering Technology, Management and Applied Sciences (ijetmas)*, 2(7).

Parameshwari, K. 2011. "An implementation of APERTIUM morphological analyzer and generator for Tamil." *Parsing in Indian Languages* 41.

Rajan, K., V. Ramalingam, and M. Ganesan. 2012. "Machine Learning for Sandhi Rules in Tamil." *Proceedings of the 11th International Conference INFITT*.

Vempaty, Phani Chaitanya, and Satish Chandra Prasad Nagalla. 2011. "Automatic sandhi splitting method for telugu, an indian language." *Procedia-Social and Behavioral Sciences* 27: 218-225.

இயல் அமை : கைபேசி மூலம் தமிழ் இணையதளம்

தங்கவேலு சின்னசாமி <ctv1957@gmail.com>

செயல்மன்றம், தமிழ் நாடு

மனிதத் தொடர் கண்டுபிடிப்பில் நம் உயிர்மெய் எழுத்துருவின் மூலத்தோற்றத்தை அறிவோம்.(1) நாம் பேசும் பேச்சு கைபேசி கருவி மூல கட்டமைப்பில் சில்லுத்தொகுதிகளின் (chipset) ஆற்றலில் இலக்குகளில், எண்ணி முறைமையில் (DIGITAL Process), 'இரும' (binary) கொள்கைகளில், இலக்கம் ஒன்றெனில்(1) திறக்கவும், சுழியமெனில்(0) மூடும் சில்லுகள் தொகுதிகளின் பெருக்கமே ஆகும். பல கருவிகளின் மதிப்பீட்டில், எண்ணி இரும கொள்கைகளும் இயற்கணக்கிலும் (Algebra) செயல்பட்டு, பேசும் தகவல்களை எல்லாம் இணைய முகவரி மூலமும் பரிமாறிக் கொள்கிறோம். ஒரு எண்ணியின் உட்பகு பணி பல கூறாகப் பகுத்து, எந்தெந்த பகுதி என்ன செய்ய வேண்டும் எனும் இலக்கத்தின் முறைமையுள் 'செய்நிரல்' (Programing Code) முறையில் செயல்படுகிறது. இவ்வாறான தகவல் பரிமாற்ற மொழி இலக்கணத்துடன், இணைய வடிவமைப்பு கட்டளைகளுக்குள்ளும், பேச்சு மொழியுடன் எழுத்துருக்களை ஒலிப்பதற்கு, எண்ணி (Digital) மூலம் கற்றல் முறை பயன்பட, பெரும் இணைய மொழி (படற உருவாகி இருக்கிறது. சொல்லின் அக்கணநேரக் கூற்றினை எழுத்துரு பகுப்பாய்வு செய்ய இக் கட்டுரை பயனுள்ளதாக இருக்கும்.

எழுத்துரு மொழியில் உருவாக்கும் இலக்கு:

இணைய வடிவமைப்பில் உள்ள ஊடாடும் குரல் பதிவு (Interactive Voice Response) மூலம் கணினி மொழியியல் மாதிரியில் பயன்படுத்துவதற்கு அந்தந்த மொழி இயலமை சொல்லமைப்பு வரையறையில் அவரவர்களின் கைபேசி திறன் தாய்மொழியிலும் எளிதில் அறியலாம். குறியீட்டுடன் எண்ணி (Digital) வடிவங்களில் இணைக்கப் பயன்படுத்தலாம். எண்ணி (Digital) அமைப்பு நான்கு அடிப்படை செயல்பாட்டுடன் , அதாவது, உள்ளீடு-வெளியீடு கருவிகள், முக்கிய நினைவக கட்டுப்பாடு, மற்றும் கணித-தருக்க அலகுகளை கொண்டு இயங்கும். தமிழ் மொழி இலக்கியத்திலேயே அகரமுதலி எழுத்துருவை பயன்படுத்தியே இரண்டு, மூன்று எழுத்திலேயே எண்ணியின் செயல்பாட்டினை கீழ்க்கண்ட முறையில் காணலாம்.

கணித தருக்க அலகு (LUN) ஒரு தனிப்பட்ட கட்டளையாக, இயல்பாகவே மெய்நிகர் சேமிப்பக கருவிகளின் தொகுப்பில் நியமித்ததை தனித்து அடையாளங்காட்டி, சிற்றியல்பான எண்ணி (Digital) தொடர்பரப்பி (805) மூலம் வரையறுக்கப்பட்டு, புரவன் நிரலுடன் (Host Computer) உள்ளீடுவெளியீடு (I/O) கட்டளைகளையுடன் செயல்படுத்தும் தரநிலை ஆகும். தமிழை , 'தம் இதழ்' என கூறலாம். எ.கா: 'த' என்ற {(ம+°)+(இ)}+ 'த' என்ற எழுத்தொலி உருபனுடன் சேர்ந்து, (ழ+°) இணையத்தில் தமிழ் எழுத்துருவாகும்.

த+மி(ம் +இ(த)ழ்), என அழைக்கப்படுவதும் எம்மொழியிலும் மனம், மெய்யியல்பில், மொட்டு விட்டு நாவில் வாய் மூலம் வெளிப்படுதல் எம்மொழிக்கும் பயன்படும். எ.கா: தமிழ், அதாவது, தம் இதழ் 'தமி' என்ற தமிழ்(ம-+°) +(இ)-எழுத்தொலி உருபனுடன் சேர்ந்து, (ழ(ழ+°)) என இணையக்கில் ஏற்படுத்தலாம். (தமி(ம் -இ)(த)ழ்), எனும் உருபனை தம் மனம், மெய்யியலில் மொட்டாக மலருதலே மொழி என்போம். இதை தமிழ் ஊடகங்களில் பணிக்கும் இணையும் இணையதளமாக எண்ணி முறைமை 'இரும (binary) இலக்கத்தினைக் கொண்ட கட்டுப்பாட்டுடன் தானியங்கி கருவியாக கைபேசி பல நாடுகளில் பல்வேறு மொழிகளில் பெருகி வருகிறது. தமிழ் இன

மக்களின் அகத்தில் தோன்றியவை கருத்தாளத்தில் மிகுந்து, தொகுப்பாக, இன்று இணையத் தளம் வரை தமிழ் மொழியிலும் உள்ளன. நமது அகத்தில் தோன்றி, எழுத்துருவில் ஊடுருவி, எண்ணி (Digital) வரை , ஓர் ஊடகமாக திகழ்கிறது.

'எண்ணி ஒலி மூலம் இலக்கத்தின் தரவுகளிலேயே, முறைப்படி இணையும் கட்டளைகளை, செய்நிரல் , கைபேசி மூலம் ' எண்ணி(Digital)' வரையறை 'இரும கொள்கை தொடர்புடன் பெருகி வருகின்றன. உயிர்மெய்யுடன் எழுப்பும் ஒலிக்குறிப்பில் புரியும்.

எ.கா: 'ப' எனும் எழுத்துரு ஒலிப்பு, பலராலும், பலகாலம் தங்களது தேவைகளை நிறைவேற்ற, பல அறிகுறிகளிலேயே சொற்களாக தமது குரல் வெளிப்பாட்டில் அந்தந்த மொழிப் பழக்கத்தில் உருவாக்கப்படுகிறது. (2) , (3), (4) தமிழ் எழுத்துரு 'ப' எனும் மெய்யெழுத்து ப(ப+அ) எனும் தமிழ் மொழி அகரவரிசையில் முறைப்படி ஒலித்தனர். பின்னர் ப (ப-அ) எனும் 'அகர வரிசை ஒலிப்பு 'ஆ' என ஆகார ஒலிப்பினில் 'பா' என நிலைத்திருக்கிறது.

கற்றலில் பா என வடிவம் அறிந்து பயன்படுத்தி எழுத்துருவில் 'ல+அம்' 'பாடம்' என எழுத்துரு வாக அறிந்து கொண்தை நூற்றாண்டு எழுத்துரு படத்தின் மூலம் அறியலாம்.

இக்காலத்தில், இணையம், பத்துப்பாட்டு, எட்டுத்தொகை, எனத் தொடங்கி, பதினெண் மேற்கணக்கு, பதினெண் கீழ்க்கணக்கு தானியங்கியில் பாவினமாக பவனியெங்கும், தமிழ் இலக்கிய வரலாற்று பதிவுகளிலும் இணையதளம் வரை பதிவிலும் தொடர்கிறது. அன்று தகவல் தொடர்பு ஒருவருக்கொருவர் ஒலிப்பு முறையில் பரிமாற்றம் செய்தததை, இன்று கைபேசி மூலம் இணையதளம் வரை தொடர்கிறோம். எண்ணி இரும கொள்கை கணித தருக்க அலகு பலரும் தேவைபடும் வகையில் இயங்கும் பல கருவிகள் நடைமுறையிலும் பெருகியுள்ளது. இந்த எண்ணி இரும கொள்கை இயற்கணக்கு அடிப்படையில், மொழி இலக்கண இணைய வடிவமைப்பு மேம்பாட்டிற்கும், உள்ளூர் பேச்சு மொழியுடன் பழைய எழுத்துருக்களை ஒலிபெயர்ப்பதற்கும், எண்ணி (Digital) மூலம் கற்றல் முறையைப் பயன்படுத்தும் இணைய மொழியாக உள்ளது. மனித உளவியல் நேர்மறை, நடுநிலை, எதிர்மறை அணுகுமுறை என மதிப்பீட்டு அளவில் கண்டறியலாம். ஆய்வு தள பகுப்பாய்வு 'செயல் மன்றம்' என சொல்லில் உள்ள எழுத்துருவை கரந்து உறைகளில் வைப்பது போல், ஒவ்வொரு சொல்லும் எழுத்தையும் அந்த சொல்லில் உள்ள எழுத்துரு பொருள் வரையறை கொண்டு "கரந்துறை" ஆய்வு கட்டுரை தமிழ் சமூகத்தில், கைபேசி துணை கொண்டே தமிழ் மொழி பரிமாற்றம் செய்து வருகிறோம். 'கரந்துறை பேச்சு' எழுத்துரு கொண்டே சொற்பொழிவு மூலம் பல்வேறு துறைகளிலும் உரையாற்றும் நிலை பெற்றுள்ளது. தமிழ் மொழியில் அறிந்து கொள்வது மிகப்பெரிய சவால் எனினும், தொடர்ந்து பதிவிடுகிறோம். ஆன்டராய்டு உள்ளீடு வெளியீட்டு கருவியம் பல்வேறு வகையான பயன்பாடு கொண்டது ஆகும். இக்கருவியத்தினை அடிப்படையாக கொண்டு இணைய முகவரி மூலம் ஒரு தகவலை பரிமாறிக் கொள்வோம்.

பட் என்றால் இணையும் இணையகளம்:

ஒரு எண்ணியின் உட்பகு பணியைச் செய்ய, அதனைப் பல கூறாகப் பகுத்து, எதன் பின் எதனைச் செய்ய வேண்டும் என இலக்கத்தின் முறைமையுள் செயலினை 'செய்நிரல்' மூலம் நிர்ணயிக்கும். நம் பகுதியில் உள்ள வரலாற்று பதிவுகள் தொடர்,

தொடர்பு நிகழ்வுகளை தொகுத்து அறிவது மிகவும் சிறந்த ஒன்றாகும். ஓர் இனம் மிகுந்து, வேறொரு வேற்றுமை இன புணர்ச்சியில் தோற்றம் தரும் இயற்கை உயிரினம் மனித உயிரின தோற்றமும், ஓர் மாற்றம் ஆகும். கடியலூர் கண்ணனார் பதித்த 'பட்டினப்பாலை' பாடல்கள் மூலம் ஒலி மற்றும் ஒலிக்குறிப்பு ஒவ்வொரு காலகட்டத்திலும் வெவ்வேறு மொழிகளில் பரவி உள்ள நிலையை அறியலாம். தொடர் அறிவு வரலாற்று நிகழ்வுகள் மூலம் அறிவதாகும். 'பட்டினம்' என்ற சொல், பெயரில் ஒவ்வொரு காலகட்டத்திலும் ஒலிக்குறிப்பு மூலம் பட்டறிவு தளமாக மொழி முறைமை வழிவழியாக ஆங்காங்கே பல்வேறு மாறுபட்ட எழுத்துருவில் நிலைத்து உள்ளதை காணலாம்.

பட்டினப்பாலை பண்டமாற்று முறை ஆரம்ப கால வாழ்க்கை வரலாற்றில் பதிவுகளுக்கு ஓர் எடுத்துக்காட்டு ஆகும். 'பட்டி' என்ற சொல் சிற்றூர் என பொருள்படும். பதிவுகளாக பாடல்களின் தொகுப்பில் இடம் பெற்றுள்ளது. Booty, 'Buddy', Port போன்ற ஆங்கில மொழி ஒலிப்பு முறை கிரேக்க மொழி, 'பட்' எனும் தமிழ் ஒலிப்பு கிட்டத்தட்ட அதே ஒலிப்பு முறைத் தொடர்பினைக் காணலாம். கிரேக்கத்தில் தொடங்கி அமெரிக்க ஆங்கில மொழி ஒலிப்பு முறை வரையறையிலும் உள்ள ஒற்றுமை தன்மையை கேட்பதில் பேசும் திறனை புரிந்து கொள்ள முடியும். பட், பட்டென், பின், ஆங்கில மொழி Port எனவும் விரிவாக்க ஒலிப்பு முறையை அறிந்து கொள்ளலாம். மனிதர்கள் வாழும் முறை ஒலிப்பியல் சைகைகளின் மூலம் இது போன்ற பாடல்கள் தெரிவிக்கின்றன. 'பட்' போன்ற சொல் கிரேக்க, லத்தீன் மொழியில் உள்ள ஒலி, பண்டைய வரலாற்று குறிப்பு ஒலிப்பு ஒன்றிய சைகை குறிப்பாகும். பட்(ட+°) என்ற சொல் முறைப்படி டி(ட+இ) என பாக்களில், பட்டி, இனமாக ஒன்றி 'பட்டினம்' என்ற சொல் 'இயல் அமை முறை எந்த மொழியியல் மூலமும் பதியலாம். 'பட்டி' எனும் சொல், அறிவுசார்ந்த தமிழ் எழுத்துரு முன்னோட்ட பின்னோட்ட சொற்களாக விரிவடைந்து ஒவ்வொரு வாழ்வியல் நிலையிலும் 'பட்', 'பட்டி', 'பட்டியல்' என்று பல எழுத்துரு அளவு முன்னும் பின்னும் சேர்த்து ஆன்றோர் பெருமக்கள் ஊர், ஊராக ஊர்ந்து சென்று ஊர், ஊராட்சி, பட்டி என சிறுசிறு கிராம அளவில் பயிர் தொழில் செய்து வருவது நாம் காணும் கண்கூடு.

பயிர் தொழில் பயிற்சி மூலமே, ஆக்கமாக ஊர், பட்டி என விரிவடைந்து நிலைத்து இருக்கிறது. (3) 'வருவாய்' எனும் சொல் 'இயல் அமை நிலையினைக் காண்போம். வருவாய், வரும்படி, பொருள் விளங்க முடியும். சூழலும் புவி ஒவ்வொருவருக்கும் அடிப்படை வாழும் தன்மை நிலைக்கப்படுகிறது. சூழலும் புவியில் ஒவ்வொருவருக்கும் ஆதாரமாக நிலைப்படுவதை, புவி தள கொள்முதலுடன் கூடிய தொடர் மனித இன அறிவாகும். வருவாய் கொண்ட பொருட்கள் பெறும் ஆற்றல் பெற பயன் படுத்தும் உழைப்பு, வேலை தொடரும் நிலைப்பாடு, பெறும் ஆற்றலும், கருப்பொருளில் நிரம்பிய பயிற்சியும் கலந்து ஈடுபடுவது தொழில்சார் நிலை ஆகிறது. பயிர் தொழில் நுட்ப அறிவும், அனுபவமும் தொடர்ந்த நிலையில், வானம் வழங்கும் நீரினை அறிந்து கொண்டு உணவுபொருளை பெறுகிறோம். ஜம்பொறிகளில் ஒவ்வொன்றும் விளங்கிய காலம் என்ற ஒன்று உண்டு. தள வரைபடம் மூலம் பல முறைப்படி வழங்க முடியும் வரை செல்லும் வல்லமை கொண்டது. தள மூலப்பொருள் வளம் தரும் வழிபாட்டுக்கு உரியவை, நல நீர்ச்சுற்றே நில அமைப்பு. தம்மிதழ் வழங்கும் மெய்யுறுப்பு தொகுப்பு, சொல்லுடன் கூடிய மொழி முறைமை ஆகும். பலரும் அறியச் செல்லும் வழியே மொழிதனில் நிலைக்கும். கூடுதல் பணியில் கிடைக்கும் வெகுமதியும், ஊருக்கு வழங்கும் தன்மையுமே, தனம் தரும். பற்றுதல் கொண்டு செயல்படும் திறன் கொண்ட மனிதர்களால் மேன்மேலும் செழிப்பது நாடு. இயல்பிலே ஈடுபட்டு, காத்து வகுத்தலில் வல்லமை கொண்டதே அரசு. அனைவருக்கும் கல்வி இயக்க முறைமையே தற்கால கருத்தாகும்.

எழுவாய்! பயனிலை அடை! பயன்படும் பொருளாகும்!

தற்கால இயக்கமுறையும் இயற்கை நிலையை பின்பற்ற வேண்டிய வழிமுறைகளே நிலைக்கும். நடைமுறை கருத்தில், இடைத் தொகுப்பு அறிவதும், கல்விப் பயிற்சி நிகழ்கால செயல் முறையுடன், வருங்கால தலைமுறை காக்கும். வருவாய் என்ற சொல்லில் வரும் வழி மெய் உயிர் வாய்மொழியிலும் வாய்ப்பு. செல்களின் இணைப்பு. உயிர் மெய் செல்களின் வரும் வழி வாய்ப்பு மொழிகளில் உண்டு. பல காலம் மனித உறுப்புகளில் 'எழுவாய்' என பெயரில் நிலைத்து, பயன்படும் பொருளை செயல் திறனை அறிந்த 'பயனிலை' 'பயன்படும்பொருள்' ஆக சொல்லாற்றல் நிலைபெற இலக்குகள், 'இலக்கியம்' ஆகவும், அக்கணமே புரிதலில் உள்ள இலக்கு, 'இலக்கணம்' ஆகவும், ஒவ்வொரு மொழியிலும் உள்ளது போல் தமிழ் மொழியில் நிறைந்து உள்ளன. மனித இன மெய்யுறுப்புகளை தொடர்புபடுத்தும் ஆறனை(RNA), அடிப்படையில் நான்கு அடிப்படைகூறுகளை(A) கோர்த்து (G), இணைத்து (C) உட்கருவினை(U) ஒவ்வொரு தாயனை(DNA) மூலமும் உட்கருவாகிறது.(5) வட்டத் தலைமுறை தாங்கும் திறன் , அறம், பொருள், இன்பநிறைவு குறியீட்டெண் கணக்கீடு அளவு பதிவாகும். இறைவரி: இறை, இயற்கை எனும் சொல் இறைந்து இயற்கையில் கிடைக்கும் அகப்புற நிலைப்பாடு ஆகும்.

"சொல்லிய முறையால் சொல்லவும் படுமே" என்கிறார், தொல்காப்பியர். ஆன்ராய்டு மற்றும் பல ஆப்பிள் உள்ளீட்டு, வெளியீட்டு கருவிளுக்கான எந்த ஒரு தளமும் பயன்படுத்த உதவும் இணையகளத்தை உருவாக்கலாம். இணைய மேம்பாடு என்பது இணைய உலாவிகளில் வேலை செய்யும் வலைத்தளங்கள் மற்றும் இணைய பயன்பாடுகளை உருவாக்கும் செயல்முறையாகும். திறன்பேசிக்கு ஏற்ற இணையதளத்தை தேடுபொறிகள் மூலம் மிகச்சிறந்த இடத்தை பெற்று இருக்கிறது. தற்கால கணிப்பின்படி 60 விழுக்காடு திறன்பேசி என்பதை நாம் கண்கூடாக காண்கிறோம். வணிக வலைத்தளம் பெரும்பாலும் திறன்பேசி மூலமே கிடைக்கிறது. இணையகள வேகத்தை மேம்படுத்தி, காட்சிகளிலும் வேகமாகவும் மிகப்பெரிய தோற்றத்தையும் ஏற்படுத்துகிறது. திறன்பேசி நேர அளவிலும் ஐந்து நொடியில் ஏற்றப்படும், ஒரு நொடியில் ஏற்றப்படும் இணையதளமாக, கிட்டத்தட்ட மூன்று மடங்கு அதிகமான மாற்று விகிதத்தைக் கொண்டுள்ளது என போர்டென்ட் நடத்திய ஆய்வில் தெரிவிக்கப்பட்டுள்ளது. திறன்பேசி படங்களை சுருக்கி, பார்ப்பதற்கு ஏற்ற வேண்டிய தரவு அளவாக குறைக்கப்பட்டு, இணையதள வேகத்தையும் அதிகரிக்கிறது. திறன்பேசி மூடுவதற்கு ஏற்ற வகையில் இருப்பதும் மிகவும் முக்கியமான ஒன்றாகும்.

முடிவுரை :

கைபேசி மூலம் இணையதள உருவாக்க நெறிமுறைகள் பல வழிகள் உள்ளன. முதலில், பயனர் திறன்பேசி கருவிகளில் இணையதளம் உருவாக்கிய பின் செயல்படுகிறதா எனப்பார்க்கவும். ஆன்ட்ராய்டு மற்றும் உள்ளீட்டு வெளியீடு கருவிகளைப் பயன்படுத்தியும் கூகுளின் திறன்பேசி நட்பு முறையில் இணையதளத்தை ஒவ்வொரு பக்கத்தையும் பார்ப்பது மிகவும் பயனுள்ளதாக இருக்கும்.

References

1.தமிழ் உயிர்மெய் எழுத்து வரலாறு

History of Tamil Script Picture Presented by Karaikudi Kamban Adipodi Sa. Ganesan

https://youtu.be/_b4KKdApORE?feature=shared.

2. உயிரின ஒலி மெய் உறுப்புகளில் எவ்வாறு நிலை பெறுகிறது.
<https://youtu.be/OOkWF27hHc0?si=Y4UGSHZag15vjTTv> 3(a), b

3. குறி அறிகுறி சைகை வடிவம் மொழியாகிறது.
<https://youtu.be/i0Jn38YbtOw?feature=shared>.

3(b) குறி, அறிகுறி தகவலறிகுறியும் நிலைகுறியாகட்டும்.
<https://youtu.be/OOkWF27hHc0?si=Y4UGSHZag15yjTTv>

4. Why அ, ஆ. A Sound in all Languages?
<https://youtu.be/OOkWF27hHc0?si=Y4UGSHZag15yjTTv>.

5. ப விளக்க படம்

6. Why அ, ஆ, A Sound in all Languages?
<https://youtu.be/ILZZKXt23BU?feature=shared>

SIGNET: Superior Intelligence for Gesture-based Neural Exploration in Tamil

Kanimozhi Suguna S1*, Prema S2, Vasanthakumari M3

¹*Assistant Professor, Department of Computer Applications

²Assistant Professor & Head, Department of Computer Applications

³Assistant Professor & Head, Department of Tamil

Arulmigu Arthanareeswarar Arts and Science College, Tiruchengode, Namakkal Dt

Abstract:

Despite significant advancements in sign language recognition, Tamil Sign Language (TSL) remains underexplored, with existing methods failing to capture its intricate nuances. This study introduces “SIGNET” - Superior Intelligence for Gesture-Based Neural Exploration in Tamil, a novel framework that employs a hybrid model of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers to address this gap. Current models often struggle to accurately recognize TSL gestures, leading to imprecise results and limited practical application. This inadequacy poses significant accessibility challenges for the Tamil-speaking deaf community, who lack a reliable method for recognizing Tamil script through sign language. SIGNET leverages state-of-the-art deep learning techniques to enhance the accuracy and robustness of TSL gesture recognition. By combining CNNs, RNNs, and Transformers, the proposed model captures the detailed spatial features of hand shapes and facial expressions (via CNNs), the temporal dynamics of gesture sequences (via RNNs), and long-range dependencies and contextual understanding (via Transformers). This hybrid approach addresses the specific challenges posed by TSL, such as the subtle variations in hand gestures and the spatial-temporal dynamics involved. The innovative integration of these neural network architectures allows SIGNET to provide a more comprehensive and sophisticated understanding of TSL gestures. The study’s findings are expected to significantly improve accessibility and inclusivity for Tamil-speaking deaf individuals, fostering a more cohesive and inclusive community. Through resolving the limitations of existing models, SIGNET establishes a robust foundation for TSL recognition, with both theoretical and practical implications.

Keywords: Tamil Sign Language (TSL), Gesture-based Recognition, Neural Exploration, Superior Intelligence for Gesture-Based Neural Exploration in Tamil (SIGNET), Tamil-speaking deaf community, TSL gestures, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers.

Introduction

Tamil Sign Language (TSL) is a crucial means of communication for the Tamil-speaking deaf community. It encompasses a rich array of gestures, facial expressions, and spatial movements, allowing individuals who are deaf or hard of hearing to communicate effectively. Despite its significance, TSL has historically received less attention in research and technological development compared to other sign languages, such as American Sign Language (ASL) or British Sign Language (BSL). This oversight has contributed to a

communication gap, limiting the accessibility and inclusivity for Tamil-speaking individuals with hearing impairments.

In recent years, the significance of TSL has gained increased recognition, driven by the growing awareness of the need for inclusive communication solutions. Advances in technology, particularly in the fields of artificial intelligence and machine learning, have opened new avenues for developing robust sign language recognition systems. These technological advancements are crucial for enhancing the quality of life for the deaf community, enabling better access to education, employment, and social interactions. The development of sophisticated models capable of accurately interpreting TSL can facilitate more effective communication, thereby promoting inclusivity and equal opportunities.

The renewed focus on TSL is also reflective of a broader societal shift towards embracing diversity and ensuring that technological progress benefits all segments of the population. By addressing the unique challenges associated with TSL, researchers and developers are working to create tools that can bridge the communication gap, fostering a more connected and empathetic society. As such, the study and advancement of TSL recognition not only contribute to technological innovation but also underscore the importance of cultural and linguistic diversity in the digital age.

Problem Statement

The existing models for Tamil Sign Language (TSL) recognition are inadequate in capturing the intricate nuances and variations inherent in TSL gestures. This inadequacy leads to inaccuracies and inefficiencies in real-world applications, limiting effective communication for the Tamil-speaking deaf community. Despite advancements in sign language recognition technologies for other languages, TSL remains underexplored, creating a significant research gap. The lack of a robust and precise recognition system for TSL hinders the inclusivity and accessibility of communication tools for the deaf community.

Objective

The primary objective of this research is to develop “SIGNET” - a Superior Intelligence Framework for Gesture-Based Neural Exploration in Tamil Sign Language. This framework aims to:

1. Enhance the accuracy and robustness of TSL gesture recognition using advanced deep learning techniques.
2. Address the unique challenges posed by TSL, such as variations in hand movements, facial expressions, and spatial dynamics.
3. Bridge the existing research gap by providing a comprehensive and efficient TSL recognition system.
4. Improve communication accessibility for the Tamil-speaking deaf community, thereby fostering a more inclusive and connected society.
5. Establish a foundation for future research and development in TSL and other underrepresented sign languages.

Recent Works on Tamil Sign Language

The study in [1] focuses on neural machine translation (NMT) for the Tamil-Telugu language pair and addressed the limitation of parallel corpora by utilizing monolingual data through pre-trained word embeddings in a transformer-based NMT model. This study [2] focuses on recognizing and categorizing South Indian Sign Language gestures based on different age groups through transfer learning models. The researchers utilized a dataset comprising 30,000 images of double-handed gestures. Within this dataset, there were 10,000 images for each considered age group: 1-7, 8-25, and 25 and above. The researchers of [3] aimed to address the communication barrier faced by individuals who use American Sign Language (ASL) by creating a real-time ASL recognition system. By bridging the gap between ASL and text, this system enhances accessibility and inclusivity for the deaf and hard of hearing community. The study by the researchers [4] focuses on bridging the communication gap between individuals who use Tamil Sign Language (TSL) and those who communicate through spoken language and focused on the challenges such as TSL involves intricate hand movements, facial expressions, and body postures. [5] study focuses on real-time hand gesture recognition in the Turkish sign language detection system. The researchers utilize the YOLOv4-CSP algorithm, which is based on a Convolutional Neural Network (CNN) and is a state-of-the-art object detection method. The goal is to achieve real-time and high-performance detection of sign language gestures

The researchers [6] propose two novel machine learning models (Vision Transformer, and Lightweight Convolutional Neural Network) for hand gesture recognition using publicly available datasets: the NUS Hand Posture Dataset I, the Turkey Ankara Ayrancı Anadolu High School's Sign Language Digits Dataset, and the American Sign Language dataset. The primary aim of [7] work is to provide real-time recognition of TSL gestures and translate them into Tamil letters. To achieve this, the researchers employed a Convolutional Neural Network (CNN) as a classifier. In the work [8], the researchers propose a novel approach for emotion analysis in the Tamil language using transformer-based models. The study in [9] aims to improve machine translation quality between Sinhala and Tamil, two languages with limited resources with the implementation of low-resource Neural Machine Translation (NMT) for the Sinhala-Tamil language pair. Transformer models can outperform existing state-of-the-art Statistical Machine Translation models by up to 3.28 BLEU points in Tamil-to-Sinhala translation scenarios. As an improvised work with Artificial Intelligence (AI) is presented in the article [10] with the adoption of a transformer-based neural network which is capable of analyzing over 500 data points from a person's gestures and face to translate sign language into text.

