



கலஞ்சியம் – சர்வதேசத் தமிழ் ஆய்விதழ்
Quarterly Peer-Reviewed Open Access Crossref Google Scholar
SJIF Impact Factor 4.0 Multidisciplinary International

DOI: <https://doi.org/10.63300/kijs06022026>



Vol. 6 No. 02, August 2026, p309-319 | Available online at: www.tamilmanam.in
Original Research Paper

கணினியில் தமிழ் வளர்ச்சியும் நவீன இயல்மொழி செயலாக்கத்

தொழில்நுட்பங்களும்: ஓர் விரிவான ஆய்வறிக்கை

Tamil Language Development in Computing and Modern Natural Language Processing
Technologies: A Comprehensive Research Report

திருமதி சு.கலைமதி,

உதவிப் பேராசிரியர், பெரி கலை மற்றும் அறிவியல் கல்லூரி, மண்ணிவாக்கம், சென்னை.

Mrs. S. Kalaimathi^{ID*},

Assistant Professor, PERI College of Arts and Science, Mannivakkam, Chennai.

Mobile: 9940206780, Email: nandhinikalaimathi8@gmail.com.

திருமதி ரா அனிதா,

உதவிப் பேராசிரியர், மின் மற்றும் மின்னணு பொறியியல் துறை (EEE), பெரி தொழில் நுட்பக் கல்லூரி,

மண்ணிவாக்கம், சென்னை-48.

Mrs. R. Anitha^{ID},

Department of Electrical and Electronics Engineering (EEE), Assistant Professor,

PERI Institute of Technology, Mannivakkam, Chennai-48.

Email: anithaananth2711@gmail.com

Article history

Received: 10-08-2026

Revised: 10-08-2026

Accepted: 12-08-2026

Published: 13-08-2026

*Corresponding Author Email:

nandhinikalaimathi8@gmail.com

Abstract: With the rapid growth of computer technology, the application of the Tamil language has expanded globally. Overcoming the early limitations of 8-bit ASCII encoding systems, the introduction of 16-bit encoding standards such as Unicode and TACE16 has enabled the seamless use of Tamil characters across all computer and mobile platforms. This paper comprehensively examines the historical evolution of Tamil printing and typing systems to its modern spread across emails, blogs, online dictionaries, digital libraries, and social media. Furthermore, it evaluates the technical advancements achieved in modern Natural Language Processing (NLP) domains, including spell checkers, morphological analyzers, syntactic treebanks, Tamil Optical Character Recognition (OCR), speech technologies (TTS & Speech Recognition), and Artificial Intelligence-driven e-governance chatbots. This study highlights how the Tamil language maintains sustainability and international stature in the digital era through contributions from the International Forum for Information Technology in Tamil (INFITT) and various research organizations.

Keywords: Tamil Computing, Unicode, TACE16, Natural Language Processing (NLP), Morphological Analyzer, Tamil Treebank, Tamil OCR, Artificial Intelligence (AI).



Copyright © 2024 by the author(s). Published by Department of Library, Nallamuthu Gounder Mahalingam College, Pollachi. This is an open access article under the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

சுருக்கம்

கணினித் தொழில்நுட்பத்தின் அசுர வளர்ச்சியுடன் தமிழ்மொழியின் பயன்பாடும் உலகளாவிய ரீதியில் விரிவடைந்துள்ளது. 8-பிட் அஸ்கி (ASCII) குறியேற்ற முறைகளின் தொடக்கக்காலக் கட்டுப்பாடுகளைக் கடந்து, ஒருங்குறி (Unicode) மற்றும் TACE16 போன்ற 16-பிட் குறியேற்ற அமைப்புகளின் அறிமுகம், தமிழ் எழுத்துக்களை அனைத்துக் கணினி மற்றும் கைப்பேசித் தளங்களிலும் தடையின்றி பயன்படுத்த வழிவகுத்துள்ளது. இக்கட்டுரை, தமிழ் அச்ச மற்றும் தட்டச்சு முறைகளின் வரலாற்று வளர்ச்சியிலிருந்து, மின்னஞ்சல், வலைப்பூக்கள், இணைய அகராதிகள், மின்னியல் நூலகங்கள் மற்றும் சமூக ஊடகங்களில் தமிழின் பரவல் வரையிலான பரிணாமத்தை விரிவாக ஆராய்கிறது. மேலும், சொற்பிழைத் திருத்திகள், உருபனியல் பகுப்பாய்விகள் (Morphological Analyzers), தொடரியல் மர வங்கிகள் (Treebanks), தமிழ் எழுத்துரு உணரி (OCR), பேச்சுத் தொழில்நுட்பங்கள் (TTS & Speech Recognition) மற்றும் செயற்கை நுண்ணறிவு சார்ந்த மின்-ஆளுமை உரையாடல் செயலிகள் (Chatbots) உள்ளிட்ட நவீன இயல்மொழி செயலாக்கப் (NLP) புலங்களில் அடைந்துள்ள தொழில்நுட்ப முன்னேற்றங்களை மதிப்பீடு செய்கிறது. உலகத் தமிழ் தகவல் தொழில்நுட்ப மன்றம் (உத்தமம் / INFITT) மற்றும் பல்வேறு ஆய்வமைப்புகளின் பங்களிப்புடன், தமிழ்மொழி டிஜிட்டல் யுகத்தில் எவ்வாறு நிலைத்தன்மையையும் பன்னாட்டு ஆளுமையையும் பெற்றுத் திகழ்கிறது என்பதை இவ்வாய்வு விளக்குகிறது.

