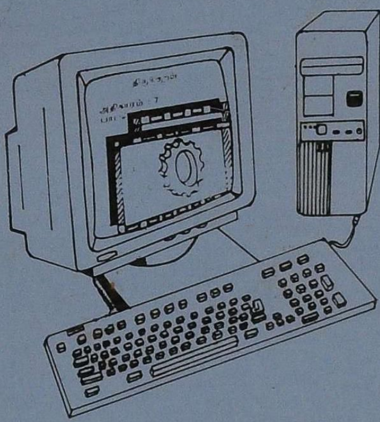
தமிழும் கணிப்பொறியும்

கருத்தரங்கு

ஆகஸ்டு 5,6 - 1994



கணிப்பொறி அறிவியல் தொழில் நுட்பத்துறை

அண்ணா பல்கலைக்கழகம்

சென்னை - 600 025

தமிழும் கணிப்பொறியும்

கருத்தரங்கு

ஆகஸ்டு 5,6 - 1994

கணிப்பொறி அறிவியல் தொழில் நுட்பத்துறை
அண்ணா பல்கலைக்கழகம்
சென்னை - 600 025

Laser typeset, Offset printing by: Page Offset, 22 & 27 Wallajah Road, Madras-2. ☎ 830 1[illegible]0

ஆலோசனைக் குழு

1. முனைவர் மு. ஆனந்த கிருஷ்ணன், துணை வேந்தர் (தலைவர்)
2. முனைவர் ஔவை நடராஜன், துணை வேந்தர்
3. முனைவர் வி. ஞானபிரகாசம், துணை வேந்தர்
4. திரு. எஸ். ரங்கராஜன் (''சுஜாதா'')
5. திரு. டி.எஸ். முத்துகிருஷ்ணன் (எஸ்.ஆர்.ஜி)

புரவலர்கள்

1. தமிழ் பல்கலைக் கழகம்
2. தமிழ்நாடு கால்நடை மருத்துவ அறிவியல் பல்கலைக் கழகம்
3. தமிழ்நாடு அறிவியல் தொழில்நுட்ப மாநில மன்றம்
4. சாஃப்ட்வேர் ரிசர்ச் குரூப்
5. இந்தியன் இன்ஸ்டிடியூட் ஆஃப் இஞ்சினியரிங் டெக்னாலஜி
6. எஸ். ஆர்.எம். பொறியியற் கல்லூரி
7. டிசி எல்காட்

ஒருங்கிணைப்பு குழு

1. முனைவர் கே. மதியழகன்
2. முனைவர் வி. கணபதி
3. முனைவர் வி. சங்கரநாராயணன்
4. முனைவர் என். சங்கரநாராயணன்
5. முனைவர் வி. கிருஷ்ணமூர்த்தி (ஒருங்கிணைப்பாளர்)
6. முனைவர் பி. ஈஸ்வர்
7. முனைவர் ஏ. பாலசுப்ரமணியம்
8. திருமதி வி. உமா மஹேஸ்வரி (பொருளாளர்)
9. முனைவர் ஏ. இளங்கோவன்
10. முனைவர் பி. ஆர். நக்கீரன்
11. முனைவர் டி.எஸ். சுப்பராமன்
12. முனைவர் எம். ஆறுமுகம்

உட்குழுக்கள்

1. அச்சு

முனைவர் எ. பாலசுப்ரமணியம்

திரு. அருள் சிரோமணி

முனவர் வி. ராமச்சந்திரன்

முனைவர் எம். ஆறுமுகம்

2. அரங்க நிர்வாகம்

முனைவர் பி. ஈஸ்வர்
முனைவர் கே.எஸ். ஸ்ரீதரன்
முனைவர் பி.ஆர். நக்கீரன்
திரு. டி.வி. கோபால்
திரு. ஸ்ரீதர் கந்தசாமி
திருமதி. ஏ.பி. சாந்தி

3. வரவேற்பு

முனைவர் (திருமதி) டி.வி. கீதா
திருமதி எஸ். சாந்தா
திருமதி டி. மஞ்சுளா
திரு. ஏ. கண்ணன்
திரு. எம். ராமசுப்ரமணியன்
செல்வி. வள்ளி

கருத்தரங்கின் நோக்கம்

நாளைய உலகில் எந்தத்துறையர்க இருந்தாலும் போட்டியைச் சமாளித்து வெற்றி பெற கணிப்பொறியைப் பயன்படுத்துவது இன்றியமையாத ஒன்று. இதற்கு மொழி ஒரு தடையாக இருக்கக்கூடாது. மேலும் மொழியின் வளர்ச்சியிலும் கணிப்பொறி ஒரு சிறந்த பணியாற்றலாம். இதற்கெல்லாம் முதற்படி தமிழ் மொழியை கணிப்பொறிக்குப் பரிச்சயப்படுத்துவதுதான். இதற்கான பல்வேறு முயற்சிகள் ஆங்காங்கே நடைபெற்று வருகின்றன. இந்த முயற்சிகளை ஒருமுகப்படுத்தி வெளிப்படுத்தி, இந்த ஆராய்ச்சியாளர்களை ஊக்குவிப்பதுடன், நாளைய தேவை என்ன - என்னென்ன ஆராய்ச்சிகள் உடனடியாக மேற்கொள்ளப்பட வேண்டும் என்பவற்றை விவாதித்து முடிவெடுக்கவும் ஒரு கருத்தரங்கும், கலந்துரையாடலும் தேவை என்ற எண்ணத்தில் இக்கருத்தரங்கு ஏற்பாடு செய்யப்பட்டுள்ளது.

இம்முயற்சிக்கு பல்வேறு இடங்களிலிருந்தும் நல்ல ஆதரவு கிடைத்துள்ளது. இக்கருத்தரங்கில் சமர்ப்பிக்கப்படும் சுமார் கால்நூறு ஆராய்ச்சிக் கட்டுரைகளே இதற்குச் சான்று. எழுத்து அறிதல், விசைப்பலகை அமைப்பு, மொழி மாற்றம், மொழி ஆய்வு, சொல் உருவாக்கம் போன்ற பலவித தலைப்புகளிலும் கட்டுரைகள் வந்துள்ளன. அவை இங்கு தொகுத்தளிக்கப்பட்டுள்ளன - மிகவும் குறுகிய காலத்தில். இக்கட்டுரைகளில் கூறப்படும் கருத்துக்கள் அந்தந்த ஆசிரியருடையவை. பல்கலைக்கழகத்தின் கருத்தாக எடுத்துக் கொள்ளக் கூடாது.

கணிப்பொறி உலகில் தமிழ் பற்றிய ஆராய்ச்சியை பொது மக்கள் அறிந்து கொள்ள வேண்டும் என்ற எண்ணத்தில் கருத்தரங்குடன் ஒரு கண்காட்சியும் அமைக்கப்பட்டுள்ளது. ஆராய்ச்சி நிலையிலிருக்கும் மென்பொருட்களும் காட்சியில் இடம்பெறுகின்றன. நாளை வரப்போகும் பொருட்களைப் பற்றியும் பொதுமக்கள் அறிய வேண்டும் என்பதற்காகவே இவ்வாறு செய்யப்படுகின்றது.

இக்கருத்தரங்கிற்கு பல்வேறு வகைகளிலும் ஆதரவளித்த புரவலர்களுக்கும் ஏனையோருக்கும் எங்கள் நன்றி. ஆய்வுக்கட்டுரைகளைத் தொகுத்து வெளியிட தமிழ்நாடு அறிவியல் தொழில்நுட்ப மாநில மன்றம் ஆதரவளித்துள்ளது என்பதை நன்றியுடன் நினைவு கூறுகிறோம்.

இக்கருத்தரங்கு ஒரு முதல் முயற்சி. ஓர் அனுபவமாகவும் வழிகாட்டியாகவும் அமையும் என்ற நம்பிக்கையுடன் செயல் புரிவோம்.

வெ. கிருஷ்ணமூர்த்தி
ஒருங்கிணைப்பாளர்

பொருளடக்கம்

1) தமிழும் கணிப்பொறியும் - தொடக்க உரை 1

2) கணித்தமிழ் - தமிழில் ஒரு கணிப்பொறி மொழி 11

3) முத்துக் கணினி - 34 17

4) கணியன் விசைப்பலகை 33

5) கணிப்பொறித்துறைக் கலைச்சொல்லாக்கம் 44

6) கணிப்பொறி - ஓர் மொழியாய்வுக் கருவி 53

7) மொழி ஆய்வில் கணிப்பான் 61

8) கணிப்பொறி வழி உரை அல்லது பாடநூல் ஆய்வு-ஆய்வறிக்கை 66

9) தமிழாய்வில் கணினியின் பங்களிப்பு சிக்கல்களும், தீர்வுகளும் 78

10) கணிப்பொறி வழி சித்த மருத்துவம் 82

11) கணிப்பொறி எண் கட்டுப்பாட்டு இயந்திரம் மூலம் தமிழ் எழுத்துக்களைப் பொறித்தல் 85

12) தமிழ் எழுத்துகளை அடையாளம் கண்டுகொள்வதில் வடிவமைப்புகளின் சிறப்பியல்புகள் 93

13) கணிப்பொறியில் தமிழ் எழுத்துக்களை உருவாக்குவதில் உள்ள சில சிக்கல்கள் 106

14) கோல இயல்வழி கையெழுத்து அறிதல் 108

15) Need for standardisation of Tamil Keyboards 113

16) கணிப்பொறியும் தமிழாக்கமும் 117

17) Tense Identification and Analysis of Tamil Verb Phrases Using Neural Networks 130

18) The possible impact of computer as a communication tool on preservation and furtherance of Tamil language in the modern era 137

19) Tamil language front-end to INGRES 139

20) Intelligent Tutoring System for Tamil Grammar 147

21) Spelling the difference - GIST Tamil Spell Caster 152

22) Recognition of handwritten Tamil Characters using Neural Networks 155

23) Automated Data Capturing Using Pattern Recognition 165

24) கண்காட்சியில் பங்கு பெறுபவர்கள்

தமிழும் கணிப்பொறியும்

தொடக்க உரை

எஸ். ரங்கராஜன் ("சுஜாதா")
பொறுப்பாசிரியர்,
குமுதம், சென்னை

கணிப்பொறியும் தமிழும் என்னும் இந்தக் கருத்தரங்கம் சரித்திர மதிப்பு மிக்கது எனக் கருதுகிறேன். இருபத்தோராம் நூற்றாண்டின் விளிம்பில் இருக்கும் போது இந்தியாவில் வழக்கிலுள்ள மிகப்பழைய மொழியான தமிழ் முன்னணியில் நிற்பது ஒரு வியப்புக்குரிய விஷயம் அல்ல.

தமிழ் காலப்போக்கில் மெல்ல மெல்ல மாறியபோது ஒரு பெயர் தெரியாத தமிழன் வட மொழி போல இத்தனை வரிவடிவங்கள் தேவையில்லை. கூட்டெழுத்துக்கள் வேண்டாம் என்று தீர்மானித்தபோது இந்த ஆச்சிரியத்தின் விதைகள் விழுந்திருக்கின்றன.

மற்றொரு முக்கிய நிகழ்ச்சி மதபோதகர்களின் அச்சுக்கலையின் வருகை. வீரமாமுனிவர் என்கிற பெஸ்கி அடிகளார் அச்சடிக்கும் போது சில புதிய வரி வடிவங்களை அமைத்ததும் முக்கிய நிகழ்ச்சி.

இதன் விளைவாக தமிழில் பிற்காலத்தில் கணிப்பொறிக்கு பயன்படும்படியாக இர்ண்டு விஷயங்கள் நிகழ்ந்தன. எழுத்துக்களின் வரிவடிவங்களின் எண்ணிக்கை குறைந்தது. தொல்காப்பியர் காலத்திலிருந்தே அகர முதன் னகர இறுவாய் முப்பஃது என்று ஆங்கிலத்துக்கு அருகிலேயே இருந்திருக்கிறது தமிழின் நெடுங்கணக்கு. மொத்த எழுத்துக்கள் என்று சொல்லும்போது இருநூற்று நாற்பத்து ஏழு என்று கணக்கிற்காக வைத்துக் கொண்டாலும் கணிப்பொறியாளர்களின் ஆராய்ச்சிப்படி அவைகளில் நடைமுறையில் பயன்படுவது சுமார் நூற்றுப்பத்து தான் என்று ஆராய்ச்சிக் கட்டுரைகளில் எழுதியிருக்கிறார்கள். உதாரணமாக ஸ்ரீ, மௌ போன்றவையெல்லாம் மிகமிக அரிதாகத்தான் தமிழில் பயன்படுகின்றன.

தமிழில் ஒலி வடிவங்கள் நாற்பதுதான் சொல்கிறார்கள். இது 'ஃபோனிம் சின்தஸிஸ்' ஒலியன் தயாரிப்புக்கு முக்கியமான செய்தியாகும்.

வரிவடிவங்கள் காலப்போக்கில் மாறி வந்திருக்கின்றன. கணிப்பொறியைப் பொறுத்தவரை அவைகளை இரண்டு வகையாக பிரிக்க வேண்டும். தனியே இயங்கும் வடிவங்கள் க, ப, த, ந போன்ற எழுத்துக்கள். தனிப்பட்டு இயங்கும் போது அர்த்தமில்லாத வடிவுகள், கொம்பு, புள்ளி, கால் போன்றவை.

கணிப்பொறிக்கு இவ்வகை பாகுபாடு முக்கியம். தனி வடிவம் இல்லாத விசைகளை இயக்கினால் அதை கணிப்பொறி நிராகரிக்க வேண்டும் அல்லது வேறுவகையில் தீர்க்க வேண்டும் என்கிற விதி அதன் நிகழ்ச்சி நிரலில் இருக்க வேண்டும்.

இப்படி பகுத்தால் தமிழில் வரிவடிவங்கள் பன்னிரண்டு உயிர் எழுத்துக்கள், பதினெட்டு மெய் எழுத்துக்கள், ஆய்த எழுத்து, கொம்பு, புள்ளி, கால், இரட்டைக் கொம்பு, ை தனி எழுத்துக்கள் கொண்ட கு பு மு ரு ரூ டு டூ பூ சு சூ . கசடதபற வல்லினங்கள் உ ஊ உடன் சேரும்போது தனி வரிவடிவம் பெறுவதால் மொத்தம் நூறுக்குள் எல்லா வரிவடிவங்களும் அடங்கி விடுகின்றன.

மேலும் ஆங்கிலம் போல அப்பர் கேசு இல்லாததால் விசைப் பலகையை அமைப்பதிலும் தமிழை கணிப்பொறித் திரையில் எழுதுவதிலும் மிகுந்த எளிமையைத் தருகிறது. கீபோர்டு லே அவுட், விசைப்பலகை இட ஒதுக்கம் பற்றி கொஞ்சம் விரிவாக பிறகு பேசலாம்.

தமிழில் புதிய வரிவடிவங்கள் தேவையா என்கிற ஆராய்ச்சி பலர் செய்திருக்கிறார்கள். தமிழ் ஒரு குறைபட்ட மொழி, க ச ட ப போன்றவற்றின் கன வடிவங்களுக்கு தனிப்பட்ட எழுத்துக்கள் அதற்கு இல்லை என்று அதற்காக அண்மையில் ஓர் அம்மையார் என்னை வந்து பார்த்து க என்பதற்கு கீழே இரட்டைப் புள்ளியும் க்க என்பதற்கு அருகே இரட்டைப் புள்ளி போன்ற வினோதமான விதிகள் அனைத்து இதை குமுதத்தில் பதிப்பியுங்கள் என்று சொன்னார்.

அவருக்கு நான் சொன்ன பதில், ஆங்கிலத்தில் 'கட்' என்பதற்கும் 'புட்' என்பதற்கும் யு எழுத்து வேறுவிதமாக ஒலிக்கிறதே அதற்காக இரண்டு 'யு' யாராவது கேட்டிருக்கிறார்களா? பெர்னார்டு ஷா தன் வாழ்நாளின் இறுதிப் பகுதியை ஆங்கிலத்துக்கு ஒரு புதிய ஃபோனெட்டிக் ஆல்ஃபபெட் அமைப்பதில் செலவழித்து ஏமாற்றத்தில் இறந்து போனார். அதற்காக தன் சொத்தின் பகுதியை எழுதி வைத்தார். யாராவது வெற்றி கண்டால் பணம் கிடைக்கும். அவரால் அதைக் கொண்டு வரமுடியவில்லை. காரணம் ஒரு மொழியின் வரிவடிவம் காலப்போக்கில்தான் மாறுகிறது. ஒரு ராத்தியியில் .அதை மாற்ற முடியாது. அதற்கான அவசியம் ஏற்பட்டு எழுத்தாளர்களும் வல்லுனர்களும் அச்சடிப்பவர்களும் ஒத்துழைத்து மெல்லத்தான் மாற்ற முடியும்.

மலையாளத்தில் அண்மையில் அச்சுக் கலைக்காக கொண்டு வந்த மாற்றங்களை மக்கள் இன்னும் கடிதங்கள் எழுதும்போது பயன்படுத்துவதில்லை. விளைவு புத்தகத்தில் ஒரு மலையாளமும் கடிதங்களில் ஒரு மலையாள வரி வடிவமும் உள்ளது. வரி வடிவ மாற்றங்கள் அங்கீகாரத்துக்கு இரண்டு தலைமுறை கழிய வேண்டும். இப்போதும் தமிழில் கடிதம் எழுதும் போது பழைய லை, னை களைத் தான் பயன்படுத்துகிறோம்.

மேலும் தமிழில் இனியும் எழுத்துக் குறைப்பு தேவையில்லை என்பது என் கருத்து. எழுத்து அதிகப்படி வேண்டுமா என்றால் அதன் தேவையை எல்வோரும் உணர்ந்து மெல்ல செயல்பட வேண்டிய காரியம் இது.

கணிப்பொறியைத் பொறுத்த வரையில் எத்தனை புதிய வரிவடிவங்கள் பொண்டு வந்தாலும் அத்தனையும் சமாளிக்கும் திறமை உள்ளது. கேளுங்கள் கொடுக்கிறோம். சீன, ஜப்பானிய மொழியின் பட எழுத்துக்களையே சமாளிக்கும் போது தமிழ் வரிவடிவங்களை சமாளிப்பது எம்மாத்திரம். அதனால் கணிப்பொறிக்கு இனி எழுத்துக் குறைவு ·தேவையில்லை. அதிகப்படி வேண்டுமென்றால் கணினியாளர்களின் பாஷையில் நோ ப்ராப்ளம்.

கணிப்பொறிக்கு ஒலியன்களை அமைக்கும்போது முக்கியமாக தேவைப்படுவது ஓர் உச்சரிப்பு அகராதி, அகரமுதலி. இதுபற்றி விவரமாகப் பின்னால் சொல்கிறேன்.

பதினெட்டு வரிவடிங்களே உள்ள மெய் எழுத்துக்கள் இருபத்தெட்டு ஒலி வடிவங்கள் பெறுகின்றன. அதை ஒரு குறை என்று சொல்ல முடியாது. உலகில் எந்த மொழியும் குறைபாடற்ற நெடுங்கணக்கை பெறவில்லை என்று சொல்லலாம். சமஸ்க்ருதத்தைச் சொல்கிறார்கள். ஆனால் சமஸ்க்ருதம் வழக்கொழிந்து விட்டது.

அடுத்து விசைப்பலகை இட ஒதுக்கீடு பற்றி கொஞ்சம் யோசிக்லாம். தற்போது கணிப்பொறி இயலின் பார்வையிலிருந்து இது முக்கியமாக கவனிக்க வேண்டிய பிரச்சினை, நிலையான, எல்லோருக்கும் பொதுவான ஓர் இட ஒதுக்கீட்டுத் திட்டம். இப்போது தமிழில் டிடிபி DTP என்னும் இயல் கண்ட மேனிக்கு வளர்ந்திருப்பதற்கு முக்கியமான காரணம் விஸ்சைப்பலகையின் லே அவுட்டை யாரும் ஸ்திரப்படுத்தாதது என்று அடித்துச் சொல்வேன்.

என் அனுபவத்தில் நான் ஏறக்குறைய எட்டு வருஷங்களாக கதை கட்டுரைகள் எழுத கணிப்பொறியை பயன்படுத்தி வருகிறேன். எல்லா பத்திரிகைகளுக்கும் எழுதுகிறேன். மிகச் சுலபமாக அந்த பத்திரிகை ஆபீசில் போய் அவரவர் கணிப்பொறியில் என்னால் கதை கட்டுரைகளை அமைத்து அடித்துக்

கொடுத்துவிட்டு வர முடியும் என்று நான் சென்னைக்கு வந்த போது எண்ணினேன். ஆனால் வந்தபோதுதான் அது சாத்தியமில்லை என்று அறிந்தேன். காரணம் ஒவ்வொரு பத்திரிகை அலுவலகத்திலும் ஒவ்வொரு விசைப்பலகை. இதுவரை ஆறு இட ஒதுக்கீடுகளைச் சந்தித்திருக்கிறேன்.

நான் என் கணிப்பொறியில் அச்சிடுவதை மீண்டும் அவர்கள் அவர் கணிப்பொறியில் அடித்துக் கொள்கிறார்கள். இது ஒரு பெரிய நேர விரயம். தமிழ் டிடிபி அமைப்பவர்கள் எல்லோரும் அரசாங்கமோ, கல்வியாளர்களோ முன் வராததால் தயாரிப்பாளர்கள் ஒன்று சேர்ந்து இதில் ஒரு தர நிர்ணயம் ஸ்டாண்டர்டைஸேஷன் செய்ய வேண்டும். இல்லையேல் கணிப்பொறியாக்கத்தின் முழுப் பயனையும் நம்மால் பெற முடியாது. ஆங்கில சூழ்நிலையில் இதை நினைத்துப் பார்த்தால் பிரச்சினையின் அபத்தம் புரியும். ஒவ்வொரு பத்திரிகை ஆபீசிலும் வெவ்வேறு விசைகளில் ASDF இருப்பது போல!

விசைப்பலகையில் எழுத்துக்கள் ஒதுக்கும்போது வரிவடிவ ஒதுக்கீடா ஒலிவடிவ ஒதுக்கீடா என்கிற முக்கியமான தீர்மானம் தேவைப்படுகிறது. 'இன்ஸ்கி' என்னும் இந்திய மொழிகளுக்கான பொது அமைப்பில் தட்டச்சு பலகை லே அவுட் இந்திய மொழிகளுக்கு இடையே செய்திப் பரிமாற்றத்துக்கு உதவும்.

ஆனால் இது இயற்கையாக தமிழை எழுதுவதற்கு புறம்பான முறை. மெய்யெழுத்தை முதலிலும் உயிர் எழுத்தை பின்பும் இயக்க வேண்டும். ஒற்றெழுத்துக்களை அறிய தனிப்பட்ட விதிமுறைகளும் உண்டு. இது பகுத்தறிவின்படி மிகவும் பாராட்டத்தக்க முறையானாலும் ஒரு மொழியின் வளர்ச்சி எப்போதும் பகுத்தறிவின் பால் ஏற்படுவதில்லை.

தமிழைப் பொறுத்தவரை லட்சக்கணக்கான தமிழர்கள் தட்டச்சு இயந்திரத்தின் லே-அவுட்டுக்கு பழகியுள்ளார்கள். அதை கணிப்பொறிக்குப் பயன்படுத்தும் போது அதிலுள்ள டெட் கீ விசைப் பலகை நகராத முறை கூட தேவையில்லை.

தமிழை நாம் எப்படி எழுதுகிறோமோ அப்படியே அச்சடிக்கும் வரிசையும் அமைக்க முடியும் எனபதற்காகவே என் கருத்தில் தட்டச்சு இயந்திர இட ஒதுக்கீட்டை வலுவாக பரிந்துரைக்கிறேன். நேரடியாக கீ போர்டு பழகிய லட்சக்கணக்கான தமிழர்களுக்கு மறுபயிற்சி தேவையில்லாமல் இருக்கும்.

இன்ஸ்கி மற்ற மொழிகளோடு தொடர்பு கொள்வது எப்படி? தமிழ் மொழிக்கு மட்டும் மற்ற இந்திய மொழிகளிலிருந்து விலகி தனியே அமைத்துக் கொண்டால் நாம் ஒதுக்கப்படுவோமே என்று கேட்கலாம். நவீன கணிப்பொறி இயலில் அதெல்லாம் பிசினையே இல்லை. இரு மைக்ரோ கண்ட்ரோலர் மூலம்

எந்த லே அவுட்டையும் பயன்படுத்தி அதை இன்ஸ்கி சங்கேதங்களுக்கு மாற்றி அமைக்க ஒரு ஆஸிக் சில்லு அமைக்க முடியும். லே அவுட் என்பது நம் சௌகரியங்களுக்கு அமைக்க வேண்டுமே தவிர மற்ற மொழியாளர் சௌகரியத்துக்காக இல்லை.

கணிப்பொறியை தமிழுக்கு பயன்படுத்தும்போது இந்த சிந்தாந்தத்தை முக்கியமாக பயன்படுத்த வேண்டும். கணிப்பொறிக்கேற்ப தமிழை மாற்றாமல் தமிழுக்கேற்ப கணிப்பொறியை மாற்றுவதுதான் முக்கியம்: தமிழை எப்படி எழுதுகிறோமோ அதன்படி லேஅவுட் அமைக்கலாம். அது தட்டச்சு இயந்திரத்தின் லே அவுட்டா அல்லது நா. கோவிந்தசாமி பரிந்துரைக்கும் லே அவுட்டா? எதாவது இரண்டிற்குள் ஒன்றை இந்த கருத்தரங்கில் பரிந்துரைத்தால் அதுவே பெரிய சாதனை என்பேன்.

அவர்கள் விவாதித்து சம்மதம் இல்லை எல்லாரும் இன்ஸ்கியைத்தான் பயன்படுத்த வேண்டும் என்றாலும் என் போன்றோர்களுக்கு கோபம் வராது. எல்லோரும் ஒரே லேஅவுட்டு பயன்படுத்துவதுதான் மிகமிக முக்கியம். தயவு செய்து எல்லோரும் ஒரே ஒரு விசைப்பலகை இட ஒதுக்கீட்டை பயன்படுத்துவோம். வாருங்கள்.

அடுத்து தமிழ் மொழியின் இலக்கணம் கணிப்பொறியின் நோக்கிலிருந்து.

தற்போது எழுதப்படும் இலக்கண நூல்களை நாம் பார்க்கும் போது ஒருவிதமான தொல்காப்பிய சுமை, தொல்காப்பிய ஹாங் ஓவர் அல்லது நன்னூல் சுமை இருப்பதைக் காண்கிறோம். எல்லா இலக்கணங்களும் அவைகூறும் விதிகள் தொல்காப்பியத்திலோ நன்னூலிலோ இருக்கும் விதிகளைத்தான் சொல்கின்றன. சில எளிமையாக, சில கடினமாக.

தற்காலத் தமிழின் இலக்கணத்தை யாரும் சரியாக எழுதவில்லை என்று தோன்றுகிறது. உதாரணமாக வேற்றுமை உருபுகளை எட்டு என்று கணக்கிடுகிறார்கள். இவைகளில் சில வேற்றுமைகள் கால மாறுதல்களால் வழக்கொழிந்து கொண்டிருக்கின்றன. சில மறைந்து ஒலிக்கின்றன. இரண்டாம் வேற்றுமையான 'ஐ' நிலத்தைக் கடந்தான் என்பதற்குப் பதில் நிலம் கடந்தான் போன்ற பிரயோகங்கள் இன்று வருகின்றன. அதை வேற்றுமைத் தொகை என்கிறார்கள்.

வேற்றுமைகள் தொக்கு நிற்பது மட்டுமின்றி சில வேற்றுமைகள் அழிந்துபோய் விட்டன. யாரும் பயன்படுத்துவதில்லை. தற்காலத் தமிழில் ஐந்தாம் வேற்றுமைகளான இன் இல் என்பவை மறைந்துகொண்டு வருகின்றன. ஆறாம் வேற்றுமையான 'அது' 'இன்' என்பவை மறைந்து 'உடைய' என்பது

மெல்ல வழக்கொழிந்து வருகிறது. இது யார் புத்தகம்? ராமன் புத்தகம் என்கிறோமே தவிர யாருடைய புத்தகம்? ராமனுடைய புத்தகம் என்பது அருகி வருகிறது.

எட்டாம் வேற்றுமையில் விளிவேற்றுமைகளில் 'ஆ' மட்டும்தான் உள்ளது. பூவாய், ஐயவோ என்று யாரும் இப்போது எழுதுவதுகூட இல்லை.

தமிழ் பேச்சுத் தமிழுக்கும் எழுத்துத் தமிழுக்கும் கணிசமான வெறுபாடுள்ள மொழி. அங்கே போனான் என்றுதான் பேசுகிறோம். அங்கே சென்றான் என்று எழுதுகிறோம். வீரமாமுனிவர் காலத்திலேயே இதை உணர்ந்து அவர் 'கொடுந்தமிழ்' இலக்கணம் எழுதினார். இன்றும் ஒரு கொடுந்தமிழ் இலக்கணம் தேவைப்படுகிறது. பெயரை மட்டும் மாற்றி விடுவோம்.

மேடையில் பேசும்போது அதற்கென்று தனிப்பட்ட மிகையான சாரியை சந்திகளும் பயின்று 'வாழ்ந்திட்டிட' 'சென்றிட்டிட' என்றெல்லாம் அலங்கரிக்கிறோம். இதற்கெல்லாம் சமூக நிலைக்காரணங்களும் சிக்கலான அரசியல் காரணங்களும் உண்டு.

ஒரு கணிப்பொறியாளன் இதை விமர்சிக்க வேண்டியதில்லை. இந்த சூழ்நிலையில் கணிப்பொறியுடன் நாம் தமிழில் எவ்வாறு பேசப்போகிறோம் என்று எதிர்காலத்தில் தீர்மானிக்க வேண்டும். அதற்கேற்ப ஓர் எளிய இலக்கணத்தை அமைக்க வேண்டும். அதற்கென்று சுமார் ஆயிரம் வார்த்தைகளும் எளிய இலக்கண விதிகளும் எழுதப்பட வேண்டும். கணிப்பொறிக்கு ஏற்றவாறு அது பேச்சுத் தமிழ், எழுத்துத் தமிழின் கலவையாக இருக்கலாம். எப்படியும் அதற்கேற்ப ஓர் இலக்கணம் எழுதுவது முக்கியம்.

கிரியா தற்காலத் தமிழ் அகராதி செய்தது போல தற்காலத்தமிழ் இலக்கணம் அமைக்க வேண்டியது முக்கியம். கணிப்பொறி மூலம் தமிழைப் பேசவைக்கும் முயற்சிக்கு முக்கியமான தேவை.

இவ்வாறான இலக்கணம் ஒன்று ஒரு ஜெர்மானியத் தமிழறிஞர் எழுதி ஸிஎல்எஸ் வெளியிட்டிருக்கிறதாக அறிந்தேன். கணிப்பொறி மூலம் தமிழ் பேச வைக்கும் முயற்சிக்கு மற்றொரு முக்கிய தேவை ஒலிப்பு அகராதி Orthoepical அல்லது உச்சரிப்பு அகராதி தேவைப்படுகிறது. கஞ்சி, கட்சி, கட்டு, கண்டு, கத்து, கந்து, அப்பன், அம்பு போன்றவைகளில் வல்லினங்கள் வேறு வேறு ஒலி பெறுவதை நாம் இயல்பாக உச்சரிக்கிறோம். கணிப்பொறிக்கு இவைகளை விதிகளாக வெளிப்படையாகச் சொல்ல வேண்டும்.

இதற்கானவை எழுதப்படா விதிகளாக இருக்கின்றன. இவைகளைத் தெளிவாகச் சொல்ல வேண்டும். குறிப்பாக அகம், முகம், ஆகம் என்னும் சொற்களில் வரும் கவும் எஃகு போன்ற சொற்களின் கவும் வேறுபடுகிறது.

அதேபோல் பசி ஊசி கட்சி இவைகளில் உள்ள சகர ஓசையை நோக்கலாம். தமிழில் எனக்குத் தெரிந்தவரை இவ்வகையான ஒலிப்பு அகராதி அண்மையில் வரவில்லை. அங்கெடல் (Anketell) என்பவர் 1888ல் ஆங்கிலத் தமிழ் ஒலிப்பு அகராதி ஒன்றை வெளியிட்டார். மார்டினெட் என்பவர் தமிழ் ப்ரெஞ்சு ஒலி அகராதி வெளியிட்டிருக்கிறார் 1897ல். வேறு அகராதிகள் இத்துறையில் இல்லை. மு. சதாசிவம் அகராதி வெளியிட்டிருக்கிறார் எனத் தெரிகிறது.

எனவே தமிழறிஞர்களிடம் வேண்டுகோள்: எங்களுக்குத் தேவைப்படுவது தமிழின் சராசரி ஆண் பெண்கள் பேசும் வகை ஒலியன்களின் பட்டியல். அவைகளுக்கான குறியீடுகளை அமைத்து தமிழ் எழுத்துக்களின் விதவிதமான ஒலிகளை விளக்கிப் பதிவிடும் ஒலிக்குறிப்பு அகராதி இருந்தால்தான் சின்தசைஸர் என்னம் சில்லு அமைக்க முடியும். மேல்நாட்டில் விற்கும் வோட்ராக் போன்ற சின்தஸைஸர் சில்லுகளைப் பயன்படுத்தினால் அவைகள் ஆங்கில ஃபோனிம்களின் வழி உருவானதால் வெள்ளைக்காரன் தமிழ் பேசுவது போல் ஒலிக்கும்.

சென்னைப் பல்கலைக்கழக தமிழ் பேர்கராதியில் எழுத்துப் பெயர்ப்பு தரப்பட்டுள்ளது. ட்ரான்ஸலிட்டரேஷன் - ஒலிப்பு முறையை சுட்டும் நேரடி தமிழ் அகராதி இயலில் இன்னும் தோன்றில்லை எனலாம். கிரியாவின் தற்காலத் தமிழகராதி ஒலிக் குறிப்பை எப்போதாவது கொடுக்கும்போது ஆங்கில ட்ரான்ஸலிட்டரேஷன் முறையைத்தான் பயன்படுத்துகிறது.

தமிழ் எழுத்துக்களை கணிப்பொறியின் மூலம் பரிச்சயம் செய்து கொள்ளும் ஆப்டிக்கல் காரக்டர் ரெகக்னிஷன் முறைகளைப் பற்றி முதல் ஆராய்ச்சிக் கட்டுரைகள் எனக்குத் தெரிந்தவரை நவம்பர் 1984ல் ஐஈடிஈ இதழ் ஒன்றில் தோன்றியது. அதன் பொது எடிட்டரான ஆர்.எம்.கே. சின்ஹா இந்த இயலில் குறிப்பிட வேண்டிய ஆராய்ச்சியாளர். இதற்கு அடிகோலியான சில முயற்சிகள் அவ்வப்போது நிகழ்ந்திருந்தாலும் 1984ல் தான் முதலில் ஒரு முழு ஆராய்ச்சிக்கட்டுரை தோன்றியது. அந்தக் கட்டுரையின் மூன்று ஆராய்ச்சியாகளில் ஒருவரான கிஃப்ட் சிரோமணி என்கிற அறிஞரை இப்போது நினைவுகொள்ள வேண்டும். கணிதப் பேராசிரியரான இவர் மொழி இயலாளர் கம்ப்யூட்டர் இயலாளர் கல்வெட்டியல் எபிக்ரஃபி எல்லாத் துறைகளிலும் ஆராய்ச்சிக் கட்டுரைகள் எழுதியுள்ளார். அண்மையில் காலமான அவர் வாழ்நாளில் அவருடைய ஆராய்ச்சிப்பணியை மேலைநாட்டில் பலர் அறிந்திருக்கிறார்கள் சிரோமணி. அவர்கள் தமிழ் உரைநடையின் எண்ட்ரப்பி செய்தி மாறும் அளவைப் பற்றியும் தமிழ் மொழியின் கெர்னல் என்னும் விதை அமைப்புக்களைப் பற்றியும் தமிழ் ப்ரம்மி வரிவடிவங்களைப் பற்றியும் பல அரிய கட்டுரைகள் எழுதியுள்ளார். இன்ஃபர்மேஷன் தியரி certain applications of information theory என்னும் அவருடைய முனைவர் பட்டத்துக்கான ஆராய்ச்சிக்

கட்டுரை சிறந்தது. பல்லவ கிரந்த எழுத்துக்களை கம்ப்யூட்டர் முறைகள்படி வடிவு தெளிவாக்குதல் பற்றியும் ஆராய்ந்திருக்கிறார். தமிழ் எழுத்துக்கள் ஆளுமை எண்ணிக்கையை கொண்டு ஃப்ரிக்வென்ஸி அனாலிஸிஸ் செய்து அதற்கான 'ஹஃப்மன் கோடை' யும் அமைத்திருக்கிறார். இந்த விந்தை மனிதரின் பணியை நாம் தொடர்ந்து செய்ய வேண்டும்.

தமிழில் கன்ஃப்யுஷன் மாட்ரிக்ஸ் ஒன்று அமைக்க வேண்டும். தமிழில் உள்ள அத்தனை எழுத்துக்களிலும் ஒசிஆர் மூலம் பரிச்சயம் செய்துகொள்ள வேண்டியவை கீழ்க்காணும் வரி வடிவங்களே.

அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ ஓ ஃ	-	(12)
க ங ச ஞ ட ண த ந ப ம ய ர ல வ ழ ள ற ன	-	(18)
கி சி டி ணி தி நி பி மி யி ரி லி வி ழி ளி றி னி	-	(16)
கீ சீ டீ ணீ தீ நீ பீ மீ யீ ரீ லீ வீ ழீ ளீ றீ னீ	-	(16)
கூ சூ டூ ணூ தூ பூ நூ மூ யூ ரூ லூ வூ ளூ தூ றூ னூ	-	(16)
கு சு டு ணு து நு பு மு யு ரு லு வு ழு ளு று னு	-	(16)
ா ெ ே ை.	-	(5)

இந்த 99 வரிவடிவங்களை அடையாளம் கண்டுகொள்ள முடிந்தால் அச்சடித்த தமிழ்ப் பக்கத்தை கணிப்பொறியால் படிக்க முடியும். 99 x 99 சதுரம் மாட்ரிக்ஸ் அமைத்து எந்த எழுத்து எந்த எழுத்துடன் குழப்பமாகிறதோ அவைகளை எல்லாம் குறிப்பிட வேண்டும். உதாரணமாக ெ கொம்பும் எ எழுத்தும் குழம்பலாம் சவும் சுவும் கவும் குழம்பலாம். இம்மாதிரியான சதுரமும் எழுத்துப்பொருத்தங்களைக் கொண்ட டை கிராம் ட்ரை கிராம் என் - கிராம்களும் அமைக்கவேண்டும். உதாரணமாக கவும் ஞ்ம் பொருந்தும் இது டைகிராம். கஞ் என்று அடையாளம் கண்டுகொண்டபின் அதையடுத்து பல எழுத்துக்கள் பொருந்தா. பெரும்பாலும் சகரத்துடன் சம்பந்தப்பட்ட எழுத்தாகத் தான் இருக்கும் அது ட்ரைகிராம் இவ்வாறு அமைக்கப்படும் விதிகளை Disambiguation விதிகள் என்று சொல்வர்.

'குழப்ப நீக்க' விதிகள் சந்தேக நீக்க விதிகள் என்று சொல்லலாம். இவைகளனைத்தையும் சளைக்காது பட்டியலிட்டால் 'ஒஸிஆர்' இயலுக்கு மிகவும் உதவும். தமிழில் எழுதப்படாத விதிகள் பல உள்ளன. வார்த்தைகளின் முதலில் சில எழுத்துக்கள் வரவே வராது. வார்த்தைகளின் மத்தியில் சில

எழுத்துக்கள் வாரா. உதாரணமாக உயிர் எழுத்துக்கள், அளபெடை போன்ற அரிய சந்தர்ப்பங்களைத் தவிர அதேபோல் வார்த்தைகளின் இறுதியிலும் சில வாரா விதிகளைக் கொண்டு இவ்வாறாக குழப்ப நீக்கம் செய்ய முடியும். இவைகள் மூலம் 'ஸ்பெல் செக்' எழுத்துப்பிழை நீக்கம் செய்ய முடியும்.

இந்த விதிகளை எழுத தமிழறிஞர்கள் உதவி கணிப்பொறியார்களுக்கு தேவைப்படுகிறது. இந்த விதிகளைப் பயன்படுத்தி 'அடோபி' போன்ற சில மென் பொருள் அமைப்பின் மூலம் தமிழ் எழுத்துக்களை ஒரு 'ஸ்கானர்' மூலம் வருடி அதை ஒவ்வொரு எழுத்தையும் காரக்டர் மேட்ரிக்ஸில் அடக்கி அவைகளின் சாரத்தை எடுத்து இந்த எழுத்தாக இருக்கலாம் என்று கண்டு சந்தேகமிருந்தால் உபரியான விதிகளை பயன்படுத்தி ஏறத்தாழ 100 விழுக்காடு கணிப்பொறி மூலம் தமிழ் எழுத்து பரிச்சயம் முடியும் என்று தோன்றுகிறது. தமிழ் அச்செழுத்து 'நான்கர்சிவ்' தனித்தனியாக எழுதப்படுவது ஒரு வரப்பிசாதமாக இருக்கிறது. கணிப்பொறி இயலில் கிடைக்கும் பிஸி ஆட் ஆன் கார்டுகளை ஓஸிஆர் கார்டுகளை பயன்படுத்தி கணிப்பொறி மூலம் தமிழ் படிக்க முடியும் என்பது என் கணிப்பு.

அடுத்து தமிழில் தகவல் தளங்கள் அதில் பல முயற்சிகள் நடந்து வருவது பற்றி அறியமுடிகிறது. திருக்குறள் முழுவதையும் சிலப்பதிகாரம் முழுவதையும் தகவல் தளம் அமைத்து வார்த்தைகளை வகைப்படுத்தி எத்தனை முறை அவை பயன்படுத்தப்படுகின்றன, கன்கார்டன்ஸ் போன்ற பலவகை சாத்தியங்கள் மற்ற மொழி இலக்கியத்தில் உள்ளது போல் தமிழிலும் செய்ய முடியும். தகவல் தளம் டேட்டாபேஸ் என்பது பெரிய விஷயம். எதிர்காலத்தில் இந்த தகவல் தளங்கள் ஒரு பொது சொத்தாக சொல்வார்கள். இப்போதே தமிழில் இருக்கும் தகவல் தளங்கள் தமிழ்நாட்டு மூலிகைகள், மருந்துகள் பழைய வீரமாமுனிவர் போன்றோரின் சொற்கள், நிகண்டுகள், பொருளகராதி, தொடையகராதி, போன்றவையிலிருந்து பிரபலமானவர் விலாசங்கள் வரை பல வேறு தகவல் தள அமைப்புக்கள் வரமுடியும். அது நெட்வொர்க் என்னும் பொது தகவல் வங்கியில் இவைகள் கிடைத்தால் பல்வேறு துறை ஆராய்ச்சியாளர்களுக்கு பயன்படும்.

அடுத்து நவீன கணிப்பொறி இயலில் ஒரு சொல் பல பொருள், ஒரு பொருள் பல சொல் வகை பொருளகராதிகள் தேவை. சுமார் நாற்பது நிகண்டு வகைகள் இருந்தன. அவை அனைத்தும் பட்டியலிடும் சொற்கள் வழக்கொழிந்த சொற்களாகவே பாதிக்கு மேல் இருக்கின்றன. உதாரணத்துக்கு திவாகர நிகண்டில் உண்பதை அர்த்தமாகக் கொண்ட சொற்களை பட்டியலிடுகிறபோது 'துய்த்தல், அருந்தல், ஆர்தல், துவ்வுதல், தூற்றல், நுகர்தல், மாந்தல், அயிறல், உண்டல்,

தின்றல், பருகுதல்' இவை உணும் பண்பிற் தொழிலைப் பகருங்கிளவி, என்கிற அருமையான நூற்பாவில் தற்போது அதிகமாகப் பயன்படும் 'சாப்பிடுதல்' இல்லை. இப்போது எவரும் துவ்வுதல், மாந்தல் போன்றவற்றை பயன்படுத்துவதில்லை. பொதுவாக நமக்கு கிடைக்கும் அகராதிகளும் இலக்கணங்களும் சரித்திர வகைப்பட்டதே தவிர (Historic) சமகால Synchrowoos வகை அல்ல. நமக்குத் தேவை சமகால நிகண்டுகளும் சமகால இலக்கண புத்தகங்களும். திவாகர நிகண்டு வேண்டாம் கணிப்பொறி நிகண்டு வேண்டும். எதற்காக இந்த நிகண்டு?

'ஸ்டைல் செக்' என்று ஒரு மென்பொருள் உள்ளது. தமிழில் ஒரு கடிதமோ கட்டுரையோ எழுதும்போது ஒரே சொல் பலமுறை திரும்ப வந்தால் அழகாக இருக்காது என்று மாற்று வார்த்தைகளை யோசனை கூற கணிப்பொறியைப் பயன்படுத்துகிறார்கள். ஒரு முழு 'தெஸாரஸ்' என்னும் ஒரு சொல் பல பொருள் அகராதி கணிப்பொறியின் சொற்தொகுப்பு மென்பொருளுடன் இலவச இணைப்பாக கொடுக்க முடியும்.

இந்தக் கட்டுரையின் முடிவுரையாக நான் கூற விரும்புவது நான் கணிப்பொறியாளன் ஓர் எழுத்தாளன் என்கிற, இரண்டு சுமாரான தகுதி படைத்த, எனக்கு தமிழின் கணிப்பொறியாக்கத்திலுள்ள சவால்களையும் அதை இந்த நூற்றாண்டில் செய்யமுடியும் என்கிற திடமான நம்பிக்கையும் கொண்ட எனக்கு நான் படித்த வரை கீழ்க்காணும் வசதிகள் தேவைப்படுகின்றன.

1. ஒரு நல்ல தற்கால தமிழ் இலக்கணம்.

2. ஒரு புதிய நிகண்டு.

3. அடையாள விதிகள், சந்தேக நீக்க விதிகள், குழப்ப மேட்ரிக்ஸ் டைகிராம், ட்ரைகராம் என்ற என்கிராம்கள் தமிழின் எழுத்துக்கள் பயன்படும் எண்ணிக்கை அலசலின் பாற்பட்ட ஒரு ஹஃப்மன்கோட் பரிந்துரைக்கு ஏற்ற ஒரு தமிழ் எழுத்து பயன்பாட்டு எண்ணிக்கை எண்ட்ரபி ஆர்தபிக் அகராதி என்ற உச்சரிப்பு அகராதி, தமிழின் ஃபோனிம்களை வகைப்படுத்தி அவைகளுக்கான குறியீடு.

இவை அனைத்தும் இருந்தால்தான் தமிழை அடுத்த நூற்றாண்டுக்கு அழைத்துச் செல்ல முடியும்.

நன்றி வணக்கம்.

"கணித்தமிழ் - தமிழில் ஒரு கணிப்பொறி மொழி

க.கோபிநாத், கம்ப்யூட்டர் டிவிஷன்.
ஆர்.வி.எஸ். பொறியியற் கல்லூரி. திண்டுக்கல்,

எம். அய்யாதுரை, தலைவர், கணிப்பொறித்துறை,
தியாகராசர் பொறியியற் கல்லூரி, மதுரை - 625 015.

கணிப்பொறியில் இயற்கை மொழியில் உரையாடல் என்பது தற்போதைய ஐந்தாவது தலைமுறை கணிப்பொறியாளர்களின் முயற்சியாகும். அந்த தத்துவத்தை நோக்கிச் செல்லும்பொழுது, நமது மாநில மொழிகளிலேயே கணிப்புகள் செய்வதற்கு ஆய்வுரை அறிவுறுத்துகிறது.

மிகப் பழமையான மொழியான தமிழ்மொழியின் மூலமாக 'புரோகிராம்' எழுதுவதற்கு ஏற்றவாறு ஒரு 'கம்பைலரை' தயாரிக்க இந்த உரை முனைகிறது. அதற்கு 'கணித்தமிழ்' என்ற பெயரிட்டு முதற்கட்டமாக ஆய்வுப்பணி தொடங்கப்படுகின்றது. மற்ற கணிப்பொறி மொழிகளைப் போலவே., கணித்தமிழும் "காண்டக்ஸ்ட்ஃபிரி" இலக்கணத்தைச் சார்ந்துள்ளது. மேலும், இந்த ஆய்வு எந்த மாநில மொழிகளிலும் புரோகிராம் எழுதுவதற்கு ஏதுவாக சிம்பல்டேபிள், ரிசர்வ்ட் வார்த்தைகள் போன்றவற்றில் சில மாறுதல்கள் செய்யவும் முன் வருகின்றது.

தற்போதுள்ள கணிப்பொறிகளின் மூலம் தமிழ் எழுத்துக்களை கையாள்வது இயலாததால் (பாரதி - போன்ற கணிமங்கள் இருந்தாலும் கூட) ஆங்கில தட்டெழுத்துக்களையே (டிரான்ஸ்லிட்டரேஷன் முறையில்) புரோகிராம் எழுத வேண்டியுள்ளது. எனவே, 'கணித்தமிழுக்கென்றே' எடிட்டர் ஒன்றை இணைக்கவும் இந்த ஆய்வுரை வழி காண்கிறது. அப்படி செய்யும்பொழுது நேரடியாகவே தமிழ்மொழியில் எழுதிய புரோகிராம்களை உட்புகுத்தல் வழியை இந்த ஆய்வு வலியுறுத்துகிறது.

1.0 முன்னுரை:

கணிப்பொறிகள் ஆய்வுக் கூடங்களிலிருந்து ஒரு சாமானிய மனிதரை சென்றடைய வேண்டும் என்ற நோக்கத்தோடு செயற்படும் இக்காலத்தில், அந்த மனிதனின் தாய்மொழியிலேயே கணிப்பொறியில் கணிப்பு செய்ய ஏதுவாக ஒரு கணிப்பு மொழியினை உருவாக்க இந்த ஆய்வுக் கட்டுரையில் முனைந்துள்ளோம். ஏனெனில், இன்றும் பலருக்கு தாய்மொழி ஒன்றே தொடர்பு மொழியாக இருக்கின்றது.

தாய்மொழியிலேயே அல்லது பிராந்திய மொழியிலேயே கணிப்பொறியுடன் உரையாடுவது என்பது எதிர்காலத்தில் தவிர்க்க முடியாத வளர்ச்சி ஆகும். அந்த இலக்கை நோக்கி பயணிக்கும் போது, முதல் அடியாக ஒரு கணிப்பு மொழியினை நம் தாய் மொழியான தமிழிலேயே உருவாக்கி, அதற்கு கணித்தமிழ் என்றும் பெயரிட்டுள்ளோம். மேலும், இதன் மூலமாக கணிப்பு மொழிகள் அல்லது ஆங்கிலம் அறியாதவர்கள். கணிப்பு செய்ய ஏதுவாகின்றது.

இந்த அமைப்பு, ஒரு மொழிமாற்றி (கம்பைலர்)க்குரிய அனைத்து கட்டங்களையும் (படிகளையும்) ஏற்றதாக உருவாக்கி முனைந்துள்ளோம். இம்மொழி மாற்றி தமிழ் புரோகிராம்களை கணிப்பொறிக்கு ஏற்றவாறு மாற்றம் செய்கின்றது. இதில் இரண்டு வகையான முறைகளில் புரோகிராம்களை தயாரிக்கலாம். ஒன்று, புரோகிராம்களை தமிழ் எடிட்டர் மூலமாக உட்புகுத்தி, அவற்றை டிரான்ஸ்லிட்டரேஷன் முறையில் ஆங்கில எழுத்துக்களாக மாற்றிய பின் "கணித்தமிழ்" மொழிமாற்றிக்கு கொடுக்கலாம். இந்த முயற்சியில் நிறைய ஆய்வுகள் நடந்துள்ளன. இரண்டாவது வகையில் நாம் நேராகவே டிரான்ஸ்லிட்டரேஷன் செய்யப்பட்ட புரோகிராம்களை 'கணித்தமிழ்' மொழிமாற்றிக்கு கொடுக்கலாம். இதற்கான அட்டவணையை உபயோகித்து தமிழ் வார்த்தைகளுக்கு ஆங்கில எழுத்துக்களில் ஒலிமாற்றம் (டிரான்ஸ்லிட்டரேஷன்) செய்தல் வேண்டும். இது சற்று கடினமான காரியமே. ஆனால், இந்த மொழிமாற்றியை மலையாளம் அல்லது வேறு எந்த மொழியிலும் உபயோகிக்க, சிம்பல் டேபிள் கருவை சற்று மாற்றம் செய்தால் போதுமானது. தமிழுக்கு பதிலாக வேறு ஒரு மொழியின் புரோகிராம்களை ஆங்கிலத்திற்கு (டிரான்ஸ்லிட்டரேஷன்) செய்தபின் அவற்றை மொழிமாற்றிக்கு கொடுக்கலாம். இப்படி செய்வதன் மூலம் பொதுவான மொழி மாற்றி ஒன்று உருவாக்கப்படுகிறது.

இந்த ஆய்வு கட்டுரையில் முக்கிய நோக்கம் கணித்தமிழ் மொழியின் செயற்பாட்டினை விளக்குவதாக அமைக்கப்பட்டுள்ளது. மொழி மாற்றியினுடைய நுட்பங்களை விவரிப்பது என்பது இரண்டாம் பட்சமாக கொண்டுள்ளோம்.

2.0 கணித்தமிழ் விளக்கம்

கணித்தமிழில் தயாரிக்கப்பட்ட புரோகிராம்கள் தனித்தனியே அமையப்பட்ட தொடர்ச்சியான கட்டளைகள் அல்லது வாக்கியங்கள் ஆகியவற்றின் தொகுப்பு ஆகும். ஒவ்வொரு வாக்கியமும் தனிப்பட்ட முழு எண் என்றால் தொடங்கப்படுகிறது. அந்த முழு எண்ணே வாக்கிய எண் ஆகும். இந்த புரோகிராமிற்கு பெயரிடப்படும் முறை அதிகப்படியாக எட்டு ஆங்கில எழுத்துக்கள் இருக்க வேண்டும். கணித்தமிழின் எழுத்து வடிவங்கள் மற்ற கணிப்பொறி எழுத்து வடிவங்கள் போல் (குறிப்பாக பேஸிக் மொழியினுடைய இலக்கணத்தை ஒட்டி) அமைக்கப்பட்டுள்ளன.

2.1 கட்டளை வாக்கியங்கள் விபரம்

எந்த ஒரு புரோகிராமிற்கும் அதன் இயக்கத்தை அல்லது அதன் வரைமுறையை பற்றி விவரித்தல் அவசியமாகின்றது. அதை எழுதியவர்க்கும் உபயோகிப்பவர்க்கும் பிற்காலத்தில் புரோகிராமின் செயற்பாட்டினை விளக்க இந்த் விவரங்கள் உதவி புரிகின்றன. இந்த காரணங்களுக்காக 'விவரம்' என்ற ஒரு கட்டளை பயன் படுத்தப்படுகிறது. இந்த கட்டளை, "இயக்கம் சாரா" வாக்கியமாகும்.

(உம்)

விவரம் இது ஒரு எடுத்துக்காட்டு

50 விவரம் நாள் 20.9.1966

'இடு' கட்டளை

'இடு' அல்லது 'கொடு' அல்லது 'படி' போன்ற கட்டளைகள் புரோகிராமிற்கு தேவையான இன்புட் மாறிகளை அளிக்க உதவுகின்றன.

(உம்)

20 இடு க, ம, பரப்பு

90 இடு ஆரம், விட்டம்

இந்த மாறிகளுக்கு தேவையான மதிப்புகள் புரோகிராம் இயக்கப்படும் பொழுது உபயோகிப்பாளரால் கொடுக்கப்பட வேண்டும்.

“விடையளி” கட்டளை

கணிப்பொறியில் இடப்பட்ட புரோகிராம்கள் இயக்கப்படும்பொழுது அவற்றின் விடைகள் அல்லது தீர்வுகள் வெளியுலகிற்க அளிக்கப்பட்டாக வேண்டும். அதற்காக விடையளி என்ற கட்டளை கணித்தமிழ் மொழியில் இணைக்கப்பட்டுள்ளது.

(உம்) 100 விடையளி பெயர், தீர்வு, பரப்பு

110 விடையளி தனிவட்டி

120 விடையளி “வணக்கம்”

கட்டுப்பாட்டு வடிவங்கள்

“என்றால்” அல்லது “எனில்”

கணித்தமிழ், வாக்கியங்களின் தொடர்ச்சியை சில காரணங்களுக்காக மாற்ற, “என்றால்” அல்லது “எனில்” என்ற வாக்கியங்களை அறிமுகப்படுத்துகின்றது. அதை ‘முடிவு சார்ந்த பிரிவு’ அல்லது ‘கட்டாய கிளை’ என்ற இரு வகைகளாக பிரிக்கலாம். எனில் ‘அல்லது’ என்றால் வாக்கியங்கள் கீழ்க்கண்ட இலக்கணத்தில் அமைந்திருக்கும். இது முதல் வகையாகச் சார்ந்தது.

(உம்) 1) 20 மதிப்பு 100 என்றால்செல்க 430

30 விடையளி மதிப்பு

2) வருடம் என்றால் வட்டி முதல் வட்டி

செல்க

புரோகிராம்மின் செயற்பாடு, வாக்கியங்களின் வரிசையைச் சார்ந்தது. அதனை மாற்றவோ அல்லது புரோகிராம் கட்டுப்பாட்டை வேறு ஒரு வாக்கியத்திற்கு திசை திருப்பவோ ஏற்ற வகையில் ‘கட்டாயக்கிளை’ வடிவங்கள் உதவுகின்றன. இந்தக் கருத்தை ‘செல்க’ என்ற கட்டளை மூலம் நிறைவேற்றலாம். இது இரண்டாம் வகையைச் சார்ந்தது.

(உம்) 100 செல்க 430

இந்த வாக்கியத்தின் மூலம், கட்டுப்பாடு வாக்கிய எண் 100லிருந்து வாக்கிய எண் 430க்கு சென்றுவிடுகிறது.

2.3 'தொடர் இயக்கம்'

ஒரு குறிப்பிட்ட புரோகிராம் பகுதியை அல்லது வாக்கியத் தொகுப்பை பலமுறை செயற்படுத்தும் அவசியத்தை 'தொடர்' இயக்கம் என்ற தத்துவம் நிறைவேற்றுகிறது. அந்த தொடர் இயக்க அமைப்பின் முதல் வாக்கியமாக 'தொடர்' என்ற வினைச்சொல் பயன்படுத்தப்படுகின்றது.

இறுதி வாக்கியமாக 'மறுபடியும்' என்ற சொல் அமைகிறது.

(உம்)	510	தொடர்	.	மா 5.50 அதிகரி 2
			.	(உடற்பகுதி)
	600	மறுபடியும்	.	மா

மேற்கண்ட எடுத்துக்காட்டில், 'மா' என்ற மாறி 5லிருந்து 50 வரை தனது மதிப்பை இரண்டிரண்டாக ஏற்றிக் கொள்கிறது. அந்த ஒவ்வொரு முறையும் தொடர் இயக்கத்தின் உடற்பகுதி செயற்படுத்தபடுகிறது. 'மறுபடியும்' வாக்கியம் 'மா' என்ற மாறியின் மதிப்பை மாற்றியும், அடுத்த முறை உடற்பகுதியை இயக்கவும், கட்டுப்பாட்டை மீண்டும் வாக்கிய எண் 510க்கு திருப்பவும் வழி செய்கின்றது.

2.4 'முடிவு' வாக்கியம்

கணித்தமிழில் எழுதப்பட்ட புரோக்ராம்களின் கடைசி வாக்கியம் 'முடிவு' என்ற வாக்கியமாக அமைகிறது. இது புரோகிராமின் கட்டுப்பாட்டை நிறுத்தி விடுகிறது. இதற்கு மேல், வாக்கியங்கள் தவறுதலாக எழுதப்பட்டாலும் அவை செயற்படுத்த மாட்டா.

(உம்)	230	விடையளி வருவாய், லாபம்
	240	விவரம் கடைசி வாக்கியம்
	250	முடிவு

2.5 கணித்தமிழ் மொழிமாற்றி

கணித்தமிழ் மொழிமாற்றி காண்டக்ஸ்ட் - ஃப்ரி இலக்கணத்தை சார்ந்து அமைக்கப்பட்டது. மற்ற கணிப்பு மொழிகளின் இலக்கணம் போலவே, கணித்தமிழின் புரோகிராம்களும் மேலிருந்து கீழாக உள்ளமைக்கப்பட்ட அகராதியின் இலக்கண விதிகளின் படியும், அர்த்தி விதிகளின்படியும் ஆய்வு செய்யப்பட்டு அட்டவனைகளாக மாற்றப்படுகின்றன. அவைகள் மீண்டும் இடைப்பட்ட குறியீடுகளாக (கணிப்பொறிக்கு நெருக்கமான குறியீடுகளாக) உருமாற்றப்பட்டு செயற்பாட்டு தொகுப்பாக மாற்றப்படுகின்றன.

3. முடிவுரை

'கணித்தமிழ்' மூலம் ஒரு வலுவான அல்லது முறையான கணிப்பொறி மொழியை ஆக்கி விட இயலாது. இதை மேலும் செம்மைப்படுத்தும் முயற்சியில் ஆய்வு நடந்து கொண்டிருக்கிறது. இது போன்று பலரும் முயற்சி செய்கிறார்கள். அவர்கள் அனைவரும் ஒரு குடையினுள் அமர்ந்து, தமிழில் நல்லதொரு கணிப்பொறியை உருவாக்க வேண்டும். குறிப்பாக சில இந்த மொழிக்கு ஊட்டம் அளிக்கும் வகையில் மேலும் சில கணிப்புத் தத்துவங்களை இணைக்க வேண்டும். உதாரணமாக, அகராதிச் செயற்பாடுகள், துணை கணிப்பு தொகுப்புகள், எண் வரிசை அல்லது எழுத்து வரிசை போன்ற தகவல் அமைப்புகள் ஆகியவற்றை இணைத்தல் வேண்டும்.

இணைப்பு

முன் பதிவான வார்த்தைகள்

அதிகரி

செல்க

என்றால்

எனில்

இடு

கொடு

மறுபடியும்

முடிவு

தொடர்

விடையளி

விவரம்

முத்துக் கணினி - 34

- நெல்லை சு.முத்து

1.0 முன்னுரை

அறிவியல் தமிழ் வெறும் பேச்சுத் தமிழாக அல்லது அச்சுத் தமிழாக மட்டுமின்றி செயல்படு தமிழாக அறிவியலுக்கும் பயன்பட வேண்டுமாயின் எழுத்துச் சீர்திருத்தம் தேவை.

பேச்சொலி குறியீடுயியல் (Phonetics), பேச்சு மொழி ஒலியியல் (Phonemics) ஆகிய இவ்விரண்டின் சரிவிகிதக் கலப்பே ஒலிநுட்பவியல் (Phonology) எனலாம்.

ஒலியியலில் மூன்று துறைகள் உள்ளன. ஒலிப்புமுறை ஒலியியல் (Articulatory Phonetics), இயல்புமுறை ஒலியியல் (Acoustic Phonetics), கேட்புமுறை ஒலியியல் (Auditory Phonetics).

ஒலி உறுப்புகளிலிருந்து ஒலிகள் பிறப்பதை அடிப்படையாகக் கொண்டு உயிர், மெய் ஒலி இயல்புகளை இனங்கண்டு அவற்றைப் புரிந்து கொள்ளும் திறனைக் கணிப்பொறிக்குள் ஊட்ட வேண்டுமாயின் ஒலிஅலைகளை அலசி ஆராய்தல் வேண்டும். இதன் வழி செயற்கைப் பேச்சை உருவாக்கவும் முடியும் என்பது அறிவோம். இன்றைக்கு இத்தகைய கணினி வித்தைகள் ஓரளவு நிறைவேறி வருகின்றன.

கணினி தமிழ் கேட்கும்; பேசும்; அச்செழுதும்; கேட்பதும், பேசுவதும் ஒலியியல் சார்ந்தவை. அச்செழுத்து என்பதோ வரிவடிவம் சார்ந்தது. அனைத்துமே மின்னணு விவகாரத்தினுள் அடைபடும்.

இன்று - பத்திரிகைகள், நூல்கள், துண்டு பிரசுரங்கள், விளம்பர சுவரொட்டிகள், திருமண அழைப்பிதழ், சிற்றுண்டி, பலசரக்குக்கடை விலைச் சீட்டுகள், அலுவலகச் சுற்றறிக்கைகள், கடிதங்கள், நாட்காட்டிகள், எல்லாமே கணிப்பொறி இயந்திர ஊழியனால் வளைத்துச் சுழித்து எழுதப்படலாம். இது மட்டும் போதுமா?

அறிவியல் தமிழ் எழுதவும், படிக்கவும், கேட்கவும், பேசவும் உரிய வழக்குத் தமிழாக மட்டுமின்றி கணிப்பொறியினை இயக்கும் தமிழாகப் பயனளிக்க வேண்டும் அல்லவா?

அன்றைக்கு குகைக்குள், கல்லில் தமிழ்; இன்றைக்கோ கணிப் பொறிக்குள் சிலிக்கான் சில்லில் தமிழ்; இரண்டுமே கல், மணல், சமாச்சாரங்கள்.

நிலையான கல்வெட்டுக்களில் கொத்தி எழுதப் பெற்ற தமிழுக்கும், கையிலடங்கும் ஓலைச் சுவடிகளில் எழுத்தாணியினால் குத்தி எழுதப்பெற்ற தமிழுக்கும், இன்று காகிதத்தில் தீட்டப் பெறும் தொடர்ச்சியான வரிவடிவத்திற்கும் இடையே எத்தனையோ மாற்றங்கள், இருந்தும் கன்னித்தமிழ் நிலைத்து வாழ்கிறது.

கல்வெட்டு உளிக்கொத்தல் மற்றும் ஓலை மடல் ஆணிக்குத்தல் யாவும் இன்றைக்கோ கணிப்பொறியில் படக்கூறுகளாகத் தோன்றியும், அச்சுத்தலை (Printer head) மின்வெப்ப புள்ளிகளினால் அல்லது கரிநாடாப் பதிவுகளினால் எழுதப் பெறும் புள்ளி மாத்திரைக்கு பதிலாக "புள்ளி அணி (dot matrix)" தொழில் நுட்பமாகவும் உருவெடுத்துள்ளது.

காலத்திற்கேற்பவும், தொழில் நுட்ப வளர்ச்சிக்கு உகப்பாகவும் தமிழின் தனித்தனி எழுத்துக்கள் தம் வரிவடிவில் மாற்றங்கள் நிகழ்ந்து வந்துள்ளதை எவரும் மறுதலிக்க இயலாது. அவ்வாறே எழுத்துக்களின் எண்ணிக்கையும், துணைக்குறியீடுகளின் எண்ணிக்கையும் அச்சு வசதிக்கேற்ப கூடியும், குறைந்தும் வந்துள்ளன.