The following Table 1 provides the consolidated overview of the recent models.

Table 1: Overview of the recent research on Sign Languages

Ref.	Study	Methodology	Datasets	Results
[1]	Neural Machine Translation (NMT) for Tamil-Telugu Language Pair	Utilized monolingual data with pre-trained word embeddings in a transformer-based NMT model to address the limitation of parallel corpora.	Not specified	Improved translation quality using transformer-based model leveraging monolingual data.
[2]	Recognizing South Indian Sign Language Gestures Across Age Groups	Applied transfer learning models on a dataset of 30,000 images, categorized by age groups (1-7, 8-25, 25 and above).	30,000 images of double-handed gestures (10,000 per age group)	Successful categorization of gestures based on different age groups, enhancing recognition accuracy.
[3]	Real-Time ASL Recognition System	Developed a real-time recognition system for ASL using machine learning to bridge the gap between ASL and text.	Not specified	Enhanced accessibility and inclusivity for the deaf and hard of hearing community through real-time ASL recognition.
[4]	Bridging Communication Gap with Tamil Sign Language (TSL)	Addressed challenges in TSL, including intricate hand movements, facial expressions, and body postures, to improve communication between TSL users and spoken language users.	Not specified	Improved understanding and translation of TSL gestures, reducing the communication gap.
[5]	Real-Time Turkish Sign Language Detection	Implemented YOLOv4-CSP algorithm based on CNN for high-performance, real-time detection of sign language gestures.	Not specified	Achieved real-time and high-performance detection of Turkish sign language gestures.

[6]	Novel Machine Learning Models for Hand Gesture Recognition	Proposed Vision Transformer and Lightweight Convolutional Neural Network for hand gesture recognition.	NUS Hand Posture Dataset I, Turkey Ankara Ayrancı Anadolu High School's Sign Language Digits Dataset, American Sign Language dataset	Improved hand gesture recognition accuracy using novel machine learning models.
[7]	Real-Time TSL Gesture Recognition and Translation	Employed a CNN as a classifier to provide real-time recognition of TSL gestures and translate them into Tamil letters.	Not specified	Achieved real-time translation of TSL gestures into Tamil script, enhancing communication for TSL users.
[8]	Emotion Analysis in Tamil Language	Proposed a novel approach for emotion analysis using transformer-based models.	Not specified	Improved accuracy in emotion analysis for the Tamil language using transformer-based techniques.

Facilitating Efficient Web Search Engine for Language Community

Prema S^{1*}, Kanimozhi Suguna S², Vasanthakumari M³

^{1*}Assistant Professor & Head, Department of Computer Applications,
premashanmuga11@gmail.com ; ²Assistant Professor, Department of Computer Applications
dr.kanimozhisuguna@gmail.com; ³Assistant Professor & Head, Department of Tamil
ilavenil2010tamil@gmail.com

Arulmigu Arthanareeswarar Arts and Science College, Tiruchengode, Namakkal Dt

Search engines are playing a major role in effective retrieval of data. In this research paper it is planned to facilitate an effective search engine for Tamil language community. Organizing the huge quantity of Web data in a coherent and precise way is momentous for using it as an information source. Nearly eighty percent of the users anticipate the best results in the first two pages to speed up the research. BookShelf Data Structure, a new original explore effort proposed in this paper combines the procedure of mutually the resizable array list and doubly-linked lists. Keyword is playing a major role in this research paper. Relevant documents in Tamil search are clustered using the agglomerative hierarchical clustering and arranged in BookShelf Data Structure. B3-Vis (Branch and Bound BookShelf structure for Visualization) technique is proposed for visualizing the results retrieved from the Branch and Bound BookShelf structure. The main involvement of this research paper is to save the efforts and time by helping the Tamil language users to find more related results fit with their interest arranged in BookShelf Data Structure.

KEYWORDS Web Mining, BookShelf Data Structure, Visualization, Tamil Community, Personalization.

1 INTRODUCTION

Web mining is the application of data mining techniques to discover patterns from the Web. Web Resource includes text, image, video and multimedia. Web mining can be categorized as Web usage mining, Web content mining and Web structure mining. Web usage mining is the process of digging out valuable information from server and also it is the follow up of finding out what clients are staring for on the Internet. Web usage mining itself can be grouped further based on the type of conventional data considered like Web server data, Application server data and Application level data.

Web structure mining is the practice of using graph presumption to examine the node and association structure of a Web site. Based on the type of Web structural data, Web structure mining be able to done as follows like extracting patterns from hyperlinks in the Web and mining the document structure. Web content mining is the mining and incorporation of valuable information and awareness from Web page content. In the present situation,

Web page result personalization is in concert. One of the most significant benefits of modified explore is the development in the excellence of decisions users completed. The semantic Web is the annexure of World Wide Web that facilitates public to allocate content ahead of the limitations of functions and websites. It aims at translating the current Web, conquered by unstructured and semi-structured credentials into a Web of data.

Information retrieval (IR) is a concept of choosing the significant information from a manuscript database in response to search questions given by an end user. Current WebIR is a regulation which has broken some of the traditional consequences of IR increasing novel models of information access. Crawlers are disseminated agents which collect information from the Web. They creep the Web pages according to the specified query and accumulate the folios in a restricted page warehouse.

In the proposed work, categorization can be done automatically by separate classifiers learning from training samples of text documents. The main aim of the classifier is to obtain a set of characteristics that remain relatively constant for separate categories of text documents and to classify the huge number of text documents into some particular categories (or folders) containing multiple related text documents.

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In clustering, it is the data distribution that will determine the cluster membership.

Data structure is a specialized format for organizing and storing data. It determines how data is recorded, manipulated, stored, and presented by a database. Analysing the importance of data arrangement in search engine BookShelf Data Structure has been introduced for community formation, which stores the inverse indices of Web pages (Jayanthi and Prema 2010). The main objective is to encourage communication among disciplines. BookShelf Data Structure, a new innovative research work proposed in this thesis combines the usage of both the resizable array list and doubly-linked lists

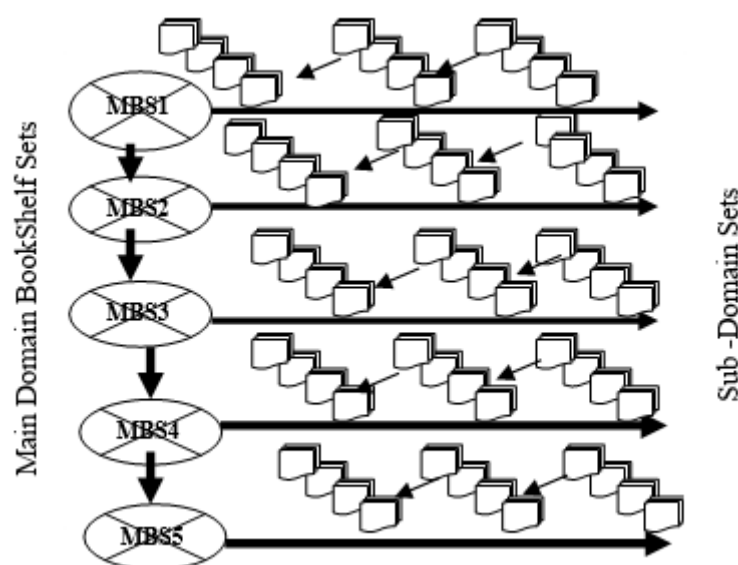


Figure 1 BookShelf Data Structure

2 LITERATURE REVIEW

Qingtian Han and Xiaoyan Gao (2009) examined a Web mining algorithm based on usage mining. This manuscript is useful to recognize the customer's performance and complete the website plan. Bar-Ilan (2005) talks about a variety of events such as database exposure, query reply time, consumer report and retrieval effectiveness for

evaluating the performance of examine engines. Akilandeswari.J and Gopalan N.P. (2008) examine an architectural frame of a crawler for locate deep Web repositories by learning multi-agent scheme. Alfredo Cuzzocrea and Carlo Mastroianni (2003) focused on the plan and development of (KM-bWS) information Management-based Web schemes. Scheming a new advance for the recovery of excellence information from the internet is a difficult assignment Feng Li (2008) argue an algorithm to mine the formation of a website mechanically support on hyperlink examination.

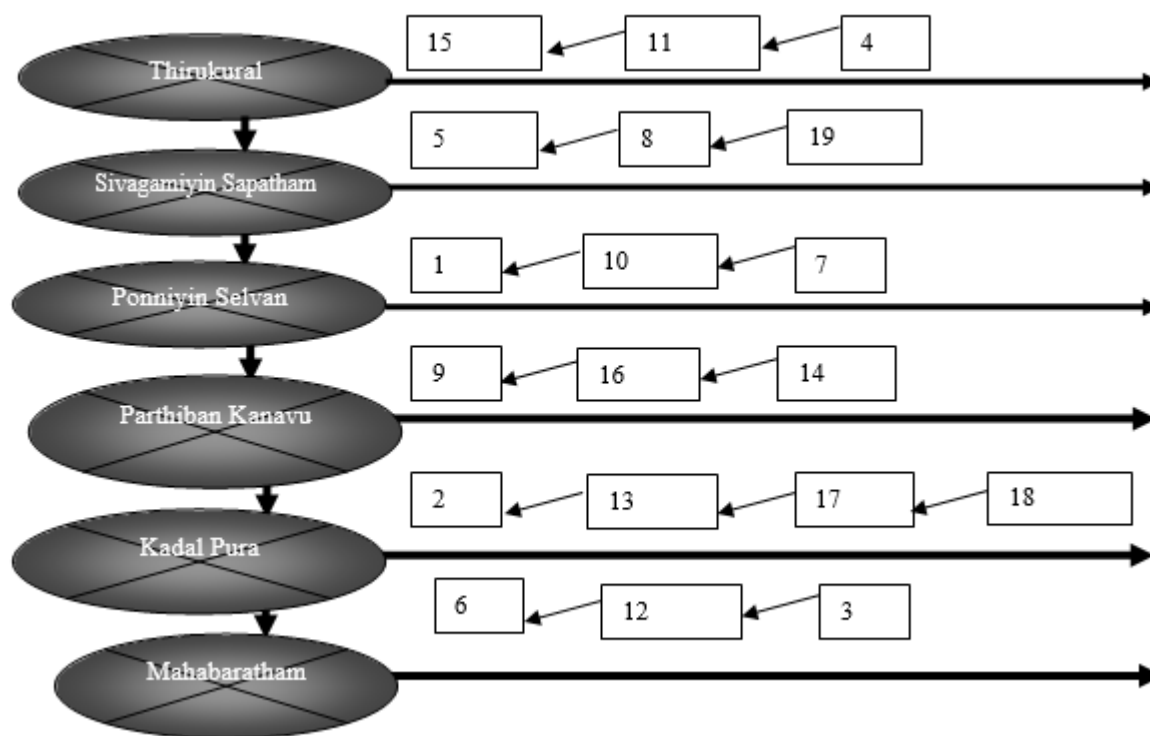


Figure 2 BSDSC with Descriptions

Dr S K Jayanthi and Prema intended a new structure of semantic similarity form on exploiting the sequential characteristics of historical click-through information. Jaime Teevan et al. (2010) discussed the techniques that leverage implicit information about the user's interests. This information is used to re-rank Web search outcomes in a significance response structure. Rich symbol of the consumer and the corpus are significant for personalization. In this research work an special data structure called BookShelf is argued for enhancing recovery of results. The retrieved consequences are accessible in visual mode for efficient access by the end user.

3 WEB SEARCH RESULTS PERSONALIZATION

Semantic Web search designed in this research work focuses on the Tamil Language domain based search through offline (Jayanthi and Prema 2013). Since the field is considered only for Tamil language community, Tamil related documents are focused. Apart from Tamil language professionals if there is any general search, then non- scholar community related documents are also suggested in the planned search engine. Non-scholar community

professionals are categorized as “others” in the proposed design. The steps involved in NLP are:

- Word sense disambiguation is achieved by creating semantic relations using WordNet
- Automatic thesaurus construction is performed using semantic classes
- Part of Speech (POS) Tagging is done using Tree-Tagger
- Phrase chunking finds all non-recursive noun phrases (NPs) and verb phrases (VPs)
- Syntax concerns the proper ordering of words and its effect on meaning
- Semantic concerns the (literal) meaning of words, phrases, and sentences
- Pragmatics concerns the overall communicative and social context.
-

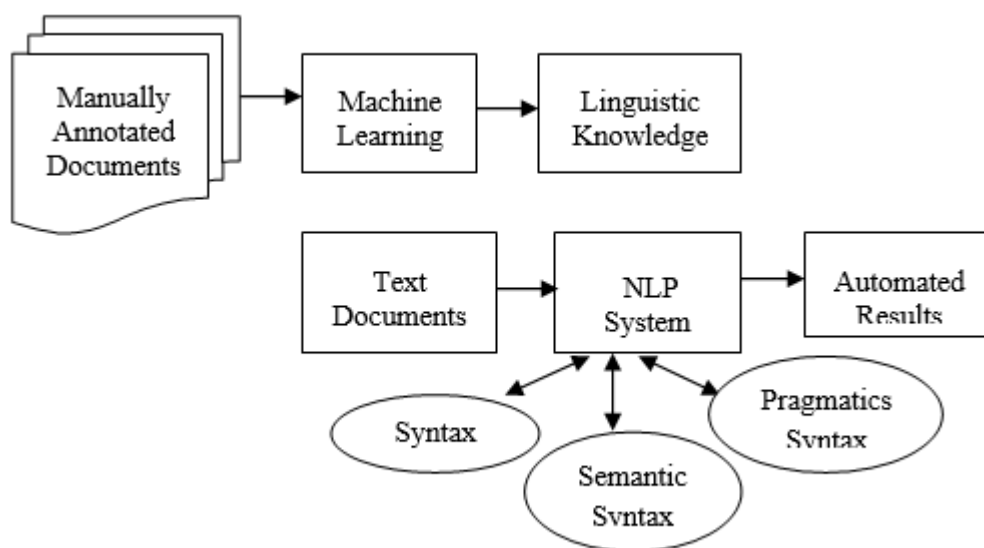


Figure 3 Learning Approach using NLP

4 BOOKSHELF DATA STRUCTURE

Document similarity is measured using cosine similarity propagation.

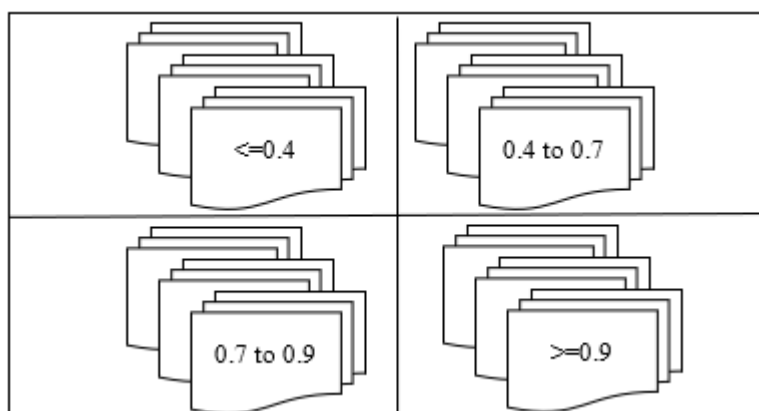


Figure 4 BSDS with Similarity R

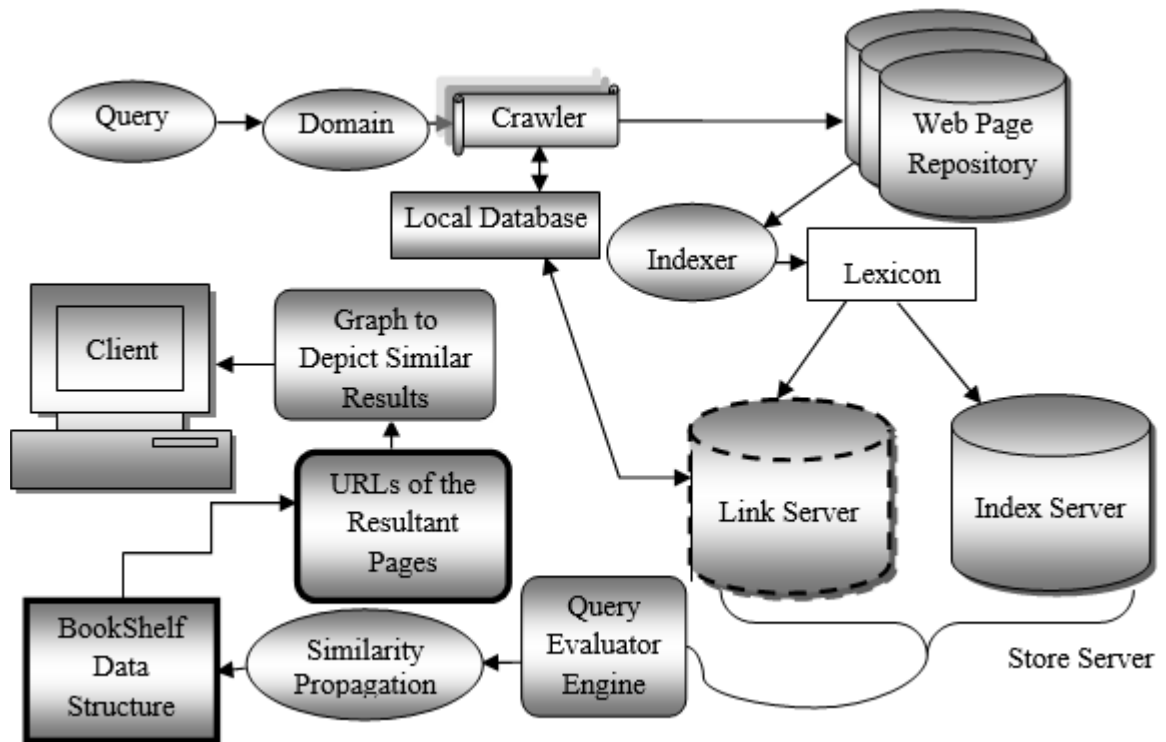


Figure 5 Domain Expert Search Engine

Two types of modes are analyzed in an offline search. One is a specialized explore based on Tamil Community domain-centric system. Document assortment and keywords focused in this segment are all based on Tamil Community domain. Main and sub domain model documents composed for specialized domain-centric search and the other is a non-professional search.

Table 1 Professional Domain-Centric Clustering

Main Domain	Sub-Domain Keywords
Ettu Thogai	"nattrinai", " kurunthogai", " agananuru", " purananuru"
Pathu Pattu	" madhurai kanchi", " kurunchi pattu", " pattinapalai", "Lover"
Paathinenkeilkanakku	" naladiyar", " thinaimaalai", "Thirukkural"
Sanga Illakkiyam	"Aagathiyam", "Tholgapiyam"
Sanga Noolgal	"Mazhargal", "Sangam", "Thinaigal"

Table 2 Non-Professional Documents Clustering

Main domain	Sub-domain keywords
Window	"Door", "Window", "Design", "Interior"
Apple	"History", "Apple", "Fruit", "Season"
Mouse	"History", "Mouse", "Home"

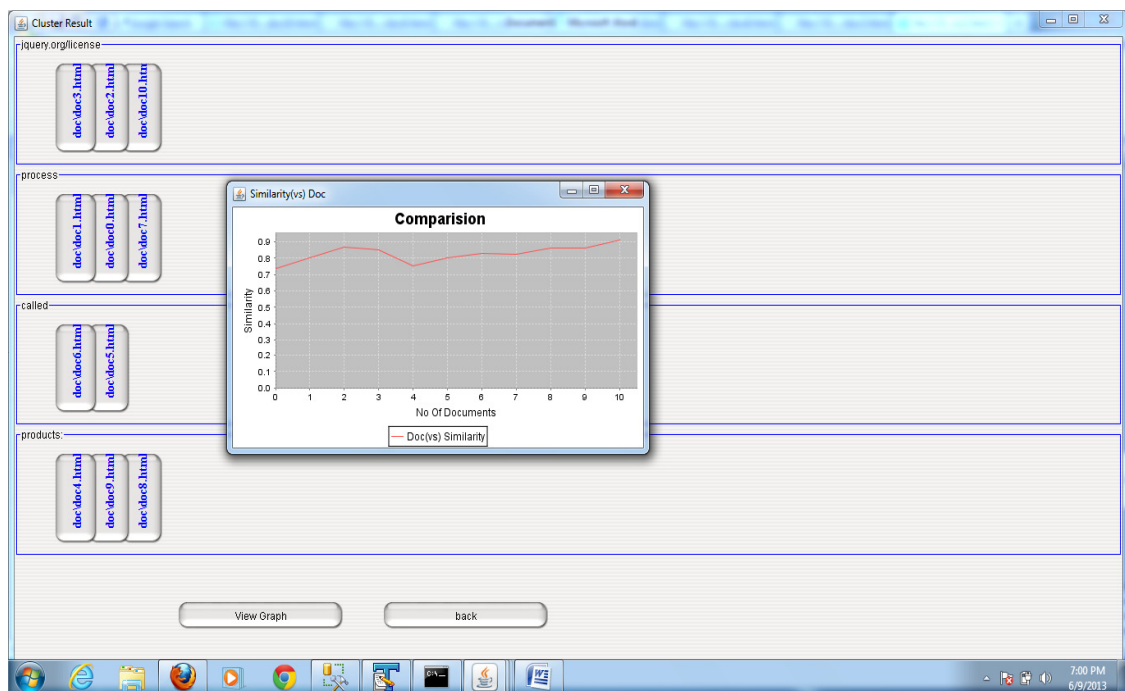


Figure 6 Implementation of Domain-Centric Search

5 CONCLUSION

In this research, the most difficult factor is data collection. Even though numerous Web pages have been composed during Google Scholar, it ought to be reframed to contest with the planned work. Ultimately, it has been concluded that this Tamil Community domain-centric search engine skilled with academic record and experienced with Google Scholar dataset is one of the techno-scientific development to the Tamil society.

REFERENCES

- 1) **Akilandeswari J., Gopalan N.P.**, 2008, "An Architectural Framework of a Crawler for Locating Deep Web Repositories using Learning Multi-agent Systems", Third International Conference on Internet and Web Applications and Services.
- 2) **Bar-Ilan J.**, 2005, "Comparing Rankings of Search Results on the Web", Information Processing and Management, Special Issue on Infometrics, Elsevier Ltd, Vol. 41, No. 6, pp. 1511-1519.
- 3) **Alfredo Teyseyre R., Marcelo Campo R.**, 2009, "An Overview of 3D Software Visualization", IEEE Transactions on Visualization and Computer Graphics, Vol. 15, No. 1, January-February.
- 4) **Feng Li**, 2008, "Extracting Structure of Web Site based on Hyperlink Analysis", IEEE International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1-4.
- 5) **Jayanthi S.K., Prema S.**, 2011, "CIMG-BSDS: Image Clustering based on BookShelf Data Structure in Web Search Engine Visualization", International Conference on Recent

Trends in Computing, Communication and Information Technologies, Chennai, Published in Springer Link Valley Series, ISSN:1865-0929 Part I, CCIS 269, pp. 457-466, December.

- 6) **Jayanthi S.K., Prema S.**, 2012, “Segregating Unique Service Object from Multi-Web Sources for Effective Visualization”, International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012), Periyar University, Salem-11, pp. 30-35, Published in IEEE Xplore Print ISBN: 978-1-4673-1037-6, 21-23 March.
- 7) **Jayanthi S.K., Prema S.**, 2012, “Improving Personalized Web Search using BookShelf Data Structure”, ICTACT Journal on Soft Computing, ISSN: 2229-6956(Online), ISSN: 0976-6561(Print), Vol. 3, No. 1, pp. 434-439, October.
- 8) **Jayanthi S.K., Prema S.**, 2012, “An NLP based Approach for Facilitating Efficient Web Search Results using BSDS”, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT 2012, Srinivasa Ramanujan Institute of Technology, Rotarypuram Village, Andhra Pradesh, India, Published in Elsevier and received the best paper award, 21-23 December.
- 9) **Jayanthi S.K., Prema S.**, 2013, “Measuring the Performance of Similarity Propagation in a Semantic Search Engine”, ICTACT Journal on Soft Computing, ISSN: 2229-6956(Online), ISSN: 0976-6561 (Print), Vol. 04, No. 1, pp. 667-672, October.
- 10) **Jayanthi S.K., Prema S.**, 2014, “Inculcating Efficient Web Search using BookShelf Data Structure”, 101th Indian Science Congress at Jammu University, Jammu, February 3-7.
- 11) **Jaime Teevan, Susan Dumais T., Eric Horvitz**, 2010, “Potential for Personalization”, ACM Transactions on Computer-Human Interaction special issue on Data Mining for Understanding User Needs, Vol. 17, No. 1, March.

Author Profile



Dr. S. Prema, currently working as an Assistant Professor and Head of Computer Applications in Arulmigu Arthanareeswarar Arts and Science College, Tiruchengode has received Ph.D., from the Bharathiar University in 2015. She has been involved in the teaching for the past 18 years. She secured the 1st Rank in B.Sc under Periyar University, Salem. She has totally 52 publications and one of her research paper entitled “An NLP based Approach for Facilitating Efficient Web Search Results using BSDS” received the best paper award. Her papers are cited at various publications (IEEE Xplore, Elsevier, Springer and International Conference Proceedings). She has h-index value: 6, i-10 index: 4, Citations: 173 and her profile is listed in Marquis Who is Who in World, International Biography Center, London, UK, 2011. She has produced 12 M.Phil and 4 Ph.D Scholars for doing their research.

**கணினி மொழிபெயர்ப்பில் முன்னேற்றங்கள்: தமிழ் - ஆங்கில இயந்திர
மொழிபெயர்ப்பு நோக்கு
(Advancements in Computational Linguistics : A Focus on Tamil - English Machine Translation)**

முனைவர் மா. வசந்தகுமாரி¹, < ilavenil2010tamil@gmail.com >

முனைவர் செ. பிரேமா², < premashanmuga11@gmail.com > &

முனைவர் செ. கனிமொழி சுகுணா³, < dr.kanizozhisuguna@gmail.com >

¹ உதவிப்பேராசியர் மற்றும் தலைவர், தமிழ்த்துறை,

² உதவிப்பேராசியர் மற்றும் தலைவர், கணினிப் பயன்பாட்டியல் துறை.

³ உதவிப்பேராசியர், கணினிப் பயன்பாட்டியல் துறை.

அருள்மிகு அர்த்தநாரீசுவரர் கலை மற்றும் அறிவியல் கல்லூரி, திருச்செங்கோடு.

நாமக்கல் மாவட்டம் - 657 201. தமிழ்நாடு, இந்தியா.

ஆய்வுச்சுருக்கம்

இந்த ஆராய்ச்சிக் கட்டுரையில் இயந்திர மொழிபெயர்ப்பு என்பது கணினியைப் பயன்படுத்தி ஒரு இயற்கை மொழியின் பனுவலை மற்றொரு மொழிக்குத் தானாக மாற்றுவதைக் குறிக்கும். இந்த ஆராய்ச்சி தமிழ்க் கணினி மொழித் தொகுப்புகளின் முன்னேற்றங்கள் பற்றியும் ஆங்கிலம் - தமிழ் இயந்திர மொழிபெயர்ப்பு முறைகளின் வளர்ச்சியைப் பற்றியும் குறிப்பிடுகின்றது. மேலும் தமிழ் எழுத்துமொழிக் குறிச்-சொற்களின் தொகுப்பு, எழுத்துமொழியினை அறிந்து கொள்ளும் செயல்முறைகள், இயற்கை மொழி உருவாக்கம் போன்ற தமிழ் மொழியின் செயல்பாட்டுக் களத்தின் அடிப்படை ஆதாரங்களை வெளிப்படுத்துகின்றது. புள்ளிவிவர இயந்திர மொழி-பெயர்ப்பு அமைப்பு - Surface Mount Technology) மொழிபெயர்ப்புச் சிக்கலை இயந்திர கற்றலின் சிக்கலாகக் கருதுகிறது. கற்றல் வழிமுறைகள் மொழியின் இணையான வாக்கியங்களில் இருந்து ஒரு மாதிரியை உருவாக்கி இம்மாதிரியை பயன்படுத்திப் புதிய வாக்கியங்கள் மொழிபெயர்க்கப் படுகின்றன. இவ்வகையினைக் கணினிக் கற்றுக்கொள்ள ஒரு பொருளைப் பற்றிய முழுமையான இலக்கியத் தரவுத் தொகுப்பு தேவை.

