அறி

அகரமுதலிச் சொற்களில் புள்ளியியல்

சூர்யா முருகேசன்

கட்டுரை எண் 001
அறி – 03/ 2021

கார்க்கி ஆராய்ச்சி நிறுவனம்

அகரமுதலிச் சொற்களில் புள்ளியியல்

சூர்யா முருகேசன்

கட்டுரைச் சுருக்கம்

பண்டைய காலம் முதல், மொழிகளின் எழுத்துப் புள்ளிவிவரங்களை ஆராய்வதில் பலர் ஆர்வம் கொண்டிருந்தனர். கி.பி 300களில் மறைகுறியீட்டியலில் இத்தகவல்கள் பயன்படுத்தப்பட்டன. தொடக்கக் காலப் பதிப்புத் தொழில்நுட்பம் முதல் தற்காலத் தொடுதிரைக் கருவிகளின் உள்ளீட்டு அமைவு வரை இவ்விவரங்கள் பயன்படுத்தப்படுகின்றன. அளவறி மொழியியல், சொல் விளையாட்டு போன்ற வேறுபட்ட பல தளங்களிலும் இத்தகவல்கள் மிகுந்த பயனளிக்கக்கூடியவை. இவ்வுரையில், இணையவழித் தமிழ் அகராதியான "சொல்"இல் உள்ள சொற்களின் எழுத்துப் புள்ளியியல் விவரங்கள் ஆராயப்படுகின்றன. சொல் விளையாட்டுக்கள் வடிவமைப்பு, புள்ளியியல் ஆராய்ச்சிகள், கல்வி போன்ற பல துறைகளில் இத்தரவுகள் பயனுள்ளவையாக இருக்கும்.

குறிப்புச் சொற்கள்

தமிழ் எழுத்துக்கள் - Tamil Alphabet, புள்ளியியல் - statistics, எழுத்து வகை - letter type

அறிமுகம்

உலகின் தொன்மையான மொழிகளுள் ஒன்றாகிய தமிழ்மொழி, தற்காலத்தில் பயன்பாட்டிலிருக்கும் மொழிகளிலே மிகப் பழமையானதாகும். தமிழின் இலக்கணம், 3500 ஆண்டு பழமையானது. தமிழிலக்கியத்தின் காலக்கோடு, கி.மு 500இல் இருந்து, தற்காலம் வரை நீள்கிறது.

2004ஆம் ஆண்டு, இந்திய அரசு தமிழைச் செம்மொழியாக அறிவித்தது. ஒரு மொழி, கீழ்க்காணும் மூன்று வரையறைகளை நிறைவு செய்யும் போது, செம்மொழியாக ஏற்கப்படுகிறது.

i. அம்மொழியின் தோற்றம் மிகத் தொன்மையானதாக இருக்க வேண்டும்.
ii. மொழி மரபுகள், பிறமொழி மரபுகளின் சார்பற்று அமைந்திருக்க வேண்டும்.
iii. கணிசமான அளவில் தொல்லிலக்கியங்களை அம்மொழி கொண்டிருக்க வேண்டும். [1]

இந்திய மாநிலங்களான தமிழ்நாடு மற்றும் புதுச்சேரியிலும், இலங்கை, சிங்கப்பூர் போன்ற நாடுகளிலும் தமிழ் ஆட்சி மொழியாக உள்ளது. தென் ஆப்பிரிக்கா, மொரீசியசு, பிசி, மலேசியா போன்ற நாடுகளில் தமிழ் பேசும் மக்கள் கணிசமான அளவு உள்ளனர் [1].

இரட்டைவழக்கு என்பது, ஒரு மொழி, இரண்டு வடிவங்களில் பயன்படுத்தப்படுவதைக் குறிக்கின்றது. தமிழின் இரு வழக்குகள் - உரைநடை வழக்கு, பேச்சு

வழக்கு என்பனவாகும். உரைநடை வழக்கு, பெரும்பாலும் முறைசார் பயன்பாடுகளில் பயன்படுத்தப்படுகிறது. இயல்பான வாழ்க்கையில் மக்கள் தங்களுக்குள் உரையாட பேச்சுத்தமிழைப் பயன்படுத்துகின்றனர்.