முக்கியச் சொற்கள்

தமிழ் கணினியியல், ஒருங்குறி (Unicode), TACE16, இயல்மொழி செயலாக்கம் (NLP), உருபனியல் பகுப்பி, தமிழ் மர வங்கி (Treebank), தமிழ் OCR, செயற்கை நுண்ணறிவு (AI).

1. அறிமுகமும் ஆய்வு நோக்கமும்

1.1 வரலாற்றுப் பின்னணியும் தமிழ் மொழியின் தொன்மையும்

மனித மூளையின் சிறப்புத் திறனான மொழித்திறனை வளர்த்தெடுக்க மனித சமுதாயம் தொடர்ந்து முயன்று வந்ததன் வெளிப்பாடே மொழிகளுக்கான வரிவடிவங்கள் ஆகும். மொழியின் எழுத்துவடிவம் என்பது மனிதன் உருவாக்கிய முதல் மொழித் தொழில்நுட்பமாகக் கருதப்படுகிறது. உலகில் வாழும் மிகவும் பழமையான செம்மொழிகளுள் தமிழ்மொழி முதன்மையானதாகும். குமரிக்கண்டத்தில் தோன்றியதாகக் கருதப்படும் தமிழ்மொழி, சிந்துவெளி நாகரிக சான்றுகள் மற்றும் பண்டைய இலக்கியங்களின் மூலம் தன் தொன்மையை நிறுவியுள்ளது. பழந்தமிழர்கள் வணிகம் மற்றும் பண்பாட்டுத் தொடர்புகள் காரணமாக உலகெங்கிலும் புலம் பெயர்ந்ததால், தமிழின் வேர்ச்சொற்கள் உலகின் பல மொழிகளிலும் பரவியுள்ளன என்பதை பரிதிமாற்கலைஞர், தேவநேயப் பாவாணர், கால்டுவெல் போன்ற மொழியியல் அறிஞர்கள் பலரும் சுட்டிக்காட்டியுள்ளனர். ரிக், யசூர்,

சாமம், அதர்வணம் போன்ற வேதங்களுக்கு எல்லாம் முன்னோடியான செறிவான செய்திகள் தமிழ் இலக்கியங்களில் பொதிந்துள்ளன.

1.2 கணித்தமிழின் தோற்றமும் ஆய்வு நோக்கமும்

அச்சுப் பொறிக்காலத்தைக் கடந்து, தகவலியல் மற்றும் கணினி யுகத்திற்குள் தமிழ் நுழையும் போது, அதன் நெகிழ்வுத் தன்மையும் நவீனப்படக்கூடிய ஆளுமையும் மீண்டும் நிரூபிக்கப்பட்டுள்ளன. ஆங்கிலத்திற்கு அடுத்தபடியாக கணிப்பொறியிலும் இணையத்திலும் பரவலாகப் பயன்படுத்தப்படும் மொழிகளில் தமிழ் முன்னணியில் திகழ்கிறது. இயல், இசை, நாடகம் என்னும் முத்தமிழுடன் "கணினித்தமிழ்" என்பது நான்காம் தமிழாக உருவெடுத்துள்ளது.

உலகில் வாழும் பல்லாயிரக்கணக்கான தமிழர்கள் தங்களின் இலக்கியம், வரலாறு, கோவில் கட்டிடக்கலை, சிற்பக்கலை, கல்வெட்டியல், சுவடியியல், அகராதி மற்றும் மொழிபெயர்ப்பு சார்ந்த தரவுகளை டிஜிட்டல் மயமாக்கிப் பாதுகாக்கவும், உலகின் பிற மக்களுடன் பகிர்ந்துகொள்ளவும் கணினித்தமிழ் வழிவகுத்துள்ளது. கணினியில் தமிழ் அடைந்துள்ள வரலாற்றுப் பரிணாம வளர்ச்சி, குறியேற்றச் சவால்கள் மற்றும் தீர்வுகளிலிருந்து நவீன இயல்மொழி செயலாக்கத் (NLP) தொழில்நுட்பங்கள் வரையிலான முழுமையான நிலப்பரப்பை ஆராய்வதே இக்கட்டுரையின் முதன்மை நோக்கமாக அமைகிறது.