தமிழ் - ஆங்கிலம் போன்ற இரு வேறுபட்ட மொழிகளுக்கு SMT (Surface Mount Technology) நிரல் நெறிமுறைகளை நேரடியாகப் பயன்படுத்துவது கடினம். இதனைத் தீர்க்கவும் மொழிபெயர்ப்புச் செயல்திறனை மேம்படுத்தவும் கூடுதல் உருவவியல்(நா0௦௭௮00௦௦) முன்செயலாக்கம் தேவைப்படுகிறது. இந்த முன் செயலாக்கம் தமிழ் மொழியியல் கருவிகளைப் பயன்படுத்திச் செய்யப்படுகிறது. இது மொழியியல் தரவுகளை ஒரு பொருளைப் பற்றிய முழுமையான இலக்கியத் தொகுப்பு) உயர்த்த விரும்புகின்றது. இந்த ஆராய்ச்சிக் கட்டுரை முந்தைய மொழிபெயர்ப்பு வரலாற்றினைக் கூறுவதுடன் தற்போதுள்ள மொழிபெயர்ப்புப் பரிசோதனைகளும் உள்ளீடு செய்யப்பட்ட நிலையில் பிற மொழிகளுக்கான மொழிபெயர்ப்பு முன்னோடியாகத் திகழும்.

திறவுச் சொற்கள்

இயந்திர மொழிபெயர்ப்பு, பனுவல், எழுத்துமொழி, குறிச்சொற்கள், இயற்கை மொழி, இலக்கியத் தரவு, உருவவியல், மொழியியல் தரவு.

முன்னுரை

தொன்மை வாய்ந்த மற்றும் செம்மையான தமிழ்மொழியானது தென்னிந்தியப் பகுதிகள், இலங்கை பகுதிகளில் அதிகளவிலும் அமெரிக்கா, இலண்டன், கனடா, மலேசியா, சிங்கப்பூர் உள்ளிட்ட பல்வேறு மேலை நாடுகளில் குறிப்பிடத்தக்க எண்ணிக்கையிலும் பேசப்பட்டு வருகின்றது. உலகில் தமிழும் ஆங்கிலமும் முக்கியத்துவம் வாய்ந்த மொழிகளாகும். இன்றைய உலகமயமாக்கலில் பயனளிக்கும் வகையிலான மொழிபெயர்ப்பின் தேவையானது அதிகரித்துள்ளது. தமிழ்மொழித் தமிழரின் பண்பாடு மற்றும் வரலாற்றுச் செல்வங்களைத் தன்னகத்தே கொண்டுள்ளது. உலகளாவிய அளவில் தமிழர்களின் செல்வாக்கு அதிகரித்து வரும்

இன்றைய சூழலில், தமிழில் இருந்து ஆங்கில மொழிபெயர்ப்பிற்கான தேவை அதிகரித்துள்ளது. இந்த ஆய்வில் தமிழை ஆங்கிலத்திற்கு மொழிபெயர்ப்பதில் எதிர்கொள்ளும் சவால்களைப் பற்றி ஆராய்வதுடன், பயனுள்ள மொழி-பெயர்ப்பிற்கான குறிப்புகள் மற்றும் வழிமுறைகள் கூறப்பட்டுள்ளன. பொதுப் பயன்பாட்டில் அமையப்பெற்றுள்ள சொற்றொடர்கள் மற்றும் வெளிப்பாடுகளைப் பற்றி எடுத்துரைக்கப்படும். மொழிபெயர்ப்புக் கருவிகள் மற்றும் ஆதாரங்கள். முதன்மையாக்கப்படும். முதன்மை இலக்கணமும் மொழியியல் வெளிப்பாடும் நுட்பமான மொழிபெயர்ப்பிற்குத் தமிழ் மற்றும் ஆங்கில மொழிகளின் இலக்கணம் மற்றும் மொழியியல் வெளிப்பாடுகள் பற்றிய முழுமையான புலமைத் தேவைப்படுகிறது. இருமொழி இலக்கணத்தின் நுணுக்கங்களை அறிந்து மொழியியல் வெளிப்பாடுகளை மொழிப்பெயர்ப்புப் பயிற்சி செய்தல் அவசியம்.

தமிழ் - ஆங்கில மொழிபெயர்ப்பிற்கான உத்திகள்

தமிழிலிருந்து ஆங்கிலத்திற்கு மொழிபெயர்ப்பதற்கு முறையான அணுகுமுறை விவரங்கள் பெற்றிருப்பது தேவையாகிறது. மொழிபெயர்ப்புகளின் தரத்தையும் நுட்பத்தையும் தரப்படுத்த சில குறிப்புகள் பின்வருமாறு

செம்மைமிகு சொற்களஞ்சிய உருவாக்கம்

தரமான மொழிபெயர்ப்பிற்குத் தமிழ் மற்றும் ஆங்கிலம் இரண்டிலும் செம்மைமிகு சொற்களஞ்சியங்கள் உருவாக்குவது தேவையாகிறது. மொழியின் சொல் வங்கியை விரிவாக்கப் பல்வேறு தமிழ் அகராதிகள், மொழிக் கற்றல் பயன்பாடுகள் மற்றும் இலக்கியங்களை ஆராய்வது போன்றவை செம்மையான சொற்களஞ்சிய உருவாக்கப் பணிக்கு வழிவகை செய்கிறது.

சமூகச் சூழலைப் புரிந்து கொள்ளல்

மொழிபெயர்ப்பில் சமூகச் சூழல் முக்கியமானதாகக் கருதப்படுகிறது. ஆங்கிலப் பிரதி ஒன்றில் இடம்பெற்றுள்ள உண்மைப் பொருளைக் கண்டறிய அதிலுள்ள உரையின் பண்பாடு, வரலாறு மற்றும் சமூகச் சூழலுடன் பழக்கிக் கொள்ளுதல் வேண்டும்.

மொழிபெயர்ப்புக் கருவிகளைப் பயன்படுத்தல்

இணைய அகராதிகள், இயந்திர மொழிபெயர்ப்பு மென்பொருள் மற்றும் ஊயுவு (கணினி. உதவி மொழிபெயர்ப்பு) கருவிகள் போன்ற மொழிபெயர்ப்புக் கருவிகளைப் பயன்படுத்துவது தேவை. இந்தக் கருவிகள் குறிப்புகளை விரைவாக வழங்குவதுடன் முழுமையான மொழி-பெயர்ப்பினைப் பின்பற்ற பயன்படும்.

மெய்ப்புத் திருத்தம்

பிரதியினை மொழிபெயர்த்த பின்னர் அதிலுள்ள பிழைகள் மற்றும் முரண்பாடுகளை நீக்குவதற்கு நுணுக்கமான இறுதி மொழிபெயர்ப்பை உறுதி-செய்வதற்கு இலக்கணம், நிறுத்தற்குறிகள் மற்றும் தொடரமைப்பு ஆகியவற்றினைக் கவனிக்க மெய்ப்புத் திருத்தம் மேற்கொள்ளல் தேவையாகிறது. இவற்றினால் தமிழிலிருந்து ஆங்கில மொழிபெயர்ப்புகளின் தரத்தையும் நுட்பத்தையும் பயனுள்ளதாக உயர்த்த வழிவகை செய்யமுடிகிறது.

பொதுப் பயன்பாட்டில் அமையப்பெற்றுள்ள தமிழ் ஆங்கில் சொற்றொடர்களின் மொழிபெயர்ப்பு வெளிப்பாட்டின் நடைமுறையினை ஆழ்ந்து பார்க்கும்பொழுது பொதுவாகப் பயன்படுத்தப்படும் தமிழ் சொற்றொடர்கள் மற்றும் அதனின் வெளிப்பாட்டுத் திறன்களை அறிந்துகொள்வது அவசியமாகிறது. அதுகுறித்த எடுத்துக்காட்டுகளாக,

கடவுள் வாழ்த்து – God bless you
வாழ்க வளமுடன் - Live Prosperously
காலை வணக்கம் - Good Morning
நான் பேசுகிறேன் - I Speak
நீ என்ன செய்கிறாய் - What are you doing
வாருங்கள் - Come on
போய் வாருங்கள் - Come and Go
நன்றி - Thanks

மேற்கண்ட சொற்றொடர்கள் போன்று அமையப்பெற்றுள்ள பொதுவான சொற்றொடர்களுடன் மொழிபொப்பாளர் தன்னை பழக்கப்படுத்திக் கொள்வதன் மூலம் தமிழிலிருந்து ஆங்கில் மொழிபெயர்ப்புப் பணிக்குத் தன்னை இணைத்துக் கொள்வதற்கான அடித்தளமாக இருக்கும்.

தமிழ் - ஆங்கில் மொழிபெயர்ப்புக் கருவிகள் மற்றும் வளங்கள்

இன்றைய எண்ணிமனனுபவையட) உலகில் தமிழிலிருந்து ஆங்கிலத்தில் மொழிபெயர்க்க ஏராளமான கருவிகள் மற்றும் ஆதாரங்கள் இருக்கின்றன. அவற்றுள் சிறப்பு வாய்ந்தவைகளாக,

இணையக் தமிழ்-ஆங்கில அகராதிகள், தமிழ் லெக்சிகன் போன்ற இணைய தளங்கள் சொற்களுக்கான பொருள், பொருத்தமான பிறசொற்கள் மற்றும் பயன்பாட்டு எடுத்துக்-காட்டுகளை வழங்கும் விரிவானதொரு அகராதிகளை வழங்குகின்றன. இயந்திர மொழிபெயர்ப்பு மென்பொருள்கள்

Google Translate மற்றும் Machine Translate போன்ற இணைய இயங்குதளங்கள் விரைவான மொழிபெயர்ப்புகளுக்கு உதவும் வகையிலான இயந்திர மொழி-பெயர்ப்புச் சேவைகளை வழங்குகின்றன. எனினும் இயந்திர மொழிபெயர்ப்புகள் நுட்பமான மொழிபெயர்ப்பு அமைப்பினைக் கொண்டிருக்கவில்லை என்பதைக் கவனிக்க வேண்டியுள்ளது. எனவே அவற்றை மொழிபெயர்ப்பின் இறுதி வடிவத்திற்குப் பயன்படுத்திக் கொள்ளாமல் மேலோட்டமாக அதனைப் பயன்படுத்தல் வேண்டும்.

கணினி உதவி மொழிபெயர்ப்புக் கருவிகள் (CAT - Assisted Translation)

SDI, Trados மற்றும் MemoQ போன்ற CAT கருவிகள் தொழில்முறை மொழி-பெயர்ப்பாளர்களால் பெரும்பான்மை பயன்படுத்தப்படுகின்றன. இந்தக் கருவிகள் ஒரு தரவுத்தளத்தில் மொழிபெயர்ப்புகளைச் சேமித்திருப்பதன் உற்பத்தித் திறன் மேம்படுத்தப்பட்ட நிலையில் பல திட்டங்களில் நிலைத்த சொற்களின் பயன்பாடு அனுமதிக்கப்படுகின்றன.

தமிழ் - ஆங்கில மொழிபெயர்ப்புக் கருவிகள் மற்றும் வளங்கள்

இன்றைய எண்ணிம(னுபவையட) உலகில் தமிழிலிருந்து ஆங்கிலத்தில் மொழி-பெயர்க்க ஏராளமான கருவிகள் மற்றும் ஆதாரங்கள் இருக்கின்றன. அவற்றுள் சிறப்பு வாய்ந்தவைகளாக, இணையக் தமிழ்-ஆங்கில அகராதிகள், தமிழ் லெக்சிகன் போன்ற இணைய தளங்கள் சொற்களுக்கான பொருள், பொருத்தமான பிறசொற்கள் மற்றும் பயன்பாட்டு எடுத்துக்-காட்டுகளை வழங்கும் விரிவானதொரு அகராதிகளை வழங்குகின்றன.

இயந்திர மொழிபெயர்ப்பு மென்பொருள்கள்

Google Translate மற்றும் Machine Translate போன்ற இணைய இயங்குதளங்கள் விரைவான மொழிபெயர்ப்புகளுக்கு உதவும் வகையிலான இயந்திர மொழி-பெயர்ப்புச் சேவைகளை வழங்குகின்றன. எனினும் இயந்திர மொழிபெயர்ப்புகள் நுட்பமான மொழிபெயர்ப்பு அமைப்பினைக் கொண்டிருக்கவில்லை என்பதைக் கவனிக்க வேண்டியுள்ளது. எனவே அவற்றை மொழிபெயர்ப்பின் இறுதி வடிவத்திற்குப் பயன்படுத்திக் கொள்ளாமல் மேலோட்டமாக அதனைப் பயன்படுத்தல் வேண்டும்.

கணினி உதவி மொழிபெயர்ப்புக் கருவிகள் (CAT - Computer Assisted Translation)

SDI, Trados மற்றும் MemoQ போன்ற CAT கருவிகள் தொழில்முறை மொழி-பெயர்ப்பாளர்களால் பெரும்பான்மை பயன்படுத்தப்படுகின்றன. இந்தக் கருவிகள் ஒரு தரவுத்தளத்தில் மொழிபெயர்ப்புகளைச் சேமித்திருப்பதன் உற்பத்தித் திறன் மேம்படுத்தப்பட்ட நிலையில் பல திட்டங்களில் நிலைத்த சொற்களின் பயன்பாடு அனுமதிக்கப்படுகின்றன.

தமிழ் - ஆங்கில மொழிபெயர்ப்புச் சவால்கள்

தமிழிலிருந்து ஆங்கிலத்திற்கு மொழிபெயர்ப்பது என்பது மொழிக் கட்டமைப்பு, இலக்கண மற்றும் பண்பாட்டு வேறுபாடுகள் காரணமாகத் தனித்துவமான சவால்களை முன்வைக்கிறது. தமிழ்மொழி சூழலுக்கேற்ப தன்னைத் தகவமைத்துக் கொள்ளும் சிறப்புடையது. அதனால் தமிழ்மொழியானது ஒரு தகவல் திரட்டியாக இருக்கும் நிலையில் வேர்ச்சொற்கள் மற்றும் பின்னொட்டுகளை இணைத்து வார்த்தைகளை உருவாக்குகிறது. அதேநேரத்தில் ஆங்கிலம் நிலையான சொல்வரிசை மற்றும் முன்மொழிவுகளை நம்பியுள்ளது. இந்தக் கட்டமைப்பு இடைவெளிதான் பொருத்தமான பொருளினைக் கண்டுபிடிப்பதற்குச் சிக்கலை உண்டாக்கும். தமிழின் பெயாச்சொல் வழக்குகள் மற்றும் வினைச்சொற்கள் மொழிபெயர்ப்பில் சிக்கலைச் சோக்கின்றன. பண்பாடு மற்றும் கழல் வேறுபாடுகள் இருமொழிகளின் ஆழமான புரிதலைக் கேட்கின்றன.

ஒருசில தமிழ்மொழித் தொடர்புடைய கருத்துகளுக்குப் பொருத்தமானவை ஆங்கிலத்தில் காணப்படவில்லை. உண்மைக் கருத்தின் பாதுகாப்பை எதிர்கொள்ளும் திறன் முன்வைக்கப் படுகின்றன. இந்தச் சவால்களை எதிர்கொள்வதற்கு மொழிபெயர்ப்பாளர்கள் தமிழ் மற்றும் ஆங்கிலம் ஆகிய இரண்டு மொழிகளிலும் புலமைப் பெற்றிருப்பது தேவையாகிறது. இரு மொழிகளின் பண்பாடு மற்றும் சமூகச் கழல் நுணுக்கங்களைப் பற்றிய ஆழ்நோக்குப் புரிதல் அவசியமாகிறது. மேலும் தரப்படுத்தப்பட்ட மொழிபெயர்ப்புக் கருவிகள் மற்றும் மூலங்களைப் பயன்படுத்துவது என்பது மொழிபெயர்ப்புகளில் அதிகளவு உதவி செய்கிறது. அதுமட்டுமல்லாமல் மொழிபெயர்ப்பின் நுட்பம் மற்றும் செயல்திறனை உறுதிசெய்வதற்கும் உதவுகிறது.

முடிவுகள்

தமிழ் - ஆங்கில இயந்திர மொழிபெயாப்பில் அந்தந்த மொழிகளின் கணினி மொழியியல் மேம்பாடுகளும் மொழியாக்கங்களும் பண்பாட்டு வேறுபாடுகளைக் கண்டறிய உதவுகின்றன இன்றைய தொழில்நுட்ப வளர்ச்சியால் தமிழ் - ஆங்கில இயந்திர மொழியாக்கம் பல்வேறு முன்னேற்றங்கள் அடையப்பெற்றுள்ளன. என்றாலும் முழுமைபெற்ற நுட்பமான மொழிபெயாப்பினை எட்டவில்லை என்பது காணத்தக்கது.

மொழிபெயர்ப்பின் தரத்தினை உயர்த்துவதற்கு அன்றாட தொழில்நுட்பப் பயனாளிகள், மொழியியல் அறிஞர்கள், கல்வியாளர்கள், ஆகியோரின் அறநெறிக் கோட்பாட்டின் வழி மொழிபெயர்ப்பு நடைமுறைகளை ஒஉறுதிப்படுத்துவது தேவையாகிறது.

துணை நூல்கள்

1. Fredric CC. Gey, Prospects of Machine' Translation of the Tamil Language, <https://www.researchgate.et/publication/228610428>, January 2002.
2. Santosh Kumar T.S., Word Sense Disambiguation Using Semantic Web for Tamil to English Statistical Machine Translation, IRA-International Journal of Technology & Engineering, 2016.

Aspect-Based Sentiment Analysis of Movie Reviews in Tamil- A Study On The Effectiveness of the BERT with MADTRAS Dataset

M. Arunmozhi¹ arunmozhi.me@pondiuni.ac.in, E. Syam Mohan¹
syammohane@pondiuni.ac.in, Dr. R. Sunitha¹ rsunitha@pondiuni.ac.in,
 Dr. V. Dhanalakshmi² dhanagiri@pondiuni.ac.in & Dr. K. Pajanivelou²
 <kpajanivelou@pondiuni.ac.in>

¹ Dept of Computer Science, ² Subramania Bharathi School of Tamil Language and Literature, Pondicherry University, Puducherry, India

Natural Language Processing (NLP) is emerging in diverse fields such as information retrieval, machine translation, sentiment analysis, and text summarization, due to the exponential growth of textual data on the Internet. The models and algorithms of NLP utilize a combination of linguistic models, statistical models, Machine Learning (ML) and Deep Learning (DL) models to process and analyze natural language data.

Sentiment analysis is a popular application of NLP, where the goal is to identify the sentiment or the emotional tone that is communicated in texts like social media posts, reviews, or comments and to classify the sentiments into any one of the sentiment types viz. positive, negative, or neutral.

Sentiment analysis is computationally carried out at three granularities of content viz. document, sentence and aspect. Document-level sentiment analysis involves determining the sentiment that is expressed in the entire content which might be a review or an article. This approach provides a high-level understanding of the overall sentiment towards the subject matter and is useful for tasks like analyzing product reviews, movie critiques, or feedback surveys. Sentence-level sentiment analysis concentrates on examining the sentiment articulated within each sentence of a document, assigning a sentiment polarity to each sentence independently. This fine-grained analysis allows for a more detailed examination of sentiment within the text. It enables organizations to understand customer sentiment, track public opinion, make data-driven decisions, and improve service quality across various domains. Aspect Based Sentiment Analysis (ABSA) proposes to analyze the sentiments that are stated towards various aspects present in a document, for example, opinions expressed about the different features of a still camera in a review. Compared to the analysis of English texts, analyzing sentiments in texts expressed in low-resource languages like regional languages of India can be challenging because of the inherent complexities of the language like inflexion, ambiguity, pragmatics and thrust complexities like dialectal variations, cultural references, and colloquial expressions.

With the rise of social media, people express their opinions and emotions in texts through their native language more frequently and publicly than ever before. These opinions or sentiments are computationally used in various applications like targeted advertisements, recommendation systems, community detection etc. Also, a lot of text in one's language expressing opinions on a movie or a review on the different aspects of a movie are shared on

social media. These reviews offer insights into the quality, themes, overall reflection, etc., about a movie, saving viewers' time and money by steering them towards experiences they're likely to enjoy. Thus understanding the sentiments expressed in the reviews on the whole becomes necessary to contribute to the critical discourse surrounding cinema.

ABSA focuses on every aspect of movies which serves as a guide for audiences, helping them to decide the films to watch based on their preferences and interests.

The study of the literature reveals that very few works have been done on Tamil movie review sentiment analysis at the document and sentence level. These works have primarily focused on utilizing methods such as rule-based approaches, ML techniques, and DL architectures. Rule-based methods often leverage lexicons and linguistic patterns to classify sentiment in Tamil text, while ML techniques employ features extracted from the texts for classifying the sentiment. DL models use representation learning that learns the feature automatically from the data. DL approaches offer advantages such as scalability, flexibility, and, allowing them to capture complex patterns and relationships from the data.

As DL models are proven for efficiently handling textual data, this work proposes the study of the effectiveness of various DL models for ABSA of Movie reviews in the Indian regional language Tamil. The two challenges in ABSA of movie reviews in Tamil are the non-availability of a well-annotated movie review dataset in Tamil and the difficulty in handling the inherent complexities of the Tamil language. In this work, we have curated and annotated an aspect-based movie review dataset MADTRAS (Dataset for Aspect-based Sentiment Analysis of Movie Reviews in Tamil) for training and evaluating the DL models. The dataset MADTRAS consists of 1760 reviews out of which 880 are positive and 880 reviews are negative. This dataset has been curated from YouTube, Twitter and Google Forms. The reviews provide insights into aspects such as Direction, Acting, Songs/Music, Screenplay, Story, Background music and Editing. These seven target aspects are short-listed from a set of most frequently discussed aspects of Tamil movies. The average length of the reviews in the dataset is four sentences.

We have employed models like Long Short-Term Memory (LSTM), Bi-Directional LSTM, Transformers, Gated Recurrent Unit (GRU), Bi-Directional GRU and Bi-directional Encoder Representations from Transformers (BERT) to perform ABSA of the Movie Reviews in Tamil. The main objective is to find the suitability of the BERT in efficiently classifying the positive and negative sentiments expressed in Tamil movie reviews compared to the aforementioned models. The GRU architecture offers faster training and a simpler structure but struggles with long-term dependencies and capturing complex relationships. LSTM, on the other hand, is more computationally expensive but excels in remembering long sequences and mitigating vanishing gradients, although it suffers from overfitting. Bi-LSTM combines the strengths of LSTM with bidirectional processing, capturing contextual information from both past and future words, making it suitable for tasks where bidirectional context is crucial. However, Bi-LSTM requires more parameters and computational resources, resulting in longer training times and potential overfitting with limited data. BERT is a pre-trained model and its transfer learning capabilities enable adaptation to specific sentiment analysis tasks

even with relatively small annotated datasets, thereby enhancing its utility in sentiment analysis applications.

The experiments carried out with the different DL models over the MADTRAS dataset show that the BERT model outperforms all the other DL models with respect to the classification accuracy for all aspects. This is because of the BERT's ability to handle the Tamil text being a pre-trained model. As Bi-GRU can learn efficiently with smaller datasets, it ranks next. Among all these models the performance of LSTM and its variants are less comparatively. Hence, the study shows that BERT is well suited for Aspect Based Sentiment Analysis of Tamil Text.

**கூகுள் லென்ஸ்: காட்சித்தேடலும் படச் சொற்களை மொழிபெயர்ப்பதில்
ஏற்படும் சிக்கல்களும்**

முனைவர் சு. பிரபாவதி,
உதவிப்பேராசிரியர் (தமிழ்),
அருள்மிகு கபாலீசுவரர் கலை மற்றும் அறிவியல் கல்லூரி,
கொளத்தூர், சென்னை-99

சமூக வளர்ச்சி மாற்றங்களுக்கு ஏற்ப மனிதனின் தேவைகள் பெருகுகின்றன. தேவைகளுக்கு ஏற்ப நவீன தொழில்நுட்பம் வளர்ச்சியை நோக்கிச் சென்று கொண்டிருக்கின்றது. தொடர்பு ஊடகங்களின் வருகை எங்கோ ஒரு மூலையில் இருப்பவரை நம்முடன் தொடர்புகொள்ளச் செய்கிறது. அதே ஊடகம் தான் நம் அருகில் இருப்பவரை நம்மிடமிருந்து அந்நியப் படுத்துகிறது. ஒரு குடும்பத்தில் இருக்கும் ஒவ்வொருவரிடமும் தனித்தனி அலைபேசிகள் இருக்கும் நிலையில் அக்குடும்பத்தில் உள்ள அனைவரும் ஒருவருக்கொருவர் அந்நியராக ஒரே வீட்டில் வாழ்கின்ற சாத்தியங்களை இன்றைய தொழில்நுட்பம் நம்மிடையே உருவாக்கியுள்ளது. இதரில் இருப்பவரைப் பார்க்காமல் சமூக ஊடகங்களிலும் விளையாட்டுச் செயலிகளிலும் இன்றைய இளம் தலைமுறை மூழ்கிக் கிடப்பது வருத்தத்திற்குரியது.

கூகுள் குரோம், பயர்பாக்ஸ், ஓப்ரா, மைக்ரோசாப்ட் பிங் முதலான இணைய தேடு பொறிகளின் வாயிலாக நமக்குத் தேவையான அனைத்தையும் உள்ளீடு செய்து தேடிக் கண்டடைய முடியும். இவற்றுள் கூகுள் நிறுவனம் நவீனத் தேவைகளுக்கும் சூழல்களுக்கும் ஏற்பத் தனது தனது தேடுபொறியின் செயல்பாட்டினை மேம்படுத்திக் கொண்டு வருகிறது. அந்த வகையில், எழுதப் படிக்கத் தெரிந்தவர்களால் மட்டுமே உள்ளீடு செய்து பயன்படுத்த முடிந்த தேடுபொறியில் எழுதப் படிக்கத் தெரியாமல் ஒரு மொழியைப் பேசத் தெரிந்திருந்தால் மட்டுமே போதும் அம்மொழியில் நமக்கான தேவைகளைக் கேட்டுப் பெறமுடியும் என்னும் சாத்தியங்களை 'குரல் தேடல்' வழியாக சாத்தியப்படுத்தியது. அதற்கு அடுத்த நிலையில் ஒருவருக்கு எழுதவோ, படிக்கவோ தெரியாது; அந்த மொழியைப் பேசவும் தெரியாது; ஆனால் தான் ஒன்றைத் தெரிந்துகொள்ள வேண்டும் என்றால் தன் தேவையைப் படம் பிடித்து அப்படத்தினை உள்ளீடு செய்தாலே போதுமானது. நாம் உள்ளீடு செய்த படத்திற்குப் பொருத்தமான தகவல்கள் உடனே நமக்குக் கிடைத்துவிடுகின்றன. இதனை முதலில் கூகுள் லென்ஸ் என்ற தனி செயலியாக உருவாக்கிய கூகுள் நிறுவனம், அதை கூகுள் தேடுபொறியில் இணைத்திருப்பதன் மூலம் எளிமையான பயன்பாட்டுச் சூழலை உருவாக்கித் தந்திருக்கிறது.

ஓசியூர் தொழில்நுட்பம்

ஐஓஎஸ் மற்றும் ஆண்ட்ராய்டுகளுக்கான கூகுள் ஆப்ஸ், கூகுள் படங்கள் மற்றும் கூகுள் கேமரா போன்ற பல்வேறு சேவைகளிலும் இயங்குதளங்களிலும் கூகுள் பயன்படுத்தப்படுகிறது. இது ஆப்டிகல் கேரக்டர் ரெகக்னிஷன் (OCR) மூலம் செயல்படுகிறது. கூகுள் லென்ஸின் படங்களிலிருந்து உரையாக மாற்றுவது இந்த OCR அடிப்படையிலேயே நிகழ்கிறது. OCR என்பது படங்களாக இருக்கும் கையால்

எழுதப்பட்ட உரைகள், அச்சடிக்கப்பட்ட எழுத்துக்களை அடையாளம் கண்டு திருத்தக்கூடிய உரையாக மாற்றும் தொழில்நுட்பமாகும்.

கூகுள் லென்ஸ் செயல்பாடுகள்

பூதக் கண்ணாடி என்று தமிழில் நாம் குறிப்பிடுவதையே லென்ஸ் என்ற பெயரில் பயன்படுத்துகிறது. கண்ணுக்குத் தெரியாமல் இருக்கும் மிகச்சிறிய எழுத்துக்களையோ, பொருட்களையோ பூதக்கண்ணாடியின் துணையுடன் கண்டுபிடித்துவிட முடியும். அதுபோல் இந்த லென்ஸ் நமது தேவைகள் அனைத்தையும் கண்டறிய உதவும் என்பதைக் குறிக்க கூகுள் நிறுவனம் கூகுள் லென்ஸ் என்னும் சொல்லைப் பயன்படுத்தியுள்ளது. இதன் பயன்பாட்டினை, Search What you see, Shop What you see, Translate What you see, Identify What you see என்னும் அறிமுகத்தோடு வார்த்தைகளால் விவரிக்க முடியவில்லையா? உங்களுடைய கேமரா அல்லது படங்களின் மூலமாகத் தேடுங்கள் என்னும் முகப்புப் பட்டையோடு இந்தச் செயலி செயல்படுகிறது.