புள்ளி மெய்யெழுத்துக்கள், கொம்புகள், துணையெழுத்துக்கள் அனைத்துமே மறுபரிசீலனைக்குள்ளாயின. தமிழில் அத்தனைப் பரிசோதனைகளின் அக்கினிக் குண்டங்களிலும் நெருப்புப் குளித்து நிமிர்ந்துள்ள இனிப்புத்தமிழ், கனித்தமிழ் அடுத்த நூற்றாண்டில் கணிப்பொறித் தமிழாக பரிணமித்து ஓங்குவது திண்ணம்.

2.0 ஒலி வடிவம்

2.1 மெய்யொலிப்புகள்

தமிழ் எழுதத் தெரியாதவர் கூட கணிப்பொறியுடன் உரையாட முடியும். உலகின் மற்ற எந்த மொழியினை விடவும் தமிழில் மட்டுமே ஒலியன்கள் முறையாக வகைப்படுத்தப் பெற்றுள்ளன. குரல்மொழி (tone language) யாகிய சீனத்தில் ஒரே ஒலியன் உச்சரிப்புச் சுரத்திற்கு (Pitch) ஏற்ப வெவ்வேறு பொருள் தரும் என்பர். ஆங்கிலமோ ஒரே ஒலிக்கு வெவ்வெறு எழுத்துக்கள் வரைந்துக் காட்டும் தூய ஒலியன் மொழியாகிய (Phonemic language) தமிழில் இத்தகைய சிக்கல்கள் அதிகம் இல்லை. ஒருசில ஆப்பிரிக்க மொழிகளினைப் போல மூச்சு உள்ளிழுக்கும் (ingressive) போது தமிழ் பிறப்பதில்லை. நம் தாய் மொழியில் இயல்பாகவே மூச்சு வெளியிடும் போது (egressive) பேச்ச பிறக்கிறது.

வல்லினம், மெல்லினம், இடையினம் ஆகியவை முறையே மூச்சுக்காற்றுத் தொழில் (air-stream process) நிகழ்வினால் மார்பு, கழுத்து, மூக்கிலிருந்து பிறப்பன. (வங்காள மொழி உச்சரிப்புகள் பெரும்பாலும் மார்பிலிருந்தும், மலையாளம் அனேகமாக மூக்கிலிருந்தும் பிறப்பதை உணர்வோம்)

ஆயின் தமிழ் மெய்யினங்களின் ஒலிப்புத் தொழில் (articulatory process) ஏறத்தாழ ஒரே மாதிரி நிகழ்கின்றது எனலாம். மூவின மெய்களின் உச்சரிப்பின் போதும் மேலண்ணம், நாக்கு, இதழ் அசைவுகளாலான ஒலிப்பு முறைகள் (Manners of articulation) ஒரே மாதிரி அமைவதை இங்கு சுட்டிக் காட்டலாம்.

"க - ச - ட - ற- த- ப" என நிதானமாக உச்சரித்தால் இவ்வுண்மை புரியும். 'க' என்னும் போது வாய் திறந்திருக்க, நாக்கு நுனி உள்நோக்கி தொட்டு உள்வாங்கிய நிலையிலிருந்து மேல் அண்ணத்தில் ஒட்டி படிப்படியாக முன்னேறி மேல்வாய்ப்பல்லின் பின்புறம் தட்டி 'ப' எனும் போது இதழ்மூடித் திறப்பதை அறிகிறோம்.

"ங - ஞ - ண - ன - ந - ம" என்றவாறு மெல்லினத்தை உச்சரித்தால் முன் போலவே வாய் திறந்திருக்க, நாக்கு நுனி மேல் அண்ணம் வழி நகர்ந்து பல்லின் பின் வாக்கில் தட்டி 'ம' எனும் போது இதழ் மூடித் திறக்கிறதல்லவா?

"ழ - ய - ள - ல - ர - வ" என்கிறபடி இடையின வரிசையினை உச்சரிக்கும் போது நாக்கின் சலனம் வல்லின, மெல்லின நிரல் ஒலியன்களை ஒத்தசைவதை உணர்கிறோம். 'வ' எனும் போது மட்டும் இதழ்குவித்து கீழுதட்டை உள்ளுக்குள் அடக்கி மேல் வாய்ப்பற்களால் நெறித்துத் திறக்கிறோம்.

இந்த மூவின ஒலிப்பிடங்கள் (Places of articulation)-கடையண்ணம் (Velvar)/ இடையண்ணம் (palatal), வளைநா (retroflex), நுனியண்ணம் (alveolar), பல் (dental) மற்றும் ஈரிதழ் (bilabial) எனபன அறிகிறோம். ஆயின் வல்லின மெய்கள் - அடைப்பொலிகள் (Stops) ஆகும். இவை காற்று வாய்வழி செல்லும் போது மேல் குறித்த ஆறு ஒலிப்பிடங்களில் ஏதாவது ஒரு ஒலிப்பானால் (articulator) சற்றுத் தடைப்பட்டு பிறப்பன. அவற்றுள்ளும் 'ற' கரமானது நுனியண்ண ஒரு பகுதியைத் தொட்டும் தொடாமலும் நிற்கும் நுனிநா - மூச்சுக்காற்றின் இயக்கத்தினால் ஆடும் போது ஆடொலி (trill) யாகப் பிறக்கிறது.

மூச்சுக்காற்று முழுவதும் மூக்கின் வழி செல்கையில் எழும் அடைப்பொலிகள் தமிழில் மெய்யின ஒலிகள் அனைத்தும் மூக்கொலிகளாக (nasal sounds) த் தோன்றும். ஆயின் அவற்றின் ஒலிப்பிடங்கள் வல்லின ஆறு ஒலிகளின் அதே ஆறு பிறப்பிடங்களையே நெருங்கியுள்ளன. ஆதலின் ஆறு அடைப்பொலிகளை வல்லினம், மெல்லினமாக்க கணினிக்கு இரண்டு ஆணைகள் போதும்.

இடையின ஒலிகளோ குரல்வளையில் நிகழும் ஒலிப்புடைமைத் தொழிலினால் (Phonation Process) பிறக்கின்றன. எனினும் 'ய' எனும் ஒலி கழுத்திலிருந்து மருங்கொலி (lateral) யாகத் தோன்றுகிறது. அதாவது மூச்சுக்காற்று தடுக்கப்பட்டு இடையண்ண ஒலிப்பானின் ஒருப்பக்கமாக (மருங்காக) ஒலி வெளிப்படுகிறது. 'ள'கரமோ வளைநா ஒலிப்பானின் இருமருங்காகவெளிப்படுகிறது.'ல' கரம் நுனியண்ண நுனிநா மருங்கொலியாக பிறக்கும் போது - நுனிநா தட்டியத் தட்டில் அடியொலி (tap) யாக ரகரம் பிறக்கிறது. பல்லிதழ் உரசொலியாக (fricative) யாக வருவது வகரமும் ஃபகரமும். (முன்னதில் காற்று கழுத்திலிருந்தும், பின்னதில் மார்பிலிருந்தும் வெளிப்படுகின்றன) நாவளை தொடர் மருங்கொலியாக வெளிப்படும் சிறப்பு 'ழ' கரம் கடையண்ணத்தையே ஒலிப்பிடமாக கொண்டுள்ளது.

மெய்யொலிகள் தவிர வடமொழி எழுத்துக்களான ஷ, ச, ஹ, ஜ ஆகியனவற்றை கையாளதலும், அவற்றினை ஒலிப்பிடம் (place of articulation) மற்றும் ஒலிப்பு முறை (Process of articulation) ஆகியவற்றுக்கேற்ப வகைப்படுத்தலும் கணிப்பொறித் தமிழின் தேவை.

நுனியண்ண ஒலிப்பில்லா உரசொலி (voiceless fricative) - ஃச(ss)

நுனியண்ண உரசொலி - ஸ (s,z)

இவைஇரண்டும் தட்டை யுரசொலிகள் (Surface fricative)

வளைநா ஒலிப்பில்லா உரசொலி - ஷ (sh) அதுவும்

குழிந்துரசொலி (groove fricative).

ஒலிநாண் (vocal chord) ஒலிப்பில்லா உரசொலி - ஹ (h)

இவ்விதம் தமிழின் மெய்யெழுத்துக்களை ஒலிப்பிடம் மற்றும் ஒலிப்பு முறைகளுக்கேற்ப நிரைப்படுத்துதல் கணினித் தமிழுக்குரிய ஒலிவடிவக்குரல் உருவாக்கத்திற்கு (Voice Synthesiser) உதவும்.

அட்டவணை (1)ல் குறிப்பிட்டுள்ளபடி இப்புதிய மெய் நிரலினை உற்று நொக்கிட, க் - ங், ச்-ஞ், த் - ந் ஆகிய இணைகள் மற்றும் ட்-ண்-ள், ற்-ன்-ல் ஆகிய இனக் கோர்வைகள் பற்றிய தொடர்பும் எளிதில் விளங்கும். வல்லின அடைப்பொலிகள் போன்றே மெல்லின மூக்கொலிகளும் அடைப்பொலிகளே. ஆயின், மூச்சுக்காற்றுப் பிறப்பிடம் மட்டுமே வேறுபடும். வல்லின அடைப்பொலிகளில் மார்பு மூச்சுக் காற்று கனத்து வெளிப்பட்டால் கனவொலிகள் பிறக்கம். இத்தகைய ஒலிவகைப்பாடுகளை கணிப்பொறி எளிதில் புரிந்து கொள்ளும்.

சான்றாக, நிலைமொழியீற்றில் ளகரமும், வருமொழி முதலில் வல்லெழுத்து (வல்லின ஒலியன்) வர 'ள' கரம் 'ட' கரமாகவும், மெல்லெழுத்து (மெல்லின ஒலியன்) வர 'ள' கரம் 'ண' கரமாகவும், மாறுவதற்கு தமிழில் இலக்கண விதிகள் உள்ளன.

கள்குடம்	..	கட்குடம்
உள்நாக்கு	..	உண்ணாக்கு

இச்சொற்களில் கள்குடமும், கட்குடமும் மூச்சுக் காற்று ஒலிப்பிறப்பினால் வேறுபட்டு ளகர, டகரங்களை இனங்காட்டுமே அன்றி இவை இரண்டுமே ஒலிப்புமுறையினால் ஓரினத்தவை. இரண்டின் உச்சரிப்பும் கணினிக்கு புரியும்.

2.2 கிரந்த ஒலியன்கள்

அறிவியல் தமிழில் இன்று ஆங்கில மொழியிலுள்ள கன ஒலியன்கள் பலவும் தேவை. ஆயின் தமிழில் பயன்படுத்தப் பெறாமல் துருப்பிடித்துக் கிடக்கும் 'ஆயுத' எழுத்தினை மினுக்கி எடுத்துக் கையாளலாம். இந்த ஆய்த எழுத்தின் உபயத்தினை 'தமிழுக்குக் கூடுதல் ஒலியன்கள் தேவையா?' (தின மணிசுடர், 10.6.1989) என்கிற எதிரொலியில் ஏற்கெனவே நாம் குறிப்பிட்டிருந்தோம்.

க (K)	-	ஃக (G)
ச (Ch)	-	ஃச (ss)
ட (T)	-	ஃட (D)
த (Th)	-	ஃத (Dh)
ப (P)	-	ஃப (B)

இங்கு வழக்கத்திலுள்ள 'ஃப' (F) ஒலிப்பு முறை மாற்றப்பட்டாக வேண்டும்; (காண்க: அட்டவணை - 2) ஆயின் 'ஃப' என்பது ஏற்கனவே ஃபிரான்சு (France) முதலான சொல்லாட்சிகளில் சிறப்பாகக் கையாளப்பட்டு நிலைத்தும் விட்டது என வாதிடலாம். ஆதலினால் 'F' பல்லிதழ் உரசொலியினைத் தனிக்குறியீட்டினால் இனி சுட்ட இயலாது. தனித்த ஒலியன் உருவாக்கவும் தேவையில்லை. 'ஃப' என்றெழுதினால் போதும். ஆயின் தேசம், சிவன் போன்ற சொற்களில் வரும் சகரத்திற்கு 'உ' கரத்திற்கும் இடையிலான விசேட நுனியண்ண ஒலிப்பில்லா உரசொலியினை (Voiceless fricative) 'ஃச' கரத்தினால் பெறலாம். மலையாளத்தில் இந்த ஒலிக்கு தமிழின் 'ல' எழுத்தில் கண்ணாடி இடவல 'ശ' பிம்பமான என்கிற மாதிரி வடிவமும் இந்திமொழியில் '51' எண் மாதிரியான 'श' வடிவமும் வழக்கிலுண்டு.

தவிர ஜ, ஷ, ஸ, ஹா முதலான ஃகிரந்த ஒலிகளையும் தமிழில் அப்படியே எடுத்தாளலாம்.

கணிப்பொறி வாசிப்பதற்கு எளிமையான வல்லின மாற்றோலி வரி வடிவங்களின் அட்டவணை - 2ல் இருந்து கீழ்காணும் சில விதிமுறைகளை வகைப்படுத்தலாம்.

விதி: 1

வல்லின உயிர் மெய்கள் மொழி ஈற்றில் வரும் போது க (K) - ஹ (H) உரசொலியும், ச(Ch) - ஃச(ss) உரசொலியும், ட((T) என்பது ஃட(D) மாற்றொலி (allophone) யும், த(Th) - ஃத(Dh) மாற்றொலியும், ப(P) என்பது ஃப(B) மாற்றொலியும் பெறுகின்றன.

விதி : 2.1

புணர்ச்சிச் சொல்லாக அன்றி தனிச்சொல்லாக வரும் மொழி நடுவிலும் இந்த வல்லின உயிர் மெய்களில் மேற்குறித்த விதிமுறையே தாக்கம் பெறுகிறது.

விதி: 2.2

மொழி நடுவில் வல்லொற்றினைத் தொடர்ந்து வரும் வல்லின மெய் ஒலிகள் இயல்பான ஒலியிலேயே வெளிப்படுகின்றன.

விதி: 2.3

மொழி நடுவில் இடையொற்றினைச் சார்ந்து வரும் வல்லின மெய்கள் விதி 1.ல் குறிப்பிட்ட ஒலிகளைப் பிறப்பிக்கின்றன.

விதி: 2.4

மொழி நடுவில் மெல்லொற்றினைச் சார்ந்து ஒலிக்கும் வல்லினங்கள் வேறு தனி ஒலியன்கள் ஆகின்றன.

க - 'ஃக' மாற்றொலியும், ச - 'ஜ' மாற்றொலியும், 'ட' - ஃட மாற்றொலியும், த - ஃத மாற்றொலியும், ப(P) என்பது ஃப (B) மாற்றொலியும் வருகின்றன.

விதி: 3

இதனால் மொழி முதலில் வரும் வல்லினங்கள் தத்தம் ஏனைய புதுமாற்றொலிகளை அறிவியல் தமிழில் வெளிப்படுத்தலாம்.

க - ஃக (G), ஹ(H) ஒலிகளும்

ச- ஃச (ss), ஸ, ஷ, ஜ ஆகிய ஒலிகளும்

ட-ஃட (D)

த-ஃத (Dh)

ப-ஃப (B)

ஒலிகளும் அறிவியலில் தேவையாகிறது. கணிப்பொறியினுள் மேற்குறித்த விதிகளின் அடிப்படையினை ஆணைகளாகத் தொகுத்தளித்தால் நம் தமிழை அது எளிதாக வாசிக்கவும் ஓ.சி.ஆர். (OCR) எனப்படும் ஒளிவரிவினப் புரிதல் (Optical Character Recognition) முறைக்கும் இந்த அடிப்படைகள் போதும்.

3. வரி வடிவம்

3.1 உயிரெழுத்துகள்

தமிழ் நெடுங்கணக்கில் ஐ. ஔ ஆகிய இரண்டு எழுத்துக்களுமே அதிகப் பற்று என்றொரு எண்ணம் மொழியியல் அறிஞரிடை நிலவுகிறது.

“ஐ, ஔ - இரண்டையும் ஒழிப்பது தேவை தானா அல்லது தேவய் தானா?” என்று பேராசிரியர் தமிழண்ணல் போன்றோர் பெரும் விவாதத்தையே தொடங்கி வைத்தனர். (தினமணி சுடர் 18-7-1992)

ஐ, ஔ - ஒழிப்பது தேவை தான் எனக் கருத்து வெளியிட்ட பொறிஞர் கொடுமுடி க.சண்முகன், “வீரநடய் அறிவியக்கம் என்ற பெயரில் முனனவர் சாலய் இலாந்திரயன் பல்லாண்டாய் ஒரு திங்களிதழ் நடத்தி வருகிறார். முழுக்க முழுக்க ஐ, ஔ வடிவங்களை நீக்கியே அச்சிடுகிறார்” என்றும் அறிவித்தார். (தினமணி சுடர் 12-9-1992)

அகரம், இகரம் ஐகாரம் ஆகும் (அ+இ = ஐ) அகரம் உகரம் ஔகாரம் ஆகும்.(அ+உ = ஔ)

இவை மாதிரியே, மேலும் பல தனி நெடில்கள் வருவாகியிருக்கலாமோ என்றெண்ணத் தோன்றும்.

உ + இ - உய்

உ + உ - உவ்

எ + இ - எண்

எ	+	உ	-	எவ்
ஒ	+	இ	-	ஒய்
ஒ	+	உ	-	ஒவ்
இ	+	இ	-	இய்
இ	+	உ	-	இவ்

இவ்வித ஒலிகளுக்கு மட்டும் தனி நெடில்கள் தேவையில்லையா? தமிழ் எழுத்து முறையில் இதற்கு இடமில்லை. எனவே அய், அவ் என் ஒலிக்கும் ஐ, ஔ ஆகிய தனி நெடில்களை உயிரெழுத்து நிரலிலிருந்து அகற்றிவிடலாம்.

பழமலை என்பதனை ' பழமலய்' என்பது போலவே புதையல் என்பதனை 'புதயல்' என்று வழங்குவதும் மாத்திரை அளவு அடிப்படையில் சமம்; (மலையாள மொழியிலும் கல (கலை) மல (மலை) பற (பறை) சிற (சிறை) என்பதே வழக்கு. இவற்றுடன் வேற்றுமை உருபுகள் புணரும் போது கலய், மலய், பறய், சிறய் என்கிறபடி யகர ஒற்று மிகும்)

ஆங்கில உயிரொலிகள் (vowels) எனப்படும் a,e,i,o,u போல தமிழிலும் அ, எ, இ, ஒ, உ ஆகிய ஐந்து உயிர்களே போதும். இவற்றை நெடிலாக்கி எழுத கால் (ா) எழுத்தின் துணையினால் முடியும். கணிப் பொறிக்கும் இல்விலக்கணம் எளிதிற் புரியும்.

அா	-	ஆ
இா	-	ஈ
உா	-	ஊ
எா	-	ஏ
ஒா	-	ஓ

3.2 உயிர்க் குறியீடுகள்

உயிர்க்குறிலை உயிர் நெடிலாக்க உதவும் கால் (ா) துணையெழுத்தே மெய்க்குறில்களை நெடில்களாக்கப் போதுமானது.

"தமிழ் எழுத்துச் சீரமைப்பு"எனும் நூலில் இது தவிர ஏனைய ஏழு உயிர்க் குறியீடுகளைப் பரிந்துரைக்கிறார் அறிவியலர், முனைவர் வா.செ.குழந்தைசாமி.

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ
–	ா	ി	ീ	ു	ൂ	ெ	ே	ை

இவற்றிலுள்ள உகர, ஊகாரக் குறியீடுகள் இன்று மலையாளப் புது லிபியில் அங்கீகரிக்கப் பெற்றவை என அறிவியல் அறிஞர் ஐராவதம் மகாதேவன் (தினமணி சடர் 19-4-1989) எடுத்துரைக்கிறார், இஃது உண்மையே.

தவிர, இந்த எட்டு உயிர்க்குறிகள் அறிஞர் தெ.பொ.மீனாட்சி சுந்தரனார் பல ஆண்டுகளுக்கு முன்னர் வழங்கிய பரிந்துரையே (The Tamil Script Reforms. பக்கம் 159)என்று கொடுமுடி சண்முகப் பிரகதம் தமது "தேவைக்கேற்ற தமிழ்" எனும் மொழியியல் ஆய்வு நூலில் காட்டும் மேற்கோளினால் அறிய முடிகிறது.

கும்பகோணம் கோ. ஜெயபாலன் என்கிற மொழியியல் "மக்கள் தமிழ்" அறிஞர் கி. ராஜாநாராயணன் வழி எமக்கு அறிவித்த கருத்துரையில் ஒகர, ஓகாரத்திற்கும் இரண்டு புதிய குறியீடுகளையும் கூட்டிச் சேர்க்கின்றார். தவிர, இதர ஈகாரத்திற்கு மலையாளக் குறியீடுகளையும் எடுத்தாள விழைகிறார்.

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஒ	ஓ
–	ா	ി	ീ	ു	ൂ	ெ	ே	)	[illegible]

தமிழ்நாடு தமிழாசிரியர் பேரவைத் தலைவரும், தமிழ் எழுத்துச் சீர்திருத்த இயக்கச் செயலாளருமான புலவர் வே.புகழேந்தி கூட உயிர்க்குறியீடுகள் 11 என்று பெரும்பாலும் மலையாள நிரலைப் பின்பற்றுகிறார். ஒகர, ஓகாரம் விதிவிலக்குகள்.

அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
–	ா	ി	ീ	ു	ൂ	െ	േ	ൈ	ம	[illegible]	[illegible]

இனி கணித்தமிழுக்கு வருவோம்.

தமிழின் உயிர்மெய்கள் அனைத்தும் ஒரே விதியின் கீழ் வடிவமைக்கப் பெற்றவை அல்ல.

ஆகார நெடில்கள் (கா, சா, டா போன்றவை), இகர மெய்கள் (கி, சி,டி....), ஈகார நெடில்கள் (கீ, சீ, டீ....), எகர மெய்கள் (கெ,செ,டெ), ஏகாரம் (கே,சே,டே....), ஒகரம் (கொ, சொ, டொ..), ஓகாரம் (கோ, சோ, டோ...), ஔகாரம் நெடில்கள் (கௌ, சௌ, டௌ..) ஆகியன மட்டும், முறையே ா, ி, ீ, ெ, ே, ொ, ோ, ௌ - ஆகிய துணையெழுத்துக்களுடன் ஒரே மாதிரி எழுதப் பெறுகின்றன. இவற்றுள்ளும் 5 குறியீடுகள் பொது.

மற்றபடி, ஐகார நெடில்கள் பலவிதமாக எழுதப் பெற்றுவந்ததால் அவை 'ஸ' என்ற துணை எழுத்தால் சீராக்கப்பெற்றன. எதிர்ப்போர் எதிர்க்கட்டும்; இனி ஐகாரக் கொம்பு வேண்டாம் என அச்சு ஊடகம் ஏகமனதாக இச் சீரமைப்பினை ஏற்றுக் கொண்டும் விட்டது.

லை - லை - லய்

னை - னை - னய்

ணை - ணை - ணய்

ளை - ளை - ளய்

அவ்வாறே ஔகார நெடில்களும் தேவையில்லை

ஔவையார் - அவ்வையார்

ஐயப்பன் - அய்யப்பன்

மிஞ்சியிருப்பது உகரம், ஊகாரம் ஏறிய மெய்கள் ஒவ்வொன்றும் ஒவ்வொரு விதமாக வரிவமைக்கப்படுகின்றன. உகர மெய்களைக் கவனிப்போம்.

க + உ - கு

ச + உ - சு

ட + உ - டு

த + உ - து

ப + உ - பு

ங + உ - ஙு

ம + உ - மு

ர + உ - ரு

இவ்வாறே ஊகார மெய்களும், கூ, சூ, டு, தூ, பூ, ஙூ போன்ற பல வரி வடிவங்களின் துணையால் வெளிப்படுகின்றன.

கணிப்பொறியில் இத்தகைய இரண்டு எழுத்துக் கூட்டல்களை மூன்றாவது எழுத்தாக வரிவடிவில் எழுதச் செய்யலாம். சான்றாக 'உ' என்ற எழுத்தினை நிலைத்த எழுத்தாக (dead key)த் தட்டி, அதே இடத்தில் 'க' எழுத்தை அழுத்தி 'கு' பெறலாம். இப்படியே வெவ்வேறு எழுத்துக்களுக்கு உகர, ஊகார மெய்யெழுத்துக்களைக் கணினிக்குள் நினைவுக் கிடங்கில் நிரப்பி வைக்க வேண்டிவரும், கணினி ஆற்றல் விரயம்.

கணிப்பொறிக்கும் கொஞ்சம் மரபுக் கவிதைக்கும் தொடர்பிருக்கிறது. இ, ஈ, எ, ஏ, - ஆகிய ஒரே ஒலிச்சாயல் கொண்ட 'இகர' மோனைகள் மொழியியலார் இவற்றை இதழ்விரி உயிர்கள் (unrounded vowels) என்பர். அவ்வாறே உ, ஊ, ஒ, ஓ ஆகிய 'உகர' மோனைகளை இதழ் குவி உயிர்கள் (rounded vowels) என்பர்.

இனி, இவற்றின் மெய்யெழுத்து வரிவடிவங்களைச் சற்று கவனிப்போம். எகர, ஊகாரங்களுடன் துணையெழுத்து (ா) சேர்த்து ஒகரம், ஓகாரம் பிறக்கின்றன. விரித்த இதழ்களுக்கு 'கால்' எழுத்துச் சேர்த்து இதழ்குவியச் செய்கிறோம். ('கெ' 'கொ' என்றாகிறது)

கெ - கொ

கே - கோ

இம்மாதிரி இதழ்வரிக்கும் இகர, ஈகாரங்களுடன் அதே துணையெழுத்துச் சேர்த்து உகர, ஊகாரங்களைப் பெறலாமே!

கி - கிா (கு)

கீ - கீா (கூ)

இதில் இன்னொரு விசேஷம் பாருங்கள்.

கணிப்பொறி தன் எழுத்தெண்பலகை (Alpha numeric board) யாகிய விசைப் பலகையில் ெ, ே, ி, ீ ஆகிய உயிர்க்குறிகளை முதலில் தட்டிய பிறகு 'க' போன்ற மெய்யைத் தட்டினாலே கெ, கே, கி, கீ ஆகிய 'இகர' மோனைகள் சின்னத் திரையில் தோன்றும். இதே வரிசையில் அவற்றைத் தொடர்ந்து 'ா' என்கிற துணையெழுத்தைத் தட்டினால் கொ, கோ, கிா(கு), கீா (கூ) ஆகிய 'உகர' மோனைகள் தோன்றுமென்பது நடைமுறை லாபம் அல்லவா?

ஆயின் 'இகர' மோனைகளை துணையெழுத்துக் காட்டி 'உகர' மோனைகளாக கணினியில் எழுதும் போது ஒரு சிறுவேறுபாடு - ெ, ே ஆகியன தனி உயிர்க்குறியீடுகள். ி, ீ ஆகியன விசைப் பலகையில் இறந்த உயிர்க்குறிகள் (dead keys) அவ்வளவே.

அட்டவணை : 1. மெய்நிரல் ஒலியன்கள்

ஒலிப்பிறப்பு முறை (நாக்கு, இதழ், அண்ண அசைவு)	ஒலிப்பான்	ஒலிப்பிட ஒலியன்கள்							
		மார்பு (வல்லினம்)				மூக்கு (மெல்லினம்)	கழுத்து இடையினம்		
		அடைப்பொலி	ஆடொலி	உரசொலி	கனவொலி	மூக்கொலி (அடைப்பொலி)	மருங்கொலி	உரசொலி	அடியொலி
நாக்கு கடையண்ணத்தினைத் தொட்டு உள்வாங்கிய நிலை	கடையண்ணம்	க்	-	ஹ் (ஒலிநாண் ஒலிப்பின்னம்)	ஃக்	ங்	ழ்	-	-
நாக்கு படர்ந்து இடையண்ணக் கடைவாய்ப் பற்களை ஒட்டிய நிலை	இடையண்ணம்	ச்	-	-	ஜ்	ஞ்	ய்	-	-
நாக்கு நுனி வளைந்து அண்ணத்தினை உள்மடக்கி அழுத்திய நிலை	வளைநா	ட்	-	ஷ் (ஒலிப்பின்னம்)	ஃட்	ண்	ள்	-	-
நுனியண்ணத்தில் நாக்கு நுனி இயங்கிய நிலை	நுனியண்ணம்	-	ற்	ஃச (ஒலிப்பின்னம்) ஸ (நுனிநா)	-	ன்	ல்	-	-
நாக்கு நுனி மேலும் முன்னேறி முன்பற்களின் பின்புறம் தட்டிய நிலை	பல்	த்	-	-	ஃத்	ந்	-	-	ர்
நாக்கு சலனமற்று, உதடுகள் மட்டும் குவிந்து திறந்த நிலை	ஈரிதழ்	ப்	-	-	ஃப்	ம்	-	வ்	-

அட்டவணை –2 : வல்லின மாற்றொலிகள்

வல்லினம்	மொழியீறு	மொழிநடு (இருசொற்புணர்ச்சியன்றி)				மொழி முதல்
		உயிர்மெய் நிலை	ல்லொற்றுச் சார்நிலை	இடையொற்றுச் சார்நிலை.	மெல்லொற்றுச் சார்நிலை	
க (K)	ஹ(H) உரசொலி (நகு, நகை)	ஹ(H) உரசொலி (நகர், நாகர்)	க(K) அடைப்பொலி (தக்கை, பக்கம்)	ஹ (H) உரசொலி (செல்க, ஆய்க)	ஃக (G) கனவொலி (நான்கு, வாங்கு)	க (K), ஹ (H), ஃக (G)
ச (Ch)	ஃச(SS) உரசொலி (பசை, தோசை)	ஃச(SS) உரசொலி (தேசம், பாசம்)	ச (Ch) அடைப்பொலி (பச்சை, உச்சி)	ஃச (SS) உரசொலி	ஜ (J) கனவொலி (பஞ்சம், மஞ்சள்)	ச (Ch), ஃச (SS), ஜ (J)
ட (T)	ஃட(D) கனவொலி (குடை, சேறு)	ஃட(D) கனவொலி (பாடம், நாடகம்)	ட(T) அடைப்பொலி (சட்டை, முட்டை)	ட (D) கனவொலி (வார்டன், ஆர்டர்)	ஃட (D) கனவொலி (மண்டு, வண்டு)	ட(T) ஃட (D)
ற	ற ஆடொலி (கூறு, சேறு)	ற ஆடொலி (கூறுதல்)	ற ஆடொலி (கூற்று)	ர் – ஒலி	ர – ஒலி	ற, (ர)
த (Th)	ஃத(Dh) கனவொலி (சூது, வாது)	ஃத(Dh) கனவொலி (பாதகம்)	த் (Th) அடைப்பொலி (மெத்தை)	ஃத (Dh) கனவொலி (ஊர்தி, ஆள்தல்)	ஃத (Dh) கனவொலி (மந்தை)	த (Th), ஃத (Dh)
ப (p)	ஃப(B) கனவொலி (உருபு, சொருபி)	ப(B) கனவொலி (கோபம், சாபம்)	ப(P) அடைப்பொலி (குப்பை, அப்பு)	ஃப (B) கனவொலி (சார்பு, சால்பு)	ஃப (B) கனவொலி (அன்பு, பண்பு)	ப (P) ஃப (B)

ஆக, நம் புதிய முறையில் ஐந்து உயிர்க்குறிகளே போதும். அடுத்த நூற்றாண்டின் தமிழ்க் கணினியில் 247 எழுத்துக்களும் அவற்றுக்குரிய அனைத்துச் சிறப்புக் குறிகளும் தேவையில்லை. (அட்டவணை - 3) தற்போது வழக்கிலுள்ள எழுத்துக்களில் குறியீடுகளின் வரிவடிவத்தில் அதிக மாற்றமும் இல்லாமலேயே குறைந்த எழுத்துக்களில் நிறைந்த தமிழ் தீட்டலாம். (காண்க: அட்டவணை - 4) இந்தக் கணினி எழுத்துச் சீர்மையினை "முத்துக் கணினி-34" (அட்டவணை - 5) என வழங்கலாம்.

அட்டவணை : 3 கனித் தமிழ் வரிவினங்கள்

உயிர்	=	12	
ஆய்தம்	=	1	
மெய்	=	18	
ஒற்றுப்புள்ளி	=	1	
குறியீடுகள்:			
ஆகாரம்	=	1	(ா)
இகரம்	=	1	(ி)
ஈகாரம்	=	1	(ீ)
எகரம்	=	1	(ெ)
ஏகாரம்	=	1	(ே)
உகரம்	=	2	(ஞு, ஙு)
ஊகாரம்	=	2	(ஞூ, ஙூ)
உகர ஒட்டுக் குறி	=	3	(பு, யு, வு)
ஊகார ஒட்டுக்குறி	=	3	(பூ, யூ, வூ)
தனி உகர மெய்கள் (கு - வடிவச் சாயல்)	=	7	(கு,சு,டு,மு,ரு,ளு,ழு)
தனி உகர மெய்கள் (து-வடிவச் சாயல்)	=	6	(து,று,ணு,னு,லு,நு)
ஊகார மெய்கள் (சூ-வடிவச் சாயல்)	=	6	(சூ, டூ, மூ, ளூ, மூ)
ஊகார மெய்கள் (தூ - வடிவச் சாயல்)	=	6	(தூ, றூ, ணூ, னூ, லூ, நூ)
சிறப்பு கூகாரம்	=	1	(கூ)
மொத்த வரிவின எண்ணிக்கை	=	73	

அட்டவணை - 4 கணித்தமிழ் வரிவினங்கள்

உயிரெழுத்து	--	5
ஆய்தம்	--	1
உயிர்க்குறியீடு	--	5
மெய்யெழுத்து	--	18
மெய்ப்புள்ளி	--	1
கிரந்த எழுத்துக்கள்	--	4
		34

அறிவியல் தமிழ் வெறும் செய்தித்தமிழாகவும், பாடத் 'தமிழாகவும் அன்றி, தொழில் நுட்ப உயிர்த் தமிழாகவும் இயல் மாறும் என்பது உறுதி.

பயன்பாடு

1. முத்துக்கணினியின் எழுத்தெண்பலகையில் தமிழெழுத்துக்கள் முப்பத்து நான்கு மட்டுமே இருப்பதால் - விசைப்பலகையின் அனைத்துப் பொத்தான்களிலும் வெவ்வேறு புதிய பல கட்டளைகள் பொதிந்து வைக்கலாம்.

2. தமிழில் நம்மோடு உரையாடும் உத்திகளை ஆணைகளாக உள்ளிடலாம்.

3. ஆங்கிலத்தில் 26 எழுத்துக்களிலும் 40 ஒலியன்களாலும் பெறப்படாத சிறப்பு ழகரம், ணகர, னகர வேறுபாடு, றகர - ரகர வேறுபாடு, லகர - ளகர வேறுபாடு, ஞகரம், ஙகரம் போன்ற பல தனித்தன்மைகள் இப்புதிய கணினி எழுத்துச் சீர்மையில் உண்டு.

4. வடமொழி ஃகிரந்த எழுத்தொலிகளும் அறிவியல் தமிழில் இலகுவாகக் கையாளலாம்.

அட்டவணை 5: 'முத்து கணினி - 34' எழுத்தெண்பலகை

மேல்தளம்

"	1	2	3	4	5	6	7	8	9	0	-	+	/
"	∞	Σ	Δ	π	Λ	∩	|	%	(	)	÷	=	\

கீழ்த்தளம்

ர	ந	எ	ற	த	ய	உ	இ	ஒ	ப	ு	ீ	மேல்தளம்
Q	W	E	R	T	Y	U	I	O	P	!	?	கீழ்த்தளம்

அ	ஷ	ட	ஞ	ங	ஹ	ஜ	க	ல	ெ	ே	மேல்தளம்
A	S	D	F	G	H	J	K	L	:	;	கீழ்த்தளம்

ழ	ா	ச	வ	ஃ	ண	ம	ஸ	ள	ன	மேல்தளம்
Z	X	C	V	B	N	M	,	.	°	கீழ்த்தளம்

5. சொல்லெழுதி (Word Processor) அமைப்பில் தமிழ் எழுத்துக்களினூடே ஆங்கில அகர முதலிகளையும் விசைப்பலகையின் மேல் தளத்திலும் (Upper case), கீழ்த் தளத்திலுமாக (Lower case) பதிந்து வைக்கலாம். மொத்தம் 128 வரிவுகள் (Characters) பதிக்கக் கூடிய எழுத்தெண்பலகையில் 47 பொத்தான்களை "முத்துக் கணினி - 34" ல் குறிப்பிட்டுள்ளவாறு (மேல், கீழ்த்தளங்களிலாக 94 வரிவுகள்) அமைக்கிறோம். ஏனைய பதினேழு பொத்தான்களையும் கணிப்பொறி தன் 'இடக்குறி' (Cursor)யினை இடம், வலம், மேல், கீழாக நகர்த்தவும் வேறு பல கட்டளைத் தொகுப்பு ஆணைகளுக்குமாக வழக்கம் போல் கையாளலாம். (இந்த எழுத்தெண் பலகை அமைப்பே தட்டச்சுப் பொறிக்கும் உதவும் என்பது மற்றொரு நடைமுறை சௌகரியம்). இம்முறையில் பல்வேறு நவீன கணிதக் குறியீடுகளும் இடம் பெறுவது அறிவியல் தமிழப் படைப்பாக்கத்திற்கு வரப் பிரசாதம்.

கணியன் விசைப்பலகை

நா. கோவிந்தசாமி
நன்யாங் தொழில்நுட்பப் பல்கலைக் கழகம்
தேசிய கல்விக் கழகம்
சிங்கப்பூர்

அண்ணா பல்கலைக்கழகத்தில் 1994ம் ஆண்டு ஆகஸ்டு 5,6ம் நாள்களில் நடைபெற்ற தமிழும் கணிப்பொறியும் என்ற ஆய்வரங்கில் படிக்கப்பட்ட கட்டுரை.

அ. முன்னுரை

ரோமன் எழுத்துக்களைப் பயன்படுத்தும் மொழிகளுக்கு, எழுத்து வடிவிலான தகவல்களைக் (Data) கணினிக்குள் உட்புகுத்துவது, மிகவும் எளிதான செயல். ஆனால் சீனம் ஜப்பானியம் போன்ற Ideagram வடிவிலான மொழிகளுக்கு Data input மிகவும் சிரமமான ஒரு செயல். இந்த மொழிகள் அனைத்திற்கும் உள்ள அதே சிரமம், இந்திய மொழிகள் அனைத்திற்கும் உள்ளது.

சிக்கல் நிறைந்த சீனம் ஜப்பானியம் போன்ற ஆசிய மொழிகள் கணினியுகத்தினுள் அடியெடுத்து வைத்தவுடனேயே, மின்னியல் கருவியில் செய்திகளை, எளிதில் உட்புகுத்துவதற்கான ஆய்வுகளில் ஈடுபடத் தொடங்கின. அந்த ஆய்வுகளின் வழி கிடைத்த கருவிகளே, இன்று நடைமுறையில் இருக்கின்றன.

அத்தகைய சிக்கல் நிறைந்த சீன மொழி அதிகம் புழங்கும் ஒரு நாட்டைச் சேர்ந்தவன் நான். எளிதான ஒரு சீன மொழி விசைப்பலகை உருவாக்கத்திற்கு, சீனாவும் தைவானும் அதிகம் பங்காற்றியுள்ளன. இவ்விரு நாடுகளிலுள்ள ஆய்வாளர்கள் உருவாக்கிய எளிதான சீன விசைப்பலகை அமைப்புகள் சிங்கப்பூர்ப் பள்ளிகளில் அதிகமான பயன்பாட்டில் உள்ளன.

சீன மொழி விசைப்பலகை ஆய்வுகள் தொடங்கப்பட்ட அதே கால கட்டத்தில் தான், எங்கள் கல்வி நிறுவனத்தின் சார்பில் நான் எளிதான தமிழ் விசைப்பலகைத் தொடர்பிலான ஓர் ஆய்வைத் தொடங்கினேன். 1988ம் ஆண்டு ஜனவரியில் தொடங்கப்பட்ட ஆய்வு, அதே ஆண்டு டிசம்பரில் முடிவுற்றது. தமிழ்க் கணினிக்கான எளிதான ஒரு விசைப்பலகை, சிங்கப்பூர் (le) விசைப்பலகை என்ற பெயரில் அறிமுகம் கண்டது.

சிங்கப்பூர் மலேசியாவில் விற்பனையில் இருக்கும் பாரதி, வீனஸ், துணைவன், முரசு, இந்தோவொட் (Indoword) போன்ற மென்பொருள்கள் (Softwares) தங்கள் சொற் தொகுப்பில் இவ்விசைப்பலகையை இணைத்துள்ளன. இன்று எங்கள் பள்ளிகளில் 200 வீனஸ் மென் பொருள் வாங்கப்பட்டுள்ளது. 500 பதிவு பெற்ற

தனிநபர்கள்(individual users), தமிழ்க் கணினியைச் (Individual users) சிங்கப்பூரில் பயன்படுத்துகின்றனர். இவர்களில் 95% விழுக்காட்டினர் சிங்கப்பூர் விசைப்பலகை யைத் தான் பயன்படுத்துகின்றனர்.

மலேசியாவில் சுமார் 80 பதிப்பகங்கள் தமிழ்க் க்ணியைப் பயன்படுத்துகின்றன. அதோடு 450 தனிநபர்களும் தமிழ்க் கணினியைப் பய்ன்படுத்துகின்றனர். பதிப்பகங்களில் 95% விழுக்காட்டினர் சிங்கப்பூர் விசைப்பலகை யைப் பயன்படுத்துகின்றனர். தனி நபர்களில் 80% விழுக்காட்டினர் இந்த விசைப்பலகையை பயன்படுத்துகின்றனர்.

ஆக, கடந்த ஐந்தரை ஆண்டுகளாகச் சிங்கப்பூர் - மலேசியாவில் அதிகப் பயன்பாட்டில் இருக்கும் சிங்கப்பூர் விசைப்பலகை யை இந்தியாவில் அறிமுகம் செய்வது இக்கட்டுரையின் நோக்கம். விசைப்பலகையின் பயன்பாடு குறித்து, அறிவியல் நிலையிலிருந்தும், மொழியியல் நிலையிலிருந்தும் விவாதிப்பது முதன்மையான நோக்கம்.

ஆ. பயன்பாட்டில் இருக்கும் மூன்று விசைப்பலகைகள்

எண்பதுகளின் மத்தியில் 'வணிக ரீதியாகத் தமிழ்ச் சொற் தொகுப்பு மென்பொருள்கள் உருவாக்கப்பட்ட போது, 'மென்பொருள் எழுத்தாளர்கள்' விசைப்பலகை தொடர்பான ஆய்வை மேற்கொள்ள முனையவில்லை. வணிக ரீதியாக ஒரு பொருள் சந்தைக்கு வரும் போது, முன்னால் ஸ்தாபிக்கப்பட்ட ஏதோ ஒன்றைத் துணையாகக் கொள்வது. வணிக ரீதியில் வெற்றியளிக்கக்கூடிய ஒன்றாகும். ஆகவே தான் 1988ம் ஆண்டுக்கு முன்பு தமிழ்ச் சொற் தொகுப்பு மென்பொருள்களை உருவாக்கியோர் தமிழ் டைப் ரைட்டர் விசைப்பலகை அமைப்பைப் பயன்படுத்திக் கொண்டனர்.

1958ம் ஆண்ட சென்னை மாநில அரசு அங்கீகரித்த தமிழ் டைப் ரைட்டர் விசைப்பலகை அமைப்பு, இரும்பு, எஃகு போன்ற உலோகங்களால் செய்யப்பட்ட ஒரு கருவிக்கு மிகவும் பொருத்தமானதாக இருக்கலாம். ஆனால் அது இன்றைய மின்னியல் கருவியான கணினிக்கு எந்த வகையில் ஏற்புடையது என்று சிந்தித்தால், அக்கருவியை நாம் அப்படியே ஒதுக்கிவிடுவோம். நாம் பயன்படுத்தும் ஒரு கருவி நவீன தொழில்நுட்பத்திற்கு ஏற்புடையதாக இருக்கவேண்டும். இல்லையென்றால் காலமே அதை ஒதுக்கித் தள்ளிவிடும். டைப்ரைட்டர் காலத்திலிருந்து கணினியுகத்தினுள் அடியெடுத்து, மெல்ல நடந்து கொண்டிருக்கும் நமக்கு, இடைக்காலத்தில், சில வருடங்களுக்கு இக்கருவி தேவைப்படலாம். இதை ஒரு Transitional period கருவி என்று நாம் கொள்ளலாம்.

அப்படி என்றால் தமிழ்க் கணினிக்கு ஏற்ற ஒரு விசைப்பலகை எது?

ஒலியியல் வடிவிலான விசைப்பலகையே தமிழ்க் கணினிக்கு ஏற்ற விசைப்பலகை.

டெல்லியில் இருக்கும் Department of Electronics (DOE) 1988ம் ஆண்டில் தெளிவான ஒரு தீர்வை அறிவித்தது சிக்கல் நிறைந்த இந்திய மொழிகளைக் கணினி மயமாக்குவதற்கு ஏதுவாக ISCII குறியீட்டு முறையையும், ஒலியியல் வடிவிலான ஒரு விசைப்பலகையையும் அறிவித்தது. இவ்விரு திட்டங்களும் தொலைநோக்குடன் அறிவிக்கப்பட்ட திட்டங்களாகும்.

ISCII குறியீட்டு முறை, இந்திய மொழிகளுக்கிடையே தகவல் பரிமாற்றம் நிகழ்வதை உறுதி செய்கிறது. ஒலியியல் விசைப்பலகை தகவல் உட்புகுத்துவதை எளிதாக்குகிறது.

ஆனால் DOE அறிமுகம் செய்த ஒலியியல் விசைப்பலகை தமிழ் மொழிக்குப் பொருத்தமில்லாதது.

தேவநாகிரி எழுத்துக்களின் புழக்கத்தை ஆய்வு செய்து, அதன் அடிப்படையில் ஒரு விசைப்பலகையை உருவாக்கி "அதை மற்ற இந்திய மொழிகள் அப்படியே பின்பற்றவேண்டும்"என்று கூறுவது, இந்திய ஒருமைப்பாட்டிற்கு உகந்ததாக இருக்கலாம். ஆனால் அது ஒரு மொழியின் சிறப்புக் கூறுகளை மனத்தில் கொண்டதாக இருக்க முடியாது. அதிலும் குறிப்பாக நமது தமிழ் மொழிக்கு DOE விசைப்பலகை அமைப்பு எற்புடையதன்று.

DOE விசைப்பலகை ஆய்வில் ஈடுபட்ட அதே கால கட்டத்தில் தான், எங்கள் கல்வி நிறுவனத்தின் சார்பில், ஒலியியல் வடிவிலான தமிழ் விசைப்பலகை ஆய்வில் நான் ஈடுபட்டேன். 24-1-1988ல் எங்கள் கல்வி நிறுவனத்தில் நடைபெற்ற "கணினியும் தமிழ் கற்பித்தலும்" என்ற ஆய்வரங்கில், ஒலியியல் வடிவிலான ஒரு முன்னோட்ட விசைப்பலகை அமைப்பு அறிமுகம் கண்டது. 1988 டிசம்பரில் மாற்றம் பெற்று இன்று சிங்கப்பூர் மலேசியாவில் 90 விழுக்காட்டுக்கு மேல் பயன்பாட்டில்இருக்கும் சிங்கப்பூர் விசைப்பலகை அமைப்பு முறையாக அறிமுகம் கண்டது.

மொரீசியசில் 1989 டிசம்பரில் நடைபெற்ற ஏழாவது உலகத் தமிழ் ஆராய்ச்சி மாநாட்டில் இந்த விசைப்பலகையின் அமைப்புப் பற்றிப் பேசப்பட்டதெனினும், கணினி வல்லுநர்கள் நிறைந்துள்ள இவ்வாய்வரங்கத்தில் அக்கட்டுரையின் சாரத்தைக் கூறுவது பொருத்தமுடைமை என்பதால், அதை வேறொரு நிலையிலிருந்து விளக்க முயல்கிறேன்.

இ. சிங்கப்பூர் விசைப்பலகை அமைப்பின் நோக்கங்கள்

ஒலியியல் வடிவிலான விசைப்பலகையை உருவாக்கும்போது மூன்று தெளிவான நோக்கங்கள் என்னிடம் இருந்தன.

1. அறிவியல் ஆய்வு நிலையிலேயே விசைப்பலகையின் அறிமுகம் நிகழ்தல் வேண்டும்.
2. எளிதில் இயக்கும் வகையில் விசைப்பலகையின் அமைப்பு இருக்க வேண்டும்.
3. தமிழ் இலக்கணத்தை மனத்தில் கொண்டு விசைப்பலகையின் அமைப்பு அமைதல் வேண்டும்.

இனி இந்த மூன்று நோக்கங்கள் விசைப்பலகை அமைப்பிற்கு எவ்வாறு துணை நின்றுள்ளன என்பதைச் சுருக்கமாகப் பார்ப்போம்.

1. அறிவியல் ஆய்வு - தமிழ் எழுத்துகளின் புழக்கம்

விசைப்பலகை அமைப்பதற்காக நான் 1988ம் ஆண்டில் ஓர் ஆய்வில் ஈடுபட்டேன். தமிழ் மொழியில் புழங்கும் உயிர், மெய் ஒலிகளின் விழுக்காட்டளவைக் கண்டறிவது நோக்கம். 27 எழுத்தாளர்களின் 27 உரைநடைப்பகுதிகளில் இடம்பெற்றிருந்த 3150 சொற்களில் உள்ள 25372 உயிர் மெய் ஒலிகள் தனித் தனியாகப் பிரித்து வகைப்படுத்தப்பட்டன.

கீழ்காணும் இரண்டு அட்டவணைகள் உயிர் மெய் எழுத்துகளின் புழக்கத்தை விளக்குகின்றன.

அட்டவணை 1

உயிர் எழுத்துகளின் புழக்கம்

எழுத்துக்கள்	எண்ணிக்கை	விழுக்காடு	நிலை
அ	3630	32.68%	1
உ	2084	18.76%	2
இ	1987	17.89%	3
ஆ	1218	10.96%	4
ஐ	720	6.48%	5
எ	470	4.32%	6
ஏ	292	2.62%	7
ஒ	254	2.28%	8
ஓ	246	2.21%	9
ஈ	100	0.97%	10
ஊ	95	0.85%	11
ஔ	02	0.018%	12

அட்டவணை 2

மெய் எழுத்துகளின் புழக்கம்

எழுத்துக்கள்	எண்ணிக்கை	விழுக்காடு	நிலை
க	2336	16.37%	1
த	1608	11.27%	2
ப	1188	8.32%	3
ம	1142	8.00%	4
ட	944	6.62%	5
ன	884	6.20%	6
ல	849	6.00%	7
ர	819	5.74%	8
வ	815	5.71%	9
ய	735	5.15%	10
ற	735	5.15%	11
ந	499	3.50%	12
ச	485	3.40%	13
ள	466	3.26%	14
ண	355	2.48%	15
ழ	199	1.39%	16
ங	164	1.14%	17
ஞ	43	0.30%	18

2. எளிதாக இயக்கும் வகையில் (உடலியல் கூறுகளை மனத்தில் கொண்டு) விசைப்பலகையின் அமைப்பு அமைதல் வேண்டும்.

1. 90% மேற்பட்டோர் வலது கைப்பழக்கம் உள்ளவர்கள்.
2. வலது, இடது கைகளின் ஆள்காட்டி விரல்களால் அதிகமான வேலைகளைச் செய்ய முடியும்.
3. அதற்கு அடுத்த நிலையில் வலது இடது கைகளின் நடுவிரல்கள் வேலைகளைச் செய்யும்.
4. மோதிர விரல்கள் அதற்கு அடுத்த நிலையிலேயே வேலைகள் செய்யும்.
5. சுண்டு விரல்கள் மிகவும் குறைந்த வேலைகளை மட்டுமே செய்யும். அதிகமான வேலைகளைச் செய்தால் அவை களைப்படைந்துவிடும்.
6. மாற்று விசைகள் மிகவும் குறைந்த அளவில் இருப்பது (Data input) தகவல் உட்புகுத்தலை எளிதாக்கும்.

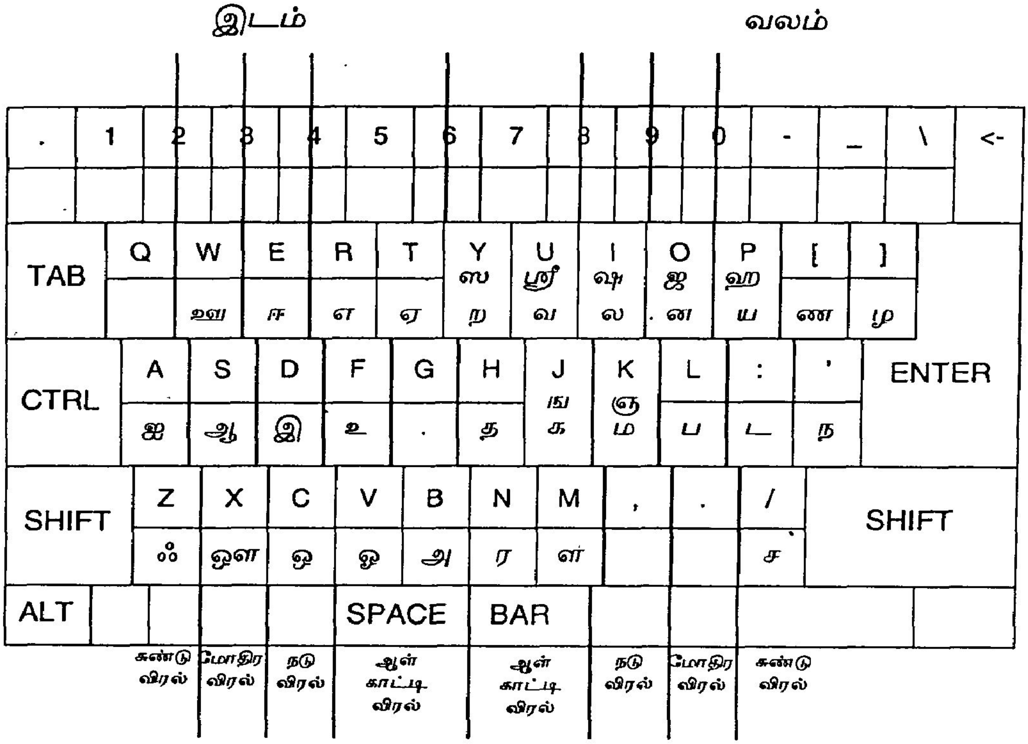
விரல்களின் இந்தச் செயற்பாடு விசைப்பலகை அமைப்பில் கவனமாகக் கருத்தில் கொள்ளப்பட்டது.

அகரம் ஏறிய மெய் எழுத்துகள் வலது கைப்பக்கத்தில் வைக்கப்பட்டன. உயிர் எழுத்துகளும் மெய்ப்புள்ளியும், இடது கைப்பக்கத்தில் வைக்கப்பட்டன.

3. சொல் உருவாக்கத்தில் தமிழ் இலக்கணம் கருத்தில் கொள்ளப்பட வேண்டும்.

1. அகரமேறிய மெய் எழுத்துகள் உயிர் எழுத்துகளுடன் சேர்ந்து உயிர்மெய் எழுத்துகளை உருவாக்குகின்றன.
2. ஒரு சொல்லுக்கு நடுவில் தனி உயிர் வராது· என்ற இலக்கணம் பின்பற்றப்படுகிறது. செய்யுளில் ஓசை நீட்டிப்புக்காக உயிர் வரவேண்டுமானால், Shift விசையின் மூலம் சொல்லுக்கு இடையில் உருவாக்கலாம்.(எ.கா துப்பாய தூஉம் மழை).
3. சொல்லுக்கு முதலில் உயிர் எழுத்து வரும் என்ற விதியும் பின்பற்றப்படுகிறது. இடைவெளிக்கான Space bar விசையைத் தட்டிய பிறகு, உயிர் எழுத்தை Shift இல்லாமலேயே அடிக்கலாம். 1988 ஆண்டு டிசம்பரில் துணைவன் சொற்தொகுப்பு சாதனமும், மே 1989ல் பாரதி சொற்தொகுப்புச் சாதனமும், 1989 பிற்பகுதியில் முரசு சொற்தொகுப்புச் சாதனமும், 90ல் வீனஸ் சொற்தொகுப்புச் சாதனமும், 1994ல் இந்தோவொட் சொற்தொகுப்புச் சாதனமும் தங்கள் மென்பொருள்களில் சிங்கப்பூர் விசைப்பலகை வசதியை ஏற்படுத்திக்கொடுத்தன.

விசைப்பலகை அமைப்பு இது தான்



ஈ. அச்சடிப்பதற்கான பயிற்சி (Type Writing Manual) உருவாக்கம்

1989ம் ஆண்டு ஜூலையிலிருந்து எங்கள் பாடத்திட்டத்தில் 30 மணி நேரம் application Software Training *பாடம் நடைபெறுகிறது. முதல் 6 மணி நேரம் அச்சடிப்பதற்கான பயிற்சியும், மீதமுள்ள 24 மணி நேரத்தில் வீனஸ், பாரதி, துணைவன், முரசு போன்ற சொற்தொகுப்புச் சாதனங்களை இயக்குவதற்கான முறையான பயிற்சிகள் எங்கள் பல்கலைக் கழகத்தில் அறிமுகமாயின. நான்கு மணி நேரப் பயிற்சிக்குப் பிறகு, எழுத்துகளைப் பார்க்காமல் அடிப்பதற்கான பயிற்சியை* Touch Method Practice *மூலமாக பயிற்சியாளர்கள் பெற்றனர். இந்த* Touch Method Praactice *என்று சொல்லப்படும் ''விரல்களின் உணர்வு நிலைப் பயிற்சிக்கென்று நான்கு தொகுதிப் பாடங்கள் உருவாக்கப்பட*

விரைவுப் பயிற்சிக்கென்று ஒரு தொகுதிப் பாடம் உருவாக்கப்பட்டது.

முதல் நான்கு தொகுதிப் பாடங்களில், நான்கு மணி நேரப்பயிற்சியைப் பெறும் ஒருவர், உணர்வு நிலையிலேயே எழுத்துகளை அடிக்கும் முறையில் பயிற்சிகள் உருவாக்கப்பட்டன. தமிழ் இலக்கணத்தை அடிப்படையாக கொண்டு இந்த விசைப்பலகை உருவாக்கப்பட்டதால், இது எளிதில் சாத்தியமாயிற்று.

1989, 1990ம் ஆண்டுகளில் விசைப்பலகைப் பயிற்சிக்கென்று நான் உருவாக்கிய பயிற்சிகளை முரசு TOOLS மூலமாகக் கணினி வழியாக்கினேன்.

1991ம் ஆண்டு முதல் கணினி வழியில் பயிற்சியாளர்கள், இப்பயிற்சியைப் பெறத் தொடங்கினர்.

சிங்கப்பூர் விசைப்பலகை விசைகளின் புழக்க நிலை

இவ்வாரங்க விவாதத்திற்காக, விசைப்பலகையில் உள்ள விசைகளின் புழக்க நிலையை அறிந்துகொள்வதற்கு, நான் உருவாக்கிய 16 விரைவுப் பயிற்சிக்கான பாடங்களில் இடம்பெற்றுள்ள மொத்த பதிவுகளை வகைப்படுத்திப் பார்த்தேன். கீழ்க்காணும் அட்டவணைகள், விசைகளின் புழக்க நிலையை மிக நுட்பமாகப் பார்க்கின்றன.

அட்டவணை 3

இடது கைவிரல் விசைகளின் புழக்கம்

விசைகள்	எண்ணிக்கை	விழுக்காடு	நிலை
புள்ளி	3905	37.714%	1
உ	1856	17.925%	2
இ	1578	15.240%	3
ஆ	906	8.750%	4
ஐ	653	6.306%	5
எ	352	3.399%	6
ஏ	249	2.405%	7
ஒ	248	2.395%	8
அ	246	2.375%	9
ஓ	222	2.144%	10
ஊ	84	0.811%	11
ஈ	54	0.521%	12
ஔ	01	0.009%	13

அட்டவணை 4

வலது கைவிரல் விசைகளின் புழக்கம்

விசைகள்	எண்ணிக்கை	விழுக்காடு	நிலை
க	1945	15.356%	1
த	1505	11.882%	2
ம	1188	8.676%	3
ப	1017	8.029%	4
ட	848	6.695%	5
வ	819	6.466%	6
ர	752	5.937%	7
ன	748	5.905%	8
ல	706	5.573%	9
ற	682	5.384%	10
ய	612	4.831%	11
ள	562	4.437%	12
ச	399	3.150%	13
ந	376	2.968%	14
ண	253	1.997%	15
ழ	219	1.729%	16
ங	94	0.742%	17
ஞ	21	0.165%	18

அட்டவணை 5

இடது கைவிரல்களின் புழக்கம்		
1.	இடது கை ஆள்காட்டி விரல்	66.213%
2.	இடது கை நடு விரல்	17.905%
3.	இடது கை மோதிர விரல்	09.570%
4.	இடது கை சுண்டுவிரல்	06.306%

அட்டவணை 6

வலது கைவிரல்களின் புழக்கம்		
1.	வலது கை ஆள்காட்டி விரல்	49.462%
2.	வலது கை நடு விரல்	14.249%
3.	வலது கை மோதிர விரல்	13.934%
4.	வலது கை சுண்டுவிரல்	21.37%

விசைகளின் புழக்கம் – முடிவுகள் (Findings)

பொதுவாகப் பார்த்தால்

1. Home Keys விசைகள் 68.66% புழக்கத்தில் உள்ளன.
2. Home Keys விசைகளுக்கு மேல் உள்ள விசைகள் 20.74% புழக்கத்தில் உள்ளன.
3. Home Keys விசைகளுக்கு கீழ் உள்ள விசைகள் 10.55% விழுக்காட்டுப் புழக்கத்தில் உள்ளன.
4. மாற்று விசையின் புழக்கம் 0.512% மட்டுமே.

இடது கை, வலது கைப்புழக்கத்தைத் தனித்தனியாகப் பார்த்தால்

5. இடது கை ஆள்காட்டி விரல் மிக அதிகமாக, 66.213% விழுக்காடு புழங்குகிறது.
6. இடது கை நடுவிரல் 17.905% விழுக்காடு பயன்படுத்தப்படுகிறது.
7. இடது கை மோதிரவிரல் 9.570% விழுக்காடு புழங்குகிறது.
8. இடது கை சுண்டுவிரல் 06.306% விழுக்காடு மட்டுமே பயன்படுத்தப்படுகிறது.
9. வலது கை ஆள்காட்டி விரல் வலப் புறத்தில் மிக அதிகமாக 49.462% விழுக்காடு பயன்படுத்தப்படுகிறது.
10. வலது கை நடுவிரல் 14.249% விழுக்காடு புழங்குகிறது.
11. வலது கை மோதிரவிரல் 13.934% விழுக்காடு புழங்குகிறது.
12. இடது கை சுண்டுவிரல் போன்று அல்லாமல், வலது கை சுண்டு விரல் 21.37% விழுக்காடு பயன்படுத்தப்படுகிறது.

ஊ. தொகுப்புரை

இது வரையில் என் கட்டுரையில் நான் விவாதித்த செய்திகளை இவ்வாறு தொகுத்துக் கூறலாம்.

1. டைப் ரைட்டர் விசைப்பலகை தற்காலிகமான ஓர் ஏற்பாடு போன்ற ஒரு நிலையிலேயே பயன்பாட்டில் இருந்துகொண்டிருக்கும். எஃகு உலோகத்திற்காக உருவாக்கப்பட்ட இந்த விசைப்பலகை காலப்போக்கில் ஒலியியல் விசைப்பலகைக்கு இடம் கொடுத்து ஒதுங்கிவிடும்.
2. தமிழ் கணினி மென் பொருள்களை உருவாக்குவோர், DOE அறிவித்துள்ள ISCII குறியீடுகளின் வழி இயங்கும் ஒலியியல் விசைப்பலகையை உருவாக்குவது முக்கியமான ஒரு செயற்பாடாக விளங்கும்.
3. ஒலியியல் வடிவிலான DOE விசைப்பலகை அமைப்புள்ள ஒரு மென் பொருளில் சிங்கப்பூர் விசைப்பலகையை அமைப்பது மிகவும் எளிதான செயல். தகவல் பரிமாற்றத்தில் எந்த விதப்பாதிப்பும் இல்லாமல், சிங்கப்பூர் விசைப்பலகை அமைப்பை அமைக்க முடியும்.

4. DOE விசைப்பலகை, சிங்கப்பூர் விசைப்பலகை ஆகிய இரண்டு விசைப்பலகைகள் மட்டுமே காலப்போக்கில் நிலைத்து நிற்கும். இந்த இரண்டு விசைப்பலகைகளில், பயன்படுத்துவோர்க்கு எளிதில் கைவரக்கூடிய ஒரு விசைப்பலகையே நின்று நிலைக்கும்.

5. இந்த என் கட்டுரை, சிங்கப்பூர் விசைப்பலகை நிலைத்து நிற்பதற்கான வாய்ப்புகளை அறிவியல் நிலையிலும், மொழியியல் நிலையிலும் பெற்றிருக்கிறது என்பதை விவாதித்தது.

எ. பரிந்துரை

கீழ்க்காணும் ஒரு பரிந்துரையை நான் முன் வைக்கிறேன்.

1. தமிழ் நாட்டில் தற்போது பயன்பாட்டில் இருக்கும் சொற்தொகுப்பு மென்பொருள்கள், சிங்கப்பூர் விசைப்பலகை அமைப்பை ஏற்படுத்திக்கொடுக்கவேண்டும்.

ஏ. முடிவுரை

எழுத்தாளர் சுஜாதா அவர்கள் சிங்கபூருக்கு வந்திருந்த போது இதன் செயற்பாட்டை நேரில் கண்டார். இந்த விசைப்பலகையைத் தமிழ் புழங்கும் நாடுகளில் பரவலாக்க இதன் பெயரை கணியன் விசைப்பலகை என்று மாற்றலாம் என்ற ஒரு கருத்தை வெளியிட்டார். இந்த விசைப்பலகை தமிழ் நாட்டில் பரவலாக ஏற்றுக் கொள்ளப்படுவதற்கான நடவடிக்கைகள் எடுக்கப்படுமானால்,எங்கள் கல்வி நிறுவனத்தில் கலந்தாலோசித்து இதன் பெயரைக் கணியன் விசைப்பலகை என்று மாற்றி அமைப்பதற்கு முயற்சியில் நான் ஈடுபடுவேன். ‘‘யாதும் ஊரே, யாவரும் கேளிர்’’ என்று பாடிய சங்க புலவன் கணியன் பூங்குன்றனார்க்குச் சிறப்புச்சேர்க்கும் முயற்சியில் நாம் ஈடுபடுவோம்.

மேற்கோள் கட்டுரைகள்

1. எஸ் சீனிவாசன், தமிழிலேயே தந்திச் செய்திகளை அனுப்பலாம். தமிழ் நேசன்.(மறுபதிப்பு)8.11.1987

2. நா.கோவிந்தசாமி, ஒருமைப்படுத்தப்பட்ட கணினி விசைப்பலகை அமைப்பு 24-1.1988ம் நாள் சிங்கப்பூர் கல்விக் கழகத்தில் நடைபெற்ற ‘‘கணினியும் தமிழ் கற்பித்தலும்’’ என்ற ஆய்வரங்கில் படிக்கப்பட்ட கட்டுரை.