இவ்வுரையில், "சொல்" அகராதியிலுள்ள சொற்களுக்கான எழுத்துப் புள்ளியியல் விவரங்கள் விவரிக்கப்படுகின்றன. சொல் விளையாட்டுக்கள் வடிவமைக்கவும், விளையாடவும், பல மொழிகள் ஒரே எழுத்தமைப்பைப் பகிர்ந்து கொள்ளும் பொழுது உரைநடை எம்மொழியைச் சார்ந்தது என்பதைக் கண்டறியவும் எழுத்துப் புள்ளியியல் விவரங்கள் மிகுந்த பயனளிக்கக்கூடியவை. தரவுப் பாதுகாப்புக்கான (Data security) மறைகுறியீட்டியல் (Cryptography) தொழில்நுட்பத்தில் தரவு மறையாக்குதலிலும் (Encrypt) மறைவிலக்குதலிலும் (Decrypt) எழுத்துப் புள்ளியியல் தரவுகள் பயனளிக்கும் [2]. தனிநபர்களுக்கான கையடக்கக் கருவிகள் பல அறிமுகப்படுத்தப்படும் நிலையில், அவற்றில் உள்ளீடு செய்வதற்கான பல்மொழி மெய்நிகர் தட்டச்சுப்பலகைகள் வடிவமைக்க வேண்டியுள்ளது. அவ்வாறான வடிவமைப்புக்களில் எழுத்துப் பயன்பாட்டுப் புள்ளிவிவரங்கள் இன்றியமையாதவையாகின்றன.

பின்னணி

பண்டைய காலத்திலிருந்தே மறைக்குறியீட்டியல் தொழில்நுட்பத்தில் எழுத்துப் புள்ளியியல் தரவுகள் பயன்படுத்தப்பட்டிருக்கின்றன. அரேபிய கணித அறிஞர் அல் கிண்டி (Al-Kindi), சொற்களில் எழுத்துக்கள் பயன்படுத்தப்படும் விகிதத்தை ஆராய்ந்து (frequency analysis) ஒரெழுத்து மறையீட்டை (monoalphabetic ciphers) மறைவிலக்கும் முறையைக் கண்டறிந்தார். கி.பி 800இல் அறிமுகப்படுத்தப்பட்ட இம்முறை இரண்டாம் உலகப் போர் வரை பயன்பாட்டிலிருந்தது [3]. "Frequency analysis" என்றழைக்கப்படும் இம்முறையின்படி ஒரு மொழியின் எழுத்துப் பயன்பாட்டு விகிதத்தைக் கணக்கிட்ட பின்னர் மறையாக்கப்பட்ட (encrypted) எழுத்துரையில் வரும் எழுத்துக்களின் விகிதத்தைக் கணக்கிட வேண்டும். இவ்விகிதங்களைக் கொண்டு மறையாக்கப்பட்ட எழுத்துக்களுக்குப் பதிலாகப் பிற எழுத்துக்களைப் பதிலீடு (substitute) செய்து மறைவிலக்க(decrypt) முயல வேண்டும் [4].

1992ஆம் ஆண்டில், அமெரிக்க மறைக்குறியீட்டியல் அறிஞர் வில்லியம் ஃரீட்மான் (William Friedman), ஃரீட்மான் அல்லது கப்பா (Friedman test or Kappa test) என்று அழைக்கப்படும் சோதனையைக் கண்டறிந்தார். அல்-கிண்டி [3,4] போன்று எழுத்துப் பயன்பாட்டு விகிதத்தைப் பயன்படுத்தும் இச்சோதனை, மறையாக்கப்பட்ட எழுத்துரை, ஒரெழுத்து அல்லது பல்லெழுத்து (monoalphabetic or polyalphabetic) மறையீடா என்று தீர்மானிக்கப் பயன்படும். மேலும், இச்சோதனையைக் கொண்டு பல்லெழுத்து மறையீட்டான விஜெனெர் (Vigenère) முறையில் பயன்படுத்தப்படும் திறவுச்சொல்லின் (keyword) நீளத்தையும் கண்டுபிடிக்கலாம் [5].