- ஒரு படத்தை கூகுள் லென்ஸின் தேடு பாதையில் பொருத்தி அப்படத்தில் உள்ள சொற்களை அப்படியே உங்களுக்குத் தேவையான மொழியில் மொழிபெயர்க்கலாம் (image to translate),
- படமாக உள்ள சொற்களைத் தனித்தனிச் சொற்களாக மாற்றலாம். (image to text),
- படங்களைக் கொண்டு அதே போன்ற படங்கள் குறித்த தகவல்களை அறிய முடியும்
- கல்வி நிலையங்களில் கொடுக்கப்படும் வீட்டுப் பாடங்களைக் கேள்விகளாகப் படம் எடுத்துத் தேடினால் அதற்கான விடைகளை அறிந்துகொள்ள முடியும்.
- ஒரு பொருளை வாங்க விரும்பினால் அதற்கான படத்தை கூகுள் லென்ஸில் பதிவேற்றி அதற்கான விலை வேறுபாடுகளை அறிந்து இணைய வழியாக அப்பொருளை குறைந்த விலையில் வாங்க முடியும்.
- ஒரு இடத்தைப் படம் எடுத்துத் தேடும் பொழுது அவ்விடம் குறித்த அனைத்துத் தகவல்களையும் நம்மால் அறிந்துகொள்ள முடியும்.

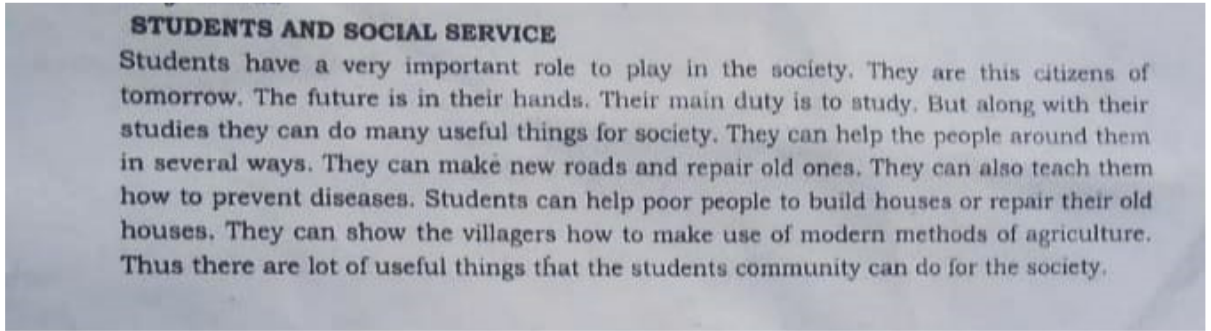
மேற்குறிப்பிட்ட பயன்பாடுகள் யாவும் கூகுள் லென்ஸ் வழியாக நாம் பெறக்கூடியவை. இவற்றில் படங்களில் உள்ள சொற்களை தேர்ந்தெடுக்கக்கூடிய சொற்களாக (text) மாற்றுவது மற்றும் படச்சொற்கள் எவ்வாறு மொழிபெயர்க்கப்படுகின்றன என்ற கேள்விக்கு விடைதேடும் வகையில் இக்கட்டுரை அமைகிறது. உதாரணமாக, ஒரு மருத்துவ அறிக்கையை நீங்கள் படம்பிடித்து அதில் என்ன கூறப்பட்டுள்ளது என்று அறிய விரும்பினால் அதனை கூகுள் லென்ஸ் வழியாக எளிமையாக அறிந்துகொள்ள முடியும். ஒருதாளில் கையெழுத்தில் இருந்தாலும், அச்செழுத்தில் இருந்தாலும் அதே தாளிலே நமக்குத் தேவையான மொழியில் மாற்றிப் படித்துப் புரிந்துகொள்ளமுடியும். மேலும், இதனை நேரடியாக கூகிள் மொழிபெயர்ப்புக்கும் மாற்றிப் பார்க்கவும் முடியும். நவீன தொழில்நுட்ப வளர்ச்சி படச்சொற்களையும் (text image) மொழிபெயர்க்கும் சாத்தியங்களை ஏற்படுத்தியுள்ளது.

படச்சொற்களை மொழிபெயர்த்தல்

கூகுள் டிரான்ஸ்லேட்டர் ஏப்ரல் 2006ஆம் ஆண்டு அறிமுகப் படுத்தப்பட்டபோது சொற்களை அப்படியே நமக்குத் தேவையான மொழியில் நேரடியாக

மொழிபெயர்ப்பு செய்து தந்தது. மேலும், ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு நேரடியாக மொழிபெயர்க்காமல் முதலில் ஆங்கிலத்திற்கும் பின்னர் நமது இலக்கு மொழிக்கும் மொழிபெயர்த்தது. இவ்வாறு மொழிபெயர்க்கப்படும்போது பல சிக்கல் ஏற்பட்டன. ஒவ்வொரு மொழியின் அமைப்பும் வேறுபட்ட பல தன்மைகளைக் கொண்டிருக்கும் போது அவற்றை அத்தன்மை மாறாமல் மொழிபெயர்ப்பது என்பது ஒரு சவாலான காரியமாக இருந்தது. 2016 நவம்பரில் கூகுள் டிரான்ஸ்லேட் நரம்பியல் இயந்திர மொழிபெயர்ப்புக்கு மாறுவதாக அறிவித்தது. இது கூகுள் நியூரல் மெஷின் டிரான்ஸ்லேஷன் (GNMT) என்ற பெயரில் ஒரு முழு வாக்கியத்தைத் துண்டு துண்டாக மொழிபெயர்க்காமல் ஒரே நேரத்தில் முழுவதுமாக மொழிபெயர்ப்பு செய்தது. இதிலும் துல்லியத் தன்மை என்பது குறைவு. காரணம் ஒரு மொழி அதன் பண்பாட்டுச் சூழலில் எவ்வாறெல்லாம் பொருள் கொள்ளப்படுகிறது என்பதை அறிந்தால் மட்டுமே துல்லியமான மொழிபெயர்ப்பினைச் செய்ய முடியும். எனினும் கூகுள் லென்ஸ் என்பது இல்லாத ஊருக்கு இலுப்பைப் பூ சர்க்கரை என்பதைப் போல மொழி தெரியாத ஊருக்குச் சென்றாலும் நம்மால் அடிப்படைத் தேவைகளைக் கண்டடைய முடியும் என்னும் சாத்தியங்களை ஏற்படுத்தி இருக்கிறது.

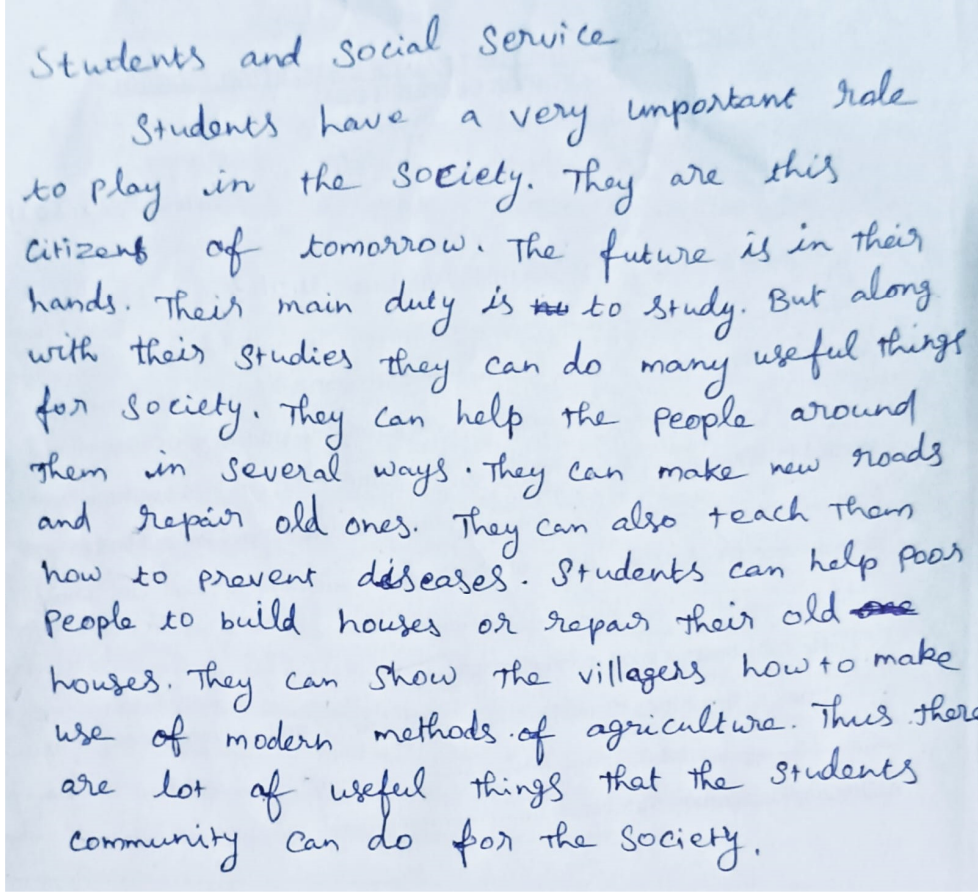
அச்ச உரையின் படம்



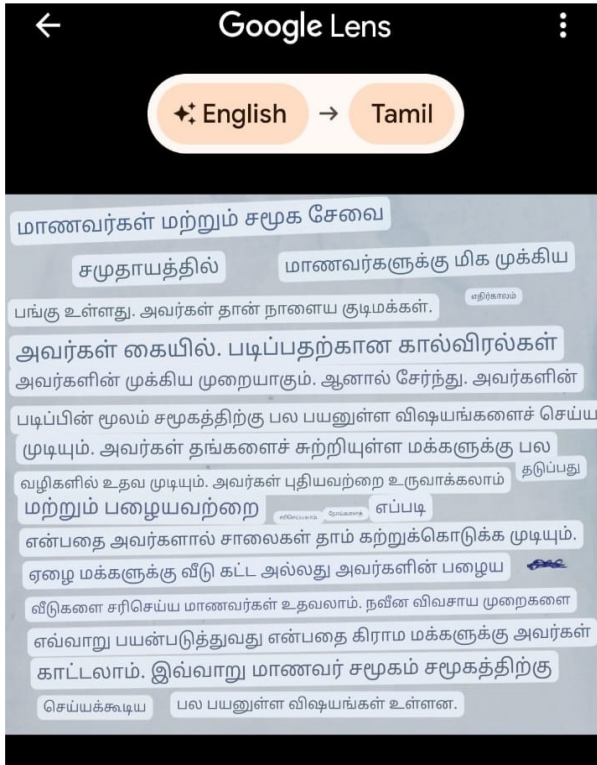
அச்ச உரையின் பட மொழிபெயர்ப்பு



கையால் எழுதப்பட்ட உரையின் படம்



கையால் எழுதப்பட்ட உரையின் மொழிபெயர்ப்பு



கையெழுத்துப் பட மொழிபெயர்ப்பு	பிடிஎப் பட மொழிபெயர்ப்பு
<p>மாணவர்கள் மற்றும் சமூக சேவை</p> <p>சமுதாயத்தில் மாணவர்களுக்கு மிக முக்கிய பங்கு உள்ளது. அவர்கள் தான் நாளைய குடிமக்கள். எதிர்காலம் அவர்கள் கையில். படிப்பதற்கான கால்விரல்கள் அவர்களின் முக்கிய முறையாகும். ஆனால் சேர்ந்து. அவர்களின் படிப்பின் மூலம் சமூகத்திற்கு பல பயனுள்ள விஷயங்களைச் செய்ய முடியும். அவர்கள் தங்களைச் சுற்றியுள்ள மக்களுக்கு பல வழிகளில் உதவ முடியும். அவர்கள் புதியவற்றை உருவாக்கலாம் மற்றும் பழையவற்றை சரிசெய்யலாம். நோய்களைத் தடுப்பது எப்படி என்பதை அவர்களால் சாலைகள் தாம் கற்றுக்கொடுக்க முடியும். ஏழை மக்களுக்கு வீடு கட்ட அல்லது அவர்களின் பழைய வீடுகளை சரிசெய்ய மாணவர்கள் உதவலாம். நவீன விவசாய முறைகளை எவ்வாறு பயன்படுத்துவது என்பதை கிராம மக்களுக்கு அவர்கள் காட்டலாம். இவ்வாறு மாணவர் சமூகம் சமூகத்திற்கு செய்யக்கூடிய பல பயனுள்ள விஷயங்கள் உள்ளன.</p>	<p>மாணவர்கள் மற்றும் சமூக சேவை</p> <p>சமுதாயத்தில் மாணவர்களுக்கு மிக முக்கிய பங்கு உள்ளது. அவர்கள் நாளைய குடிமக்கள். எதிர்காலம் அவர்கள் கையில். படிப்பதே இவர்களின் முக்கிய கடமை. ஆனால் அவர்கள் படிப்போடு சமூகத்திற்கு பல பயனுள்ள விஷயங்களைச் செய்ய முடியும். அவர்கள் தங்களைச் சுற்றியுள்ளவர்களுக்கு பல வழிகளில் உதவ முடியும். அவர்கள் புதிய சாலைகளை உருவாக்கலாம், பழைய சாலைகளை சரிசெய்யலாம். நோய்களைத் தடுப்பது குறித்தும் அவர்களுக்குக் கற்றுக் கொடுக்கலாம். ஏழை மக்களுக்கு வீடு கட்ட அல்லது அவர்களின் பழைய வீடுகளை சரிசெய்ய மாணவர்கள் உதவலாம். நவீன விவசாய முறைகளை எவ்வாறு பயன்படுத்துவது என்பதை கிராம மக்களுக்கு அவர்கள் காட்டலாம். இவ்வாறு மாணவர் சமுதாயம் சமுதாயத்திற்கு செய்யக்கூடிய பயனுள்ள விஷயங்கள் ஏராளம்.</p>

மேற்காட்டப்பட்டுள்ள மாதிரிகளைப் பார்க்கும் பொழுது கூகுள் லென்ஸ் மூலமாக ஒரு படத்தை மொழிபெயர்க்கும் போது அது அச்செழுத்துக்களால் இருந்தால் தெளிவான மொழிபெயர்ப்பும், கையெழுத்தாக இருந்தால் பல தவறுகளும் நடைபெற்றுள்ளதைக் காணமுடிகிறது.

கூகுள் லென்ஸ் மொழிபெயர்ப்பு

மேற்காட்டப்பட்டுள்ள மொழிபெயர்ப்பில் "But, along with their studies they can do many useful things for society" என்னும் தொடர் கையெழுத்தாக இருந்த பட மொழிபெயர்ப்பில் "**அவர்களின் படிப்பின் மூலம் சமூகத்திற்கு பல பயனுள்ள விஷயங்களைச் செய்ய முடியும்**" என்றும், "அச்செழுத்துப் பட மொழிபெயர்ப்பில் **ஆனால் அவர்கள் படிப்போடு சமூகத்திற்கு பல பயனுள்ள விஷயங்களைச் செய்ய முடியும்**" என்றும் மொழிபெயர்க்கப்பட்டுள்ளது. இதில் நாம் கவனிக்க வேண்டியது இதன் தொடரமைப்பு. கையெழுத்துப் படத்தின் தொடரமைப்பில் எந்த வித சிக்கலும் இல்லாமல் எழுவாய், பயனிலை, செயப்படுபொருள் என்ற அமைப்பில் தெளிவான ஒரு முழுமையான பொருள் வெளிப்படுகிறது. அதேபோல் அச்சுப் பட உரையின் மொழிபெயர்ப்பு ஆங்கிலத் தொடருக்கு இணையான மொழிபெயர்ப்பாக

உள்ளது. இதில் கூகுள் மொழிபெயர்ப்பு சொல்லுக்குச் சொல் மொழிபெயர்ப்பு என்று இல்லாமல் அச்சொல்லை முழுமையான தொடராக மாற்றும் செயல்பாடுகளையும் கொண்டுள்ளது. இது மொழிபெயர்க்கப்படும் மொழியைப் பொருத்து மாறுபடும்.

மேலும்,

They can help the people around them in several ways. They can make new roads and repair old ones. They can also teach them how to prevent diseases. என்னும் தொடரை,

அச்செழுத்துப்பட உரை:

அவர்கள் புதிய சாலைகளை உருவாக்கலாம், பழைய சாலைகளை சரிசெய்யலாம். நோய்களைத் தடுப்பது குறித்தும் அவர்களுக்குக் கற்றுக் கொடுக்கலாம்.

கையெழுத்துப் பட உரை

அவர்கள் புதியவற்றை உருவாக்கலாம் மற்றும் பழையவற்றை சரிசெய்யலாம். நோய்களைத் தடுப்பது எப்படி என்பதை அவர்களால் சாலைகள் தாம் கற்றுக்கொடுக்க முடியும்.

கையெழுத்தில் road என்ற சொல் தெளிவாக இல்லாததால் மொழிபெயர்க்கும் ஒரு முழுமை பெற்ற தொடரை கூகுள் மொழிபெயர்ப்புச் செயலி உருவாக்குகிறது. புதிய சாலைகளை உருவாக்கலாம், பழைய சாலைகளைச் சரி செய்யலாம் என்பது புதியவற்றை உருவாக்கலாம் மற்றும் பழையவற்றை சரி செய்யலாம் என்றும் மொழிபெயர்க்கப்பட்டுள்ளது. இதற்கு அடுத்த அடியில்,

“நோய்களைத் தடுப்பது குறித்தும் அவர்களுக்குக் கற்றுக் கொடுக்கலாம்” என்று அச்சுப் பட உரையில் மொழிபெயர்க்கப்பட்டுள்ள தொடர், கையெழுத்துப் பட உரையில்,

“நோய்களைத் தடுப்பது எப்படி என்பதை அவர்களால் சாலைகள் தாம் கற்றுக்கொடுக்க முடியும்” என்று மொழிபெயர்க்கப்பட்டுள்ளது. இதில் எழுவாய் பயனிலை செயப்படுபொருள் என்ற அமைப்பில் மொழிபெயர்க்கப்பட்டிருந்தாலும் பொருள் நிலையில் தவறானதாகவே இத்தொடர் மொழிபெயர்க்கப்பட்டுள்ளது.

முடிவாக,

- தேடு பொறியில் சொற்களுக்கு மாற்றாக படங்களைப் பயன்படுத்த முடியும் என்ற சாத்தியங்களை கூகுள் நிறுவனம் ஏற்படுத்தியுள்ளது.
- படங்கள் மூலமாகத் தேடுவதற்கு கூகுள் லென்ஸ் என்ற பெயரில் தனி செயலியை உருவாக்கியுள்ளது.
- படங்களை உள்ளீடு செய்து, மொழிபெயர்த்தல், இடத்தைக் கண்டறிதல், பொருட்களை வாங்குதல் என பலவற்றைச் செய்ய முடியும்.
- கூகுள் மொழிபெயர்ப்பே கூகுள் லென்ஸிலும் படங்களை மொழிபெயர்க்கிறது.
- சொல்லுக்குச் சொல் மொழிபெயர்த்தல் என்ற நிலையில் இல்லாமல், தொடரமைப்பை ஒட்டி ஒரு தொடர் மொழிபெயர்க்கப்படுகிறது.
- கூகுள் மொழிபெயர்ப்பைப் படங்களைப் பயன்படுத்திச் செய்வதில் அச்செழுத்தாக இருந்தால் ஓரளவு சரியான மொழிபெயர்ப்பினையும், கையெழுத்தாக இருந்தால் பல இடங்களில் தவறான மொழிபெயர்ப்பினையும் கொடுக்கிறது.

- கூகுள் மொழிபெயர்ப்பின் துல்லியத்தன்மை மேலும் ஆராயப்பட வேண்டிய ஒன்று. அதற்கு நாம் மேலும் பல தமிழ்ச் சொற்களையும் அவற்றின் பயன்பாட்டு விதிகளையும் இணையத்தில் தொடர்ந்து பதிவேற்றி நம் மொழியின் மேன்மையை நம்மொழியை அறியாதவர்களும் அறியும் வண்ணம் செய்ய வேண்டும்.

Improving Tamil-Telugu Neural Machine Translation using Morpheme-based Tokenization

**Parameswari Krishnamurthy <parameshkrishnaa@gmail.com>,
Sushvin Marimuthu <sushvinmarimuthu@gmail.com> &
Nagaraju Vuppala <nagaraju.vuppala@research.iiit.ac.in>
International Inst. of Information Technology, Hyderabad, India**

Neural Machine Translation (NMT) systems face significant challenges when translating between morphologically complex languages like Telugu and Tamil due to their rich morphological structures. This paper proposes the integration of morpheme-based tokenization techniques to enhance the performance of Telugu-Tamil NMT systems.

Morpheme-based tokenization allows for the decomposition of words into their smallest meaningful units, aligning well with the agglutinative nature of Telugu and Tamil. By preserving morphological information, this approach addresses issues related to out-of-vocabulary words and captures fine-grained semantic nuances during translation.

Tokenization is a fundamental step in natural language processing (NLP) tasks, where text is divided into smaller units called tokens. Each token represents a distinct unit of meaning within the sentence. These tokens could be words, punctuation marks, numbers etc., By breaking down text into tokens, it becomes possible to analyse the language in various ways, such as identifying parts of speech, named entity recognition, training machine learning models etc.

In morpheme-based tokenization, words are segmented into morphemes based on linguistic rules rather than simply splitting them by spaces or punctuation marks. A morpheme is the smallest unit of language that carries meaning. It can be a word or a part of a word, such as prefixes, suffixes, roots, or grammatical markers. In agglutinative languages, morpheme-based tokenization is particularly important due to the complex nature of word formation. This paper examines the efficacy of morpheme-based tokenization for Tamil in comparison to Byte Pair Encoding (BPE), Sub-Word tokenization, WordPiece tokenization, and customized tokenization methods for Tamil.

By leveraging this approach, the language models can better capture the intricate structure and semantics, leading to improved performance in various natural language processing tasks. Evaluation and refinement techniques are also discussed to enhance the accuracy and effectiveness of the language models. The integration of these models into applications requiring Tamil text processing especially in machine translation by demonstrating the practical relevance and utility of the proposed methodology. Our findings underscore the importance of linguistically informed tokenization techniques in improving the effectiveness of NMT systems, particularly for language pairs characterized by complex morphology like Telugu and Tamil.

**Tokenization, training and fine-tuning strategies for large language models
for Tamil to English text translation task**

Siddharth Krishna Kumar, <kksiddharth@gmail.com>

RingCentral Innovation India Private Limited, Bangalore &

Madhavaraj, A, <madhavarajaa@gmail.com>

IndiaSpeaks Research Labs Private Limited, Madurai

In this paper, we present our experiments involving different strategies for training and fine-tuning large language model (LLM) for the downstream task of text translation from Tamil to English. We propose a novel tokenizer design which uses a combination of (i) byte pair encoding (BPE) technique, (ii) Tamil word morphological and pronunciation rules. Further, we use various open-source monolingual Tamil text corpora containing 5 million sentences for pretraining the baseline English LLM model. We use another corpus of 50 million tokens of bilingual text which is designed such that alternate sentences are in English and Tamil to train the LLM. This is to enable the model to jointly learn to predict next tokens in both languages, at the same time to learn the correspondence and relation between successive sentences which are in English and Tamil. Finally, the trained model is finetuned using 1 million pairs of Tamil-English translated sentence pairs for the downstream task of Tamil to English text translation. This finetuned model is then used to evaluate the translation capability of 50000 Tamil test sentences and we report the BLEU scores.

We have independently performed our experiments on two different baseline LLMs namely LLAMA-7b and MISTRAL-7b models and compare the results. We achieved BLEU scores of 39 and 41 for LLAMA and MISTRAL trained models respectively, when testing the translation of 50000 Tamil test sentences.

பழந்தமிழ் இலக்கியத்தில் மெய்ம்மயக்கத்தின் பாங்கு

இராம்பிரசாந்த் வெங்கடக்கிருஷ்ணன்¹, &
பாலசுந்தரராமன் இலக்குவன்²,

¹ramprashanthvenkatakrishnan@gmail.com, மதுரை காமராசர் பல்கலைக்கழகம், இந்தியா ;

² sundar@arizona.edu, இண்டிடு சப்பான், தோக்கியோ

ஆய்வுச்சுருக்கம்

சங்கத்தமிழ் இலக்கியங்களில் மெய்யொலித் தொடர்களை (consonant clusters) அடையாளங்கண்டு அவற்றின் எண்ணிக்கையைப் அட்டவணையிலிடும் மென்பொருளொன்றை பைத்தானில் எழுதியுள்ளோம். அவ் அட்டவணை பழந்தமிழின் ஒலியன்வருதலை வெளிச்சமிட்டுக் காட்டுகிறது. மொழிமுதல் ஒலிகள், மொழியீற்று ஒலிகள், அசைகள் ஆகியவற்றுடன் மெய்ம்மயக்கம், ஒலியன்வருவியலில் முக்கிய பங்காற்றுகிறது. சங்கத்தமிழின் எட்டுத்தொகை பத்துப்பாட்டு நூல்கள் அனைத்தையும் ஆய்வுக்கு எடுத்துக் கொண்டோம். மெய்ம்மயக்கங்களின் எண்ணிக்கையில் தெளிவாகத் தெரிந்த பாங்கை நிகழ்தகவுக் கோட்பாட்டின்வழி அலசி (probabilistic analysis) அதன்பின்னால் இருக்கக்கூடிய மொழியியற் கூறுகளை இக்கட்டுரையில் தந்துள்ளோம். காட்டாக, எதனால் வல்லொலி இரட்டிக்கிறது அல்லது இனமான மூக்கொலியைத் தொடர்ந்து வருகிறது போன்ற கேள்விகளுக்கான விடைகாண விழைந்துள்ளோம். அளவறி ஒலியியல் (quantitative phonology) முறையில் மெய்ம்மயக்கத்தை ஆய்வுசெய்யும் முதல் முயற்சி இது.

1. ஆய்வுமுறை

சங்க இலக்கியப்பாடல்கள் யாப்பிலக்கணத்துக்குட்பட்ட மரபுப்பாடல்கள். ஆகையால் பாவகைக்கேற்ப வரக்கூடிய சீர்கள் தளைத்தாதவாறு பிரிக்கப்பட்டிருக்கின்றன. ஒரு சொல் இருவேறு சீர்களாகப் பிரிந்தும் இருவேறு சொற்களின் பகுதிகள் ஒரே சீராக இணைந்தும் வரலாகும். அவ்வாறிருக்கையில் மெய்ம்மயக்கங்களை அளவிடும்போது சில மெய்யெழுத்துத் தொடர்கள் பிரிந்தும் வரலாம். சொல் எல்லையைத் தாண்டிய தொடர்கள் இணைந்தும் வரலாம். இவ்வாய்வுக்கு நாங்கள் இருவகைகளிலான உரைகளை எடுத்துக்கொண்டுள்ளோம்.

முதலாவதாக யாப்பிலக்கணத்துக்குப் பொருத்தமான வடிவிலமைந்த உரையில் இடைவெளிகளை நீக்கிவிட்டு முழுநீள உரையில் மெய்யொலித் தொடர்களின் எண்ணிக்கையை அளவிட்டோம். இதை அ வகை எனக் கொள்வோம். இரண்டாவதாக, யாப்பு வடிவிலன்றி ஒவ்வொரு சொல்லாகச் சீர்பிரித்த உரை. இதை ஆ வகை எனலாம். முந்தையதில் சொல் எல்லை கடந்த சில தொடர்கள் கூட்டப்படும். பிந்தையதில் சொல்லெல்லைக்குட்பட்ட எண்ணிக்கை மட்டுமே வரும். அதேவேளை புணர்ச்சிநிமித்தம் வரக் கூடிய தொடர்கள் அறுந்துவிடுகின்றன. இவ்விரண்டு எண்ணிக்கைகளின் கூட்டலிடை, மெய்ம்மயக்கத்தின் பாங்கில் சாய்வின்றிப் பயன்படுகிறது.

2. மெய்ம்மயக்கத்தின் எண்ணிக்கைகள் CLUSTER FREQUENCIES

சங்க நூல்களான எட்டுத்தொகையின் எட்டு நூல்களிலும் பத்துப்பாட்டின் பத்து நூல்களிலும் வந்துள்ள மெய்ம்மயக்கங்களின் (மெய் + மெய்) எண்ணிக்கைகளை பின்வரும் அட்டவணைகளில் காணலாம். அட்டவணைகளில் கிடை வாரியாகவும் செங்குத்து வரிசையிலும் வரும் அதிகபட்ச எண் கொண்ட பெட்டிகள் நிறத்தால் சுட்டிக்காட்டப்பட்டுள்ளன. ஒரு பெட்டி சாம்பல் நிறத்தால் குறிக்கப்பட்டிருந்தால் அது அந்த அட்டவணையில், அந்த வரிசையில் உள்ள அதிகபட்ச எண் என்பதைக் குறிப்பதற்காகவே. அதேபோல் நெடு வரிசையில் உள்ள அதிகபட்ச எண் கொண்ட பெட்டி கருப்பு நிறத்தால் குறிக்கப்பட்டிருக்கிறது. ஒரு பெட்டியில் உள்ள எண் கிடைவாரியாகவும் நெட்டுக்குத்தாகவும் இரண்டிலும் சேர்ந்து அதிகபட்ச எண்ணாய் இருந்தால் அந்தப் பெட்டி கருஞ்சாம்பல் நிறத்தால் குறிக்கப்பட்டுள்ளது. இவ்வாறு எண்ணிக்கை மிகுந்த மெய்த்தொடர்களை நிறம்பிரித்துக் காட்டும்போது ஆர்வூட்டும் பாங்கொன்று தெளிவாகிறது. இது தமிழ் ஒலியன்வருவியலின் அழகினைக் கூறுவதாய் அமைகிறது. சாம்பல் நிற பெட்டிகள் அனைத்தும் வல்லின-வல்லின PPs (geminate plosive) மெய்ம்மயக்கங்களை கொண்டவை. கருப்பு நிற பெட்டிகளோ மெல்லின-மெல்லின NNs (geminate nasal stops). மெய்ம்மயக்கங்களைக் கொண்டவை. கருஞ்சாம்பல் நிறம் கொண்ட பெட்டிகள் மெல்லின-இணைவல்லின NPs (homorganic nasal-oral stop clusters) மெய்ம்மயக்கங்களைக் கொண்டவை. தமிழ் எழுத்து முறையானது வலிந்த மெலிந்த வல்லொலிகளை வரிவடிவத்தில் பிரித்துக் காட்டாது. உயிரிடை வல்லொலிகள் மெலிவது இயல்பு. இவ்வாய்வில் நாங்கள் அவ்வேறுபாட்டைக் கணக்கில் கொள்ளவில்லை. மேலும் இந்த மெய்ம்மயக்க எண்ணிக்கைக் கணக்கெடுப்பு, அவை ஒரு உருபனுக்கு உள்ளே இருப்பவையா அல்லது இரண்டு உருபன்களுக்கு நடுவே உள்ளவையா எனக் கருத்தில் கொள்ளாமல் செய்யப்பட்டது.