3. நா. கோவிந்தசாமி . கணினி வழி, தமிழ் கற்றல் இரண்டாவது உலகத்தமிழ் ஆசிரியர் மாநாட்டு ஆய்வடங்கல், கோலாலம்பூர், மலேசியா, ஜூன் 1994.

பி.கு.இந்த விசைப்பலகையைத்தங்கள் கணினிச்சாதனங்களில்பயன்படுத்த விரும்புவோர் அதற்கான அனுமதியை இலவசமாகக் கீழ்க்காணும் முகவரியில் தொடர்பு கொண்டு பெறலாம்.

N. Govindasamy, Lecturer, Nanyang Technological University, National Institute of Education, 469 Bukit Timah Road, Singapore - 1025. E.Mail : Govin@am.nie.ac.sg@pmdf

கணிப்பொறித்துறைக் கலைச் சொல்லாக்கம்

சொல்லெல்லாம் கலைச் சொல்லாகுமா?

சு. சீனிவாசன்,
அறிவியல் அதிகாரி, கணிப்பொறி மையம்,
இந்திராகாந்தி அணுவாராய்ச்சி மையம்,
கல்பாக்கம் - 603 102

நித்தம் வளர்ந்து வரும் அறிவியல் துறையில் ஆங்கிலத்தின் தாக்கம் பெருகி வருவது கண்கூடு. கணிப்பொறி அறிவியல் என்பது ஒரு பல்துறை சார் ஆய்வாக உருவெடுத்திருக்கிறது. கணிதவியல், மின்னணுவியல், இயற்பியல் ஆகியவற்றின் சங்கமமாக கணிப்பொறித் துறை விளங்குகிறது.

இத்துறையில் கலைச் சொல்லாக்கத்திற்கு தேவையான சொற்களைத் திரட்டுவது பிரச்சினை அன்று. க்ணிப்பொறித்துறைக்கென்று சொற்பொருள் விளக்கம் கூறும் (annotated dictionary) அகராதிகள் பல இன்று ஆங்கிலத்தில் எழுந்துள்ளன; பல முளைத்தும் வருகின்றன. இவற்றை அடிப்படையாகக் கொண்டு கலைச் சொல்லாக்கத்திற்குத் தேவையான ஆயிரமோ பதினாயிரமோ சொற்களைத் தெரிவு செய்து விடலாம். அடுத்து, தமிழில் கலைச்சொல் சமைப்பதற்கு ஒரு நல்ல வழிகாட்டுதல் நெறி தேவைப்படுகிறது. அதற்கென நால்வகை நெறிகளை இங்கே காண்போம்.

1. *எச்சொற்களுக்கு கலைச்சொல்லாக்கம் தேவை? எவற்றை ஆங்கிலத்திலுள்ளது போல் ஏற்றுக்கொள்ளலாம்? என்பதற்கு விடை அறிதல் வேண்டும்.*

2. *மொழிக்குடும்பத்தால் ஆங்கிலத்துடன் ஒன்றுபட்ட பிரெஞ்சு, ஜெர்மானியரின் போக்கையும் பட்டறிவையும் பயன்படுத்திக் கொள்வது.*

3. *வழக்கொழிந்த சொற்களை விடுத்து, தேவை எனபவற்றுக்கும் இன்றியமையாதவை என்று கருதுபவற்றுக்கும் உடனடியாகக் கலைச்சொல்லாக்கம் மேற்கொள்வது.*

4. *அடிப்படை கலைச்சொல் புனைவிற்கு எங்ஙனம் தமிழர்தம் நுண்ணறிவைப் பயன்படுத்திக் கொள்வது.*

இனி, இவற்றை ஒவ்வொன்றாய் விரிந்து நோக்கவோம்.

கற்றல்(learning), கற்றுக்களைதல் (unlearning) மீண்டும் கற்றல் (relearning) என்ற மனப்போக்கை நாம் அறவே விட்டொழிக்க வேண்டும். பிராணவாயு என்ற சொல் தமிழ்வயப்பட்டு உயிர்வளி ஆயிற்று. ஆக்சைடு என்ற சொல்லை உருவாக்க உயிர்வளியைத் திரிக்கும் முயற்சி தோற்றுப்போய் அதை அப்படியே கையாள வேண்டியதாயிற்று. பிறகு பொதுமைப்பண்பு (Universality Principle)அடிப்படையில் உயிர்வளி என்ற சொல்லை நீக்கி, ஆக்ஸிஜன் புகுத்த வேண்டியிருந்தது. இதற்கு நாம் எடுத்துக் கொண்ட காலம் 30 ஆண்டுகள். இனி, இத்தகு காலவிரய முயற்சியில் நாம் இறங்குதல் கூடாது என்பதை மனதில் நிறுத்திக் கொள்ளவேண்டும்.

மேலை மயமாக்கம்

ஆங்கிலத்திலுள்ள அடிப்படைக் கலைச்சொற்கள் கணிதவியல், இயற்பியல், மின்னணுவியல், வேதியியல், உயிரியல், மருத்துவம் முதலிய துறைகளில் நிறைவை எட்டி விட்டன. Force/ Power/ energy ஆகிய அடிப்படை கலைச்சொற்கள் முறையே விசை, திறன், ஆற்றல் எனத் தமிழில் செந்தரப்(Standardiation) படுத்தப்பட்டுள்ளன. இனி, உருவாகி வருபவை பெரும்பாலும் புனைவின் (Innovation) அடிப்படையில் இடுகுறிப் பெயர்களாகவே அமைக்கப்பட்டு வருவதால் பிறமொழியாரும் நாட்டாரும் அவற்றை அப்படியே ஆளத்தலைப்பட்டுள்ளனர். இவற்றுக்கு எடுத்துக்காட்டாக டிரான்சிஸ்டர், லேசர், ஹைப்பர்டெக்ஸ்ட் முதலிய சொற்களைக் குறிப்பிடலாம். இந்த நிலையில் நாமும் அவர்கள் போக்கைக் கடைபிடிப்பது பயனுள்ளதாகவே தோன்றுகிறது.

ஆங்கிலம், ஜெர்மன், பிரெஞ்சு ஆகிய மொழிகளில் மேற்கொள்ளப்பட்டுள்ள கலைச்சொல் பணிகளிலிருந்து நாம் கற்றுக்கொள்ள வேண்டியது ஏராளம். இம்மொழிகளின் மற்றொரு சிறப்பு, இவை பேசும் நாடுகளில் அறிவியல், ஆய்வுப்பணியும் அவ்வண் மொழிகளிலேயே நடைபெறுவதாகும்.

ஆங்கிலம் நமக்கு கண்ணாடி

அறிவியல் துறையில் நிகழும் கண்டுபிடிப்புகளை உடனுக்குடன் அறிய ஆங்கிலம் நமக்கு ஒரு கண்ணாடியாகவே துணை நிற்கிறது. உலகத்தாருக்கு நம் கருத்துகளைத் தெரிவிக்கவும் ஆங்கிலும் பயன்படுகிறது. ஆங்கிலம் இன்று ஒரு முக்கிய அறிவியல் தகவல் தொடர்பு மொழியாக உருவாகி வருகிறது. இந்த நிலையில் புதிய கலைச்சொற்கள் பல ஆங்கிலத்திலேயே முதலில் படைக்கப்படுகின்றன.

ஜெர்மன் நமக்கு வழிகாட்டி

தமிழ் மொழியைப் போலவே ஜெர்மன் மொழியிலும் பல கலைச்சொற்கள் வேர்ச்சொல் அடிப்படையிலேயே உருவாக்கப்படுகின்றன. எடுத்துக்காட்டாக நீர்வளி (Wasserstoff) உயிர்வளி (Sauerstoff). நீராவி (Wasserdampf), தொலைக்காட்சி (Fernsenen) ஆகிய சொற்களை நாம் உருவாக்கும் முறையும் ஜெர்மானியர் படைக்கும் முறையும் ஒத்திருக்கின்றன. தமிழரைப் போலவே ஜெர்மானியரும் மொழித் தூய்மையைப் பேணுபவர்கள். இவற்றுக்கு எடுத்துக்காட்டாக ஜெர்மன் மொழியிலுள்ள வெப்பம் (Warme), திறன் (leistung), மின்னூட்டம் (Ladung) ஆகிய சொற்களைக் குறிப்பிடலாம்.

பிறமொழிச் சொற்களை ஏற்றுக்கொள்வதில் தயக்கம் காட்டுபவர்கள் ஜெர்மானியர். தமிழர் வடசொற்களை ஆள்வதை விரும்பார். இதுபோல கிரேக்க, லத்தீன் மொழி மூலங்களை ஜெர்மானியர் எடுத்து ஆள்வதில்லை. இவர்கள் தமக்கென ஒன்றை உருவாக்கிக் கொள்வர். எடுத்துக்காட்டாக television/ telescope ஆகியவற்றுக்கு இணையான சொற்களைக் குறிப்பிடலாம். தொலைக்காட்சியை Fernsenen என்றே ஜெர்மானியர் அழைக்கின்றனர். சொற்றொடரில் இடம் பெறும் பெயர்ச் சொற்களின் முதல் எழுத்து ஜெர்மன் மொழியில் எல்லாவிடங்களிலும் பெரிய எழுத்தில் (Capital) குறிக்கப்படுகிறது.

தமிழைப் போலவே ஜெர்மன் மொழியிலும் கூட்டுச் சொற்களும் சந்திகள் உபயோகமும் மிகுதி. இதனால் பிறமொழியர் ஜெர்மன் மொழியிலுள்ள அறிவியல் கட்டுரைகளைப் படித்துப் புரிந்து கொள்ள கூடுதல் சிரமப்படுவர். அணிமைக்காலத்தில் ஜெர்மானியர் போக்கிலும் பெரிய மாற்றம் ஏற்பட்டு வருகிறது. பல அறிவியல் துறைச் சொற்களை ஒலிபெயர்க்க துவங்கியுள்ளனர். கணிப்பொறியை Rechner என அழைக்கும் நிலையிலிருந்து Computer* எனவே ஆள முற்பட்டுள்ளனர்.

பிரெஞ்சு நமக்கு முன்னோடி

லத்தீன் மொழியின் சிதைவிலிருந்து உருவாகியது பிரெஞ்சு மொழியாகும். எனவே பிரெஞ்சுக்காரர்கள் லத்தீன் அடைமொழிகளை ஏற்கத் தயங்கார். பல கலைச்சொற்கள் பிரெஞ்சு மொழியிலும் ஆங்கிலத்திலும் ஒன்றாகவே அமைந்துள்ளன. இவற்றுக்கு எடுத்துக்காட்டாக agriculture, Calculator, Charge, radio, reflection, solution, முதலிய சொற்களைக் குறிப்பிடலாம்.

ஆங்கிலத்தைக் காட்டிலும் பிரெஞ்சு எழுத்து நடையில் ஒலியன் (Phoneme) தெளிவு அதிகம். ஜெர்மன் மொழிபோல் சொற்களைப் பிரித்துப் பொருள் கொள்ள வேண்டிய தேவை இல்லை.

* பிரெஞ்சு மொழியில் ordinateur என அழைக்கப்படுகிறது

நம் அணுகு முறை

தமிழ், ஐரோப்பிய மொழி குடும்பத்திலிருந்து முற்றிலும் மாறுபட்ட திராவிட மொழிக் குடும்பத்தைச் சார்ந்திருக்கிறது. எனவே ஆங்கிலத்திலுள்ள கணிப்பொறிச்சொற்கள் அனைத்தையும் மொழியாக்கம் செய்வது தேவை எனத் தோன்றுவது இயல்பு. ஆனால் இது நடைமுறை சாத்தியமன்று.

ஆங்கிலச் சொல்லுக்கு மொழியாக்கம் ஜெர்மன், பிரெஞ்சு மொழிகளில் வேறுபட்டு இருப்பின், நாமும் அவ்விடங்களில் மொழியாக்கம் மேற்கொள்ளலாம். அடிப்படை கலைச்சொல் நீங்கலாக, பிற புனைவுப் பெயர்கள் இம் மொழிகளில் ஒன்றாகவே ஆளப்படின் நாமும் அவற்றைப் பின்பற்றலாம். எடுத்துக்காட்டாக helium, transister, vector முதலிய சொற்கள் ஒன்றாகவே ஆளப்படுவதைக் கவனிக்கலாம்.

Hypertext, Hologram முதலிய கலைச்சொற்களின் உட்பொருளை ஒரு வார்த்தையில் மொழிபெயர்ப்பு செய்வது எல்லா மொழிகளிலும் கடினம். கிரேக்க, லத்தீன்மொழிக் கூறுகளைப் பயன்படுத்தியே ஆங்கிலத்திலும் இச்சொற்கள் உருவாக்கப்படுகின்றன.

Greek welds words where English has to string them

- Richmond Lattimore

இந்த மரபுக்கு உலகளவில் எதிர்ப்பு இருப்பதாகத் தெரியவில்லை. புனைவுச் சொற்களின் உட்பொருளை உயர் அறிவியல் பயிலா மொழிப் பண்டிதர்களாலும் மொழியாக்கம் செய்வது இயலா.

உயிர்வளி, நீர்வளி, கரியமிலவாயு, ஆகிய சொற்களை முறையே ஆக்ஸிஜன், ஹைட்ரஜன், கார்பன்-டை-ஆக்ஸைடு எனத் தமிழரும் உலக செந்தரச் சொல் வழக்கில் ஆளத் துவங்கி இருக்கின்றனர். Photograph என்பதை புகைப்படம், நிழற்படம் என்று சொல்வது மறைந்து ஒளிப்படம் என்றாகி இருக்கிறது. Photo என்ற முன்னொட்டுக்கு 'ஒளி' என்ற பதம் தமிழில் உருவாகி ஆட்சி செய்யத் துவங்கி இருக்கிறது. இருப்பினும் தமிழில் அறிவியல் கலைச் சொற்களை நெறிப்படுத்தவும் செந்தரப்படுத்தவும் முயற்சிகள் பெரிய அளவில் துவங்கப்படவில்லை.

வான்ஊர்தி வேகத்தில் வளர்ந்து வரும் கணிப்பொறித் துறையில் சொற்கள் பல புதிதாக உருவாகின்றன. இவற்றில் சில நிலைக்கின்றன. சில மறைகின்றன, Hollarith Card, Punch, Pause ஆகிய சொற்கள் வழக்கொழிந்து விட்டன. எனவே தேவை எனப்படுவற்றுக்கும் இன்றியமையாதன என கருதப்படும் அடிப்படைச் சொற்களுக்கும் உடனடியாக கலைச்சொல்லாக்கம் மேறகொள்வது சால்பு (பிற்சேர்க்கை 1 & 2-ஐக் காண்க).

அடிப்படைச் சொற்களுக்கு கலைச்சொல்லாக்கம் காணும் போது தமிழர் தம் நுண்ணறிவையும் போற்றுதல் வேண்டும். ஈராயிரம் ஆண்டு பழமை உடைத்த தமிழ் மொழியில் அமைந்துள்ள 'யாப்பிலக்கணக்கோட்பாடு' கணித அடிப்படையில் வேரூன்றி மிளிர்வதைக் காணலாம். பைனரி இலக்கம் என்று நாம் கூறும் உண்டு/ இல்லை பண்பை நேர்/நிரை என்ற "அசை" உறுப்பாக நம் முன்னோர்கள் கையாண்டிருக்கிறார்கள். இச்செய்யுள் இலக்கண வழிநின்று bit/ byte, word,* short, long, parity ஆகியவற்றுக்கு முறையே அசை∂, சீர், அடி, குறளடி, நெடிலடி, தொடை எனப் பெயர்ச் சூட்டவும் இயலும். இம்முயற்சி வரவேற்கத்தக்கதா என அறிஞர் பெருமக்கள் தம் கருத்தை முன்வைத்தல் வேண்டும். P.T.O (பிறசேர்க்கை 2-ஐ காண்க).

தையற்பொறி, கடைசல்பொறி, அரவை எந்திரம் போல கணிப்பொறியும் ஒரு கருவியே, ஆனால் முன் விளம்பியவை போல் ஒரு குறிப்பிட்ட பணிக்கென அல்லாமல் பல்துறையிலும் பயன்படவல்லது , கணிப்பொறி. இதன் பயன்பாட்டுக்கு எல்லை விரித்துரைக்க இயலாது.

இக்கட்டுரையின் கண் மொழிந்த வழிமுறைகளைப் பின்பற்றி கலைச்சொல் சமைத்தால்மட்டும்நம்பணி முற்றுப்பெற்று விடும்என கூறுவதற்கில்லை. சமைத்தல் வேறு;அமைதல் வேறு. சமைத்த கலைச் சொற்களை நூல்வடிவில் எழுதிப் பயன்படுத்த வேண்டும். அவ்வேளையில் இடம் பொருள் காரணமாக வழுவிய சொற்கள் மறையலாம். செவ்விய சொற்கள் தழைக்கலாம். இவ்வாறு அமையும் சொற்களே 'கலைச்சொல்' தகுதி பெறும். இச்சூழல் உருவாக தமிழில் கணிப்பொறி அறிவியரைப் பயிற்றுவிப்பதற்கு முயற்சிகள் மேற்கொள்ளப்பட வேண்டும் என செப்பவும் வேண்டுமோ!

---

* Bit-the smallest unit of storage. Byte - fixed number of bits. Word - Collection of bytes

∂ எழுத்தினாலாவது அசை. அசையினாலாவது சீர். பல சீர்களால் அமைவது அடி.

பிற்சேர்க்கை – 1
அஸ்கீ உருக்களைப் பெயர்கள்

!	Exclamation	வியங்கோள்
"	Quotation Mark	மேற்கோள்
#	Number Sign	எண் குறியீடு
$	Currency Symbol	நாணயக் குறி
%	Percent	விழுக்காடு
&	Ampersand	உம்மைக் குறி
`	Apostrophe	உடைமைக் குறி
(	Opening Parenthesis	பிறைத்துவக்கம்
)	Closing Parenthesis	பிறை முடிவு
*	Asterisk	உடுக் குறி
+	Plus	கூட்டல்
,	Comma	காற்புள்ளி
-	Hyphen or Minus	இணைப்பு (அ) கழித்தல்
.	Period	முற்றுப்புள்ளி
/	Slant	சாய்கோடு
:	Colon	முக்காற்புள்ளி
;	Semi Colon	அரைப்புள்ளி
<	Less than	வீழ் குறி
=	Equal	நிகர் குறி
>	Greater than	வெற்றிக் குறி
?	Question mark	வினாக் குறி
@	Commercial at	வீதக் குறி
[	Opening Bracket	பகரத் துவக்கம்
\	Reverse Slant	இடம் சாய்கோடு
]	Closing Bracket	பகர முடிவு
^	Circumflex	கூரைக் குறி
_	Underline	அடிக்கோடு
`	Grave accent	பலுக்கக் குறி
{	Opening Brace	காப்புத் துவக்கம்
|	Vertical line	குத்துக்கோடு
}	Closing Brace	காப்பு முடிவு
~	Tilde	தலை பெய் குறி
Blank Space		வெளி

பிற்சேர்க்கை - 2
அடிப்படை கலைச் சொற்கள்

Hardware - வன்கலன்
Software - மென்கலன்
Firmware - இடைக்கலன்
Bit - அசை
Byte - சீர்
Word - அடி
Nibble - குறுஞ்சீர்
Long word - நெடிலடி
Short word - குறளடி
Parity - தொடை
Boundary - தளை
Binary - இருமம
Octal - எண்மம
Decimal - தசமம
Hexadecimal - பதினறுமம
Binary tree - மரக்கிளை
Flag - பறை
Letter - எழுத்து
Number - எண்
Digit - இலக்கம்
Character - உரு
Special Character - சிறப்புரு
Character set - உருக்கணம்
Alphabet - எழுத்துரு
Alphanumeric - அகர எண்ணுரு
Sign - குறி
Symbol - இடு குறி, சின்னம்
Code - குறியீடு
Mode - பாணி
Mnemonic - குழூஉக்குறி
Notation - குறிமானம்
Cyptogram - மறைக் குறியீடு
Integer - முழுவெண்
Real - மெய்யெண்
Fraction - பின்னம்
Constant - மாறிலி

Variable - மாறி
Mantissa - வருமொழி இலக்கம்
Characteristic - நிலைமொழி இலக்கம்
Exponent - அடுக்கு
Radix - அடிமூலம்
Fixed Point - இடம்பெயரா
Floating point - இடம்பெயர்
Modulus - மீயெண்
Least Significant digit - குறு வளஞ்சார் இலக்கம்
Most Significant digit - கழிவளஞ்சார் இலக்கம்
Pseudo - போலி
Random - பிறப்பொக்கு எண்கள்
Operation - வினை, செயல்
Power - படி
Truncate - ஈறுஅறு
Rounding Off - ஈறுதிரி
Static Variables - சலனமிகா மாறிகள்
Dynamic Variables - சலனமிகு மாறிகள்
Automatic Variables - படைப்பு மாறிகள்
Global Variables - நிலைப்பு மாறிகள்
Instruction - ஆணை
Statement - கூற்று
Program - நிரல் தொடர்
Programming - நிரல் தொடராக்கம்
Source program - சுனை நிரல் தொடர்

Object program - மறை நிரல் தொடர்
Executable program - செயல் நிரல் தொடர்
Routine - வழித்தொடர்
Subroutine - துணை வழித்தொடர்
Execute - செயற்படுத்து
Run - ஓட்டு
Process - வழிப்படுத்து
Step - படிநிலை
Expression - கோவை
Control - கட்டுறு
Condition - நிபந்தனை
Comment - குறிப்புரை
Call - விளி, அழை
String - சரம்
Stream - மடை
Nesting - அணை
Block - திரள்
Iteration - மீள்தொடர்
Do loop - வினைச்சுழல்
Sequence - வரிசை
Serial - தொடர்
Queue - சாரை
List - பட்டியல்
Table - அட்டவணை
Scheduling - வரையறு
Array - அணி
Matrix - அணிக்கோவை
Row - நிரையணி
Column - நிரலணி
Keyboard - கைஅகம், விசைப்பலகை
Disk - வட்டு
Input - உள்ளீடு
Output - வெளியீடு
Sink - மடு
Source - சுனை
Jump - தாவிச்செல்

Tabbing - தத்திச்செல்
Compatible - தகவமை, ஏற்புடை
Corruption - கறை, களங்கம்
Scroll - வரிஉலா
Spooler - திரட்டி
Split Screen - பகுதிரை
Terminal - முனையம்
Exit - வெளியேறு
Quit - விடுபடு
Halt - தங்கு
Wait - காத்திரு
Hold - தாமதி
Stop - நிறுத்து
Return - திரும்பு
Terminate - துண்டி
Memory - நினைவகம்
Storage - கிடங்கு
Register - பதிவு நினைவகம்
RAM - இச்சை அணுகு நினைவகம்
Stack - கூப்பு
Heap - குவியல்
Cache - (உள்)ளுறை நினைவகம்
Buffer - இடைப்படு (அ) கைமாறு நினைவகம்
Scratch Pad - இரவல் ஏடு
Editor - வரிபேணி
Word Process - சொல்பேணி
Emulator - வழிபேணி
Simulator - பாவனை பேணி
Compiler - தொகுத்துரைப்பி
Intrepreter - பகுத்துரைப்பி
Assembler - இணைந்துரைப்பி
Operating system - வினைக்கலன்

Data base	- விபரக்கலன்	Redudancy	- கழிமிகை
Spread Sheet	- விரிமடல்	Sort	- ஒழுக்கு
Package	- பொதி	Merge	- ஒருங்கு
Data	- விபரம்	Colour	- வண்ணம்
Information	- தகவல்	Hue	- நிறச்சாயல்
Communication	- தொடர்பு, செய்தி	Palette	- கிண்ணி
Message	- கருத்து, சேதி	Pixel	- படக்கூறு
Maths	- கணிதம்	Point	- புள்ளி, முனை
Mathematics	- கணிதவியல்	Element	- கூறு
Arithmetic	- கணக்கு	Picture	- படம்
Algorithm	- நெறிமுறை	Chart	- வரைபடம்
Computation	- கணிப்பு	Dot Matrix	- இம்மி அணிக்கோவை
Record	- ஏடு		
Item	- உறுப்பு	Graphics	- சித்திரம்
Menu	- உறுப்படி	Animation	- அசை ஓவியம்
Access	- அணுகு	Visual	- கட்புலன்
Retrieve	- மீட்டல்	Textual	- உரைசார்
Free	- கட்டுறா	Tactile	- உணர்வுசார்
Format	- உருவளவு	Audio	- செவிசார்
Unit	- அலகு	Video	- விழிசார்
Pack	- அடுக்கு, திணி	Mark	- மையிடு
Fault	- பிறழ்வு	High light	- மிக்கொளிர்
Garbage	- கூளம்	Scanner	- வருடி
Query	- வினா	Display	- காட்டி
Packet	- சிப்பம்	Cursor	- இடஞ்சுட்டி
Compression	- சுருக்கம், குறுக்கம்	Pointer	- சுட்டி
		Prompt	- உசுப்பி
Reduction	- குறைப்பு	Clear	- துலக்கு
Satelite	- செய்மதி	Window	- பலகனி
Network	- வலைப்பின்னல்	Medium	- ஊடகம்

கணிப்பொறி - ஓர் மொழியாய்வுக் கருவி

(COMPUTER AS A LANGUAGE RESEARCH TOOL)

ச. பாஸ்கரன்
வழியமைப்பாளர், கணிப்பு மையம்,
தமிழ்ப் பல்கலைக்கழகம், தஞ்சாவூர்

1. முன்னுரை

சுமார் நாற்பது ஆண்டுகளுக்கு முன்னர் மனிதனை விட பன்மடங்கு வேகத்திறன் கொண்ட ஒரு நல்ல "கணக்கீட்டுக் கருவியாக" மட்டும் விளங்கிய கணிப்பொறி தற்பொழுது நம் வாழ்வின் ஒவ்வொரு நிலையிலும் ஏதோ ஒரு வகையில் பயன்படும் "பல்-பயன்பாட்டு கருவியாக" வளர்ச்சிப் பெற்றுள்ளது. உயர் அறிவியல் ஆய்வுப்பணிகளுக்கு விலை மதிப்புப் பெற்ற தொழில் நுட்ப மிக்க ஆய்வுக்கூடக் கருவியாக மட்டுமே பயன்படுத்தப்பட்டு வந்த கணிப்பொறி தற்பொழுது எவரும் எத்துறைக்கும் பயன்படுத்தக்கூடிய கருவியாக வடிவமைக்கப்படுகிறது.

இந்நிலையில், மொழியாய்வில் கணிப்பொறியின் பயன்பாடு, மொழியாய்விற்கு ஏற்ற வகையில் கணிப்பொறியின் தொழில் நுட்ப வசதி. வடிவமைப்பு ஆகியவைகள் பற்றி தொகுத்தாய்வதே இக்கட்டுரையின் நோக்கமாகும்.

2. மொழி ஆய்வில் கணிப்பொறி

அறிவியல் துறை ஆய்வுப் பணிகளுக்குக் கணிப்பொறியைப் பயன்படுத்துவது போன்று மொழியாய்வுப் பணிகளுக்கும் கணிப்பொறியைப் பயன்படுத்தலாம். 1950-ஆம் ஆண்டு முதற்கொண்டே மேலை நாடுகளில் கணிப்பொறிகள் மொழியாய்விக்கு பயன்படுத்தப்பட்டு வந்துள்ளன.

உள்ளக் கருத்தை வெளியிடுவதற்கு எண்ணப் பரிமாற்றங்களுக்கு மொழி ஒரு கருவியாக பயன்படுகிறது. பேச்சு நடை இன்னொருவரிடமிருந்து வேறுபடுகிறது. இவ்வாறே ஒவ்வொருவருடைய எழுத்து நடையும் தனித்துவம் பெறுகின்றன. எனவே, "மொழி நடையியலின் ஆய்வு" இலக்கியத் திறனாய்விற்குப் பெரிதும் துணை புரிகிறது. சொற்களையும் இலக்கண கூறுகளையும் சொற்றொடர் அமைப்புக்களையும் கணக்கிட்டு ஒப்பிடுவதன் மூலம் ஒரு நூலின் அல்லது படைப்பின் படைப்பாளியை இனம் கண்டு கொள்ளலாம் என ஆய்ந்தறியப்பட்டுள்ளன.

IBM-704 என்ற கணிப்பொறி மூலம் மாத்யூஸ் அர்னால்ட் (MATHEW'S ARNOLD) என்ற ஆங்கில கவியின் செய்யுள்களுக்கு பேராசிரியர் ஸ்டீபன் பாரிஸ் (Prof. Stephen Parish) என்பவரால் சொல்லடைவு தயாரிக்கப் பட்டது. கணிப்பொறி மூலம் முதன்முதலாக சொல்லடைவு தயாரிக்கப்பட்டது இதுவேயாகும். இதற்குப் பின்னர் ஆக்சுபோர்டு அகராதி (Oxford Dictionary) தயாரிப்பதற்கு உதவியாக சொல்லடைவு கணிப்பொறி வழி உருவாக்கப்பட்டது.

உரைநடைச் செய்திகளையோ, செய்யுள் வடிவில் அமைந்த செய்திகளையோ கணிப்பொறிக்கு உள்ளீடு செய்து உரிய வழியமைப்புகளின் மூலம் சொல்லடைவுகள் ஆக்கப்படுகின்றன. இச்சொல்லடைவுகளைக் கொண்டு இலக்கியச்சொற்கள், காலத்திற்கேற்ற சொற்கள், பாத்திரங்களுக்கேற்ற சொற்கள் என வகைப்படுத்தி சொற்கள் ஆராயப்படுகின்றன. மேலும், சொற்களின் தேர்வு, பயன்படுத்தும் முறை, புதிய சொற்களை ஆக்கும் திறன், சொற்களுக்குப்புதிய பொருள் அளித்தல் ஆகியவைகளின் மூலம் ஒரு படைப்பாளி அடுத்தவரிடமிருந்து வேறுபடுவது ஆராயப்படுகிறது.

பெயரெச்சங்கள், வினையெச்சங்கள், உம்மை உருபு, தொடர் இணைப்பான்கள் முதலியன சொற்றொடர்களை (Phrases) விரிப்பதற்கு (Expansion) எவ்வாறு பயன்படுகின்றன என்பது ஆராயப்படுகிறது. சொற்றோடரே மொழியின் கட்டமைப்பைப் புலப்படுத்துவதற்கு அடிப்படை எல்லையாக அமைவதை அறிவியல் அடிப்படையில் நோக்கப்படுகிறது. இவ்வாறு, மொழிக் கட்டமைப்பை கணிப்பொறி வழி ஆய்வது இரு வேறு மொழிகளின் கட்டமைப்புகள் ஒன்றோடொன்று எவ்வாறு ஒத்துள்ளன, உறவு கொண்டுள்ளன, வேறுபடுகின்றன போன்றவைகளை எளிதில், விரைவாக, அறிவியல் கண்ணோட்டத்தில் கண்டறியப் பயன்படுகிறது. இத்தகைய ஆய்வுப் பணிகள் கணிப்பொறி வழி மொழிபெயர்ப்பிற்குப் பெரிதும் துணை செய்கின்றன.

மேலும் ஒரு மொழியின் பயன்பாடு காலத்திற்கு காலம் எவ்வாறு வேறுபடுகின்றது, வளர்ச்சியடைகின்றது, அறிவியல் முன்னேற்றங்கள் ஒரு மொழியின் வளத்தை எவ்வாறு தாக்குகிறது என்பன ஆராயப்படுகின்றன. மேலும், மொழிக் கருவி ஆளுக்கு ஆள் வேறுபடுகிறது. எனவே, மொழி நடையினை ஆய்வதின் மூலம் ஒரு படைப்பின் கால நிர்ணயம் (Dating) படைப்பிற்குரிய ஆசிரியரின் பெயர் அறிதல் (Authorship Identification) ஆகியவற்றை கணிதவியல் (Mathematics) புள்ளியியல் (Statistics) கோட்பாடுகளின் மூலம் அறிவியல் கண்ணோட்டத்தில் கணிப்பொறி மூலம் கண்டறியலாம். எனவே, கணிக்கும் திறன் உள்ள கணிப்பொறி மூலம் இத்தகைய ஆய்வுப் பணிகள் மேலை நாடுகளிலும் நம் நாட்டிலும் மேற்கொள்ளப்பட்டுள்ளன.

மொழிசார்ந்த எந்த ஆய்வுப் பணிகளுக்கும் கணிப்பொறியை பயன்படுத்தலாம் என்றாலும், இதுவரை கணிப்பொறி மூலம் வெற்றிகரமாக செய்து முடிக்கப்பட்ட ஒரு சில மொழியாய்வுப் பணிகள் கீழே கொடுக்கப்பட்டுள்ளன.

1. கணிப்பொறி வழிச் சொல்லடைவு (Concordance)
2. மொழிநடை ஆய்வு (Stylistics study)
3. ஆசிரியர் பெயர் தெரியாத படைப்புகளுக்கு மொழிநடையை ஒப்புநோக்கி ஆசிரியர் பெயர் கண்டறிதல் (authorship identification)
4. மொழி இலக்கணம் (Syntax) ஆய்தல்
5. பேச்சு மொழிக்கும், எழுத்து மொழிக்கும் (இலக்கிய மொழி) உள்ள வேறுபாட்டினை ஆய்ந்தறிதல் (studies in Oral and written style)
6. இலக்கண அமைப்பு நிலையில் இயந்திர மொழிபெயர்ப்பு (Machine translation on syntactic level)
7. கணிப்பொறி வழி தொலைக்காட்சிச் செய்தி (Television News) நடை ஆய்வு அறிதல், அதன் மூலம், செய்திகளை எளிதில் புரிந்துகொள்ளும் வண்ணம் மாற்றியமைத்தல். இத்தகைய ஆய்வுகள் நிறையச் செய்திகளைச் சுருக்கமாகவும் தெளிவாகவும் படித்தவர், பாமரர் அனைவரும் எளிதில் புரிந்துகொள்வதற்கு ஏற்பத் தயாரிக்கச் செய்தி ஆசிரியர்களுக்கு உதவுகிறது.
8. தொலைக்காட்சிச் செய்திப் போன்று செய்தித்தாள் மொழிநடை ஆய்வு
9. படைப்புகளுக்கு கால நிர்ணயம் செய்தல் (Dating)
10. கணிப்பொறி வழி மொழி கற்றல், கற்பித்தல்

2.1. தமிழ்ப் பல்கலைக்கழகத்தில் கணிப்பொறி மொழியாய்வு

அறிவியல் அணுகுமுறையில் (Scientific View) தமிழ் மொழியைச் செம்மைப்படுத்தும் நோக்கத்துடன் தமிழ்ப் பல்கலைக்கழகக் கணிப்பு மையத்தில் கீழ்க்கண்ட ஆய்வுப்பணிகள் மேற்கொள்ளப்பட்டன.

1. இயந்திர மொழிபெயர்ப்பு: உருசியன் - தமிழ்
(Machine Translation System: Russian - Tamil)
2. கணிப்பொறி வழி கலைச்சொல் வங்கி
(Computerised Scientific Data Bank)

3. தமிழ்ச் செய்தித்தாள்களின் உரைநடை ஆய்வு
(Newspaper Text Analysis)

4. கணிப்பொறி வழி நூலக ஆய்வுப்பணி
(Tamil/SDI/ Tamil/Marc)

5. கணிப்பொறி வழி அனைத்துலகத் தமிழ் ஓலைச்சுவடிகளின் பட்டியல் தயாரித்தல்
(Prepration of International Catalogue of Tamil palmleaf Manuscripts)

6. கணிப்பொறி வழி திருக்குறள் ஆய்வு (Computer Analysis of Tirukkural)

7. கணிப்பொறி வழி மொழிகளுக்கான ஒருங்கிணைந்த வரிவடிவ மாற்று அமைப்பு (Integrated Script Conversion System)

8. தமிழ்ச் சொல்லாய்வி (Tamil Word Processor)

9. கணிப்பொறி வழி சங்க இலக்கிய அகராதித் திட்டம்
(Computerised Sangam Literature Dictionary Making)

இத்திட்டப் பணிகளில் சிலவற்றை சுருக்கமாகக் கீழ்க் காண்போம்.

கணிப்பொறி வழி மொழிபெயர்ப்புப் பணியை கணிப்பு மையம் தலையாயப் பணியாக மேற்கொண்டுள்ளது. உருசிய மொழியிலிருந்து தமிழில் மொழிபெயர்க்கும் முயற்சியில் ஓரளவு வெற்றி பெற்றுள்ளது. இத்திட்டத்தில் இதுவரை கண்ட முன்னேற்றத்தை தமிழ்ப் பல்கலைக்கழக கணிப்பொறி மொழி பெயர்ப்பு : உருசியன்-தமிழ் என்ற ஆய்வு நூலாக வெளியிட்டுள்ளது.

எந்த மொழிச் செய்திகளையும் கணிப்பொறி மூலம் ஆய்வு செய்யலாமென்றாலும் செய்திகளைக் கணிப்பொறியில் உள்ளீடு செய்வதற்கும் வெளியீடு செய்வதற்கும் ஆங்கில எழுத்துக்களை மட்டுமே கொண்ட கணிப்பொறி இடுவரல்,விடுவரல் (Input/output Devices)கருவிகள் இருந்தன. எனவே, மொழிச் செய்திகளை உள்ளீடு செய்ய ஒலிப்பெயர்ப்பு (Transliteration) செய்தல் அவசியமாகிறது. உள்ளீடு செய்பவர் செய்திகளை ஒலிபெயர்ப்பு செய்து உள்ளீடு செய்ய வேண்டும். தமிழ் மொழி ஆய்வதற்கு கணிப்பொறியைப் பயன்படுத்த இயலும் என்பதை நிறுவுவதற்கேற்ப சோதனைச் செயலாக "கணிப்பொறி வழி திருக்குறள் ஆய்வு" திட்டம் கணிப்பு மையத்தில் மேற்கொள்ளப்பட்டது. இந்த ஆய்விற்காக 1330 குறட்பாக்களும் ஒலிமாற்றம் செய்யப்பெற்று கணிப்பொறிக்கு ஊட்டப்பட்டன. குறள் வெண்பாக்களுக்குரிய இலக்கண விதிமுறைகள் வழியமைப்பு மூலம் கணிப்பொறிக்கு ஊட்டப்பட்டது. இந்த வழியமைப்பைக் கொண்டு கீழ்க்கண்ட முடிவுகள் அறியப்பட்டன.

அ. தேமா, புளிமா, கருவிளம், கூவிளம், தேமாங்காய், புளிமாங்காய், கருவிளங்காய், கூவிளங்காய் ஆகிய இலக்கண அமைப்பில் அமைந்துள்ள வார்த்தைகளின் எண்ணிக்கை.

ஆ. மேற்கண்ட இலக்கண அமைப்புகளில் தொடங்கும் குறட்பாக்களின் எண்ணிக்கை குறட்பாக்களின் பட்டியல்

இ. நாள், மலர், காசு, பிறப்பு இவைகளில் முடியும் குறட்பாக்கள், அவைகளின் பட்டியல்

ஈ. இயற்சீர், வெண்சீர், இயற்சீர் மற்றும் வெண்சீர் அடிப்படையில் அமைந்துள்ள குறட்பாக்களின் எண்ணிக்கைப் பட்டியல்.

உ. அளபெடைகளின் புழக்கம்

ஊ. எழுத்துக்கள் புழக்கத்தின் நிகழ்வெண்ணிக்கை

எ. சொல் புழக்கத்தின் நிகழ்வெண்ணிக்கை

ஏ. சொல்லடைவு

ஐ. இடம் பெற்ற மொத்தச் சொற்கள், வெவ்வேறான சொற்கள் ஆகியச் செய்திகள், இந்த ஆய்வு முடிவுகள் Computer Analysis of Tirukkural என்ற தலைப்பில் ஆங்கிலத்தில் எழுதப்பட்ட நூலாக வெளிவந்துள்ளது.

இப்பணியினை மேற்கொண்டபொழுது உள்ளீடு, வெளியீடு செய்வதற்கு ஆங்கில எழுத்துக்களும், ஒலிபெயர்ப்பும் தேவைப்பட்டதால் பல நடைமுறைச் சிக்கல்களைச் சந்திக்க நேர்ந்தது. எனவே, தமிழ்ச் செய்திகளைத் தமிழிலேயே உள்ளீடு,வெளியீடு செய்யும் வழிகளைக் கணிப்பொறியில் உருவாக்க முயற்சிக்கப்பட்டது. இதன் விளைவாக டாட்மேட்ரிக்ஸ் (Dot Matrix) அச்சுப் பொறிகளில் படங்கள் வரையக்கூடிய (Graphics) வசதிகளைக் கொண்டு தமிழ் எழுத்துகள் வடிவமைக்கப்பட்டன. இதன் விளைவாக வெளியீடுகளை தமிழ் எழுத்துக்களில் அச்சிடும் வசதி உருவாக்கப்பட்டது. சுமார் 16 வெவ்வேறான அளவுகளில் தமிழ் எழுத்துக்களை அச்சிடும் வசதி செய்யப்பட்டது. இப்பணியை அடிப்படையாகக் கொண்டு கணிப்பொறி மூலம் இந்திய மொழிகளுக்கான ஒருங்கிணைந்த (Indian Script Conversion System) வரிவடிவ மாற்ற அமைப்பு என்ற பணி மேற்கொள்ளப்பட்டது. இந்த அமைப்பின் மூலம் மலையாளம், வங்காளம், தெலுங்கு, கன்னடம், மராட்டி, இந்தி ஆகிய மொழிகளில் எழுத்துக்கள் அச்சிட வழியமைப்புகள் உருவாக்கப்பட்டன. இந்த அமைப்பின் மூலம் மேலே குறிப்பிடப்பட்டுள்ள இந்திய மொழி எழுத்துக்களில் செய்திகளை அச்சிட இயலும். எனினும், உள்ளீடு செய்வதற்கு ஆங்கில எழுத்துக்களும், ஒலிபெயர்ப்பு முறையுமே ப்யன்படுத்தப்பட்டு வந்தது. இச்சிக்கலைத் தவிர்க்கும் பொருட்டு தமிழ்ச் செய்திகளைத் தமிழ் எழுத்துக்களிலேயே உள்ளீடு செய்ய தமிழ்

தட்டச்சு அமைப்பு போன்ற தட்டச்சுப் பலகையொன்றையும் (Key Board) காட்சித்திரையில் (Display Unit) தமிழ் எழுத்துக்களை உருவாக்கும் அமைப்பும் வடிவமைக்கப்பட்டது. இவ்வாறு தமிழ்ச் செய்திகளைத் தமிழ் எழுத்துக்களிலேயே உள்ளீடு செய்யவும், ஆய்வு செய்யவும், அச்சிட்டு வெளியிட செய்யவும் வசதிகளைக் கொண்டமைந்த தமிழ்ச் சொல்லாய்வி (Tamil Word Processor) என்ற வழியமைப்பும் உருவாக்கப்பட்டது.

கணிப்பொறி வழி திருக்குறள் ஆய்வு, இந்திய மொழிகளுக்கான வரிவடிவ மாற்றம் மேலும் தமிழ்ச் சொல்லாய்வி ஆகியப் பணிகளின் அடிப்படைத் தத்துவத்தைக் கொண்டும், வசதியைக் கொண்டும், 'கணிப்பொறி வழி சங்க இலக்கியச் சொல்லடைவு மற்றும் அகராதித் திட்டம்' மேற் கொள்ளப்பட்டுள்ளது.

3. இயல்பு மொழியும் கணிப்பொறியும்

கணிப்பொறிக்குக் கட்டளையிடுவதற்குக் கணிப்பொறி மொழிகள் (Computer Languages) உள்ளன. கணிப்பொறி மொழி என்பது கணிப்பொறியும் மனிதனும் புரிந்துக் கொள்ளக்கூடிய மிகக் குறைந்த எண்ணிக்கையுள்ள சிறப்பு மிக்க வார்த்தைகள், எண்கள், ,சிறப்புக் குறியீடுகள், குறிப்பிட்ட இலக்கண அமைப்புகள் ஆகியவற்றைக் கொண்டமைந்துள்ளது. கணிப்பொறி மொழிகளை அவைகளின் பயன்பாட்டிற்கேற்ப வணிகத் துறைகளுக்குத் தேவையான (Commercial Application) வழியமைப்புகளை எழுத உதவும் மொழிகள், அறிவியல் ஆய்விற்குப் பயன்படுத்துவதற்கேற்ற மொழிகள், மொழியாய்விற்குப் பயன்படும் மொழிகள், பல துறைகளுக்கும் பயன்படுத்துவதற்கேற்ற பொதுவான மொழிகள் எனப் பா கு படுத்தப்படுகின்றன. சுமார் 120 கணிப்பொறி மொழிகள் உள்ளன. எனவே, கணிப்பொறியை எந்த துறைக்குப் பயன்படுத்த வேண்டுமோ அந்த துறைக்கான வழியமைப்புக்களை உருவாக்க வசதியுள்ள கணிப்பொறி மொழியினைத் தெரிந்திருப்பது அவசியமாகிறது. வழியமைப்புக்களைக் கொண்டு ஆய்வு செய்ய வேண்டிய செய்திகளை உள்ளீடு செய்யவும் (Inputing), ஆய்வு செய்யவும் (Processing), வெளியீடு செய்யவும் (Outputing) இயலும். கணிப்பொறியில் சேமித்து வைத்துள்ள செய்திகளிலிருந்து (Data) ஏதாவது தகவல் பெற வேண்டுமெனில் அதற்குரிய கணிப்பொறி கட்டளைகளைப் பயன்பாட்டாளர்(User) தெரிந்திருத்தல் வேண்டும். இவ்வாறு கணிப்பொறி மொழி: ஒரு குறிப்பிட்ட தேவைக்காக, சூழலுக்காக, செயலுக்காக ஒரு பொறியை இயக்கப் பயன்படுத்தப்படும் செயற்கை வடிவில் அமைந்த மொழியாக அமைந்துள்ளது. எனவே. கணிப்பொறியைக் கொண்டு ஆய்வினை மேற்கொள்ள முற்படுபவர்கள் கணிப்பொறி தொழில் நுட்ப வல்லுநர்களிடம், பணியாளர்களிடம் எதை செய்யவேண்டுமோ அதைத் தெளிவாகச் சொல்ல

வேண்டும். கணிப்பொறி துணையின்றி இப்பணிகளை எவ்வாறு நடைமுறையில் செய்கிறார்கள் என்ற செயற்பாட்டு தத்துவத்தை தெளிவாக்க வேண்டும். இது சிக்கலற்றது. ஆனால் எளிமையானதல்ல, மேலும் கணிப்பொறியை எந்த துறையின் பணிக்குக் கையாளப்படவேண்டுமோ, அத்துறையில் கையாளப்படும் மிகச்சிறிய துறைச் சொற்களைக்கூட தெளிவாக்க வேண்டும். எடுத்துக்காட்டாக, ஒரு மொழியியல் வல்லுநர் அவர் துறையில் ஏதாவது ஒரு பணிக்கு கணிப்பொறியைப் பயன்படுத்திக் கொள்ள விரும்பினால், அப்பணிக்குத் தேவையான மொழியியல் துறைச் சொற்களை அப்பணியைக் க்ணிப்பொறி வழி மேற்கொள்ளும் தொழில்நுட்ப வல்லுநரிடம் தெளிவாக்கவேண்டும். இதனால் தொழில் நுட்ப அறிவுமிக்க கணிப்பொறி வல்லுநரிடம் தெளிவாக்கவேண்டும். இதனால் தொழில் நுட்ப அறிவுமிக்க கணிப்பொறி வல்லுநர்களின் நேரங்கள், அவர்கள் துறைக்குச் சற்றும் பொருந்தாத ஏனைய துறைகளைப் பற்றிய செய்திகளைச் சேகரிப்பதில் விரயமாகிறது. இச்சிக்கலுக்கு காரணம் மக்கள் பேசும் இயல்பான (Natural) மொழியிலிருந்து விலகி செயற்கையாய் கையாளப்படும், ஒரு சிறப்பு மிக்க மொழியாய் கணிப்பொறி மொழி அமைவதே ஆகும். கணிப்பொறியுடன் செய்தித் தொடர்பு கொள்ள கணிப்பொறி மொழி தெரிந்திருத்தல் அவசியமாகிறது. அதே சமயத்தில் ஏனைய துறை வல்லுநர்கள் கணிப்பொறி மொழி பயின்று கணிப்பொறியை திறம்படப் பயன்படுத்த முயல்வதும் தேவையற்றது

தற்பொழுது கணிப்பொறி என்பது இத்துறை வல்லுநர்களால் தொழில் நுட்பப் பணியாளர்களால் மட்டுமே இயக்கக்கூடிய ஓர் உயர் நுட்பக் கருவி என்ற நிலையிலிருந்து யாவரும் பயன்படுத்துவதற்கேற்ப உயர் தொழில் நுட்பத்துடன் கூடிய; மிக எளிதாக இயக்கக் கூடிய ஒரு கருவியாக வடிவமைக்கப்படுகிறது. கணிப்பொறியில் செய்திகளை உள்ளீடு, வெளியீடு செய்வதற்குப் பயன்படும் இடுவரல்,.விடுவரல் கருவிகள் (I/O Devices) அவைகளை இயக்க எவ்வித சிறப்பு அறிவியல் தொழில்நுட்ப பயிற்சி தேவையில்லை என்ற ரீதியில் வடிவமைக்கப்படுகின்றன. தட்டச்சுப் பலகையில் தட்டச்சு செய்யத் தெரிந்தவர்கள் கூட கணிப்பொறியுடன் தொடர்பு கொள்ள இயலும். மேலும், நமது வாய்மொழி கட்டளைகளை ஏற்றுச் (Voice Recognition) செயற்படும் கருவிகளும் வடிவமைக்கப்படுகின்றன. இருப்பினும் 'கணிப்பொறி மொழி' என்னும் "ஒரு மொழிக்கருவி" கணிப்பொறிக்கும்; பயன்பாட்டாளருக்கும் ஒரு தடையாய் விளங்குகிறது. இக்குறைப்பாட்டினை நீக்கும் வண்ணம் கணிப்பொறிக்கும், பயன்பாட்டாளருக்கும் இடையே நிகழ்கின்ற செய்தித் தொடர்பு மக்கள் பேசும் எழுதும் இயல்பான மொழியில் இருக்க வேண்டுமென உரைப்பட்டு ஆய்வுகள் மேற்கொள்ளப்பட்டுள்ளன. இவ்வாய்வுகளில் கறிப்பிடத்தக்க முன்னேற்றமும் வெற்றியும் கிட்டியுள்ளன.

இவ்வாறு, கணிப்பொறியைப் பயன்படுத்த கணிப்பொறி மொழி அறிவு பயன்பாட்டாளருக்குத் தேவை என்ற நிலையிலிருந்து; கணிப்பொறிக்கு மக்கள் பேசும் இயற்கை மொழி அறிவு அவசியம் என்ற நிலைக்கு வரப்பட்டுள்ளது. இதற்கு இயற்கை மொழி அறிவையும் மொழி நிபுணத்தத்துவையும் கணிப்பொறிக்கு உட்பட வேண்டும். இதற்கு ஒரு மொழியின் இசை அமைப்பு (Rhythm), எழுத்து, சொல் உச்சரிக்கும் கலை (Phonology), (Morphology) சொற்றொடர் அமைப்பு, வாக்கிய அமைப்பிற்கான இலக்கணம், சொல், சொற்றொடர்களுக்கான பொருள், கருத்தோட்டம் போன்ற செய்திகள் கணிப்பொறிக்கு ஊட்ட வேண்டும். இப்படிச் செய்வதின் மூலம் இயற்கை மொழி உணர்தல் (Natural Language) (Understanding), இயற்கை மொழியில் விடையளித்தல் (Natural Language Generation) ஆகிய வசதிகளைக் கணிப்பொறியில் உருவாக்கலாம். இத்தகையப் பணிகள் கணிப்பொறியில் மொழி ஆய்வு (Natural Language Processing) என்ற திட்டத்தின் அடிப்படையில் ஆய்வுப் பணிகளாக மேற்கொள்ளப்பட்டுள்ளன.

4. முடிவுரை

மனித ஆற்றலுக்கு முற்றிலும் நிகரான கணிப்பொறியை உருவாக்க ஆய்வுகள் மேற்கொள்ளப்படுகின்றன. கணிப்பொறியை ஒரு மருத்துவராக, பொறியியலாளராக, ஆசிரியராக, மொழியியல் ஆசிரியராக ஆக்கும் முயற்சிகளில் ஓரளவு வெற்றி கிடைத்துள்ளது. இவ்வாறு மனித சிறப்பு அம்சங்களான செய்திகளைத் திரட்டுவது, சேமித்துள்ள செய்திகளைக் கொண்டு ஓர் சிக்கலுக்கு சிந்தித்துத் தீர்மானம் காண்பது எதிர்கால தீர்மானங்களுக்குத் திட்டமிடுவது, பல செய்திகளை ஒருமைப்படுத்துவது, ஒருங்கமைப்பது, சோதிப்பது, துறை அறிவில் பக்குவப்படுவது போன்ற பண்புகள் கணிப்பொறிக்குக் கற்பிக்கப்படுகின்றன. மேலும் அன்றாட வாழ்வில் மனிதனுக்கு ஏற்படுகின்ற அனுபவ அறிவு ஆற்றலையும், கணிப்பொறிக்கு அளிப்பதில் முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. இம்முயற்சிகளின் பலனாக தொழில் நுட்பமிக்க பல வசதிகளைக் கொண்டு வடிவமைக்கப்பட்டுள்ள கணிப்பொறிகள் மொழியாய்விற்கு பெரும் அளவிற்கு பயன்படும் என்பதில் ஒரு சிறிதும் ஐயமில்லை.

இவ்வாறு கணிப்பொறிகள் தொழில் நுட்ப ரீதியில் வளர்ச்சி பெற்றாலும் யாவரும் அவைகளுடன் தொடர்பு கொள்ளும் வண்ணம் மொழி அறிவைக் கணிப்பொறிக்கு ஊட்டுவது அவசியமாகிறது. இதற்காக பல நிறுவனங்கள் கணிப்பொறி வழியமைப்புக்களை உருவாக்குகின்றன. இத்தகைய பணிகள் கல்வி நிறுவனங்களில் ஆய்வுக்கூடங்களில் ஆராய்ச்சி நிலையிலேயே மேற்கொள்ளப்பட்டு வருகின்றன. இவ்வாறான ஆராய்ச்சி முடிவுகள் பொருளாதார ரீதியில் வெற்றி பெறுமென வியாபார நோக்கத்தில் செயல்படும் பல தனியார் நிறுவனங்கள் கருதும்போது பெரும் வெற்றியை எளிதில் பெறலாம்.

மொழி ஆய்வில் கணிப்பான்

க. சாருகேசி
பேராசிரியர் மற்றும் தலைவர்

பெ.தம்பிதுரை
இணை பேராசிரியர்

பாரதியார் பல்கலைக்கழகம்
கோயம்புத்தூர் - 641 046

1. முன்னுரை

கணிப்பான் என்பது இன்றைய அளவில் பல துறைகளில் பயன்படுத்தப்பட்டு வருகிறது. இயற்கை மொழியையும் கணிப்பானின் உதவியால் ஆராய முடியும் என்பதையும், மொழி பெயர்ப்புக்கு கணிப்பானை பயன்படுத்த முடியும் என்ற கொள்கையின் அடிப்படையில் பல மொழியைச் சார்ந்த கணிப்பான் குறித்த ஆய்வுகள் நடந்து வருகின்றன. இவைகளில் ஒரு சிலவற்றை குறிப்பாக உணர்த்த இந்த கட்டுரை முயல்கிறது.

2. செயற்கை மதிநுட்பம்

மனிதர்கள் இயற்கையிலேயே சூழ்நிலைக்குத் தகுந்தவாறு தனது அனுபவத்தின் வாயிலாக பெற்ற அறிவை பயன்படுத்துகின்ற தன்மையுடையவர்களாதலால், அவர்களுக்கு இயற்கை மதிநுட்பம் (Natural Intelligence) உள்ளது எனக் கருதலாம். கணிப்பான்களையும் மதிநுட்பம் உடைய ஒரு பொறியாக ஆக்கமுடியுமா? அவைகளை சூழ்நிலைக்கு ஏற்றவாறு தங்கள் அறிவை பயன்படுத்தி செய்திகளை படைத்து, மாற்றி வெளியிடும் வகையில் உருவாக்க முடியுமா? என்ற கேள்விகளின் அடிப்படையில் தான் கணிப்பானைச் சார்ந்த செயற்கை மதிநுட்ப (Artificial Intelligence) ஆய்வுகள் இன்று பெருமளவில் உலகில் நடந்து வருகின்றன. இதனோடு தொடர்பாக கணிப்பானுடன் கருத்துப் பரிமாற்றத்தை இயற்கை மொழியை பயன்படுத்தியே நிகழ்த்த முடியுமா? என்ற கேள்வியும் எழுகிறது. ஆகவே கணிப்பானை மனிதர்களைப்போல் உணர்தல், இணைத்தல், பல செயல்களை செய்தல், மொழியை பயன்படுத்தல், செய்திகளைச் சேமித்து வைத்தல் போன்றவற்றை இயற்கையிலேயே செய்யும் வகையில் கணிப்பானுக்கு ஏற்ற அறிவை உருவாக்குவது என்பதுதான் செயற்கை மதிநுட்ப ஆய்வுகளின் அடிப்படை குறிக்கோளாகும்.

செயற்கை மதிநுட்ப ஆய்வு எழுவதற்கு முன்னோடியாக அமைந்த மொழியியல் சார்ந்த ஆய்வுகளில் இயற்கைமொழி ஆய்வு, கணிப்பான் அறிவுக்குத்

தேவையான வடிவமைப்பு, இலக்கண அறிவுக்கான ஆய்வுகள், மொழி அலகுகளின் பொருள் குறித்த ஆய்வு, மொழி அமைப்புக்களை படைத்தல் மற்றும் மொழி அமைப்புக்களை ஆய்ந்தறிதல் போன்ற பல ஆய்வுகள் முன்னோடியாக அமைந்தன.

3. இயற்கை மொழி ஆய்வு

இயற்கை மொழி ஆய்வு என்பது கணிப்பானின் உதவியால் மொழியின் இலக்கண, பொருண்மை அமைப்புக்களை விளக்கும் வகையிலான சில செயல்முறை திட்டங்களை வகுப்பது ஆகும். அதாவது ஒரு மொழியில் காணப்படும் அமைப்புக்களின் பொதுமையான வடிவம், அவ்வடிவத்தில் பங்குபெற தகுந்த அலகுகள், அவ்வலகுகளின் வரிசைப்பாடு போன்றவற்றை கணிப்பானே விளக்கும் வகையில், கணிப்பானை செயல்படுத்தும் முறைகள் இயற்கை மொழி ஆய்வில் முக்கியத்துவம் பெறுகின்றன. மொழியியலார் மொழியை, மொழி மனதில் நிலைப்பெற்றிருக்கும் விதத்தை விளக்குவதற்காக பல இலக்கண வடிவங்களை உருவாக்குகிறார்கள். இவைகளை இலக்கண உருக்கள் (Grammar formalisms) என்பர். இவ்விலக்கண உருக்களை அடிப்படையாகக் கொண்டு மொழியின் அமைப்பை விதிகளின் உதவியாலோ, சூத்திரங்களின் உதவியாலோ விளக்கும் வகையில் கணிப்பானுக்கு சில வரைமுறைகளை வழங்கமுடியும். இத்தகைய கணிப்பானுக்கே உரிய முறையில், கணிப்பானை மொழியமைப்புக்களை விளக்கும் தன்மையடையச் செய்வதற்காக படைக்கப்படும் வரைமுறைகளை கணிப்பான் தொடர்பான இலக்கண உரு (Computer Formalism) என அழைக்கலாம்.

கணிப்பானுக்கேற்ற இலக்கண வடிவங்கள் பல இன்றைய சூழலில் உருவாக்கப்பட்டுள்ளன. TEAM (1986, KING (1987), ABSITY (1988), ELIZA (1966), NETalk (1990) போன்ற இயற்கை மொழி ஆய்வில் உருவாக்கப்பட்ட சில வடிவங்களாகும்.

4. மொழி அறிவு நிலைநாட்டும் வடிவங்கள்

மொழியையும், மொழியின் அமைப்புக்களான வாக்கியம், தொடர் போன்றவற்றையும் அவைகளின் பொருளையும் ஆய்ந்தறிவதற்கு மொழியின் அமைப்புப்பற்றிய அறிவு தேவையாகிறது. கணிப்பானை இயற்கையாகவே மொழியை ஆராய்ந்து அறியும் திறனுடையதாக ஆக்குவதற்கு, அதனுடைய நினைவுப்பகுதியில் மொழிபற்றிய அறிவைப் புகட்ட வேண்டும். இந்த மொழி அறிவை கணிப்பானுக்கு எந்த வடிவத்தில் கொடுக்க வேண்டும் என்பதை நிர்ணயிப்பதின் மூலம் கணிப்பானிடமிருந்து சிறந்த விளக்கங்களை பெற

ஏதுவாகும். இத்தகைய அறிவை விளக்கும் வடிவங்கள் (Knowledge Representation Schemes) பல ஆராய்ச்சியாளர்களால் நிறுவப்பட்டுள்ளன.

இயக்க விளக்கம் (Procedural representation), நிலைபேற்று வடிவம் (Declarative representation), வேற்றுமை சட்டம் (Case frames), கருத்துச் சார்பு வடிவம் (Conceptual dependency), கருத்துச் சட்டம் (frame) போன்ற மொழி அறிவு நிலைநாட்டு வடிவங்கள் வழக்கில் உள்ளன. மொழி அறிவு என்பதில் இலக்கண அறிவு, சொல் அறிவு, பொருண்மைப்பற்றிய அறிவு போன்றவை அடங்குகின்றன. இந்த அறிவு வகைகளுக்கு அமைப்பைக் கொடுக்கும் வண்ணம் மேற்குறிப்பிட்ட அறிவு விளக்க வடிவங்கள் பயன்படுகின்றன. இவ்வடிவங்களை அடிப்படையாகக் கொண்டு கட்டளைத் தொகுப்புக்களை (Programs) உருவாக்குவதன் மூலம் கணிப்பானில் மொழி அறிவை சேமித்துவைக்க முடிவதோடு, தேவைப்படும்போது கருத்துக்களை இந்த வடிவங்களில் கணிப்பானிலிருந்து வெளியீடாக பெறமுடியும்.

5. இயற்கைமொழி அமைப்பு ஆய்வு

இயற்கை மொழி ஆய்வில் முக்கியமாக மொழியின் இலக்கண அமைப்புக்களை ஆய்ந்தறிந்து, அவ்வமைப்புகளுக்கு உரிய வடிவங்களை பெறும் வகையில் கணிப்பான்கள் பயன்படுத்தப்படுகின்றன. இத்தகைய அமைப்பு ஆய்வு என்பதன்மூலம் உரைக்கோவைகளை ஆய்தல் (text processing), உரைக்கோவைகளின் அங்க அமைப்புக்களை உணர்தல் (text comprehension), மொழி பெயர்ப்பு (translation) போன்ற செயல்களை கணிப்பானின் உதவியால் நிகழ்த்த முடியும்.

அமைப்பு ஆய்வு என்பதில் பெரும் அலகுகளைப் பகுத்து, அதன் அங்க அலகுகளைக் காணல் என்ற பகுத்தல் செயலும் (Parsing), அங்க அலகுகளின் உதவியால் பெரும் அலகுகளை உருவாக்குதல் (Generation) போன்ற செயல்கள் அடங்குகின்றன. இந்த இரண்டு செயல்களில் பகுத்தல் என்ற செயல்குறித்த கட்டளைத் தொகுப்புக்கள் வழக்கில் உள்ளன. பகுத்தல் என்ற செயலை கணிப்பான் உதவியால் நிகழ்த்த பல வழிமுறைகள் பயன்பாட்டில் உள்ளன. கொடுக்கப்பட்ட தொடரில் தோன்றும் சொற்களை கணிப்பானின் நினைவுப்பகுதியிலிருக்கும் சொல் பட்டியலில் காணப்படும், சொல்லோடு ஒப்புமைப்படுத்தி சொற்களை பிரித்து வகைப்படுத்தல் (template matching) பெயற்சி சங்கிலி வடிவம் (transition network), மற்றும் கட்டுப்பாட்டிற்குட்பட்ட பெயற்சி சங்கிலி வடிவம் (augmented transition network) போன்ற உபகரணங்களின் உதவியால் கணிப்பானை மொழித்தொடர்களை பகுத்து வெளியிடும் வண்ணம் செய்ய இயலும்.

மொழியின் தொடர்களை கணிப்பானின் உதவியால் படைக்கவும் இயலும். கணிப்பானை படைப்பாற்றல் உள்ள ஒரு பொறியாக மாற்றுவதற்கு அதன் நினைவுப்பகுதியில் சில அடிப்படை பொருண்மை அலகுகளையும், சொற்களையும் சேமிக்க வேண்டும். பின்னர் தொடர்களை உருவாக்குவதற்கு தேவையான விதிகளையும், சூத்திரங்களையும் கட்டளைத் தொகுப்பின் உதவியால் கொடுப்பதன்மூலம் கணிப்பானிலிருந்து தேவையான தொடர்களை உருவாக்கிப் பெறமுடியும். இத்தகைய செயல்பாட்டை அமைப்பு படைத்தல் (generation of structures) என்று அழைக்கலாம். அமைப்பு படைத்தல் என்பதில் வாக்கியங்களை படைத்தல் (Sentence generation), உரைக்கோவை படைத்தல் (text generation), கருத்து படைத்தல் (theme generation) போன்ற செயல்களை நிகழ்த்தமுடியும். பகுப்பாய்வையும், அமைப்பு படைத்தலையும் ஒப்பிடும்போது பகுப்பாய்வு ஒரு எளிய செயலாக தோன்றுகிறது. இதனால்தான் பகுப்பாய்வு, தொடர்பான ஆய்வுகள் பல தோன்றியுள்ளன. படைத்தல் என்பது மிக அரிதான நிலையில் உள்ளது.

6. பொருண்மை ஆய்வு

இயற்கை மொழிகளில் காணப்படுகின்ற அலகுகளான சொற்கள், தொடர்கள், வாக்கியம் என்பன பொருள் கொண்ட அலகுகள். கணிப்பானின் உதவியால் ஒரு தொடரின் அமைப்பையும், அத்தொடர் உணர்த்தும் பொருளின் அமைப்பையும் பெறமுடியும். தொடரின் இலக்கண அமைப்பை ஒருவகை கிளைவடிவம் (tree structure) எனக் கூறலாம். அதைப்போல் தொடரின் பொருள் வடிவத்தையும் ஒரு வகை கிளை அமைப்பே எனக்கொண்டு, அதனை தருக்க சூத்திரங்களின் உதவியால் விளக்கமுடியும். கணிப்பானை, தொடர்களின் தருக்க அமைப்பை வெளியிடும் தன்மையுள்ள ஒரு பொறியாக மாற்றினால், அதன்மூலம் தொடர்களின் பொருள்களையும், தொடர்களின் பொருள் மயக்கத்தையும் (ambiguity) விளக்கலாம். அதாவது ஒரு தொடரில் இருபொருள் கொண்ட சொல் தோன்றினால், கணிப்பானை இரு பொருள்களை அத்தொடருக்கு வழங்கச் செய்ய முடியும்.