மார்ச் 21, கட்டுரை எண் 0001

பழைய பதிப்புத் தொழில்நுட்பத்தில், எழுத்துரையிலிருந்த ஒவ்வொரு எழுத்துக்கும் எண்ணிக்கை கண்டறிய வேண்டியிருந்தது [6]. தொலைத்தொடர்பில் பயன்படுத்தப்படும் மோர்ஸ் குறியீட்டில் (Morse Code), மிகுதியாகப் பயன்படும் எழுத்துக்களுக்குச் சிறிய குறியீட்டுத் தொடர் (sequence) ஒதுக்கப்பட்டிருப்பது [7], எழுத்துப் புள்ளியியல் தகவல்களின் பயன்பாட்டிற்குச் சான்றாகும். எழுத்துக்களுக்குகான குறியீடுகளைக் கண்டறிய, சாமுவேல் மோர்ஸ் (Samuel Morse), பதிப்பகங்களில் எழுத்துக்களின் பயன்பாட்டை ஆய்வு செய்தார் [7]. ஆங்கில எழுத்துக்களின் பயன்பாட்டுத் தகவல்களையும் மோர்ஸ் குறியீட்டினையும் அடிப்படையாகக் கொண்டு ஆங்கில QWERTY விசைப்பலகை வடிவமைக்கப்பட்டது [6].

அளவறி மொழியியலில் (Quantitative linguistics), ஒரு எழுத்துரையின் மெய் மற்றும் உயிர் எழுத்துக்களின் விகிதத்தைக் கொண்டு, அவ்வுரை, கவிதையா அல்லது உரைநடையா என்று கண்டறிவதை 1467ஆம் ஆண்டு, ஆல்பெர்டி (Alberti) என்ற அறிஞர் எடுத்துரைத்தார் [8]. ஜாக் க்ரீவ் (Jack Grieve) என்பவர், எழுத்தாளர்களின் பண்புக்கூறுகளைக் (Authorship attribution) கண்டறியும் பல படிமுறைகளைப் பரிசோதித்திருக்கிறார். அந்தப் படிமுறைகளில், மற்றக் கூறுகளுடன், எழுத்துக்களின் ஒப்பீட்டுப் பயன்பாட்டு விகிதங்களையும் ஒரு கூறாகக் கையாண்டிருக்கிறார் [9].

[10], [11], [12] போன்ற விசைப்பலகை அமைவு குறித்த பல ஆய்வுகளில் ஈரெழுத்தொருவொலிகளின் பயன்பாட்டு விகிதம் பயன்படுத்தப்பட்டிருக்கிறது. ஆகஸ்ட் ட்வோராக் (August Dvorak), ட்வோராக் விசைப்பலகையை வடிவமைக்க, ஆங்கிலத்தில் மிகுதியாகப் பயன்பாட்டிலிருக்கும் எழுத்துக்களையும், ஈரெழுத்தொருவொலிகளையும், மனிதக் கையின் இயக்கவியலையும் ஆராய்ந்தார் [13]. அடிக்கடிப் பயன்படுத்தப்படும் எழுத்துக்களின் விசைகளை வலிமையான விரல்களால் இயக்குவதாகவும், மிகுதியாகப் பயன்பாட்டிலிருக்கும் ஈரெழுத்தொருவொலிகளின் இரு எழுத்துக்களும் இரு வேறு கைகளிலுள்ள விரல்களால் இயக்குவதாகவும் இவ்விசைப்பலகை வடிவமைக்கப்பட்டுள்ளது [13].

சொல் விளையாட்டுக்களை விளையாட எழுத்துக்களின் பயன்பாட்டு விகிதங்கள் உதவியாக இருக்கும் என்று குறிப்பிடுவதுடன், ஆங்கில எழுத்துக்களின் பயன்பாட்டு விகிதங்களை [14] இன் ஆசிரியர் பட்டியலிட்டிருக்கிறார். ஆங்கிலச் சொல் விளையாட்டான ஸ்காராபில் (Scrabble)இல் எழுத்துக்களுக்கான மதிப்பெண்களை ஒதுக்க, அதனைக் கண்டுபிடித்த ஆல்ஃப்ரெட் மோஷர் பட்ஸ் (Alfred Mosher Butts), செய்தித்தாள்களிலுள்ள எழுத்துக்களில் Frequency analysis முறையை மேற்கொண்டார் [15].