2. கணினியில் தொடக்ககாலத் தமிழ் வளர்ச்சியும் குறியீட்டு முறைகளும்

2.1 தொடக்ககால விசைப்பலகைகளும் எழுத்துருக்களும்

சுமார் முப்பது ஆண்டுகளுக்கு முன்னர், கணினியில் தமிழை உள்ளீடு செய்வதற்கான முதற்கட்ட முயற்சிகள் தொடங்கின. தொடக்கத்தில் விசைப்பலகைகளில் நேரடியாக தமிழ் எழுத்துக்கள் இல்லாத சூழலில், ரோமன் (ஆங்கில) எழுத்துக்களில் தட்டச்சு செய்து, பின்னர் அவற்றை தமிழ் எழுத்துக்களாக மாற்றும் ஒலிபெயர்ப்பு (Transliteration) முறைகள் பயன்படுத்தப்பட்டன. "ஆதமி" (Adhami) போன்ற தொடக்ககால மென்பொருட்களும், "மயிலை" (Mylai) போன்ற எழுத்துருக்களும் ஆங்கில விசைப்பலகை அமைப்பைப் பயன்படுத்தி தமிழ் உரையைப் பெற வழிவகுத்தன. படிப்படியாக, தமிழுக்கெனத் தனித்துவமான விசைப்பலகை அமைப்புகளும் அச்சுப் பணிக்கு ஏற்ற எழுத்துருக்களும் மென்பொருள் வல்லுநர்களால் உருவாக்கப்பட்டன.

2.2 8-பிட் அஸ்கி (ASCII) குறியேற்றங்களும் அவற்றின் வரையறைகளும்

தொடக்கத்தில் வேர்டு (MS Word), வேர்டு ஸ்டார் (WordStar) போன்ற இயங்குதளப் பயன்பாடுகளில் தமிழைப் பயன்படுத்த TAB, TAM, TISCII/TSCII, டிஸ்கி (DISKI) மற்றும் டாம்/டாப் போன்ற 8-பிட் அஸ்கி (ASCII) அடிப்படையிலான குறியேற்ற முறைகள் பயன்படுத்தப்பட்டன. எனினும், இக்குறியேற்ற முறைகளில் பெரும் இடர்பாடுகள் நீடித்தன. ஒரு குறிப்பிட்ட மென்பொருள் அல்லது எழுத்துருவைக் கொண்டு தட்டச்சு செய்யப்பட்ட

ஆவணத்தை, மற்றொரு கணினியிலோ அல்லது வேறு மென்பொருளிலோ மாற்றியமைக்கவோ வாசிக்கவோ இயலாத நிலை இருந்தது. குறியீட்டு எண்கள் தரப்படுத்தப்படாமல் ஒவ்வொரு நிறுவனமும் தனித்தனி எழுத்துருக் குறியீடுகளைப் பயன்படுத்தியதால், தரவுப் பரிமாற்றமும் ஆவணக் காப்பக உருவாக்கமும் கடுமையாகப் பாதிக்கப்பட்டன.

2.3 ஒருங்குறி (Unicode) மற்றும் TACE16 குறியேற்றச் சீர்திருத்தங்கள்

கணினித் தமிழ் சந்தித்து வந்த குறியேற்றச் சிக்கல்களுக்கு முற்றுப்புள்ளி வைக்கும் வகையில் பன்னாட்டுத் தரநிலையான "ஒருங்குறி" (Unicode Encoding) அறிமுகப்படுத்தப்பட்டது. ஒருங்குறியில் தமிழின் உயிர் எழுத்துக்கள், மெய் எழுத்துக்கள், ஆய்த எழுத்து, கிரந்த எழுத்துக்கள் மற்றும் குறியீடுகளுக்கு 0B80 முதல் 0BFF வரையிலான தொகுதி எண்கள் (Code block) ஒதுக்கப்பட்டன. இதன் மூலம் கணினிகள், கைபேசிகள் மற்றும் அனைத்து மின்னணு சாதனங்களிலும் தமிழ் உரையை தடையின்றி உள்ளீடு செய்யவும் பெறவும் இயலும் நிலை உருவானது.