உதாரணம்

நான் <u>பள்ளிக்குச்</u> சென்றேன்

அவன் <u>மூட்டையைப்</u> பார்த்தான்

கொடுக்கப்பட்ட ஒரு தொடர் இருவகையான தொடர் அமைப்பைக் கொண்டிருந்தால், கணிப்பானை இரு பொருள்களை அத் தொடருக்கு வழங்கச் செய்ய முடியு ம்.

உதாரணம்

(சென்னை மகளிர்) கல்லூரி

சென்னை (மகளிர் கல்லூரி)

ஆக ஒரு தொடருக்கு இரண்டு வகையான தருக்க அமைப்புக்களை கணிப்பான் வழங்கும் நிலையிலிருந்தால், அத்தொடர்கள் பொருள்மயக்கம் கொண்டவை என கணிக்கப்படும்.

கணிப்பானை தொடர்களின் பொருள் ஒற்றுமையை விளக்கும் தன்மைப் பெற்றதாக ஆக்குவதற்கு உரிய ஆய்வுகளையும் மேற்கொள்ளலாம். தகுந்த கட்டளைத் தொகுப்புகளை உருவாக்குவதன் மூலம் தொடர்களை ஒரே பொருள் கொண்டவை என கணிக்கும் திறனை கணிப்பானுக்கு வழங்க முடியும்.

உதாரணம்

நான் புத்தகத்தைப் படித்தேன்

புத்தகம் என்னால் படிக்கப்பட்டது

7. ஆய்விற்கு அறைகூவல்

பொதுவாக இயற்கை மொழிகள் கடினமான கட்டமைப்புக்களைக் கொண்டிருப்பதாலும், அவைகளில் படிந்து கிடக்கின்ற அடிப்படை நியதிகள் இன்றும் தெளிவாக அறுதியிடப்படாத காரணத்தினாலும் இயற்கை மொழியை கணிப்பான் உதவியால் முழுமையாக ஆராய முடியாது என்ற ஒரு கொள்கை பரவலாக காணப்படுகிறது. மேலும் மனிதர்கள் மொழியைப் பயன்படுத்தும்போது அவர்களது மனம் செயல்படும் விதத்தையும், அவர்களது மனதில் மொழியானது அறிவு வடிவில் நிலைபெறுவதையும் வெளிப்படையான உருக்களால் (formal models) வளக்கினால்தான் மொழியின் செயல்பாட்டை விளக்கும் வடிவத்தை பெறமுடியும். எனவே கணிப்பான் வழி இயல்மொழி ஆய்வு என்பது ஒரு சிம்மசொப்பனமாகவே இருக்கிறது. இருந்தாலும் பல கோணங்களில் ஆய்வுகள் நடந்து வருகின்றன. அதன்விளைவாக பலவகை கணிப்பான் கட்டமைப்புக்கள் வெளியிடப்படுகின்றன. மொழியானது செயல்பாடும், செயல்பாட்டைப்பற்றிய விளக்கமும், கணிப்பான்வழி மொழி ஆய்வும் வளர்ச்சி அடைந்து கொண்டிருக்கின்றன.

8. முடிவுரை

இந்தக் கட்டுரையில் கணிப்பான் வழியிலான மொழியில் ஆய்வுக்குறித்த சில கருத்துக்கள் பொதுவாக விளக்கப்பட்டுள்ளன. பகுப்பாய்வு போன்ற சில துறைகளில் பலவிதமான வளர்ச்சியை காணமுடிகிறது. ஆனால் அறிவு நிலைநாட்டு வழிமுறைகள், மொழியமைப்பு படைத்தல், மனித இயந்திர கருத்துப்பரிமாற்றம் போன்றவைகளில் தீவிரமான ஆராய்ச்சிகள் மேற்கொள்ளப்பட்ட வருகின்றன.

பார்வை நூல்கள்

1. Ralph Grishman, 'Computational Linguistics', Cambridge University Press, 1986.
2. Artificial Intelligence, An International Journal, Vol. Nos. 32,33,34 Netherlands

கணிப்பொறி வழி உரை அல்லது பாடநூல் ஆய்வு ஆய்வறிக்கை

(A Research Report on Computerised Textual Analysis)

ச. பாஸ்கரன்
வழியமைப்பாளர், கணிப்பு மையம்,
தமிழ்ப் பல்கலைக்கழகம், தஞ்சாவூர்

1. முன்னுரை

உரை அல்லது பாடநூல் ஆய்விற்கு 'கிராஃப்போலாஜிக்கல்' பகுப்பாய்வு (Graphological Analysis), அகராதி தொகுக்கும் கலைக்கேற்ற பகுப்பாய்வு (Lexical Analysis), ஒலிக்கட்டு பகுப்பாய்வு (Phonological Analysis), தொடரியல் பகுப்பாய்வு (Syntactic Analysis) மற்றும் சொல்பொருள் பகுப்பாய்வு (Sementic Analysis) தேவைப்படுகின்றன. இவைகள் மொழி ஆய்வுப் பணிகளுக்கு பலவாறு பெரிதும் உதவுகின்றன.

இவைகளின் பயன்களாய் கணிப்பொறிவழிச் சொல்லடைவுகளையும் (concordances), அகராதிகளையும் (Dictionaries) உருவாக்கலாம். மேலும், மொழிநடை ஆய்வு (Stylistics study) ஆசிரியர் பெயர் தெரியாத படைப்புகளுக்கு மொழிநடையை ஒப்புநோக்கி ஆசிரியர் பெயர் கண்டறிதல் (Authorship Identification), மொழி இலக்கணம் (Syntax) ஆய்தல் மற்றும் படைப்புகளுக்கு கால நிர்ணயம் (Dating) செய்தல் போன்ற பணிகளை கணிப்பொறி வழி மேற்கொள்ளலாம். மேலே கூறப்பட்டுள்ள பகுப்பாய்வுகள் இத்தகைய ஆய்வுப் பணிகளுக்கு அடிப்படையாக அமைந்துள்ளன. துணை புரிகின்றன. மேலும் கணிப்பொறிக்கு மொழி அறிவை ஊட்ட உதவுகின்றன.

இந்த பகுப்பாய்வுகளை கணிப்பொறிவழி மேற்கொள்ளும் முயற்சி இக்கட்டுரையில் சுருக்கமாக விளக்கப்பட்டுள்ளது.

2. கணிப்பொறி வழி மொழியாய்வு - முதல் தேவை

கணிப்பொறியில் மொழி ஆய்வுப் பணிகளை கணிப்பொறி வல்லுநர்களும் மொழியியல் வல்லுநர்களும் இணைந்து கூட்டுமுயற்சியாக மேற்கொள்கின்றனர். ஆனால், இவ்விரண்டு துறை வல்லுநர்களின் பார்வையும் (Perspective) இலக்கும் (objective) வெவ்வேறாக உள்ளன. கணிப்பொறி வல்லுநர்களின் குறிக்கோள் தொழில் நுட்ப வளர்ச்சியை மையமாகக் கொண்டது. மொழியியல் வல்லுநர்களின் குறிக்கோள் மொழியியல் வளர்ச்சியை மையமாகக் கொண்டது.

கணிப்பொறி வல்லுநர்கள் மக்கள் பேசும் எழுதும் இயல்பு மொழியை (Natural Language) புரிந்து கொண்டு செயலாற்றும் கணிப்பொறியை உருவாக்க முயல்கின்றனர். இவர்கள் கணிப்பொறிக்கும் மனிதனுக்கும் நடைபெறுகின்ற உரையாடல் (conversation) ஒரு மனிதனுக்கும் மற்றொரு மனிதனுக்கும் நடைபெறுகின்ற உரையாடல் போல அமைகின்ற வகையில் கணிப்பொறி நிரல்களையும் (Programs) அமைவுகளையும் (Systems) உருவாக்க முயல்கின்றனர். எனவே, கருத்துப் பரிமாற்றங்களுக்கு மொழி பயன்படும் முறைகளை ஆராய்ந்து மனிதனின் 'மொழியறிவை' கணிப்பொறிக்கு ஊட்டும் வண்ணம் மொழியாய்வுப் பணிகளை மேற்கொள்கின்றனர்.

கணிப்பொறியில் தகவல் தொகுப்பு (Data base) உருவாக்கவும் பயன்படுத்தவும் பயன்பாட்டாளருக்கும் (User) கணிப்பொறிக்கும் நிகழ்கின்ற உரையாடலை இயல்பு மொழியில் அமைப்பதற்கும்; கணிப்பொறி வழி மொழி பெயர்ப்புப்பணிகளை மேற்கொள்வதற்கும்; கணிப்பொறி வழி மொழி கற்றல்/ கற்பித்தல் அமைப்பை உருவாக்குவதற்கும்; பேசும்/கேட்கும் கணிப்பொறிகளை உருவாக்குவதற்கும் கணிப்பொறி வல்லுநர்கள் முனைந்துள்ளனர்.

கணிப்பொறியின் செயல்வேகம், நினைவக கொள்திறன், பதிவு செய்யப்பெற்ற செய்திகளிலிருந்து தேவையான தகவல்களை நொடிப்பொழுதில் திரும்பப் பெறும் வசதி, தேவையான தகவல்களை தேவையான படிவத்தில் மாற்றி மாற்றி தயாரிப்பதற்கான வசதி ஆகியன மொழியிலாளர்களை (Linguists) கணிப்பொறிவழி ஆய்வுப்பணிகளை மேற்கொள்ளத் தூண்டுகிறது. கவர்கிறது. அவர்களை பொருத்தவரையில் கணிப்பொறி என்பது சொற்களை அகரவரிசைப்படுத்த (Alphabetical Order) உதவும் ஓர் கருவி (Sorting Device) ஆகும்.

மேலும், ஓர் உரையில் இடம்பெறும் வாக்கியங்களின் (Sentence) எண்ணிக்கை சொற்களின் (Words) எண்ணிக்கை, சொற்களின் நிகழ்வெண்ணிக்கை (Frequency), எழுத்துக்களின் (characters) எண்ணிக்கை, எழுத்துக்களின் நிகழ்வெண்ணிக்கை போன்றவற்றை கணிப்பொறி கணக்கிட்டு தருவாதால் மொழியியலாளர்கள் பார்வையில் கணிப்பொறி ஓர் எண்ணும் கருவி (Counting Device) ஆகும்.

ஆனால், இவ்விரண்டு துறை வல்லுநர்களும் இணைந்து செயல்படும்பொழுது தான், ஒவ்வொரு மொழியின் இலக்கண அமைப்பில் (Grammatical system) வாய்ப்பாடு, சூத்திரம் என்கிற வகையில் பொதிந்து கிடக்கும் கணிதப்பண்புகளை (Mathemetical Properties) அடையாளம் காண முடிகிறது. எந்தவொரு மொழியாவுப் பணிகளுக்கும் ஏற்றவகையில் கணிதப்படுத்தப்பட்ட முறையான வழிமுறைகளை திட்ட அமைப்புகளை உருவாக்க உதவுகிறது.

எனவே, கணிப்பொறிவழி மொழியாய்வுப் பணிகளுக்கு கணிப்பொறியாளர்களின் மொழியியாலாளர்களின் கூட்டுமுயற்சி தேவையாகும்.

3. உரையாய்வு (Text Analysis)

பொதுவாக உரையில் (Text) அல்லது பாடத்தில் இடம்பெறும் செய்திகளை (அது உரைநாடையாகவோ அல்லது செய்யுள் வடிவமாகவோ இருக்கலாம்) கீழ்க்கண்ட தலைப்புக்களில் கணிப்பொறிவழி ஆய்வுகளை மேற்கொள்ளலாம்.

1. 'கிரஃப்போலாஜிக்கல்' பகுப்பாய்வு (Graphological Analysis)
2. அகராதி தொகுக்கும் கலைக்கேற்ற பகுப்பாய்வு (Lexical Analysis)
3. ஒலிக்கட்டு பகுப்பாய்வு (Phonological Analysis)
4. தொடரியல் பகுப்பாய்வு (Syntactic Analysis)
5. சொல்பொருள் பகுப்பாய்வு (Semantic Analysis)

3.1 கிரஃப் போலாஜிக்கல் பகுப்பாய்வு (Graphological Analysis)

கணிப்பொறிவழி மேற்கொள்ளப்பெறும் உரையாய்வில் உரையில் இடம்பெறும் வாக்கிய வகைகள் (Sentence Types), வாக்கியத்தின் படித்தலுக்கேற்ற தன்மை (Readability), உரையில் எழுத்துப்புழக்கம் (Freuqncy of Characters), சொல் அட்டவணை (Index) மற்றும் சொல்லடைவு (Concordance) போன்ற தகவல்களும்/ பட்டியல்களும் பெறப்படுகின்றன/உருவாக்கப்படுகின்றன. இவைகள் பற்றி ஒவ்வொன்றாக காண்போம்.

3.1.1. வாக்கிய வகைகள் (Sentence Types)

உரையில் இடம்பெறும் காற்புள்ளி, அரைப்புள்ளி, புள்ளி முதலான நிறுத்தல் குறிகள் (Punctuation Marks) கணக்கிடப்படுகின்றன. இது, உரையிலுள்ள வாக்கிய வகைகளை கணக்கிட உதவுகிறது. சாதாரண வாக்கியங்கள், வினா வாக்கியங்கள் (Interrogative Sentence), கட்டளை வாக்கியங்கள், வியப்பு வாக்கியங்கள் ஆகிய பற்றிய எண்ணிக்கைகள் கணக்கிடப்படுகின்றன. இதன்மூலம் உரையாசிரியர் எந்தெந்த சூழலில் எந்தெந்த வாக்கிய அமைப்புகளை கையாள்கிறார் என்பதைக் கண்டறியலாம்.

3.2.2 வாக்கியத்தின் படித்தலுக்கேற்ற தன்மை (Readability)

வாக்கியத்தின் நீளங்கள் (Sentence Length) கண்டறியப்படுகின்றன. இது இரண்டு வகைகளில் கணக்கிடப்படுகிறது. முதல்முறையில் ஒரு வாக்கியத்தில் இடம்பெறும் சொற்களில் எண்ணிக்கையைக் கொண்டு வாக்கியத்தின் நீளம் கணக்கிடப்படுகிறது. இரண்டாவது முறையில் ஒரு வாக்கியத்தில் இடம்பெறும் எழுத்துக்களின் (Characters) எண்ணிக்கையைக் கொண்டு கணக்கிடப்படுகிறது.

வாக்கியத்தின் நீளத்திற்கும் (Sentence Length) அது தாங்கி நிற்கும் கருத்தை தெளிவாக புரியவைக்கின்ற தன்மைக்கும் (அல்லது படித்தலுக்க உகந்ததாக அமையும் தன்மைக்கும் - Readability) உள்ள தொடர்பு கண்டறியப்படுகிறது. பொதுவாக, ஒரு வாக்கியத்தில் அதிக சொற்கள் இடம்பெற்றிருப்பின் அது படித்தலுக்கு உகந்ததாக அமையாது எனக் கருதப்படுகிறது. இத்தகைய வாக்கியங்களை கணிப்பொறி அடையாளம் காட்டினால் அந்த வாக்கியங்களை உடைத்து சிறுசிறு வாக்கியங்களாக மாற்றியமைக்க உதவும்.

3.2.3 எழுத்துப்புழக்கம் (Characters Frequency)

வாக்கியங்களின் நீளங்களைப் போலவே சொற்களின் நீளங்களும் (Word Lengh) கண்டறியப்படுகின்றன. இதையும் இரண்டு வகையாகக் கணக்கிடலாம். முதல்வகை, சொல்லில் இடம்பெறும் எழுத்துக்களின் எண்ணிக்கையைப் பொருத்தது. உதாரணமாக, 'கணிப்பொறி' என்ற சொல்லில் ஐந்து எழுத்துக்கள் உள்ளன. இரண்டாவது வகை, சொல்லிலுள்ள அசைகளின் (Syllables) எண்ணிக்கையைப் பொருத்தது. 'கணிப்பொறி' என்ற சொல்லில் 'கணிப்', 'பொறி' என்ற இரண்டு அசைகள் உள்ளன.

முதல்முறை, உரையில் இடம்பெறும் எழுத்துக்களின் புழக்கத்தையும் நிகழ்வெண்ணிக்கையையும் கண்டறியப் பயன்படுகிறது. உதாரணமாக, திருக்குறளில் 43 எழுத்துக்கள் பயன்படுத்தப்படவில்லை. 'ல்', 'ற்', 'ர்', 'க', 'ன்' ஆகிய ஐந்து எழுத்துக்கள் அதிகமாக பயன்படுத்தப்பட்டு (1000க்கு மேல்) எழுத்துக்களின் நிகழ்வெண்ணிக்கைப் பட்டியலில் முதல் ஐந்து இடங்களைப் பெறுகின்றன.

ஒரு பொருளை (Object) அல்லது வினையைக் (Verb) குறிக்க ஒன்றிற்கு மேற்பட்ட சொற்கள் இருப்பதாகக் கொள்வோம். அவைகளில் எந்தச் சொல் அதிக புழக்கத்தில் உள்ளது என்பதைக் காணலாம். சொற்களின் நீளத்திற்கும் அதன் புழக்கத்திற்கும் தொடர்பு ஏதாவது இருக்கிறதா என்பதைக் கண்டறியலாம்.

ஒரு சொல்லிலுள்ள அசைவைக் (Syllable) கொண்டு அதன் நீளத்தைக் கண்டறிவது மொழி கற்பவருக்கு உச்சரிப்புக் கலையை கற்பிக்க பயன்படுகிறது. மேலும் இதன் பயன்பாடு உருபனியல் பகுப்பாய்வில் (Morphological Analysis) பலவாகும். சேர்ச் சொல்லோடு முன்ஒட்டு (Prefix), பின்ஒட்டு (Postfix) ஆகியவைகளைச் சேர்த்து சொற்களை கணிப்பொறிவழி உருவாக்கப்பயன்படுகிறது.

3.2.4. சொல் அட்டவணை (Index)

உரையிலுள்ள சொற்களின் பட்டியல் கீழ்க்கண்ட படிவத்தில் உருவாக்கப்படுகின்றன.

சொல் (Word)	நிகழ்வெண்ணிக்கை (Frequency of Occurrence)

இப்பட்டியலில் சொல்லும் அதன் நிகழ்வெண்ணிக்கையும் உள்ளன. ஓர் உரைக்கான சொற்பட்டியலில் சொற்களை மூன்று வகையாக வரிசைப்படுத்தலாம். அவைகளாவன:,

அ) சொற்களை அவைகளில் அகரவரிசையில் (Alphabetical Order) தொகுத்தல்

ஆ) சொற்களை அவைகளின் நிகழ்வெண்ணிக்கையைக் கொண்டு (Frequency of Occurrence) தொகுத்தல்

இ) உரையில் சொற்கள் இடம்பெறும் வரிசைமுறையில் (in the order of Appearence) தொகுத்தல்

ஒவ்வொரு வகை பட்டியலும் அதற்கென உரிய தனிச் சிறப்புகளையும் பயன்களையும் பெற்று விளங்குகின்றன.

உரையில் ஒரு சொல் எத்தனை தடவை வேண்டுமானாலும் இடம் பெற்றிருக்கலாம். ஆனால் உரையிலுள்ள வெவ்வேறான சொற்களின் (Distinct Words) எண்ணிக்கையை கணக்கிடுவதற்கு ஒரு நிக்ழ்வே எடுத்துக் கொள்ளப்படும் உரையிலுள்ள வெவ்வேறான சொற்களின் மொத்த எண்ணிக்கையால் வகுத்தால் கிடைப்பது டிடிஆர் (T.T.R - Type Token Ratio) ஆகும். அதாவது,

$$\text{டிடிஆர்} = \frac{\text{வெவ்வேறு சொற்களின் கூட்டு எண்ணிக்கை}}{\text{சொற்களின் மொத்த எண்ணிக்கை}}$$

டிடிஆர் மதிப்பு பூஜ்ஜியத்திலிருந்து ஒன்றுக்கள் இருக்கும். இந்த மதிப்பு உரையாசிரியரின் சொல்வளத்தைக் (Vocabulary Richness) குறிக்கும்.

உதாரணமாக, திருக்குறளில் மொத்தம் 11597 சொற்களும், 4888 வெவ்வேறான சொற்களும் இடம் பெற்றுள்ளன. எனவே, திருவள்ளுவரின் சொல்வளத்தைக் குறிக்கும் டிடிஆர் மதிப்பு 0.4215 (4888 / 11597) ஆகும்.

இவ்வாறு சொல்பட்டியலால் பயன்கள் நிறைய உள்ளன. மேலும் ஒரு பயனைக் காண்போம்.

எடுத்துக்காட்டாக, சொல்லடைவில் சொற்கள் நிகழ்வெண்ணிக்கையின் அடிப்படையில் இறங்கு வரிசையில் (Descending Order) தொகுக்கப் பட்டிருப்பதாகக் கொள்வோம். இப்பட்டியலிலுள்ள முதல் 50 அதாவது அதிக நிகழ்வெண்களைக் கொண்ட முதல் 50 சொற்களின் பொருள்கள் (meaning) இலக்கணக் கூறு (Grammatical Category), கையாளும் விதம் போன்ற செய்திகளை நன்கு பயின்றோமானால், அந்த உரையின் பொருளை 34 விழுக்காடு புரிந்து கொள்ளலாம். முதல் 100 சொற்கள் தெரியுமானால் 45 விழுக்காடு பொருளை தெரிந்துகொள்ளலாம். முதல் 2750 சொற்கள் தெரியுமானால் 96 விழுக்காடு பொருளை தெரிந்து கொள்ளலாம். மேலும் 1000 சொற்கள் தெரியுமானால் ஏறக்குறைய ஒவ்வொரு வாக்கியத்திலும் ஒரு சொல்தான் படிப்பவருக்குத் தெரியாத சொல்லாக இருக்கும். அதாவது, புதிய சொல்லாக இருக்கும். ஆனால் உரையில் இடம் பெற்ற அதிக நிகழ்வெண்ணைக் கொண்ட முதல் 3000 சொற்கள் தெரியுமானால் ஒரு பத்தியில் ஒரு சொல் மட்டும்தான் அறிமுகமாகாத புதியச் சொல்லாக இருக்கும். இது ஆய்வினால் நிறுவப்பட்ட உண்மையாகும். இத்தகையப்பட்டியல், ஒருவருக்கு தெரியாத மொழியிலுள்ள செய்திகளை இதில் கையாளப்படும் சொற்களை மட்டும் தெரிந்து கொண்டு கூடிய வகையில் கூடிய விரைவில் யாருடைய உதவியுமின்றி புரிந்து கொள்ள உதவுகிறது.

3.2.5 சொல்லடைவு (Concordance)

மேலே விவரித்த சொல் அட்டவணையிலிருந்து சொல்லடைவை (Concordance)கணிப்பொறி வழி உருவாக்கலாம். பொதுவாக சொல்லடைவு கீழ்க்கண்ட படிவத்தில் அமையும்.

சொல்	நிகழ்வெண்ணிக்கை	சொல் இடம் பெறும் இடங்கள்	சொல் இடம் பெறும் வாக்கியம் / அடி
(Word)	(Frequence of Occurence)	(Place of Occurrence)	(Citation)

இதிலிருந்து ஒரு சொல், உரையில் எந்தெந்தப் பொருளில் (Meaning) கையாளப்பட்டுள்ளது எந்தெந்த இலக்கணக்கூறுகளில் பயன்படுத்தப்பட்டுள்ளது போன்ற செய்திகளை எளிதில் பெறலாம். கணிப்பொறிவழி உருவாக்கப்பட்ட சொல்லடைவுகளில் இந்த கூடுதல் செய்திகள் சேர்க்கப்படுகின்றன. பின்னர், சொல்லடைவில் கடைசி காலத்திலுள்ள (Column) வாக்கிய வரிசைகளை சொல்லின் பொருள்களுக்குத்

தகுந்தாற்போல் மாற்றி அமைக்கலாம். அதாவது, ஒரு சொல் ஒன்றுக்கு மேற்பட்ட பொருளில் இடம் பெற்றால், இந்தெந்த வாக்கியங்களில் முதல் பொருளை ஏற்கிறது, இந்தெந்த வாக்கியங்களில் இரண்டாவது பொருளை ஏற்கிறது போன்ற வகையில் அமைக்கலாம்.

3.2 அகராதி தொகுக்கும் கலைக்கேற்ற பகுப்பாய்வு (Lexical Analysis)

சொல்பட்டியல் (Index), சொல்லடைவு (Concordance), ஆகியவைகளை கணிப்பொறிவழி உருவாக்கும் முறைகளைப் பின்பற்றி உரைகளிலிருந்து அகராதிகளை (Dictionary from the text) உருவாக்கலாம். சொல் பட்டியலில் (Index) வேர்ச்சொல்லும் (Root), இதிலிருந்து உருவாகும் பிற சொற்களும் தனித்தனிச் சொற்களாக இருக்கும். சொல்பட்டியலில் தனிச்சொல் அல்லது முழுச்சொல் என்பது வாக்கியத்தில் வரும் சொல் என எடுத்துக் கொள்ளப்படுகிறது. அதாவது தனிப்பெயராகவும் வேற்றுமை ஏற்ற பெயராகவும், தனி வினையாகவும், விகுதிகளை ஏற்று வரும் வினையாகவும் வரும் சொற்கள் என்று பொருள் கொள்ளலாம். மகன், மகனுக்கு, மகனை, வீடு, வீட்டில் போன்றவைகளும் வா, வர, வந்து, வந்தான், படித்தான், படித்தாள் போன்றவைகளும் முழுச்சொற்கள் எனக் கொள்ளப்படுகின்றன. பொதுவாக இவைகள் வார்த்தைகள் (Words) என்றாலும், இவைகள் வார்த்தை வடிவங்கள் (Word forms) என்ற சிறப்பான பொருளில் வருவதாக மொழியியலாளர் கொள்வர். ஆனால் அகராதியில் (Dictionary) சொற்கள் இவ்வாறு தொகுப்பதில்லை. அகராதியில் வேர்ச்சொல் (Root), சொல்வகைபாடு (Parts of Speech), பொருள் (Meaning) பிரயோகம் போன்ற தகவல்கள் கொடுக்கப்பட்டிருக்கும். இங்கு 11 சொல்வடிவங்கள் இருந்தாலும் அகராதியில் மகன், வீடு, வா, படி ஆகிய நான்கு சொற்களே இருக்கும். எனவே, இம்மாதிரியான சொல் அட்டவணை, சொல்லடைவு ஆகியவைகளிலிருந்து அகராதி உருவாக்க மனித முயற்சி அல்லது மனித உதவி கணிப்பொறிக்கு தேவைப்படுகிறது. தற்பொழுது, கணிப்பொறி வழி அகராதிகள் இம்முறையில்தான் உருவாக்கப்படுகின்றன.

ஆனால், இதை முற்றிலும் கணிப்பொறி வழி மேற்கொள்வதற்கு அதாவது சொல்லடைவிலிருந்து அகராதிச் சொற்களை (Dictionary Words) உருவாக்குவதற்கு உருபனியல் பகுப்பாய்வு (Morphological Analysis) தேவைப்படுகிறது. இது இப்பணியை எளிமையாக்கும். மொழியியலாளர்கள் அகராதிச் சொற்களை 'அகராதியன்' (LEXEMES) என குறிப்பிடுகின்றனர். சொல் வடிவங்களிலிருந்து (Word forms) அகராதியினை (LEXEMES) உருவாக்குவதற்கு 'லெம்மாடிஷேசன்' (Lemmatisation) என்று பெயர். இதை முற்றிலும் கணிப்பொறி வழி மேற்கொள்வதற்கு நன்கு வரையறுக்கப்பட்ட உருவனியல் விதிமுறைகள் (Morphological Rules) தேவைப்படுகிறது. பின்னர்தான். இந்த விதிமுறைகளை

நிரலுக்கேற்ற (Programs) வகையில் கணிதப்படுத்த இயலும். ஆனால் இதற்கு தேவைப்படும் சொல்லின் எழுத்துக்கூட்டுருவுக்கான விதிகள் (Spelling Rules) சிக்கலானதாக உள்ளது. இருப்பினும், இந்த முயற்சி அகராதி தயாரிப்பதற்கு மட்டுமல்லாது கணிப்பொறி வழி மொழிபெயர்ப்பிற்கும் பெரிதும் உதவும் என்பதால் தமிழ்ப் பல்கலைக்கழகத்தில் கணிப்பொறி அறிவியல் துறையும் மொழியியல் துறையும் இணைந்து இப்பணியை மேற்கொண்டுள்ளது.

3.3. ஒலிக்கட்டு பகுப்பாய்வு (Phonological Analysis)

சொல் அட்டவணை, சொல்லடைவிலிருந்து ஒலிக்கட்டு பகுப்பாய்வை (Phonological Analysis) மேற்கொள்ளலாம். ஒவ்வொரு மொழியிலும் அதற்கென்று ஒலிகள் உண்டு. அந்த ஒலிகளைக் குறிக்க வரிவடிவத்தைப் பயன்படுத்துகின்றோம். இந்தப் பட்டியல் மூலம் ஒலிப்பு அலகுக்கும் (Functional Sound) எழுத்துக்கூட்டுருக்களுக்கும் (Letter Units) உள்ள தொடர்பை நிறுவலாம். அதாவது ஒலியனுக்கும் (Phonemes) வரிவடிவத்திற்கும் (Graphemes) இடையிலுள்ள தொடர்பை நிறுவலாம். இவ்வாறு நிறுவினால் உரையிலிருந்து பேச்சையும் (Text to Speech), பேச்சிலிருந்து உரையையும் (Speech to Text) கணிப்பொறி வழி உருவாக்க உதவும். இதற்கு கணிப்பொறிப்படுத்தப்பட்ட மொழிக்கான ஒலியன் அகராதி (Phonemic Dictionary) தேவை. இவற்றின் உதவியால் பேசும், கேட்கும் கணிப்பொறிகளை உருவாக்கலாம். தமிழ் பேசும், கேட்கும் கணிப்பொறிகளை உருவாக்க முற்படுபவர்கள் இதற்கான மொழியியல் தகவல்களுக்கு டி. வசந்தகுமாரியின் 'Generative Phonology of Tamil' என்ற நூலை நோக்கலாம். இந்நூலில் கணிப்பொறிக்கு ஏற்ற வகையில் ஒலியன் முறைகள் சூத்திரங்களாக விவரிக்கப்பட்டிருக்கின்றன.

3.4 தொடரியல் பகுப்பாய்வு (Syntactic Analysis)

வாக்கிய கட்டமைப்பை (Sentence Structure) அறிவதற்கு தொடரியல் பகுப்பாய்வு தேவைப்படுகிறது. இதற்கு ப்யன்படும் விதிமுறைகளுக்கான அமைப்பை 'பார்சர்' என்கிறோம். தொடரியல் பகுப்பாய்வினை கணிப்பொறிபடுத்த நல்ல அகராதியும் மாற்றிலக்கண விதிகளும் (Transformational Rules) தேவை. 'பார்சர்' அல்லது 'தொடரியல் பகுப்பி' (Syntax Analyser) வாக்கியத்திலுள்ள ஒவ்வொரு சொல்லின் சொல் வகைபாடுகளை (Parts of Speech) மட்டும் தருவதோடு அல்லாமல் சொற்கள் தொடரியலில் எவ்வாறு குழுமமாகின்றன (Groups) என்பதையும் கண்டறிவதாக அமைய வேண்டும். இவ்வாறு உருவாக்கப்படும் 'தொடரியல் பகுப்பாய்வி' கொடுக்கப்பட்ட வாக்கியம் இலக்கணப் பொருத்தமுடைய வாக்கியமா (Grammatical Sentence) அல்லது வழு வாக்கியமா (Ungrammatical Sentence)எனக்

கண்டறியும். இலக்கணப் பொருத்தமுடைய வாக்கியமானால் அதன் கட்டமைப்பை அதாவது எழுவாய், பயனிலை போன்ற தகவல்களை 'Tree' அமைப்பில் தரும். தமிழ் வாக்கியங்களின் கட்டமைப்பை ஆய்வதற்கு பல தொடரியல் பகுப்பாய்வு முறைகள் உள்ளன. தமிழிலுள்ள சில வாக்கிய வகைகளின் கட்டமைப்புகளை கணிப்பொறிவழி காண சில நிரல்களை இக்கட்டுரையாசிரியர் உருவாக்கியுள்ளார்.

3.5. சொல்பொருள் பகுப்பாய்வு (Sementic Analysis)

தொடரியல் பகுப்பாய்வில் (Syntactic Analysis) வாக்கியத்திலுள்ள அல்லது சொற்றொடரிலுள்ள சொற்களின் பயன்பாட்டையொட்டியே சொல்வகை (Parts of Speech) நிர்ணயிக்கப்படுகிறது. இங்கு சொல்லின் பொருள் (Meaning) தேவை இல்லை. ஆனால் சொல்பொருள் (Meaning) தேவை இல்லை. ஆனால் சொல்பொருள் பகுப்பாய்வு (Sementic Analysis) சொற்களின் பொருளை ஒட்டியே அமைகிறது. இதற்காக சொற்பொருள் பகுப்பாய்வி (Sementic Analyser) ஒரு வாக்கியத்திலுள்ள ஒவ்வொரு சொல்லிற்கும் அகராதி சொற்பொருள் (Lexical Meaning), இலக்கணப்பொருள் (Grammatic Meaning) ஆகியவைகளைக் கண்டறிந்து கொடுக்கப் பட்ட வாக்கியத்தில் புதைந்து கிடக்கும் கருத்தை அல்லது பொருளை விடையாக தருகிறது. தொடரியல் பகுப்பாய்விற்கு பயன்படும் அகராதியில் ஒவ்வொரு சொல்லிற்கான சொல்வகை அதாவது அடிப்படை சொற்பகுப்பாடுகளான பெயர் (Noun). மாற்றுப்பெயர் (Pronoun), பெயரிடை (Adjective), வினை (Verb), விடையடை (Adverb), முன்னிலை (Pre position) மற்றும் வியப்பு (Interjection) போன்ற விவரம் போதுமானது. ஆனால், சொல்பொருள் பகுப்பாய்விற்கு சொல்லிற்கான உட்பொருள் பொதிந்த கருத்து (Concept type) அவசியமாகும். அதாவது ஒரு சொல் பொருளைக் (Object) குறிக்கிறதா, மனிதனைக் குறிக்கிறதா, ஊரைக் குறிக்கிறதா, செயலைக் குறிக்கிறதா என்ற வகைபாடும் இது பற்றிய விளக்கமும் தேவை. உதாரணமாக, 'மரம்' என்ற பொருளை (Object) விளக்க வேண்டுமானால் அதன்பாகங்களின் (வேர், தண்டு, கிளை) மூலம் விளக்க வேண்டும். ஒரு மரத்திற்கு ஒரு தண்டு உண்டு. மரம் கிளையற்றதாக அல்லது கிளையுள்ளதாக இருக்கலாம். இலையுள்ளதாக இலையற்றதாக இருக்கலாம். மரத்தின் உயரம், அகலம் போன்ற செய்திகளை கணிப்பொறிக்கு ஊட்ட வேண்டும் . சுருங்கக்கூறின் 'மரம்' என்ற பொருள் பற்றி உலக அறிவை (World knowledge) கணிப்பொறிக்கு ஊட்ட வேண்டும். கணிப்பொறிக்கு 'பொருள்' (object) பற்றி விளக்குவதற்கு 'Frame' என்ற தகவல் அமைப்பும், நிகழ்ச்சித்தொடரை விளக்க 'Script' என்ற அமைப்பும் உள்ளன. இத்தகைய வசதிகளுடன், தொழில்நுட்பத்துடன் வடிவமைக்கப்படும் சொல்பொருள் பகுப்பாய்விகள் (Sementic Analyser) பல திறமைகளைப் பெற்று விளக்கும். இக்கருத்தை விளக்குவதற்கு, கீழ்க்கண்ட மூன்று வாக்கியங்களை கணிப்பொறியில் உருவாக்கப்பட்ட சொல்பகுப்பாய்விக்கு உள்ளீடு (input) தருவாதகக் கொள்வோம்.

ராமன் அயோத்திக்கு வந்தான்.

அங்கு சீதையைக் கண்டான்.

காதல் கொண்டான்.

இவைகளை ஏற்றுக் கொண்ட சொல்பொருள் பகுப்பாய்வி கீழ்க்கண்ட வகைகளிளெல்லாம் செயல்படும்.

அ) ராமன் எங்கு வந்தான்? ராமன் எங்கு சீதையை கண்டான்? ராமன் யாரிடம் காதல் கொண்டான்? போன்ற கேள்விகளுக்கு பதில்களைத் தரும்.

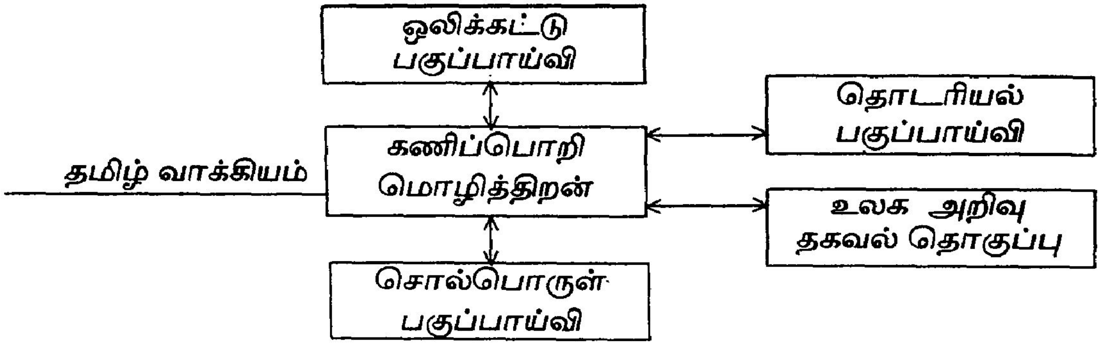
ஆ) முன்பதிவீட்டுச் சொற்களை (References) புரிந்து கொள்ளும். எடுத்துக்காட்டில் இரண்டாவது வாக்கியத்தில் ‘அங்கு’ என்பது அயாத்தியைக்குறிக்கிறது. எனவே, ‘அங்கு’ என்றசொல்லை ‘அயோத்தி’ எனக் கொள்ளும். அதபோலவே, இந்த வாக்கியத்தில் வரும் ‘கண்டான்’ என்பதை ‘இராமன் கண்டான்’ எனப் பொருள் கொள்ளும்.

இ) கொடுக்கப்பட்ட வாக்கியத்தை அதன் பொருள் மாறாமல் விரித்தோ (Expand) அல்லது சுருக்கியோ (Precise) தரும். உதாரணமாக, இரண்டாவது வாக்கியத்தை அதாவது, ‘அங்கு சீதையைக் கண்டான்’ என்பதை ‘அயோத்தியில் சீதையை இராமன் கண்டான்’ என்று விரித்து தரும்.

இங்கு தொடரியல் பகுப்பாய்வும் (Syntactic Analysis) சொல் பொருள் பகுப்பாய்வும் (Sementic Analysis) தனித்தனியாக விளக்கப்பட்டுள்ளது. ஆனால், நடைமுறையிலும், கணிப்பொறிப்படுத்தும் பொழுதும் (Computerised) இவைகளின் செயலாக்கமும் ஒன்று மற்றோன்றை சார்ந்து (Depend) நிற்கும். அதாவது, தொடரியல் பகுப்பாய்விற்கு சொல்பொருள் பகுப்பாய்வியின் உதவியும், சொல்பொருள் பகுப்பாய்விற்கு தொடரியல் பகுப்பாய்வின் உதவியும் தேவைப்படும்.

4. கணிப்பொறி மொழித்திறன்

உரையாய்விற்கு பயன்படும் பல பகுப்பாய்விகளின் ஒன்றுபட்ட இணைப்புகள் கணிப்பொறிக்கு மொழித்திறனை ஊட்ட பயன்படும். கிராஃப்போலாஜிக்கல் பகுப்பாய்வி (Graphological Analyser), உருபனியல் பகுப்பாய்வி (Morphological Analyser), ஒலிக்கட்டு பகுப்பாய்வி (Phonological Analyser), தொடரியல் பகுப்பாய்வி (Syntactic Analyser) மற்றும் சொல்பொருள் பகுப்பாய்வி (Sementic Analyser) போன்றவைகள் ஒன்றுக்கொன்றும் தொடர்பு கொண்டு செயல்கொண்டு கணிப்பொறிக்கு மொழியறிவை ஊட்டும். இந்த தத்துவம் கீழ்க்கண்ட படம் மூலம் விளக்கப்பட்டுள்ளது.



5. முடிவுரை

இக்கட்டுரையில் உரையாய்வில் (Text Analysis) என்னென்ன பணிகளை கணிப்பொறி மேற்கொள்ளலாமென விளக்கப்பட்டுள்ளது. இதன் பயன்களில் சில குறிப்பிடப்பட்டுள்ளன. இதில் கூறப்பட்டுள்ள கருத்துக்களை இதன் ஆசிரியர் பல நூல்களிலிருந்து பெற்று கணிப்பொறிவழி மேற்கொள்ளும் முயற்சிகளின் அனுபவமாகும். எதிர்கால ஆய்வுத் திட்டபணிகளின் நோக்கமாகும். உரைநடை ஆய்வில் தொடரியல் பகுப்பாய்வு (Syntactic Analysis), பொருள்நிலை பகுப்பாய்வு (Sementic Analysis) இவைகளுக்கான தொடர்பு போன்றவைகளை கணிப்பொறி படுத்த இயலும். இந்த ஆய்வுகள் வெற்றிபெறின், 'கணிப்பொறி வழி மொழி பெயர்ப்பு' மிக எளிதாகும். இந்த தகவல் யுகத்தில் மொழியியலாளர்கள் தமிழுக்கு ஆற்ற வேண்டிய சிறப்புப்பணி கணிப்பொறிக்கு ஏற்றவகையில் இலக்கணத்தை மாற்று அமைப்பில் எழுதுவதாகும். இக்கருத்தில் தமிழ்ப் பல்கலைக்கழக கணிப்பொறி அறிவியல் துறையும் மொழியியல் துறையும் ஒன்றுபட்டுள்ளது. அறிவியலையும் கலையையும் இணைக்கும் பாலம் கணிதமொழியியல் (Comtational Linguists) என்ற கருத்தோடு இந்த புதிய துறைக்கு தன்னால் இயன்றதை செய்ய முனைந்துள்ளது.

Reference Books:

1. Alain Bonnet, Artificial Intelligence; Promise and Performance, Prentice Hall International, New Delhi.
2. Daniel Hunt. V. Artificial Intelligence and Expert Systems Source Book, Chapman & Hall, London.
3. John Krutch, Experiments in Artificial Intelligence for Micro Computers, Howard W.Sams & co., USA.
4. Keith Hudson, Introducing CAL. A Practical Guide to Writing computer-Assisted Learning Programs, Chapman & Hall, London.
5. Masoud Yazdani [Editor], Artificial Intelligence Principles and Applications, Chapman & Hall, London.
6. Chellamuthu K.C. Baskaran S. Computer Analysis of TirukkuRaL, Tamil University, Thanjavur.
7. Vasanthakumari T., Generative phonology of Tamil
8. Shanmugham, collilakkaNak kooTpaaTu, aNaittindtiya tamizh moozhiyiyaR kazhakam, Annamalai Nagar.
9. Muthushamugham, ikkaala mozhiyiyazh, tirudelveezhi tennindtiya caivacittaandta nduuRpatippuk khazhakam limiTeT, cennai.

Reference Paper:

10. Baskaran S., Experiment in implementing State-Machine-Parser for Tamil Sentences, Feb 1993, CSI Communication, Bombay.

தமிழாய்வில் கணினியின் பங்களிப்பு
சிக்கல்களும், தீர்வுகளும்

முனைவர் (திருமதி) தி. கமலி
விரிவுரையாளர் - தமிழ்த்துறை
அன்னை தெரசா மகளிர் பல்கலைக்கழகம்
கொடைக்கானல் - 624 102.

இலக்கியத் தமிழ் செந்தமிழ் என்றும், பேச்சுத்தமிழ் கொடுந்தமிழ் என்றும், இயற்றமிழ், இசைத்தமிழ், நாடகத்தமிழ் என்று முத்தமிழ் ஆக, இருந்த தமிழ் அறிவியல் தமிழாகவும், தொழில்நுட்பத் தமிழாகவும், கணினித் தமிழாகவும் உருவெடுக்கத் தொடங்கியிருப்பது இனிய விளைவாகும்.

கணினியும் மனிதரும் புரிந்து கொள்ளக்கூடிய மிகக்குறைந்த எண்ணிக்கையுள்ள சிறந்த சொற்கள், எண்கள், சிறப்புக் குறியூடுகள், குறிப்பிட்ட இலக்கண அமைப்புகள் ஆகியவற்றைக் கொண்டு அமைவது கணினி மொழி (Computer language) ஆகும். சுமார் 120 கணினி மொழிகள் உள்ளன. கணினியை எந்தத்துறைக்குப் பயன்படுத்த வேண்டுமோ அந்த துறைக்கான வழியமைப்புகளை (Programs) உருவாக்கக் கூடிய கணினி மொழியறிவு இன்றியமையாதது. வழியமைப்பின் மூலம் தரவுகளைக் கணினியில் உள்ளீடு (Input) செய்யவும், பரிசீலிக்கவும் (Processing) வெளியீடு (Output) செய்யவும் இயலும். தமிழைப் பொறுத்தவரை,

- தமிழ்ச் செய்திகளை கணினியில் உள்ளீடு, வெளியீடு செய்வதற்குரிய ஆங்கில எழுத்துகளுக்கேற்ப ஒலிபெயர்ப்பு (Transliteration)செய்து பயன்படுத்துதல்.

  (உ.ம்) க - K

- அல்லது தமிழ் எழுத்திற்கு ஏற்ப ஏதோ ஒரு ஆங்கில எழுத்தைப் பயன்படுத்தும் வகையில் வழி செய்தல்.

  (உ.ம்) அ - D
  ஆ - E
  இ - F
  ஈ - R

- தமிழ்ச் சொற்களைத் தமிழிலேயே உள்ளீடு ,வெளியீடு செய்தற்கேற்ற ‘தமிழ்ச் சொல்லாய்வி’ (Tamil Word Processor)விரிவாக்குதல்

என்ற முறைகள் கையாளப்பட்டுள்ளன. தமிழ் எழுத்துக்களைக் கொண்ட கணினி தட்டச்சுப் பலகையும் (Computer Key board) காட்சித் திரையில் (Display Unit)தமிழ் எழுத்துக்களை உருவாக்கும் அமைப்பும் இன்றியமையாதது.

தமிழ்ச் சொல்லாய்விகளும், தமிழ் வழியமைப்புகளும், கணினித் தமிழ் வழியமைப்பு மொழிகளும் (Computer Tamil Programming languages) அதிக அளவில் தோற்றம் பெற்றால்தான், பயன்படுத்தப்படும் நிலை உருவானால்தான் 'கணினித் தமிழ்' வரமுடியும்.

எம்மொழியையும் கணினி மயமாக்க முடியும். தொடக்க நாளில் கணினிகள், உயர் ஆய்வுக் கருவியாக மட்டும் பயன்படுத்தப்பட்டன. இன்று இந்நிலை மாறி வருகிறது.

கணினி மூலம் கீழ்க்காணும் மொழியாய்வுப் பணிகள் வெற்றிகரமாகச் செய்யப்பட்டுள்ளன.

- கணினிச் சொல்லடைவுகள் (Concordances)
- மொழி நடை ஆய்வுகள் (Stylistic Studies)
- இலக்க ஆய்வுகள் (Syntactical Studies)
- இலக்கண அமைப்பு நிலையில் இயந்திர மொழிபெயர்ப்பு. (Machine translation on syntactic level)
- இலக்கியப் படைப்புகளின் காலத்தை நிர்ணயித்தல்(Dating of literary works)
- கணினி வழி மொழி கற்றல் - கற்பித்தல்(language learning and Teaching)
- ஆசிரியர் பெயர் தெரியாத படைப்புகளுக்க மொழிநடையை ஒப்பிடுவதன் மூலம் ஆசிரியர் பெயரைக்கண்டறிதல் (Author ship Identification)
- தமிழகத்தைப் பொறுத்தவரை, கணினி வழி மொழியாய்வுப் பணிகளை ஏற்றுச் செய்து வருவது, 'தஞ்சைத் தமிழ்ப் பல்கலைக்கழகக் கணினி மையமாகும்'

தமிழ், தெலுங்கு, மலையாளம், கன்னடம், மராட்டி, வங்காளம், இந்தி மொழிகளுக்கான ஒருங்கிணைந்த வரி வடிவ மாற்றுமுயற்சி (Integrated Script Conversion System) மேற்கொள்ளப்பட்டுள்ளது. கணினி வழி திருக்குறள் ஆய்வு (Computer Analysis of Tirukkural)வெளிவந்துள்ளது. இயந்திர மொழிபெயர்ப்பில் உருசியன் - தமிழ் (Machine Translation: Russian - Tamil) செய்யப்பட்டு வருகிறது.

இயல்பு மொழியில் கணினி பயிற்சியளிக்க சோதனை முறையில் 'ஆசிரியர் பயிற்சித் திட்டம்' (Natural language processing teacher training programme) கல்லூரி ஆசிரியர்களுக்காகச் செயற்படுத்தப்பட்டு வருகிறது. பயிற்சியின் தொடர் நிகழ்ச்சிகள் (Follow-up of rammes) மேற்கொள்ள இயலாத நிலை உள்ளது.

முன்னோடி முயற்சிகளை இத்துறையில் செயற்படுத்தியிருக்கும் தமிழ்ப் பல்கலைக்கழகக் கணினித்துறை பாராட்டுக்குரியது. ஒரு துறை அளவிலான முயற்சி மட்டும் போதுமானதா? திட்டமிட்ட இலக்குகளை அடைவதற்கும், புதுப்புது செயல்திட்டங்களை மேற்கொள்ளவும் வெவ்வேறு அணுகுமுறைகளைக் கையாளவும் பரவலான நிலையில் உழைக்க வேண்டாமா? அதற்குச் செய்ய வேண்டுவதென்ன? காணலாம்.

எவரும் எளிதில் பயன்படுத்துதற்கேற்ற வகையில், விலையில் கணினிகள் தற்சமயம் வடிவமைக்கப்படுவது மகிழ்ச்சியான செய்தியாகும். தட்டச்சுப் பயிற்சி பெற்றோர் கணினியைக் கையாள முடியும் என்கையில். 'தமிழ்க் கணினிப் பயிற்சி' மாணவர் அனைவருக்கும் கிடைக்கும் வகையில் திட்டம் தீட்டிச் செயலாக்குதல் வேண்டும்.

- அடிநிலை மாணவர்களது 'கணினி அறியாமை' அகற்றல்
- கணினி எனும் ஆங்கில வரிவடிவ மின்னணுக் கருவியைத் தத்தெடுத்து தமிழ்மயமாக்கி இம்மண்ணின் இயல்பொருள் (Indigenous Product) ஆக்குதல்.
- அனைத்துத் துறைகளிலும் கணினித் தமிழ் கால்கோள் கொள்ளச் செய்தல்

எனுமிப் பூர்வாங்கப் பணிகளில் கவனம் செலுத்தினால், இவற்றின் விளைவாக உயர்மட்டச் சாதனைகள் தன்னில் உருவாகும். அச்சாதனைகளை இனங்காணவும். ஊக்குவிக்கவும் ஏற்ற வகையில்,

- கணினித் தமிழில் சாதனை புரிவோருக்கு விருதுகள், பட்டங்கள், பதக்கங்கள், உதவித்தொகைகள் வழங்குதல்
- 'கணினித் தமிழ் மையம்' அமைத்து, வளர்ச்சிப் பணிகளை ஒருங்கிணைத்து நெறிப்படுத்துதல்.
- அயலகக் கணினி மொழி வல்லுநர் மற்றும் கணினித் தமிழ் வல்லுநர் ஆண்டுக்கொரு முறை கூடிக்கருத்துகளைப் பரிமாறுதற்கேற்ற வகையில், கருத்தரங்குகள், மாநாடுகள், பணிமனைகள் நடத்துதல்.

- அவர்களுக்கிடையே பரிமாற்றமும் (Inter Exchange) ஒருங்கிணைவும் (Co -ordination) நிலைநிறுத்தப்படுதல்

ஆகிய செயல்களை அரசு, கல்வியாளர், ஆட்சியாளர் ஆகியோர் இணைந்து மேற்கொள்ள வேண்டும்.

- கணினியறிவு கணினிப் பொறியாளரிடமும் (Comter Engineers) மொழியறிவு மொழியியலாரிடமும் (linguists) இலக்கிய ஆய்வு இலக்கிய ஆய்வாளரிடமும் (literary researchers) கல்வி பரப்பும் பணி கல்வியாளரிடமும் (Academicians) தனித்தனியாகச் சிறைப்பட்டு கிடக்கையில் மேற்கூறிய செயல்களின் சாத்தியங்கள் ஐயுறுதற்குரியன.
- வரையறைத் தடைகள் அகற்றப்பட்டு 'தமிழைத் தாய் மொழியாகக் கொண்டோர் அனைவரும் தமிழை நவீனப்படுத்த - கணினித் தமிழாக்கக் கடமைப்பட்டோர்' என்ற உணர்வுடன் ஒன்று கூடினால்தான் முடியும்.
- 'களஞ்சியம்' எனும் வளர்தமிழ்க் கையேட்டினைத் தொடங்கி, சிறப்புற நடத்திக் கொண்டும், 'கணிப்பொறியும் தமிழும்' என்ற தலைப்பில் 'சிந்தனையைத் தூண்டும் கருத்தரங்கை வடிவமைத்தும், கணினி (நவீன மின்னணுக்கருவி) - தமிழ் (பழம்பெருசெவ்வியல் மொழி) இணைக்க முற்படும் செயல்கள் நல் உதயத்தின் அறிகுறிகள்: இம்முயற்சிகள் ஒருங்கிணைந்த பணியாகத் தொடரப்பட்டு, ஒட்டு மொத்த தமிழரது வெற்றிச் செயல்பாடாக உருவெடுத்தால் வெற்றி நிச்சயம்.

இருபத்தோரம் நூற்றாண்டில், 'தமிழறிந்த எளிய பாமரரும் கணினித் தமிழறிந்தோராய் - கையாளுந் திறனுடையோராய் - அதிவேக வளர்ச்சிக்கு துணை நிற்போராய் ஆதல் கூடும். உலக மொழிகளிலே 'கணினித் தமிழ்' மொழி உன்னத மொழியாய், உயரிடம் பெறக்கூடும்.

கணிப்பொறி வழி சித்தமருத்துவம்

மரு. சே. பிரேமா, எம்.டி. (சி)
இணைப்பேராசிரியர் மற்றும் தலைவர்,
தமிழ்ப் பல்கலைக்கழகம், தஞ்சாவூர்

கட்டுரைச் சுருக்கம்

அறிவியல் மாற்றங்கள் தமிழகத்தில் பலவகையிலும் நடைபெற்று வருகின்றன. இம்மாற்றங்களுக்குக் கணிப்பொறியும் மிகப் பெரிய அளவில் தன் பங்கினை வகித்து வருகின்றது. தமிழகத்தின் தொன்மையான மருத்துவமாம் சித்த மருத்துவமும் இவ்வறிவியல் மாற்றங்களுக்கு ஆட்படுத்தப்பட வேண்டியது அவசியமாகிறது. சித்த மருத்துவக் குறிப்புகள் அனைத்தும் ஏடுகளிலும் அச்சு நூல்களிலும் மட்டுமே அமைந்துள்ளன. நோய் கணிப்பு முறையும் கணிப்பொறிவழி மேம்பாடடையச் செய்தல் வேண்டும். இதுவரை அச்சியற்றப்படாமல் உள்ள நாட்டுபுறத்தில் வழக்கில் உள்ள மருத்துவமுறைகள், பழங்குடிமக்களின் மருத்துவ முறைகள் அனைத்தும் கணிப்பொறி வழி ஒழுங்குபடுத்தப்பட்டு உலகோர் அனைவரும் அறியும் வகையில் வெளிக்கொணரப்படுதல் வேண்டும். சித்தமருத்துவ ஆய்வாளர்கள் அடிப்படை ஆய்வுகளில் காலத்தை கழித்துக்கொண்டிராமல் பல புதிய அறிவியல் ஆய்வுகளுக்கு வழிவகுக்க கணிப்பொறி வழி சித்தமருத்துவம் சீர்படுத்தப்பட்டு வெளிக்கொணரப்பட வேண்டும்.

அறிவியல் மாற்றங்கள் தமிழகத்தில் பலவகையிலும் நடைபெற்று வருகின்றன. இம்மாற்றங்களுக்குக் கணிப்பொறியும் மிகப்பெரிய அளவில் தன் பங்கினை வகித்து வருகின்றது. தமிழகத்தின் தொன்மையான மருத்துவமாம் சித்தமருத்துவம். இவ்வறிவியல் ஆய்வுகட்கு ஆட்படுத்தப்பட வேண்டியது அவசியமாகிறது. எங்கும் கணிப்பொறி எதிலும் கணிப்பொறி என்ற இன்றைய நிலையில் சித்த மருத்துவத்தைக் கணிப்பொறி வழி சீர்படுத்தி மேம்பாடடையச் செய்தல் தொடர்பான சில கருத்துக்களை இக்கட்டுரை விளக்குகிறது.

சித்தமருத்துவச் சுவடிகளும் கணிப்பொறியும்

செவி வழிச் செய்தியாக வழிவழியாக வந்த சித்த மருத்துவ முறை மன்னராட்சி காலத்தில் ஏட்டுச் சுவடிகளை அடைந்துள்ளன. ஆடிப்பெருக்கிற்கு ஆற்றில் விடப்பட்டும் போகிப்பண்டிகைக்கு நெருப்பிலிடப்பட்டும் கரையான் அரிப்பிற்கும் ஆளானது போக மீதம் உள்ள தற்போது பாதுகாக்கப்பட்டு வரும் தமிழ்ச் சுவடிகளில் 60 விழுக்காடு சித்த மருத்துவச் சுவடிகளாகும். இச்சுவடிகள்

பல்வேறு சித்தர்கள் பெயரில் பாடல்களாகவும் உரைநடையில் எழுதப்பட்ட அனுபவமுறைகளும் ஆகும். பல்வேறு சுவடிகள் ஒரே பெயரிலும் ஒரே சுவடி பல்வேறு பெயரிலும் காணப்படுகின்றன. இதனால் நூலாக வெளிவந்த சுவடிகளே மீண்டும் அச்சியற்றப்படுதலும் அச்சியற்றப்படாத சுவடிகள் வெளியுலகிற்குத் தெரியவராமலும் உள்ளன.

சுவடி நூலகங்களில் உள்ள சித்தமருத்துவச் சுவடிகள் பற்றிய விவரங்கள் கணிப்பொறி வழி ஆவணப்படுத்தி அட்டவணைப்படுத்தி வெளியிடுதல் மிகு அவசரமான செயல்பாடாகும். இதற்குரிய திட்டம் ஒன்றும் தமிழக அரசிற்கு எங்களால் அளிக்கப்பட்டுள்ளது. பல புதிய சித்த மருத்துவ நூல்களை அச்சியற்ற இது வழிவகை செய்யும். உரைநடையில் உள்ள சித்த மருத்துவச் சுவடிகள் ஏறத்தாழ 30 விழுக்காடாகும். இவற்றை கணிப்பொறிவழி அட்டவணைப்படுத்தி வெளியிட சித்த மருந்துகள் பல புதிய பரிணாமங்களை அடைய இயலும்.

சித்த மருத்துவ அச்சு நூல்களும் கணிப்பொறியும்

சித்த மருத்துவ நூல்கள் பல அச்சியற்றப்பட்டு வெளிவந்துள்ளன. ஆயினும் அவை இன்றைய அறிவியல் ஆய்வாளர்களின் ஆய்வுகளுக்கு ஏற்ற வகையில் நெறிமுறைப்படுத்தி அட்டவணைப்படுத்தப்படுதல் அவசியமாகும். இதுவரை வெளிவந்துள்ள சித்த மருத்துவ நூல்கள் பற்றி அட்டவணை ஒன்று 'சித்த மருத்துவ நூலோதி' என்ற பெயரால் எங்களால் வெளியிடப்பட்டுள்ளது. இதனை அடிப்படையாகக் கொண்டு அச்சியற்றப்பட்ட நூல்களில் உள்ள நோய்க்குறிப்புகள், மருத்துவ முறைகள் அனைத்தும் கணிப்பொறி வழி அட்டவணைப்படுத்தப்பட வேண்டும். இதனால் வெளிவந்த நூல்களே மீண்டும் மீண்டும் வெளிவராமல் தடுக்கவும் பல புதிய நூல்கள் வெளியாகவும் வெளிவந்துள்ள மருத்துவக் குறிப்புகளைக் கொண்டு அடுத்த கட்ட ஆய்விற்கு ஆய்வாளர்கள் எடுத்துச் செல்லவும் இது வழிகோலும்.

சித்த மருத்துவ நோய் கணிப்புமுறையும் கணிப்பொறியும்

நோய் கணிப்பு முறையில் எண்வகைத் தேர்வினை அடிப்படையாகக் கொண்டுள்ளது சித்த மருத்துவம், இதனை அறிவியல் முறையில் எளிதாக நோய்களைக் கணிக்க கணிப்பொறிவழி முறைப்படுத்தி வெளியிடப்படவேண்டும். எடுத்துக்காட்டாக சித்தர் நூல்கள் பலவற்றிலும் கூறப்பட்டுள்ள நாடிபற்றிய குறிப்புகளை கணிப்பொறியில் செலுத்தி நாடியைக் கொண்டு நோய் கணிப்பு முறை குறித்து செம்மைப்படுத்தலாம். இதனால் சித்த மருத்துவர்கள்

பயன்படுத்தக் கூடிய பல நோய்கணிப்புக் கருவிகள் உருவாக்கவும் நோய்கணிப்பு முறைகளை அனைத்து சித்த மருத்துவர்களும் வேறுபாடின்றி பயன்படுத்திக் கொள்ளவும் காரணமாக அமையும்.

சித்த மருந்துகளும் கணிப் பொறியும்

திரு. மெக்கன்ஸி அவர்கள் கூறியது போல் சித்த மருத்துவ நூல்களில் நோய்கள் பற்றிய குறிப்புகளை விடவும் மருந்துகள் பற்றிய குறிப்புகள் மிக அதிக அளவில் காணப்படுகின்றன. சித்த மருந்துகள் தயாரிப்பதற்கு இன்றைக்கு பல நிறுவனங்கள் முன்வந்துள்ளன. டங்கல் திட்டத்தினால் இது மேலும் பலமடங்கு வளர்ச்சியடையவும் சாத்தியக் கூறுகள் எற்பட்டுள்ளன. இலேகியம், நெய், குளிகை என்றிருந்த சித்த மருந்துகள் கேப்சூல், சிரப், டாப்லட் எனப் பல புதிய வடிவங்களைப் பெற்று வருகின்றன. ஆய்வுக் கருவிகள் எதுவுமின்றி அன்றைய சித்தர்கள் அனுபவத்தில் கண்டறிந்தவற்றை இன்றைய அறிவியல் ஆய்வுகளுக்கு உட்படுத்தி அவற்றின் செயல்திறனை வெளிப்படுத்துகிறோம். சித்த மருந்துகள் மற்றும் அவற்றின் மூலக்கூறுகளான தாது தாவர பாடாணம் சீவப்பிரிவுகள் பற்றிய அறிவியல் ஆய்வுகள் கணிப்பொறி வழி நெறிப்படுத்தி அனைவரும் அறிய வழி செய்யப்பட வேண்டும். பல்வேறு மருந்து செய் நிறுவனங்கள் தயார் செய்யும் மருந்துகள் பற்றிய குறிப்புகள் கணிப்பொறி வழி எங்களால் தற்போது அட்டவணைப்படுத்தப்பட்டு வருகிறது.

சித்த மருத்துவம் மேலும் பல சிறப்பினைப் பெற்றிட கணிப்பொறி மிகவும் பயனுள்ளதாக அமையும் என்பதில் ஐயமில்லை. கணிப்பொறி அறிஞர்களும் சித்த மருத்துவர்களும் இணைந்து இதில் ஈடுபடவேண்டியது

கணிப்பொறி எண் கட்டுப்பாட்டு இயந்திரம் மூலம் தமிழ் எழுத்துக்களைப் பொறித்தல்

அ. ஜோதிலிங்கம்,
முனைவர் சொ. கணபதி
உற்பத்தி பொறியியல் துறை,
எம். ஐ. டி வளாகம்,
அண்ணாபல்கலைக் கழகம், சென்னை-44.

கட்டுரைச் சுருக்கம்

தற்பொழுது தொழிற்சாலைகளில் கணிப்பொறி எண் கட்டுப்பாட்டு இயந்திரங்கள் மிகுதியாகப் பயன்படுத்தப்படுகின்றன. கொடுக்கப்படுகின்ற வரைபடங்களுக்கேற்ப கணிப்புக்கோவை உருவாக்கப்பட்டு அதனைக்கொண்டு இயந்திரம் இயக்கப்பட்டு பலவகையான பாகங்கள் தயாரிக்கப்படுகின்றன. நாம் பொதுவாக பயன்படுத்தும் பொறித்தல் இயந்திரம் மூலம் ஒரு சமபகுதியில் எழுத்துக்கள் பொறிக்கப்பட வேண்டுமானால் அதிக நேரம் தேவைப்படுகிறது. மேலும் ஒவ்வொரு முறையும் இயந்திரத்தை எழுத்திற்கேற்ப முறைப்படுத்திய பின்தான் இயக்க முடியும். அது மட்டுமல்லாமல் ஒரு குறிப்பிட்ட அளவு வரைதான் எழுத்துக்களைப் பொறிக்க முடியும். எனவே இக்கட்டுரையில் இந்த கால விரயத்தை குறைக்கும் முகத்தான் புழக்கத்தில் உள்ள கணிப்பொறி எண் கட்டுப்பாட்டு துருவல் இயந்திரத்தைக் கொண்டு எளிதாக எழுத்துக்களைப் பொறிக்கும் முறை பற்றி கூறியுள்ளோம். இதற்காக தமிழ் எழுத்துக்கள் அனைத்திற்கும் துருவல் இயந்திரத்திற்கேற்ப கணிப்புக் கோவை ஒன்றினை உருவாக்கியுள்ளோம். இதனைக் கொண்டு நமது விருப்பத்திற்கேற்ற பல்வேறு அளவுகளிலும் எழுத்துக்களைப் பொறிக்க முடியும். அது மட்டுமின்றி இயந்திர கட்டுப்பாட்டு முறைக்கேற்ப மாற்றக்கூடிய தன்மையும் கொண்டது. இது எல்லோராலும் எளிதாக உபயோகிக்கூடிய கணிப்புக் கோவையாகும்.