உள்ளீட்டுக் கருவிகள் வடிவமைப்பு, மறைகுறியீட்டியல், சொல் விளையாட்டுக்கள், அளவறி மொழியியல் என்று பல தளங்களில் எழுத்துப் புள்ளியியல் விவரங்கள் பயன்படுத்தப்படுகின்றன.

நோக்கம் மற்றும் செயற்பரப்பு

தமிழ்ச்சொற்களிலுள்ள எழுத்துக்களின் புள்ளியியல் விவரங்களை ஆராய்வது இக்கட்டுரையின் நோக்கமாகும். தமிழின் பல்வேறு எழுத்து வகைப்பாடுகளையும் சொற்களில் எழுத்துக்களின் இடங்களையும் அடிப்படையாகக் கொண்டு புள்ளியியல் விவரங்கள் கணக்கிடப்பட்டுள்ளன.

"சொல்" இருமொழி அகராதியிலுள்ள தமிழ்ச்சொற்களைத் தரவுகளாகக் கொண்டு புள்ளியியல் கணக்கீடுகள் மேற்கொள்ளப்பட்டுள்ளன. பேச்சுத் தமிழிலும், மற்ற முறைசார் பயன்பாடுகளிலும், சொற்கள் எவ்வளவு பரவலாகப் பயன்படுத்தப்படுகின்றன என்பதை இக்கட்டுரை கவனத்தில் கொள்ளவில்லை. பலதரப்பட்ட வகைப்பாடுகளின் கீழுள்ள எழுத்துக்களின் பயன்பாட்டு விகிதத்தை மட்டுமே இக்கட்டுரை விவரிக்கின்றது. தமிழில் பயன்படுத்தப்படுகின்ற கிரந்த எழுத்துக்களும் இவ்வாராய்ச்சியில் இணைக்கப்பட்டிருக்கின்றன.

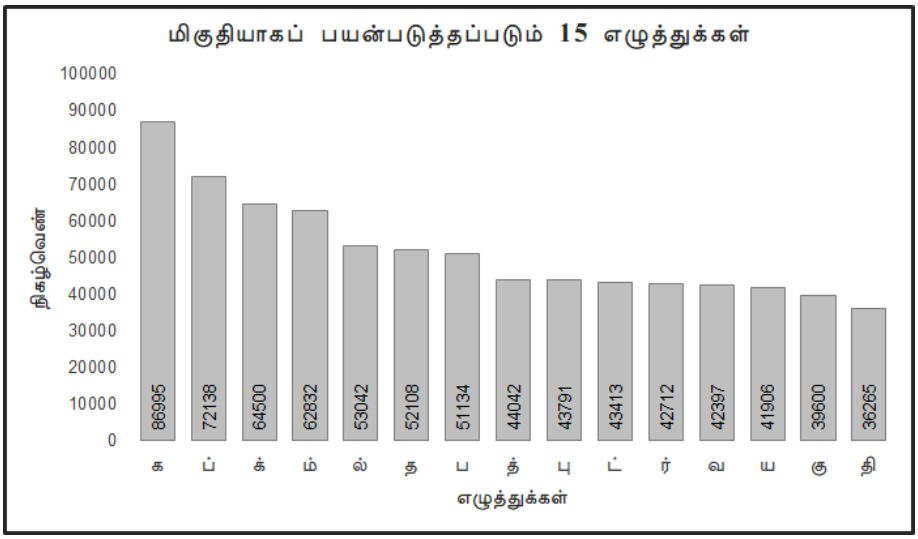
விரித்துரை

சொல் அகராதியிலுள்ள 240665 சொற்களில் பயன்படுத்தப்பட்டுள்ள எழுத்துக்களின் புள்ளியியல் விவரங்களை இக்கட்டுரை விளக்குகிறது. கீழ்காணும் விவரங்கள் பின்வரும் பகுதிகளில் விரித்துரைக்கப்படும்.