3. மெய்ம்மயக்க எண்ணிக்கைகளை காட்சிப்படுத்தல்

3.1 எட்டுத்தொகை முழுவதும்

	k	ñ	c	ñ	t	ñ	r	ñ	t	n	p	m	y	v	r	i	l
k	8838	0	5	0	1	0	0	0	8	0	4	1	0	0	1	1	0
ñ	6730	9	5	0	0	0	0	0	1	0	0	0	0	0	0	0	0
c	5	0	2300	0	0	0	0	0	3	1	3	0	0	1	0	0	0
ñ	2	0	2377	58	0	0	0	0	0	0	1	0	0	1	0	0	0
t	325	0	120	0	2585	0	0	0	12	1	298	1	1	0	0	0	0
ñ	877	1	173	11	2662	1380	1	2	567	234	858	606	37	198	1	1	0
r	1027	1	201	0	0	0	3234	0	18	3	810	2	0	2	2	0	0
ñ	1992	1	866	27	1	6	4397	3016	1570	1235	2418	1952	236	1248	2	0	0
t	2	0	3	0	0	0	0	1	8944	2	0	0	0	0	1	0	0
n	1	0	0	1	0	0	2	1	9603	218	0	0	0	0	0	0	0
p	1	0	2	0	0	0	1	0	0	0	8160	1	0	0	0	1	0
m	2332	0	976	20	0	0	0	1	1565	1329	6036	2009	172	1298	1	2	0
y	535	1	197	4	0	1	0	1	1185	525	466	339	437	380	5	0	0
v	1	0	0	0	0	0	0	0	1	0	0	0	4	526	0	0	1
r	2649	10	763	60	0	1	0	0	1598	1990	2711	1456	114	1380	1	0	0
l	1809	0	633	28	0	0	0	0	1106	1051	1401	1199	247	1952	1	3218	0
l	452	0	160	13	0	0	1	0	342	424	498	399	65	720	1	6	1376
l	352	0	122	3	0	0	0	0	380	555	332	170	6	180	0	1	0

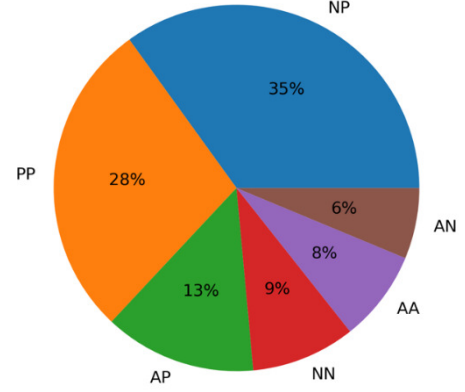
படம் 1. அ வகை

	k	ñ	c	ñ	t	ñ	r	ñ	t	n	p	m	y	v	r	i	l
k	4505	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ñ	5375	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	430	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ñ	0	0	1421	54	0	0	0	0	0	0	0	0	0	0	0	0	0
t	211	0	49	0	2306	0	0	0	0	0	140	0	0	0	0	0	0
ñ	351	0	0	0	2712	1237	0	0	18	0	219	198	0	23	0	0	0
r	458	0	49	0	0	0	2704	0	0	0	219	0	0	0	0	0	0
ñ	225	0	6	0	0	0	4568	2845	33	0	470	280	19	2	0	0	0
t	0	0	0	0	0	0	0	6387	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	9323	85	0	0	0	0	0	0	0	0
p	0	0	0	0	0	0	0	0	0	3321	0	0	0	0	0	0	0
m	6	0	0	0	0	0	0	0	0	2411	742	0	4	0	0	0	0
y	198	0	2	0	0	0	0	923	240	105	118	400	122	0	0	0	0
v	0	0	0	0	0	0	0	0	0	0	0	0	0	297	0	0	0
r	761	0	48	1	0	0	0	463	1025	854	57	0	189	0	0	0	0
l	599	0	36	0	0	0	0	1	1	59	1	6	416	0	2793	0	0
l	71	0	0	0	0	0	0	3	0	52	2	1	174	0	0	1273	0
l	139	0	14	0	0	0	0	218	425	106	2	0	52	0	0	0	0

படம் 2. ஆ வகை

	k	ñ	c	ñ	t	ɾ	ɽ	ɽ	t	n	p	m	y	v	r	l	l
k	6672	0	2	0	0	0	0	0	4	0	2	0	0	0	0	0	0
ñ	6052	8	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	2	0	1365	0	0	0	0	0	2	0	2	0	0	0	0	0	0
ñ	1	0	1899	56	0	0	0	0	0	0	0	0	0	0	0	0	0
t	268	0	84	0	2446	0	0	0	6	0	219	0	0	0	0	0	0
ɾ	614	0	86	6	2687	1308	0	1	292	117	538	402	18	110	0	0	0
ɽ	742	0	125	0	0	0	0	2969	0	9	2	514	1	0	1	1	0
ɽ	1108	0	436	14	0	3	4482	2930	802	618	1444	1116	128	625	1	0	0
t	1	0	2	0	0	0	0	0	7666	1	0	0	0	0	0	0	0
n	0	0	0	0	0	0	1	0	9463	152	0	0	0	0	0	0	0
p	0	0	1	0	0	0	0	0	0	0	5740	0	0	0	0	0	0
m	1169	0	488	10	0	0	0	0	782	664	4224	1378	86	651	0	1	0
y	366	0	100	2	0	0	0	0	1054	382	286	228	418	251	2	0	0
v	0	0	0	0	0	0	0	0	0	0	0	0	2	412	0	0	0
r	1705	5	406	30	0	0	0	0	1030	1508	1782	756	57	784	0	0	0
l	1204	0	334	14	0	0	0	0	554	526	730	600	126	1184	0	3006	0
l	262	0	80	6	0	0	0	0	172	212	275	200	33	447	0	3	1324
l	246	0	68	2	0	0	0	0	299	490	219	86	3	116	0	0	0

படம் 3. அ வகை ஆ வகை ஆகியவற்றின் கூட்டலிடை



படம் 4. மெய்ம்மயக்க வீத வட்டப்படம்

3.2 பத்துப்பாட்டு முழுவதும்

	k	ñ	c	ñ	t	ɾ	ɽ	ɽ	t	n	p	m	y	v	r	l	l
k	1560	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0
ñ	1540	1	1	0	0	0	0	0	1	1	1	1	0	1	0	0	0
c	1	0	303	0	0	0	0	0	1	2	0	0	0	1	0	0	0
ñ	1	0	373	23	0	0	0	0	1	0	0	0	1	0	0	0	0
t	59	0	39	0	495	0	0	0	1	62	0	0	1	0	0	0	0
ɾ	122	0	43	7	545	278	1	0	34	27	129	131	7	58	0	0	0
ɽ	198	0	74	0	0	0	584	0	1	192	2	0	0	0	0	0	0
ɽ	343	0	146	8	0	0	764	698	114	68	337	374	15	225	0	0	0
t	1	0	2	0	0	0	0	0	1640	0	1	3	0	3	0	0	0
n	0	0	0	0	0	0	0	0	1997	54	0	1	0	1	0	0	0
p	2	0	0	0	0	1	0	0	2	2	1360	2	0	1	0	0	0
m	369	0	143	3	0	0	0	0	205	113	1357	348	3	284	0	0	0
y	110	0	36	1	0	0	0	0	173	84	88	78	93	66	0	0	0
v	0	0	0	0	0	0	0	0	0	0	0	0	0	80	0	0	0
r	457	2	148	17	0	0	0	0	338	446	515	266	12	267	42	0	0
l	383	0	122	1	0	0	13	11	93	87	280	159	45	432	0	402	0
l	111	0	34	1	7	6	0	0	44	35	84	69	0	101	0	240	0
l	107	0	50	2	0	0	0	0	96	163	77	48	1	52	0	0	7

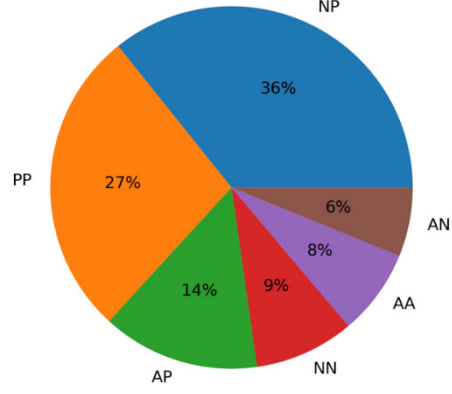
படம் 5. அ வகை

	k	ñ	c	ñ	t	ɾ	ɽ	ɽ	t	n	p	m	y	v	r	l	l
k	515	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ñ	590	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	65	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ñ	0	0	102	12	0	0	0	0	0	0	0	0	0	0	0	0	0
t	18	0	13	0	270	0	0	0	0	0	16	0	0	0	0	0	0
ɾ	12	0	0	0	273	115	0	0	5	1	14	17	0	5	0	0	0
ɽ	25	0	2	0	0	0	282	0	0	0	24	0	0	0	0	0	0
ɽ	21	0	0	0	0	0	354	286	3	0	16	20	4	2	0	0	0
t	0	0	0	0	0	0	0	0	740	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	922	15	0	0	0	0	0	0	0
p	0	0	0	0	0	0	0	0	0	0	437	0	0	0	0	0	0
m	3	0	0	0	0	0	0	0	8	0	334	41	0	0	0	0	0
y	19	0	0	0	0	0	0	0	71	28	11	20	42	18	0	0	0
v	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0
r	67	0	2	0	0	0	0	0	59	135	85	5	0	21	0	0	0
l	40	0	8	0	0	0	0	0	4	1	13	2	2	41	0	152	0
l	14	0	1	0	0	0	0	0	6	1	6	6	0	13	0	100	0
l	13	0	1	0	0	0	0	0	16	56	9	0	0	5	0	0	0

படம் 6. ஆ வகை

இவ்வாய்வுக்கான மென்பொருளை பைத்தான் நிரல்மொழியில் NLTK, Matplotlib, Pandas முதலிய மென்மியப்பொதிகளின் துணையுடன் எழுதியுள்ளோம். அந்நிரலை பின்வரும் இணைப்பில் பெறலாம்: <https://github.com/oligoglot/mayal>. ஆய்வுக்குப் பயன்படுத்திய உரைகளை பின்வரும் இணைப்பில் பெறலாம்: <https://github.com/oligoglot/mayal/tree/main/corpora>. சங்க இலக்கிய நூல் ஒவ்வொன்றுக்குமான மெய்ம்மயக்க அட்டவணைகளைப் பின்வரும் இணைப்பின்வழி பெற முடியும்: <https://github.com/oligoglot/mayal/tree/main/out>.

	k	ñ	c	ñ	t	ɳ	ɳ	t	n	p	m	y	v	r	l	l
k	1038	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ñ	1065	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	184	0	0	0	0	0	0	1	0	0	0	0	0	0
ñ	0	0	238	16	0	0	0	0	0	0	0	0	0	0	0	0
t	38	0	26	0	382	0	0	0	0	39	0	0	0	0	0	0
ɳ	67	0	22	4	409	196	0	0	20	14	72	74	4	32	0	0
ɳ	112	0	38	0	0	0	433	0	0	108	1	0	0	0	0	0
n	182	0	73	4	0	0	559	490	58	34	176	197	10	114	0	0
t	0	0	1	0	0	0	0	0	1190	0	0	2	0	2	0	0
n	0	0	0	0	0	0	0	0	1460	34	0	0	0	0	0	0
p	1	0	0	0	0	0	0	0	1	1	888	1	0	0	0	0
m	186	0	72	2	0	0	0	0	106	56	846	194	2	142	0	0
y	64	0	18	0	0	0	0	0	122	56	50	49	68	42	0	0
v	0	0	0	0	0	0	0	0	0	0	0	0	54	0	0	0
r	262	1	75	8	0	0	0	0	198	290	300	136	6	144	21	0
l	212	0	65	0	0	0	6	6	48	44	146	80	24	236	0	277
l	62	0	18	0	4	3	0	0	25	18	45	38	0	57	0	170
l	60	0	26	1	0	0	0	0	56	110	43	24	0	28	0	0



படம் 7. அ வகை ஆ வகை ஆகியவற்றின் கூட்டலிடை

படம் 8. மெய்ம்மயக்க வீத வட்டப்படம்

4. ஓரொற்றடுத்து மெய்யொலி வரும் தொடர்களின் வகைகள் TYPES OF BICONSONANTAL CLUSTERS

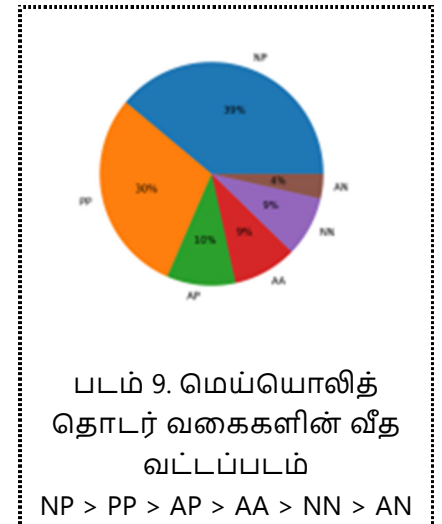
மேலேயுள்ள அட்டவணைகளில் காணும் ஓர் ஒற்றெழுத்தை அடுத்து மெய்யொலி வரும் தொடர்களைப் பின்வருமாறு வகைப்படுத்தலாம்.

1. இனமான மெய்யெழுத்தைத் தொடர்ந்து ஒத்த வல்லெழுத்து வருவது. இதை **NP** (Nasal plus Plosive cluster) எனக்குறித்துள்ளோம்.
2. வல்லொற்றிரட்டல். இதை **PP** (a geminate Plosive) எனக் குறித்துள்ளோம்.
3. மெல்லொற்றிரட்டல் **NN** clusters (a geminate Nasal stop). இது தமிழில் அரிதாக வரும்.
4. இடையொற்றைத் தொடர்ந்து வல்லொலி வருவதும் **AP** (Approximant plus Plosive) இரு இடையொலிகள் தொடர்ந்து வருவதும் **AA** (geminate Approximant or Hetero-organic Approximant cluster), இடையொற்றைத் தொடர்ந்து மெல்லொலி வருவதும் **AN** (Approximant plus Nasal) குறைந்த எண்ணிக்கையில் நிகழ்கின்றன.

5. எண்ணிக்கையின் போக்கு

5.1 கூடுதல் எண்ணிக்கையில் காணும் மெய்த்தொடர் வகைகள்

எட்டுத்தொகையிலும் பத்துப்பாட்டிலும் மேலே குறிப்பிட்ட கிடைவாரி, நெடுவாரி, இருவாரிப் பார்வைகளில் எண்ணிக்கையில் மிகுதியாக வரும் தொடர்களையும் அவற்றின் வகைகளையும் பின்வரும் அட்டவணையில் காணலாம்.



அட்டவணை யில் பெட்டி நிறம்	மிகுதியான தொடர் வகைகள்	எட்டுத்தொகையிலும் பத்துப்பாட்டிலும் காணும் மெய்ம்மயக்கங்கள்
	வல்லொற்றிரட்டல் (PP) மிகுந்தது	க்க kk, ச்ச cc, ட்ட tt, ற்ற rr, த்த tt, ம்ப mp, ய்த yt, வ்வ vv, ர்ப rp, ழ்ந் ln
	மெல்லொலி- வல்லொலி தொடர் (NP) மிகுந்தது	ங்க nk, ஞ்ச ஞ்ச, ண்ட nt, ன்ற nr, ந்த nt, ப்ப pp, ல்ல ll, ள்ள !!
	மெல்லொலித் தொடர் (NN) மிகுந்தது	ஞ்ஞ ற்ற, ங்ங ற்ற, ன்ன ற்ற, ம்ம mm, ய்ய yy, ய்ர yr, ம்ழ் m, ர்ங் n, ர்ந் n, ல்வ lv

எதனால் இப்படியோர் ஒழுங்கான பாங்கு வெளிப்படுகிறது? நிகழ்தகவின் அடிப்படையில் இதை அணுகிப் பார்க்கலாம். ககர வரிசையை முதலில் எடுத்துக் கொள்வோம். ககர ஒற்றைத் (க் k) தொடர்ந்து அதிக எண்ணிக்கையில் வந்துள்ள ஒற்று எது? ககரமே. வல்லொற்றைத் தொடர்ந்து வேறு மெய்யொலி வருவதில்லை. அதனாலேயே வெளிர்சாம்பல் நிறப்பெட்டிகள் இத்தகையனவாக உள்ளன. அதேபோல மெல்லொலிகளை எடுத்துக் கொள்வோம். ணகர ஒற்றைத் (ண் n) தொடர்ந்து எது மிகுதியாக வருகிறது எனப் பார்த்தால் டகரமே. வேற்றுநிலை மெய்ம்மயக்கத்தில் மெல்லொற்றைத் தொடர்ந்து இனமான வல்லொலி மட்டுமே பெரும்பாலும் வரும். இதுவே கருநிறப் பெட்டிகளில் அமைகிறது.

கிடைவாரியாக இல்லாமல் நெடுவாரியாகப் பார்த்தால் மெல்லொலிக்கு முந்தைய ஒற்றுகள் வேறு எதுவுமிருக்க முடியாது. மெல்லொற்றிரட்டல் மட்டுமே அங்கு வாய்ப்புள்ளது. இந்தப்பாங்கு பழந்தமிழின் ஒலிப்பிலக்கணத்தின் முக்கிய கூறாக விளங்குகிறது.² இன்றும் தமிழில் இப்பாங்கு நீடித்து வருகிறது.

5.2 எழுத்தர்களின் பிழைகள்

நாங்கள் கணக்கிட்ட அட்டவணைகளில் மிகக் குறைவான எண்ணிக்கை கொண்ட மெய்ம்மயக்கங்கள் சிலவும் இடம்பெற்றுள்ளன. இவை எழுத்தர்களின் பிழைகளெனக் கருதுகிறோம். இச்செய்யுள்களில் காணப்படும் இம் மெய்ம்மயக்கங்கள் பழந்தமிழின் ஒலியன்சேர்க்கை விதிகளுக்கே புறம்பானவை. ஆதலால் இவை எழுத்தர்களின் கவனக்குறைவினால் கடத்தப்பட்டவை என தோன்றுகிறது. ஒலியன்சேர்க்கை விதிகளில் அனுமதிக்கப்படாத மற்றும் எழுத்தர் பிழையால் ஏற்பட்ட மெய்ம்மயக்கங்கள் சில - க்ட் kt, க்ர் kr, க்ல் kl, ங்த் nt, ச்த் ct, ச்ந் cn.

5.3 அதிக எண்ணிக்கை கொண்ட மெய்ம்மயக்கங்களுக்கான உருபன்- ஒலியனியல் காரணங்கள்

1. தன்வினை - பிறவினை³ உறழ்ச்சி கொண்ட வினைச்சொற்களில், மெல்லின-இணைவல்லின NP மற்றும் வல்லின-வல்லின PP மெய்ம்மயக்க உறழ்ச்சி (contrast) உருபனியல் அடிப்படையிலும் மிகவும் செழுமையானது. தன்வினைச் சொற்கள் மெல்லின-இணைவல்லின NP மெய்ம்மயக்கங்கள் கொண்டவை. அதேபோல்

பிறவினைச் சொற்கள் வல்லின-வல்லின PP மெய்ம்மயக்கங்கள் கொண்டவை. இந்த வினைச் சொற்களின் பயன்பாட்டால் செய்யுளில் மெல்லின-இணைவல்லின NP மற்றும் வல்லின-வல்லின PP மெய்ம்மயக்கங்கள் அதிக எண்ணிக்கைகளில் காணப்படலாம். எ.கா. - பி.வி. அடக்கு aṭakku vs த.வி. அடங்கு aṭanku.

2. கூட்டுப் பெயர்ச்சொற்களில் (compound nouns) புணர்ச்சி காரணமாக உருபங்களின் நடுவே வல்லின-வல்லின PP (தங்கம் taṅkam + கிண்ணம் kiṇṇam > தங்கக்கிண்ணம் taṅkakiṇṇam) வகையிலான மெய்ம்மயக்கங்களும் மெல்லின-இணைவல்லின NP (மா mā + பழம் paḷam > மாம்பழம் māmpaḷam) வகையிலான மெய்ம்மயக்கங்களும் வருவதுண்டு. இது அவற்றின் எண்ணிக்கையை அதிகப்படுத்தலாம்.
3. வல்லினம் இரட்டித்தல் என்பது பெயர்ச் சொல்லிலிருந்து கிழமைப் பொருள் (genitive) காட்டுவதற்கு அடிக்கடிப் பயன்படும் ஓர் உருபன்-ஒலியனியல் உத்தி. எ.கா. - வீடு vīṭu > வீட்டு vīṭṭu. இது வல்லின-வல்லின PP மெய்ம்மயக்கங்களின் எண்ணிக்கையைக் கூட்டுகிறது.
4. பல பெயர், வினைச் சொற்களின் வேர்களிலேயே இந்த வல்லின-வல்லின PP, மெல்லின-இணைவல்லின NP-யும் மெல்லின-மெல்லின NN மெய்ம்மயக்கங்களும் உள்ளன . எ.கா. - பக்கம் paṅkam, தங்கு taṅku, அம்மா amma.
5. “தனிக்குறில் முன் ஒற்று உயிர்வரின் இரட்டும்” என்றொரு நன்னூல் விதி உண்டு. இதன்படி இந்த குறித்த உருப-ஒலியனியல் சூழலில் இடையின மற்றும் மெல்லின மெய்யொலிகள் இரட்டிக்கின்றன. எ.கா. - கல் kaḷ + உ u > கல்லு kaḷḷu, கண் kaṇ + இல் il > கண்ணில் kaṇṇil. இவ்வாறு இரட்டித்தல் மெல்லின-மெல்லின NN மற்றும் இடையின-இடையின AA மெய்ம்மயக்கங்களின் எண்ணிக்கைகளை கூட்டும்.
6. தொல்காப்பியத்தின் குற்றியலுகரப் புணரியல் பிரிவில் ஒரு விதி (Rangan K. 2012) குறிப்பிடப்பட்டுள்ளது. அதன்படி ஒரு பெயர்ச்சொல்லின் இறுதியிலுள்ள மெல்லின-இணைவல்லின NP மெய்ம்மயக்கம் இன்னொரு பெயர்ச்சொல்லின் வருகையால் வல்லின-வல்லின PP மெய்ம்மயக்கமாக மாறும் என்பதே. எ.கா. - குரங்கு kuṛaṅku + கால் kāḷ > குரக்குக் கால் kuṛaṅkuṅ kāḷ. இம்மாதிரியான உருபன்-ஒலியனியல் விதிகளின் செழுமையான பயன்பாட்டால் சில மெய்ம்மயக்கங்கள் அதிக எண்ணிக்கையில் காணப்படுகின்றன.

5.4 NP மெய்ம்மயக்கங்களின் அதிக எண்ணிக்கைக்கு சில காரணங்கள்

புகழ்பெற்ற மொழியியலாளரான இராபட்டு கால்டுவெல்லின் திராவிட மொழிகளின் ஒப்பிலக்கணம் என்னும் நூலில் ‘ஒலித்துணை மெல்லினம் சேர்தல்’ (Euphonic Nunnation or Nasalisation) என்னும் ஒரு கருத்தை முன்வைக்கிறார். அவருடைய இந்த நோக்கீட்டின்படி திராவிட மொழிகள் சிலவற்றின் வேர்ச்சொற்களில் சேர்க்கப்படும் பின்னொட்டுகளின் முதல் மெய்யொலியின் முன் ஒரு மெல்லின ஒலி சேர்கிறது (Caldwell 1875).

எடுத்துக்காட்டாக, அது, இது என்னும் சுட்டுச் சொற்களின் வல்லின மெய்யொலிகளின் முன்னர் மெல்லின ஒலி சேர்த்து அந்த, இந்த என்னும் சுட்டுப் பெயரெச்சச் சொற்கள் பிறக்கின்றன.

அத்(உ) + அ > அந்த் + அ > அந்த

ஒலித்துணை மெல்லினம் சேர்தல் (euphonic nunnation) எனும் இவ்விளைவே மெல்லின-இணைவல்லின NP மெய்ம்மயக்கங்களின் அதிக எண்ணிக்கைக்குக் காரணமாக இருக்கலாம். இந்த மெல்லினம் சேர்தல் கன்னடத்திலும் தெலுங்கிலும் அதிகம் காணப்படுவதில்லை. ஆதலால், ஒப்பீட்டில் ஒரு தமிழ்ச் சொல்லின் மெல்லின-இணைவல்லின NP மெய்ம்மயக்கத்திற்கு இணையாக கன்னடத்தில் வெறும் ஒற்றை வல்லொலி தான் காணப்படுகிறது.

எ.கா. -	கன்னடம்	எரடு	eraḍu
	தமிழ்	இரண்டு	iraṇḍu
	கன்னடம்	மூறு	mūru
	தமிழ்	மூன்று	mūṇru

கால்டுவெல் 'ந்த்' மற்றும் 'ம்ப்' என்னும் மெல்லினம் சேர்ந்த வல்லொலிகள் தமிழ்ப் பெயர்ச் சொற்களின் இறுதிகளில் அதிகம் காணப்படுகின்றன என்று கூறுகிறார். இந்த மெல்லினம் சேர்தல் தமிழை அதன் உறவு மொழிகளிடமிருந்து வேறுபடுத்திக் காட்டுகிறது. எ.கா. - தமிழ் எறு-ம்பு eru-**mbu**, கன்னடம் இறு-வே iru-**ve**. மலையாளம் தமிழுக்கு மிக நெருக்கமான மொழியாதலால் இந்த 'மெல்லினம் சேர்த்தல்' வழங்கிவருகிறது. அம்மொழியில் சில இடங்களில் மெல்லினம் இரட்டித்தலாக மூக்கொலி சேர்தலின் அடுத்த நிலையை எட்டியுள்ளது:

மலையாளம்	மூந்து	mūnnu
----------	--------	-------

5.5 மெய்ம்மயக்கங்களின் பொதுப் பாங்கு

அதிக எண்ணிக்கையில் வரும் மெய்ம்மயக்கங்கள் பழந்தமிழில் என்னென்ன மெய்யொலிகள் சேர்க்கை கொள்ளக்கூடியவையாக இருந்துள்ளன என்பதையும் சொல் உருவாதலில் அவற்றின் இன்றியமையாத் தன்மையையும் செழுமையையும் (productivity) சுட்டிக்காட்டுகின்றது. பழந்தமிழ் மெய்ம்மயக்கங்களை பொதுமைப்படுத்தி இரண்டு மெய்யொலிகளின் சேர்க்கைக்கு ஒரு வாய்ப்பாட்டைக் கண்டடையலாம்:

NP, PP, NN, AA, AP, AN

இங்கு P என்பது வல்லொலிகளை குறிக்கும் = {**k, c, t, p, r**} = {க், ச், ட், த், ப், ற்}

N என்பது மெல்லின ஒலிகள் = {**ṇ, ñ, ṇ, n, m, ṇ**} = {ங், ஞ், ண், ந், ம், ன்}

A என்பது இடையின ஒலிகள் = {**(ḷ), (l, l), (r), (v, y)**}
= {(ழ்), (ல், ள்), (ர்), (வ், ய்)}

6. முடிவுரை

நாங்கள் கணினி செய்நிரல் மூலமாக பழந்தமிழ் செய்யுள்களில் வரக்காணும் எல்லா மெய்ம்மயக்கங்களின் எண்ணிக்கைகளையும் கணக்கிட்டு அட்டவணை-

யிலிட்டுள்ளோம். இதன்மூலமாகக் கிடைத்த அட்டவணைகளில் வந்துள்ள அதிக எண்ணிக்கை கொண்ட மெய்ம்மயக்கங்களின் பொதுப் பாங்கினைக் கண்டறிந்தோம். இந்த பாங்கின் பின்னுள்ள உருபன் உறழ்ச்சியையும் (contrast) உருபன்-ஒலியனியல் விதிகளையும் விளக்கியுள்ளோம். இந்தக் காரணிகளே NP, PP, NN மற்றும் AA மெய்ம்மயக்கங்களின் அதிக எண்ணிக்கைக்குக் காரணம் என்பது எங்கள் நிலைப்பாடு. அம்பு, அப்பா, அம்மா, அய்யா போன்ற சொற்களில் வரும் தற்செயலாக அமைந்த NP, PP, NN, AA மெய்ம்மயக்கங்களை மட்டும் கொண்டு இந்த மிகுதியான வரவை விளக்கவியலாது.