1. முன்னுரை

தமிழ் எழுத்துக்களை உலோகத்தின் சமபகுதி/சமதளத்தில் பொறிக்க நாம் பலவகையான உத்திகளைக் கையாண்டு வருகிறோம். முதன் முதலில் மனிதன் வெட்டு உளியல், சுத்தியின் உதவி கொண்டு தன்னுடைய பலத்தினால் எழுத்துக்களை பொறிக்கத் தொடங்கினான். இந்த முறை இன்னும் பல இடங்களில் புழக்கத்தில் உள்ளது. இதை அடுத்து மனிதன் இயந்திரங்களின் உதவியை நாடினான். இதனால் பலவிதமான பொறித்தல் இயந்திரங்கள் உருவாக்கப்பட்டன. இவைகளும் இன்று அதிக அளவில் உபயோகத்தில் உள்ளது. இது போன்று பலவிதமான முறைகள் இருப்பினும் ஒவ்வொன்றிலும் ஏதேனும் ஒரு குறை அல்லது முழுமை/நிறைவு இல்லாது உள்ளது. இது போன்ற குறைபாடுகளைக் குறைத்து, மிகவும் எளிதாக, சுலபமாக, விரைவாக, நம் விருப்பத்திற்கேற்ப செய்யக்கூடிய முறையை இக்கட்டுரையில் எடுத்துக் கூறியுள்ளோம். இதற்காக கணிப்பொறி எண் கட்டுப்பாட்டு துருவல் இயந்திரத்தைக் கொண்டு தமிழ் எழுத்துக்களைப் பொறிக்க கணிப்புக்கோவை ஒன்று உருவாக்கப்பட்டு அதனை செய்து பா-ார்த்த முறையினையும் விளக்கியுள்ளோம்.

2.0 புழக்கத்தில் உள்ள முறைகள்

இன்று நடைமுறையில் பலவிதமான முறைகள் இருப்பினும் ஒவ்வொன்றையும், நமது இன்றைய தேவைக்கேற்ப உள்ளதா என்று ஆராய்ந்து பார்த்தால்,இல்லை என்ற விடையே முடிவாகக் கிடைக்கிறது. எடுத்துக்காட்டாக, நாம் புழக்கத்தில் உள்ள இரண்டு முறைகளை ஆராய்ந்து பார்த்தால், பின் கண்ட குறைபாடுகள் உள்ளன.

2.1 மனிதன் வெட்டு உளி, சுத்தி கொண்டு எழுத்துக்களைப் பொறிக்கும் முறை

இந்த முறையில் எழுத்துக்களைப் பொறிக்க மிக அதிக நேரமும், பலமும் தேவைப்படுகிறது. இது மட்டுமன்றி பொறிப்பவரின் முழுகவனமும், விருப்பமும் இணைந்து இருக்க வேண்டும். அப்பொழுது தான் எழுத்துக்கள், சிறப்பாக அழகாக இருக்கும். எழுத்துக்களை திரும்ப திரும்ப பொறிப்பதால் செய்பவர் அதில் வல்லுநர் ஆகிறார். ஆகவே இது முழுவதும் செய்பவரின் திறமையையும், செயல் திறனையும் சார்ந்ததாகும். ஒரே மாதிரி செயலைச் செய்வதால் மிகவும்

சோர்ந்து விட வாய்ப்பு உண்டு. கவனக்குறைவாக செய்தால் காயம் ஏற்படும். அதுமட்டுமல்லாமல், எழுத்துக்களின் அளவுகளை அடிக்கடி மாற்றுவது சற்று சிரமம். எனவே இந்த முறை சிரமமானது. அனைவராலும் எளிதாக பயன்படுத்த இயலாது.

2.2. பொறித்தல் இயந்திரம் மூலம் எழுத்துக்களைப் பொறித்தல்

மேற்கண்ட முறையின் குறைபாடுகளை நிவர்த்தி செய்து, மனிதன் விரைவாக செய்ய இயந்திரங்களைப் பயன்படுத்தத் தொடங்கினான். அவ்வகையில் உருவாக்கப்பட்டது தான் பொறித்தல் இயந்திரம். பொறிக்கப்பட வேண்டிய எழுத்தின் வடிவத்திற்கேற்ப ஒரு மாதிரி ஒன்று உருவாக்கப்பட வேண்டும். பின் அந்த மாதிரியின் மேல் 'நகரும் முள்' ஒன்றைப் பொருத்தி, அதன் துணை கொண்டு எழுத்தின் அளவிற்கேற்ப, வடிவத்திற்கேற்ப இயந்திரத்தை மனித சக்தி அல்லது தானியங்கி சக்தி கொண்டு எழுத்துக்களைப் பொறிக்கலாம். இம்முறையில் வடிவத்தின் அழுகு குறையாமல், சிறப்பான தோற்றத்துடன் பொறிக்க முடிந்தது. எனினும் இம்முறையில் பொறிக்கும் நேரத்தை விட அந்த இயந்திரத்தை முறைப்படுத்தும் நேரம் அதிகமாக உள்ளது. மேலும் ஒரே மாதிரி அளவிற்கு தான் மிகவும் உகந்தது. எழுத்துக்களின் அளவுகளை அடிக்கடி மாற்ற முயன்றால் கால விரயம் அதிகமாகும். அதுமட்டுமன்றி எழுத்துக்களின் அளவுகளை ஓரளவுக்குத் தான் அதிகரிக்க, அல்லது குறைக்க முடியும். பன்மடங்கு மாற்ற முனைந்தால் இயந்திரம் மிக பெரியதாக வடிவமைக்கப் படவேண்டும். இதனால் இயந்திரத்தின் முதலீட்டுத் தொகையும், எடையும் அதிகமாகும்.

எனவே எழுத்துக்களைப் பொறிக்க, தனியாக ஒரு இயந்திரத்தை உருவாக்குவதை விட, இருக்கின்ற இயந்திரங்களில் ஏற்றதாக ஒன்றினை கண்டறிந்தால் அது பலவிதத்திலும் அனைவருக்கும் பலனளிக்கும். அவ்வகையில் கணிப்பொறி எண் கட்டுப்பாட்டு துருவல் இயந்திரம் மிகவும் ஏற்றதாகும்.

3.0. எழுத்துக்களைப் பொறிக்க ஏற்ற முறை

இன்றைய அளவில் தொழிற்சாலைகளில் நாம் அதிகமாக கணிப்பொறி எண் கட்டுப்பாட்டு துருவல் இயந்திரத்தை உபயோகிக்கத் தொடங்கியுள்ளோம். இதனைக் கொண்டு உலோகம் அல்லது உலோகமில்லா எந்த பொருளிலும் சமபகுதியை உருவாக்க முடியும். ஒரு வரைபடத்திற்கு ஏற்பவும், இயந்திரத்தின் கட்டுப்பாட்டு முறைகளுக்கு ஏற்பவும் , கணிப்புக் கோவை ஒன்றை முதலில்

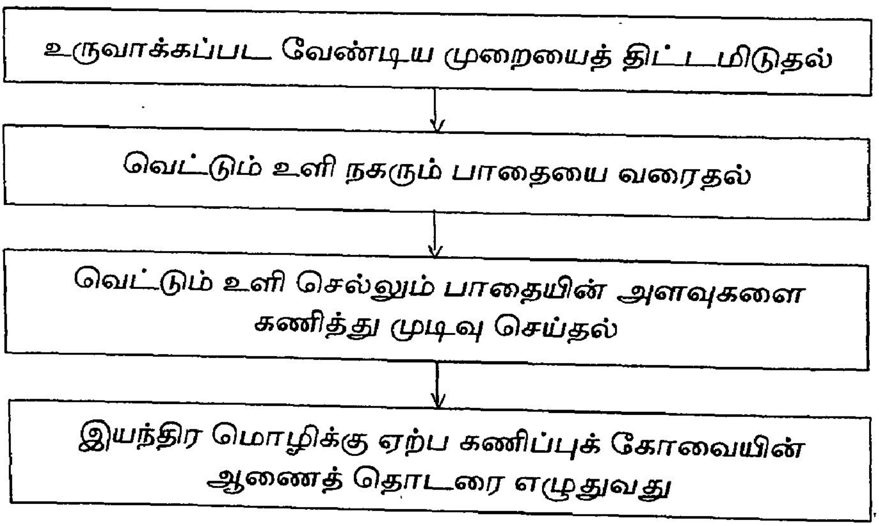
உருவாக்க வேண்டும். இது முழுக்க முழுக்க இயந்திர மொழியைச் சார்ந்தது. இதற்கென தனியான இலக்கண முறைகளும் ஆணைத் தொடர்களும் உள்ளன. இம்முறையினால் எளிதாக விரைவாக மிகவும் அழகாக பாகங்களைத் தயாரிக்க முடிகிறது. இந்த இயந்திரத்தில், கணிப்பொறி வசதியும் உள்ளதால் அதன் முழுபயனையும் அடைய முடிகிறது. எனவே செய்ய வேண்டியதை மிகவும் சிறப்பாகவும், தரமாகவும் செய்ய முடிகிறது. இந்த இயந்திமே தமிழ் எழுத்துக்களைப் பொறிக்க ஏற்றதாகும். வேண்டிய விதத்தில், அளவில், அடிக்கடி அல்லது எண்ணிக்கையைப் பொறுத்து செய்யக்கூடிய, எல்லாவித நன்மைகளையும் ஒருங்கே கொண்டது. அனைவராலும் எளிதில் கற்றுக்கொண்டு உபயோகிக்கக்கூடியது.

நாம் கணிப்புக் கோவையை மட்டும் உருவாக்க வேண்டும். பின் வெட்டும் உளியை மட்டும் தெரிவு செய்து இயந்திரத்தை இயக்க வேண்டும். இது போன்ற எளிதான முறை வேறு எதுவும் இல்லை என்பது எங்களின் முடிவு.

3.1. கணிப்புக் கோவையை உருவாக்கும் முறை

கணிப்பொறி எண் கட்டுப்பாட்டு துருவல் இயந்திரத்தின் கணிப்புக் கோவையை நாம் ஆணைத் தொடராக எழுதி இயக்கலாம் அல்லது கணிப்பொறி கணிப்புக் கோவை மொழியில் வெளிவந்துள்ளவைகளைக் கொண்டும் உபயோகிக்கலாம். இதில் நாங்கள் முதல் வகையைப் பயன்படுத்தியுள்ளோம்.

வரை படத்திற்கேற்ப ஆணைத் தொடர்களை எழுத வேண்டும். இதற்கென கொடுக்கப்பட்ட வழிமுறை:



இந்த வழிமுறையைக் கொண்டு, எல்லாவித வடிவங்களையும் நாம் உருவாக்கலாம். இதையே அடிப்படையாகக் கொண்டு நாங்கள் தமிழ் எழுத்துக்களுக்கு கணிப்புக் கோவையை உருவாக்கியுள்ளோம்.

ஆணைத் தொடரை எழுதுவதற்கான முறை மிகவும் எளிதானது. எந்த ஒரு ஆணைத் தொடரை எடுத்துக் கொண்டாலும், அதில் சில குறிப்பிட்ட ஆங்கில எழுத்தும், அதனைத் தொடர்ந்து எண்களும் தான் இருக்கும்.

எடுத்துக்காட்டாக, ஒரு ஆணைத் தொடரின் அமைப்பு

(உம்).....என் (N).......10. ஜி (G) 01 எக்ஸ் (X) 10 ஒய் (Y) 10. எஸ் (S)100. எப் (F) 50 டி (T) 1 எம் (M) 01. என்று குறிப்பிட்டு இருந்தால் அதன் விளக்கம் கீழ்கண்டவாறு இருக்கும்.

(N) என் 10 - என்பது கணிப்புக் கோவையின் விலாசம்.

(G) ஜி 01 - செய்ய வேண்டிய வேலையின் குறிப்பு

(X) எக்ஸ் 10 & (Y) ஒய் 10 - புறப்படும்,சேரும் இடத்தின் தூர அளவு.

(S) எஸ் 100 - சுழலும் வேகம்

(F) எப் 50 - ஒரு வினாடிக்கு நகரும் வேகம்

(T) டி - 1 - வெட்டும் கருவியின் எண்

(M) எம்.01 - பொதுவான குறியீடு (உம்) இயந்திரத்தை நிறுத்த,துவக்க

இது போன்று ஒவ்வொரு எழுத்திற்கும் திட்டமிட்டு, அதன் பாதைக்கேற்ப ஆணைத்தொடரை எழுதி, அதை விசைப் பலகை மூலம், கணிப்பொறியில் சேமிக்க வேண்டும். பின்னர் அதைக் கொண்டு, இயந்திரத்தை இயக்கினால், கொடுக்கப்பட்ட எழுத்தைப் பொறிக்க முடியும். இந்த வகையில் நாங்கள் எல்லா தமிழ் எழுத்துக்களுக்கும் கணிப்புக் கோவையை உருவாக்கியுள்ளோம்.

இந்த கட்டுரையில் கணிப்புக் கோவையை நாங்கள் 'டென்போர்டு' கம்பெனியின் கட்டுப்பாட்டு முறையில் இயங்கும் துருவல் இயந்திரத்திற்கேற்ப உருவாக்கியுள்ளோம். இந்த கணிப்புக் கோவையைக் கொண்டு, ஜப்பானின் 'பானுக்' கம்பெனி கட்டுப்பாட்டு முறைக்கும், ஜெர்மனியின் 'சினுமரிக்' கம்பெனியின் கட்டுப்பாட்டு முறைக்கும், இந்தியாவின் 'இனுமரிக்' கம்பெனியின் கட்டுப்பாட்டு முறைக்கும் ஏற்றதாக மாற்றும் தன்மையைக் கொண்டுள்ளது. தொழிற்சாலைகளில் இந்த வகையான இயந்திரங்களைத்தான் அதிக அளவில் உபயோகிக்கிறோம்.

நாங்கள் உருவாக்கியுள்ள மாதிரி கணிப்புக் கோவை

OFFLINE PROGRAMMING (SAMPLE ONLY)

Program - Milling - Tamil Alphabet

N001	M06	TOOL CHANGE;	TOOL 1		
N002	M03	SPINDLE FWD;	1500 RPM		
N003	G00	RAPID TRAVERSE;	X 30.000	Y 25.000	Z 1.000
N004	G01	LINEAR; F 100.00	X - Y -	F -	Z - 1.000
N005	G91	INCR'TAL FORMAT			
N006	G02	CIRCULAR CLW; Z - CENTRE	X 5.0000 F - X 0.000	Y - 5.000 Y - 5.000	F -
N007	G01	LINEAR; Z -	X - Y - 5.000 F -		F -
N008	G02	CIRCULAR CLW; X - CENTRE	X - 5.000 F - X - 5.000	Y - 5.000 Y 0.000	F -
N009	G01	LINEAR; Z -	X - 10.000 F -	Y -	F -
N010	G02	CIRCULAR CLW CENTRE	 X 0.000	 Y 5.000	
N011	G01	LINEAR; Z -	X - F -	Y 15.000	F -
N012	G02	CIRCULAR CLW; Z - CENTRE	X 10.000 F - X 10.000	Y 10.000 Y 0.000	F -
N013	G01	LINEAR; Z -	X 45.000 F -	Y -	F -
N014	G02	CIRCULAR CLW; Z - CENTRE	X 10.000 F - X 0.000	Y - 10.000 Y - 10.000	F -
N015	G01	LINEAR; Z -	X - F -	Y - 15.000	F -
N016	G02	CIRCULAR CLW; Z - CENTRE	X - 10.000 F - X - 10.000	Y - 10.000 Y 0.000	F -
N017	G01	LINEAR; Z -	X - 20.000 F -	Y -	F -
N018	G02	CIRCULAR CLW; Z -	X - 5.000 F -	Y 5.000	F -

		CENTRE	X 0.000	Y 5.000	
N019	G01	LINEAR;	X -	Y - 10.000	F -
		Z -	F -		
N020	GO3	CIRCULAR CCLW;	X 5.000	Y - 5.000	F -
		Z -	F -		
		CENTRE	X 5.000	Y 0.000	
N021	G01	LINEAR;	X 20.000	Y -	F -
		Z -	F -		
N022	G03	CIRCULAR CCLW;	X 5.000	Y 5.000	F -
		Z -	F -		
		CENTRE	X 0.000	Y 5.000	
N023	G01	LINEAR;	X -	Y - 30.000	F -
		Z -	F -		
N024	G02	CIRCULAR CLW;	X - 10.000	Y - 10.000	F -
		Z -	F -		
		CENTRE	X - 10.000	Y 0.000	
N025	G01	LINEAR;	X - 35.000	Y -	F -
		Z -	F -		
N026	G02	CIRCULAR CLW;	X - 10.000	Y 10.000	F -
		Z -	F -		
		CENTRE	X 0.000	Y 10.000	
N027	G01	LINEAR;	X -	Y 5.000	F -
		Z -	F -		
N028	G02	CIRCULAR CLW;	X 5.000	Y 5.00	F -
		Z -	F -		
		CENTRE	X 5.000	Y 0.000	
N029	G01	LINEAR;	X 5.000	Y -	F -
		Z -	F -		
N030	G02	CIRCULAR CLW;	X 5.000	Y - 5.000	F -
		Z -	F -		
		CENTRE	X 0.000	Y - 5.000	
N031	G01	LINEAR;	X -	Y - 5.000	F -
		Z -	F -		
N032	M05	SPINDLE STOP			
N033	M02	END PROGRAM			

இது போல நாம் உருவாக்கும் கணிப்புக் கோவையைக் கொண்டு, ஒரு எழுத்தின் அளவினைக் குறைக்க, அல்லது அதிகரிக்க முடியும். எழுத்தின் பிம்பத்தையும் பொறிக்க முடியும். இதற்கான ஆணைத்தொடரை இதனுடன் இணைக்க வேண்டும். அதுமட்டுமல்லாமல், பொறிக்க வேண்டிய எழுத்தின் பள்ளம், அதிகமாக இருப்பின் அதற்கான ஆணைத் தொடரை உபயோகித்து, ஒரே பாதையில் வெட்டும் உளி, திரும்பத் திரும்ப சென்று குறைந்த நேரத்தில் பொறிக்கும் வசதி உள்ளது.

இந்த கணிப்புக் கோவையை இயக்க, இதற்கான அடிப்படை இயந்திர மொழி/கட்டுப்பாட்டு மொழியுள்ள கணிப்புக்கோவை இருந்தால் போதும். குறைந்தபட்ச தேவையான கணிப்பொறி Pc 83/86 வகையாகும். மற்றபடி வேறு எந்த வகை உதிரிபாகங்களும் வேண்டாம். தமிழ் எழுத்தைப் பொறிக்க, நாங்கள் எடுத்துக் கொண்ட நேரம் 42 வினாடிகளே. ஆகவே இந்த முறை நமது தொழிற்சாலைகளிலும், விளம்பரப்பலகை, பெயர் பலகைகளிலும் தமிழ் எழுத்துக்களைப் பொறிக்க மிகவும் ஏற்றதாகும்.

4.0 முடிவுரை

நாங்கள் உருவாக்கியுள்ள தமிழ் எழுத்துக்களைப் பொறிக்கும் கணிப்புக்கோவை, மிகவும் எளிதானது, அனைவரும் உபயோகிக்கக் கூடியது. எந்தவித உலோகமானாலும், விதவிதமான அளவுகளில் நம்மால் எழுத்துக்களைப் பொறிக்க முடியும். அதுபோல் பொறிப்பதற்கு முன், அது எவ்வாறு இருக்கும் என்பதையும் கணிப்பொறி மூலம் அறிய முடியும். அந்த தோற்றத்தை முன்னரே அறிவதால், அதன் அழகையும், சிறப்பையும் நாம் அறிய முடியும். ஒரு முறை உருவாக்கப்பட்ட எழுத்தின் வடிவத்தையும், அளவையும் குறைக்கவும், அதிகரிக்கவும் முடியும். மொத்தத்தில், நாம் சிரமமின்றி, ஆபத்தில்லாமல், எளிதாக, விரைவாக, மன நிறைவுடன் அழகாக தமிழ் எழுத்துக்களைப் பொறிக்கலாம்.

தமிழ் எழுத்துக்களை அடையாளம் கண்டுகொள்வதில் வடிவமைப்புகளின் சிறப்பியல்புகள்

பேராசிரியர் சு.வைத்திய நாதன்
முன்னாள் கணிதத் துறைத்தலைவர்
அண்ணா பல்கலைக் கழகம்
சென்னை - 600 025

உரைச்சுருக்கம்

மொழி ஒவ்வொன்றிலும் எடுக்கப்படும் எழுத்துக்கள் காலவேறுபாடுகளுக்கு உகந்தவாறு, அவற்றின் வடிவமைப்புகளில் மாறிக் கொண்டே வருவதை யாவரும் அறிவர். இக்கட்டுரையில் ஆசிரியர் தமிழ் எழுத்துக்களின் சிறப்பான வடிவமைப்புகளிலுள்ள மாறுபாடுகளையும் (Variabilities), வெளித்தோற்றங்களின் வகைபாடுகளையும், பரிமாணங்கள்(Fonts), பருமன்கள்(Sizes), மற்றும் நிலைமைகள்(positions)- ஆகியவற்றின் அடிபபடையில் விவரித்துக் கூறியுள்ளார். தமிழ் எழுத்துகளை எளிதாக ஒளியியலையொட்டி அடையாளம் காண்பதில் (optical character recognition) தீவிரமான ஆராய்ச்சிகள் நடைபெற்றுக் கொண்டு வருகின்றன. தமிழ் எழுத்துகளின் படிநிலைக் குழும விவரிப்பில் (Hierarchical Representation) அமையும் துவக்கக் கோடு மட்டம் (Stroke Level), பகுதிக்கூறு மட்டம் (Component Level) மற்றும் முழு எழுத்து மட்டம் (Character Level)- ஆகியவற்றின் வரைபட அமைப்புகளில் சில இறுக்கிகளைப் (Constraints) பயன்படுத்தி, எழுத்துக்களை இனம் கண்டு கொள்ளும் நோக்கத்தைக் கருத்திற்கொண்டு எந்திரமாகக் கவருவதில் (Capture Mechanism) இக்கட்டுரை முனைந்துள்ளது.

நடைமுறையில் செயல்படுத்துவதற்குரிய உருவப்படிவத்தின் நிலைமைகளை விளக்கி, நற்பயனை அளிக்கிற வகையில் முயற்சிகள் இக்கட்டுரையில் எடுக்கப்பட்டுள்ளன. வடிவமைப்பு மற்றும் வடிவகணிதத்திற்குரிய மாற்றமின்மைகள் (Structural and Geometric Invariances), அண்மை மற்றும் வடிவகணித இறுக்கிகள் (Tological and Geometric Constraints), எழுத்துப் படிவங்களின் (Character patterns) மூலங்களின் (primitives) மேல் சுமத்தப்பட்டு விளக்கப்பட்டுள்ளன.

ஏறக்குறைய முழுமையாக எழுத்துக்களை அடையாளம் கண்டு கொள்வதற்கு உகந்ததும், நெறிப்படுத்தியதும், மிகச் சிறப்புடையதுமான பிரச்சினை ஒன்று

(Constrained-Based Optimization problem) உருவாக்கப்பட்டு தகுந்த முடிவுகளும் இக்கட்டுரையில் எடுக்கப்பட்டுள்ளன.

முன்னுரை

மொழி ஒன்றிற்குரிய எழுத்துக்களின் உருவமைப்பையும், பரிமாணத்தையும் பிறப்பிப்பதற்கே கணிப்பொறி வளைவரை மாதிரி (Computer Graph Model) விரிவாகப் பயன்படுத்தப்படுகிறது. இம்முறையில் கையெழுத்துப் படிவம் சார்ந்த சிக்கலான பிரச்சினைக்குரிய அநேக நூல்களும் வெளிவந்துள்ளன. சில மொழிகளில் இணைப்பு வடிவமைப்புகள் (Compound Structures) கிடையா. இருப்பினும், சில மொழிகளில், குறிப்பாக, இந்திய நாட்டு மொழிகளில், கூட்டெழுத்து அமைப்புகள் இன்றியமையாததாகக் காணப்படுகின்றன. சிக்கலான படிவமைப்பின் உட்கருத்தை விளக்குவதற்கும், அமைப்பிற்குரிய படிவத்தை இனம் கண்டு கொள்வதற்கும் அடிப்படைத் தொடக்க மூலங்களைத் (Basic primitive Elements) தேர்ந்தெடுப்பது அவசியமாகிறது.(4)

அச்சடிக்கப்பட்ட, மற்றும் கையெழுத்துப் படிவத்திலுள்ள எழுத்துக்கள் அடிப்படைக்கோடுகள் கணமொன்றிலிருந்து முறையாக எடுத்து வடிவமைத்துப் தோற்றுவிக்கப்பட்டுள்ளதென உறுதியாக முடிவெடுக்க முடியும்.

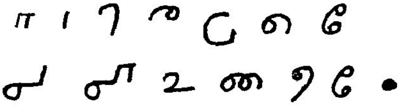
தனிப்பட்ட எழுத்து அமைப்புகளில் தோற்றம், படிவம் ஆகியவற்றின் அடிப்படையில் பலதரப்பட்டதும், வெவ்வேறானதுமான மொழிகளுக்குரிய எழுத்துகள் இனம் பிரிக்கப்பட்டு, அவற்றைப் பிறப்பிப்பதில், நியமம், நுணுக்கம், மிகைவு ஆகிய பண்புகளுக்கு ஏற்றதான குறியீட்டுப் பரிமாணங்கள் தடித்ததாகவோ, சீரானதாகவோ அல்லது சீரற்றதாகவோ உள்ள வழி முறைகளை மேற்கொண்டு உருவாக்கப்பட்டுள்ளன.

2. தமிழ் எழுத்துகளின் சிறப்பு அமைப்புகள்

நவீன தமிழ்மொழியில் 12 உயிர், 18 மெய், மற்றும் "ஃ" போன்ற தனித்த பண்புடைய எழுத்துக்கள் வரையறுக்கப்பட்டுள்ளன. அவற்றுடன் கூட்டுச் சேர்க்கையால் அமையும் எழுத்துக்களின் எண்ணிக்கை 12 x 18 = 216 ஆகும்.

முதலாவதாக, 18 மெய் எழுத்துகள் ஒலியியிலைச் சார்ந்து வல்லினம், மெல்லினம், இடையினம் என்ற மூன்று பிரிவுகளைக் கொண்டுள்ளன. இரண்டாவதாக, தமிழ் எழுத்துக்கள் அவற்றின் துவக்கப் புள்ளியைச் சார்ந்து இடது மேல்பாகம் (Left - Top) இடது கீழ் பாகம் (Left - Bottom), மற்றும் இடது மைய பாகம் (Left-Middle) என மூன்று வகைகளாகப் பிரிக்கப்பட்டு வரையறுக்கப்படுகின்றன. மூன்றாவதாக, கூட்டெழுத்துக்களை, மெய்யெழுத்து

களிலிருந்து பிறப்பிக்கும் போது, 14 தனிச் சிறப்பான குறியீடுகள் பயன்படுத்தப்படுகின்றன. அக்குறியீடுகள் வருமாறு:



3. அடிப்படை மூலங்களின் வாயிலாகத் தமிழ் எழுத்துக்களின் வடிவமைப்புகளை உருவாக்குதல்

தமிழ் எழுத்துகளை உருவாக்க, அடிப்படை மூலங்களின் கணம் Vn = { |, ⌒, ⌒, ◦} முறையே, நிலைக்குத்தான அடிப்படைக் கோடு - Vertical Stroke- (V), வட்டத்தின் ஒரு பகுதி- Portion of A Circle -(C), நீள்வட்டத்தின் ஒரு பகுதி- (A portion of an Ellipse) - (E), மற்றும் சிறிய வட்டப்புள்ளி-(A Dot) - (D) ஆகியவற்றால் வரையறுக்கப்படும்.

நிலைக்குத்தான அடிப்படைக்கோடு - V - மேலிருந்து கீழ் நோக்கியவாறு எழுதப்படும். சுழற்சிகளுக்குரிய செயலி - (Rotation Operator) - (R), V-ன் மேல் பயன்படுத்தப்படும் போது, நிலைக்கோடு 45^0 அளவில் வலஞ்சுழியாக சுழல் இயக்கத்தை நிகழ்த்துகிறது. இவ்வாறே மற்ற அடிப்படை மூலங்களும், சுழற்சியின் வாயிலாக தேவையான இருப்பிடங்களை மேற்கொள்ள முடியும்.(5)

இங்கு 31 (30+1*) தமிழ் எழுத்துக்கள் அடிப்படை மூலங்களைச் சார்ந்து, பின்வரும் தொடர் முறையில் தொகுப்பட்டுள்ளன.

(i) [ட,ப,ய,ம,ழ]; [க,ச,த]; [ஈ,ர,ற,ந,ங]]

(ii) [ல. எ,ஏ,ஞ,ள,ஐ]; [உ,ஊ,வ,ஒ,ஓ,ன,ண]]

(iii) [அ,ஆ,இ]

(iv) [ஃ]

1* (ஔ - என்ற எழுத்து ஒ,ள சேர்க்கையால் பிறப்பிக்கப்படும் என்பதால், இத்தொடர் முறையில் ஒதுக்கப்பட்டுள்ளது)

மேலும் தமிழ் எழுத்துக்களை வெளிப்படுத்துவதில் மூன்று வகையான படிநிலைக்குழும மட்டங்கள் (Three-Level Representation Hierarchy) பின் வருமாறு வரையறுக்கப்படும்.

1. அடிப்படைக்கோடு மட்டம் (Stroke Level)

2. பகுதிக்குரிய மட்டம் (Component Level)

3. முழு எழுத்திற்குரிய மட்டம் (Character Level)

மேற்கூறிய மட்டங்களின் அமைப்பில் தமிழ் எழுத்துக்களை அறிவிக்கும் போது, நிகழும் கடிமம் (Task) எளிதாக்கப்படும். இம்மூன்று மட்டங்கள், தனிப்பட்டதோரு நிலையான (Stable) உறவுகளை, அவற்றினிடையே, இட அமைப்பில் (Spatial), வல-இடது உறவு (Right - Left Relation), கீழ்-மேல் உறவு (Below-Above relation) மற்றும் சுற்றுப்புறம் சார்ந்த உறவு (Surrounding Relation) - என்ற வகைகளில், நிகழ்த்துகின்றன. அடிப்படைக்கோடுகள் உறவு பொருத்துமாறு மியல்பற்றதாக (Relatively Invariant) க்காணப்படுவதால், அக்கோடுகள் (அதாவது அடிப்படை மூலங்கள்) உறுதி நிலைப் படிவத்திற்கு மிக அண்மையில் (Close to Rigid Pattern) இருப்பதாகக் கருதப்படும். பகுதிகள் ஒவ்வொன்றும், அடிப்படை மூலங்களால் உருவாக்கப்பட்டுள்ளதால், அவை மிகுந்த மாறும் தன்மையை (Greater Variability) வெளிப்படுத்துவதாகக் கொள்ளப்படும். முழு எழுத்துகள் ஒவ்வொன்றும், அதற்கு ஒத்துள்ள அனைத்துப் பகுதிகளையும் ஒன்று சேர்த்து நிர்ணயிக்கப்பட்டுள்ளதால், அவை மீப்பெரும மாறுபாடுகளை (Greatest Variability) வெளிப்படுத்துவதாகக் கொள்ளப்படும்.(3)

4. எழுத்து வடிவமைப்பை வெளிக்கொணர்வதற்குரிய கட்டுப்பட்ட உருவ மாதிரி ஒன்றை அமைத்தல்

கட்டுப்படுத்தப்பட்ட உருவவளை வரை (Constrained Graph) ஒன்றைப் பயன்படுத்தி ஓர் எழுத்தின் அமைப்பை வெளிப்படுத்துவதற்கு, அவ்வெழுத்தின் எல்லைக்கோடு வரம்பில் (Skeleton or Outline or Boundary of the Character) மாதிரிப்புள்ளிகளின் கணம் (Set of Sampling points) ஒன்றை, அடிப்படை மூலங்களின் பண்புகளை வெளிப்படுத்தக் கூடிய வகையில் எடுக்க வேண்டும். இவ்விதம் தேர்ந்தெடுக்கப்பட்டுள்ள மாதிரிப் புள்ளிகள், கோட்டுப் பகுதிகளைக் காட்டிலும் (Line Segments) ஒரு சில நற்பயன்களைப் பெற்றுள்ளன. இவ்வகை மாதிரிப் புள்ளிகள், நடைமுறையில், பெரும்பாலும், எல்லை (இறுதி)ப் புள்ளிகள் (End points), திருப்பத்தில் அமையும் புள்ளிகள் (Turning points), T- வடிவத்தில் அமையும் சந்திப் புள்ளிகள் (T-Junctions), குறுக்கு வெட்டுப் புள்ளிகள் (Cross points) ஆகியவற்றால் வரையறுக்கப்படும். மீதமுள்ள மற்றபுள்ளிகள், அடிப்படை மூலங்களுக்குரிய கோடுகள் அல்லது வளை கோடுகளின் மேல் அடுத்தடுத்து அமையும் சீரான இடைவெளிகளை (Finite intervals of Equal Lengths) வரையறுக்கும் இடங்களில் தேர்ந்தெடுக்கப்படும். இவ்வகையான மாதிரிப் புள்ளிகளை, வளைவரையின் பிரிமுனைப் புள்ளிகளாகப் (Nodes or vertices)ப்பயன்

படுத்தப்பட்டு எழுத்துடன் உறவுடைய உருவப்படம் (Relational Graph) ஒன்றை விளக்குவதற்காக நிறுவப்பட்டுள்ளதாகக் கொள்ளப்படும். மேலும் மாதிரிப் புள்ளிகள் ஒரு சில இயல்புப் பண்புகளையும் (Properties), மற்றும் பிரிமுனைப் புள்ளிகள் ஒன்றிற்கொன்று தொடர்புத் தன்மைகளையும் (Relations) பெற்றுள்ளதாகவும் கருதப்படும். இங்கு மாதிரிப் புள்ளி மூலங்களின் இயல்புகள் அனைத்தும், அப்புள்ளிகளின் மூலங்களின் இயல்புகள் அனைத்தும், அப்புள்ளிகளின் இருப்பிடங்கள் (positions) மற்றும் இணைப்புத் தன்மைகளனைத்தையும் (Conncections) துல்லியமாக வரையறுக்கக் தக்கனவாகவும் உள்ளன. தவிரவும், இவ்வித இயல்புப் பண்புகளனைத்தையும், கட்டுப்பட்ட உருவ மாதிரி வளைவரை ஒன்றின் பிரிமுனைப் புள்ளிகள் எடுத்த எழுத்தை விளக்குவதற்கு ஏற்றவாறு பெற்றுள்ளன எனவும் கொள்ளப்படும். தொடர்பைப் பொருத்தமட்டில், இவ்வகை மாதிரிப் புள்ளிகளில் உள்ள உறவினங்கள் பின்வருமாறு கொள்ளப்படும்.

1. இரு மாதிரிப் புள்ளிகளின் இருப்பிடத்தைச் சார்ந்த உறவு

2. அவற்றினிடையேயுள்ள இணைப்புத்தன்மை

இவ்வித உறவுகள், கட்டுப்பட்டவளை வரையின் விளிம்புகளாகக் கொள்ளப்படும்.(3)

5. கட்டுப்படுத்தப்பட்டுள்ள வளைவரையில் இருக்கதக்க ஒரு சில பொதுவான உடைமைகளும் உறவுகளும்

கட்டுள்ள வளைவரையில் (Constraint graph) சில பொதுவான உடைமைகளும் உறவுகளும் இருக்க வாய்ப்புள்ளது. இத்தகைய பொது உடைமைகளும், உறவுகளும் வளைவரையில் எடுக்கப்பட்டுள்ள பிரிமுனைப் புள்ளிகளின் மேல் கட்டாயப்படுத்தப்பட்ட கட்டுமானங்களாகக் கருதப்படும். இக்கட்டுமானங்கள் எழுத்துப் படிவங்களில் அமைப்பு மாறாத் தன்மையை (Structural Invariance) துல்லியமாகக் கவரக்கூடியதாக அமைகின்றன. மாதிரிப் புள்ளிகளின் மேல் சுமத்தப்பட்ட கட்டுமானங்களின் எண்ணிக்கை குறைவாகவுள்ள போது, ஒரே மாதிரியான இனப்பிரிவுகளின் எழுத்தமைப்புப் படிவங்களுக்கு இடையே அதிக மாறும் தன்மையை (High Variability) அனுமதிக்கக் கூடியதாகவும், மாறாக, அதிக எண்ணிக்கைகளைக் கொண்ட கட்டுமானங்கள், மாதிரிப்புள்ளிகளில் உள்ள போது, குறைவுள்ள மாறும் தன்மை (Low Variability) புகுத்தப்பட்டுள்ளதாகவும் கொண்டு, கட்டுமானங்கள் இருவகையாகப் பிரிக்கப்பட்டுள்ளதாகக் கொள்ளப்படும்.

சேர்க்கைக்குரிய கட்டுமானங்கள் (Connection Constraints) - C^c - பின்வருமாறு வரையறுக்கப்படும்:-

$C^c_{ij}(O_i, O_j) =$ { உண்மை, பிரிமுனைப்புள்ளி O_i ஆனது, - O_j உடன் இலக்கு நேராகச் சேர்க்கப்பட்டுள்ள போது - அதாவது கட்டுமானம் நிறைவு செய்யப்படும் போது,

பொய், இலக்கு நேராகச் சேர்க்கப்படவில்லை என்ற நிலையாகில். அதாவது கட்டுமானம் நிறைவு செய்யப்படாமலுள்ள போது

எழுத்துப்படிவங்களின் அண்மை அமைப்புகள் (Topological Structures) தொடர்ந்து ஒரே மாதிரியாகவுள்ள போது, சேர்க்கைக்கட்டுமானங்கள் எழுத்து ஒன்றின் அண்மை மாற்றலின்மை (Topological Invariances)களைக் குறிக்கின்றன.நிலைமைக்குரிய கட்டுமானங்கள் (Position Constraints) C^b பின்வருமாறு வரையறுக்கப்படும்;

$$C^p_{123}(O_1, O_2, O_3) = \begin{cases} \text{உண்மை, } \dfrac{|O_1 - O_2|}{d_1} = \dfrac{|O_2 - O_3|}{d_2} = \dfrac{|O_3 - O_1|}{d_3} \text{ என்றுள்ளதாகில்} \\ \text{பொய், மேற்கண்டவாறு இல்லையெனில் (Otherwise)} \end{cases}$$

இங்கு $|O_1 - O_2| = O_1, O_2$ - பிரிமுனைப்புள்ளிகளுக்கு இடையிலமையும் (Euclidean - Distance) - தொலைவு ஆகும்.

d_1, d_2, d_3 - ஆகியவை, மாதிரிப்புள்ளிகளுக்குமிடையே ஏற்றதான மாறிலிகளாக (constants) த் தேர்ந்தெடுக்கப்பட்டுள்ள தொலைவுளைக் குறிக்கும்.

நிலைமைக் கட்டுமானங்கள், எழுத்து ஒன்றின் வடிவ கணிதத்திற்குரிய மாற்றலின்மைகளை (Geometrical Invariances)க் குறிக்கின்றன.

எனவே, நிலைமைக் கட்டுமானங்கள் நிலையானதாக (Rigid) கொள்ளப்படும்.

எழுத்துப் படிவமைப்பில் மிகையான மாற்றம் அனுமதிக்கப்பட வேண்டுமென்பதற்குகந்தவாறு, இவ்வித நிலைக்கட்டுமானங்களைத் தளர்த்தி விட வேண்டும். அதாவது, நிலையற்ற நிலையில் ஏற்படும் மாற்றத்தை (Nonrigid variation) அடிப்படை மூலங்கள் மட்டம், பகுதிக்குரிய மட்டம், முழு எழுத்திற்குரிய மட்டம் - ஆகிய நிலைமைகளில் அனுமதிக்க வேண்டிவரும். பூலியன் மதிப்புகளுக்குப் (Boolean values) பதிலாக - (ஆம், இல்லை) - அளவுபடுத்துவதற்குரிய பிழைகள் (Quantization Errors), இரைச்சல் (Noise), மற்றும் வெவ்வேறு விளக்க மட்டங்களில் நிகழும் மாற்றங்களின் வேறுபாடுகள் (Differences of Variability at Different Representation Levels) - ஆகியவற்றிற்குகந்தவாறு அளவின அளவைகள் (Quantitative Measures) எடுக்கப்படவேண்டும்.(1), (2), (3).

6. வளைவரையின் உருவப்படிவம் ஒன்றை அமைப்பதற்குரிய செய்முறையின் முக்கியத்துவம்

முப்படி நிலைமைக் கட்டுமானங்கள் (Ternary position Constraints), இடப்பெயற்சி மாற்றமின்மை (Translation Invariance), சுழற்சி மாற்றமின்மை (Rotation Invariance) அளவு அலகில் மாற்றமின்மை (Scaling Invariance) - ஆகியவற்றை விளக்குகின்றன. எனவே முப்படி நிலைமைக் கட்டுமானங்களைப் பயன்படுத்தப்பட வேண்டிய நிர்பந்தம் ஏற்படுகிறது. இதற்குத் தேவையான கணிப்பொறி நினைவு நிரப்பிடம் (Memory Space) அதிக அளவில் தேவைப்படுகிறது. அப்போது இதற்காக ஆகும் செலவும் மிகுதியாகிறது. எனவே அதிகப்படி நிலைமைக் கட்டுமானங்களைத் தோராயமாக அறிவிக்க, இருபடி நிலைமைக் கட்டுமானங்கள் பயன்படுத்தப்படும் போது அலகில் மாற்றமின்மையை இழக்க நேரிடுகிறது. இவ்வித இழப்பை ஈடுசெய்வதற்கு பருமத்தை (Size) இயல் நிலைத்தர அளவிற்கு மாற்றி அமைக்க வேண்டி வருகிறது. இரைச்சலால் அடிப்படை மூலங்களில் உண்டு பண்ணப்பட்ட முறிவுகளை மற்றும் தவறான இணைப்புகள் - ஆகியவற்றின் நிலைமைகளை கணக்கில் எடுத்துக் கொள்வதற்காகச் சேர்க்கைக்குரிய கட்டுமானங்கள் பின்வருமாறு மாற்றி அமைக்கப்படும்.

$$C^{c}_{ij}(O_i, O_j) = \begin{cases} O, & O_i \text{ -ஆனது } O_j \text{ உடன் நேராக சேர்க்கப்பட்டுள்ள போது} \\ \Gamma, & \text{சேர்க்கப்படாமலுள்ள போது} \end{cases}$$

இங்கு Γ, ஒரு நேர் மிகை மாறிலியாகும். இவ்வித முறிவுகள் அல்லது தவறான இணைப்புகளின் எண்ணிக்கை அதிகமாகவுள்ள போது, Γ-ன் மதிப்பைச் சிறியதாகவும், இதற்க மாறாகவுள்ள போது, Γ- ன் மதிப்பை அதிகமாகவும் கொள்ள வேண்டும். எழுத்திற்குரிய வளைரையில் m மாதிரிப் புள்ளிகளை எடுக்கும் போது, m (m-1) இணைப்புக் கட்டுமானங்கள் கிடைக்கின்றன.

இரு மாதிரிப் புள்ளிகளுக்கு - (O_i ,O_j களுக்கு) - இடையிலமையும் தொலைவு d_{ij} ஆனது இருபடி நிலைமைக் கட்டமானத்தை அறிவிப்பதாகக் கொள்ளப்படும். d_{ij}- களின் கூட்டுச்சராசரியை $\mu(d_{ij})$ எனக்கொள்ளவும். d_{ij}-களில் நிகழும் மாற்றத்தை, (பெரும d_{ij}- சிறும d_{ij}) என்ற வீச்சு அளவு வரையறுக்கிறது. வீச்சின் மதிப்பளவையை R_{ij}- எனக்குறிக்க இருபடி நிலைமைக்கட்டுமானம் பின்வருமாறு கணக்கிடப்படும்.

$$C^{p}_{ij}(O_i, O_j) = \frac{1}{R_{ij} + \varepsilon} \left| d_{ij} - \mu(d_{ij}) \right|$$

C^{p}_{ij}- - ஆனது தேரப்பெறாத அமைப்பை (Indeterminate Form) த் தவிர்க்க, $\varepsilon > 0$ என்ற ஒரு சிறிய நேர் மிகை மாறிலி எடுக்கப்படும்.

இணைப்புக்கட்டுப்பாடுகள் மற்றும் சாருகின்ற நிலைமைக் கட்டுப்பாடுகள் ஆகியவற்றின் சேர்வு (Combination), வெவ்வேறான எழுத்து விளக்கப்படி நிலைக்குழும மட்டங்களில், ஒரு எழுத்தின் வடிவமைப்பின் சிறப்பியல்பு மாற்றங்களை ஓரளவிற்கு விவரிக்க வல்லமை பெற்றுள்ளது. Rij ($1 \leq i, j \leq m, i \neq j$) என்ற வீச்சு அளவு (Range) மேற்கூறிய மாற்றங்களை, ஒரே சீரான முறையில், விளக்கக்கூடிய வகையில் அமைகிறது. இதைப் பற்றிப் பின் வரும் குறிப்புகள் (Observations) அறிவிக்கத்தக்கனவாகவுள்ளன.

(i) ஏதேனுமொரு சோடிப்புள்ளிகள் O_i, O_j களுக்கு. $R_{ij} = 0$ என்றுள்ள போது, நிலைமை சார்ந்த மாறுபாடுகள். அப்புள்ளிகளினிடையே, இல்லை என்று கொள்ளப்படும். ஒரே மாதிரியான எழுத்தின் படிவங்களுக்கு இடப்பெயர்சி செயலி ஒன்று (Translation) மற்றும் சுழற்சி (Rotation) செயலி ஒன்று ஆகியவற்றிலிருந்து பிறப்பிக்கப்படும் மாற்றங்கள் மட்டுமே இருக்கத் தக்கனவாகும்.

(ii) ஏதேனுமொரு சோடிப்புள்ளிகள் O_i, O_j - களுக்கு, R_{ij} - ஆனது மிக அதிக மதிப்பைப் பெற்றுள்ள போது, ஒரே மாதிரியான இனத்தின் படிவங்களுக்கு, அவ்வின அண்மையின் அமைப்புகள் (Topological Structures) அதுவேயானதாகத் தொடர்ந்து இருப்பதற்கு ஒப்பாக அமையும் சந்தர்பங்களில், வடிவ

கணிதத்திற்குரிய மாற்றங்களில் (Geometrical Variations) அதிக நெகிழ்மைகள் (Flexibilities) இருக்கத்தக்கனவாகும்.

(iii) கட்டுப்பாடு வளைவரை ஒன்றில்,. பல்வேறு பகுதிகள் (Components) உள்ளனவெனில்,

ஒவ்வொரு பகுதியினுள்ளேயும் $R_{ij} = 0$ என்று இருக்கும் போதும், மற்றும் $R_{ij} >> 0$ என்று அனைத்து மற்ற நிலைமைகளில் உள்ள போதும், பகுதி ஒவ்வொன்றும், நெகிழ்வுக்கு உட்படா நிலைப்பொருள் (Rigid Object) ஒன்றாகக் கொள்ளப்படும். மேலும் அனைத்துப் பகுதிகளின் மாறுபாடுகள், அவற்றினிடையே, ஒன்றிற்கொன்று சார்பற்றதாகவும் (Relatively Independent) கருதப்படும்.

கட்டுப்பாடு வளைவரையை உருவாக்கும் நேரத்திலேயே

$$O \leq R_{ij}\Big|_{\substack{\text{அடிப்படை} \\ \text{மூலம்}}} < R_{ij}\Big|_{\text{பகுதி}} < R_{ij}\Big|_{\text{எழுத்து}} << \infty$$

என்ற அமைப்பை உள்ளடக்கியதாகக் கொள்ளப்படும்.

7. எழுத்து ஒன்றை அடையாளம் காண்பதின் விதியாக்கம் இறுதிச் சிறப்புத் தீர்வுப் பிரச்சினையாகக் கருதுதல்

$N_o = \{O_1, O_2, \ldots\ldots\ldots\ldots O_m\}$ என்பது , m - மாதிரிப் புள்ளிகளை, முன் கூட்டியே தெரிந்துள்ள எழுத்து ஒன்றிற்குரிய ஒரு கட்டுப்பட்ட வளைவரையில் எடுத்து அமைக்கப்படும் கணமொன்றைக் குறிக்கட்டும்.

ஒரு தெரியாத உள்ளிடும் எழுத்துப்படிவ அமைப்பு வரம்பின் மேல் (On the Skeleton of an Unknown input character pattern) எடுக்கப்படும் n-புள்ளிகளின் $\{p_1, p_2, \ldots p_n\}$ கணத்தை

$N_p = \{p_1, p_2, \ldots p_n\}$ என்பது குறிக்கிறது. வரம்பின் மேல் அமையும்

p_j , j=1,2, . . . n புள்ளிகள் ஒவ்வொன்றையும் O_i, i=1,2,......m புள்ளிகள் ஒவ்வொன்றினுடனும் தொடர் வரிசையாக ஒப்பிடவும் O_i க்கு மிக அருகில் வரம்பிலமையும் புள்ளியை $p_{k(i)}$ எனக்கொள்ளவும் (k=1,2, ...j,.........n) அதாவது மாதிரிப் புள்ளி, O_i= யை;, P_j புள்ளிக்கு அமைப்பு மாற்றம் செய்வதாகில், $O_i \rightarrow P_j$ என்ற ஒத்திசைவு பெறப்படும்.

உள்ளிடும் எழுத்துப்படிவ அமைப்புப் புள்ளி $P_{j=}P_{k(i)}$) க்கும், கட்டுப்பட்ட வளைவரையிலமையும் மிகக் குறைவான தொலைவிலுள்ள O_i , i = 1, 2, ... m க்குமிடையே உள்ள தூரத்தை (distance) $E(O_1, O_2 .... O_m)$ என்ற விலைச் சார்பு (Cost function) வரையறுப்பதாகக் கொள்ளப்படும். அப்போது

$$\text{தூரம்} = \text{மீச்சிறுமம்}\ E(O_1, O_2, \ldots\ldots O_m) \quad (7.1)$$

(ஒத்திசைவு அமைப்பு மாற்றங்கள் பொருத்து)

$$= \underset{O_i \to p_{j=k(i)}}{\text{சிறுமம்}} \left[\sum_{i \neq j} C^c_{kj}(O_i, O_j) + \sum_{i \neq j} C^p_{ij}(O_i, O_j) \right]$$

$$= \sum_{i=1}^{m} C^c_{ik(i)}\left(p_{k(i)}, p_{k(j)}\right) + \sum_{i=1}^{m} C^p_{ik(i)}\left(p_{k(i)} - p_{k(j)}\right)$$

என்றாகும் [3]

இங்கு எழுத்து ஒன்றை அடையாளம் கண்டு கொள்வது, ஒரு தெரியாத இன வகையின் எழுத்து ஒன்றிற்கும், அவ்வெழுத்திற்குரிய கணிப்பொறி அகராதி ஒன்றில் நிரந்தரமாகப் பதிவு செய்து வைக்கப்பட்டுள்ள (Stored in a Dictionary) தெரிந்த வகைகளுக்குரிய வளைவரைப் படிவ அமைப்பு மாதிரிகள் ஒவ்வொன்றிற்கும் இடையேயுள்ள தூரத்தை ஒப்பீடு செய்வதால் வரையறுக்கப்படும். பதிவு செய்து வைக்கப்பட்டுள்ள மாதிரிகளின் கணத்தை, $L = \{1, 2, \ldots\ldots \ell\}$ எனக் கொள்ளவும்.

ஒரு குறிப்பிட்ட மாதிரி ℓ க்கு உரிய கணிப்புகள் வருமாறு

சேர்க்கைக்குரிய கட்டுமானங்கள்

$$C^c_{ij}(O_i, O_j) = C^c_{ik(i)}\left(p^{\ell}_{k(i)}, p^{\ell}_{k(j)}\right)$$

$$= \begin{cases} 0, & p^{\ell}_{k(i)}, p^{\ell}_{k(j)} \ \text{ஆகியவை நேராகச் சேர்க்கப்பட்டுள்ளபோது} \\ \Gamma^{\ell} & \text{சேர்க்கப்படாமலுள்ளபோது} \end{cases}$$

நிலைமைக்கட்டுமானங்கள்

$$C^{p}_{ij}(O_i, O_j) = C^{p}_{ik(i)}\left(p^{l}_{k(i)}, p^{l}_{k(j)}\right)$$

$$= \frac{1}{R^{l}_{ik(i)} + \varepsilon} \left| d^{l}_{ik(i)} - \mu(d^{l}_{ik(i)}) \right|$$

இங்கு

$$d^{l}_{ik(i)} = \sqrt{\left[x^{l}_{k(i)r} - x^{l}_{k(i)r}\right]^2 + \left[y^{l}_{k(i)s} - y^{l}_{k(j)s}\right]^2}$$

l = 1,2,................l

r = 1,2,................m

s = 1,2,................m

k(i) = 1,2,................m

k(j) = 1,2,................m

$R^{l}ik(i)$ = பெருமம $d^{l}_{ik(i)}$ - சிறுமம $d^{l}_{ik(i)}$

i = 1,2, ...m

k(i) = 1,2,..........m

8. முடிவுரை

மாதிரிக்கணம் L= {1,2,,..........l} -ல் அடங்கிய ஒவ்வொரு வரைவரைப்படிவ அமைப்பிற்கும்,மேற்கூறிய தூரத்தைக் கணக்கிட்டு,எந்த ஒரு மாதிரி அமைப்பிற்கு இத்தூரம் மிகக்குறைவாகவுள்ளதோ அதுவே தெரியாத உள்ளிடும் எழுத்துப் படிவ அமைப்பை அறிவிப்பதாக முடிவு செய்யப்படும்.

இக்கட்டுரையில் கட்டுப்பட்ட வளைவரை ஒன்று,பொதுவாக எழுத்து ஒன்றை அடையாளம்கண்டுகொள்ளஉருவாக்கப்பட்டு அறிமுகம்செய்யப்பட்டுள்ளது.எழுத்து இனம் ஒவ்வொன்றும், கட்டுப்பட்ட வளைவரை மாதிரி அமைப்பில் விளக்கம்

செய்யப்பட்டுள்ளது. எழுத்து வரம்பின் மேல் எடுக்கப்படும். மாதிரிப்புள்ளிகள், வளைவரையின் பிரிமுனைப் புள்ளிகள் (Nodes) அல்லது உச்சிப்புள்ளிகள் (Vertices) எனக் கொள்ளப்பட்டுள்ளன. நிலைமைக் கட்டுமானங்கள் மற்றும் சேர்க்கைக் கட்டுமானங்கள் வளைவரையின் விளிம்புகளாக (Edges) அமைகின்றன. இவ்வித மாதிரி எழுத்து ஒன்றின் அண்மை மற்றும் வடிவ கணித மாற்றலின்மை பொதுவாகவும், சீரான முறையிலும் கவர்ந்து வெளிப்படுத்தப்பட்டுள்ளது.

எழுத்து ஒன்றை அடையாளம் கண்டு கொள்வதற்கு உரியதான விதியாக்க முறை கட்டுமான அடிப்படையில் இறுதிச் சிறப்புத் தீர்வுப் பிரச்சினையாக வரையறுக்கப்பட்டுள்ளது. உள்ளிடும் எழுத்துப் படிவ அமைப்பில் உள்ள புள்ளி P_j ஐ கட்டுப்பட்ட வளைவரையிலமையும் புள்ளி O_i உடன் ஒப்பீடு செய்யப்பட்டு, அவற்றிற்கு இடையேயுள்ள தொலைவுத்தூரத்தை மீச்சிறுமமாக்கும் பிரச்சினையாக உருவாக்கம் செய்யப்பட்டு, நடைமுறையில் பிரச்சினைக்கு விடைகள் காண்பதற்கு உரியதான கணித முறை வழியமைப்புகள் எழுதப்பட்டுள்ளன.

ஆதாரங்கள்

1. K. FUKUNAGA — Introduction to Statistical Pattern Recognition, New York, Academic Press, 1982
2. V.K.GOVINDAN — "Character Recognition, - A review", Pattern Recognition, Vol.23, No.7 PP.671-683, 1990.
3. XIAOFEI HUANG, JUNGU and YOUSHOU WU — A Constrained approach to Multifont Chinese Character Recognition, IEEE Transactions in pattern Analysis and Machine Intelligence, Vol.15. No.8. PP.838-843, 1993
4. S.VAIDYANATHAN, B. MUTHURAMAN and J.V. MANI — Letter Structure in Eight dimensional frame Proceedings of the First Tamil Nadu science Congress, Tiruchirapalli - pp. 271 - 279, 1992
5. S. VAIDYANATHAN and B. MUTHUKUMARAN — Recognition of Tamil Characters in an AI environment - pp. 289 - 293 International computing congress, Dec.1993, Hyderabad

கலைச் சொற்கள்

அண்மை இறுக்கிகள்	:	Topological Constraints
அண்மை மாற்றலின்மை	:	Topological Invariance
இறுக்கிகள்	:	Constraints
இணைப்பு வடிவமைப்புகள்	:	Compound Structures
எழுத்துப் படிவம்	:	Character Pattern
எந்திரமாகக் கவருதல்	:	Capture Mechanism
கட்டுள்ள வளைவரை	:	Constrained Graph
விலை சார்பலன்	:	Cost Function
சேர்க்கைக் கட்டுமானம்	:	Connection constraint
துவக்கக் கோடு மட்டம்	:	Stroke Level
பகுதிக்குரிய மட்டம்	:	Component Level
பரிமாணங்கள்	:	Fonts
பருமன்கள்	:	Sizes
மிகச் சிறப்புடைய பிரச்சினை	:	Optimization Problem
முழு எழுத்து மட்டம்	:	Character Level
மூலங்கள்	:	Primitives
வடிவமைப்பு மாற்றமின்மை	:	Structural Invariance
வடிவ கணித மாற்றமின்மை	:	Geometrical Invariance
வடிவ கணித இறுக்கிகள்	:	Geometrical Constraints

கணிப்பொறியில் தமிழ் எழுத்துக்களை உருவாக்குவதில் உள்ள சில சிக்கல்கள்

ஆண்டோ பீட்டர்
ஸாஃப்ட் வியூ கம்ப்யூட்டர்ஸ்

கணிப்பொறியில் இந்திய மொழிகளைப் பொருத்த வரை வாடிக்கையாளர்களுக்கும், எங்களுக்கும் உள்ள பெரிய குறையே எழுத்தின் உருவம் தான். சில வாடிக்கையாளர் DTP-யில் தமிழ் எழுத்தின் பிரிண்ட் அவுட்டை பார்த்தால் கொம்பு (பு), கொக்கி (கி), புள்ளி (க்), சுழி (சூ), ஓட்டுக்கால் (தூ), தும்பிக்கை (ணே), ஆகிய எழுத்துக்கள் நகர்ந்துள்ளதா? என்று கூர்ந்து கவனிப்பார்கள். இவையெல்லாம் இரு எழுத்துக்களின் கூட்டாக உருவாகுபவை. இதனை சர்வதேச அளவில் பயன்படுத்தும் மென் பொருட்களில் கூட இக்குறையைச் சரிசெய்ய முடியவில்லையே என்பார்கள். ஆனால் இவைகளை WINDOWS 3.1 & 4.0 (BETA VERSION) களில் சமாளிக்க முடிந்தது எங்கள் மனதிற்கு மகிழ்ச்சியாக உள்ளது. அயல்நாட்டு மொழிகள் சதுரமானவை. ஆதலால் அவைகளில் கொம்பு, சுழி, புள்ளி ஆகியவை இல்லை. சதுர உருவம் உள்ளதால் வலது புறம், இடது புறம் கொடுக்கப்படும் காலி இடம் சமமாக இருக்கும். அதனால் எழுத்துக்கும், எழுத்துக்குமிடையே உள்ள இடம் (INTER CHARACTER SPACING) சமமாக இருக்கம். நம்முடைய எழுத்துக்கள் மேடும், பள்ளமுமாக உள்ளதால் பார்த்துப் பார்த்து பொறுமையாகச் செய்ய வேண்டியுள்ளது. இதனால் தயாரிப்பளவு குறைகிறது.

தென்னிந்திய மொழிகளில் மலையாளத்தை எடுத்துக் கொண்டால் கொக்கி, சுழி, புள்ளிகள் இருந்தாலும் அவை எழுத்தோடு எழுத்து ஓட்டாது. பிரச்சினை இல்லை. தெலுங்கு, கன்னடம், எழுத்துக்களில் கொக்கி, சுழிகள் எழுத்தோடு எழுத்தாக பிணைந்திருக்கும். அதனைப் பார்க்கும் போது தமிழைப் பார்க்க ஆறுதலாயிருக்கும்.

சிலர் கணிப்பொறியில் கிரந்த (   ) எழுத்துக்களைக் கேட்பார்கள். சிலர் தும்பிக்கை ஈன எழுத்துக்களைக் கேட்பார்கள். ஆதலால் நாங்களும் அவற்றை உருவாக்கி கூட்டாகக் கொடுத்து விடுகிறோம். மேலும் நம் பண்பாட்டுக்கு ஒத்துப் போகும் குத்து விளக்கு, அகல், கும்பம், கலசம், பிள்ளையார், ஈட்டி, வேல், மயில், (இரட்டை இலை, உதய சூரியன்) ஆகியவற்றையும் அளிக்க வேண்டியுள்ளது. இவற்றையும் பலர் எழுத்துக்கள் போல் கருதிக் கேட்கிறார்கள்.

Inscrypt Ms DOS Word Processor யில் எந்த மொழியிலும் விசைப்பலகை மூலம் தான் உள்ளீடு செய்வார்கள். இப்போது தயாரிக்கப்படும் மென் பொருளில் எலிப்பொறி (Mouse) மூலமும் கூட உள்ளீடு செய்யலாம். இதனால் கணிப்பொறியில் தமிழ் தட்டச்சு தெரியாதவர்கள் கூட இந்த சொற்றொகுப்பானை பயன்படுத்தலாம்.

எது சரி, எது தவறு?

சில வாடிக்கையாளர்கள், ஐ, ஜ போன்ற எழுத்துக்கள் Base Line விட்டு கீழே இருக்க வேண்டும் அல்லது மேலே இருக்க வேண்டும் என குழப்புவார்கள். எது சரி, எது தவறு என புரியவில்லை.

ஜகன் ஜகன்

கணிப்பொறியில் சிலர் க்ஷ, ஷ ஆகிய எழுத்துக்களில் புள்ளி குறிப்பிட்ட இடத்தில் இருக்கவேண்டும் என்று சொல்வார்கள். எது, எது தவறு என புரியவில்லை.

க்ஷ, ஷ க்ஷ, ஷ க்ஷ, ஷ

கணிப்பொறியில் தி,சி ஆகிய எழுத்துக்களில் கொக்கி எங்கிருந்து தொடங்க வேண்டும்? என்பதும் சரியாக முடிவு செய்யப்படாத ஒன்று.

1. தி, சி 2. தி, சி

கோல இயல் வழி கையெழுத்து அறிதல்

முனைவர் வெ. கிருஷ்ணமூர்த்தி
கு. நாகசுப்ரமணியன் ச. கார்த்திகேயன்
தி. இரா.வேங்கடராகவன் கோ. கணேசன்
கணிப்பொறி அறிவியல் தொழில் நுட்பத்துறை,
அண்ணா பல்கலைக் கழகம், சென்னை - 25.

கட்டுரைச் சுருக்கம்

ஓர் எழுத்து பல நூறு புள்ளிகளின் கோர்வை. இவற்றில் சில புள்ளிகளை மட்டும் தேர்ந்தெடுத்து, அவற்றின் இடையே உள்ள உறவுகளை ஆராய்வதன் மூலம் அது எந்த எழுத்து என்று கண்டுபிடிப்பது இந்த முறையின் நோக்கம். புள்ளிகளின் எண்ணிக்கை மிகச்சிலவாகக் குறைக்கப்படும் போது எழுத்தின் பல தகவல்கள் குறைந்து அல்லது அழிந்து போகும். தேவையான முக்கிய தகவல்கள் சிதையாமல் காக்கப்பட வேண்டும். அடுத்தபடியாக, இந்த குறைந்த புள்ளிகளின் தொகுப்பு சில பகுதிகளாகப் பிரிக்கப்படுகின்றன. இந்தப்பகுதிகள் ஒவ்வொன்றாக தரப்படுத்தப்பட்ட பகுதிகளுடன் ஒப்பு நோக்கப்பட்டு, அதிலிருந்து எழுத்து எதுவென்று முடிவு செய்யப்படும். இம்முறையில் ஆராயத் தேவைப்படும் நேரம், மிகக்குறைவாக இருக்கும். கையெத்துக்களில் தோன்றும் வேறுபாடுகளை எளிதில் நீக்கி விட்டு எழுத்தை அறியலாம். இம் முறையில் சிலபுள்ளிகளையும் அவற்றைச் சேர்க்கும் சில கோடுகளையும் கொண்ட கோலம் என்னும் கணித அமைப்பு அடிப்படையாக அமைகிறது. (இம்முறையை சோதிக்க) தொடக்க நிலைத் தயாரிப்பில் உள்ள ஒரு கட்டளை நிரல் தற்போது சுமார் 15 எழுத்துக்களை அடையாளம் காண்கிறது. இம்முறை வெற்றி பெறும் என்பதற்கு இது சான்று. நிரல் முழுமை பெறும் போது எல்லாத் தமிழ் எழுத்துக்களும் அடையாளம் காணப்படும்.

1. கோலம் உருவாக்குதல்

எழுதப்பட்டுக் கொண்டிருக்கும் கையெழுத்தில், ஒரு எழுத்திற்கு பல புள்ளிகள் வரிசையாக, ஒன்றன் பின் ஒன்றாகக் கிடைக்கும். இந்தப் புள்ளிகளை ஆராய்ந்து, அவற்றில் சிறு தொலைவிற்குள் மேற்புறப்புள்ளி எனக் கூறக்கூடியவை முதலில் கணிக்கப்படுகின்றன. இதே போல் சிறு தொலைவிற்குள் கீழ்ப்புறப்புள்ளி, வலப் பக்கப் புள்ளி, இடப்பக்கப் புள்ளி என்பவையும் கணக்கிடப்படும். இத்துடன் முதற்புள்ளி, கடைசிப்புள்ளி, கோடுகள் குறுக்கிடும் புள்ளி முதலியவையும் கணக்கிடப்படுகின்றன. இம்மாதிரிப் புள்ளிகள் ஒரு எழுத்தில் மிகச் சிலவே இருக்கும். இவற்றில் எந்தப் புள்ளியிலிருந்து எந்தப்புள்ளிக்குச் செல்கிறோம் என்பதை எளிதில் கணக்கிடலாம். இதனால் கிடைப்பது ஒரு கோலம் [1,2]

கோலம் என்பது ஒரு கணித வடிவம். இது பல புள்ளிகளையும், அவற்றில் சிலவற்றை இணைக்கும் (சில கோலங்களை படம் - 1 இல் காணலாம்.) ஒரு எழுத்திற்கான கோலத்தில் ஒவ்வொரு புள்ளிக்கும் உள்ள குணங்கள் ஒரு பதிவேட்டில் எழுதிவைக்கப்படும். இந்த மாதிரிக் கோலங்களில் சில புள்ளிகள் மட்டுமே இருப்பதால் இவற்றைக் கொண்டு கணிப்பது விரைவாக இருக்கும்.