இருப்பினும், தமிழ் மொழியின் 247 எழுத்துக்களையும் நேரடியாக 16-பிட் அமைப்பில் கையாள "TACE16" (Tamil All Character Encoding) எனும் மாற்றுத் தரநிலையை தமிழ்நாடு அரசின் தொழில்நுட்பக் குழுவும் கணினித் தமிழ்ச் சங்கமும் பரிந்துரைத்தன. TACE16 முறையானது, தமிழ் உயிர்மெய் எழுத்துக்கள் ஒவ்வொன்றிற்கும் தனித்தனி குறியீட்டு புள்ளிகளை PUA (Private Use Area) பகுதியில் ஒதுக்கி, இயற்கை மொழி செயலாக்கம் மற்றும் அச்சுத் துறை பயன்பாடுகளை எளிமைப்படுத்தியது. தமிழ்நாடு அரசு மற்றும் உத்தமம் (INFITT) அமைப்பு ஆகியவற்றின் கூட்டிணைவால், ஒருங்குறி முதன்மைத் தரநிலையாகவும் TACE16 மாற்றுத் தரநிலையாகவும் ஏற்றுக்கொள்ளப்பட்டு அரசாணைகள் வெளியிடப்பட்டன.

குறியேற்ற தரநிலை	பிட் அளவு	எழுத்து ஒதுக்கீடு	நன்மைகள்	முதன்மை பயன்பாடு
8-Bit ASCII (TAB / TAM / TSCII)	8-Bit	128 க்கும் குறைவான எழுத்துருக் குறியீடுகள் (Glyphs)	தொடக்கக்கால அச்சுப்பதிப்பிற் கு எளிதாக இருந்தது.	தொடக்கக்கால கணினித் தட்டச்சு மற்றும் அச்சுத்துறை.

குறியேற்ற தரநிலை	பிட் அளவு	எழுத்து ஒதுக்கீடு	நன்மைகள்	முதன்மை பயன்பாடு
Unicode (ஒருங்குறி)	16-Bit	0B80 - 0BFF தொகுதி (உயிர், மெய், குறியீடுகள் மட்டும்)	உலகளாவிய இயங்குதளப் பொருத்தம்; அனைத்து சாதனங்களிலும் இயங்கும்.	இணையம், சமூக ஊடகங்கள், மின்னஞ்சல், கைபேசிகள்.
TACE16	16-Bit	384 குறியீட்டு புள்ளிகள் (அனைத்து உயிர்மெய் எழுத்துக்களுக்கு ம்தனித்தனி)	மொழிச் செயலாக்கம் (NLP) மற்றும் அச்சுப் பதிப்பிற்கு மிக அதிவேகச் செயல்பாடு.	இயற்கை மொழி செயலாக்கம், தொழில்முறை அச்சு, உட்பொதிக்கப்பட் ட சாதனங்கள்.

3. இணையத்திலும் மின்னணுத் தொடர்பாடலிலும் தமிழின் விரிவாக்கம்

3.1 மின்னஞ்சல், வலைப்பூக்கள் மற்றும் இணைய அகராதிகள்

அமெரிக்காவில் இணையத் தொழில்நுட்பம் விரிவடைந்ததைத் தொடர்ந்து, கூகுள் (Google) போன்ற தேடுபொறிகளின் வரவு தகவல்களைத் தேடிப் பெறுவதை எளிதாக்கியது. உலகெங்கும் சிதறி வாழும் தமிழர்கள் தங்களுக்குள் தகவல்களைப் பரிமாறிக்கொள்ளவும் பண்பாட்டுச் செய்திகளைப் பகிர்ந்துகொள்ளவும் இணையம் ஒரு சக்திவாய்ந்த ஊடகமாக அமைந்தது.

மின்னஞ்சல் (E-mail) தொழில்நுட்பம் மூலம் எந்தவொரு கூடுதல் கட்டணமுமின்றி நொடிப் பொழுதில் தமிழில் செய்திகளை அனுப்புவது சாத்தியமானது. மேலும், கணினி, அறிவியல், பொருளாதாரம், கல்வி மற்றும் வணிகத் துறைகளில் பயன்படுத்தப்படும் ஆங்கிலச் சொற்களுக்கு இணையான கலைச்சொற்களை உருவாக்கும் நோக்கோடு உமர் என்பவரின் முயற்சியாலும் பிற தமிழறிஞர்களின் பங்களிப்பாலும் "தமிழ் இணைய அகராதி" உருவாக்கப்பட்டது.