(i) மிகுதியாகப் பயன்படுத்தப்படும் எழுத்துக்கள்
(ii) சொற்களின் முதலில் மிகுதியாக வரும் எழுத்துக்கள்
(iii) சொற்களின் ஈற்றில் மிகுதியாக எழுத்துக்கள்
(iv) மூவின எழுத்துக்களின் பயன்பாடு
(v) சொற்களின் முதலில் வரும் மூவின எழுத்துக்களின் பயன்பாடு
(vi) சொற்களின் ஈற்றில் வரும் மூவின எழுத்துக்களின் பயன்பாடு
(vii) குறில், நெடில், மெய் எழுத்துக்களின் பயன்பாடு
(viii) சொற்களின் முதலில் வரும் குறில், நெடில், மெய் எழுத்துக்களின் பயன்பாடு
(ix) சொற்களின் ஈற்றில் வரும் குறில், நெடில், மெய் எழுத்துக்களின் பயன்பாடு
(x) கிரந்த எழுத்துக்களின் பயன்பாடு
(xi) குறைவாகப் பயன்படுத்தப்படும் எழுத்துக்கள்

(i) மிகுதியாகப் பயன்படுத்தப்படும் எழுத்துக்கள்

சொல் அகராதியின் 240665 தனிச் சொற்களில் 2069095 எழுத்துப் பயன்பாடுகள் உள்ளன. இவற்றுள் மிகுதியாகப் பயன்படுத்தப்படும் முதல் பதினைந்து எழுத்துக்களின் விவரங்கள் பின்வரும் கோட்டுருவில் (படம் 1) உள்ளன. கோட்டுருவின் கிடை அச்சு, எழுத்துக்களையும், செங்குத்து அச்சு, அவ்வெழுத்துக்கள் "சொல்" அகராதியிலுள்ள சொற்களில் எத்தனை முறை பயன்படுத்தப்பட்டிருக்கின்றன என்ற நிகழ்வெண்களையும் (frequencies) குறிக்கின்றன.



படம் 1: சொற்களில் மிகுதியாகப் பயன்படுத்தப்படும் 15 எழுத்துக்கள்

இப்படத்திலுள்ள ஒவ்வொரு எழுத்தின் பயன்பாட்டு விழுக்காடும் 4% முதல் 2% வரை உள்ளதைத் தரவுகள் கொண்டு அறியலாம். மிகுதியாகப் பயன்படுத்தப்படும் 15 எழுத்துக்களின் மொத்தப் பயன்பாடு 37.5% ஆக உள்ளது. 89% (268) எழுத்துக்களின் பயன்பாடு, 0.99% முதல் 4.83303 X 10^{-07} வரை என, மிகக் குறைவாகவே உள்ளது.

கீழுள்ள தரவுப்பட்டியலில், நிகழ்வெண்கள் (frequencies) சீரற்ற வரம்பெல்லைகளாக (non-uniform intervals) வகைப்படுத்தப்பட்டுள்ளன; அவற்றுள் நிகழ்வெண்கள் கொண்ட எழுத்துக்களின் எண்ணிக்கையையும் பட்டியலில் காணலாம்.

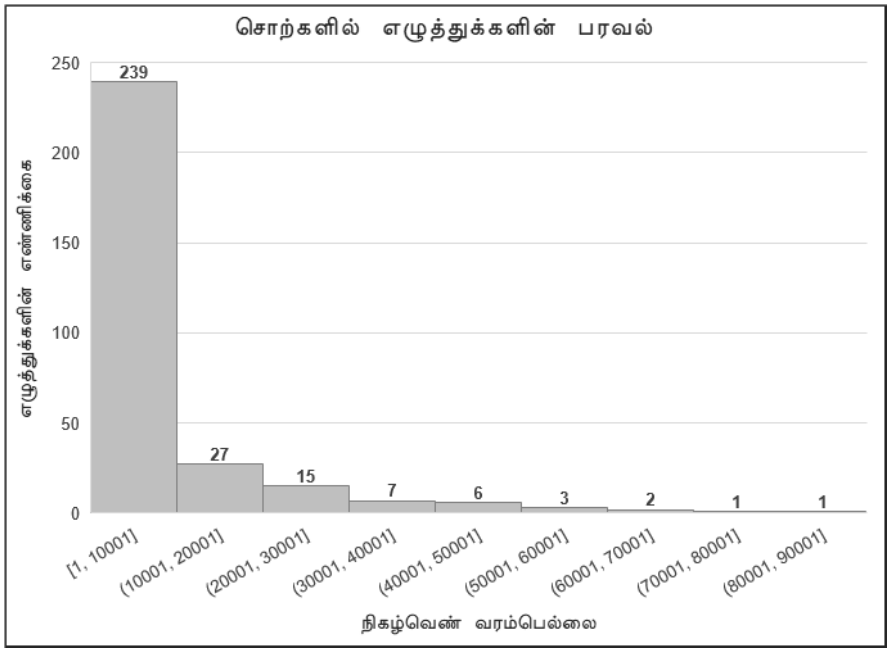
பட்டியல் 1: சொற்களில் பயன்படுத்தப்படும் எழுத்துக்களின் பரவல்