2. கோலத்தைப் பிரித்தல்

அடுத்த நிலையில், எழுத்திற்கான ஒரு கோலம் பல பகுதிகளாகப் பிரிக்கப்படுகிறது. குறுகிய எல்லைக்குள் மேற்புறப்புள்ளிகளாக உள்ள புள்ளிகளை இரண்டாகப் பிரிப்பதன்மூலம் இது செய்யப்படுகிறது. இப்படிப் பிரிக்கப்பட்டவுடன் ஒவ்வொரு பகுதியும் ஒவ்வொரு கோலமாக மாறுகிறது. இந்தக் கோலங்களில் முதற்பகுதிக்கான கோலம் ஆராயப்பட்டு, அது எம்மாதிரியான எழுத்தின் முதற்பாகத்தைக் குறிக்கிறது என்பது கண்டறியப்படுகிறது. இவ்வாறான முதற் பாகங்கள் மொத்தத்தில் 5 மட்டுமே உள்ளன! இவை அ, ட, ர, ற, ே என்ற எழுத்துக்களில் உள்ள ஆரம்பப்பகுதிகளைக் குறிப்பவை.

முதற்பகுதி கண்டுபிடிக்கப்பட்டவுடன், இரண்டாம் பகுதியை அடையாளம் காண்பது துவங்குகிறது. முதற்பகுதியைப் பொறுத்த இரண்டாம் பகுதி குறிப்பிட்ட சிலவற்றுக்குள்தான் இருக்க முடியும். இரண்டாம் பகுதி என்னவென்று முடிவு செய்யப்பட்டவுடன் அடுத்த பகுதி ஆராயப்படுகிறது. இப்படி ஒன்று முதல் சுமார் நான்கு பகுதிகள் வரை முடிவு செய்யப்பட்டவுடன் எழுத்து புலனாகிறது. அறியப்பட்ட எழுத்து சரிதானா என்பதை மேலும் சில சோதனைகள் மூலம் உறுதி செய்து கொள்ள வேண்டும்.

எந்தெந்த முதற்பகுதிக்கு எந்தெந்த இரண்டாம் பகுதிகள் வரவேண்டும், அவை எந்த எழுத்துக்களுக்கு உரியவை என்ற தகவல்கள் படம் 2-இல் காட்டப்பட்டுள்ளன.

3. ஆராய்ச்சியின் இன்றைய நிலை

தற்போது எழுதப்பட்டு வரும் நிரலில் மேலே கூறப்பட்ட செயல்பாடுகளில் ஓரளவு செயல்படுத்தப்பட்டுள்ளன. கோலத்தை உருவாக்குதல், அதைப் பிரித்து பல கோலங்களாக்குதல், முதற்பகுதிகளை அடையாளம் கண்டு எழுத்துக்களை அறியும் வேலையில் சிறிதும் முடிந்துள்ளது. இதனால் தற்போது சுமார் 15 எழுத்துக்களை அடையாளம் காண முடியும். இதே வழியில் எல்லா எழுத்துக்களையும் அடையாளம் காண முடியும் என்பது இதனால் உறுதிசெய்யப்பட்டுள்ளது.

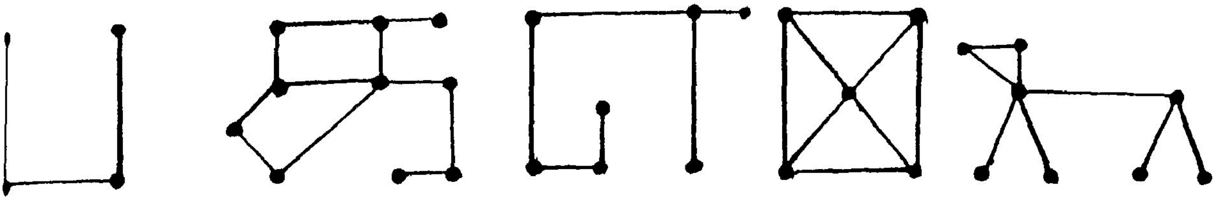
கையினால் ஒரு எழுத்தினை எழுதும் போது எழுத்தின் வடிவில் பல மாறுதல்கள் இருக்கும். கோலத்தை உருவாக்கும்போது இந் மாறுதல்களினால் ஏற்படும் வித்தியாசங்கள் களையப்படுகின்றன. ஆனால் தேவையான தகவல்கள் விட்டுப்போகாமல் கோலத்தை உருவாக்குவதுதான் இம்முறையின் நுட்பமாகும். ஒரு பகுதியைச் சரியாக்க கணிப்பதில் கையாளப்படும் முறைகளும், அதற்குத் தேவைப்படும்போது ஒரு கோலத்தை பல சிறு கோலங்களாகப் பிரித்து ஆராய்வதும் இந்த முறையின் மற்ற சிறப்பு அம்சங்களாகும்.

வளைந்த கோடுகள் உள்ள எழுத்துக்களை கோலமாக மாற்றி ஆராய்வதும், தமிழ் எழுத்துக்களுக்கு இம்முறையைப் பயன்படுத்துவதும் சிறந்த பயனளிக்கும் என எதிர்பார்க்கப்படுகிறது. தமிழ் எழுத்துக்களின் நீள அகலங்களில் உள்ள வித்தியாசங்களை ஒரு குறையாகக் கருதாமல், அதைத் தகவல்களை அதிகமாகத் தரும் ஒரு நிறையாகக் கருதுவதும் ஒரு சிறப்பு.

நூற்பட்டியல்

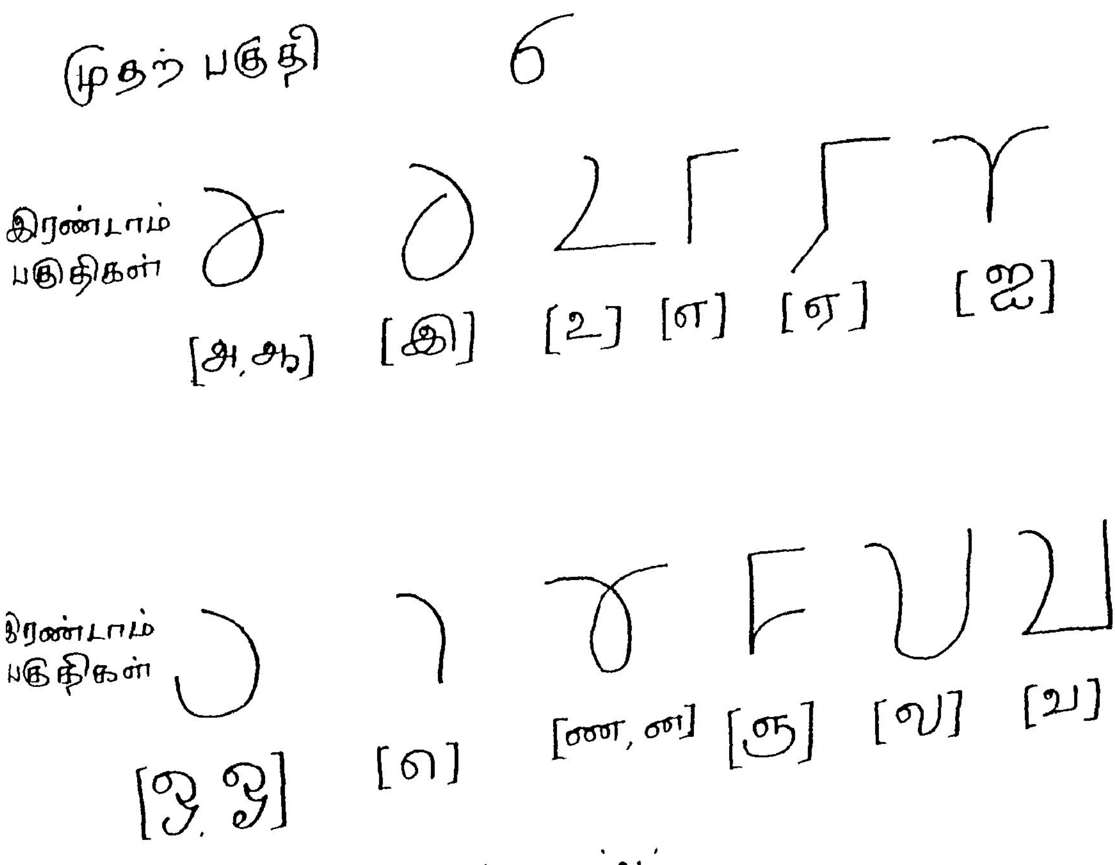
1. வெ. கிருஷ்ணமூர்த்தி, புதிர், சிந்தனை, கணிதம், நியூ செஞ்சுரி புக் ஹவுஸ், சென்னை, 1986

2. வெ. கிருஷ்ணமூர்த்தி, எஸ். கேசவன், கார்த்திகாதேவி, தமிழ் எழுத்துக்களை அடையாளம் காண்பதில் கோல இயல் முறை, தமிழக அறிவியல் பேரவை - முதற்கூட்டம், 1992

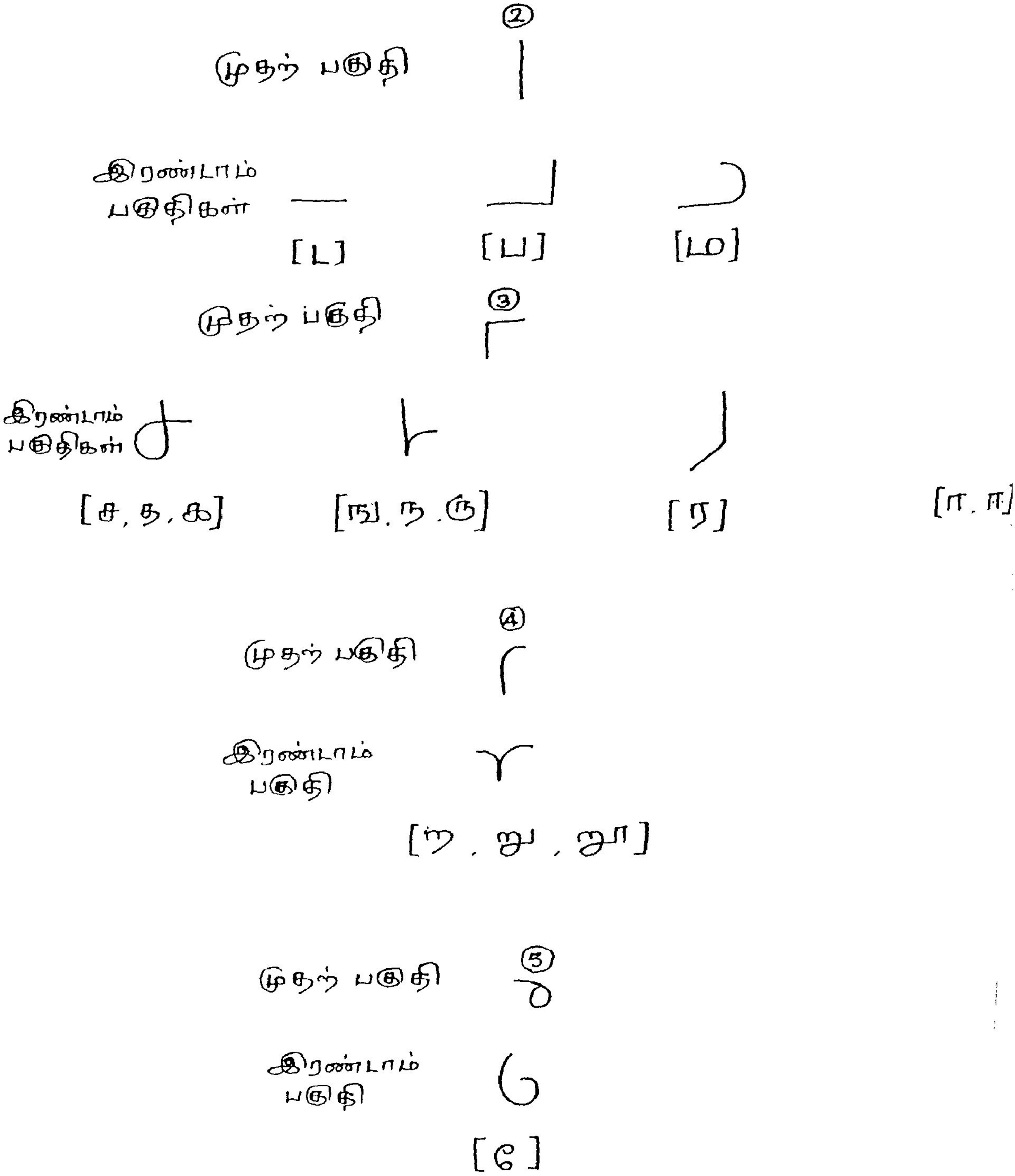


படம் I

①



படம் ௨அ


முதற் பகுதி
இரண்டாம் பகுதிகள்
[ட] [ப] [ம]
முதற் பகுதி
இரண்டாம் பகுதிகள்
[ச, த, க] [ஞ, ந, ரு] [ர]
முதற் பகுதி
இரண்டாம் பகுதி
முதற் பகுதி
இரண்டாம் பகுதி
[ே]

படம் உ 'அ'

NEED FOR STANDARDISATION OF TAMIL KEYBOARDS

M. MADHAN & A. SENTHIL
MCA Students
School of Computer Science and Engineering,
ANNA UNIVERSITY.

Tamil and computers - another wierd pair like `Dinosaurs and man'?

Subramania Bharathi says

"இவள் என்று பிறந்தவள் என்றுணராத
இயல்பினளாம் எங்கள் தாய்"

"உன் சீரிளமைத் திறம் வியந்து
செயல் மறந்து வாழ்த்துதுமே"

says Manonmaniyam Sundaram Pillai.

These lines puts forth the ever juvenile nature of Tamil. So why not Tamil and Computers be integrated for the development of each other?

One major problem that hurdles the integration is the layout of Tamil Keyboards. Here again, the question of 'Why the problem of layout?'. arises since for many years now there as been the existence of Tamil mechanical type writers. Then why not just use the mechanical keyboard layout? The answer is in an electronic device with speeds in Megahertz, the modus operandi of a mechanical typewriter layout is highly inefficient. The inefficiency is mainly due to the presence of dead keys on the typewriter.

When you want to generate characters with projections like கி, கு, கூ, பூ, etc., we have to type the projections, ி, ு, ூ etc first which come under the dead keys (i.e) the keys that do not advance the carriage and then type the character 'க' to finish the character, after which the carriage advances. This makes it clear about the unnecessary extra keys on the keyboard and the extra-carriage maintenance problems, that can be easily over looked by an electronic device.

In the conventional computer keyboard, it is a very common practice and even a highly preferable practice to use combinations for often used actions. We use the Ctl, Alt, and Shift Keys as much as we use the character keys. So why don't we use some comibinations to generate characters that need a dead-key pressing? This would make the carriage-maintenance easier or rather completely wipes out this problem.

Taking these details into account, the Department of Electronics (DoE) has prescribed a keyboard layout for Tamil. The Tamil Character set has 12 *உயிர் எழுத்துக்கள்* and 18 *மெய் எழுத்துக்கள்*. The remaining 216 (12x18) characters are generated as a combination of one *உயிர் எழுத்து* and one *மெய்யெழுத்து*. DOE's standard requires you to press a *அகர வரிசை உயிர்மெய்* followed by its counter part *உயிர் எழுத்து* to generate the *உயிர்மெய்* character. If need arises to follow a *உயிர்மெய்* with a *உயிர்*, there needs to be typed a'/' in between to disable the conjunction.

Inspite of the DoE's keyboard layout, there are many other keyboard layouts in use. 'MULTIWORD' from SONATA use the DoE's layout, but of the DTP Centres we visited, we ended up collecting a different keyboard layouts. The other layouts all share the property that, they have the projections keys separately. But each layout has a remarkable difference in placement of the keys. Further, there are people who use the Mechanical typewriter layout, backed by the argument that recruiting operators and DTP professionals gets easier this way and if an employee is trained on a new layout and for some reasons he quits, then the new employee also needs to be trained before he can take up the job.

As long as the number of Softwares for DTP in Tamil is limited, the layout difference may not matter much, but as the packages are increasing in number, the need for standardisation becomes an urgency. Even if the characters storage formats of all the packages are the same and one can edit the files created using another, it is not possible for the same user to work on the same document with different packages, if thus follow different keyboard layouts. Hence any DTP publisher has to either stick to a single package, or employ multiple operators for multiple packages.

The different criterion based on which a keyboard layout has to be designed are

(i) Frequency of usage of the character

(ii) Physical use for the user (i.e) giving more job to the right hand etc.

(iii) The grammar of the language (i.e) the rules dictating which character will not follow a character and which is most likely to follow etc.

Considering the different number of layouts we have come across, it is impossible for one to tryout all the layouts and decide on which satisfies the criterion given above. Also it is not easy to pick out the salient features of each layout and discuss. Hence it becomes necessary to design a new layout, which should be comfortable to use and the layout of which is to be declared as a standard, before the market is flooded with more and more incompatible keyboard layouts. In designing such a keyboard it also helps to take into account the difference in speed between the human fingers and the processor. Between two keystrokes, the processor is idle, which time can be used to do a fair amount of internal calculations in generating a character. Hence the keyboard should be designed such that the keystrokes are minimal for even complex characters and the generation is done internally even if some calculations are required.

One of the complex multilingual keyboard layout is supplied along with the Graphical Intelligent Script Technology Card (GIST). This card is able to translate between languages by following the sounds produced for each word/rather than following any characterwise relationship. For example 'ANNA' typed in English is being converted to 'அண்ணா' in Tamil directly by scanning the English text, without any further information being needed. However, it also supports direct Tamil keying in of characters, whose layout has all the projection characters separately and complex characters like 'கொ'/ 'கோ' are generated by the combination of 'ொ' and 'க' keys etc. By substituting other characters for 'க' the whole series is generated. The layout and usage is further complicated by the existence of Sanskrit characters like ஹ, ஸ்ரீ, க்ஷ etc. in Tamil. To generate it is required to type key followed by the 'ஹ' key which has no connection with the sound of 'ஸ்ரீ'. Also the உயிர் எழுத்துக்கள் are placed as upper case letters, that require the activation of the SHIFT key.

In the 'கணியன் விசைப்பலகை' layout designed by Dr.N. Govindasamy of Singapore, the rule that each Tamil word starts with a உயிர் எழுத்து is used as an advantage in character generation. After a spacebar is pressed this keyboard doesn't expect a SHIFT key but generates the உயிர் எழுத்து automatically. Dr.Govindasamy even forecasts that only this Singapore keyboard and the DoE's layout will survive in the market.

The ACES Trumpcard Tamil keyboard layout supported by VENTURA, uses a different layout, although its working is similar to that of the GIST keyboard, the character are placed in different positions. Also the ACES layout provoides all the Sanskrit Characters as separate

characters and doesn't need to be generated using any complex combinations. Due to these extra keyboard positions needed. some of the numeric keys are shared with the Tamil characters.

The Inscript Tamil keyboard layout from Softview computers follows the mechanical keyboard layout without any changes made and has to take care of dead keys and ordinary keys that advance the carriage automatically.

Another major difference arise in the layout designed by Software Research Group for their package 'BHARATHI'. In this layout the projection characters have been moved to the numberic keys and to the right half of the board while the other layouts have placed then at the Central rows in their left half.

The Monotype India (Ltd) DESKJET TAMIL Keyboard layout differs in that each மெய்யெழுத்து is given a full key unlike other layouts that use double key combinations to generate a மெய்யெழுத்து. Also these keys are shared by the அகரமேறிய மெய் யெழுத்துக்கள் when the SHIFT key is not activated. In this keyboard layout, the projection character have been moved down to the last row of the keyboard and characters like 'கொ' 'கௌ' etc. need three (3) keystrokes to be completed. This layout has even placed the much less used ஆய்த எழுத்து 'ஃ' at the central left hand side position.

The Keyboard layout for the package 'ANNAI TAMIL' has distributed the projection characters to the numeric keys, bottom row and right half of the middle rows. The layout also forces the user to type multiple keys for some characters like கௌ and கோ etc.

The simple numeric relation between the number of DTP packages looked at and the number of different layouts will make clear the absolute necessity for standardisation of the Tamil keyboard layouts and the need to alter the different packages as soon as possible to match the decided standard and avoid further complications in Tamil word processing which is a hurdle to the progress of Tamil word processing in Computers.

கணிப்பொறியும் தமிழாக்கமும்

SUDHA
Sri Venkateswara Engineering College

With the advent of Artificial Intelligence, Text translation Systems have gathered momentum. Henceforth, English to Tamil Text Translation systems are still no more a myth. This paper deals with a part of English to Tamil Text translation systems (i.e.) Translation of an English verb sequence into a Tamil word.

Initially, translation was done word by word, by consulting dictionaries. This consumed more time and space. The dictionaries were not only stuffed with english words and their equivalent Tamil forms, but also the word inflexions and their equivalent forms. Hence a better approach probably form a linguistic view, was sought after. Hence the need for a generalised translation of an English verb sequence into its Tamil equivalent.

The paper begins with the Text analysis of two languages: English and Tamil. The Text Analysis of the former is termed Morphological Analysis as it also deals with the Morpheme Identification of the English verb. The Text analysis of the latter is termed as Etymological Analysis by the linguists, as it deals with the various components involved in the formation of a Tamil word.

Apart from Text Analysis, the paper also deals with the implementation aspects of the translation system. Implementation can be discussed under two main modules namely, the Morphological Analyser and the Word Generator. The two modules are supported by a knowledge base structure termed the INFORMATION BASE - a key structure aiding translation process from English to any language in future !

TEXT ANALYSIS

Morphological Analysis

Morphological Analysis deals with the verb and its sequence. Morphological analysis, also discusses the morphemes, which form the smallest unit of any word. An English word is to contain one or more morphemes with or without affixes. Affixes may be of two types namely, prefix and suffix. Hence in general a Verb in English is said to have the following structure:

[Prefix (es)]   Morpheme(s)   [Suffix(es)]

Note that [ ] mean that the enclosures are optionals. The following table gives a generalised spelling rules for most common cases which are not the exceptions (took, got, ran). The notation of capitalised C stands for a part of the word. For example Cx refers to a word ending with x.

SUFFIX	TRANSFORMATION	EXAMPLE
-s	Cy -> Cie-	Lies, Flies (but pays)
	Co -> Coe-	Potatoes, Goes
-ed,-en	e- -> O	Tasted, Taken Oxen
-est,-er	Cy -> Ci-	Driest,Fastest, Hottest,Bolder,Faster,Drier

Example:

Decomposability

Morpheme I	:	Compose	Prefix	:	de-
Morpheme II	:	able	Suffix	:	-ty

Unhappy

Morpheme	:	Happy	Prefix	:	un-

Fastest

Morpheme	:	Fast	Suffix	:	-est

Now the information that can be retrieved from a given verb is limited. For example if the word itself is a root without any affixes, as in went, it means the past tense.

Since Tamil is a context-sensitive language, formation of the equivalent word require more information, in addition to the one extracted from the root. For example, consider the sentence **he WAS WALKING,** the factors retrieved from the verb **WALKING** *is that is a* continous form irrespective of the tense. But formation of Tamil word requires tense person, gender etc. Hence, its better to consider the verb sequence rather than the verb alone.

A clause contains atleast one verb and can include a complex verb sequence such as **would have been sleeping**, which conveys information about time & modality. The final word of this sequence is the main verb of the clause & one but the last have the feature - Auxilary. In transformational grammar there is a verb phrase (VP) which encompasses the verb sequence along with the objects & complements (Winograd 1983). In other words, a clause is described to contain a verb sequence made up of a main verb, auxilaries & modifiers.

The verb sequence has the marks of tense & other aspects of the clause, embedded in it. It is important to distinguish between the tense of an individual verb & of the clause as a whole, as determined by the verb sequence. Individual verbs includes modals, such as will, can and could (which have only one form), and ordinary verbs, which five basic forms INFINITIVE (go, be, write), PRESENT (GOES, IS, WRITES), PAST (*went, was, wrote*), PRESENT PARTICIPLE (going, being, writing) and PAST PARTICIPLE (gone, been, written).

The following is the pattern specifying a set order of the verb sequence with optionals in the parentheses.

(Modal)	**(Have)**	**(Be1)**	**(Be2)**	**Main verb**
will	have			gone
	has	been		writing
	had	been		corrected

The symbols HAVE and BE stand for forms of the verbs have and be. The verb sequence must have a VERB. Each element determines the form according to one of the following rules (Winograd 1983).

If the sequence contains	**The form of following element**
Modal	Infinitive
Have	Past participle
Be1	Present participle
Be2	Past participle

Given a complete verb sequence, it seems reasonable to ask what the tense of the clause as a whole is. There is no such thing as tense of the clause. There is only a sequence of tense.

Thus the factors that can be retrieved from a verb sequence can be classified in to two kinds : Root - Associated Factor and Auxiliary Factor.

For example:

Consider the sentence 'he WAS WALKING', the factors retrieved are as follows.

Root-Associated Factors

Root : Walk

Suffix : -ing

Tense : Past

Voice : Active

Number : Singular

Auxiliary Factors

Gender : Masculine

Person : Third

Etymological Analysis :

Etymology is an important branch of the linguistic science relating to formation of word & derivation of its meaning (Sykes 1983). In tamil we call it சொல்லியல். It is usually discussed in 3 parts namely classification, Inflection & derivation. Classification deals with the word classes & their nature. Inflection talks about the various changes which words undergo as they appeal in the sentences, while derivation describes the manner in which the individual words are formed and derived. Tamil words are categorised into four classes viz., Name (*பெயர்*), verb (*வினை*), Qualifying words (*உரிச்சொல்*) and particles (இடைச்சொல்). Of these the verb and its inflections are of our interest.

Inflection treats, changes which words undergo in order to express various relations and their meanings. Of the four parts of speech, the verb and noun are alone inflected. Figure(2) shows the schematic representaion of factors influencing the inflection of verbs.

A Tamil Word may be classified into two types namely, the simple words *(பகாபதம்)*, having indivisible roots which cannot be further divided into simple forms and compound words *(பகுபதம்)*, which are derived from the simple words. A compound word may have the following constituents namely, Root *(பகுதி)*, Termination *(விகுதி)*, Medial *(இடைநிலை)*, Sandhi *(சந்தி)*,Increment *(சாரியை)*.

A Compound word may be formed in the following manner. The elements in ( ) are optionals.

Root + (Sandhi) + (Medial) + (Increment) + Termination

The constituents of the Tamil word are termed as

Example

படி + த் + த் + அன் + அன் = படித்தனன்

Mood may be Imperative (*ஏவல்*), Optative *(வியங்கோல்)* or indicative. Indicative Mood may imply affirmative meaning or negative meaning. Considering the affirmative indicative mood, in which the verbs more are inflected by Voice, Tense, Gender, Number & Person. The twenty seven variations of the affirmative mood are thus distributed:

First person	:	2 Number * 3 Tenses	=	6 Variations
Second person	:	2 Numbers * 3 Tenses	=	6 Variations
Third person	:	3 Tenses * 5 Genders	=	15 Variations

Note that the genders classification includes the numbers.

Among the factors causing the variations the tense affects the middle particle of the word i.e. Medials *(இடைநிலை)* and the Gender affects the Termination of the word i.e *(விகுதி)*. The following tables depict the different tense and gender types with examples.

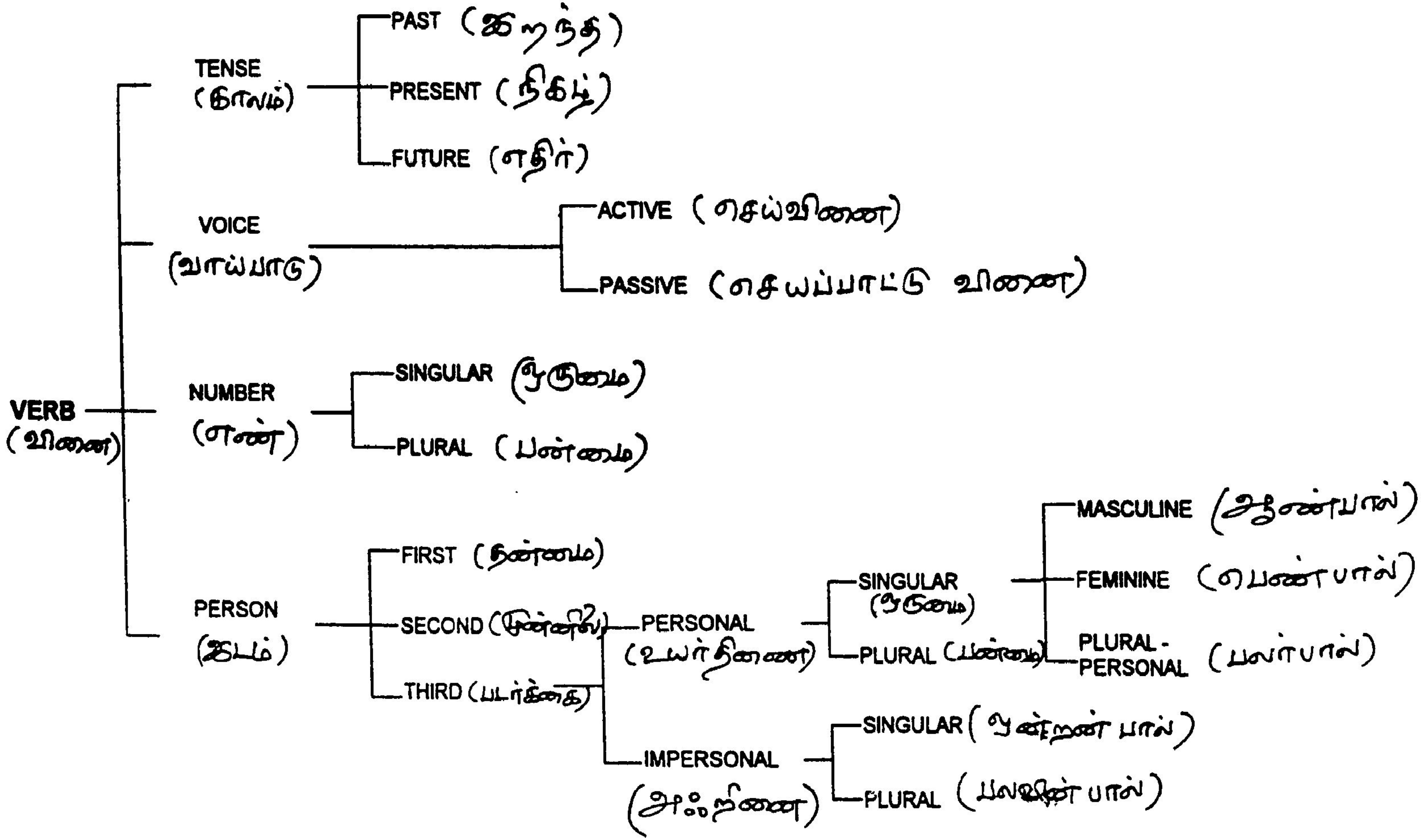


FIGURE 2 INFLECTION OF TAMIL VERBS

Tense	காலம்	Example
Past	இறந்த காலம்	படித்தான்
Present	நிகழ் காலம்	படிக்கிறான்
Future	எதிர் காலம்	படிப்பான்

TENSES AND ITS KINDS

Gender	பால்	Example
Masculine	ஆண்பால்	உண்டான்
Feminine	பெண்பால்	உண்டாள்
Plural Personal	பலர்பால்	உண்டார்கள்
Singular Impersonal	ஒன்றன் பால்	உண்டது
Plural Impersonal	பலவின் பால்	உண்டன

GENDER AND ITS KINDS

Having discussed the constituents of the word mentioned. The constituents are contined by a process called combination of letters (புணர்ச்சி), which may be natural combination (இயல்பு) or mutational combination (விகாரம்). In natural combination no change occurs in the union of final and initial letters as with பொன் + மணி = பொன்மணி

If the combination leads to the change in the word, then it is termed as Mutational Combination. Mutational combination is further divided into four types, based on the kind of the initial letter of the forth coming word (வருமொழி முதல்) and the final letter of the first word (நிலைமொழி ஈறு), as shown below

S.No.	நிலைமொழி ஈறு	வருமொழி முதல்	Example
1.	Vowel-final	Vowel-final	அது + ஐ
2.	Vowel-final	Consonant-final	பாட + சொன்னான்
3.	Consonant-final	Vowel-final	கல் + எறிந்தான்
4.	Consonant-final	Consonant-final	மரம் + கிளை

IMPLEMENTATION

The general flow structure of the Translation System is shown in following figure 3. As the figure depicts two modules have been implemented namely Morphological Analyser and Word Generator. The Morphological Analyser is an independent module that deals with English Language. The verb constituent is the input to the Morphological Analyser. The Morphological Information such as Tense, Voice, Person, Number, State are retrieved from the verb constituent which is stored as the INFORMATION BASE structure. The Word Generator takes the input from the Information base and generates the equivalent Tamil Word.

Two static structures, the LEXICON and the INFORMATION BASE are used in the process of translation. Translation is impossible without a dictionary. Only thing we can do is to frame a generalised rule as in this case, so as to reduce the number of words stuffed. Two dictionaries are used, one for the word and its exceptional forms and the other for the word and its equivalent Tamil forms. For example, in the former 'go went past' is stored, while in the later, go = போ is stored.

Morphological Analyser :

The function of the Morphological Analyser is to extract the semantic information from the given verb sequence. The input is the verb constituent, whose last token is always a verb. It finds the length of the verb sequence and identifies the last word as the verb or verb form.

The Morphological Analyser the does the token test based on the length. If the length is one, then the tense is the tense of the verb which can be only a simple past or present. The number and person are determined by the appropriate predicates while the voice is always Active and State a simple form. If the length is two then the initial word can be a modal which will determine the number, person, voice etc. For example in **am going** the first token **am** always represents present, first person, singular. In such cases as **was going** which is a past tense verb sequence, **was** may occur for first, second and third person which is left to the user to enter his choice depending on the context he refers to. Thus the length of the verb sequence can be used to select different modules to build the Information Base. The Analyser also takes care of the suffixes along with the other tokens of the verb sequence. The following figure gives the syntax of verb sequence of all possible lengths. Verb sequence can be of length 8 utmost. Those of length 7 and 8 always have their first token beginning with the Is-forms.

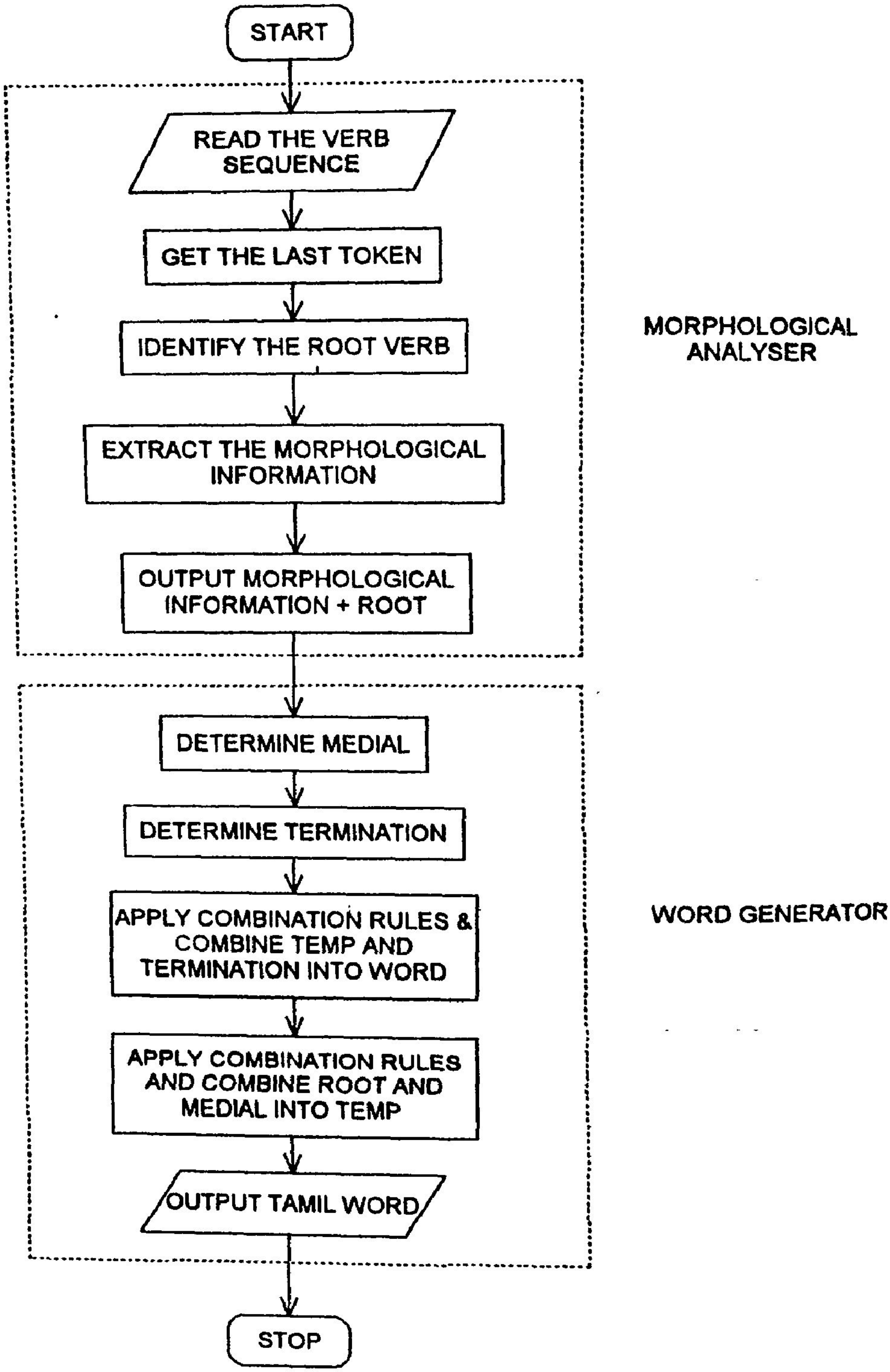


FIGURE 3 FLOW STRUCTURE OF TRANSLATION SYSTEM

Word Generator :

The Word Generator, being an independent module gets the input from the Information Base. The Information Base may be in the form of a file or can be directly from the keyboard. The factors which mainly influence the generation of Tamil word are the tense, number, person. The generation is carried out in four steps :

* **Determination of Termination**
* **Determination of Medial**
* **Combination of Root and Medial**
* **Formation of Tamil Word**

. Terminations are inflected by the person, number and gender. The medials change with respect to tense forms of the verb. The Sandhi and Medial are combined by means of casual combination. The combination of the Root with the result is based on the last character of the first word and the first character of the second word. The combination may be of following kinds :

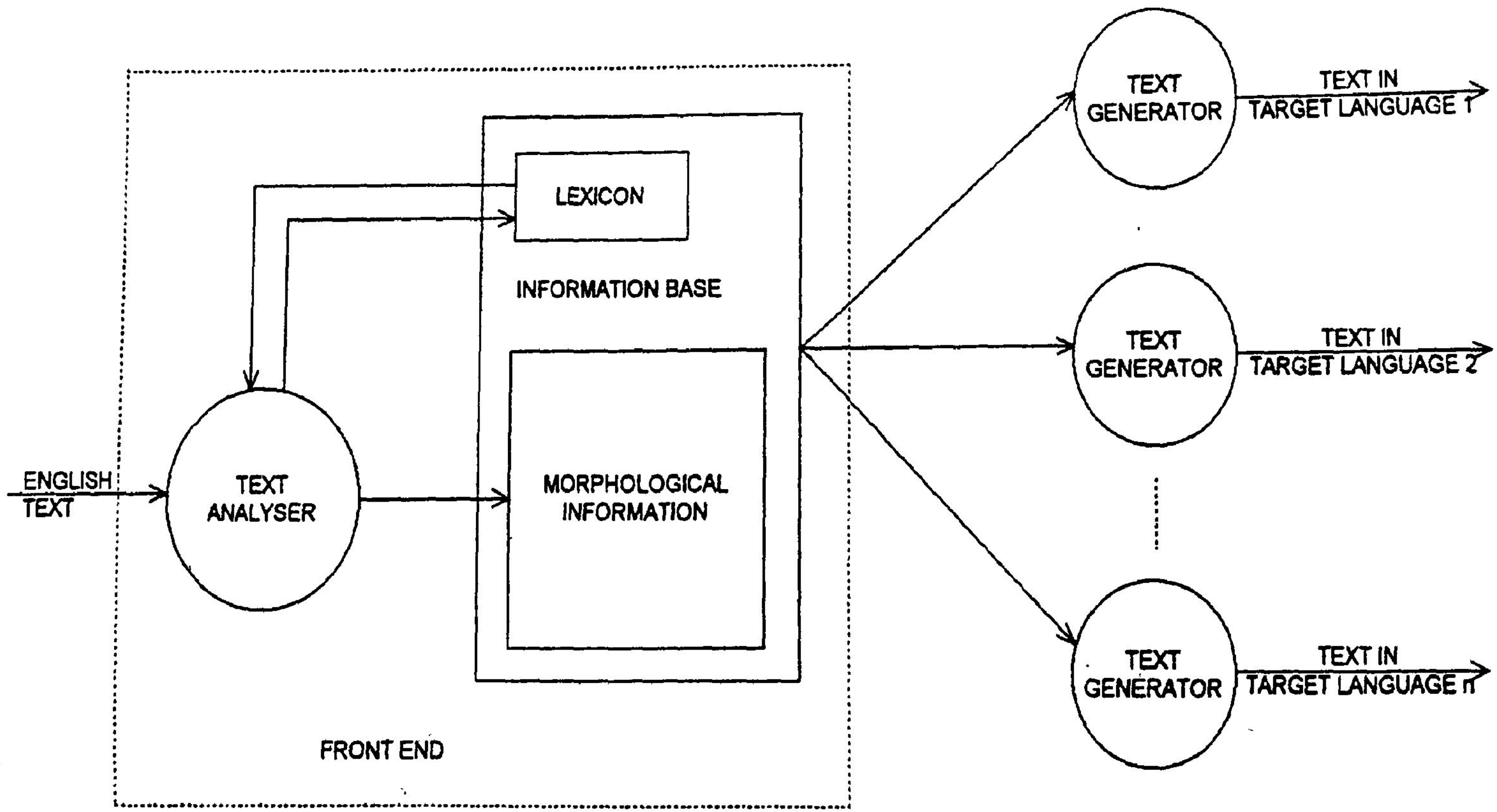
1. **Consonant Final with Consonant Initial**
2. **Consonant Final with Vowel Initial**
3. **Vowel Final with Consonant Initial**
4. **Vowel Final with Vowel Initial**

The combination of the termination with the rest of the word also based on the nature of the last character of the first word and the first character of the second word. Here only two kinds are possible namely the Consonant Final with Vowel Initial and Vowel Final with Vowel Initial.

CONCLUSION

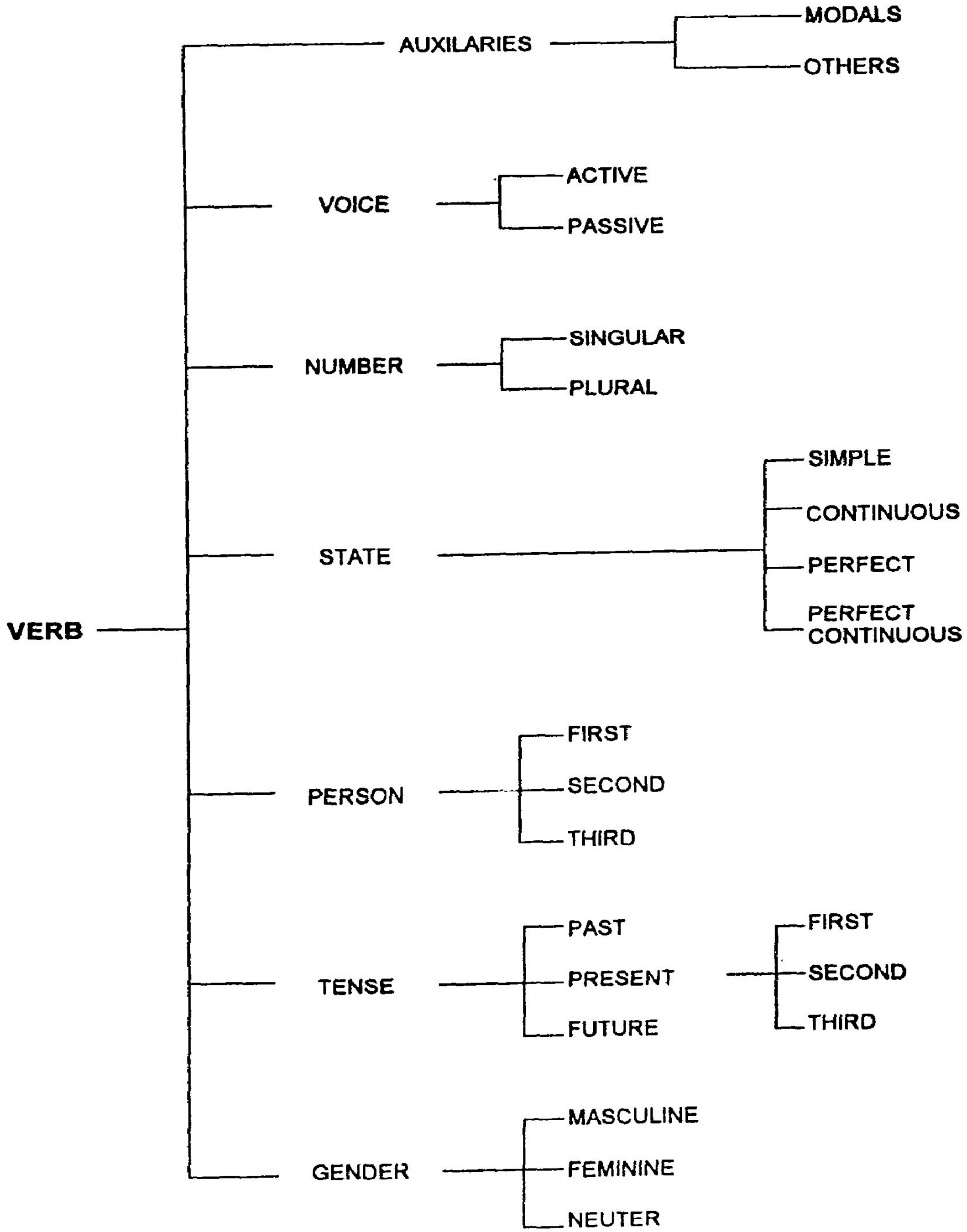
The paper has lit the fire of translation from English to Tamil, which has to be fuelled now and then, by improving it further. Future Works include translation of other parts of speech, apart from the verb and finally translation of the English Sentence in to a Tamil Sentence, which is no further than our reach !

APPENDIX I



OUTLINE OF BASIC MODULES

APPENDIX II



SYSTEM NETWORK FOR ENGLISH VERBS

LENGTH SYNTAX

APPENDIX III

1. Verb [nil]
   Verb [d/ed]
   Verb [form]

2. Future-form + Verb [nil]
   Modal + Verb [nil]
   Have-form + Verb [d/ed/nil]
   Is-form + Verb [ing]
   Is-form + Verb [d/ed]

3. Future-form + Have + Verb [d/ed/nil]
   Future-form + be + Verb [ing]
   Have-form + To + Verb [nil]
   Have-form + Be-form + Verb [ing]
   Is-form + Being + Verb [ing]

4. Is-form + Verb [ing] + To + Verb [nil]
   Modal + Have + To + Verb [nil]
   Future-form + Have + To + Verb [nil]
   Have-form + To + Be + Verb [ing]
   Future-form + Have + Been + Verb [d/ed/nil]
   Future-form + Have + Been + Verb [ing]

5. Is-form + Verb [ing] + To + Have + Be-form + Verb [ing]
   Future-form + Have + To + Have + Be-form + Verb [ing]

Syntax of verb sequence

APPENDIX IV

Factor No. / Kind of Word	1	2	3	4	5	6	7	8	9
Noun	Gender Masculine Feminine Neuter	Number Singular Plural	Root Quality Action State	Case Accusative Nominative Possessive	Type Proper Pronoun Abstract Common	Common Count Mass	Form Personal Impersonal		
Verb	Voice Active Passive	Number Singular Plural	State Simple Continuous Perfect Perfect Continuous	Person First Second Third	Tense Past Present Future	Gender Masculine Feminine Neuter	Form Personal Impersonal		
Adjective	Quantity True False	Number Singular Plural	Quality True False	Demonstrative True False	Interrogative True False	Comparison Positive Comparative Superlative			
Adverb	Quantity True False	Time True False	Manner True False	Frequency True False	Place True False	Comparison Positive Comparative Superlative	Affirmative True False	Reasoning True False	
Pronoun	Gender Masculine Feminine Neuter	Number Singular Plural	Action Reflexive Emphatic	Demonstrative True False	Case Nominative Possessive Accusative	Person First Second Third	Relative True False	Interrogative True False	Form Personal Impersonal
Preposition	Place True False	Time True False	Manner True False	A-I Agency Instrumentality	C-R-P Cause Reason Purpose	P-C-C Possession Contrast Concession	M-S-R-V Measure Standard Rate Value	I-M-S-O Inference Motive Source Origin	
Article	Definite True False	Indefinite True False							

Information Base

TENSE INDENTIFICATION AND ANALYSIS OF TAMIL VERB PHRASES USING NEURAL NETWORKS

B. Chandra
Department of Mathematics
Indian Institute of Technology
Hauz Khas, New Delhi - 110 016

and

B. Krishna Murthy
Department of Electronics
Government of India
Electronics Nikethan
New Delhi - 110 003

Abstract

Neuro Computing technology is very useful to develop computer systems for applications where there is little information of how to develop/design an algorithm. Natural Language Processing is one such domain to test the potential of this emerging technology. In this paper an attempt has been made for the tense identification and Analysis of Tamil verb phrases. Verb phrase analysis is an important issue in Natural Language Processing. A multi-layared neural network has been designed for the classification of Tamil words Traditional Back-propagation algorithm has been used to train the neural network.

1. Introduction

Ever since the term Artificial Intelligence (AI) had been coined in as early as 1950, AI has become one of the frontiers of computer technology. Scientists have been trying to solve various socio-economic problems using this technology and developing various systems for different applications. As a result of this research, various areas of AI have been evolved Natural Language Processing (NLP) is one of the areas [1] of AI which is still an active area of research in many universities and labs in India and abroad. NLP, which is also refered as computational linguistics, brings computer scientists and linguists into one arena. The tasks before these two distinguished specialists are a) Machine Translation b) Natural language interface to database access c) Spell checkers d) Electronics lexicons etc.

In a multilingual society like ours, it is necessary to be able to communicate in more than one language with equal ease. In a country like India, where there are as many as 18 official languages and more than 10 scripts for these languages, human communication is really a difficult task. To overcome this language barrier, NLP research could be of great help.

NLP has recently emerged as a major area of research. This research concerns with syntactic, semantic and pragmatic analysis of phrases/sentences of the language. Parsing of the sentences, which involves analysis of sentences and generate its syntactic stuructures, is a primary stage in an NLP system. A good NLP system requires an efficient parser.

Morphological Analyser [1,8] is one of the basic modules in any Parser. A single meaningful unit of any language is called Morpheme. Morphological analysis of a language concerns with the study of the formation of various declensions from the root-words or morphemes. The versatality of Indian languages is that most of them are agglutinative languages i.e. the Case markers, Parsargs and many other added to the root words or morphemes to generate the declensions. These suffixes are embedded into the word itself. This complicates the morphology of a language [3]. As in many Indian languages the declensions and root words are different words. Storing all these words is redundant and reduces the efficiency of the parser. A strong morphological analyser should be capable of identifying the case markers, root words and also the declensions from the same word. Verb-phrase analysis is a part of morphological analysis [8], which deals with the identification of tense, gender, number, voice etc.

In this paper, we have made an attempt for the verb-pharse analysis. We have taken Tamil as language for our implementation. Tamil is an ancient Indian language like Sanskrit. Tamil is also a free-word order and phonetic language. We have used neurocomputing technology [4,5,6] for our implementation. Neuro computing offers distinct advantage over conventional computing where there is little information of how to develop an algorithm and where there is ambiguous or missing data. A multi-layered feedforward network has been designed for the identification of the gender, tense number etc. for the Tamil verbs.

The rest of the paper is organised in the following manner: Section 2 gives an introduction to neural networks, Section 3 presents some of the linguistic features of Tamil language and the linguistic aspects of Tamil verb phrases, section 4 gives the design of the neural network, its training & implementation, and section 5 gives discussions & conclusions.

2. Introduction to Artificial Neural Networks

Artificial Neural Network (ANN) is an information or signal processing system composed of a large number of simple computing/processing elements called Artificial Neurons [4,5] which are interconnected by direct links. One of the features of ANNs is their capability to adapt themselves to a new environment by changing their connection weights. Eachcomputing element has several inputs but will produce a single output which is equal to the weighted sum of the input signals. The single output can be fanned out to other neurons as input in the network. The final output is a complicated signal which is related to the initial input signals and weights of the network connections. The weights are adjusted using training algorithms [2,5,6] to obtain the desired output. Even complicated networks, which are designed to perform difficult tasks, may even consists of several layers neurons. Such networks are called multi-layered networks. Various research studies, carried out regarding the number of hidden layers, reveal that a single hidden layer is sufficient [8]. Such a network is depicted in the following figure.

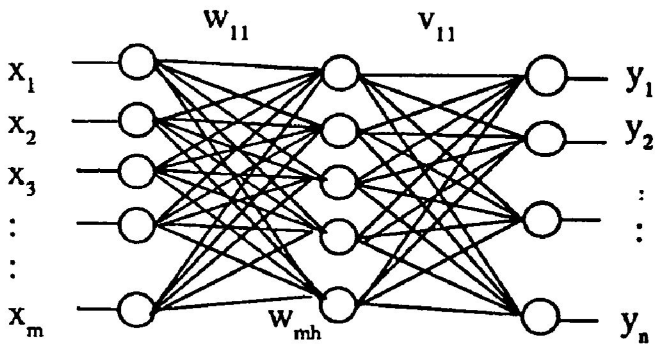


Fig. 1: Muti-layered artificial neural network

In a typical multi-layered network (MLN) let the input signals be x_1^{μ}, x_2^{μ} ..x_i^{μ}..x_m^{μ} and y_1^{μ}, y_2^{μ} ......y_k^{μ} ....y_n^{μ} be the output signals for the μ th (where $\mu = 1,2,.....$) input-output pattern in the training set. The weights of the connections between the input and hidden layers be w_{11}, w_{12}...... v_{ij}.......v_{mh} and the weights between hidden and output layer v_{11}, v_{12}...... v_{jk}.......v_{hn}

Given a pattern μ, each hidden unit receives a net input:

$$I_j^{\mu} = \sum_i w_{ij} x_i^{\mu} \qquad (1)$$

and produces an output

$$h_j^{\mu} = f(I_j^{\mu}) = f(\sum_i w_{ij}\, x_i^{\mu}) \qquad (2)$$

Where f is called transfer function or activation function which is used to limit the output values. The sigmoid function is often chosen as the transfer function i.e.

$$f(x) = 1/(1+e^{-x}) \qquad (3)$$

Thus each output unit receives

$$O_k^{\mu} = \sum_j v_{jk}\, h_j^{\mu} \qquad (4)$$

and produces a final output

$$y_k^{\mu} = f(\sum_j v_{jk}\, h_j^{\mu})$$

$$= f(\sum_j v_{jk}\; f(\sum_i w_{ij}\, X_i^{\mu})) \qquad (5)$$

If D_k^{μ} (where k = 1,2,.............) are the desired outputs then the error measure or the cost function could be

$$\Gamma(w) = \Sigma [y_k^{\mu} - D_k^{\mu}]^2 \qquad (6)$$

A network is trained so that application of a set of inputs produces the desired output. Learning is accomplished by sequentially applying the input vectors, while adjusting the network weights according to a predetermined procedure called learning algorithm. During this process the network weights gradually converge to values such that the network can perform the desired task. Training or learning of such networks is of two types: namely, supervised and unsupervised. In supervised learning the desired an input vector is applied, the network output is compared with the desired output and the difference is fed back and is used to adjust the weights. This process is repeated for all the vectors in the training set and until the error is minimum. In the unsupervised learning mode, the training set consists of only input vectors. In this mode the input vectors are applied and the work is allowed to stabilize while adjusting the weights.

Backpropagation algorithm [2,6] is one of the best algorithms which scientists have applied for supervised training. The algorithm gives a prescription for modifying the weights in any feed forward multi-layered networks to learn a set of input-output vector pairs. The error between the desired outputs and actual outputs will be used to modify the weights. This algorithm uses the gradient descent rule which says that the change in the weights is proportional to the negative of the derivative of the least mean sqare error as measured on the current input-output pattern. As the output error is fed back to modify the weight parameter, this algorithm is called as error-backpropagation or simply backpropagation algorithm.

3. Some linguistic features of Tamil language

Tamil is one of the ancient Dravidian languages. Tamil has about 35 alphabets(13 Vowels and 22 consonents). These are coded by English alphabets for ease of representation as shown in the table 1. A syllable can be a vowel (v), consonent (c) or combination of these (cv&ccv). Like many other Indian languages Tamil,is a free-word order and phonetic language. A word is formed from a root word or a morpheme by the addition of Parsargs, tense tags, or other tags to generate declinsions. In most of the cases, these tags are embedded into the word itself. However, in certain cases the verb phrases may even contain two or three words. The verb phrase can give the information regarding the tense & Voice of a sentence; gender, number, person of the subject in the sentence etc. The following are the different forms for these classes.

Category			No.of types
Voice:			
	Active(A)		
	Passive(P)		2
Tense			
	Present(Pr)	simple(Si)	
	Past (Pt)	continuous (Co)	
	Future (Fu)	perfect (Pf)	
		perfect continuous (Pc)	12
Gender			
	Masculine (M)		
	Feminine (F)		
	Neutral (N)		3
Number			
	Singular(S)		
	Plural (Pl)		2
Person			
	First person (Fp)		
	Second person (Sp)		
	Third person (Tp)		3

Any system for the analysis of verb phrases must give the above information. Consider the following examples:

I teach lessons	nAn pA tn k L	*so lli ko tu kki rEn*	
	(I lessons	teach)	
we teach lessons	nAn k L pA tn k L	*so lli ko tu*	*kki rOm*
She is teaching lessons	a v L pA tn k L	*so lli ko tu*	*i Ri kki rA L*
You have taught lessons	nin k L pA tn k L	*so lli ko tu*	*ti Ri kki rir h L*

Notice that the verb phrases in the above sentences shown in underline, are our main concern. The input-output patterns for the above four verb phrases in the system could be as given below.

Input phrase	Output classifications (Voice Tense Gender No.Person)
so lli ko tu kki rEn	A PrSi M S Fp
so lli ko tu kki rOm	A PrSi M Pl Fp
so lli ko tu i Ri kki rA L	A PrCo F S Tp
so lli ko tu ti Ri kki rir h L	A PrPf M Pl Sp

4. Design & Implementation of the System

The design of a multi-layered net work involves finding out the number of neurons in each of the three layers(Input, Output & Hidden) of the network assuming that there is one hidden layer. Obviously the input to the network is the total verb phrase. Hence the maximum number of words in a verb phrase will decide the number of input neurons. The output should be the tense, person, gender, voice and number. These classes decide the number of neurons in the output layer.

A verb phrase may consist of at most 3 words, without loss of generality we can assume that there are atmost 10 syllables in a word. Hence the maximum number of syllables in the verb-phrase would be 30. As a Tamil syllable can be represented by 15 bits (binary digits). The input layer of the network should consist of 450 neurons.

As the total number of classes in Tense, gender, number, voice and person are about 22 classes, 22 neurons are sufficient in the output layer. But we would like to access the word/ phrase also in the output layer. Hence a vocabulary index for each word has been introduced. The vocabulary index gives the position of the verb phrase in our lexicon or the training set for the neural network. We expect to train the network for about 2.5 lakhs of Tamil-verb-phrases that includes the root words and their declinsions. Hence 18 bits have been allocated for this vocabulary index. Therefore the total number of neurons for the network is 40. There is no specific formula for determining the number of neurons for the hidden layer. Various experiments reveal that to achieve good performance of the neural network, the hidden layer must contain the average number of input/output neurons. The hidden layer in this network consists of 250 neurons.

The back-propagation algorithm has been implemented in C language on a PC AT/486 computer with 50 MHZ clock speed. The training network requires about 25 minutes to train 100 Tamil verb phrases. Though a lot of improvements are required to increase the performance of the system and speed for training, the main advantage of this approach is that all the words/verbphrases are stored in terms of the weights of the network and network is capable of recalling the appropriate outputs given the proper inputs.

5. Conclusions

A novel approach for the identification of Number, gender, tense etc. for Tamil verb phrases have been presented. Though the system has been developed for Tamil language only, it can be easily tailored to any language without any major modifications in the configuration of the network. Neuro-computing is a promising tool in many areas of information technology. Back-propagation algorithm has been widely applied to many applications in Pattern Classification/Recognition, Optical Character recognition, Signal/ Speech/Image Processing etc. In this paper an attempt has been made to apply this new technology in natural Language Processing. NLP is a fertile area to test the potential of artificial neural networks.

TABLE 1: Representation of Tamil Characters

Vowels

Tamil character	:	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ	ஃ
English repr.	:	a	A	i	I	u	U	e	E	Y	o	O	V	H

Consonants

Tamil Character	:	க	ங	ச	ஜ	ஞ	ட	ண	த	ந	ன	ப	ம	ய	ர	ற	ல	ள	ழ	வ	ஷ	ஸ	ஹ
English repr.	:	k	f	c	j	F	t	N	T	n	M	P	m	y	r	R	l	L	Z	v	s	S	h

REFERENCES

1. Allen J. "Natural Language Understanding" Benjamin/Cumings Pub. Co.Inc.1987
2. Blum A. "Neural Networks in C++: An object oriented Framework for building Connectionist Models" Pub. Willey Professional Computing 1993.
3. Chandra B. & Krishna Murthy B. "Neural Networks for Parsing of Indian Languages" **AKSHARA 94**: International Conference on Information Technology Applications in South Asian Languages 25-26th Feb. New Delhi 1994.
4. Chandra B.& Krishna Murthy B. "A Neural Network Approach for Indian Language Processing" J. of Neural Computing - to appear.
5. Hopfield J.John "Artificial Neural Networks" in IEEE Circuits and Devices Sept. 1988 pp 3-10.
6. Lipman P. Richard "An Introduction to Computing with neural Networks" in IEEE ASSP Apr. 1987 pp 4 to 22.
7. Sengupta P. & Chaudhuri B.B."Natural Language Processing in Indian Language (Bengali)-I:Verb Phrase Analysis" **IETE** Technical Review Vol.10 No.1 Jan-Feb 1993 pp 27-41.
8. Villiers J. de & Barnard E. "Backpropagation Neural Nets with One and Two Hidden Layers"IEEE Trans. on Neural Networks Vol.4 No.1 jan.1992 pp 136-141.

THE POSSIBLE IMPACT OF COMPUTER AS A COMMUNICATION TOOL ON PRESERVATION AND FURTHERANCE OF TAMIL LANGUAGE IN THE MODERN ERA.

V. Gnanaprakasam*, S. Jayasundar and T. Mohanasundari****

Tamil Nadu Veterinary & Animal Sciences University
*Vice chancellor **Bioinformatics Centre

ABSTRACT

This seminar on "Tamil and Computer" is practically an attempt to link two important communication tools namely Tamil language and computer. This presentation suggests creation of databases on historical and modern literary works in Tamil. These databases should cover scholars' names, literary works, year, their publication details, major literary awards, and awardees. All these literary works, past and present, should obtain ISBN numbers for easy access. Another important area, this presentation emphasises, is on the Tamil equivalents of Scientific and technical words. Whatever be the source of the Tamil technical words, a specified agency formed by appropriate authorities should database these technical words, remove redundancies, authenticate their usage, integrate them into CD-ROMs and international networks for ready and wide accessibility.

The age of computer is entering into a new phase namely communication. The computer has now become a synonym for communication. For ages, we have been accustomed to another tool of communication namely, language. Amongst languages, Tamil is one of the oldest and the only one which is the earliest language being practiced in this modern era. This seminar is rightly marriaging the two important tools of communication namely the Tamil language, a qualitative tool and the computer, a quantitative tool. No doubt, this marriage is going to be mutually beneficiary.

Having a wide experience of involving computers in databases creation, retreival and dissemination and scientific communication worldwide, we wish to underline a few areas unexplored so far. The basis for this presentation are a few leaves from the current status of scientific literature dissemination which are largely transforming from print medium to electronic medium.

Tamil Nadu being the Principal homeland of Tamil language, this state of India has got the atmost duty of creation of database on all the historical literature starting from "Tholkappium". We should not leave behind the modern literature; hence they should also be included in these databases. These databases should include names of all Tamil poets, Scholars, Writers, their works, addresses, major and other ancillary awards available for Tamil literary works and Tamil promotion and names and addresses of awardees.

The authors suggest here that a agency or a conglomeration of agencies is to be created to monitor development of above mentioned databases in compact discs.

Tamil equivalents of scientific and Technical words are being created in various disciplines by a variety of individuals or groups ranging from the scientists of a particular discipline to common Tamil scholars. There are chances for variants, overlappings, redundants,etc. The agency suggested above should be made mandatory for all publishing houses, large or samll, to send information on all publications all the technical words created should be cut into compact discs for easy access, for putting them into network and for automatic elimination of redundant technical words.

This presentation looks largely at the possible impact of computer as a communication tool on Tamil. The authors have deliberately, avoided the subject of usage of Tamil in computer world such as Tamilisation of common softwares viz dBase, BASIC, Spreadsheet etc which, it is assumed, would be taken up elswhere. The possibility of computers lending helping hand to Tamil development is enormous. The above suggestions are only thetip of icebergs. However these could be a start and these by themselves, are works in enormity to be executed with utmost care, devotion, scientific accuracy and thoughts for the sake of our dear language, Tamil.

TAMIL LANGUAGE FRONT END TO INGRES

S. CHITRA
T. DEVI
K. SARUKESI
Department of Computer Science
Bharathiar University, Coimbatore-46.

ABSTRACT

The construction of a computer system which understands natural language is an important enterprise. Such systems enable people to converse easily with a computer system and extract useful work from it without the restrictions imposed by the use of any programming language.

Natural language front end is one of the fastest growing area in natural language Processing (NLP). This paper deals with the development of a Tamil front end to INGRES. Graphics and Intelligent based Script Technology (GIST) is to be made use of to recognize Tamil scripts.

Tamil input text will be given using a GIST card. The front end to be developed will convert the input text into database queries. These queries when given to the Relational Database Management System - INGRES, will produce the required output.

INTRODUCTION

Language is a system for encoding and transmitting ideas. The history of communication between man and machine has followed a path of increasing provision for the convenience and ease of communication on the part of the human. From raw binary through various symbolic, assembly, scientific, business and higher level languages, programming languages have increasingly adopted notations that are more natural and meaningful to a human user.

Natural Language Processing (NLP) is one of the fastest growing area in the field of Artificial Intelligence (AI). It allows the computer to communicate with the world in natural language. by this it is easier for the unscrupulous person to invade the privacy of computers. It is extremely elusive to develop a fully functional natural language processing system. This paper deals with the methodology of developing a Natural language Front End (NLFE) to an RDBMS.