இலக்கிய வளர்ச்சிக்கும் கருத்து வெளிப்பாட்டிற்கும் "வலைப்பூக்கள்" (Blogs) பெரும் பங்காற்றின. உரை, ஒலி, ஒளி மற்றும் புகைப்படக் கோப்புகளை இணையம் வழியே உலகிற்குப் பகிர்வதற்கு வலைப்பூக்கள் உதவின. "வேர்களைத் தேடி" போன்ற வலைப்பூக்களில் தமிழாசிரியர்கள் இலக்கிய ஆய்வுகளைப் பதிவேற்றம் செய்த அதேவேளையில், அறிவியல், கணிதம், விண்வெளி, மருத்துவம், பெண்கள் சுதந்திரம் மற்றும் வேலைவாய்ப்பு சார்ந்த பல்சுவை வலைப்பூக்களும் தமிழில் பெருமளவில் தோன்றின.

3.2 தமிழ் மின்னியல் நூலகங்களும் மின்நூல்களும்

இயந்திரமயமான நவீன உலகில் நூலகங்களுக்கு நேரடியாகச் சென்று படிக்கும் ஆர்வம் குறைந்து வரும் சூழலில், தமிழ் மின்னியல் நூலகங்களும் (Digital Libraries) இணைய நூலகங்களும் முக்கியப் பங்காற்றுகின்றன. அச்சு நூல்களை டிஜிட்டல் வடிவத்திற்கு மாற்றி ஆவணப்படுத்த இவை உதவுகின்றன. யுனெஸ்கோ (UNESCO) மற்றும் அமெரிக்க காங்கிரஸ் நூலகம் இணைந்து உலக மின்னியல் நூலகத்தை வழிநடத்தும் வேளையில், தமிழ் மொழியிலும் "தமிழ் இணையக் கல்விக்கழகம்" (Tamil Virtual Academy) மற்றும் "திட்டம் மதுரை" (Project Madurai) போன்ற அமைப்புகள் பல்லாயிரக்கணக்கான சங்க இலக்கியங்கள், காப்பியங்கள் மற்றும் நவீன நூல்களை இலவச மின்நூல்களாக (E-books) இணையத்தில் வெளியிட்டு முன்னணியில் திகழ்கின்றன.

3.3 சமூக ஊடகங்களும் பன்னாட்டுச் செய்தி நிறுவனங்களும்

முகநூல் (Facebook), கட்செவி அஞ்சல் (WhatsApp), குறுஞ்செயலிகள் (Messaging Apps), X (ட்விட்டர்) போன்ற சமூக ஊடகங்களின் வருகை, தமிழ் மொழியின் சாதாரண மக்கள் பயன்பாட்டைப் பல மடங்கு உயர்த்தியுள்ளது. ஒலி, ஒளி, காணொலி மற்றும் உரை வடிவில் கருத்துக்களைப் பரிமாறவும், சமூகப் பிரச்சினைகளை விவாதிக்கவும் சமூக ஊடகங்கள் தமிழில் தளமமைத்துக் கொடுத்துள்ளன.

மேலும், பன்னாட்டுச் செய்தி நிறுவனங்களான பிபிசி (BBC Tamil) தனது செய்திச் சேவைகளிலும், சீன வானொலி (China Radio International) தனது 43 உலக மொழி ஒலிபரப்பில் உருது மற்றும் வங்காள மொழிகளுக்கு அடுத்தபடியாக இந்திய மொழிகளில் தமிழிலும் இந்தியிலும் செய்திகளை ஒலிபரப்பி வருகிறது.

டிஜிட்டல் தமிழ் வளம் / நிறுவனம்	உருவாக்கம் / ஒருங்கிணைப்பு	முதன்மை நோக்கம் மற்றும் சேவைகள்	தொழில்நுட்ப சிறப்பம்சங்கள்
தமிழ் இணையக்	தமிழ்நாடு அரசு	பழங்கால நூல்கள், கல்வெட்டுகள்,	மாபெரும் தமிழ் உரைத்

டிஜிட்டல் தமிழ் வளம் / நிறுவனம்	உருவாக்கம் / ஒருங்கிணைப்பு	முதன்மை நோக்கம் மற்றும் சேவைகள்	தொழில்நுட்ப சிறப்பம்சங்கள்
கல்விக்கழகம் (TVA)		சுவடிகள் காப்பகம் மற்றும் தமிழ் கற்றல்.	தரவுத்தொகுப்பு (Corpora) மற்றும் மின்-நூலகம்.
திட்டம் மதுரை (Project Madurai)	தன்னார்வலர் இயக்கம்	சங்க இலக்கியம் முதல் நவீன இலக்கியம் வரையிலான இலவச மின்நூல்கள்.	Unicode/UTF-8 வடிவில் திறந்தவெளி அணுகல் (Open Access).
பிபிசி தமிழ் (BBC Tamil)	பிபிசி (BBC World Service)	உலகளாவிய செய்திகளைத் தமிழில் வழங்குதல்.	இணையம், பாட்காஸ்ட் மற்றும் சமூக ஊடக ஒருங்கிணைப்பு.
தமிழ் விக்கிப்பீடியா	விக்கிமீடியா அறக்கட்டளை	திறந்தவெளித் தமிழ்க் கலைக்களஞ்சியக் கட்டுரைகள்.	48,000 க்கும் மேற்பட்ட அறிவுசார் கட்டுரைகள் மற்றும் கூட்டுப் பங்களிப்பு.