நன்றி நவிலல்

இவ்வாய்வுக்குத் தேவையான நூல்சான்றுகளை எங்களுக்குப் பெற்றுத்தந்து உறுதுணையாயிருந்த தஞ்சைத் தமிழ்ப் பல்கலைக்கழகத்தில் ஆய்வு மேற்கொண்டுள்ள கிருட்டிணகுமாருக்கும் வாட்டர்லூ பல்கலைக்கழகப் பேராசிரியர் செ.இரா. செல்வக்குமாருக்கும் நன்றி தெரிவிக்க விரும்புகிறோம்.

சான்றுகள்

1. Agesthalingom S. 1977. *A Grammar of Old Tamil with Special Reference to Patirrupattu : Phonology & Verb Morphology*. 1st ed. Annamalainagar: Annamalai University.
2. Caldwell, Robert. *A comparative grammar of the Dravidian or South-Indian family of languages*. Trübner, 1875.
3. Devine, A.M. and Laurence D. Stephens (1977) Two Studies in Latin Phonology, Anma Libri, Saratoga, CA.
4. Hayes, Bruce and Tanya Stivers (1996) "The phonetics of postnasal voicing," ms., Dept. of Linguistics, UCLA, Los Angeles, CA.
5. Hayes, Bruce P. "Phonetically driven phonology." *Functionalism and formalism in linguistics* 1 (1999): 243-285.
6. Ilakkuvanar Singaravel and Tolkāppiyar. 1963. *Tholkāppiyam (In English)* First ed. Madurai: "Kuraḷ Neri" Publishing House.
7. Krishnamurti, Bhadriraju. *The Dravidian Languages*. Cambridge University Press, 2003.
8. Kumaraswami Raja, N. *Post-nasal Voiceless Plosives in Dravidian*. India, Annamalai University, 1969.
9. "மெய்ம் மயக்கம் (meym mayakkam)." *Tamil Wikipedia*. 16 Oct 2022 <https://ta.wikipedia.org/w/index.php?title=%E0%AE%AE%E0%AF%86%E0%AE%AF%E0%AF%8D%E0%AE%AE%E0%AF%8D_%E0%AE%AE%E0%AE%AF%E0%AE%95%E0%AF%8D%E0%AE%95%E0%AE%AE%E0%AF%8D&oldid=3323839>.
10. Shankara Bhat, D. N. *Sound change*. India, Motilal Banarsidass Publishers, 2001.
11. Rangan, K. Toward Formulating Formal Phonological Rules of *Tolkāppiyam - Ēuttatikāram*. Central Institute of Classical Tamil, Chennai. 2012.

**இயந்திர மொழிபெயர்ப்பு: சிக்கல்கள் - தமிழ்ச் செவ்விலக்கியப்
பரவலாக்கத்தை முன்வைத்து)**

Suja Suyambu kurinjiipirai@gmail.com

Assistant professor, DDGD Vaishnav College, Chennai

உலகில் பேசப்படும் மொழிகளில் சில மொழிகள் மட்டுமே செம்மொழிக்கான முழுமையான தகுதிகளையும் பெற்றுச் சிறந்து விளங்குகின்றன. அந்த மொழிகளுள் முதன்மையான இடம் பெற்ற மொழியாகத் தலைசிறந்த அறிஞர்கள் பலரால் ஏற்றுக்கொள்ளப்பட்ட மொழியாகத் தமிழ் மொழி விளங்குகிறது. அதற்கு மிக முக்கியமான அடிப்படைச் சான்றுகளாக விளங்குபவை சங்க இலக்கியங்கள் என்று அழைக்கப்படும் பாட்டும் தொகையும். உலக அளவில் தமிழர்களின் பண்பாட்டுமேன்மைகளை எடுத்துக்காட்டும் மிக முக்கியமான சான்றுகளை- யெல்லாம் தனக்குள்ளே கொண்டு விளங்குபவையாக அவை விளங்குகின்றன. சி.வை. தாமோதரம் பிள்ளை, உ.வே. சாமிநாதையர் உள்ளிட்ட சிறந்த பதிப்பாளர்களால் தமிழுக்குக் கையளிக்கப்பட்ட இந்த இலக்கியங்களின் பெருமையை உலகம் முழுவதும் கொண்டு சேர்க்கவேண்டியது தமிழர்களின் மிகமுக்கியமான கடமையாகும். இதனைக் கருத்தில் கொண்டு ஆய்வறிஞர் உலகமும் பல அரசு நிறுவனங்களும் பல்வேறு முன்னேற்றச் செயல்பாடுகளைச் செய்துகொண்டு வருகின்றன. என்றாலும் உலக வளர்ச்சியின் வேகத்தோடு நாமும் பயணப்படவேண்டிய அவசியம் நமக்கு உள்ளது. இந்தச் செயலை நாம் கணினியின் துணைகொண்டே செய்யவேண்டியுள்ளது.

திராவிட மொழிகளுள் முதன்முதலில் கணினிக்குள் இடம்பெற்ற மொழி என்கிற சிறப்பு கொண்ட நமது தமிழ் மொழியில் பல நுட்பங்களைப் பயன்படுத்தும் நிலைக்கு நாம் வளர்ந்துள்ளோம். கல்வி பயிலும் மாணவர்களுக்கான பல ஏராளமான தகவல்களைக் கொட்டிக் கொடுக்கும் தளமாகக் கணினி விளங்குகிறது. குரல்வடிவக் குறிப்புகளை எழுத்துவடிவ ஆவணங்களாகச் சேமித்தல், எழுத்துவடிவ அறிக்கை-களிலிருந்தும் தரவுகளைப் பிரித்தெடுத்தல், தரவுத்தளங்களை உருவாக்குதல் என்று நாம் செய்யவேண்டிய பணிகள் ஏராளம் உள்ளன. தமிழர்களின் அறிவு மரபைப் பிற மொழியினரும் நாட்டினரும் உணர்ந்துகொண்டு போற்றும் வகை செய்ய-வேண்டுமானால் நமது வளங்களைப் பிறமொழியில் பெயர்த்து அளிக்கவேண்டும். உலகில் மூவாயிரத்திற்கும் மேற்பட்ட மொழிகள் உள்ளன. அத்தனை மொழிகளிலும் நமது இலக்கியங்களை மொழிபெயர்ப்பது மிகுந்த பொருட்செலவுக்குரிய செயல் என்றாலும் அதனைச் செய்வதற்கான அறிஞர்களையும் நேரத்தையும் கண்டடைவது கடினமாக உள்ளது.

ஆனால் இத்தகைய அரிய செயலைக் கணினியின் துணைகொண்டு நாம் சிறப்பாகச் செய்யமுடியும். இலக்கியங்களை மொழிபெயர்ப்பதற்கு நாம் இயந்திரங்களைத் துணைகொள்ளும்போது பல சிக்கல்களைச் சந்திக்க நேர்கிறது.

சான்றாக,

1 ஆர்ப்பு எழு கடலினும் பெரிது அவன் களிறே

கார்ப்பெயல் உருமின் முழங்கல் ஆனாவே

யார்கொல் அளியர் தாமே ஆர்நார்ச்

செறியத் தொடுத்த கண்ணிக்
கவிகை மள்ளன் கைப்பட் டோரே? (புறம்.81- சாத்தந்தையார்)

Desire is greater than seven seas, He is
Carpal tunnel became the knee
yarkol Aliyar himself is Ornarch Touched mesh
Kavikai mallan keyboard dore?

இந்தப் பாடலின் எந்த ஒரு வரியும் சரியாக மொழிபெயர்க்கப்படவில்லை. இயந்திர மொழிபெயர்ப்பில் எளிய தமிழ்த்தொடரே பல வகையான முரண்பட்ட பொருளுடன் மொழிபெயர்க்கப்படுகிறது.

சான்றாக,
2 பனை மரத்தின் கீழ் இருந்து பால் குடித்தாலும் உலகம் கள் என்று கூறும்.

என்கிற தொடரைக் கூகுள் இயந்திர மொழிபெயர்ப்புக்குள் கொடுத்தால் அது

Even if you drink milk from under a palm tree, the world will say s.
என்று மொழிபெயர்க்கிறது. கள் என்பதைத் தமிழ்மொழியில் உள்ள பன்மை விசுவாசம் மட்டுமே கொள்கிறது.

3 பாலுக்கும் கள்ளுக்கும் வண்ணம் ஒன்று.
பார்க்கும் கண்கள் ஒன்று.
உண்டால் இரண்டும் வேறு.
என்கிற தொடர்,

Milk and black are the same color.
Seeing eyes are one.
Undal both are different.

என்றவாறு மொழிபெயர்க்கப்பட்டுள்ளது. உண்டால் என்கிற சொல் அப்படி மொழியாக்கம் செய்யப்பட்டிருக்கிறது. மேற்குறிப்பிட்ட எடுத்துக்காட்டுகள் தமிழ் மொழிக்கு நாம் இன்னும் எவ்வளவு தரவுத் தளங்களை உருவாக்கி அளிக்கவேண்டிய தேவை உள்ளது என்பதைக் காட்டுகிறது.

இந்தக் கட்டுரை,
- தமிழ் இயந்திர மொழிபெயர்ப்பில் எதிர்கொள்ளும் சிக்கல்களையும்,
- அதனை எதிர்கொள்வதற்கு நாம் உருவாக்கவேண்டிய தரவுத்தளத்தின் தன்மையையும் கட்டளையையும் பற்றி எடுத்துக்கூறுவதாக அமைகிறது.
- இதன்மூலம் தமிழர்களின் அறிவுப் பாரம்பரியத்தையும்
- பண்பாட்டுச் செழுமையையும் உலகம் உணரச் செய்வதற்கான வழிகளைச் உருவாக்க முடியும்.

ChatGpt 4 : படைப்பாக்க சிந்தனை

Kasthuri, M

Assistant Professor, Department of Tamil, DRBCCC Hindu College, India

kasthuri@drbccchinducollege.edu.in

மனிதன் உருவாக்கும் விடயங்கள் அறிவார்ந்த தேடல்களை முன்வைத்து எழுதப்படும் படைப்பாக்கங்கள் என்றால் மிகையல்ல. அத்தகைய அறிவுசார் படைப்பிலக்கியம் தற்போது கணினியின் தொழில்நுட்ப வளர்ச்சியில் இடம்பெறும் நிலை உருவாகியுள்ளது.

நாம் கணினியில் தர வேண்டிய தரவுகள் சரியான மெய்மை தன்மையில் இயங்குவது இதில் முக்கியம். தரவுத்தளம் சார்ந்த புரிதல்கள் நம்பகத் தன்மை போன்றவை இதில் முதன்மை இடம்பெறுகிறது.

கணினி வளர்ச்சி வியத்தகு மாற்றங்களைக் கொண்டு இலங்குகின்றது.

அதன் பல்பரிணாமங்களில் மொழி வளர்ச்சியும் ஒன்று. இன்றைய தொழில்நுட்ப உலகில் மொழிக்கான தேவை காலம் தாண்டி மொழி தன்னை நிலைநிறுத்திக் கொள்ள வேண்டிய அவசியம். தமிழ் மொழி தொன்மையான மொழி என்ற போதிலும் கால மாற்றத்திற்கு ஏற்ப தன்னைத் தகவமைத்துக் கொள்ளும் வல்லமை படைத்த மொழியாக விளங்குவது கண்கூடு.

ChatGpt என்பது புதிய செயற்கை நுண்ணறிவுத் திறன் மூலம் கிடைக்கப் பெற்றது. இவ்வகை அமைப்புகளில் தமிழ் மொழி கால் பதிக்கும் சூழலில் மொழியின் தரவுத்தளம் என்பது கணினியில் தேவையான விவரங்களை முழுமையாக தமிழ் மொழி அறிவுசார்ந்து இயங்க வேண்டும்

இவற்றை எல்லாம் கருத்தில் கொண்டு ஆய்வு செய்வதாக இக்கட்டுரை அமைகிறது.

படைப்பாளிகளுக்கு சாவலாக மாறும் நிலை உருவாகும் போது தமிழ் படைப்புத் திறன் என்பது எத்தகைய அறிவு நிலைச் சார்ந்து கட்டமைக்கப்படும் என்பது கேள்விக்குறியே. இவற்றில் சில நிறை குறைகள் உள்ளதைத் தவிர்க்க இயலாது. வளர்ச்சி நோக்கிய இக்கணினி யுகத்தில் தமிழ் தனக்கான இருப்பைத் தக்கவைத்துக் கொள்ள எதிர்ப்படும் சவால்களை நோக்கி பயணிப்பது அத்தியாவசியமானது.

செயற்கை நுண்ணறிவு மூலம் தமிழ்ப் படைப்பாக்கம்

நித்திஷ் செந்தூர்

அனுதினமும் தொழில்நுட்பம் அதிகவேகத்தில் வளர்ந்து வருகிறது. அதில் தற்போது செயற்கை நுண்ணறிவின் அசுர வளர்ச்சி உலகத்தை ஆட்டிப்படைத்து வருகிறது. செயற்கை நுண்ணறிவு என்பது மனித மூளை சிந்தித்து செயல்படுவது போல, கணினிச் செயலநிரலகளை உருவாக்கி, அவற்றைக் கணினியில் உள்ளீடு செய்து, அதன் ஊடாக ஓர் இயந்திரத்தைச் சிந்தித்துச் செயல்பட வைக்கும் முறை எனலாம். செயற்கை நுண்ணறிவை அடிப்படையாகக் கொண்ட புதுபுது இணையத் தளங்கள் முளைத்துக்கொண்டு வருகின்றன. பெருமொழி மாதிரி வரைவை (உ06 (லாரப806 Model) அடிப்படையாகக் கொண்டு இயங்கிவரும் சோளம், 60உ00./4/, 0-0 முதலிய செயற்கை நுண்ணறிவு இணையத்தளங்கள் தமிழ்ப் படைப்பாக்கத்திற்குத் துணையாக நிற்கின்றன. கற்றல், கற்பித்தலுக்கு அவற்றைப் பயன்படுத்தி தமிழ் வகுப்பை உயிரோட்டமுள்ளதாக ஆக்கலாம். என்னென்ன வழிகளில் அவற்றைப் பயன்படுத்தலாம், கற்றல், கற்பித்தலை இன்னும் எவ்வாறு செறியூட்டலாம், தமிழ்ப் பண்பாட்டு அம்சங்களைத் துல்லியமாக உருவாக்குவதில் உள்ள சிக்கல்களும் சவால்களும் என்னென்ன என்பதை ஆராய்கிறது இந்தக் கட்டுரை.

செயற்கை நுண்ணறிவு இணையத்தளங்கள்

நூற்றுக்கணக்கான செயற்கை நுண்ணறிவு இணையத்தளங்கள் தற்போது இருந்தாலும் படைப்பாக்கத்திற்குப் பரவலாகப் பயன்படுத்தப்படும் மூன்று இணையத்தளங்கள் மட்டுமே இந்த ஆய்வில் உட்படுத்தியுள்ளோம். அவை முறையே சேலம், 60800./4/, டிட் என்பன ஆகும்.

ChatGPT

OpenAI நிறுவனத்தால் உருவாக்கப்பட்டது ChatGPT. Gu@molwom words ampona அடிப்படையாகக் கொண்டு உருவாக்கும் திறன் சார்ந்த செயற்கை நுண்ணறிவுத் (உோஎ2ல1//6 AI) தொழிலுட்பத்தை அது கொண்டுள்ளது. பயனர்கள் தங்கள் கேள்விக் கணைகளை சேல 1- இடம் தொடுக்கலாம். ரோல் தனக்குப் பயிற்சியளிக்கப்பட்ட தரவிடம் கலந்தாலோசித்து தனித்துவமான பதிவை வழங்குகிறது. இன்று சிறுவர் முதல் பெரியோர் வரை சோல் பரவலாகப் பயன்படுத்தப்படுகிறது. பள்ளி ஒப்படைப்புகளைச் செய்தவதிலிருந்து மின்னஞ்சலுக்குத் தகுத்த பதிவைச் செம்மையாக எழுதுவது வரையிலும் சோல்மே அதிகமாக உபயோகிக்கப்படுகிறது எனலாம்.

ChatGPT 3.5, சோல்மேர 4 என இரண்டு பதிப்புகள் ஈஎ8105) உள்ளன. சோல்] 3.5, 175 பிலலியன் அளவுருக்களைக் (8எ5) கொண்டுள்ளது. ஒப்புநோக்க, ChatGPT 4, 1.76 டிரிலியன் அளவுருக்களைக் கொண்டுள்ளதாக நம்பப்படுகிறது. அளவுருக்களின் எண்ணிக்கை அதிகமாக இருக்கும்போது, ோல்்1-இன் பதில்களின் துல்லியமும் நம்பகத்தன்மையும் சிறப்பாக இருக்கும். அதோடு சோல்மோ 4 எழுத்துகள் (ல) உட்பட நிழற்படங்கள், ஒளிக்காட் சிகள், குரல்பதிவுகள் முதலியவற்றைப் பகுத்தாய்ந்து பதில்களைப் பயனர்கள் வழங்கமுடியும். பதில்களை படங்களாகவும்

ஓவியங்களாகவும் பெறலாம். ChatGPT 3.5-இல் எழுத்துகளை மட்டுமே பகுத்தாய்ந்து எழுத்துபூர்வ பதில்களை மட்டுமே அளிக்க இயலும். ChatGPT 3.5 பயனர்கள் இலவசமாகப் பயன்படுத்தலாம். ஆனால் ChatGPT 4-ஐ பயன்படுத்த, மாதந்தோறும் 20 அமெரிக்க டாலர் செலுத்தேவண்டும்.

தமிழ்ப் பைடப்பாக்கத்திற்கு ChatGPT 4 பெரிய அளவில் பயன்படுத்த முடியும். ஆனால் அதற்கு உரிய தூண்டல் கேள்விகளை (Prompt Questions) முறையாகக் கேட்க வேண்டும். இல்லாவிடில், எதிர்பார்க்கும் பதில் கிடைக்காது. உதாரணத்திற்கு சிங்கப்பூரின் கரையோரப் பூந்தோட்டத்தின் படத்தை உருவாக்கேவண்டும்.

ChatGPT-இடம் 'கரையோரப் பூந்தோட்டத்தை உருவாக்குக' எனத் தமிழில் உள்ளீடு செய்தால், ஒரு கரையோரம் அமைந்துள்ள பூந்தோட்டத்தின் படத்தை அது உருவாக்கித் தருகிறது. ஆனால், பயனர் எதிர்பார்ப்பது அதுவன்று. எனவே கேள்வியை இன்னும் செம்மைப்படுத்தி 'சிங்கப்பூரின் கரையோரப் பூந்தோட்ட பின்னணியில் ஒரு படத்தை உருவாக்குக' என உள்ளீடு செய்யும்போது, பயனர் எதிர்பார்ப்பிற்கு நெருங்கி படம் உருவாக்கப்பட்டுள்ளதைப் படம் 1-இல் காண முடிகிறது. இதில் கரையோரப் பூந்தோட்டம் என்பது சிங்கப்பூர் வட்டார வழக்குச் சொல். 'Gardens by the Bay' என்பதன் தமிழ்ப் பதம் தான் கரையோரப் பூந்தோட்டம். வட்டார வழக்குச் சொற்களை உள்ளீடு செய்யும்போது, நாட்டின் பெயரையோ வட்டாரத்தின் பெயரையோ குறிப்பிடுவது முக்கியம். அப்போது தான், பயனர் மனத்தில் எண்ணியுள்ள ஒன்றுக்கு நெருக்கமாக ChatGPT-இன் பதில் அமையும்.



கேள்வி: கரையோரப் பூந்தோட்டத்தை உருவாக்குக



கேள்வி: சிங்கப்பூரின் கரையோரப் பூந்தோட்ட பின்னணியில் ஒரு படத்தை உருவாக்குக

படம் 1: தொடுக்கப்பட்ட இரண்டு வினாக்களுக்கு ChatGPT 4 உருவாக்கிய படங்கள்

தமிழர்ப் பண்பாட்டு அம்சங்களைப் பறைசாற்றும் படங்களை உருவாக்குவதற்கு ChatGPT 4-Queor உதவியை நாடலாம். படம் 2-இல் "சிங்கப்பூர் சூழலில் பொங்கல் கொண்டாட்டங்கள்" எனத் தூண்டல் கேள்வியின் ஊடாகப் படத்தை உருவாக்க சொல்லும்போது, படத்தில் உள்ள பிழையைப் பார்க்கலாம். சிங்கப்பூர்ச் சூழலைச் செயற்கை நுண்ணறிவு சரியாகப் புரிந்திருந்தாலும் நடுவில் கூம்புபோன்ற வடிவில் இருப்பது பொங்கல் பாணையே இல்லை. பொங்கலை விளக்குவதற்கு அப்படத்தைப் பயன்படுத்தினால் அது பண்பாட்டுப் பிழையாகிவிடும். தூண்டல் கேள்வியைச் சற்று சீர்படுத்தி, 'சிங்கப்பூரின் நகர மையப் பகுதியில், பொங்கல் பாணையை நடுவில் வைத்து, தமிழர்ப் பாரம்பரிய உடையில் மக்கள் கூடி நிற்கும் வகையில் ஒரு படத்தை உருவாக்குக' எனக் கேட்டால், சோளம் உருவாக்கியுள்ள படத்தில் எந்தப் பண்பாட்டுப் பிழையும் காண இயலாது. வேட்டி, சேலை மக்கள் அணிந்தவாறு சிங்கப்பூரின் நகர மையப் பகுதியில் பொங்கல் கொண்டாட்டங்கள் மகிழ்ச்சி பொங்க இடம்பெறுவதைப் படம் புதுமையாகவும் பண்பாட்டு அம்சங்களைப் பிறழாமலும் காட்டுகிறது.



கேள்வி: சிங்கப்பூர் சூழலில் பொங்கல் கொண்டாட்டங்கள்



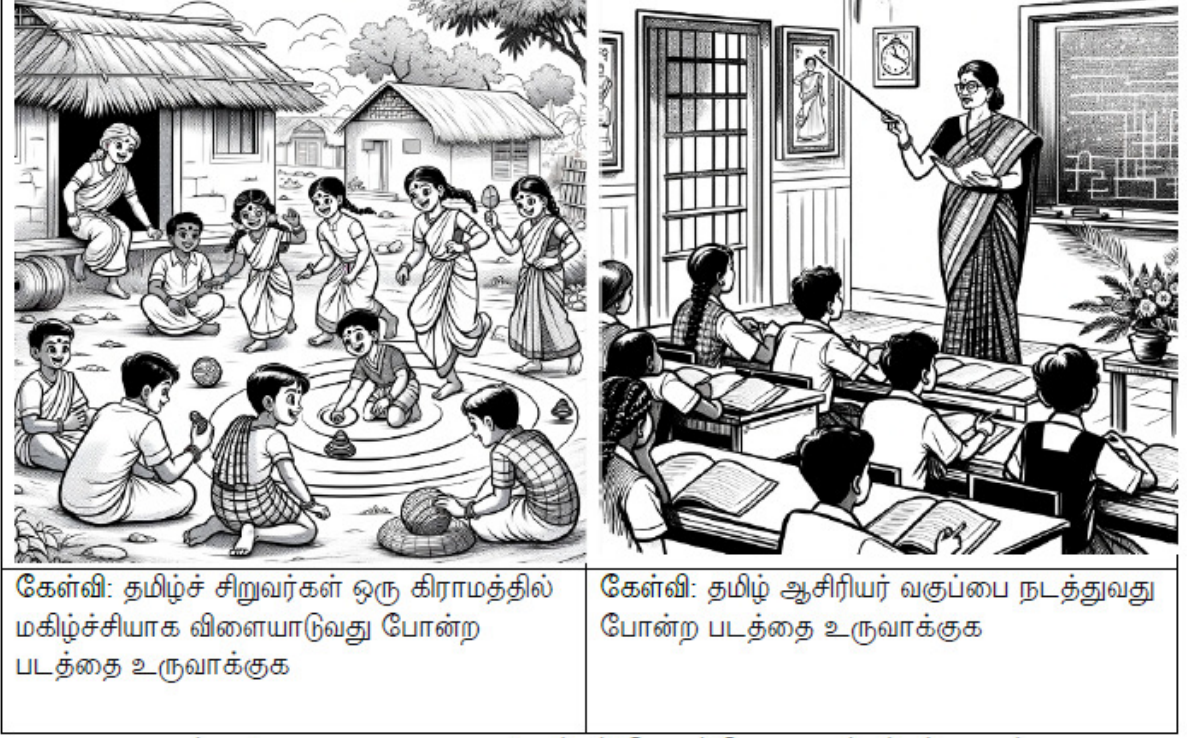
கேள்வி: சிங்கப்பூரின் நகர மையப் பகுதியில், பொங்கல் பாணையை நடுவில் வைத்து, தமிழர்ப் பாரம்பரிய உடையில் மக்கள் கூடி நிற்கும் வகையில் ஒரு படத்தை உருவாக்குக

படம் 2: சிங்கப்பூரில் பொங்கல் கொண்டாட்டங்கள் தொடர்பில் ChatGPT உருவாக்கிய படங்கள்

ChatGPT-இன் கைவண்ணத்தில் உருவாக்கப்படும் படங்களைக் கொண்டு, ஆசிரியர்களும் கல்வியாளர்களும் தங்கள் மாணவர்களுக்குரிய கற்றல் வளங்களை உருவாக்கமுடியும். ஓர் ஓவியரின் உதவி இல்லாமலும் ஒரு வடிவமைப்பாளர் நாடாமலும் ஆசிரியர்கள் வகுப்பறைகளில் பயன்படுத்தவதற்கான வளங்களை உருவாக்க செயற்கை நுண்ணறிவு வழிவகுத்துள்ளது.

ChatGPT 4-இல் 'Colouring Book Hero' என்ற அம்சம் உள்ளது. அதைனக் கொண்டு வண்ணம் தீட்டுவதற்கான படங்களையும் ஓவியங்களையும் உருவாக்கேவண்டும். அதைனக் கொண்டு பாலர்ப்பள்ளி ஆசிரியர்கள் தங்கள் மாணவர்களுக்குரிய கற்றல் வளங்களை உருவாக்கேவண்டும்.

படம் 3-இல் இருப்பது படங்களை உருவாக்கி, மாணவர்களை அவற்றுக்கு வண்ணம் தீட்ட சொல்லலாம்.



கேள்வி: தமிழ்ச் சிறுவர்கள் ஒரு கிராமத்தில் மகிழ்ச்சியாக விளையாடுவது போன்ற படத்தை உருவாக்குக

கேள்வி: தமிழ் ஆசிரியர் வகுப்பை நடத்துவது போன்ற படத்தை உருவாக்குக

படம் 3: 'Colouring Book Hero' அம்சம் கொண்டு உருவாக்கப்பட்ட படம்

Leonardo.Ai

உயர்தரமான படங்களைச் செயற்கை நுண்ணறிவின் மூலம் உருவாக்கும் இணையத்தளம் Leonardo.Ai. வார்த்தைகளை உள்ளீடு செய்தால் உயர்தரமான படங்களை அது உருவாக்கித் தந்துவிடும். ஒவ்வொரு தூண்டல் கேள்விக்கு எட்டு வெவ்வேறு படங்களை வைர உருவாக்கித் தருவது இதன் சிறப்பு அம்சம். அதோடு ஒவ்வொரு நாளும் குறிப்பிட்ட அளவிலான படங்களை இலவசமாகப் பயனர்கள் உருவாக்கலாம். ஆனால் தமிழில் வார்த்தைகளை உள்ளீடு செய்வதற்குப் பதிலாக ஆங்கிலத்தில் உள்ளீடு செய்தால் படங்கள் எதிர்பார்ப்பிற்கு நெருக்கமாய் உள்ளத்தை உணர முடியும். படம் 4-ஐ நோக்கினால் 'தமிழ்ப் பாரம்பரிய விளையாட்டுகளை விளையாட்டும் சிறுவர்களை உருவாக்குக' என்று வினவியேபாது, Leonardo.Ai உருவாக்கிய படங்கள் பயனர்களின் எதிர்பார்ப்பிற்கு அப்பால் இருப்பதைக் காணலாம். இரயில் தொடர்பான படங்களை அது உருவாக்கியுள்ளது வியப்புக்குரியது. அதே வினாவை ஆங்கிலத்தில் தொடடுத்தால் (படம் 5) பயனர்களின் மனத்திரையில் தோன்றியதைக் கணினித்திரையில் ஓரளவிற்குக் கொண்டு வரலாம். படம் 5-இல் பண்டைக் காலத்தில் தமிழ்ச் சிறுவர்கள் பாரம்பரிய விளையாட்டும் காட்சியைப் பழைம மாறாமல் புதுமையாக Leonardo.Ai பைறசாற்றியுள்ளது.