For a naive, inexperienced user, almost every transaction with current computer systems requires considerable mental effort deciding how to express the request in machine language. hence there is a need for Natural Language Processing.

Indian language usage in computers should not be more difficult than that for English, otherwise the usage will get affected. A Tamil-to-English, database, an English-to-Tamil database, a spelling database, along with grammatical information and Phrase/Idiom databases could be made use of for natural language processing. The usage of Indian scripts in computers would get encouraged by ensuring that the script remains as easy to use as English.

HUMAN LANGUAGE AND COMPUTERS

Human languages must be efficiently understood by a computer while processing them. Understanding is the process of interpreting an input and assigning it meaning. It is important to keep in mind that the success or failure of "understanding" a program can rarely be measured in an absolute sense but must instead be measured with respect to a particular task to be performed.

The human linguistic communication occurs by means of speech and written languages. The entire language understanding process has two steps:

1. Written language understanding.
2. Spoken language understanding.

Written text is processed, using lexical, syntactic, and semantic knowledge of the language as well as the required real world information. Spoken language understanding requires all the facilities of a written language as well as enough additional knowledge to handle all the noise and ambiguities of the audio signal.

PROBLEMS IN UNDERSTANDING HUMAN LANGUAGES

1. The complexity of target representation.
2. Type of mapping.
3. Level of interaction among components of the source representation.
4. The presence of noise in the input.

FEATURES OF LANGUAGE THAT MAKE IT BOTH DIFFICULT AND USEFUL:

1. No natural language program can be complete because new words, expressions, and meanings can be generated quite freely.
2. In a natural language the same expression could mean different things in different contexts.
3. There are lots of ways to say the same thing.
4. Natural language sentences are incomplete descriptions of the information that they intended to convey.

STEPS IN THE PROCESS

The process of natural language understanding can be broken down into the following pieces.

1. Morphological Analysis.
2. Syntactic Analysis.
3. Semantic Analysis.
4. Discourse Integration.
5. Pragmatic Analysis.

Morphological Analysis

To construct the appropriate representation for a sentence, it is necessary to understand each of the words in the sentence and to put these words together to form a structure that represents the meaning of the entire sentence.

Understanding of words involves looking up each word in a dictionary (lexicon), which contains entries of words in target representation for the meaning of the word, and determining the correct meaning of an individual word. This is called word sense disambiguation or lexical disambiguation. This requires the use of the other words in the sentence which can serve as part of the context to decide the particular meaning of the word.

Syntactic Analysis

Syntactic processing is the step in which a flat input sentence is converted into a hierarchical structure that corresponds to the units of meaning in the sentence. This process is called parsing.

To parse a sentence, a grammar that describes the structure of the string in a particular language should be used. Given such a grammar, a parser can assign a structure to each grammatical sentence it parses. This structure is called a parse tree.

Syntactic analysis plays an important role in many natural language understanding systems for the following reasons :

- Semantic processing must operate on sentence constituents. Parsing constrains the number of constituents that semantics can consider.
- Syntactic processing is computationally less expensive than semantic processing. Thus it can play a significant role in reducing overall system complexity.
- Syntactic parsing of a sentence helps in extracting the meaning of the sentence.

Semantic Analysis

Semantic analysis involves assigning meaning to the structures created by the syntactic analyser. In other words, it must map individual words into appropriate objects in the knowledge base. The language into which mapping is to be done should be defined first. There are two categories:

- When natural language is being considered as a phenomenon on its own.
- When natural language is being used as an interface language to another program (such as a database query or an expert system), then the target language must be a legal input to the other program.

Discourse Integration

To understand even a single sentence, it is necessary to consider the context in which the sentence was uttered. The meaning of an individual sentence may depend on the sentences that precede it and may influence the meanings of the sentences that follow it. There are a number of important relationships that may hold between phrases and parts of their discourse contexts, including:

- Identical entities.
- Parts of entities.
- Parts of actions.
- Entities involved in actions.

Pragmatic Analysis

Pragmatic analysis involves reinterpretation of structures to determine what was actually meant. For example, the sentence "Can I know the time ?" should be interpreted as a request to be told the time. Pragmatic analysis is the final step towards effective understanding and it is only here that it is decided what to do next.

"It is fuzzy to find out the boundaries between these five phases. The phases are sometimes performed in sequence and they are sometimes performed all at once."

NATURAL LANGUAGE FRONT END

Among various applications of NLP, Natural Language Front End is one of the fastest growing areas. NLFE is a user friendly interface which will accept a user's command in natural language and performs the specified operation on a computer. The natural language allowed will be restricted, but it will be sufficient to illustrate queries on the specified application.

The natural language (input text) input will be translated into an internal text (internal representation to perform necessary operation). The input text is constrained to confirm to a strict and limited grammar and the meaning of the text is represented in the form of internal text. The front end can be developed for many applications and this paper deals with the development of front end to INGRES-RDBMS.

TYPES OF NATURAL LANGUAGE FRONT ENDS

The NLFE can be developed in the following methods:

1.The system interrogates the user via a menu of options.

2.The user interrogates the system by asking questions in a database query language.

The first strategy is simple and the user does not require any special knowledge of the structure of the database. But it is inflexible. The second strategy is flexible. But it requires some structural knowledge about the database.

NLFE TO DATABASE PACKAGES

A computer database stores information about some domain. The front end allows access to the database by using a natural language. Here the input text will be converted into Database Query language (DBQ - internal text).

Therefore the meaning of the input text is identified by means of the converted internal text. If the statement cannot be translated into a statement in DBQ language then it is considered to have no meaning. The conversion of input text to internal text is represented in the following figure.

Fig 1. Natural Language Front End

THE FUNCTIONAL STRUCTURE OF NLFE

The following figure shows the front end and the back end structure of the system.

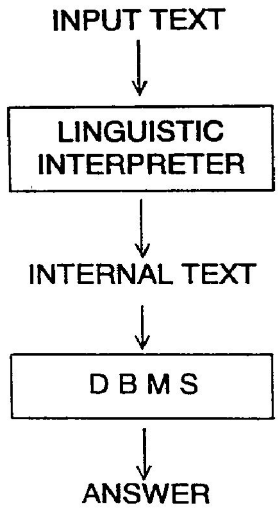


Fig 2. Overall Structure of NLFE

The front end determines the lexical structural meaning of input and maps this onto terms and structure of database query language. The back end is DBMS package and it executes the internal text. The front end structure has three phases: the syntactic processor supported by a grammar and lexicon, the semantic processor supported by domain model and lexicon and the conversion processor supported by a database schema. The following figure shows the front end structure.

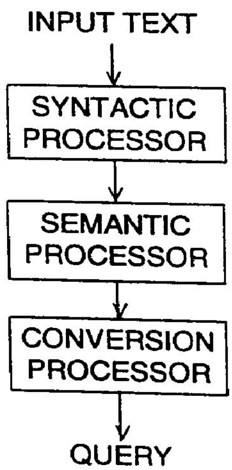


Fig 3. Block Structure of NLFE

TAMIL LANGUAGE FRONT END TO DATABASE PACKAGES

The Tamil front end will allow a common man, to access the computer, without having any special training on computers. Processing Tamil language on computers requires a system that supports the processing of Tamil scripts. The GIST card which has been developed by C-DAC can be used to process Indian scripts on computers using any computer language.

DEVELOPMENT OF TAMIL NLFE

In order to develop a NLFE, the following processes should be carried out :

-- GIST processor can be made use of, which allows the user to enter the input queries in Tamil.

-- The output from the GIST processor will be inputted to the translator, which will convert the input text into DBQ language (internal text).

-- The DBQ will be given as input to the database system.

-- The output of the database system will be given to the GIST card to produce screen output.

The above functions of the NLFE has been represented diagrammatically below:

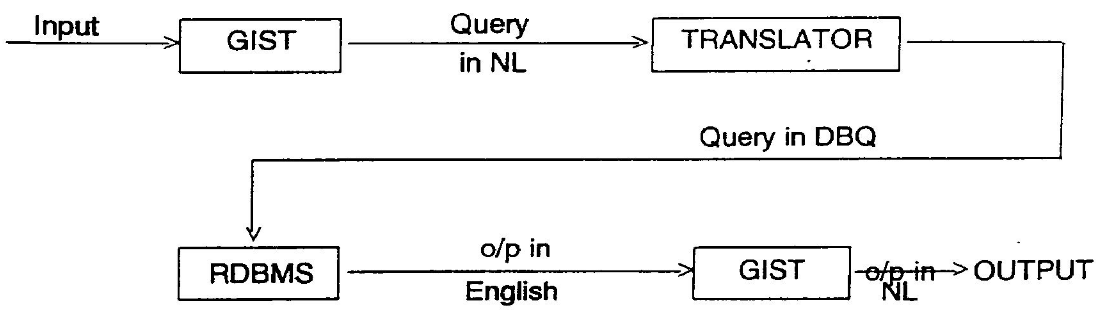


Fig 4. Functional Structure of NLFE

ADVANTAGES

Some of the advantages of using NLFE are :

1. The user of the application program needs only to know what the application is capable of doing. The user need not know how to make application program. Hence it reduces over-head involved in the process.
2. The users population for an application with a NLFE will expand tremendously. This is because, the users require much less "Computer Technical Skills" in order to make use of the application.

CONCLUSION

If Language Front End is implemented, the computer medium would become language independent. Indian languages because of their phonetic nature, *may even turn out to have* an edge over other languages of the world.

BIBLIOGRAPHY

1. **AKSHAR BHARATHI, VINEET CHAITANYA, RANEEV SANGAL**, Natural Language Processing for Linguistics, Vol. 1, A computational Framework for Indian Languages IIT, Kanpur, pp 2.1 - 2.3, 1990.

2. **BANSAL R K**, An outline of General Phonetics, Oxford University Press, London, 1971.

3. **CHRISTOPHER BUTLER**, Computers in Linguistics, Basil Blackwell Inc. New York, USA, pp 12 - 35, 1985.

4. **EUGENE CHARNIAK**, A Common Representation for Problem-Solving and Language - Comprehension Information, Artificial Intelligence, Vol. 16, pp 225 - 255, 1981.

5. **JERROLD KAPLAN S**, Cooperative Responses from a Portable Natural Language Query System, Artificial Intelligence, Vol. 15, North - Holland Publishing Company, pp 165 - 187, 1982.

6. **JOHN EDWARDS**, Language, Society and Identity, Basil Black well in Association with Andre Deutsch, pp 3 - 10, 1947.

7. **LOUIS E, FRENZEL J R**, Crash Course in Artificial Intelligence, Howard W Sams & Co., A division of Macmillan, Inc. USA., pp 145 - 182, 1987.

8. **PATNAIK B N, RAJEEV SANGAL**, NLP for Linguistics, Vol. 2, Some Computational Models of Languages, IIT, Kanpur, pp 7.2, 1990.

9. **PATTANAYAK D P, THIRUMALAI M S**, An introduction to Tamil Script Reading & Writing, Central Institute of Indian Languages, Mysore, 1980.

10. **RAJEEV SANGAL**, Machine Translation - Language Processing, Science Today, 2001 (January), pp 22 - 24, 1989.

11. **SUBRAMANIAN N**, ISFOC - Tamil, M.C.A. Project Report, Bharathiar University, 1991.

12. **VICTORIA FROMKIN, ROBERT RODMAN**, An Introduction to Language, Holt - Saunders International Editors, pp 54 - 59, 1983.

INTELLIGENT TUTORING SYSTEM FOR TAMIL GRAMMAR

S.Sai Lakshmi
T.Devi
K.Sarukesi
Department of Computer Science,
Bharathiar University, Coimbatore - 46.

Abstract

Recent trends in Computer Science clearly indicate a *strong impact of AI technology in* the utilization of Computers to effectively process Indian languages. In that context, this paper presents *an overview of an Intelligent Tutoring System model for teaching* the basic concepts of the Tamil language Grammar. Computer analysis of Tamil in recent years have been handled only from the language translation point of view and hence the primary focus of this paper is on teaching the sentential constructs of Tamil language, based on the age-old standard Tamil grammar literatures.

A simple Tamil sentence comprises of noun phrases, a verb phrase and optionally, the adjective and adverb phrases. The present formalism is based on the concept of grammatical relations according to a standard set of rules The proposed system consists of a knowledge base consisting of base rules and description of words in the lexicon which put together *described a sentence in the language.*

The tutor is to be designed in such a way that it operates in two modes; the teaching and the evaluation mode. The system needs to assess the student's understanding by the performance of the student in solving prescribed exercises related to specific grammatical constructs. This feedback is then used by the tutor to guide the student in his subsequent learning of the language.

1 Introduction

In today's Artificial Intelligence (AI) scenario, research is primarily motivated by the promise of building powerful teaching systems with greater knowledge about a domain, increased ability to make inferences about student behaviour, and increased reasoning ability about topic selection and response generation. The main goal of such a research is to use AI and cognitive science techniques to model tutors for teaching and students for learning in that domain. Computational models facilitate a high degree of explicitness about learning theories and teaching strategies that is difficult to attain in class rooms.

Computer programs that use artificial intelligence techniques to help a student learn are called intelligent tutoring system(ITS). ITS is an application of AI technology to education. The primary goal of ITS is to improve the quality of education and training. The present paper deals with an intelligent system for teaching/learning Tamil, based on Tholkapiam, the standard grammatical literature. The study of a language starts primarily with the structure of a simple sentence, based on the morphological (word formation) level, syntactic (structural) level, semantic (meaning) level, and pragmatic(contextual) level. Each of these levels has its own set of rules (grammar) which are often augmented with constraints.

2 Morphology of Tamil

Ancient Tamil Grammarians have regarded only Noun and Verb as the two basic component parts of speech in Tamil. Hence a simple Tamil sentence includes noun phrases, verb phrases and optionally, adverb phrases. A phrase is either a category word (such as noun, verb etc.) or a category word followed by other phrases and its type is the type of the head word. Thus, there are noun phrases, adjective phrases, adverb phrases and verb phrases. The process of breaking a sentence into phrases and phrase into words and constructing a sentence out of the phrases and phrases out of words is known as **parsing**.

A **word** is either morpheme (meaningful unit in a language) or combination of morphemes e.g. puthakam, puthakangal. The process of combining the morphemes to get a word or breaking a word into morphemes is called morphology.

2.1 Structure of a Sentence

Breaking a sentence down into its component parts is a vital step in getting the computer to understand and teach the language. The rules of grammar and syntax must be translated into a form that the computer can handle. An artificial intelligence system designed to manipulate such a rule-based knowledge in a particular domain is called an expert system. The grammar proposed by Tholkappiar is essentially a rule based system. Hence this can be easily implemented as an expert system.

The core of the Intelligent Tutoring System (ITS) is the rule-based expert system, wherein the inference mechanism guides the reasoning process through the knowledge base by attempting to match the facts in the database to the rule condition. The rules proposed to describe the structure of a Tamil sentence are typically the following:

SENT 1	->	AD_VP SENT2/SENT2
SENT 2	->	SUB AD_VP /SEQ AD_VP
SUB	->	AD_JP NP / NP
AD_VP	->	AD_VP VP / VP

In the above rules symbols SENT1 and SENT2 denote a sentence. AD_VP, SUB, AD_JP,VP,NP denote an adverb phrase, a subject, an adjective phrase, verb phrase and a noun phrase respectively. SEQ denotes the specific sequence of phrases but have no grammatical significance.

For example,

```
vegamaga        odugiraan
    |               |
  AD_VP             VP

raman   nanraaga   padikiraan
  |        |           |
  NP     AD_VP         VP
```

2.2 Noun Phrase and Verb phrase

A noun phrase is either a noun, a pronoun or a noun preceded by a modifier, which can either be an adjective phrase, an indicative pronoun, a numeral, a participle or a noun with the combination of these. The verb phrase, consisting of a verb root controls the presence and role of the noun phrases. An adverb phrase is either an adverb,(representing time, place, direction, manner, reason, purpose, instrument) or a phrase derived from a noun or a verb phrase.

3. Organisation and use of knowledge

The knowledge base of the ITS consists of a set of rules described for each phrase, noun, verb, adverb and adjective. The knowledge stored in the expert system can be used for generating various types of word forms, sentence types and exercises.

A Computational lexicon consisting of the terminal symbols of the grammar (lexical items) can be implemented to store descriptive (morphological, inflectional and semantic) information. Thus the proposed system consists of a knowledge base consisting of base rules and description of words in the lexicon which put together describe a sentence in the language.

The other knowledge modules of the system are:

★ The lessons module
(Reasoning about Teaching Knowledge)

★ The evaluation module
(Reasoning about Student knowledge)

Lessons may be placed on a prerequisite-post requisite hierarchy. A student can be allowed to access any lesson in the hierarchy after assessing the performance of the user in the prerequisites of the lesson. The student's performance in a lesson can be the function of scores obtained by the student in exercises associated with the lesson.

The tutoring system should provide the student a user-friendly environment. The student can converse with the system in the learning mode or in the evaluation mode.

3.1 Learning Mode

In this mode, the student first selects the type of word formation - whether verb, noun or other phrases. Once the necessary input is given, the complete information relating to the word formation stored in the knowledge base should be displayed on the screen. Thus the system must guide the student in the sentential constructs of the language step-by-step.

3.2 Evaluation Mode

In the evaluation mode, the system should generate exercises pertaining to the lesson the student has learnt. The performance of the student in solving these exercises determines the level of learning. After ascertaining the student's performance, the system should point out the mistakes he has made and display the descriptive information on the screen for that particular word formation from the lexicon. The system thus needs to support individualized evaluation to assess the performance of the student.

The proposed architecture of the ITS is shown below

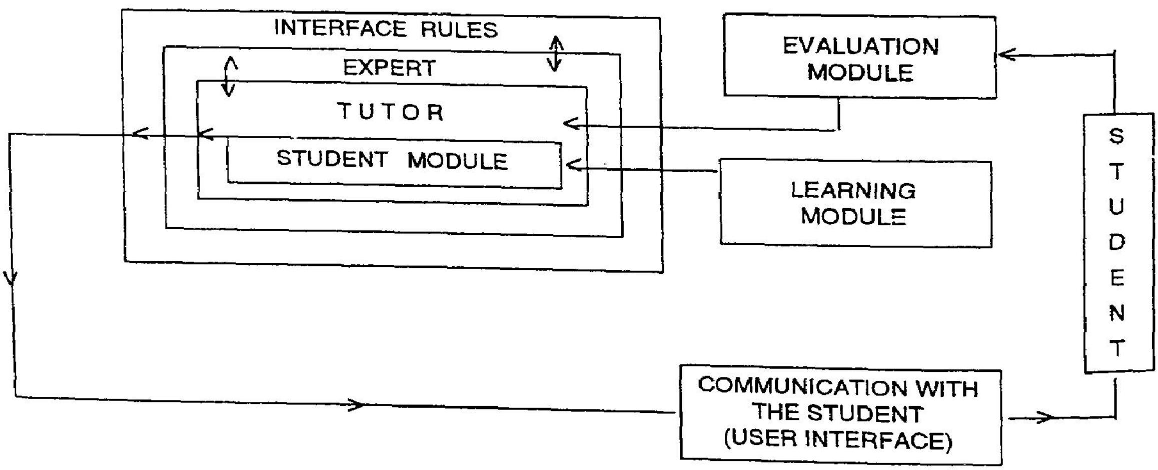


Fig .1 Architecture of ITS

4 Conclusion

This paper presents the methodology to implement an intelligent tutoring system for teaching Tamil grammar. The proposed system compared to other conventional instruction system is expected to facilitate a far more comprehensive and complete learning with elaborate descriptive information about each rules and exhaustive practice exercises. The present grammar accounts only for the set of rule describing the simple sentence formation in Tamil language according to Tholkappiam. This idea can be further enhanced to account for all the grammatical rules already available in Tholkappiam which will computerize the teaching of the entire language.

References

1. **Arun Kumar B and Sarukesi K**
   "A model to maintain user level in CAL using KBS", Department of Computer Science, Bharathiyar University.

2. **Arun Kumar B and Sarukesi K**
   "Flexible personalized Intelligent Tutoring System", Department of Computer Science, Bharathiyar University.

3. **Balaguruswamy E and Sushila B**
   "Artificial Intelligence Technology - Applications and Management", Proceedings of the 4th International Computing Congress, December 15-18, 1993, Hyderabad.

4. **Dr. Manickam V**
   "A Study of Tamil Verbs",Annamalai University.

5. **Stuart C Shapiro**
   "Encyclopedia of Artificial Intelligence", Second Edition, Volume 1, Wiley - Interscience Publication.

SPELLING THE DIFFERENCE
GIST Tamil SpellCaster

BALAJI. V.

Introduction

With the advent of desk-top publishing and word processing in Indian Languages, it is essential to have a spelling checker to complement the facilities and reduce the tedium associated with proofing a text. Thus, a spell-checker's primary function is to verify the CORRECTNESS of a word, no matter what language it is!. With this idea as the back-bone, GIST Tamil SpellCaster has been developed.

Assumptions

Certain rigid assumptions are made at the definition stage of the GIST Tamil SpellCaster. The SpellCaster will not operate at the sentence level. It will find only errors in the spelling of a word. Commonly written forms of Tamil are considered to form the basic dictionary.

Who will use the SpellCaster & Why?

Whoever wants to spell-check a document can use it. This big user domain includes Publishers, Editors, Proof-readers, Authors for their Editorial work and even a common man for personal documentation. The need for the usage of SpellCaster is to get rid of all spelling errors at the word level. To have a consistency of writing in documents.

Future trends of the SpellCaster

The basic SpellCaster which has the hard core dictionary and all the language specific rules built-in can emerge into a Thesaurus. This need not be specific for a language alone but can be interfaced with other similar GIST based Thesaurus to have a cross-referencing of words in all Indian languages. It can also serve as a Language Teaching Tool for children. Computer Based Training (CBT) Packages can be developed which will help an individual to improve one's proficiency in the language. With Natural Language Processing (NLP) concepts in it, a language Grammar Parser can also be derived at, thereby a context sensitive checker of similar kind can be thought of.

The Linguistic approach

Development of the SpellCaster comprises of the listed steps to be followed.

Part - A : Casting the Spell - treats of the corpus design.

Part - B : Spelling (out) the Caster - procedures to mould the corpus and interface it with GIST based word processors.

Casting the SPELL

To treat the corpus design, written forms in Tamil are considered. This process involves the Language Standardisation and Language Description. Language standardisation is essential to a certain extent and is still in the evolving phase and hence represent a large number of irregularities and inconsistencies. It is indispensible that some viable norms be prescribed which will rule out such inconsistencies and ensure that the language represents a certain consistent whole. Very often these norms are suggestions made by apex bodies within the respective itself and from proceedings of conferences in regional languages.

Simultaneously, the need for a basic lexicon also makes itself felt. This lexicon has to be a representative of the language as it is written and used today ie., current praxis of the language. Once the norms and the choice of words is determined, the language has to be exhaustively defined to ensure that its morphology is correctly represented. To implement, the traditional grammar in computer certain compromise has to be done. No strict grammer rules are applicable when a well-defined and categorised corpus is ported to computer in "computer-understandable" form. Hence this phase takes a considerable amount of time for its initial outcome. The steps involved in the basic design of Corpus is listed below.

- Parts of speech in Tamil are identified, categorised and their behaviour is studied

Basis	Part of Speech	Example
Pluralisation	NOUNs & PRONOUNs	பெண் → பெண்கள் மரம் → மரங்கள்
Declension	NOUNs & PRONOUNs	பெண் → பெண்ணை ஆண் → ஆணை மரம் → மரத்தை
Conjugation	VERBs & Auxiliary VERBs	செய் → செய்கிறேன், செய்வேன் வா → வருகிறேன், வருவேன்
Standalones	Adjectives, Adverbs	
Order of Precedence	POST-POSITIONs & Auxiliary VERBs	ஐ, ஆல், ஒடு, கூட

- All Sandhi rules (External & Internal) which account to any morphological change at the word level are listed down.
- Suitable oblique and normal forms for the identified sub-categories taking into account, the listed sandhi rules are formed.
- Appropriate tags for the oblique and normal forms are assigned.

- All rules which permit the derivation of new forms from the base words are defined. These rules are all fitted within a table which permits the generation of admissible and only valid forms, termed as the **GENERATION TABLE**

Once the tags are all assigned and the Generation Table is built-in, then is the process of building the dictionary in binary form which is "computer-understandable".

Spelling (out) the Caster : Software Implementation

1) To get the list of words in a PC-ASCII file which will contain only the raw unprocessed Tamill words. This can be done automatically by culling words from running text also. (Program : TMCULWRD)
2) Based on set of rules for categorisation, all parts of speech are grouped. (Program : TMPOSCAT)
3) The grouped parts of speech are tagged. For this an exhaustive pre-defined tag table is built-in with the program. This tag table will assign the right tag which can follow a particular suffix path. (Program : TMPOSTAG)
4) Given the suffix precedence list, a suffix path generator will generate all possible suffix combinations. (Program: TMSUFCMB)
5) To eliminate the suffix combinations which are not commonly used an user-friendly system is developed. Linguists can browse the combination entries and remove the invalid ones from the combination using this program. (Program : TMVALSUF)
6) To incorporate the Sandhi between any two suffixal elements, a separate utility which will read the suffix combination list and output the sandhi-applied string is developed. (Program : TMSNDWRD)
7) Suffix path are identified for each base suffix. This is done manually as it is extremely language dependent. Once this is done, the same path can be applied for their derivatives. This process is automated. (Program : TMSUFPTH)

After these steps are carried out, Two separate files - TAMILPOS.LST and TAMILSUF.LST will be available which makes the derived corpus for the Spell Checker Engine to build the "computer understandable" dictionary. The process by which the SpellChecker Engine builds the dictionary is out of the scope of this article. Once the dictionary is built, this can be accessed through the GIST based APEX LANGUAGE PROCESOR (ALP).

Extensions of the SpellCaster

Current version of the SpellCaster is interfaced with APEX Language Processor which is available in DOS and UNIX platforms. However, the SpellChecker will be ported to MS-WINDOWS. Research work is on in achieving this, so as to enable the wide range of users of GIST based products to have the zeal of spell-checking in Windows applications also !

Constant updating of the word list will be carried out by apex bodies in the region which are identified and authorised by Language Tools Team, GIST Group, Centre for Development of Advanced Computing, PUNE 411 007.

Courtesy : Dr. Raymond Doctor, Language Tools Team, GIST Group,
Centre for Development of Advanced Computing, PUNE-411 007.

RECOGNITION OF HANDWRITTEN TAMIL CHARACTERS USING NEURAL NETWORKS

T. Paulpandian and V. Ganapathy
School of Computer Science and Engineering
College of Engineering, Anna University
Madras - 600 025 INDIA

ABSTRACT

A high accuracy two stage neural network recognizer for handwritten Tamil characters independently of their position, size, and writing style is presented. The new system utilizes Walsh coefficients of the central portion of characters to partition characters into four groups and moment invariants of the outer portion to classify characters in each group. The two stage system also provides an excellent expansion capability for large - volume Tamil characters recognition.

1. INTRODUCTION

Recognition of handwritten characters has been the subject of intensive research for more than three decades. The goal of handwritten character recognition research is to develop machines that have the same recognition capability of human. Various pattern recognition techniques have been applied in the past to handwritten character recognition. But there still remain many problems in the recognition of handwritten characters due tc the large degree of deformation exhibited in the data.

Usually, recognition of handwritten characters requires the capability of the recognizer to generalize, abstract and learn tolerances with various deformations. Various neural network models have shown some or all of these properties. In the past few years, neural network approaches have been applied for the recognition of handprinted Chinese characters [1], and handwritten alphanumeric characters [2]. Very little work has been done on the recognition of characters of Tamil which is one of the ancient languages of India and more than 5000 years old. The Tamil characters are more complex in structure and larger in number compared to any of the European languages. Chinnuswamy and Krishnamoorthy [3] have described a linguistic approach for the recognition of handprinted Tamil characters. Paulpandian and Ganapathy [4] have developed a hierarchical neural network scheme for translation and scale invariant recognition of handwritten Tamil characters in conjunction with a modified crossing counting technique for feature extraction. This paper describes a two stage neural network recognition system for handwritten Tamil characters in conjunction with a hybrid feature extraction technique utilizing Walsh transform and moment invariants.

There exists a substructure which exhibits similarity in the central rectangular portion of a group of Tamil characters. The central portion of a Tamil character contains the most valuable information to partition the characters into different groups. We choose two dimensional Walsh transform to partition the Tamil characters into different groups. Huang and Chung [5] have used Walsh transform for classifying complex Chinese characters. The sequency property of Walsh transform is more suitable for the analysis of the central portion of the Tamil characters as shown in Fig.1.

In the second stage, the characters in each group are classified using the moment invariant features [6] extracted from the remaining portions of each character. The classifiers in each stage are trained by using the backpropagation learning algorithm [7].

2. INPUT PATTERN

The Tamil alphabet has thirty basic letters of which eighteen are consonants and twelve are vowels. In addition there are 216 combinations of consonants and vowels and a special symbol, and this brings the total to 247 combinations. In practice some of these combinations are not used and the printers have types for only about 125 symbols. The Tamil characters can be divided into a number of groups, most characters in each group having the same characteristics in the central rectangular portion. In our experiment, we use four groups of Tamil characters (see Fig.2), each group consisting of 3-7 characters for a total of 22 characters.

3. THE TWO STAGE RECOGNITION SYSTEM

The proposed two stage system for the recognition of handwritten Tamil characters invariant under scale, translation, and writing syle is described in this section. The schematic of the two stage recognition system is shown in Fig.3. The first stage consists of one neural network module which classifies the 22 Tamil characters into four groups using the Walsh transform of the central rectangular portions of all the characters. Each neural network module in the second stage is trained to recognize the characters in a group using the moment invariant features of the remaining portion of a character. Each module in the system is a two-layer feedforward network with one hidden layer trained by using the backpropagation algorithm.

3.1 Walsh Transform of Tamil Characters

We use Walsh coefficients as features representing the central rectangular portion of Tamil characters. The reason for extracting the central portion of the input character for use of Walsh transform is that most Walsh coefficients are stable under changes in size and position of the characters. Further, Walsh transform involves only 1 or -1 operation and thus is a fast computation.

Walsh functions $Wal(m,\Theta)$, $-\frac{1}{2} \leq \Theta \leq \frac{1}{2}$ are defined recursively as follows:

$$WAL(0,\Theta) = \begin{cases} 1 \text{ if } -\frac{1}{2} \leq \Theta \leq \frac{1}{2} \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

And

$$WAL(2j + q,\theta) = (-1)^{[-j/2]+q} \, [WAL\,(j,2\,(\theta + \tfrac{1}{4}\,))$$

$$+ (-1)^{j+q}\, WAL(j,2(\theta - \tfrac{1}{4}\,))],\ q=0 \text{ or } 1,\ j=0,1,2\ldots\ldots, n) \quad (2)$$

And if Θ is not in $(-\frac{1}{2},\frac{1}{2})$, then $WAL\,(j,\Theta) = 0$. The ordering of Walsh functions is sequency (Walsh) ordering; $WAL\,(0,\Theta)$, $WAL\,(1,\Theta)$, $WAL\,(2,\Theta)$, $WAL\,(3,\Theta)$,........

Let the input character image be matrix $[y_{ij}]$ i=1,......N; j=1,....,N, and let the origin of the coordinate system be translated to the center of the image. We extract the central rectangular portion of the input image, which contains $N_1 \times N_2$ (truncated to integers) submatrix centred at the origin where N_1=N/2 and N_2 = N/1.4. We extract N_1 equally spaced points from interval $[-\frac{1}{2},\frac{1}{2}]$ and substitute them into Θ of $WAL(m,\Theta)$ given above to get sequence of values $WAL\,(m,i)$, i=0,1,2,.......N_1-1. Similarly we extract N_2 equally spaced points from $[-\frac{1}{2},\frac{1}{2}]$ and use them to get sequence of values $WAL\,(n,j)$, =0,1,2,.........,N_2-1.

Now the extracted central portion is renamed as [Xij], i=0,........, N_1-1; j=0,..., N_2-1. Then the two dimensional Walsh transform of [Xij] is

$$C_{mn} = \frac{1}{N_1 N_2} \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} X_{ij}\, WAL\,(m,i)\, QAL\,(n,i), \quad (3)$$

where C_{mn} is called Walsh coefficient and m=0,...N_1-1; n=0,....,N_2-1. It is to be noted that since X_{ij} and Walsh functions have values 1 cr -1 the calculation of C_{mn} can be implemented in TTL "AND OR" circuits so that the computational speed can be very fast.

Very high sequencies are not realistic in Tamil Characters, so we truncate those C_{mn}'s with m 6≥ and n ≥6. Hence we have 6x6=36 coefficients as features. Theoretically C_{mn} should be invariant under size and position changes and small variations due to writing style.

3.2 Moment Invariant Features of Tamil Characters

Moment invariants have been used as features in a wide range of pattern recognition applications. A set of nonlinear functions $\phi_1,\phi_2,\ldots\ldots\ldots\phi_6$ which are invariant to rotation, translation, and scale change has been derived [6]. Khotanzad and Lu [8] utilized them for classification of English characters.

The recognition of English characters using moment invariants is rather simple due to the existence of symmetry in their geometrical structures [9]. It is difficult to use the moment invariant features of the full images to classify the Tamil characters due to structural resemblance among a group of characters. However, the moment invariant features obtained from the outer portion, other than the central portion, of the input character have the most separability power to classify each character in each group. Therefore, we choose the moment invariants, ϕ_1 through ϕ_6 of the outer portion of the input image as features to train the neural network modules in the second stage. Another reason for extracting the outer portion of the input image for use of moment invariants is that the central portion of the image has been used by Walsh transform in the first stage of recognition process.

4. SIMULATION RESULTS AND DISCUSSIONS

The results of applying neural network classifiers to 22 Tamil characters are reported. To conduct the experiment and obtain meaningful results, we must consider all possible variations in the data. Thus the variations due to character size change, position change and writing style of different people are simulated in the experiment. Each character is drawn eight times with the simulated variations on a 64x64 grid. Training has been performed with a set of four examples of each character (for a total of 88). Testing has been carried out with the remaining data. The data sets utilized for training the testing are shown in Fig.4 and Fig.5 respectively.

4.1 General Description of the Experiments

The classifier is simulated on a 80386 based PC/AT. Backpropagation algorithm has been used to train each of the neural network module. The initial weights are chosen randomly in the range of -0.5 to0.5. The learning rate coefficient and the momentum factor are selected to be 0.3 and 0.9 respectively. The stopping criterion for the algorithm is a sum squared error value of 0.01. The classification accuracy is estimated by finding the ratio of the correctly classified test samples to the total number of test samples. Experiments are conducted to obtain the minimum number of hidden nodes and the number of cycles, (i.e., the number of times the network has been through the whole set of training patterns) required for the maximum classification accuracy.

4.2 Experiment with Two Stage System

In this experiment, the module-I of the first stage is trained to classify the four groups of characters using the 6x6 Walsh coefficients obtained from the central portion of each character. The number of nodes used in the input layer of this module is 36. The number of output nodes is four with each node representing a group, and the number of hidden nodes is varied.

Each module in the second stage is trained to classify the characters in each group using the moment invariant features obtained from the remaining portion, other than the central portion of the corresponding group of characters. Each module consists of an input layer of six nodes, and a hidden layer with varying number of nodes. The number of output nodes is the same as the number of characters to be recognized in each group.

Experiments are conducted to obtain the maximum classification accuracy with minimum number of hidden nodes in each module. The results are summarized in table 1.

TABLE 1

Performance of the Two Stage Recognition System

Module	No. of Hidden Nodes	No. of Cycles	Classification Accuracy (%)
I	8	2914	98.86
IIa	5	1573	100.00
IIb	5	2017	100.00
IIc	5	2239	100.00
IId	5	2417	100.00

The two stage system provides an overall classification accuracy of 98.86%. This means that only one of the 88 examples in the test set cannot be classified correctly. The misclassified character shown in double box in Fig.5 has a central rectangular portion similar to that of the characters in group IV due to a high degree of variation in the writing style.

5. CONCLUSIONS

A two stage neural network system for the recognition of handwritten Tamil characters independently of their position, size and writing syle has been implemented. A hybrid feature extraction technique based on the structural properties of the Tamil characters has been introduced which is found to be suitable for neural network classifier. The two stage system provided a high classification accuracy of 98.66%. The two stage system also provides an excellent expansion capability for large-volume Tamil characters recognition.

REFERENCES

[1] Yong, Y(1988). *Handprinted Chinese character recognition via neural networks. Pattern* Recognition Letters 7,19-25.

[2] Fukushima, K. and N.Wake(1991). Handwritten alphanumeric character recognition by the neocognitron. IEEE Trans. Neural Networks 2, 355-365.

[3] Chinnuswamy, P. and S.G Krishnamoorthy(1980). Recognition of handprinted Tamil characters. Pattern Recognition 12,141-152.

[4] Paulpandian, T. and V.Ganapathy(1993). Translation and scale invariant recognition of handwritten Tamil characters using a hieraıchical neural network. In: Proc.IEEE Intern.Symp. Circuits and Systems.IEEE Press.

[5] Huang, J.S. and M.L. Chung(1987). Separating similar complex Chinese characters by Walsh transform. Pattern Recognition 20,425-428.

[6] Hu, M.K.(1962). Visual pattern recognition by moment invariants. IRE Trans.Inform.Theory IT-8,179-187.

[7] Rumelhart, D.E.,G.E.Hinton and R.J.Williams(1986). Learning internal representations by error backpropagation. In: D.E.Rumelhart and J.L.McClelland, Eds., Parallel Distributed Processing; Explorations in the Microstructure of Cognition. MIT Press, Cambridge, M.A.

[8] Khotanzad, A. and J.Lu(1990). Classification of invariant image representations using a neural network. IEEE Trans. ASSP 18, 1028-1038.

[9] Li,Y.(1990). Application of moment invariants to neurocomputing for pattern recognition. Ph.D. Thesis. The Pennsylvania State University.

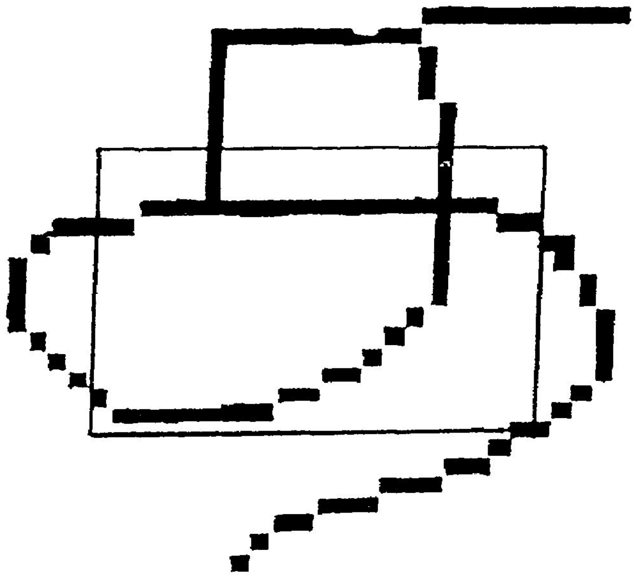


FIGURE 1. Central rectangular and outer portions of a Tamil Character

Group I	ச க த சி கி தி
Group II	ப ம ழ பி மி ழி
Group III	எ ஏ ஞ ர ந ரி நி
Group IV	அ ஆ இ

FIGURE 2. Four groups of Tamil characters.

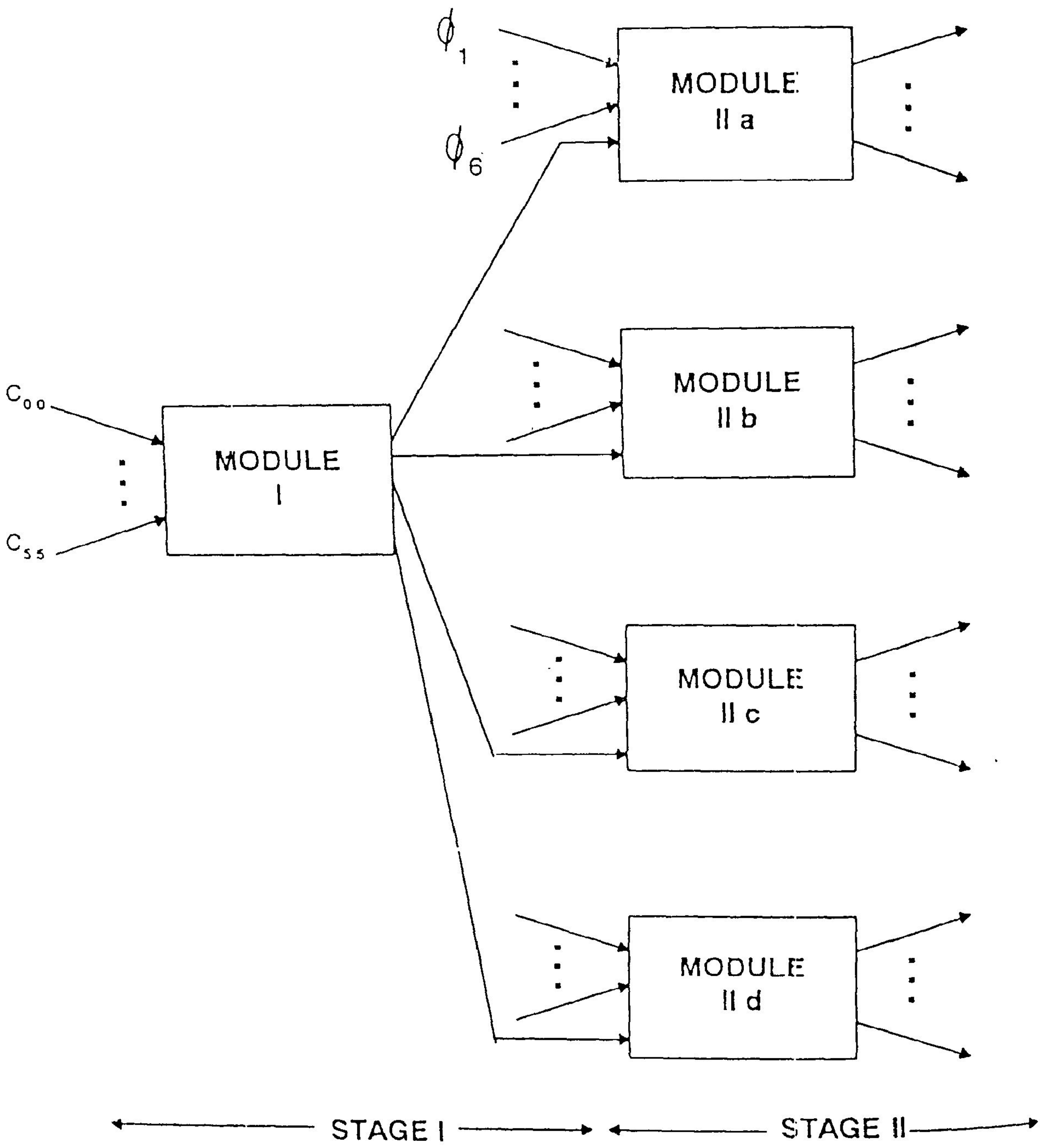


FIGURE 3. The schematic of the two stage recognition system.

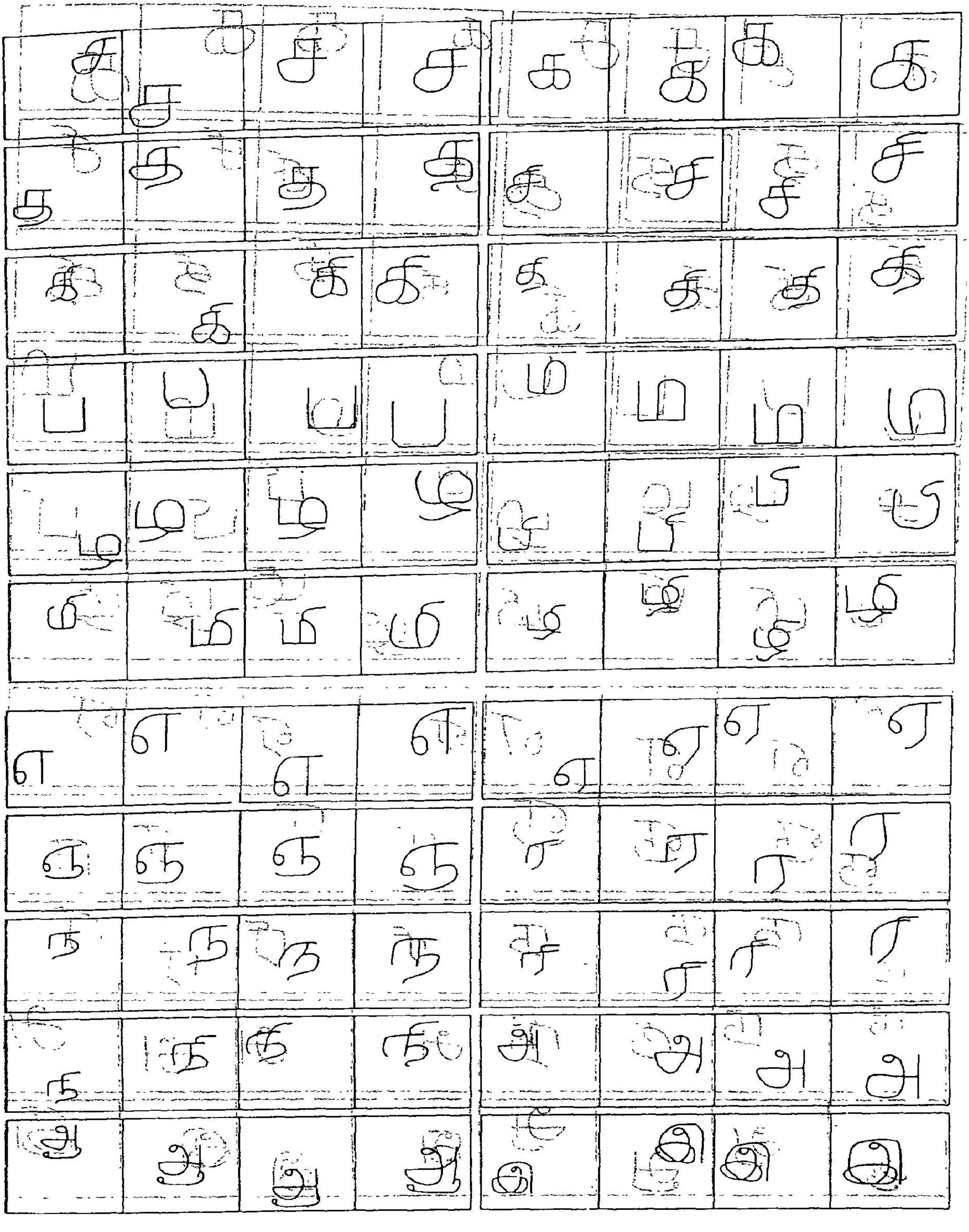


FIGURE 4. Examples of the handwritten Tamil characters used as the training patterns.

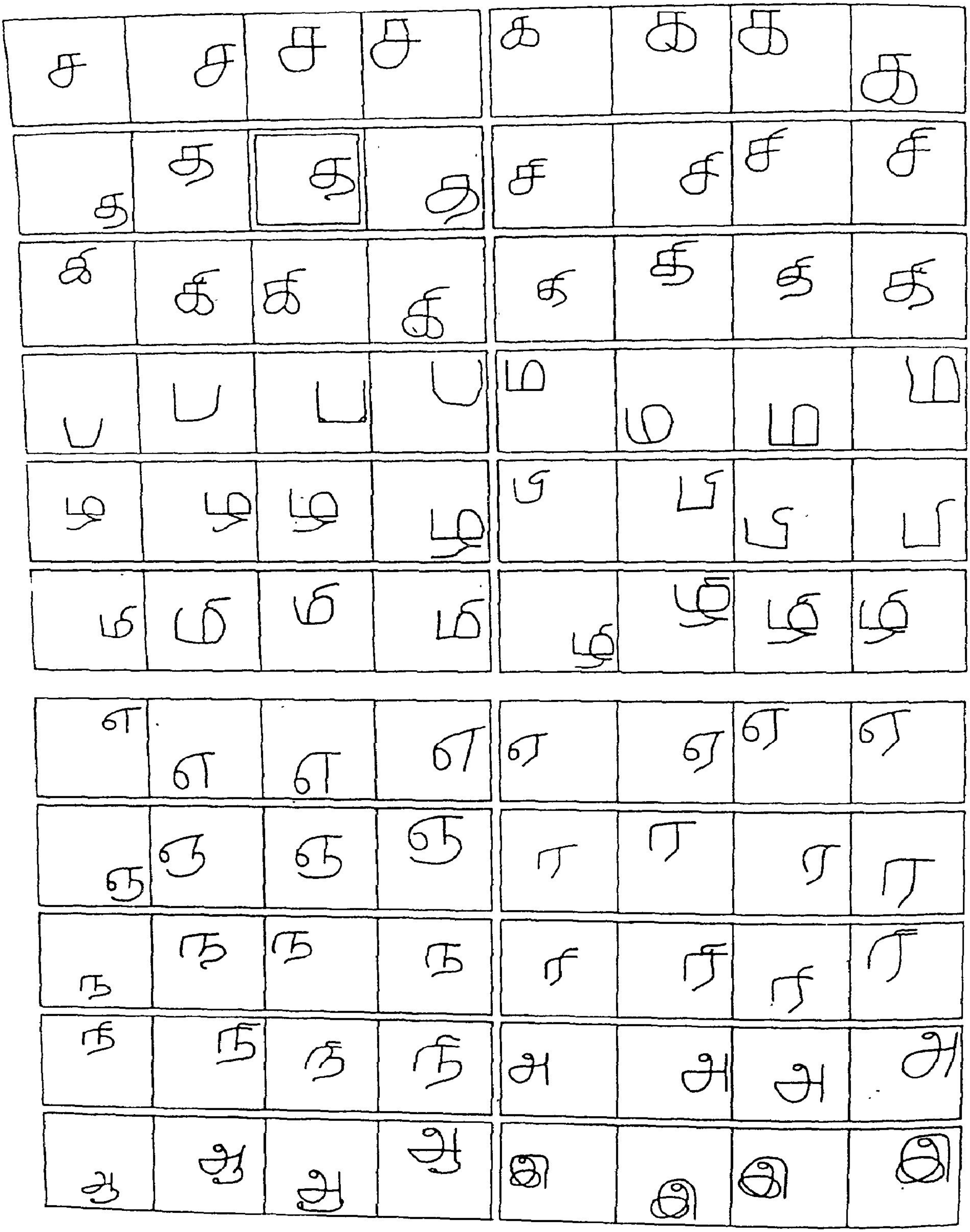


FIGURE 5. Examples of the test patterns.

AUTOMATED DATA CAPTURING USING PATTERN RECOGNITION

E.S. Samundeeswari
T. Devi
K. Sarukesi
Dept. of Computer Science,
Bharathiar University, Coimbatore-46.

ABSTRACT

The important and the time consuming event in Information Management is Data capturing and verification. It occupies a major percentage of computer time which it is important to have an efficient data capturing method which must be accurate and fast.

If there is a method in which the data in some natural language like *Tamil* can be captured at its source directly with no intervening keying, then the burden of data capture can be eliminated. Automated data capturing can be done using a scanner which scans the Tamil text (or document) photo electrically and converts into a pattern of electronic signals. These are compared with the stored pattern for all the possible characters until an exact match is found thus identifying the character. When the generated signals do not match with any character that has been stored, the input is rejected. The read data can be stored in the relevant data bases using GIST technology.

Automated data capturing completely eliminates the verification of data. When better methods are developed for recognizing imperfectly hand written characters and storing them in relevant files/data bases, this new system will no doubt take over the work of most of data entry devices. It eliminates human element in data entry and hence improves data accuracy and timeliness.

1. INTRODUCTION

All offices, irrespective of their size and kind, create and manage information that is required for making decisions or providing services to people within or outside the organization. The information is of no value if it does not satisfy the attributes like accuracy, timeliness and completeness. Many organizations handle a large volume of source data. The important and the time consuming event in Information management is data capturing and verification. It occupies a major percentage of computer time (Nearly 60%). Hence it is important to have an efficient data capturing method which must be accurate and fast.

Now-a-days computer terminals are being used extensively to capture data for subsequent processing and to enter data for immediate processing. The major drawbacks of this method of transcribing are

-- the source document must be well designed.

-- the key-operator should be a high speed, well trained person.

-- the entire document has to be keyed-in to reformat pre-printed or pre-typed documents.

-- data needs to be validated for correctness.

If there is a method in which the data can be captured at its source directly with no intervening keying, then the burden of data capture can be eliminated.

Automated data capturing can be done using a scanner which scans the text (or document) photo electrically and converts into a pattern of electronic signals. These are compared with the stored pattern for all the possible characters until an exact match is found thus identifying the character. When the generated signals do not match with any character that has been stored, the input can be rejected.

2. METHOD OF SOLUTION

The methodology of the reading algorithm is as shown in fig. 2.1.

2.1 Symbol Recognition and Interpretation

Symbol Recognition and Interpretation is the task of detecting symbols in any given image. Here, detecting of text and character in a given input image are considered. This task is carried out in 2 phases (1. Training phase 2. Detecting phase) as shown in fig. 2.2.

2.1.1. Training phase

The set of symbols used in different applications is different. So, it is necessary that the system works with a given set of symbols. In the training phase, the system is trained so that it can take a new set of symbols, i.e., new symbols can be added to the symbol library. In the training phase the document image is displayed along with a box cursor and the box cursor is used for symbol selection. The size of the box can be changed and it can be moved around in the image. Once the symbol is enclosed in a box with its centre coinciding with the symbol centre, the symbol can be selected by pressing a specific key and that symbol is added to the symbol library.

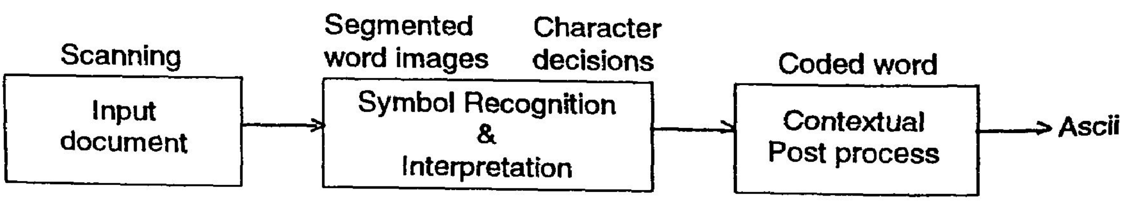


Fig 2.1 Methodology of reading algorithm

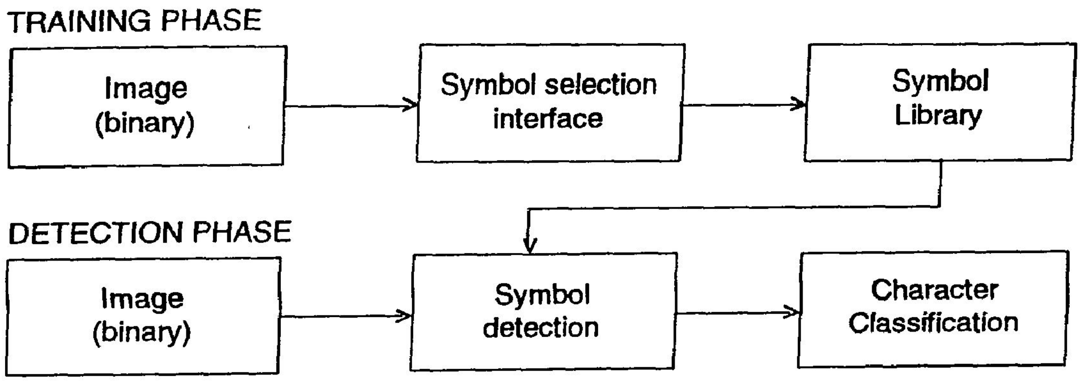


Fig 2.2 Symbol Recognition and Interpretation

2.1.2. Detection Phase

The important part in this work is symbol detection. This is carried out by segmenting the document images into block, line, word, and character.

Segmentation of Document Images

Segmentation is the process of partitioning the image into meaningful texts. It is carried out in 2 steps: detecting text areas in the document image (segmentation) and determining character in each such detected area (classification).

Block Segmentation

Segmentation of text into characters can be done by using a method called **Projection profile** method. The method is based on the observation that documents generally have rectangular block structures. A projection of the black pixel counts along the horizontal and vertical direction can be used for block segmentation. The same method can be adopted for line, word and character segmentation.

The line **segmentation** is done as follows:

Count the number of foreground pixels Fp in each row of the document image. The height of each text line is equal to the number of continuous rows whose Fp value is more than some cut off L. This process is continued upto the last row.

Word segmentation can be carried out in the following way by using line segmentation information:

Count the number of background pixels Bp (starting from the beginning of the text line up to the end of the text line) in each column of the line block.

The number of continuous columns whose Bp is greater than some threshold value can be considered as space between words. From word segmentation the information about rectangles circumscribing the words can be obtained. Using this **character segmentation** can be done. Basically the same procedure which is used for word segmentation can be used for segmenting characters except some thresholds and cutoffs will change, some characters like 'r' are treated separately while segmentation of characters are being performed.

2.3 CHARACTER CLASSIFICATION

In Character Classification the symbol library contains all characters of a given font. From the character segmentation we have the information of each character box, i.e. minimum and maximum coordinates of the circumscribing rectangles of each character. Each character in the box should be matched with all the characters in the character library and the character which matches most is stored in the output file. The matched character is found out by comparing the percentage matches of each character. While storing the segmentation information in a file some markers are stored at the end of each line. These are used to retain the sequence of words and lines in the input image.

In symbol selection there is no need to take the whole symbol as a template and it is enough to store only a part (characteristic feature), which does not resemble any other part of any symbol. All the selected symbols will be stored in a single binary file and other information like size is stored in another ascii file. The system can be trained as and when it is necessary to add a new set of symbols to the symbol library.

2.4 CONTEXTUAL POST PROCESSING

Contextual Post Processing techniques utilize a knowledge source one step above the level of individual character to resolve ambiguities and correct errors in character recognition. These methods use information about other characters that have been recognized in a word as well as knowledge about the text in which the word occurs to carry out this task. Typically, the knowledge about the text takes the form of a dictionary (a list of words that occur in the text).

2.5 INDIAN STANDARD CODE FOR INFORMATION INTERCHANGE (ISCII).

GIST (Graphics Interface Script Technology) provides a very innovative solution for Indian language computing. At the core of the technology is the concept of a phonetic keyboard which can be used for most major Indian languages like Hindi, Tamil etc. The generation of characters in the screen and printer is graphics oriented with each character, combined characters and matras being formed as per the unique feature of the specific language through combination, positioning and manipulation of the sizes of some graphic primitives.

GIST implements the Indian Standard Code for Information Interchange (ISCII) along with fonts developed as per the Indian Standard FOnt Code (ISFOC) guidelines.

3. APPLICATIONS

The read data can be stored in some relevant data bases/files. This method of data capturing aids us in

-- Consolidating texts from various documents into one document.

-- *Pre-printed or pre-typed documents can be read, reformatted and be printed on the* system printed. (serves as a link between the word processing system and the conventional type writer.)

-- Many Engineering applications which require a scanner method capable of handling a wide range of drawing sizes and a mix of line and alphabetic information.

4. CONCLUSION

Since automated data capturing reads the source documents directly, they eliminate the bottleneck of having a person transfer the data from the source document onto a computer input medium. It completely eliminates the verification of data. When better methods are developed for recognizing imperfectly hand written characters and storing them in relevant data bases/files, this new system will no doubt take over the work of most of data entry devices. It eliminates human element in data entry and hence improves data accuracy and timeliness.

BIBLIOGRAPHY

1. **HORST BUNKE**
   Structured and Syntactic Pattern Recognition, Handbook of Pattern Recognition & Computer Vision, World Scientific Publishing Company Ltd., 1993.

2. **KASTURI RANGACHAR, TRIVEDI MOHAN M**
   Image Analysis and Applications, Marcel Dekkar Inc., 1990.

3. **KEINOSUKE FUKUNAGA**
   Statistical Pattern Recognition, Handbook of Pattern Recognition & Computer Vision, World Scientific Publishing Company Ltd., 1993, pp 33 - 60.

4. **STUART C SHAPIRO**
   Encyclopedia of Artificial Intelligence, Second Edition, Vol.1, Wiley-Interscience Publication.

5. **SYAM PRAKASH B B K, MEHTRE B M**
   Automatic identification of Symbols and Text in Maps, Information Technology in Everyday life held at New Delhi, Sep 29 - Oct 2, 1991.

மகிழ் கல்வி   மகிழ் கல்வி   மகிழ் கல்வி   மகிழ் கல்வி

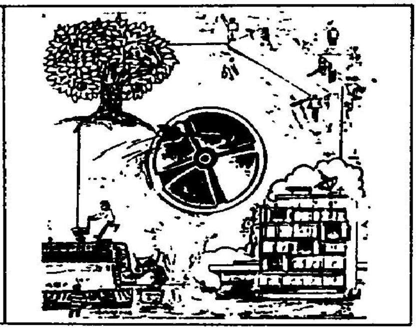
தமிழ் - இதழ் 1

தமிழ் கற்றுக் கொள்வது இனி மகிழ்ச்சியான விளையாட்டுதான்!
விளையாட்டு ஒன்றல்ல. ஒரு வரிசையே!

இந்த வரிசையில் வரும் விளையாட்டுக்கள் பலவிதமானவை. தமிழ் கற்றுக்கொள்வதற்குத் தேவையான அடிப்படைத் திறமைகளை வளர்க்க உதவும்.

எழுத்துக்களை அடையாளம் காண்பதும், அவற்றை, உயிர், மெய், உயிர்மெய் - குறில், நெடில் - வல்லினம், மெல்லினம், இடையினம் என்று வகைப்படுத்துவதும், இந்த வரிசையில் வரும் முதல் இதழில் உள்ள விளையாட்டுகளால் மிகமிக எளிதாகிவிடும்.

தேவை
பிஸி ஏடி
எம்எஸ் டாஸ் 3.1
விஜிஏ திரை
அல்லது
இவற்றுக்கு
ஒப்பானவை



முதல் இதழில் 5 விளையாட்டுகள் உள்ளன. எழுத்துக்களை அடையாளம் காண, விமான நிலையத்தில் சரியான பெட்டிகளை எடுக்கலாம். அணில்களை ஓடவிட்டு சரியான ஆப்பிள்களை உண்ண வைக்கலாம். ஜன்னல்களைத் திறந்து அவற்றுக்குப் பின்னால் ஒளிந்து கொண்டிருக்கும் எழுத்துக்களைக் கண்டுபிடிக்கலாம். கிரிக்கெட் விளையாடி, எழுத்துக்களின் வகையை அறிந்து கொள்ளலாம். இனத்தைச் சேராத எழுத்தினைக் கண்டுபிடிக்க சுறுசுறுப்புடன் தயாராகலாம்.

இவற்றிலெல்லாம் எப்படி முன்னேற்றம் ஏற்படுகிறது என்பதைக் கண்டறிய இத்துடன் வரும் திறன் மதிப்பீட்டு மென்பொருளைப் பயன்படுத்தலாம்!

அடுத்து வர இருக்கும் இதழ்களில் உள்ள விளையாட்டுகள் சொல்லையும் வாக்கியத்தையும் அடிப்படையாகக் கொண்டிருக்கும்.

இந்த விளையாட்டுகள் விரைவில் கணிதம், மற்றும் பல மொழிகளைக் கற்பதற்கும் கிடைக்கும். (கணிதமா? ஆம்! + , - முதல் போட்டித் தேர்வுகளுக்குத் தயார் செய்வது வரை!)

இத்தனையும் வியக்கத்தக்க குறைந்த விலையில்!

லேர்ன்ஃபன் சிஸ்டம்ஸ்

30-4-C, இரண்டாவது பிரதான சாலை,
காந்தி நகர், அடையாறு, சென்னை - 600 020. இந்தியா.

TRANSMATIC SYSTEMS LIMITED

75, LUZ CHURCH ROAD, MYLAPORE
MADRAS - 600 004.
TEL : 4990544, 4993927
TLX : 041-7297 TSLM IN

TRANSMATIC SYSTEMS LIMITED, is a Public Limited Company jointly sponsored by the Kerala State Industrial Development Corporation Limited, Trivandrum (KSIDC) and engineer entrepreneurs to manufacture and market Telecommunication Equipment, Computer Peripherals, Computer Systems and Micro-processor Development Tools.

Transmatic has been Licenced by CDAC to promote Indian Language Computing, using GIST technology. CDAC is a Pune-based autonomous body under Department of Electronics, Government of India. GIST breaks the monopoly of English over Computers. It also breaks the barriers between various Indian Scripts through its "Transliteration" feature. GIST is a 100% Indian Technology.

There is a wide range of products using GIST, meant for different applications and environments.

We are proud to inform you that we are the exclusive dealer for C-DAC products like GIST Cards and related Software in all the major South Indian Cities.

OUR OTHER PRODUCTS

The Products which are manufactured and marketed are Telex/Teleprinter Machines, High Speed Trimatrix Line Printers, Modems, Telephone Call Monitoring Systems, TRANDTAS (Transmatic Distributed Telex Access Systems), Image Scanners, Specialised Softwares for banking applications and C-DAC products for Vernacular Computers.

Software Research Group

COMPANY PROFILE

Software Research Group is one of the leading Research and Development oriented software organisation in India. It was formed in 1984 with Bangalore as its head quarter. The Company was promoted by Sh.T.S. Muthukrishnan (native of Tirunelveli) a B.Tech from IIT, Madras with an experience over 2 decades in the computer field. The company has started off with application software both in India and abroad which includes Hotel Management System, Production planning and control, Travel Reservation etc.

When personal computer are put to use Software Research Group has developed STENO a software controlled wordprocessor for Apple II. Wordlord a powerful wordprocessor for IBM PC's. Wordlord was introduced in Europe in 1985 and subsequently upgraded to Wordlord Super - an Integrated Office Automation software with wordprocessor, Presentation Graphics, Spell checker, Font Editor and Data Manager

Software Research Group is one of the first company to develop in Indian Language (bilingual & tri-lingual) wordprocessors and Desk Top Publishing software Venus. During Singapore Informatics in 1987 SRG launched Barathi (Tamil/English wordprocessor). The company has also participated in Efficiency fair in Holland, CEBIT '89, COMDEX '89, Singapore Informatics '90.

Software Research Group has won the Best Indian Product continously for 4 years. The company's promoter Sh.T.S. Muthukrishnan has been honoured with the following awards:

1. Udyog Excellence Award
2. Shiromani Vikas Award
3. Vikas Rattan Award

Software Research Group dealer network spreads their geographical extent of India with the branches in Singapore and U.S.A. SRG Systems (P)Ltd., a group company of Software Research Group has been formed in 1993 to execute total solution which includes the supply of required Hardware, Software after study & analysis, computer furnitures, computer stationary, power conditioning equipments and Maintenance of the above.

Software Research Group plans to go in for other interesting areas like Medical Electronics & Artifical intelligence in the near future.