4. தமிழ் இயல்மொழி செயலாக்கமும் (NLP) மொழியியல் தொழில்நுட்பங்களும்

4.1 சொற்பிழைத் திருத்திகளும் சந்திப்பிழைத் திருத்திகளும்

கணினியில் தமிழைத் தட்டச்சு செய்யும்போது ஏற்படும் எழுத்துப் பிழைகளைச் சுட்டிக்காட்ட சொற்பிழைச் சுட்டிகளும் (Spell Pointers), பிழைகளைத் திருத்த சொற்பிழைத் திருத்திகளும் (Spell Checkers) உருவாக்கப்பட்டன. தமிழுக்கே உரிய கடின இலக்கண விதிகளான சந்திப்பிழைகள் (ஒற்றுப்பிழை) மற்றும் தொடரியல் பிழைகளைத் திருத்தும் இலக்கணப் பிழைத் திருத்திகள் கணினித் தமிழ்ப் பயன்பாட்டை மேலும் துல்லியமாக்கியுள்ளன.

4.2 உருபனியல் பகுப்பாய்வியும் உரைச் சுருக்கத் தொழில்நுட்பமும்

தமிழ் ஒரு ஒட்டுநிலை (Agglutinative) மொழியாகும். ஒரு வேர்ச்சொல்லுடன் பல விசுதிகள் சேரும்போது அது பிரம்மாண்டமான சொற்களாக உருவெடுக்கிறது. எனவே, இயல்மொழி செயலாக்கத்தில் (NLP) "உருபனியல் பகுப்பாய்வி" (Morphological Analyzer) மற்றும் "உருபனியல் ஆக்கி" (Morphological Generator) மிக முக்கியப் பங்காற்றுகின்றன.

கைப்பேசிகளிலும் வலைப்பூக்களிலும் குறுஞ்செய்தி அனுப்பும்போது இடத்தையும் சேமிப்பகத்தையும் மிச்சப்படுத்த "உரைச் சுருக்கத் தொழில்நுட்பம்" (Text Compaction System) தமிழிலும் அறிமுகப்படுத்தப்பட்டுள்ளது. இத்தொழில்நுட்பம், சொற்களைப் பகுப்பாய்வு செய்து அவற்றின் வேர்ச்சொற்களைக் கண்டறிந்து (Stemming), சுருக்கக் குறியீடுகள் மற்றும் விசுதிகளை மாற்றுவதன் மூலம் உரையின் அளவைக் குறைத்து, சேமிப்புச் செலவை மிச்சப்படுத்துகிறது.

4.3 தொடரியல் மர வங்கிகளும் (Treebanks) இலக்கண பகுப்பாய்வும்

தமிழ் வாக்கியங்களின் இலக்கண அமைப்பையும் தொடரியல் உறவுகளையும் கணினி மாதிரிகளுக்குக் கற்பிக்க "மர வங்கிகள்" (Treebanks) உருவாக்கப்படுகின்றன. கையால் குறியிடுதல் (Manual Annotation), கணக்கீட்டு இலக்கணங்கள் (Computational Grammars) மற்றும் இயந்திரக் கற்றல் (Machine Learning) ஆகிய மூன்று அணுகுமுறைகளில் தமிழ் மர வங்கிகள் கட்டமைக்கப்படுகின்றன. அகில உலக உலகளாவிய சார்ந்திருத்தல் (Universal Dependencies) அமைப்பின் அடிப்படையில் உருவாக்கப்படும் இத்தகைய மர வங்கிகள், இயந்திர மொழிபெயர்ப்பு (Machine Translation) மற்றும் உணர்வுப் பகுப்பாய்வு (Sentiment Analysis) போன்ற மேம்பட்ட AI பயன்பாடுகளுக்கு அடிப்படையாக அமைவதோடு, பாகுபடுத்திகளை (Parsers) உருவாக்கவும் உதவுகின்றன.