படம் 4: தமிழ் உள்ளீடு மூலம் Leonardo.Ai உருவாக்கிய நான்கு படங்கள்



படம் 5: ஆங்கில உள்ளீடு மூலம் | (உ01800./4/ உருவாக்கிய நான்கு படங்கள்

படங்களைப் பேசும் காணொளிகளாக உருவாக்க உதவுகிறது 0-0 இணையத்தளம். அதற்கு முதலில் நாம் ஒரு படைப்பாளரை அல்லது ஒரு நபரை செயற்கை நுண்ணறிவின் மூலம் உருவாக்க வேண்டும். அதற்கு ரோல்மே1, (600800.// முதலிய இணையத்தளங்களைப் பயன்படுத்தலாம். அதன் பின்னர், 0-0 இணையத்தளத்தின் மூலம் படைப்பாளர் பேச வேண்டிய வசனத்தைத் தமிழில் உள்ளீடு செய்யவேண்டும். தமிழ், தமிழ் (இந்தியா), தமிழ் (மலேசியா), தமிழ் (சிங்கப்பூர்) என நான்கு தெரிவுகளைப் பயனர்கள் (படம் 6) தேர்ந்தெடுக்கலாம். ஒவ்வொரு தெரிவுக்கும் ஆண், பெண் குரல்கள் உள்ளன. ஒவ்வொரு நாட்டில் பேசப்படும் தமிழ்மொழியின் பாணியும் உச்சரிப்பும் வேறுபடுகிறது. அதற்கு ஏற்ப பயனர்கள் தங்களுக்கு விருப்பமான தெரிவைத் தேர்ந்தெடுக்கலாம். அதுமட்டுமின்றி, பயனர்கள் பதிவுசெய்யப்பட்ட குரலைப் பதிவேற்றலாம். அதற்கு ஏற்ப படைப்பாளரின் முகப்பாவனையும் முக அசைவும் இடம்பெறும்.



படம் 6: 010 இணையத்தளத்தில் உள்ள நான்கு தமிழ் தெரிவுகள்

'தமிழ்' தெரிவைப் பயன்படுத்தி | தமிழ் (இந்தியா)' தெரிவைப் பயன்படுத்தி உருவாக்கப்பட்ட காணொளி உருவாக்கப்பட்ட காணொளி



'தமிழ்' தெரிவைப் பயன்படுத்தி உருவாக்கப்பட்ட காணொளி



'தமிழ் (இந்தியா)' தெரிவைப் பயன்படுத்தி உருவாக்கப்பட்ட காணொளி



படம் 7: D-ID இணையத்தளத்தில் வெவ்வேறு தமிழ்த் தெரிவுகளைக் கொண்டு உருவாக்கப்பட்ட காணொளிகள்

கற்றல், கற்பித்தலுக்கு எவ்வாறு பயன்படுத்தலாம்?

D-ID இணையத்தளத்தைக் கொண்டு தமிழ் விளக்க காணொளித் தொடர்களைத் தயாரிக்கலாம். ஆசிரியர் வகுப்பறையில் பாடம் நடத்தி பழகிய மாணவர்களுக்கு செயற்கை நுண்ணறிவு ஆசிரியரின் வருகை புரிப்பையும் புத்துணர்ச்சியும் அளிக்கக்கூடும். அதோடு மாணவர்கள் நூல் அறிமுகம், நூல் கண்ணோட்டம் ஆகியவற்றை எழுதுவதற்குப் பதிலாக இதுபோன்ற இணையத்தளங்களைப் பயன்படுத்திக் காணொளிகளையும் ஒளிக்காட்சிகளையும் தயாரிக்கலாம். தமிழ்க் கற்றலும் கற்பித்தல் அடுத்த கட்டத்திற்குக் கொண்டு செல்ல அது வித்திடும். தமிழில் உள்ளீடு செய்வது சிறந்ததா அல்லது ஆங்கிலத்தில் உள்ளீடு செய்வது சிறந்ததா என்பது பயனர்கள் பயன்படுத்தும் இணையத்தளத்தைப் பொருத்தது.

முடிவுரை

செயற்கை நுண்ணறிவு தமிழ்ப் படைப்பாக்கத்திற்கு பல வாய்ப்புகளையும் வழங்கினாலும் அது செல்லவேண்டிய தூரம் நிறைய உள்ளது. பெருமொழி மாதிரி வரைவில் தமிழ்மொழி தரவுகளையும் தமிழர்ப் பண்பாட்டு அம்சங்களையும் பெரும் அளவில் சேர்க்கவேண்டியது காலத்தின் கட்டாயம். இருக்கும் தரவுகளைச் செப்பனிடும் பணியும் முக்கியம். “சட்டியில் இருந்தால்தான் அகப்பையில் வரும்” என்பது போலத் தரமான தரவுகளும் சரியான தகவல்களும் பெருமொழி மாதிரி வரைவில் இருந்தால்தான் தமிழ்ப் படைப்பாக்கத்திற்குச் செயற்கை நுண்ணறிவு செம்மையாக வருங்காலத்தில் துணைப்புரிய முடியும்.

மேற்கோள்

1. <https://www.coursera.org/articles/chat-gpt-3-vs-4#:~:text=ChatGPT%2D4%20has%20more%20mMemory%20than%20GP1%2D3.5.&text=In%20comparison%2C%20GPT%2D4%20has, text%20it%20can%20process%20simultaneously>
2. <https://chat.openai.com/>
3. <https://app.leonardo.ai/ai-generations>
4. <https://www.d-id.com/>

தமிழ்நாட்டுப் பல்கலைக்கழகங்களால் தமிழ் மொழியில் ஷோத்தங்காவிற்கு மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளின் பங்களிப்புகள்: ஓர் ஆய்வு

முனைவர் சு. கோதைநாயகி

உதவி பல்கலைக்கழக நூலகர்
பல்கலைக்கழக நூலகம், அண்ணா பல்கலைக்கழகம்
சென்னை – 25
kothai.suresh@gmail.com

ஆய்வு சுருக்கம்

இந்திய ஆராய்ச்சி நிறுவனங்கள் மற்றும் பல்கலைக்கழகங்கள் அறிவு பரப்பின் உருவாக்கத்திற்கும், பரப்புதலுக்கும் முக்கிய பங்களிப்பை வழங்குவதோடு, ஒவ்வொரு ஆண்டும் அதிக எண்ணிக்கையிலான ஆய்வறிக்கைகளை உருவாக்குகின்றன. முனைவர் பட்டத்திற்காக பரிந்துரைக்கப்படும் இத்தகைய ஆய்வறிக்கைகளை அச்சு பதிப்பிற்கு கூடுதலாக, மின்னணு பதிப்பின் மூலமாகவும் சமர்ப்பிக்க இந்தியாவின் பல்கலைக்கழக மானியக் குழு அறிக்கை வெளியிட்டது. அதனால், மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக்கட்டுரைகள் (ETDs) என்று பெயரிடப்பட்டது.

"ஷோத்தங்கா" என்பது INFLIBNET (Information & Library Network) மையத்தால் அமைக்கப்பட்ட, இந்திய மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளின், திறந்த அணுக்க, இலக்கமுறை களஞ்சியத்தைக் குறிக்க உருவாக்கப்பட்டது. "ஷோத்" என்ற சொல் சமஸ்கிருதத்திலிருந்து உருவானது.

தமிழ்நாட்டுப் பல்கலைக்கழகங்களால் தமிழ் மொழியில் ஷோத்தங்காவிற்கு மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளின் பங்களிப்பை பற்றி இக் கட்டுரையில் காண்போம்.

முன்னுரை

இந்திய ஆராய்ச்சி நிறுவனங்கள் மற்றும் பல்கலைக்கழகங்கள் அறிவுப்பரப்பின் உருவாக்கத்திற்கும், பரப்புதலுக்கும் முக்கிய பங்களிப்பை வழங்குகின்றன. இவை ஒவ்வொரு ஆண்டும் அதிக எண்ணிக்கையிலான ஆய்வறிக்கைகளை உருவாக்குகின்றன. முனைவர் பட்டத்திற்காக பரிந்துரைக்கப்படும் இத்தகைய ஆய்வறிக்கைகள், ஆய்வு தகவல்களைப் பகிரும் ஒரு தனித்துவமான, உள்ளார்ந்த சிறப்புடைய தகவல் வளமாகும். கல்வியாளர்கள் மற்றும் ஆய்வுக் குழுமங்கள், தங்களின் துறைகளில் எந்த அளவில் ஆராய்ச்சி முன்றேற்றம் அடைந்துள்ளது என்பதை இத்தகைய ஆய்வறிக்கைகள் மற்றும் ஆய்வுக்கட்டுரைகள் மூலம் அறிந்துக் கொள்கிறார்கள்.

The Networked Digital Library of Theses and Dissertations (NDLTD) என்பது மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக்கட்டுரைகளை (ETDs) தனதாக்கல், உருவாக்கம், பயன்பாடு, பரப்புதல் மற்றும் பாதுகாத்தல் ஆகியவற்றை மேம்படுத்தும் ஒரு சர்வதேச அமைப்பாகும். அதைப் போன்று, இந்தியாவில் பல்கலைக்கழக மானியக் குழுவின் முயற்சியால் மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக்கட்டுரைகளின் (Electronic Theses and Dissertations - ETDs) களஞ்சியம் உருவாக்கப்பட்டது. ஆய்வுகளின் நகலெடுப்பையும் வறியத் தெரிவுநிலையையும், அறியா காரணியாக ஆராய்ச்சி வெளியீட்டிலுள்ள குறைந்த தரத்தையும் இந்த ETDs எதிர்கொண்டு வெல்லும்.

ஷோத்தங்கா உருவாக்கம்

"ஷோத்கங்கா" என்பது INFLIBNET (Information & Library Network) மையத்தால் அமைக்கப்பட்ட, இந்திய மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளின், திறந்த அணுக்க, இலக்கமுறை களஞ்சியத்தைக் குறிக்க உருவாக்கப்பட்டது. "ஷோத்" என்ற சொல் சமஸ்கிருதத்திலிருந்து உருவானது. ஆராய்ச்சி மற்றும் கண்டுபிடிப்பைக் குறிக்கிறது. இந்திய துணைக்கண்டத்தில் உள்ள அனைத்து நதிகளிலும் "கங்கை" புனிதமானது, பெரியது மற்றும் நீளமானது. கங்கை இந்தியாவின் நீண்ட கால கலாச்சாரம் மற்றும் நாகரீகத்தின் சின்னமாக உள்ளது, ஷோத்கங்கா என்பது INFLIBNET மையத்தால் பராமரிக்கப்படும் இந்திய அறிவுசார் வெளியீட்டின் களஞ்சியமாகவும், தேக்கிடமாகவும், கருதப் படுக்கிறது. ஆராய்ச்சியாளர்களுக்கு கிடைத்த வரமாகவும் விளங்குகிறது.

மாசாச்சுசெட்சு தொழில்நுட்பக்கழகம் (MIT) மற்றும் ஹெவ்லட்-பேக்கர்ட் (HP) இடையே கூட்டாக உருவாக்கப்பட்ட DSpace என்னும் திறந்த மூல இலக்கமுறை களஞ்சிய மென்பொருளைப் பயன்படுத்தி ஷோத்கங்கா@INFLIBNET அமைக்கப்பட்டுள்ளது. ஆராய்ச்சியாளர்களால் சமர்ப்பிக்கப்பட்ட மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளைப் பெறுதல், குறியிடுதல், சேமித்தல், பரப்புதல், மற்றும் பாதுகாத்தல் போன்றவற்றை இக்களஞ்சியத்தால் செய்ய இயலும்.

ஷோத்கங்கா செயல்நோக்கம்

INFLIBNET மையத்துடன் புரிந்துணர்வு ஒப்பந்தம் செய்யும் பல்கலைக்கழகங்கள், மின்னணு பதிப்பை சமர்ப்பிப்பதால், தங்களது ஆய்வறிக்கைகளின் பழங்கோப்புகளை இலக்கமுறையாக மாற்றுவதற்கு, பல்கலைக்கழக மானியக் குழுவிடமிருந்து நிதியுதவி பெறலாம். அது மட்டுமன்றி, கருத்துக்களவை கண்டறிய உதவும் மென்பொருள் கருவிகளுக்கென உள்ள சந்தவிற்கான நிதியுதவியையும் பெற முடியும். புரிந்துணர்வு ஒப்பந்தம் செய்யும் பல்கலைக்கழகங்களுக்கு கருத்துக்கண்டறியும் மென்பொருளை இலவசமாக அணுகவும் வழி செய்யப்படுகிறது.

திறந்த அணுக்க களஞ்சியங்களால் ஆய்வறிக்கைகள் மற்றும் ஆய்வுக்கட்டுரைகளுக்கு மிகுதியான தெரிவுநிலை ஏற்பட்டுள்ளதால், இவை கருத்துக்களவு தடுப்பாக செயல்படும்.

ஆய்வின் நோக்கங்கள்

1. தமிழ்நாட்டில் உள்ள 22 மாநில பல்கலைக்கழகங்களால், ஷோத்கங்கா திட்டத்தில் மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளின் (ETDs) பங்களிப்புகளை பகுப்பாய்வு செய்தல்
2. தமிழ்நாட்டில் உள்ள 28 நிகர்நிலைப் பல்கலைக்கழகங்களால், ஷோத்கங்கா திட்டத்தில் மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளின் (ETDs) பங்களிப்புகளை பகுப்பாய்வு செய்தல்
3. தமிழ்நாட்டுப் பல்கலைக்கழகங்களால் தமிழ் துறையிலிருந்து தமிழ் மொழியில் ஷோத்கங்காவிற்கு மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளின் பங்களிப்புகள் எந்த ஆண்டு முதல் தொடங்கியது என்பது குறித்து ஆய்வு.

ஆய்வு முறையியல்

Shodhganga @ INFLIBNET இணையத்தளத்திலிருந்து (<https://shodhganga.inflibnet.ac.in/>) எடுக்கப்பட்ட தரவைப் பகுப்பாய்வதற்காக விவரண ஆய்வு (Descriptive Research) பயன்படுத்தப்பட்டுள்ளது.

தரவு திரட்டல் மற்றும் பகுப்பாய்வு

1. ஆய்வுக்கான தரவு மற்றும் புள்ளி விவரங்கள் ஷோத்கங்கா திட்டத்தில் (<https://shodhganga.inflibnet.ac.in/>) இருந்து எடுக்கப்பட்டது.

2. தரவு மூன்று முக்கிய வகைகளாக தொகுக்கப்பட்டது
அ. பல்கலைக்கழகங்களின் ஒட்டு மொத்த பங்களிப்புகள்
ஆ. பல்கலைக்கழகங்களில் உள்ள தமிழ் துறையிருந்து தமிழ் மொழியில் சமர்ப்பிக்கப்பட்ட ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகள்
இ. எந்த ஆண்டு முதல் ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகள் சமர்ப்பிக்கப்பட்டன
3. பிரித்தெடுக்கப்பட்டத் தரவுப் பகுப்பாய்வுக்காக Ms-Excel க்கு ஏற்றுமதி செய்யப்பட்டது.

ஷோத்கங்கா இணையதளத்தின் திரைப்பிடிப்பு



ஷோத்கங்காவில் பல்கலைக்கழகங்களின் பங்களிப்பு

இந்திய அளவில், ஷோத்கங்காவில் பங்களிக்கும் 10 முதன்மை பல்கலைக்கழகங்களும், அவை பதிவேற்றிய ஆய்வறிக்கைகளின் எண்ணிக்கைகளும், அட்டவணை-1இல் கொடுக்கப்பட்டுள்ளது.

அட்டவணை -1: ஷோத்கங்காவில் பங்களிக்கும் 10 முதன்மை பல்கலைக்கழகங்கள்

வரிசை எண்	பல்கலைக்கழகம்	பதிவேற்றிய ஆய்வறிக்கைகளின் எண்ணிக்கை
1.	Anna University	15906
2.	University of Madras	14874
3.	University of Calcutta	14185
4.	Savitribai Phule Pune University	12547
5.	University of Mumbai	10968
6.	Aligarh Muslim University	10194
7.	Chhatrapati Shivaji Maharaj University	10175
8.	Andhra University	9868
9.	Babasaheb Bhimrao Ambedkar Bihar University	9675
10.	Panjab University	9260

(14.03.2024 அன்று வரை)

மேற்கூறிய அட்டவணை-1இன் படி, அண்ணா பல்கலைக்கழகம் அதிகப்படியாக 15906 ஆய்வறிக்கைகளை பதிவேற்றம் செய்து முதல் இடத்தைப் பிடித்துள்ளது. சென்னைப்

பல்கலைக்கழகம் 14874 ஆய்வறிக்கைகளும், கொல்கத்தா பல்கலைக்கழகம் 14185 ஆய்வறிக்கைகளும் பதிவேற்றம் செய்து முறையே இரண்டாவது மற்றும் மூன்றாம் இடங்களை பிடித்துள்ளது.

அட்டவணை -2: தமிழ்நாடு மாநில பல்கலைக்கழகங்களின் ஆய்வறிக்கைகள் பதிவேற்றத்தின் தரவரிசை

வரிசை எண்	பல்கலைக்கழகம்	பதிவேற்றிய ஆய்வறிக்கைகளின் எண்ணிக்கை	%	தரவரிசை
1.	Anna University	15906	22.03	1
2.	University of Madras	14874	20.60	2
3.	Bharathidasan University	8300	11.50	3
4.	Bharathiar University	7393	10.24	4
5.	Manonmaniam Sundaranar University	7341	10.17	5
6.	Madurai Kamaraj University	5820	8.06	6
7.	Periyar University	3145	4.36	7
8.	Annamalai University	3038	4.21	8
9.	Alagappa University	2443	3.38	9
10.	Tamil Nadu Agricultural University	1409	1.95	10
11.	Mother Teresa Women's University	1067	1.48	11
12.	The Tamil Nadu Dr.MGR Medical University	390	0.54	12
13.	Thiruvalluvar University	289	0.40	13
14.	Tamil Nadu Physical Education and Sports University	248	0.34	14
15.	Tamil Nadu Teachers Education University	203	0.28	15
16.	Tamil University	161	0.22	16
17.	Tamil Nadu Veterinary and Animal Sciences University	59	0.08	17
18.	The Tamil Nadu Dr. Ambedkar Law University	59	0.08	17
19.	Tamil Nadu Open University	47	0.07	19
20.	Tamil Nadu Dr.J.Jayalalithaa Fisheries University	0	0.00	20
21.	Tamil Nadu National Law University	0	0.00	21
22.	The Tamil Nadu Dr.J. Jayalalithaa Music and Fine Arts University	0	0.00	22
Total		72192	100	

அட்டவணை-2இன் மூலமாக அண்ணா பல்கலைக்கழகம் 15906 ஆய்வறிக்கைகளை பதிவேற்றம் செய்து முதல் இடத்தையும், சென்னைப் பல்கலைக்கழகம் 14874 ஆய்வறிக்கைகளை பதிவேற்றம் செய்து இரண்டாவது இடத்தையும், பாரதிதாசன் பல்கலைக்கழகம் 8300 ஆய்வறிக்கைகளை பதிவேற்றம் செய்து மூன்றாவது இடத்தையும் பெற்றுள்ளதை அறிய முடிகிறது

.அட்டவணை - 3: தமிழ்நாடு மாநில பல்கலைக்கழகங்களால் தமிழ் மொழியில் ஆய்வறிக்கைகள் பதிவேற்றத்தின் தரவரிசை

வரிசை எண்	பல்கலைக்கழகம்	தமிழ் மொழியில் பதிவேற்றிய ஆய்வறிக்கைகளின் எண்ணிக்கை	%	தரவரிசை
1.	University of Madras	1684	32.83	1
2.	Bharathidasan University	1120	21.83	2
3.	Manonmaniam Sundaranar University	532	10.37	3
4.	Madurai Kamaraj University	504	9.82	4
5.	Periyar University	499	9.73	5
6.	Bharathiar University	399	7.78	6
7.	Mother Teresa Women's University	99	1.93	7
8.	Alagappa University	182	3.55	8
9.	Annamalai University	59	1.15	9
10.	Thiruvalluvar University	48	0.94	10
11.	Tamil Nadu Open University	4	0.08	11
12.	Anna University	0	0.00	12
13.	Tamil Nadu Agricultural University	0	0.00	13
14.	Tamil Nadu Dr.J.Jayalalithaa Fisheries University	0	0.00	14
15.	Tamil Nadu National Law University	0	0.00	15
16.	Tamil Nadu Physical Education and Sports University	0	0.00	16
17.	Tamil Nadu Teachers Education University	0	0.00	17
18.	Tamil Nadu Veterinary and Animal Sciences University	0	0.00	18
19.	Tamil University	0	0.00	19
20.	The Tamil Nadu Dr.Ambedkar Law University	0	0.00	20
21.	The Tamil Nadu Dr.J. Jayalalithaa Music and Fine Arts University	0	0.00	21
22.	The Tamil Nadu Dr.MGR Medical University	0	0.00	22
Total		5130	100	

இதன் மூலம், தமிழ்நாடு மாநில பல்கலைக்கழகங்களிலிருந்து தமிழ் மொழியில் சமர்ப்பிக்கப்பட்ட ஆய்வறிக்கைகளின் பதிவேற்ற தரவரிசையின் படி, சென்னைப் பல்கலைக்கழகம், பாரதிதாசன் பல்கலைக்கழகம், மனோன்மணியம் சுந்தரனார் பல்கலைக்கழகம் ஆகியவை முறையே 32.83 சதவிகிதமும், 21.83 சதவிகிதமும், 10.37 சதவிகிதமும் பெற்று முதல், இரண்டாம் மற்றும் மூன்றாம் இடங்களைப் பெற்றுள்ளன என்பதை அறியலாம்.

அட்டவணை - 4: தமிழ்நாடு மாநில நிகர்நிலைப் பல்கலைக்கழகங்களின் ஆய்வறிக்கைகள் பதிவேற்றத்தின் தரவரிசை

வரிசை எண்	பல்கலைக்கழகம்	பதிவேற்றிய ஆய்வறிக்கைகளின் எண்ணிக்கை	%	தரவரிசை
1.	Vellore Institute of Technology	2247	15.58	1

2.	SRM Institute of Science and Technology	1503	10.42	2
3.	The Gandhigram Rural Institute	1353	9.38	3
4.	SASTRA Deemed University	813	5.64	4
5.	Vels University	785	5.44	5
6.	Avinashilingam Institute for Home Science and Higher Education for Women	706	4.89	6
7.	Bharath Institute of Higher Education and Research	651	4.51	7
8.	Amrita Viswa Vidyapeetham	596	4.13	8
9.	Karunya Institute of Technology and Sciences	582	4.03	9
10.	Sathyabama Institute of Science and Technology	553	3.83	10
11.	Dr. MGR Educational and Research Institute	478	3.31	11
12.	Saveetha Institute of Medical and Technical Sciences	422	2.93	12
13.	Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya	417	2.89	13
14.	Kalasalingam Academy of Research and Education	384	2.66	14
15.	Vinayaka Mission's Research Foundation	361	2.50	15
16.	Sri Ramachandra Institute of Higher Education and Research	356	2.47	16
17.	Noorul Islam Centre for Higher Education	351	2.43	17
18.	Hindustan Institute of Technology and Science	308	2.14	18
19.	B.S. Abdur Rahman Crescent Institute of Science & Technology	305	2.11	19
20.	Karpagam University	298	2.07	20
21.	St.Peter's Institute of Higher Education and Research	273	1.89	21
22.	Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology	229	1.59	22
23.	Meenakshi Academy of Higher Education and Research	168	1.16	23
24.	Periyar Maniammai Institute of Science & Technology	157	1.09	24
25.	AMET University	130	0.90	25
26.	Chettinad Academy of Research and Education	0	0.00	26
27.	Chennai Mathematical Institute	0	0.00	27

28.	Ponnaiyah Ramalingam Institute of Science and Technology (PRIST)	0	0.00	28
Total		14426	100	

தமிழ்நாடு மாநில நிகர்நிலைப் பல்கலைக்கழகங்களின் ஆய்வறிக்கைகள் பதிவேற்றத்தின் தரவரிசைக் குறிப்பிடுவது யாதெனில், வேலூர் இன்ஸ்டிடியூட் ஆப் டெக்னாலஜி 2247 ஆய்வறிக்கைகள், எஸ்ஆர்எம் இன்ஸ்டிடியூட் ஆப் சயின்ஸ் அண்ட் டெக்னாலஜி 1503 ஆய்வறிக்கைகள், காந்தி கிராமம்- கிராமிய நிகர்நிலைப் பல்கலைக்கழகம் 1353 ஆய்வறிக்கைகளைச் சமர்ப்பித்து முறையே முதலாம், இரண்டாம் மற்றும் மூன்றாம் இடங்களைப் பெற்றுள்ளது.

அட்டவணை - 5: தமிழ்நாடு மாநில நிகர்நிலைப் பல்கலைக்கழகங்களால் தமிழ் மொழியில் ஆய்வறிக்கைகள் பதிவேற்றத்தின் தரவரிசை

வரிசை எண்	பல்கலைக்கழகம்	தமிழ் மொழியில் பதிவேற்றிய ஆய்வறிக்கைகளின் எண்ணிக்கை	%	தரவரிசை
1.	The Gandhigram Rural Institute	93	61.18	1
2.	Avinashilingam Institute for Home Science and Higher Education for Women	35	23.03	2
3.	Vels University	10	6.58	3
4.	Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya	6	3.95	4
5.	Karpagam University	5	3.29	5
6.	SRM Institute of Science and Technology	3	1.97	6
7.	AMET University	0	0.00	0
8.	Amrita Viswa Vidyapeetham	0	0.00	0
9.	B.S. Abdur Rahman Crescent Institute of Science & Technology	0	0.00	0
10.	Bharath Institute of Higher Education and Research	0	0.00	0
11.	Chennai Mathematical Institute	0	0.00	0
12.	Chettinad Academy of Research and Education	0	0.00	0
13.	Dr. MGR Educational and Research Institute	0	0.00	0
14.	Hindustan Institute of Technology and Science	0	0.00	0
15.	Kalasalingam Academy of Research and Education	0	0.00	0
16.	Karunya Institute of Technology and Sciences	0	0.00	0
17.	Meenakshi Academy of Higher Education and Research	0	0.00	0
18.	Noorul Islam Centre for Higher Education	0	0.00	0
19.	Periyar Maniammai Institute of Science & Technology	0	0.00	0
20.	Ponnaiyah Ramalingam Institute of Science and Technology (PRIST)	0	0.00	0
21.	Sathyabama Institute of Science and Technology	0	0.00	0
22.	Saveetha Institute of Medical and Technical Sciences	0	0.00	0

23.	SASTRA Deemed University	0	0.00	0
24.	Sri Ramachandra Institute of Higher Education and Research	0	0.00	0
25.	St.Peter's Institute of Higher Education and Research	0	0.00	0
26.	Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology	0	0.00	0
27.	Vellore Institute of Technology	0	0.00	0
28.	Vinayaka Mission's Research Foundation	0	0.00	0
Total		152	100	

தமிழ்நாடு மாநில நிகர்நிலைப் பல்கலைக்கழகங்களின் தமிழ் மொழியில் ஆய்வறிக்கைகள் பதிவேற்றத்தின் தரவரிசை மூலம், காந்தி கிராமம்- கிராமிய நிகர்நிலைப் பல்கலைக்கழகத்தில் தமிழ் ஆய்வறிக்கைகள் மற்ற நிகர்நிலை பல்கலைக்கழகங்களை விட அதிகம் சமர்ப்பிக்கப்பட்டுள்ளது தெரிகிறது.

அட்டவணை - 6: தமிழ்நாடு மாநில பல்கலைக்கழகங்களால் தமிழ் மொழியில் ஆய்வறிக்கைகள் ஆண்டு வாரியான பதிவேற்றத்தின் தரவரிசை

1984-90	1991-1999	2000-2009	2010-2019	2020-24
16	196	601	1256	576

தமிழ்நாடு மாநில பல்கலைக்கழகங்களால் தமிழ் மொழியில் ஆய்வறிக்கைகள் 2010-2019 க்கு இடைப்பட்ட ஆண்டுகளில் தான் அதிகமாக பதிவேற்றம் செய்யப்பட்டுள்ளது என்பதும், 1984-90 ஆம் ஆண்டுகளில் குறைந்த அளவே ஆய்வறிக்கைகள் பதிவேற்றம் செய்யப்பட்டுள்ளது என்பதும் தெரிகிறது.

அட்டவணை - 7: தமிழ்நாடு மாநில நிகர்நிலைப் பல்கலைக்கழகங்களால் ஆய்வறிக்கைகள் ஆண்டு வாரியான பதிவேற்றத்தின் தரவரிசை

1984-90	1991-1999	2000-2009	2010-2019	2020-24
0	0	9	45	29

இந்த அட்டவணை 7-ன் மூலம், தமிழ்நாடு மாநில நிகர்நிலைப் பல்கலைக்கழகங்களால் தமிழ் மொழியில் ஆய்வறிக்கைகள் 2010-2019 க்கு இடைப்பட்ட ஆண்டுகளில் தான் பதிவேற்றம் செய்யப்பட்டுள்ளது என்பதும், 1984-99 ஆம் ஆண்டுகளில் ஆய்வறிக்கைகள் எதுவும் பதிவேற்றம் செய்யப்படவில்லை என்பதும் தெரிகிறது.