MICROSENSE PATRIKA

(Windows Version)

The New Generation Indian Language Desk Top Publishing Software

Microsense Computers Pvt.Ltd. is a joint venture between Indian and Non-Resident Computer Professionals and the A.P. State Government through ANRICH (Andhra Pradesh Non-Resident Indian Investment Corporation, Hyderabad). The company is a pioneer in Indian Language DTP and for the last half a decade has been consistently offering powerful State of the Art software for Indian Language DTP users.

A solution for everyone

Whether your application is book or magazine publishing or advertising, posters or graphic design, Patrika has a solution for you.

You can use the software that best meets your need from a wide selection of Windows software including Aldus PageMaker 4.0, Xerox Ventura (Windows version), Corel Draw 3.0, Adobe illustrator, Aldus Freehand, Windows write, Word for Windows, Microsoft Excell Adobe Type Align etc.

The proof is in the printing

Patrika gives you the ultimate in printer support. Output can be taken on all Windows supported printers including 9-pin and 24-pin Dot Matrix printers, Postscript Laser Printers. HP Laser Printers, Ink Jet Printers and Postscript Typesetters.

Furthermore, you can get all the capabilities of a Postscript Laser Printer on a much cheaper Non Postscript Laser Printer.

Limitless font transformations

Whether you need crisp book fonts or exciting transformations Patrika makes it all possible. Font Outlines, Extrusions, Shading, Rotation, Skewing, Stretching or Mirroring. Aligning text to any arbitrary path and more.

Languages Supported

Patrika is currently available for Hindi, Sanskrit, Marathi, Tamil, Telugu, Kannada, Malayalam and Grantha. Also available is a version for creating English text using Diacritical marks.

Not only DTP but also Indian Language Computing under Windows

Patrika can be used not only for DTP but also for such applications like Spread Sheets and Data Base Applications using software like Lotus 123 for Windows, Microsoft Excell and Windows based data base packages.

Multi user systems(LAN versions)

Apart from providing solutions using stand-alone computers, Microsense provides LAN Versions for those requiring multiple work stations. Using the industry standard Novell Netware along with Patrika. Microsense provides high performance multi user systems.

Microsense Computers Private Limited

DTP Division	: 5D, J.P. Towers, 7/2, Nungambakkam High Road, Madras - 600 034.Tel : 476469/8256529
Secunderabad	: Emerald House, II Floor, Sarojini Devi Road, Secunderabad 500 003. Tel:846266/845008/ 844781
Bangalore	: 209, II Floor, Richmond Towers, 12, Richmond Road, Bangalore 560 025. Tel:210816/ 218037
Bombay	: 808,Embassy Centre, Nariman Point, Bombay 400 021. Tel:231489/2872948/225441
New Delhi	: Satyadeep, III Floor, 1 Community Centre, East of Kailash, New Delhi 110 065 Tel:6410134/ 6465855

WINDOWS & Inscrypt DTP Softwares

- Windows 3.1 Interface
- Typewriter & Phonetic layout
- ADOBE Type 1 Fonts
- Languages:TAMIL, HINDI, KANNADA, TELUGU, MALAYALAM, SOWRASHTRA, etc.
- Direct entry inside some of the popular windows applications - Write, Pagemaker,CorelDraw, Quark Express, Ventura. MS Excell, Adobe illustrator, Photoshop etc.
- Regional languages keyboard layout available inside all Windows applications
- ClipBoard operations facilities to import Editor files inside WINDOWS
- Scalable point Sizes - 4 to 999 pt size
- Special effects: Rotate, Mirror, Skew, Condense, Expand, Reverse, Outline, Shadow, Extrude, Blend etc.
- Supports all NON-PS, Postscript Laser Printers, Dot Matrix Printers, Plotters, Typesetters etc.
- High Quality and elegant typefaces
- Free training & moderate price
- Excellent Documentation

Inscrypt

No.1 Tamil Wordprocessor

- A highly sophisticated, cost-effective, userfriendly, Bilingual Wordprocessing Software
- Comes with an easy WordProcessor like Wordstar
- OnScreen Tamil key board lay-out
- Toggle between English & Tamil
- File and Directory listings through our window
- Editing and key caps facilities through Mouse
- Short hand Menu, Find & Replace, Block operations etc
- Olden & Periyar Tamil Characters are available
- Different font styles with user's manual
- Works under DOS based IBM compatibles (Minimum 640 KB RAM)
- Supports all Dot Matrix Printers
- Price Rs.1750/-

Softview Computers

No.10, Second Floor, Nelson Manickam Road,
Madras - 600 094.
Phone: 4836471 Fax: 621 3371

CADGRAF

DTP SYSTEM

Your in-house printing system. Instant. Fast. Immaculate boardroom quailty. Most cost effective. Light years ahead of conventional printing.

- ✓ Total Solutions for DTP - IBM and Apple
- ✓ ELANGO DTP Software for IBM and Apple
- ✓ Computer Aided Textile Designing, Printing and Weaving Systems.
- ✓ Multimedia Presentation and Educational Systems
- ✓ Imaging Solutions for Paperless Office
- ✓ Medical Imaging and Analysis

Sucessfully Installed at

IIT, HTL, Anna University, TIDCO, L&T ECC, CECRI,CIMI,
Center for Reliability, Indian Oil,Tube Investments, Directorate of Handlooms, etc.

CADGRAF SCHOOL OF DTP

State of the art training in DTP. Fully equipped. Experienced and trained faculty. Part time and regular courses. 100% placement record.

CADGRAF COMPUTERS

the DTP people

Gee Gee Plaza, III Floor, No. 1, Wheatcrofts Rd., Nungambakkam, Madras - 600 034.
TEL : 8269021, 8256812 FAX : 044 - 8522027 CAD.

குறள் - புதிய தமிழ் கணிப்பொறி மொழி

முனைவர் சு.குப்புசாமி
துறைத்தலைவர்,
கணிப்பொறித்துறை,
புதுவைப் பல்கலைக்கழகம்

திரு. வெ. பிரசன்னா வெங்கடேசன்
மாணவர் (எம்.டெக்)
கணிப்பொறித்துறை,
புதுவைப் பல்கலைக்கழகம்

முன்னுரை

மனித வளர்ச்சியின் ஒவ்வொரு கால கட்டத்திலும் அவனுக்கு உறுதுணையாய் இருந்தது,இருப்பது, இருக்கப்போவது கருவி. அன்றைய மனிதனுக்குக் கருவி கல்லும் வில்லும் என்றால் இன்றைய மனிதனுக்கு கணிப்பொறி. கணிப்பொறியின் உபயோகம் ஆங்கில மொழியை அடியொற்றி இருப்பதால் ஆங்கிலம் தெரியாத பெரும்பான்மையான தமிழருக்குக் கணிப்பொறி எட்டாக் கனியாக. சற்றே மிரட்சி தரக்கூடியதாகத் தென்படுகிறது.

இந்நிலையை மாற்ற கணிப்பொறி என்ற இன்றைய யுகக் கருவியை முழுமையாகப் பயன்படுத்த கணிப்பொறி மொழி, தமிழில் அமைய வேண்டும். தகவல்களைத் திரட்டுதல், சேமித்தல், பரப்புதல், கண்டெடுத்துக் கொணர்தல் ஆகியவற்றிற்கான அடிப்படைக் களங்கள் முழுமையாகத் தமிழில் உருவாக்கப்படவேண்டும். தமிழ் மட்டுமே தெரிந்த பெரும்பான்மையோருக்கும் கணிப்பொறி கிட்டும் கனியாக வேண்டும். இவற்றிற்கும் மேலாக தமிழ் கால வெள்ளத்தில் அழியாமல் காக்கப்பட வேண்டும்,. வளர்ச்சி என்ற பாதுகாப்பிற்கு உட்படுத்த வேண்டும். இப்பெரும் பணிக்கான சிறு முயற்சியாக குறள் என்ற புதிய கணிப்பொறி மொழியை இக்கட்டுரை அறிமுகப்படுத்துகிறது.

தமிழில் வடிமைப்பதில் உள்ள சிக்கல்கள்

தமிழில் கட்டளை மொழியினை வடிவமைப்பதில் சிக்கல்களாகச் சுட்டப்படுபவை மூன்று விதமானவை. அவையாவன,

(1) தமிழ் வரி வடிவம்

(2) தமிழில் விசைப்பலகை

(3) தமிழில் கலைச் சொற்கள்

இச்சிக்கல்களுக்கு அங்கீகரிக்கப்பட்டத் தீர்வுகள் ஏதும் இல்லாமையால் பொது நிலையிலேயே குறள் வடிவமைக்கப்பட்டுள்ளது. குறளின் தமிழ் வரிவடிவம் புதிய மாற்றங்கள் ஏதும் இல்லாதது. பெரும்பான்மையோரின் பயன் பாடு கருதி தமிழ் தட்டெழுத்து முறையை அடியொற்றியே குறளின் விசைப் பலகை அமைக்கப்பட்டுள்ளது.

தமிழில் கலைச் சொற்களை உருவாக்குவது மிகப்பெரிய சிக்கலாகக் கருதப்படுகிறது. பழந்தமிழ் இலக்கியங்களில் காணப்படும் சொற்களை இன்றைய தேவைக்கேற்ப பொருள் கொண்டு பயன்படுத்தல், ற்கால இலக்கியங்களிலிருந்து தகுந்த சொற்களைப் பெறுதல். பேச்சு மொழியிலிருந்து தேர்ந்த சொற்களை எடுத்தாளுதல், தொடர்புள்ள பிற மொழிச் சொற்களைக் கடன் பெறல், மொழிபெயர்த்தல், புதுச்சொற்களைப் படைத்தல், உலக வழக்கை அப்படியே ஏற்றுக் கொள்ளுதல் எனச் சில வழிகளைப் பின்பற்றலாம். குறளுக்கான கலைச் சொற்களை வடிவமைக்கையில் அகராதிப் பொருளுக்கேற்ப, பெற்ற பொருளைக் கொண்டு சிற்சில மாற்றங்கள் செய்து மொழியாக்கம் செய்யும் முறையே முற்றிலும் பின்பற்றப்பட்டுள்ளது.

குறள் - ஓர் அறிமுகம்

ஒன்றரையடி வடிவத்தைக் கொண்டு திருக்குறள் உலகத்திற்கே பொதுவான அனைத்துச் செய்திகளையும் சொல்லியுள்ளது, அதுவும் மிக எளிமையாக. உலகப் பொதுமறையாம் இத் திருக்குறளை முன்னுதாரணமாகக் கொண்டு தமிழில் கணிப்பொறிக்கு ஒரு மொழியை அமைத்துத்தர முயல்வதே குறளின் குறிக்கோள். எளிமைக்கும் சுருக்கத்திற்கும் முதன்மை தரப்பட்டு, அமைக்கப்பட்ட இக்கணிப்பொறிமொழி சில வரையறைகளுக்கு உட்பட்டே வடிவமைக்கப்பட்டுள்ளது.

'மொழி உருவாக்கத்தில் மிக முக்கியமான தீர்மானங்கள் அனைத்தும் எதையெல்லாம் விட்டுவிடுவது என்பதையே சார்ந்திருக்க வேண்டும்' என்கிறார் நிக்கலஸ் விர்த் (**Niklaus Writh**) . இவர் கூற்றே 'குறள்' வடிவமைப்பில் மேற்கொள்ளப்பட்ட முதல் வரையறை.

வெற்றிகரமான கணிப்பொறி மொழியை உருவாக்குவதற்கு உறாரோவிட்ஸ் (Horowitz) கூறும் படிநிலைகள் பத்து. அவையாவன.

(1) பயனுள்ள, குறிப்பிட்டப் பகுதியைத் தேர்ந்தெடுத்தல்.

(2) வடிவாக்கக் குழுவை முடிந்தவரையில் சிறியதாக அமைத்தல்.

(3) சில அரிதான விரிவாக்கக் குறிக்கோள்களை வடிவமைத்தல்.

(4) ஆர்வலர் கொண்ட சிறிய குழுக்களுக்கு மொழிவடிவம் ஒன்றை வெளியிடுதல்.

(5) மொழி வரையறுத்தலை மீள் பார்வை (Revise) செய்தல்.

(6) சிறிய அளவிலான (Prototype) மொழிபெயர்ப்பியை கட்டமைத்தல்.

(7) மொழியின் பொருளாக்கத்திற்கான முறைமையான வரையறைகளை ஏற்படுத்துதல்.

(8) மீண்டும் மொழி வரையறுத்தலை மீள்பார்வை செய்தல்.

(9) எளிமையான, தெளிவான, சுருக்கமான மொழிக்கான குறிப்பேட்டை உருவாக்கி. வெளியீடு செய்தல்.

(10) தரம்மிக்க மொழிபெயர்ப்பியை உருவாக்குதல் மற்றும் பரப்புதல்.

இப்பத்து படிநிலைகளின், முதல் மூன்றின் அடிப்படையில் குறள் வடிவமைக்கப்பட்டு நான்காவது படிநிலையான ஆர்வலர் முன் வெளியிடும் முயற்சியாக இக் கட்டுரையின் மூலம் வெளியிடப்படுகிறது. இவ்வாறு வரையறைகளுக்குட்பட்டே உருவாக்கப்பட்ட குறளின் நோக்கங்கள்.

(1) எளிய சிக்கல்களுக்குத் தீர்வு காணுதல்.

(2) கணிப்பொறி மொழியை எளிதாகக் கற்றுணர்தல்.

(3) கணிப்பொறிகளின் செயல்பாடுகளை அறிந்து கொள்ளுதல்

ஆகியவையே ஆகும். இந்நோக்கங்களுக்கு ஏற்ப குறள் தன்னகத்தே மிகச் சில ஆணைகளையும் செயல்பாடுகளையும் கொண்டு விளங்குகிறது. இன்றைக்குப் புகழ் பெற்றிருக்கும் கணிப்பொறி மொழிகளுக்கு இணையாக வளரும் எண்ணத்தோடு அமைக்கப்பட்ட குறள் தமிழில் கணிப்பொறிமொழி என்ற விண்ணகத்தைக் காண்பதற்காகத் திறக்கப்பட்ட சிறு சாளரமே.

குறள் - வடிவாக்கம்

ஒரு கணிப்பொறிமொழியை உருவாக்கும்போது பின்வரும் கூறுகளை அடிப்படையாகக் கொண்டே அமைக்க வேண்டும். அவையாவன,

(1) மொழியின் எழுத்துத் தொகுதி (Alphabets)

(2) இயக்கிகள் (Operators)

(3) மாறிலி (Constants)

(4) மாறி (Variable)

(5) எண்ணுருக் கோவை (Expression)

(6) கட்டளைகள் (Commands)

ஆகியவை ஆகும்.

குறளில் மேற்சொன்னகூறுகள் எவ்விதம் அமைந்துள்ளன என்பதை இனிக் காணலாம்.

(1) மொழியின் எழுத்துத் தொகுதி (Alphabets)

குறள், தமிழின் 247 வரி வடிவங்களையும் கொண்டதாகவே உள்ளது.

உயிரெழுத்து	-	12
மெய்யெழுத்து	-	18
உயிர்மெய்யெழுத்து	-	216
ஆயுதம்	-	1
		247

மேலும், சிறப்புக் குறியீடுகளாக,

‘,’ ‘"’ ‘$’ ‘#’ ‘%’ ‘;’ ‘<’ ‘>’ ‘=’

போன்றவற்றையும் கொண்டதாகக் குறள் அமைந்துள்ளது.

(2) இயக்கிகள் (Operators)

ஒரு மொழிக்கு இன்றியமையாதத் தேவைகளான +, -, *, / ஆகிய இயக்கிகள் குறளில் இடம் பெற்றுள்ளன.

(3) மாறிலி (Constants)

மாறிலி முழு எண்ணாகவோ. இயல் எண்ணாகவோ அல்லது ஒரு எழுத்துத் தொடராகவோ இருக்கலாம். சான்றாக,

முழு எண்	-	3
இயல் எண்	-	5.5
எழுத்துத் தொடர்	-	“குறள் - ஒரு கணிப்பொறிமொழி”

இவற்றுள் எழுத்துத் தொடர் “ ” க்குள் இடம்பெற வேண்டும். மாறிலியின் இத்தன்மைகள் குறளிலும் உண்டு.

(4) மாறி (Variables)

மாறி ஒரு முழு எண்ணையோ இயல் எண்ணையோ அல்லது எழுத்துத் தொடரையோ ஏற்று இருக்கலாம். ஆனால் மாறிக்கான பெயர் சில விதிமுறைகளுக்குட்பட்டே அமைதல் வேண்டும்.

அவ்விதி முறைகள் -

(1) மாறிக்கான பெயர் அதிக பட்சமாக ஏழு எழுத்துக்களைக் கொண்டே அமைய வேண்டும்.

(2) அதிலும் முதல் எழுத்து நெடுங்கணக்கைச் சார்ந்த எழுத்தாகவே இருக்க வேண்டும்.

(3) முதலெழுத்திற்குப் பிறகு வரும் எழுத்துக்கள் எண்ணாகவோ எழுத்தாகவோ இருக்கலாம்.

(4) இறுதி எழுத்து,

முழு எண்ணாக இருந்தால் %ஆகவும்,

இயல் எண்ணாக இருந்தால் # ஆகவும்,

எழுத்துத் தொடராக இருந்தால் $ ஆகவும்,

இருக்க வேண்டும். இவ்விதி முறைகள் குறளில் பின்பற்றப்பட்டுள்ளன.

(5) எண்ணுருக் கோவை (Expressions)

எல்லா மொழிகளிலும் உள்ள எண்ணுருக்கோவைகளைப் போலவே குறளிலும் எண்ணுருக்கோவை அமைக்கப்பட்டுள்ளது. எண்ணுருக் கோவையைக் கணக்கிடும்போது பின் வரும் விதிகள் பின்பற்றப்படுகின்றன.

(1) படிநிலை அமைப்பு விதி (Hierarchy Rule)

இவ்விதியில் பின்வரும் படிநிலை முறையிலேயே இயக்கிகள் செயல்படும்.

* , /

+, -

<, >, <=, >=, =, < >

இறுதிப்படிநிலைக் குறியீடுகள் ஒப்பீட்டு முறையில் குறளில் பயன்படுத்தப்படுகின்றன.

(2) இணைவிதி (Associativity Rule)

இணைவிதி என்பது வலது புறத்திலிருந்து இடது புறத்திலோ அல்லது இடது புறத்திலிருந்து வலது புறத்திலோ, இயக்கிகளைச் செயல்படுத்தும் முறையைப் பற்றியது. குறளைப் பொருத்தவரை இவ்விதியின் அடிப்படையில் இடது புறத்திலிருந்து வலது புறத்திற்கு இயக்கிகள் கணக்கிடப்படுகின்றன.

(3) அடைப்புக்குறி விதி (Parantheses Rule)

அடைப்புக்குறி எனும் இவ்விதி எத்தனை அடைப்புக் குறிகள் இடப்பட்டிருந்தாலும் எது மிகவும் உள்ளடங்கியதாக உள்ளதோ, அதை முதலில் எடுத்துக் கொண்டு செயல்படும், அவ்வாறே படிப்படியாக ஒவ்வொரு அடைப்புக் குறிக்குள் உள்ள கணக்கீடுகளும் மதிப்பிடப்படும். குறளும் இவ்விதி முறையிலேயே இயங்கும்.

6. ஆணைகள் (Commands)

ஒரு கணிப்பொறி மொழியில் செயல்பாடுகளை நிர்ணயிக்க உதவுபவை ஆணைகள். அவ்வாணைகள் குறளில் எவ்விதம் அமைந்துள்ளன என்பதைப் பின்வரும் பகுதி விளக்கும்.

குறளில் இடம் பெற்றுள்ள ஆணைகளை நான்கு விதமான பகுதிகளுக்குள் அடக்கலாம்.

(1) இடுவரல் விடுவரல் ஆணைகள் (Input / Output Commands)

ஒரு செயல்பாட்டைச் செய்வதற்கான இடுவரல். விடுவரல் ஆணைகள் கொண்டதாக இப்பகுதி இருக்கும்.

படி என்ற ஆணை விசைப் பலகையின் மூலமாகத் தரப்படும் இயல் எண்ணையோ. முழு எண்ணையோ, எழுத்துத் தொடரையோ குறிப்பிடப்பட்ட நினைவகத்திற்குக் கொண்டு வரும்.

(எ.கா) <u>படி</u> முதலிடு #, காலம் # , வட்டிவிகி #

<u>எழுது</u> என்ற ஆணை குறிப்பிடப்பட்ட முகவரியிலிருந்து தேவையான இயல் எண்ணையோ, முழு எண்ணையோ. எழுத்துத் தொடரையோ அல்லது எண்ணுருக் கோவையையோக் கணக்கிட்டு விடையைத் திரைக்குக் கொண்டு வந்து காட்டும்.

(எ.கா) <u>எழுது</u> "வட்டித் தொகை" = " வட்டி #

(2) வரிசைக் கட்டுப்பாட்டு ஆணைகள் (Sequence Control Commands)

பொதுவாக ஆணைகள் வரிசை கிரமமாகச் செலுத்தப்படும். தரப்பட்ட ஆணைகள் வரிசை மாறி செயல்பட வேண்டுமானால் வரிசைக் கட்டுப்பாட்டு ஆணைகள் பயன்படுத்தப்படும்.

தீர்மானி **(IF)** *இந்தஆணை, தரப்பட்ட எண்ணுருக் கோவையை தர்க்க ரீதியாக மதிப்பிட்டு அது உண்மையானதாக இருந்தால் அடுத்து கொடுக்கப்பட்டுள்ள ஆணையைச் செயற்படுத்தும்.*

(எ.கா) <u>தீர்மானி</u> (காலம் # 0.0) <u>செல்</u> 20.

செய் (Do Loop) *செய்முற்று* (Enddo)

'செய்' ஆணை தொடரில் கொடுக்கப்பட்டிருக்கும் எண்ணுருக் கோவை உண்மையாகஇருக்குமானால் தொடர்ந்து 'செய்' ஆணையிலிருந்து 'செய்முற்று' ஆணைவரைக் கொடுக்கப்பட்டுள்ள வரிசைக்கிரமமான ஆணைகளை மீண்டும் மீண்டும் செயல்படுத்திக் கொண்டேயிருக்கும். பொய்யாக ஆகுமானால் செய்முற்றுக்கு அடுத்த ஆணையைச் செயல்படுத்தத் தொடங்கிவிடும்.

(எ.கா) <u>*செய்*</u> (ஐ % < 100)

<u>*நியமி*</u> ஐ % + 1

<u>*செய்முற்று*</u>

செல் **(Go to)**

'செல்' ஆணைக்கு அருகில் தரப்பட்டுள்ள வாக்கிய எண் (Statement Number) *எதுவோ அந்த வாக்கிய எண்ணுக்குரிய ஆணையை அடுத்த ஆணையாக எடுத்துக் கொண்டு செயல்படும்*

(எ.கா) <u>செல்</u> 50

(3) நியமி ஆணை (Assignment Command)

மாறிலியையோ, மாறியையோ, எண்ணுருக்கோவை கணக்கீட்டையோ பெயர் கொடுக்கப்பட்ட மாறிக்கு நியமிக்கும் ஆணையே 'நியமி' ஆகும்.

*(எ.கா) <u>நியமி</u> வட்டி # = (முதலீடு # * காலம் # * வட்டி விகி #) / 100*

(4) பிற ஆணைகள்

குறளில் பயன்படுத்தும் பிற ஆணைகள் இப்பகுதியில் அடங்கும்.

குறிப்பு **(Comment / Remark)**

ஒரு ஆணைத் தொடருக்கான குறிப்பையோ, விளக்கத்தையோ குறிக்கப் பயன்படும் ஆணையே குறிப்பு ஆணை.

(எ.கா) குறிப்பு தனி வட்டி கணக்கிடுதல்

நிறுத்து (Stop)

'நிறுத்து' ஆணையின் மூலமாக ஓர் ஆணைத் தொடரின் செயல்பாடு நிறுத்தப்பெறும்.

(எ.கா) நிறுத்து

முற்று (End)

'முற்று' ஆணையானது மொழி பெயர்ப்பிக்கு ஆணைத்தொடர் முடிந்து விட்டது என்பதை அறிவிக்கும் பணியைச் செய்கிறது.

(எ.கா) முற்று

'குறளி'ல் மாறிகளை ஆணைத் தொடரின் ஆரம்பத்திலேயே அறிவிக்க வேண்டிய தேவையில்லை. எழுதப்படும் ஆணைத்தொடர் எப்பொழுதும் 'முற்று' ஆணையுடனேயே முடிவு பெறவேண்டும். குறளில் எழுதப்பட்ட ஆணைத் தொடரில் ஒரு வாக்கியத்தைப் பிற வாக்கியங்களிலிருந்து வரையறுத்துக் காட்ட அரைப்புள்ளி பயன்படுத்தப்படுகிறது. இவ்வாறு வடிவாக்கம் செய்யப்பட்ட குறள் கணிப்பொறி மொழியைப் பயன்படுத்தி எழுதப்பட்ட ஆணைத்தொடர் சான்றாக கீழே தரப்பட்டுள்ளது

(எ.கா)

குறிப்பு	தனிவட்டி கணக்கிடுதல்;
எழுது	"முதலீட்டுத் தொகை";
படி	முதலீடு #;
எழுது	"காலம்" ;
படி	காலம் #;
எழுது	"வட்டி விகிதம்";
படி	வட்டி விகி #;
நியமி	வட்டி (முதலீடு # * காலம் #* வட்டி விகி #) / 100;
எழுது	"வட்டித் தொகை" = * வட்டி #
நிறுத்து;	
முற்று;	

குறள் - நிர்மாணித்தல் (Implementation)

Dos தளத்தில் TURBO-C மொழியில் 'குறள்' நிர்மாணித்தல் நடைபெற்றுக் கொண்டிருக்கிறது.

முதலில் எழுத்து வடிவமைப்பான் (Font Editor) உதவியோடு தமிழ் எழுத்துக்கள் அனைத்தும் வடிவமைக்கப்பட்டன. வடிவமைக்கப்பட்ட எழுத்துக்கள் விசைப் பலகை இயக்கி (Key board driver) எழுத்துக்களோடு பொருத்திப் பார்க்கப்பட்டன. இப்பணியின் முடிவில் தமிழ்த் தொகுப்பான் (Tamil Editor) உருவாக்கப்பட்டது. கோப்பு மேலாளர், (File Manager), செயலுருவாக்கி (Interpreter) ஆகிய இரண்டையும் உருவாக்கும் பணி நடைபெற்றுக் கொண்டிருக்கிறது.

முடிவுரை.

தமிழில் கணிப்பொறி மொழி அமைக்கப்பட வேண்டும். தமிழ் மட்டுமே தெரிந்த பெரும்பான்மையினரைக் கணிப்பொறி எளிதாக சென்று சேர வேண்டும் என்ற ஆர்வத்தின் வெளிப்பாடே குறள். இக்குறள் தன்னுடைய முதன்மை நோக்கங்களை நிறைவேற்றக்கூடியது. இதனாலேயே இது முற்றிலும் முழுமை பெற்று விட்ட மொழி வடிவம் என்று தற்போது கூறிவிட முடியாது. ஆனால், வருங்காலத்தில் முழுமையான மொழியாக வடிவமைக்கப்பட வேண்டும் என்ற எண்ணத்தோடு உருவமைக்கப்பட்டுள்ளது.

> 'கானமுயல் எய்த அம்பினில். யானை
> பிழைத்த வேல் ஏந்தல் இனிது'
>
> - திருக்குறள்- 772

என்ற குறளுக்கேற்ப இம்முயற்சி மேற்கொள்ளப்பட்டது. அனைத்து கணிப்பொறி மொழித்திறன்களும் கொண்ட மொழியாக குறள் அமைய வேண்டும் என்பதே இம் மொழியின் உச்சக்கட்டக் குறிக்கோள். இக் குறிக்கோள் நிறைவேறுகையில் குறள் எதற்கும் ஈடுதரும் 'குறுகத்தறித்த குறளாக' அமையும் என்பது உறுதி.

துணை நூற்பட்டியல்

1. குழந்தைசாமி. வா.செ., - அறிவியல் தமிழ்
பாரதி பதிப்பகம்
108, உஸ்மான் சாலை
சென்னை - 17.

2. கோவிந்தசாமி. நா., (கட்டுரையாசிரியர்) - இரண்டாவது உலகத்தமிழ்
ஆசிரியர் மாநாடு ஆய்வடங்கல்
தொகுதி - 2
உமா பப்ளிகேஷன்
85, ஜலன் பெரிகன்டைன்
சென்தூல்
கோலாலம்பூர். 7

3. ELLIS HOROWITZ - Fundamentals of programming Languages
Galgotia Pub. (P) Ltd.
5, Ansari Road,
Parya Ganj
New Delhi - 110 002.

தமிழ் கற்பிக்கும் நிரல்கள், பல்-ஊடக சாதனங்கள்
(கணிப்பொறியின் வாயிலாக திருக்குறள் கற்றல்)

வி.பி. பழனியப்பன்
ஸ்ரீ ராமகிருஷ்ண மிஷன் வித்யாலயா
கல்வியியல் கல்லூரி
கோவை - 20.

அறிமுகம்

மாறிவரும் இவ்வுலகில், மாணவர்களுக்கு எளிதாகக் கற்பிக்கப் பல சாதனங்கள் உருவாகியுள்ளன. அவற்றுள் மாணவர்களுடன் தொடர்பு வைத்துக்கொள்ள உதவும் ஒரே சாதனம் கணிப்பொறியாகும். கணிப்பொறி மூலம் மாணவர்கள் தமிழைத் தவிர அனைத்துப் பாடங்களையும் கற்க முடியும்.

பெரும்பாலான கணிப்பொறிகள் ஆங்கில எழுத்துக்களையே சார்ந்திருக்கின்றன. கணிப்பொறியின் வாயிலாகக் கற்க ஆங்கில மொழியே பயன்பட்டு வருகின்றது. ஆங்கில மொழி அல்லாத அறிவியல், கணிதம், சமூகவியல் போன்ற பாடங்களையும் கணிப்பொறியின் உதவியால் கற்பதாக இருந்தால், அவற்றை ஆங்கிலம் வாயிலாகத்தான் கற்க முடியும் என்ற நிலையுள்ளது. ஆங்கில மொழி அறியாதவர்களுக்கு கணிப்பொறி வாயிலாகக் கற்றல் என்பது கனவாகவே இருந்து வருகிறது. இந்நிலையை மாற்றி, தமிழையும் கணிப்பொறியில் கற்றுக் கொடுக்க முடியும் என நிலைநாட்டவே இம்மென்பொருள் தயாரிக்கப்பட்டது.

எல்லாச் சிறிய கணிப்பொறிகளிலும் உபயோகப்படுததப்படும் ஒரு மொழி பேஸிக் (BASIC) ஆகும். இதுதான் மற்ற மொழிகளைவிட கற்பித்தலுக்கும் சிறந்தாகக் கருதப்படுகிறது. இம்மொழியில் மாணவர்களுக்குப் பயன்படும் வகையில் தமிழில் மென்பொருள் தயாரிக்க எண்ணி, உலகப் பொதுமறையாம் திருக்குறள் கணிப்பொறியாக்கம் செய்யப்பட்டுள்து. இம்மென்பொருள் மூலம், திருக்குறளை, தமிழ் படிக்கத் தெரிந்த அனைவரும் எளிதாக கற்க முடியும் என எண்ணுகிறேன்.

நுழைவு வாயில்

எழுத்து என்பது ஒருவகை வரைபடமே. கணிப்பொறி, நாம் கொடுக்கும் கட்டளைக்கேற்ப வரைப்படம் வரையும் தன்மை கொண்டது என்பது நாம் அறிந்ததே. இக்கருத்தினை அடிப்படையாகக் கொண்டு இம்மென்பொருள் தயாரிக்கப்பட்டது.

இம்மென்பொருளில், மூன்று செய்திக் கோப்புகளும் (Data Files) ஒரு இயக்கும் நிரலும் (Program) உள்ளது. இதில் உள்ள செய்திக் கோப்புகள் முறையே (1) குறளின் பாடல் (2) குறளின் தெளிவுரை மற்றும் (3) குறளின் அதிகாரப் பெயர் கொண்ட தனித்தனிக் கோப்புகளாக உள்ளன. இயக்கும் நிரல், கற்பவரின் விருப்பத்திற்கேற்ப அவர் கொடுக்கும் கட்டளைப்படி, கணிப்பொறி கோப்புகளிலிருந்து எடுத்துத் திரையில் எழுதுகிறது.

இந்த மென்பொருளில் உள்ள அனைத்து குறள்களும், அவற்றிற்குரிய விளக்க உரைகளும் பதிவு செய்யப்பட்டுள்ளன. இதில் உள்ள குறள்களும், விளக்க உரைகளும் கணிப்பொறியாக்கம் செய்ய உதவியாக இருந்தது, டாக்டர் மு.வரதராசனாரின் 'திருக்குறள் தெளிவுரை'. (112 ஆவது பதிப்பு : மார்ச் 1992) ஆகும்.

பயன்படும் முறை

கற்றல் மற்றும் கற்பித்தல் சிறந்து அமைய வேண்டுமெனில், அவை சிறிது சிறிதாகப் பொருள் தரும்படியும் , கற்பவரின் விருப்பத்திற்கு ஏதுவாகவும் இருக்க வேண்டும் என்பது நாம் அறிந்ததே. அக்கருத்து திருக்குறளுக்கும் பொருந்தும் என்பதால் திருக்குறளில் உள்ள மொத்த குறள்களையும் ஏழு பிரிவுகளாகப் பிரித்து படிக்க இம்மென்பொருள் உதவுகிறது. அவையாவன,

(1) குறளின் எண் கொடுத்துக் குறள் படிக்க :

இப்பிரிவின் அடிப்படையில், கற்பவர் எந்த எண்ணைக் கொடுக்கிறாரோ, அந்த எண்ணுக்கு உரிய குறளை கணிப்பொறி திரையில் எழுதும். அக்குறளுடன் அது சார்ந்திருக்கும் பால் (அறம் , பொருள், இன்பம்) . அக்குறளின் அதிகாரப் பெயர் மற்றும் இயல் அனைத்தையும் கணிப்பொறி கொடுக்கும். கொடுக்கப்பட்ட குறளின் விளக்க உரை வேண்டுமெனில் அப்பொருளையும், அதே குறளுக்குக் கீழ் கணிப்பொறி எழுதும். ஒவ்வொருவரின் படிக்கும் வேகம் மாறுபடும் என்பதால் கற்பவர் படித்து முடிக்கும் வரை குறளும் பொருளும் திரையில் தெரிய ஆவன செய்யப்பட்டுள்ளது.

(2)முதல் எழுத்து கொடுத்துக் குறளைப் படிக்க :

இப்பிரிவில், கற்பவர் குறளின் முதல் எழுத்தையோ அல்லது எழுத்துக்களையோ கொடுக்க வேண்டும். கொடுத்த பின், கணிப்பொறி அவ்வெழுத்துக்களில் தொடங்கும் குறள்களைத் தேடி ஒன்றன் பின் ஒன்றாகத் திரையில் எழுதும். இறுதியில், கொடுக்கப்பட்ட எழுத்தில், எழுத்துக்களில் தொடங்கும் குறள்களின் மொத்த எண்ணிக்கையைத் தரும்.

(3) இறுதி எழுத்து (கள்) கொடுத்து குறள் படிக்க :

இப்பிரிவின் அடிப்படையில், கற்பவர் குறளின் இறுதி எழுத்தையோ, எழுத்துக்களையோ கொடுக்க வேண்டும். கொடுத்த பின், கணிப்பொறி அவ்வெழுத்துக்களை இறுதி எழுத்துக்களாகக் கொண்ட அனைத்து குறள்களையும் ஒன்றன் பின் ஒன்றாகத் திரையில் எழுதும். முடிவில் கொடுக்கப்பட்ட எழுத்து, எழுத்துக்களில் முடியும் குறள்களின் மொத்த எண்ணிக்கைத் தரும்.

(4) குறளின், இரண்டாம் அடியின் முதல் எழுத்தையோ, எழுத்துக்களையோ கொடுத்து குறள் படிக்க:

இப்பிரிவில் கற்பவர், குறளின் இரண்டாம் அடியின் முதல் எழுத்தையோ, எழுத்துக்களையோ கொடுக்க, இரண்டாம் அடியில், கொடுக்கப்பட்ட எழுத்தையோ, எழுத்துக்களையோ கொண்டு தொடங்கும் அனைத்து குறள்களையும் ஒன்றன் பின் ஒன்றாக திரையில் கணிப்பொறி எழுதும். இறுதியில் மொத்தம் எண்ணிக்கையையும் தரும்.

(5)அதிகார எண் கொடுத்து குறள்களைப் படிக்க:

இப்பிரிவில் குறளில் உள்ள 133 அதிகாரங்களில் எந்த எண்ணைக் கொடுத்தாலும், அந்த எண்ணுள்ள அதிகாரத்தின் அனைத்து குறள்களையும் திரையில் காணலாம்.

(6) அதிகாரப் பெயர் கொடுத்து குறள்களைக் காண:

இப்பிரிவின் கீழ் கற்பவர் அதிகாரத்தின் பெயரைக் கொடுத்தால், கணிப்பொறி அந்த அதிகாரத்தைக் கண்டுபிடித்து, அதில் உள்ள அனைத்துக் குறள்களையும் ஒன்றன்பின் ஒன்றாக எழுதும்.

(7) உதடுகள் ஒட்டாமல் வாசிக்கக் கூடிய குறள்களைக் காண:

இப்பிரிவின் கீழ் கணிப்பொறி உதடுகள் ஒட்டாமல் வாசிக்கப்படும் குறள்களை ஒன்றன்பின் ஒன்றாக திரையில் எழுதும். உதடுகள் ஒட்டாத குறள்களை கணிப்பொறி எவ்வாறு கண்டுபிடிக்கும் என்ற சந்தேகம் எழ வாய்ப்புண்டு.எந்தக் குறளில் 'ப' மற்றும் 'ம' வரிசை எழுத்துக்கள் வரவில்லையோ அக்குறள்கள் அனைத்தும் , உதடுகள் ஒட்டாமல் வாசிக்கப்படும் குறள்களே ஆகும்..எனவே 'ப' மற்றும் 'ம' வரிசை எழுத்துக்கள் வராத குறள்களை எழுதச் சொல்லி இக்குறள்களைப் பெறலாம். திருக்குறளில் மொத்தம் 28 குறள்கள் இவ்வகையைச் சார்ந்திருக்கின்றன.

சிறப்பம்சம்

1. இம்மென்பொருளை இங்குள்ள சில கல்லூரி மற்றும் பள்ளி ஆசிரியர்கள் குறிப்பாக தமிழாசிரியர்கள் பார்த்தனர். மேலும் கல்வி விரிவுப் பணித்துறையில் நடத்தப்பட்ட பணியிடைப் பயிற்சியில் ஆசிரியர்களுக்குக் காட்டப்பட்டது. அவர்கள் கூறிய கருத்துக்களின் தொகுப்பே இப்பகுதியாகும். இந்நிரலை இயக்கத் தனியாக ஒரு விரற்கட்டைப் பட்டை (Key Board) தேவையில்லை. கணிப்பொறியில் இணைக்கப்பட்டுள்ள ஆங்கில விரற்கட்டைப் பட்டையே போதுமானது. தமிழ்த் தட்டச்சில் எந்தெந்த இடத்தில் எந்தெந்த எழுத்து இருக்கின்றதோ அதே நிலையில் இங்கு எழுத்துக்கள் அமைக்கப்பட்டிருக்கின்றன. உதாரணமாக 'm' ஐ அழுத்த 'அ' வரும். M-ஐ அழுத்த 'ஆ' வரும. தட்டச்சு தெரியாதவர்களும் இயக்க ஒரு அட்டவனை தயாரிக்கப்பட்டுள்ளது. அதில் எந்த ஆங்கில எழுத்தை அழுத்த என்ன தமிழ் எழுத்து வரும் எனக் கொடுக்கப்பட்டுள்ளது. அதன் உதவியால் நாம் எழுத்துக்களைக் கொடுக்கலாம்.

2. புத்தகங்களில் வரிசையாகப் பல குறள்கள் கொடுக்கப்பட்டிருக்கும். இது கற்பவரின் கவனச் சிதறலுக்கு வழி வகுக்கும். அதை மனதில் கொண்டு, ஒரே ஒரு குறளை மட்டும் திரையில் எழுதச் செய்யபட்டுள்ளது. கற்பவரின் கவனச் சிதறல் குறைக்கப்படுகிறது. கற்றலுக்கும் நல்ல சூழ்நிலை கிடைக்கிறது.

(3 எந்தக் குறளை எழுதும் போதும், அக்குறளின் அதிகாரப் பெயர், இயல் பால் முதலியன சேர்ந்து திரும்பத் திரும்ப வருவதால் அவை கற்பவர் மனதில் நன்கு பதிய வாய்ப்புள்ளது.

(4) எழுத்துக்களைக் கொடுக்க குறள் பார்த்தல், உதடு ஒட்டாமல் வாசிக்கப்படும் குறள்களைக் காணுதல் போன்ற புதிய கருத்துக்கள் இதில் உள்ளதால் இம்மென் பொருள் கற்பவருக்கு ஒரு தூண்டுகோலாக அமைகிறது.

(5) கற்பவர்களுக்கு ஏற்ப கால அவகாசம் கொடுக்கப்பட்டிருப்பதால் அனைவரும் குறளை கற்க ஏதுவாக அமைகிறது.

(6) புதிய கருத்துக்களை அறிந்துகொள்ள உதவுகிறது. உதாரணமாக, உதடுகள் ஒட்டமல் வாசிக்கப்படும் குறள் மொத்தம் 28 உள்ளது என்பதை இதன் மூலம் அறிய முடிகிறது.இம்மென்பொருளைப் பார்த்த ஆசிரியர்கள் கொடுத்த ஆலோசணைகள்

1. வினாக்கள் கேட்டு, மதிப்பீடு செய்யும் வகையில் இருந்தால் நல்லது.
2. குறள்கள் எழுதப்படும்போதே, அச்சொற்களும் உச்சரிக்கப்படுமானால் இன்னும் சிறப்பாக இருக்கும்.

நிறைவு:

இந்த நவீன விஞ்ஞான யுகத்தில் நாம் கணிப்பொறி உலகில் வாழ்ந்து கொண்டிருக்கிறோம் என்று கூறுவது சாலப் பொருந்தும். எங்கு நோக்கினும் சக்தியடா என்று கவி பாரதி பாடினார் அன்று; இன்றோ எங்கு நோக்கினும் கணிப்பொறியாக உள்ளது. கல்வித் துறையிலும் கணிப்பொறியின் முக்கியப் பங்கினை நாம் நன்கு அறிவோம். கணிப்பொறி துணையுடன் கற்பித்தல் மேலும் கணிப்பொறி துணையுடன் கற்றல் என்று கற்றல், கற்பித்தல் ஆகிய இரு நிலைகளிலும் மிகப் பெரிய பங்கு வகிக்கிறது.

இந்நிலையில் உலகப் பொது முறையாம் திருக்குறளைக் கணிப்பொறி மூலம் கற்றல், கற்பித்தல் என்ற இந்த மென்பொருளின் உபயோகம் கல்வி நிலையங்களில் மிகுந்த வரவேற்பு பெற்று வருகிறது. இந்தக் கருத்தரங்கிற்கு வருகைபுரிந்துள்ள கணிப்பொறியாளர்களின் மிகச் சிறந்த ஆலோசனைகளுடன் இது இன்னும் சிறப்புமிக்க மென்பொருளாக ஆக்கப்பட்டு அனைவருக்கும் பயன்படும்படி அமைக்கப்பட வேண்டும் என்ற ஆவலுடன் இதைப் படைப்பதில் மகிழ்வு கொள்கிறேன்.

தமிழில் பல்ஊடக சாதனம்
(Multimedia in Tamil)

அ. இளங்கோவன்
கேட்கிராஃப் கம்ப்யூட்டர்ஸ்,
சென்னை

பல்ஊடக சாதனத்தின் அங்கங்கள்:

நாம் அன்றாட வாழ்க்கையில் ஒரு விஷயத்தை மற்றவருக்கு எடுத்துரைக்க, விளக்க மற்றும் கற்றுக் கொடுக்க உபயோகபடுத்தும் பல்வேறு யுத்திகளும் பல்ஊடக சாதனத்தின் அகங்களாக (Input) பயன்படுத்தப்படுகின்றன. உதாரணத்திற்கு வாசகங்கள் தாய்மொழியிலோ அல்லது வேறு மொழியிலோ விளக்கவுரைகளாக (மானிட்டரி)ல் காணலாம். அவ் வாசகங்கள் அவரவர் தாய்மொழியிலேயே அதில் பொருத்தியிருக்கும் Speaker மூலம் கேட்டும் அறியலாம். அதற்கு சம்பந்தப்பட்ட பலவித வரைபடங்களையும் புகைப்படங்களையும் மானிட்டரில் கண்டு தெளிவு பெறலாம். அந்த வரை படங்களுக்கு உயிரோட்டம் (Animation) கொடுத்து அதன் இயங்கும் தன்மைகளை புரிந்து கொள்ளலாம். மேலும் உண்மைக்காட்சிகளை வீடியோ படங்கள் மூலம் மேற்கண்டவற்றுடன் மானிட்டரில் கண்டு, பன்மடங்கு தெளிவு பெறலாம். மேற்கண்ட எல்லாம் செவிக்கு தேனாகப் பாயும் சங்கீதத்துடன் இணைய ரசித்து கற்றக் கொள்ளலாம். மனதில் நிலையாக இருத்திக் கொள்ளலாம். மேலும் இவை யாவற்றையும் இயக்க தேவை கம்ப்யூட்டர் அறிவோ, பயிற்சியோ, படிப்பறிவோ அல்ல ! நமது ஒரு விரல் மூலம் மானிட்டரை தொட்டு இயக்கி (Touch Screen Monitor) நமக்கு வேண்டிய விஷயங்களை, நமக்கு வேண்டிய அளவு, நமது தாய்மொழியில், நமக்கு தேவைப்பட்ட நேரத்தில் தெரிந்து கொள்ளலாம், அதுவும் நமக்கு ஒரு ஆசிரியர் அருகாமையில் இல்லாமலும் கூட ஆச்சரியமாக உள்ளதா? அதுதான் இன்று கணிப்பொறியுகத்தில் நமக்குக் கிடைத்துள்ள வரப்பிரசாதம். - தமிழ் பல்ஊடக சாதனம் (Tamil Multimedia System).

பிரயோகங்கள்:

கல்வி

தமிழ் பல்ஊடக சாதனத்தின் பிரதான பிரயோகம் கல்வித்துறையே, மழலைக்கல்வியில் தொடங்கி, குழந்தைக்கல்வி, மாணவர் கல்வி, பன்மொழிக் கல்வி, தொழில் நுட்பக்கல்வி, மருத்துவக்கல்வி மற்றும் முதியோர் கல்வி, ஊனமுற்றோர் (செவிடு, ஊமை) கல்வி, படிப்பறிவில்லாதவர்க்கும் கல்வி எனப்படும், எல்லாவித கல்வித்துறைகளுக்கும் பல் ஊடக சாதனத்தின் பங்கு மேன்மையானதாகும். உதாரணத்திற்கு, ஒரு தாய் தன் குழந்தைக்கு அரிச்சுவடி புத்தகத்திற்குப் பதிலாக ஒரு மானிட்டரில் அம்மா படத்தைத் தொட்டால், கணிப்பொறி "அம்மா" என்று ஒலிக்கும், "அ" என்ற எழுத்தைத் தொட்டால் "அ" என்ற ஒலியும், "அம்மா" என்ற வார்த்தையைத் தொட்டால் "அ, ம், மா - அம்மா" என்று குழந்தைக்கேற்ற மெல்லிசையுடன் ஒலித்தால் அக்குழந்தையின் கற்றுக் கொள்ளும் ஆர்வம் எத்தனை மடங்காகும் என்பதில் யாருக்கேனும் சந்தேகம் உண்டோ?

மழலைக் குழந்தையையே மயக்கவல்ல பல்ஊடக சாதனம் மற்ற கல்வித் துறையினரைத் தன் வசப்படுத்தி கற்பவரின் ஆர்வத்தையும், ஊக்கத்தையும் பன்மடங்காக்கும் என்பதில் சந்தேகமில்லை.

பயிற்சி (Training)

பல் ஊடக சாதனம் பயிற்சித் துறையில் பெரும் பங்கு வகிக்க இருக்கிறது. குறிப்பாக வேலை வாய்ப்புப் பயிற்சி, விவசாயப் பயிற்சி, குடிசைத் தொழில் பயிற்சி, செய்முறைப் பயிற்சி, அறிவியல் பயிற்சி, மருத்துவப் பயிற்சி மற்றும் பல பயிற்சித்துறைகளுக்கும் பல்ஊடக சாதனம் ஒரு திருப்பு முனையாக அமைய உள்ளது.

தொழில் (Industry)

தொழில் துறையில் பல் ஊடக சாதனம் ஒரு முக்கியமான இடத்தை வகிக்க உள்ளது. குறிப்பாக எல்லா தொழில் துறைகளுக்கும் முக்கியமான விபத்துத்தடுப்பு, முதலுதவி, தொழிலாளர் பயிற்சி, பாதுகாப்பு பயிற்சி எந்திர இயக்க விளக்கங்கள், தரக்கட்டுப்பாடு (Quality Control), கழிவுப்பொருள் கட்டுப்பாடு (Waste Control) உற்பத்திச் செலவு குறைப்பு (Cost Control) போன்ற துறைகளில் பல்ஊடக சாதன விளக்கங்கள் மற்ற விளக்கங்களை விடப் பன் மடங்கு சக்தி வாய்ந்ததாக அமையும்.

வியாபாரம் (Trade)

இன்று முற்போக்குக் கம்பெனிகள் பல பல்ஊடக சாதனத்தைத் தழுவி தங்களின் நிறுவன அறிமுகத்திற்கும், பங்குமுதல் வெளியீட்டிற்கும், வியாபாரப் பொருள் அறிமுகத்திற்கும் பயன்படுத்தி பெரும் பலன் அடைந்து வருவதை நீங்களே கண்டிருக்கலாம். பல்ஊடக சாதனத்தை எல்லா பெரிய, சிறிய நிறுவனங்களும் தழுவி பலன் அடைய இருக்கின்றன.

சமூக நலன் (Social Welface)

நமது நாட்டின் பெரும் பிரச்சனைகளாகக் கருதப்படும் குடும்பக் கட்டுப்பாடு, கொள்ளை நோய்த் தடுப்பு,சத்துணவுக் குறைவு, சுற்றுப்புறச் சூழல் பாதுகாப்பு, மரம் பேணுதல், சமூக விழிப்புணர்வு, தேசிய ஒற்றுமை ஆகிற பலவற்றிற்கும் அரசாங்கம் செலவிடும் பெரும் பங்கு பிரசாரங்களுக்கு ஆகும். தாய்மொழி பல்ஊடக சாதனம் இந்த பிரசாரங்களின் சக்தியைப் பன்மடங்காக உயர்த்தி பெரும்பாலான மக்களுக்கு குறைந்த செலவில் இதன் பயன்கள் கிடைக்க வழிவகுக்கும்.

பொழுது போக்கு (Entertainment)

கணிப்பொறி விளையாட்டுகள் (Computer Games) சிறுவர் முதல் முதியவர் வரை எல்லோரையும் ஈர்த்துள்ளன என்பதில் யாருக்கும் சந்தேகம் இருக்காது என நினைக்கிறேன். விளையாடுபவர்க்கு விளையாடும் சமயத்தில் உள்ள முழு ஈடுபாடு பல் ஊடக சாதனத்திற்கு கிடைத்த முதல் வெற்றியாகும். தற்பொழுது பல்வேறுபட்ட விளையாட்டு, நடனம், சங்கீதம், சித்திரப் பயிற்சிகள், உடற்பயிற்சிகள் போன்ற துறைகளில் பல் ஊடக சாதனம் பெரும் பங்கு வகிக்க இருக்கிறது.

சிறப்புப் பலன்கள்

பல் ஊடக சாதனத்தின் முதல் சிறப்பு அம்சம் ஐம்புலன்களின் ஈடுபாடு. அவரவர் தாய்மொழியிலேயே பயிற்சியும் கல்வியும் அவரவர் தேவைக்கேற்ற அளவு அவரவர் விருப்பப்படும் பொழுது கிடைக்கச் செய்வது இதன் தனித்தன்மையாகும். இது ஒரு நல்ல ஆசிரியரை பல ஆயிரம் பள்ளிகளுக்கு ஒரே நேரத்தில் கிடைக்க செய்யும் வியக்கத்தக்க சாதனமாக திகழ இருக்கிறது.

இவற்றிற்கெல்லாம் மேலாக உபயோகிப்பவரின் தன்மைக்கேற்ப (User Interaction) ஆசிரியர் மாணவரிடையே விருப்பு வெறுப்பு இல்லாத, எத்தனை முறை கேட்டாலும் சலிப்பில்லாத விளக்கங்களை அளிக்கவல்ல சிறந்த கல்வி மற்றும் பயிற்சி சாதனமாக இந்த பல்ஊடக சாதனம் அமைய உள்ளது. சமுதாய, பொருளாதார, கல்வி, தொழில் விவசாயத்துறைகளில் கல்வியும், பயிற்சியும் தமிழிலேயே நம் மக்களைச் சென்றடைய தமிழ் பல் ஊடக சாதனம் தயாராகி விட்டது.

இச்சக்தி தமிழ்க் கிராமங்களில் ஊடுருவிச் சென்று பாமர தமிழ் மக்களிடையே ஓர் விழிப்புணர்வை ஏற்படுத்தப் போவதில் ஐயமில்லை.

தமிழில் தகவல் தளத்தின் அவசியம்

(Need for Database in Tamil)

டி.எஸ். முத்துகிருஷ்ணன்
சாஃப்ட்வேர் ரிசர்ச் குரூப்

மனிதனின் வாழ்க்கை பெரும்பாலும் தகவலின் மேல்தான் உள்ளது, எந்த ஒரு செயலுக்கும்,தொழிலுக்கும் பலவிதமான தகவல்கள் அவசியமாயுள்ளன.

அலுவலகங்கள், கடைகள், பள்ளிகள், நூல்நிலையங்கள், தொழிற்சாலைகள் போன்ற இடங்களில் பலவிதமான தகவல்கள் பெறுவதற்கும் அவற்றைச் செயல்படுத்தவும் கணிப்பொறியினை உபயோகப்படுத்துகிறார்கள்.

எல்லா தகவல்களையும் கணிப்பொறியில் போட்டு அவைகளை உபயோகத்ததிற்கு தகுந்த மாதிரி பிரித்து வைக்க வேண்டும். இவ்வாறு எல்லா தகவல்களையும் ஒருங்கிணைந்து வைத்திருப்பதே தகவல் தளமாகும்.

நமது நாட்டில், இன்றும் கிராமப் புறங்களில் இருக்கும் மக்களுக்கு தாய் மோழியான தமிழ்தான் நன்கு தெரியும் அவர்களும் உபயோகிக்கும் அளவு இத்தகவல் அடித்தளம் தமிழிலும் கணிப்பொறியில் ஏற்றவேண்டியது அவசியம். வேண்டிய தகவல்களை தமிழில் பெற அமைப்பு வேண்டும். அதுமட்டுமன்றி தகவல்களை தமிழில் இருந்து வேறு எம் மொழிக்கும் மாற்றவும் செய்யலாம். வேறுமொழியில் உள்ள தகவல்களை தமிழிலும் கணிப்பொறியில் தமிழிலும் பெறஇயலும் இந்த அளவிற்கு கணிப்போறி ஏற்றம் கண்டிருக்கிறது. எல்லா புது தகவல்களையும் தகவல் அடித்தளத்தில் சேர்க்கலாம், உபயோகமற்ற தகவல்களை நீக்கவும் செய்யலாம்.

பெரிய கடைகளில் இத்தகவல் அடித்தளத்தின் மூலம் வாடிக்கையாளர்கள் தங்களுக்கு தேவையானவைகளை கணிப்பொறியின் மூலம் தேர்ந் தெடுத்துக்கொள்ளலாம். வாங்கும் பொருட்களின் புறத்தோற்ம், உள்தோற்றம், பொருட்களின் அமைப்பு ஆகியவற்றையும் கணிப்பொறியின் மூலம் காண இயலும் (Through Multimedia). அதுமட்டுமன்றி, கணிப்பொறியே அப்பொருட்களை பற்றி விளக்கமும் கொடுக்கும் அதை நாம் கேட்கவும் செய்யலாம், இன்னும் சில இடங்களில் நாம் கேட்கும் கேள்விகளுக்கும், கணிப்பொறி பதிலும் கொடுக்கும்.

இத்தகவல்அடித்தளம் எல்லா இடங்களிலும் அவசியமாயுள்ளது எல்லோருக்கும் அவசியமாயுள்ளது. எக்காலத்திற்கும் அவசியமாயிருக்கிறது, எப்பகுதியிலும் அவசியமாயிருக்கிறது. இத்ததகவல் அடித்தளம் முறைப்படி உபயோகப்படுத்தப் படின், மக்களின் வாழ்க்கைத்தரம் உயரும், வணிகம் பெருகும், தரமான பொருட்கள் நியாயமான விலையில் கிடைக்கும் அரசாங்கம் திறம்பட செயல்பட இயலும்.

நூலகங்களில் ஏராளமான நூல்கள் உள்ளன நூல்கள் ஒவ்வொன்றினையும் அதனுடைய தரப்பிற்கு ஏற்றவாறு பிரித்து, வெளியிடும் அச்சகத்தைப் பொறுத்து, ஆசிரியரைப் பொறுத்து என பல்வேறு வகையில் அட்டவணைப்படுத்தலாம். இவ்வாறு செய்தலில் பயன்யாதெனில், நூலகத்தை உபயோகிப்பவர் தனக்கு

வேண்டியவற்றை சரியாக தேர்ந்தெடுத்து படிக்க இயலும். நூல்கள் நன்கு பயன்படுத்தப்படும். இதன் மூலம் நாட்டின் கல்வியறிவும் மேம்படையக்கூடும். மேலும் ஒரு குறிப்பிட்ட சொல் அல்லது தகவல் எந்த நூலில் உள்ளது என்பதையும் தகவல் அடித்தளத்தின் மூலம் அறியலாம். இன்று நம் நாட்டில் பல்வேறு வகையான பாடத்திட்டங்கள் உள்ளன. பாடங்களை பள்ளியில் மட்டும் அன்றி வீட்டிலும் நன்கு படித்து நல்ல மதிப்பெண்கள் வாங்குவது மேற்படிப்புக்கு அத்தியாவசியமாகியுள்ளது. இதற்கும் இந்நாளில் தகவல் அடித்தளத்தின் மூலம் எல்லா பாடங்களையும் கணிப்பொறியின் மூலம் கற்றுக்கொள்ள சாதனங்கள் உள்ளன.

எந்த வகையான கல்வி கற்றால் எந்த வகையில் வரலாம்?, அதன் குணாதிசயங்கள் என்ன போன்ற விவரங்களையும் தகவல் அடித்தளத்தின் மூலம் அறியலாம்.

இதுவரை தகவல் அடித்தளத்தின் நன்மைகளை ஆராய்ந்தோம். ஆனால், தகவல் அடித்தளம் எவ்வாறு அமைக்கப்படுகிறது? எவ்வாறு பராமரிக்கப்படுகிறது? என்பவற்றை சற்று ஆராய்வோம்.

வரும் தகவல்கள் எல்லாவற்றையும் எப்படி பிரித்து வைத்தால், உபயோகிப்பவர்களுக்கு அதிக பலனை கொடுக்க முடியும் என்பதனை ஆராய்ந்து அதற்கேற்றபடி தகவல்கள் பிரிக்கப்பட்டு அத்தகவல்களை கணிப்பொறியில் ஏற்றுகின்றனர். அம்மாதிரியே, நூல்கள் வந்தாலும் அது எம்மாதிரியான தகவல்கள் எல்லாம் கொடுக்கமுடியும் என ஆராய்ந்து அம்மாதிரியான தகவல்களை எல்லாம் வகைப்படுத்தி கணிப்பொறியில் எற்றுகின்றனர்.

அம்மாதிரியே சேகரித்த தகவல்களை எல்லாம் எப்பொழுதும் மாறாமல் இருப்பதில்லை. சில தகவல்கள் மாறிக்கொண்டேயுள்ளன. அம்மாதிரி மாறியுள்ள தகவல்களை, தகவல் அடித்தளத்தில் மாற்ற வேண்டும். மாற்றும் வேளையில், பழைய தகவல்களை அழிக்கவோ அல்லது வைத்துக் கொண்டிருக்கவோ செய்ய வேண்டும். இந்த மாறுதல்கள் தகவல் அடித்தளத்தில் செய்யப்படாவிடின் தகவல் அடித்தளம் உபயோகமற்றதாகிவிடும்.

மிகவும் சிரமப்பட்டு சேகரித்த தகவல் அடித்தளத்தினை நன்கு பராமரித்தால், அதன் மூலம் ஏற்படும் விளைவுகள் எண்ணிலடங்கா ! இது எம்மாதிரி எனில், நாம் ஒரு வாகனம் அதிக விலை கொடுத்து வாங்குகிறோம். அதிக விலை கொடுத்துள்ளோம் என்பதனால் உபயோகிக்காமலோ அல்லது பராமரிக்கப்படாமலோ இருந்தால், அவ்வாகனம் நாளடைவில் உபயோகமற்றதாகிவிடும். வாகனத்தை உபயோகிக்கவும் வேண்டும் பராமரிக்கவும் வேண்டும்.

தற்போது தகவல் அடித்தளம் பெரும்பாலும் ஆங்கிலத்திலே அமைந்துள்ளது. அதை தமிழில் தருவதற்கு பல நிறுவனங்கள் சாதனங்களை கொண்டுவந்துள்ளன. தமிழிலேயே கணிப்பொறிக்கு தேவையானவற்றை கொடுத்து, வாங்க, வசதிகள் உள்ளன. கணிப்பொறியின் வளர்ச்சிக்கு ஏற்றவாறு, அதை தமிழிலும் உபயோகப்படுத்தகூடிய அளவிற்கு ஆணைத்தொடர்கள் உள்ளன. தமிழின் அற்புத குணங்களை கொண்டு இச்சாதனைகள் படைக்கப்பட்டுள்ளன.

எதையுமே வாங்குவதை விட நன்கு பராமரித்தால்தான் நல்ல பயன்களை அடைய முடியும்.

RESEARCH IN TAMIL LITERATURE USING COMPUTERS

P.PANDIAYARAJA
Department of Computer Science
American College, Madurai

The digital computer is a great boon to the statistical study of language and literature. The computer's ability to search, sort or process even a vast amount of text material with great ease and speed is used to the fullest advantage in many areas of literary and linguistic studies. Early attempts of numerical analysis of language data received a great fillip after the advent of modern computer, thoroughly revolutionising the area of computational linguistics.

Dr. P.C. Ganeshsundaram, Dr. Namperumal and Dr.R. Shanmugam were some of the early users of computers in Tamil studies. All of them used the then available FORTRAN II language which is relatively unsuited for string processing. Yet their studies were similiar to attempts to walk on untrodden soil since they had to struggle not only with the data but also with the limitations of the Computer language they were using. The Tamil University's work on 'Computer analysis of Tirukkural' can be said to be the first major attempt on the use of computers in the analysis of Tamil literature in India. This author had made some studies on the Tamil language and literature using Computers and has presented some papers based on his study. This article aims at highlighting the methodology followed by the author and share his experiences.

The first ever attempt by the author to use computers for his study on Tamil language was in early 80 's . He was inspired by a remark made by Gustuv Herdan in his book Type-Token Mathematics. According to Herden, a language is an apparatus for the communication of thought within a speech community consisting of persons using the same phonemic system, lexicon and grammar. However, he claims that quantitative studies have shown that there is another common factor, namely a farreaching similarity between members of the speech community, not only as regards the phonemic system, lexicon and grammar they use, but also as regards the frequency of use of particular phonemes, lexicon items (words) and grammar forms and structures. In other words, a similarity not only in <u>what is used</u>, but also in <u>how often it is used</u>.

Taking this as the hypothesis, a study was undertaken to test the hypothesis in modern written Tamil. So, two authors, namely ti.Janakiraman and Kannadasan were chosen. One novel written by each, namely Amma: Vantha:l and U:maiyan ko:ttai were selected. Words were chosen from these novels using random sampling technique. Then came the problem of transliterating them in English to be read by the computer through a programming language. Initially, the usual way of writing Tamil in Roman script (K,NG,C,NJ,T., N.,TH,N,P,M,TR,N,Y,R,L,V,Z,L,A,A:,I,I:,U,U:,E,E:,AI,O,O:,AU) was followed by the author. It is evident that here that the use of diacretic marks has become almost unavoidable especially in vowels. The combination AA,II,UU,EE,OO were thought of but not used because of the possible confusion in their pronounciation. For example, EE combination in

the English words like tree, Spree, see etc., have to be pronounced as i: in English. Similarly the OO combination in the words taboo, zoo., would confuse with the O: sound in Tamil. Hence it was decided to use the usual diacretic mark used in Linguistics also. But these diacretic marks make the sound recognition in the Program more complicated. For example, the letter T should be recognised only in combination with the next symbol (asT.,TH or TR).

In spite of these difficulties, programs were written in Pascal language to count the various phonemes in the two samples. Apart from this, the texts of three works in pathuppa:ttu namely, Thirumuruka:rruppatai, netunalva:tai and Mullaippa:ttu were fed into the computer and through a program, they were split into words which were sorted to enable the study of the vocabulary content of these texts. The type / token ratios for these texts were calculated and compared with those of various chapters of New Testament in the Bible as presented by Herden. This study also enabled the author to verify the law of relative growth in literary products which in quite similar to the law of relative growth in Biology. Finally, through another program, the ci:r and talai counts in these three texts were made and their distributions were compared for their significance of difference. All these results were presented in a paper titled, 'Statistical analysis of some Linguistic features in Tamil Literature' at the Third All India Conference in Tamil Linguistics held at Thanjavur in February, 1988.

After this experience, the author drastically changed the pattern of transcription from Tamil to English and has developed his own method on one sound-one key basis. Using this method, the author has stored enormous amount of data, starting from Tholka:ppiam and upto the Modern Tamil in Magazines and Newspapers. He has also written many programs in Pascal and is actively involved in the chronological study of various features of Tamil.

In the mean time, the author has presented some papers also based on his computer analysis of data in Tamil. One such paper was at Pondicherry Institute of Language and Culture where he presented a paper on "The association between Sound and Meaning". Various pieces of texts from sangam poetry depicting different feelings like love, sorrow etc., were analysed through programs in the computer and definite preference (by the authors) for certain sounds for certain feelings were brought out by the study. The ci:r and thalai distribution of Akaval Pa: in all the texts in paththuppa:ttu were analysed and a diachronic study of their distribution in sangam period was made and presented in the fifth annual conference of the International Institute of Tamil culture and Research held in June 1994 at Trichirappalli.

All those studies would not have been possible but for the extensive use of computer. The great limitations of text processing through a programming language are

1) The creation of data files

2) The inability of the programs to read Tamil Characters.

The process of feeding the data which is the most crucial one is the excruciating one also. Even if this could be done easily by sophisticated input devices whether it would be possible to make a programming language read them is doubtful. If these two problems are solved that would really be a big leap in the researches in Tamil using computers.