4.4 தமிழ் எழுத்துரு உணரி (OCR) மற்றும் பேச்சுத் தொழில்நுட்பம்

அச்சிடப்பட்ட பழங்கால தமிழ் நூல்கள், செய்தித்தாள்கள் மற்றும் பனை ஓலைச் சுவடிகளை கணினியில் திருத்தக்கூடிய டிஜிட்டல் உரையாக மாற்ற "தமிழ் எழுத்துரு உணரி" (Tamil OCR) தொழில்நுட்பம் பயன்படுகிறது. Indian Institute of Science (IISc) பேராசிரியர் ஏ.ஜி. இராமகிருஷ்ணன் போன்ற ஆய்வாளர்கள் உருவாக்கிய "OmniFont Tamil OCR" மற்றும் கூகுளின் 'Tesseract OCR' ஆகியவை அச்சிடப்பட்ட தமிழையும் ஓலைச்சுவடிகளையும் டிஜிட்டல் மயமாக்குவதில் மைல்கல்லாக அமைந்துள்ளன.

அதேபோல், உரையை ஒலியாக மாற்றும் "உரை-ஒலி மாற்றம்" (Text-to-Speech - TTS) மற்றும் மனிதப் பேச்சை உரையாக மாற்றும் பேச்சு உணரி (Speech Recognition) அமைப்புகள் பார்வையற்றோருக்கும் கைபேசிப் பயனர்களுக்கும் பெரும் துணையாக உள்ளன. பேச்சுத் தடங்கல்களைக் (Speech Disfluencies) கண்டறியும் பல-அளவுகோல்

கவன மாதிரிகள் (Multi-scale Attention Frameworks) மற்றும் mBERT மாதிரிகள் தமிழில் 95.85% துல்லியத்துடன் வெற்றிகரமாகச் சோதிக்கப்பட்டுள்ளன.

NLP தொழில்நுட்பக் கூறு	பயன்படுத்தப்படும் மாதிரிகள் / கருவிகள்	முதன்மைச் செயல்பாடு
உருபனியல் பகுப்பாய்வி (Morphological Analyzer)	Rule-based & Finite State Transducers (FST)	சொற்களிலிருந்து விசுதிகளைப் பிரித்து வேர்ச்சொல்லைக் (Root Word) கண்டறிதல்.
தொடரியல் மர வங்கி (Tamil Treebanks)	Universal Dependencies (UD) & Machine Learning Parsers	வாக்கியங்களின் இலக்கண அமைப்பு மற்றும் சொற்களுக்கு இடையேயான சார்ந்திருத்தலைக் குறித்தல்.
தமிழ் OCR (Optical Character Recognition)	Omnifont OCR & Tesseract Tamil Engine	அச்சிடப்பட்ட தாள்கள் மற்றும் ஓலைச்சுவடிகளை கணினி உரையாக மாற்றுதல்.
பேச்சு உணரியும் பகுப்பாய்வும் (Speech & Disfluency)	Multilingual BERT & Multi-Scale Attention Frameworks	மனித பேச்சிலுள்ள தயக்கங்கள், மறுபடியும் கூறும் சொற்களைக் கண்டறிந்து உரையாடலைத் துல்லியமாக்குதல்.

5. செயற்கை நுண்ணறிவும் மின்-ஆளுமையில் தமிழின் எதிர்காலமும்

5.1 பெரும் மொழி மாதிரிகளும் (LLMs) தரவுத் தொகுப்புகளும்

செயற்கை நுண்ணறிவின் யுகத்தில் தமிழ்மொழிக்கெனப் பிரத்தியேக தரவுத் தொகுப்புகள் (Corpora) உருவாக்கப்பட்டு வருகின்றன. 'IndicCorp' போன்ற 300 கோடி சொற்களுக்கு மேலான பிரம்மாண்ட தமிழ் தரவுத் தொகுப்புகள், GPT மற்றும் BERT போன்ற பெரும் மொழி மாதிரிகளுக்குப் (LLMs) பயிற்சி அளிக்கப் பயன்படுகின்றன. இதன் மூலம் தமிழ் மொழியிலேயே இயல்பான உரை உருவாக்கம், ஆவணச் சுருக்கம் வழங்குதல் மற்றும் வினா-விடை அமைப்புகளை உருவாக்குவது சாத்தியமாகியுள்ளது.

5.2 அரசுச் சேவைகளில் தமிழ் உரையாடல் செயலிகள் மற்றும் AI தொழில்நுட்பங்கள்

தமிழ்நாடு மின்-ஆளுமை முகமை (TNeGA), அண்ணா பல்கலைக்கழகத்துடன் இணைந்து "அனில்" (Anil) என்ற தமிழ் இயல்மொழி செயலாக்கம் சார்ந்த தானியங்கி உரையாடல் செயலியை (Chatbot) உருவாக்கியுள்ளது. இச்செயலி, பொதுமக்களுடன் தமிழில் உரையாடி இருப்பிடச் சான்றிதழ், வருமானச் சான்றிதழ், சாதிச் சான்றிதழ் போன்ற அரசுச் சேவைகளைப் பெறுவதற்கு வழிகாட்டுகிறது. மேலும், "உழவன்" செயலி மூலம் விவசாயிகள் பயிர் நோய்களைப் படம்பிடித்து அனுப்பினால் தமிழில் தீர்வு வழங்கும் செயற்கை நுண்ணறிவு அமைப்புகளும் பயன்பாட்டில் உள்ளன.