முடிவுரை

உலகம் முழுவதும், சோதகங்கா மிகவும் தேவையான, அதிகம் கோரும் ஒரு களஞ்சியமாக உருவெடுத்துள்ளது என்றால் அது மிகையாகாது. சோதகங்கா ஆய்வியலார்களுக்கு மதிப்பாய்வுகளைச் சேகரிக்க, தொகுக்க உதவி செய்வதோடு மட்டுமல்லாமல், அவர்களின் ஆராய்ச்சித் துறையில் பின்னோக்கு ஆய்வுகளை பற்றி அறிந்துக் கொள்ளவும் உதவுகிறது. ஆராய்ச்சியாளர்களுக்கு தங்கள் துறையிலிருக்கும் பல்வேறு கருப்பொருள்களைப் புரிந்துக் கொள்ள வழிவகை செய்கிறது.

புரிந்துணர்வு ஒப்பந்தம் செய்யாத பல்கலைக்கழகங்கள் சோதகங்காவில் சேருவதால் தங்களின் தரநிலையை உயர்த்துவதோடு மட்டுமல்லாமல், தங்கள் ஆய்வியலாளர்களின் ஆராய்ச்சியை வேறு கட்டத்திற்கு அழைத்து சென்று அவர்களின் வெற்றிக்கும் உறுதுணை புரியலாம்.

இத்துணை சிறப்புமிக்க ETDs பற்றிய விழிப்புணர்வை நாம் ஏற்படுத்திக் கொடுத்தால், மின்னணு ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகளின் அணுக்கம் அதிகமாகும் என்பதில் எவ்வித ஐயமுமில்லை.

பகுப்பாய்வு / ஆய்வு கண்டுபிடிப்புகள் / பரிந்துரைகள்

1. தமிழ்நாட்டு மாநில பல்கலைக்கழகங்களில் அண்ணா பல்கலைக்கழகம் ஆய்வறிக்கைகளின் பதிவேற்றத்தில் முதலிடம் பெற்றுள்ளது.
2. சென்னை பல்கலைக்கழகம் தமிழ் மொழியில் அதிக ஆய்வறிக்கைகளைப் பதிவேற்றியிருக்கிறது.
3. தமிழ் மொழியில் ஆராய்ச்சிகள் அதிகம் மேற்கொள்ளப்படவேண்டும். அதற்கு ஊக்கமும் அளிக்கப்பட வேண்டும்.
4. தமிழ் நாட்டில் உள்ளப் பல்கலைக்கழகங்களில் தமிழ்த்துறை இல்லையெனில், ஆரம்பிக்கப்பட வேண்டும்.
5. ஆங்கிலத்தில் சமர்ப்பிக்கப்படும் ஆய்வறிக்கைகள் தமிழ் மொழியில் மொழிபெயர்த்து சமர்ப்பிக்கப்பட வேண்டும்.
6. தமிழ்நாடு மாநில பல்கலைக்கழகங்களில் 2010 ஆம் ஆண்டிற்கு பின்பு தான் தமிழ் மொழியில் ஆய்வறிக்கைகள் அதிக அளவில் சமர்ப்பிக்கப்பட்டுள்ளது.
7. தமிழ்நாடு மாநில நிகர்நிலைப் பல்கலைக்கழகங்களில் தமிழ் மொழியில் ஆய்வறிக்கைகள் 1984-99 ஆம் ஆண்டுகளில் எதுவும் பதிவேற்றம் செய்யப்படவில்லை.
8. ஷோத்கங்காவில் ஆய்வறிக்கை சமர்ப்பிப்பதால், பல்கலைக்கழகங்களின் தெரிவுநிலை அதிகரிக்கும் (Increase in visibility)
9. பல்கலைக்கழகங்களிடையே கூட்டு ஆய்வு அதிகரிக்கும் (Increase in Collaborative Research)
10. பல்கலைக்கழகங்களின் தோற்றரவு அதிகரிப்பால், அந்தந்த பாடப் பொருள் வல்லுநர்களை இனம் கண்டுக்கொள்வது எளிதாகிறது (Identifying the Experts in the respective subject areas)

மேற்கோள்கள்

1. <https://ndltd.org/>
2. <https://shodhganga.inflibnet.ac.in/>
3. Anbalagan, Muthuraj. (2016). Research output analysis of electronic theses and dissertations with special reference to Shodhganga. Journal of Current Trends in Library and Information Science. 3. 16-20.
4. Jeyapragash, B, Rajkumar, T and Muthuraj, A(2016).Research output analysis of electronic theses and dissertations with special reference to Shodhganga, Journal of Current Trends in Library and Information Science, 3(½), 16-20.
5. Panda, S. (2016), "Shodhganga – a national level open access ETD repository of Indian electronic theses: current status and discussions", Library Hi Tech News, 33(1), pp. 23-26. <https://doi.org/10.1108/LHTN-09-2015-0062>
6. Sheeja, N.K., & Cherukodan, S. (2011). The development and promotion of ETDs in Kerala. Available at <http://ir.inflibnet.ac.in/bitstream/handle/1944/>
7. Sivakumaren, KS and Swaminathan, S (2020). Availability of Electronic Theses and Dissertations (ETDs) of State Universities of Tamil Nadu in INFLIBNET Shodhganga Project: an analysis, Library Philosophy and Practice, article 4113.
8. Subhash Khode (2020). An analysis of contribution of Universities of Madhya Pradesh in Shodhganga, Journal of Indian Library Association, 56(3), July-Sept.

9. Vaishali., & Babasaheb, A. (2014). ETDs in India: Towards a national repository with value added e-theses service. *A Journal of Library and Information Sciences*, 2(1), 92-100.

**இணையத்தில் தமிழ்மொழியின் வளர்ச்சி
முனைவர் பி.ஆர். இலட்சுமி**

**முதுமுனைவர் பட்ட ஆய்வாளர் (வேல்ஸ் பல்கலைக்கழகம்)
என்.வைரமணி, (கணக்கியல் அதிகாரி-சென்னை)**

முன்னுரை

தமிழ்மொழி கி.பி 2010 முதல் சிறப்பாக வளர்ந்து வருகிறது. இருப்பினும், தாய்மொழித் தமிழ் தொடர்பான பல கருத்துகள் ஆய்வு நிலையிலேயே இருக்கிறது. மக்கள் தினமும் பயன்படுத்தும் அறிவியல் கருவிகளில் தமிழ்மொழி அதிகரிக்கவேண்டும். அவ்வாறு அமைந்தால்மட்டுமே தமிழ்மொழியில் பணி வாய்ப்புகள் கிராமப்புற மக்களுக்கு அமையும். தமிழ்மொழியின் பயன்பாடு அதிகரிக்கும்வழி குறித்து இவ்வாய்வுச் சுருக்கம் அமைகிறது.

குறிச்சொற்கள்- பாரதி, நாளிதழ்கள், மின்வணிகம், குறுஞ்செயலிகள், தமிழ்க்கல்வி

இணையத்தில் தமிழ் நாளிதழ்கள்

நாளிதழ் என்பது சமுதாயத்தின் நிகழ்வுகளை வெளிக்காட்டும் கண்ணாடி. கால மாற்றத்தினால் கையினால் எடுத்துப் படிக்க நேரமின்றி மக்கள் பிள்ளைகளின் படிப்புக்கென பொருள் தேடிப் பிழைக்க பல இடங்களுக்குப் புலம் பெயர்ந்து வாழ்ந்து வருகின்றனர். அந்த நேரத்தில் அந்தந்த மாவட்டத்தின் செய்திகளை அறிந்து படிக்க தினமலர் நாளிதழ் உதவி செய்கிறது. முதியோர்கள் எல்லாக் கோவில்களையும் தேடிப் பார்க்க விரும்புவது இயல்பு. அத்தகைய வசதியினையும் தினமலர் நாளிதழ் இலவசமாக அளித்து வருகிறது. தொலைக்காட்சி வடிவத்திலும் செய்திகளை அளித்து வருகிறது. 'செந்தமிழ் நாடெனும் போதினிலே, இன்பத் தேன் வந்து பாயுது காதினிலே' என்பது பாரதியின் வாக்கு. அதற்கேற்ப தினமலர் ஒலி வடிவில் நாளிதழை அளித்து வருகிறது. தினமணி,தி.இந்து, தினகரன்,மாலைமுரசு,தமிழ்முரசு,மக்கள்ஒசை,தினபூமி,முரசொலி போன்ற நாளிதழ்கள் மக்களிடம் செய்திகளை உடனுக்குடன் இணையத்தில் பகிர்ந்து வருகின்றனர். குறுஞ்செயலிகள் வழியாகவும் நாளிதழ்கள் சிறந்த பணி செய்து வருகின்றன. மின் வணிகத்தில் தமிழ்மொழி

வணிகம் என்பது வணிகங்கள் தங்கள் நடவடிக்கைகளை இணைய வழியாக நடத்துவதை உள்ளடக்கியது. மின்னணு சாதனங்கள் வழியாக தமக்குத் தேவையான பொருளை இருந்த இடத்திலிருந்து பெறுவதாகும்.

**வணிகத்திலிருந்து நுகர்வோர்
வணிகத்திலிருந்து வணிகம்**

நுகர்வோர்-வணிகம்

நுகர்வோர்-நுகர்வோர் என்ற பிரிவுகளின் அடிப்படையில் இணையத்தில் மின் வணிகம் நடந்து வருகிறது. மின் வணிகங்கள் தமிழ்மொழியில் முழுவதும் அமைந்து விட்டால் நாட்டில் வறுமை அகலும். தமிழ்மொழி வளரும். காரணம் தமிழ்நாடு விவசாயம் சார்ந்த பகுதி. இம் மாநிலத்தில் பொதுவாக விவசாயிகள் தங்கள் குழந்தைகளை அருகில் இருக்கும் அரசு பள்ளிகளில் கல்வி பயில அனுப்புகின்றனர். அவர்களுக்கு இன்னமும் ஆங்கில வழிக்கல்வி முழுமையாக கிடைக்கப் பெறவில்லை. இதனால், அவர்கள் தாய்மொழி தமிழிலேயே அனைத்து பாடங்களையும் பயின்று வருகின்றனர். பொறியியல்,மருத்துவம் போன்ற படிப்புகள் தாய்மொழியில் இருந்தாலும் இன்னமும் விவசாயிகளுக்கு உயர்படிப்புகள் எட்டாக்கனியாக இருக்கிறது. இந்நிலை மாற வேண்டுமாயின் மின்வணிகங்கள் முழுமையாக தாய்மொழியில் அமைவது சிறப்பானது. அமேசான், ஃபிளிப்கார்ட், பிக்பேஸ்கட் போன்ற அமைப்புகள் சிறப்பாகப் பங்களித்து வருகின்றனர். ஃபேஸ்புக், ட்வீட்டர், இன்ஸ்டாகிராம், டெலிகிராம் போன்றவற்றின்

வழியாகவும் தமது தயாரிப்புப் பொருட்களை மக்கள் அதிக செலவில்லாமல் விற்று தமது வாழ்வாதாரத்தை அமைத்து வருகின்றனர். இதற்கு செல்லிடபேசியில் கூகுள் பே மிகவும் பயனுள்ளதாக அமைந்துள்ளது. கிராம மக்களும் பயன்படுத்தும் வகையில் தமிழ்மொழியில் அமைந்தால் சிறப்பாக இருக்கும். கிராமங்களில் இன்னமும் படிக்கத் தெரியாத மக்கள்வாழ்ந்து வருகின்றனர். ஆனால் அவர்கள் ஆண்ட்ராய்டு செல்லிடபேசியைப் பயன்படுத்தி வருகின்றனர். அவர்கள் பயன்படுத்தும் வகையில் ஒலி அமைப்புகள் அமைக்கப்படவேண்டும்.

இணையத்தில் தமிழ்க்கல்வி

தமிழ்நாட்டில் பல பள்ளிகள் தங்களது நிறுவனம் சார்ந்த தளங்களில் கல்வி தொடர்பான செய்திகளை வெளியிட்டு வருகின்றனர். கல்வி தொடர்பான பல செய்திகளை

ஃபேஸ்புக், ட்விட்டர், வாட்ஸ்அப் போன்றவற்றில் பயன்படுத்தி தமிழ்மொழி வளர உதவி செய்கின்றனர். கூகுள் நிறுவனத்தினரின் விரலினால் எழுதினால் தட்டச்சு கிடைக்கும்

முறையினையும், பேசும் ஒலியின் வழி தட்டச்சு கிடைக்கும் முறையினையும் ஆராய்ச்சியாளர்கள் பின்பற்றி வருகின்றனர். இதனால் தட்டச்சு செய்யும் பணி எளிமையாக இருக்கிறது.

முடிவுரை

ஒவ்வொரு நாட்டிலும் தாய்மொழியில் கல்வி கற்பது சிறப்பானது என்பதை ஆராய்ச்சியாளர்கள் வலியுறுத்தினாலும் தாய்மொழியில் இணையத்தில் இயங்க பல சிக்கல்கள் காணப்படுகின்றன.

அச் சிக்கல்களைத் தீர்க்கும் காரணிகளாக மின்வணிகம், நாளிதழ்கள், தாய்மொழிக்கல்வி போன்றவை அமையும் என்பதை இவ்வாய்வுக்கட்டுரை விளக்குகிறது.

contact

Lakshmi Vairamani <rambharathivel@gmail.com> vels university India

Vairamani Natrajan <vairam61@gmail.com>

<https://vairamani-lakshmi.blogspot.com>

**காணி இன மக்களின் பேச்சு மொழியும் கணினித் தமிழ் வழி
பெற்ற வளர்ச்சி நிலையும் (கன்னியாகுமரி மாவட்டப்
பழங்குடியின மக்களை முன் வைத்து)**

**The spoken language of the Kani people and their Educational Development
Level through Computer Learning" (Presenting the tribal people of
Kanyakumari district)**

**முனைவர் மு.ஜோதிலட்சுமி
இணைப் பேராசிரியர், தமிழாய்வுத்துறை,
பிஷ்ப்ஹீபர் கல்லூரி, திருச்சி - 17, தமிழ்நாடு**

முன்னுரை:-

1912 இல் தமிழக வனத்துறையின் கட்டுப்பாட்டுக்குள் காணி குடியிருப்புகள் வந்தன .பின்பு, மேற்குதக்தொடர்ச்சி மலையில், தேக்கு,சந்தனம்,அகில்,உள்ளிட்ட மதிப்புமிகு மரங்களை அதிக அளவில் நடுவதற்கும் பராமரிப்பதற்கும் காணி காரர்களின் பங்களிப்பு பெரிதும் உதவியது.ஒந்து மாவட்ட மக்களின் குடிநீர் பாசன நீர் ஆதாரமாகத் கிகழும் பாபா நாசம் அணைக் கட்டுவதற்கு காணிக் காரர்களின் உழைப்பு அகிகம். திருநெல்வேலி மாவட்டத்தில் துவக்க காலத்தில் 13 இடங்களில் தனித்தனி குழுக்களாக காணிக்காரர்கள் வசித்து வந்துள்ளனர். இந்த இன மக்கள்,இஞ்சிக் குழிக் காணி சேப்பார் காணி,வாடி விளைக் காணி,வரட்டையாறு காணி, பெருமாள் காணி, பேயார் காணி, மேலக் கௌதலைக் காணி, சிற்றாறுக் காணி, கிடா வெட்டிப் பாறைக் காணி, கொட மாடிக் காணி ஆகிய பகுதிகளில் 30 முதல் 300 குடும்பங்கள் வரை தனித்தனி குழுக்களாக மொத்தம் 1112, காணிக் குடும்பத்தினர் வசித்துள்ளனர். அம்மை நோயால் பாகிக்கப்பட்டு பலர் இடம் பெயர்ந்து விட்டனர். காணிக் குடியிருப்பின் தலைவனாக மேட்டுக் காணியும் அடுத்த நிலையில் மூதவனும் கிகழ்கின்றனர். உடல் பலவீனம் மன பலவீனம் தர்க்கும் மருத்துவராக பிலாத்தியும், மக்களை ஒருங்கிணைத்து, பிற காணிக் குழுக்களுக்கு தகவல் தெரிவித்து பதில் அறியும் பணிகளை விளி காணியும் செய்கிறார்கள். இவர்கள் வாழ்வியல் மற்றும் கல்வி நிலை வளர்ச்சி மற்றும் கணினி அறிவின் பயன் குறித்தும் இக்கட்டுரை ஆராய முற்படுகின்றது.

வாழ்வியல் :-

காணிக்காரர் என்பதன் பொருள் நிலத்துக்குச்சொந்தக்காரர் என்பதாகும் இவர்கள் பேசும் மொழி காணிக்கார மொழி எனப்படுகிறது. இவ்வின மக்கள் குட்டையான உருவமும் சுருண்ட முடியும், கருத்த நிறமும் உடைய கோற்றம் கொண்டவர்கள். பச்சைக் குக்கிக் கொள்ளுதல் இவர் தம் பழக்க வழக்கங்களில் குறிப்பிடத்தகுந்த ஒன்றாகும். இவர்கள் தாம் வாழும் இடத்தை "காணிக்குடி " என்று அழைக்கின்றனர். மணமாகாத ஆண்களுக்குத் தனியாகக் குடியிருப்புகள் உண்டு. மணமாகாதோர் இரவில் அங்கு தான் தங்க வேண்டும் மூன்று ஆண்டுக்கு ஒரு முறை தங்கள் வாழும் இடத்தை மாற்றுகின்ற பழக்கத்தை உடையவர்களாக இவர்கள் திகழ்கின்றனர். "வேளாண்மை இவர்தம் முக்கியத் தொழிலாக கருதப்படுகிறது. " (தினமணி -

2017)|மரவள்ளிக் கிழங்கு முக்கிய உணவாகும். இவர்களிடத்தில் கொக்கரை எனும் தனித்துவம் வாய்ந்த இசைக்கருவி உண்டு. இரவு ஏழு மணி துவங்கி காலை ஏழு மணி வரை இக்கருவியை ஏழு நபர் சேர்ந்து இசைத்து காணி இன மக்களின் தீராத நோய் தர்க்கின்றனர். மரணத்தை நெருங்குகின்ற உயிர் கூட இந்த இசையால் மறு பிறவி எடுக்கிறது என்பது இவர்களின் அசைக்க முடியாத நம்பிக்கை ஆகும். இமக்களுக்கு ஆவி உலகக்கோட்பாட்டிலும் மறுபிறப்பிலும் அதிக நம்பிக்கை உண்டு இறந்தவர்களை எரிக்கவோ புதைக்கவோ செய்கின்றனர்.

தகவல் பரிமாற்றம் :

கன்னியாகுமரி மாவட்டம் கூவக்காடு மலை, வெள்ளாம்பி மலை, குராவிலை, வில்லு சாரி, ஆலிப்பாறை போன்ற மலைகளில் வாழுகின்ற காணி இன பழங்குடியின மக்கள் தமக்குள் ஒரு பேச்சு மொழியை பின்பற்றுகின்றனர். சான்றாக, சாப்டியா? என்பதற்கு திண்ணியா? என்று கேட்கின்றனர். இக்குடியிருப்பு பகுதிகளில் இருந்து, 60 கிலோமீட்டர் தூரத்தில் உள்ள பொன்முடி, போங்காவு, அகஸ்கியர் கூடம் போன்ற மலைப்பகுதியில் வாழ்கின்ற காணி இன மக்கள் சாப்டியா? என்பதற்கு, வேக்கரி கிட்டியா? என்று வினவுகின்றனர். இவ்விதம், "ஒரு மலைப் பகுதியில் வசிக்கும் இரு குடியிருப்பு மக்களுக்கிடையே பேச்சு மொழி வேறுபடுகின்றது" 2 (சுரேஷ் காணி - கள ஆய்வு)பேச்சு வழக்கில் மிகவும் வித்தியாசமான, ஒன்றுக்கொன்று தொடர்புடைய வார்த்தைகளை பயன்படுக்கிக்கொள்கிறார்கள். இப்பகுதியைச் சார்ந்த இரு வேறு நபர்கள் ஓரிடத்தில் சந்தித்துக் கொள்ள நேரிடும் போது, ஒருவருடைய கருத்தை இன்னொருவர் புரிய வைக்க சமிஞ்ஜை மொழியைப் பயன்படுத்திக் கொள்கின்றனர் இப்படித்தான், துவக்க காலத்தில் இவர்களுடைய, தகவல் பரிமாற்றம் மற்றும் உரையாடல்கள் நிகழ்ந்துள்ளது.

கணினி வழி கற்றல் அறிவு :-

தமிழக அரசின் உண்டு உறைவிடப்பள்ளியில் கல்வி கற்கும் காணி இன மாணவர்கள், மேல் நிலைப்பள்ளி வரும் பொழுது கணினி அறிவைப் பெறுகின்றனர். கள ஆய்வின் வழி பிரியா என்ற மாணவியை அணுகிய பொழுது கீழ்க்கண்ட தரவுகளைப் பெற முடிந்தது. வில்லு சாரி மலையில் இருந்து ஐந்து மாணவிகள் பள்ளிக்கூடம் செல்கிறார்கள். பிரியா, மஞ்சு, அனிதா, மகேஸ்வரி, சித்ரா இப்பகுதியில் வாழும் மக்களுக்கு ஆலம்பாறை மற்றும் பேச்சிப்பாறை ஆகிய இரண்டு இடத்தில் பள்ளிக்கூடம் உள்ளது. இது இவர்கள் விட்டில் இருந்து 4. கிலோ மீட்டர் தூரத்தில் உள்ளது. 2 கிலோ மீட்டர் பேருந்து வசதி இல்லை எனவே, நடந்து சென்று, 2 கிலோமீட்டர் க்குப் பிறகு, பேருந்தில் பயணிக்கின்றனர். மேல்நிலை வகுப்பு வரை உள்ள இந்தப்பள்ளிக் கூடத்தில் 5 ஆசிரியர்கள் பணியாற்றுகின்றனர். காலை 8 மணி துவங்கி, மாலை 5 மணி வரை, பள்ளி நடைபெறுகிறது. உண்டு உறைவிடப் பள்ளி என்பதால் விடுதியில் தங்கி விடுகின்றனர். ஆனால் கணினி அறிவு பெற வேண்டும் என்றால், இவர்கள் 25 கிலோமீட்டர் கடந்து குலசேகரம் என்ற ஊர் செல்ல வேண்டும். தமிழக அரசின் இலவச கணினிப்பயிற்சி அங்குதான் உள்ளது. இதன் வழி, இம்மாணவர்கள் 1817 கற்றுள்ளனர். கணினியில் உள்ள எழுத்துக்களை அறிவதற்கு இவர்கள் தங்கள் பேச்சு மொழியை பயன்படுத்துகின்றனர் வகைப்படுத்தப்படாத திராவிட மொழிக் குடும்பத்தைச் சார்ந்த இவர்களின் பேச்சு மொழி தமிழும் மலையாளமும் கலந்த ஒன்றாகும்.. கணினியின் மாம் செயலியில்

இருக்கும் மிக மேம்பட்ட மொழி சார் தொழில்நுட்பமாக 'குரல் உள்ளீடு " (office dictationlummiM, " இயந்திர வாசித்தல்" ஆகிய உள்ளீடு செய்யப்பட்டுள்ளன.

குரல் உள்ளீடு :-

சொற்செயலி ஒன்றைத் திறந்து கோப்பைத்தட்டச்சு செய்வதற்குரிய வகையில் தயார் செய்து கொள்ள வேண்டும். தட்டச்சு செய்ய வேண்டிய செய்தியைக் கெளிவான உச்சரிப்புடன் வாய்மொழியாக சொல்ல சொல்ல தட்டச்சு எழுதப்படும்.

இயந்திர வாசித்தல் :-

சொற்செயலியில், " நாம் தமிழில்

தட்டச்சு செய்தவை சந்திப்பிழைகள் முதலான இலக்கணப் பிழைகளின்றி அமைந்துள்ளனவா என்பதைக் கண்டறிவதற்கும் தொடர்கள் செம்மையாக அமைந்துள்ளனவா என்பதைத் தெரிந்து கொள்வதற்கும் உதவியாக இருக்கின்றது" 3 (தமிழ் வளம். கணினியின் தமிழ் பக் - 23 ,24- 2023)

இதன் பயிற்சி வழி ,காணி இனப் பழங்குடியின மக்களின் கணினி அறிவு சார்ந்து புரிந்து கொள்ளும் ,வார்த்தைகள் பின்வருமாறு:-

- கணினி கரையுமோ?- கணினி கற்கத்தெரியுமோ?
- கணினி - தட்டாம் பெட்டி
- விசைப்பலகை - கை கட்டு பலகை
- திரை - கண் பினுக்கி
- சொற்செயலி - வாக்கு
- அழிக்கல் _ மாச்சி (மாயுகல்)
- தேடுதல் - திறக்கு
- சேமி - கூட்டி
- வள்ளி
- குத்து

இது போன்று தங்கள் பேச்சு மொழியைப் பயன்படுத்தி கணினி அறிவில் ஆரம்ப நிலை வளர்ச்சி பெற்றுள்ளனார் . சுரேஷ் காணி என்பவர் இப்பகுதி மக்களிடையே முனைவர் பட்டம் பெற்ற முதல் பட்டதாரி ஆவார்.இவர் கல்வி கற்ற பிறகு தம் காணி இன மக்கள் கல்வி வளர்ச்சி ப்பெறுவதற்காக கி.பி.2000 இல் "வழிகாட்டி " என்ற அறக்கட்டளையை உருவாக்கி உள்ளார் கி.பி..2016 வரை ஆஸ்கிரேலியாவில் இருந்து நிதி உதவி பெற்ற இந்த அறக்கட்டளை தற்பொழுது நிதி உதவி பெற இயலா சூழலில் இவரின் தனித்த முயற்ச்சியால் தம் இன மக்களின் உதவியுடன் செயல்படுகின்றது. அரசின் உதவி இன்றி 168 பட்டதாரி மாணவர்களை உருவாக்கிய சாதனை போன்றுகற்குரியது. மேலும்,சுரேஷ் காணி தமது மகளை மருத்துவர் ஆக்கி உள்ளார். விழிப்புணர்வு இல்லா தம் மக்கள் வாழ்வு மேம்பட அரும்பணி ஆற்றுகிறார்.

தொகுப்புரை :-

=> தமிழக அரசு இம்மக்கள் நலன் சார்ந்து இவர்களின் கல்வி நலன் மேம்பட பேருந்து வசதி செய்து தர வேண்டும். -2சமூக நலன் கருதி , உண்டு உறைவிடப் பள்ளி மட்டும் அல்லாது "வழிகாட்டி " அறக்கட்டளை மேம்பட உதவ வேண்டும்

=> இப்பகுதியில் ,கணினி கற்க விருப்பமுள்ள மாணவிகளுக்கு இலவச கணினி வழங்கும் குலசேகரம் போன்று எளிதில் அணுகும் பயிற்ச்சி மையம் ஏற்படுத்தல் வேண்டும்.

=> குரல் உள்ளீடு, இயந்திர வாசித்தல் போன்ற சாதனங்கள் இப்பகுதியின மக்கள் வளர்ச்சி மேம்பட உதவுகிறது.

=> கல்வி அறிவு குறித்த விழிப்புணர்வு காணி இன மக்களிடம் அதிகம் வர வீடு தோறும் கல்வித் திட்டம் மலைப்பகுதி வரை செயல்முறைப்படுத்த வேண்டும்.

=> ஆரோக்கியமான இயற்கை சூழல், உணவு ,எளிதில் அணுகும் கல்வி பொருளாதார வளம் மேம்படும் தொழில்சார் அறிவு ஆகியவை இக்காணி இன மக்கள் பெற அரசு மட்டுமின்றி தன்னார்வத் தொண்டு நிறுவனங்களும் உதவ வேண்டும்.

=> வருங்கால காணி இனத் தலைமுறை மக்கள வளமாக்க மேற்கண்ட திட்டங்கள் செயல்முறைப் படுத்துதலே இக்கட்டுரை வழி பெறப்பட்ட முடிவுகளாகும்.

துணை நின்ற நூல்கள் :

1. இயற்கையோடு இயைந்த காணிக்காரர்கள் (டிசம்பர் - 2017)
2. தமிழ் வளம் - "கணினியின் தமிழ் "" கிருஷ்ணமூர்த்தி பழனியப்பன், படைப்பு பதிப்பகம், கடலூர் தமிழ்நாடு, ஆகஸ்டு- 2023
3. கள ஆய்வு - முனைவர் சுரேஷ் காணி, வில்லுசாரி மலை. -பிரியா (மாணவி)

Reference Books:

1. Iyarkkaiyodu Iyaynthu Kaanikarargal (December-2017)
2. Tamizh Valam - "Kaniniyin Tamizh", Krishnamoorthy Pazhaniyappan, Creation Publication, Cuddalore, Tamilnadu (August-2023)
3. Kala Aivu- Dr. Suresh Kaani, Villusari Malai - Priya (student)



INTERNATIONAL CONFERENCE ON TAMIL COMPUTING AND INFORMATION TECHNOLOGY


June 14 - 16, 2024 Dallas, TX, USA

ICTCIT 2024


பன்னாட்டுக் கணிததமிழ்த் தகவல் தொழில்நுட்ப மாநாடு 2024



Seats
300 People



Date
June 14 - 16, 2024



Speaker
30+ Professional