6. பன்னாட்டு அமைப்புகளும் மாநாடுகளும்

கணினித் தமிழின் பரவலாக்கத்திற்கு பன்னாட்டு அமைப்புகளும் மாநாடுகளும் முக்கியப் பங்காற்றியுள்ளன. சென்னையைச் சேர்ந்த கணினித் தமிழ்ச் சங்கம் மதுரையில் "கணினித் திருவிழா" என்ற பெயரில் மென்பொருள் கண்காட்சிகளை நடத்தி தமிழ் மென்பொருட்களை மக்களிடம் கொண்டு சேர்த்தது.

1997-ஆம் ஆண்டு முதல் உலகளவில் நடத்தப்பட்டு வரும் "உலகத் தமிழ் இணைய மாநாடுகள்" (Tamil Internet Conferences - TIC) கணினித் தமிழின் மைல்கற்களாகும். இம்மாநாடுகளை ஒருங்கிணைக்கும் "உலகத் தமிழ் தகவல் தொழில்நுட்ப மன்றம்" (உத்தமம் / INFITT), தமிழ் ஒருங்குறித் தரநிலை, TACE16 குறியேற்றம், இயல்மொழி செயலாக்கம், ரோபாட்டிக்ஸ் மற்றும் செயற்கை நுண்ணறிவுப் புலங்களில் தொடர்ச்சியான சர்வதேச ஆய்வுகளை ஊக்குவித்து வருகிறது.

7. முடிவுரையும் எதிர்கால நோக்கும்

கணினியுக் கருதும் அளவுக்கு மனித வாழ்வில் கணினியும் அலைபேசியும் தவிர்க்க முடியாத அங்கமாக மாறியுள்ளன. இத்தகைய நவீன சூழலில், தமிழ் மொழி தன்னைத் தொடர்ச்சியாகப் புதுப்பித்துக் கொண்டு, ஆங்கிலத்திற்கு இணையான ஆளுமையுடன் கணிப்பொறியிலும் இணையத்திலும் திகழ்கிறது. 8-பிட் எழுத்துருக் குழப்பங்களிலிருந்து விடுபட்டு, ஒருங்குறி, TACE16, இயல்மொழி செயலாக்கம், மர வங்கிகள், OCR மற்றும் செயற்கை நுண்ணறிவு உரையாடல் செயலிகள் வரை கணினித் தமிழ் அடைந்துள்ள வளர்ச்சி வியக்கத்தக்கதாகும்.

எதிர்காலத்தில், குறைந்த வளங்களைக் கொண்ட மொழிகளுக்கான (Low-resource languages) நரம்பியல் இயந்திர மொழிபெயர்ப்பு (NMT) மாதிரிகளை மேலும் வலுப்படுத்துவதும், தமிழ் ஓலைச்சுவடிகள் மற்றும் அரிய இலக்கியங்களை முழுமையாக டிஜிட்டல் மயமாக்கி AI மாதிரிகளுக்குப் பயிற்றுவிப்பதும் கணினித் தமிழின் அடுத்தகட்ட சவால்களாகவும் வாய்ப்புகளாகவும் அமையும். காலத்திற்கு ஏற்பத் தன்னை

மறுவடிவமைத்துக் கொள்ளும் தமிழின் தகவமைப்புத் திறன், டிஜிட்டல் யுகத்திலும் அதன் நிலைத்தன்மையையும் உலகளாவிய சீர்மையையும் என்றும் உறுதி செய்யும்.

8. ஆசிரியர் பிரகடனம் மற்றும் நலன்களின் முரண்பாட்டின்மை

இவ்வாய்வுக் கட்டுரையின் ஆசிரியர்கள், இக்கட்டுரையில் இடம்பெற்றுள்ள தகவல்கள், ஆய்வுகள் மற்றும் முடிவுகள் அனைத்தும் முற்றிலும் கல்வி மற்றும் ஆய்வு நோக்கத்திற்காகவே சமர்ப்பிக்கப்பட்டுள்ளன என்பதைப் பிரகடனம் செய்கின்றனர். இவ்வாய்வினை மேற்கொள்வதிலோ அல்லது இதன் முடிவுகளை வெளியிடுவதிலோ எவ்விதமான நிதி சார்ந்த அல்லது தனிநபர் சார்ந்த நலன்களின் முரண்பாடும் (Conflict of Interest) இல்லை என உறுதி அளிக்கின்றனர்.