நிகழ்வெண் வரம்பெல்லை	எழுத்துக்களின் எண்ணிக்கை
0 - 10	58
11 to 100	44
101-1000	54
1001- 10000	83
10001-20000	27
20001-30000	15
30001 - 40000	7
40001 - 50000	6
50001 - 60000	3
60001 - 70000	2
70001 - 80000	1
80001 - 90000	1

பட்டியலில் கொடுக்கப்பட்டுள்ள தரவு, காட்சியாகப் புலனாகும்படி படம் 2இல் பரவல் செவ்வகப்படம் கொடுக்கப்பட்டுள்ளது. எழுத்துக்கள், "சொல்" அகராதிச் சொற்களில் எத்தனை முறை இடம்பெறுகின்றன என்பதைக் குறிக்கும் நிகழ்வெண்கள் (frequencies), சீரற்ற வரம்பெல்லைகளாக வகைப்படுத்தப்பட்டு பரவல் படத்தின் கிடை அச்சில் (X axis) கொடுக்கப்பட்டுள்ளன. ஒவ்வொரு வரம்பெல்லைக்குள்ளும் நிகழ்வெண்கள் கொண்ட எழுத்துக்களின் எண்ணிக்கை, செங்குத்து அச்சில் (Y axis) கொடுக்கப்பட்டுள்ளது.

பெரும்பாலான எழுத்துக்களின் நிகழ்வெண்கள் 1001-10000 என்ற வரம்பெல்லைக்குள் உள்ளன. 30000-90000 என்ற வரம்பெல்லையில் 20 எழுத்துக்கள் உள்ளன. இவ்விருபது எழுத்துக்கள் சொல் அகராதியில் உள்ள சொற்களில் 33% இடம்பெற்று இருக்கின்றன.

மிகக் குறைவான எழுத்துக்களே சொற்களில் மிகுதியாகப் பயன்படுத்தப்பட்டிருப்பதை இத்தரவுகளின் உட்கிடையாகக் கொள்ளலாம்.

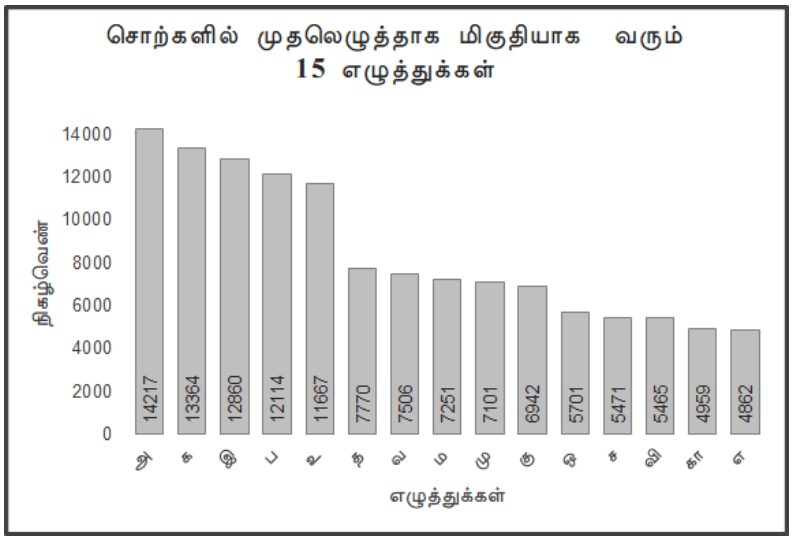


படம் 2: சொல் அகராதிச் சொற்களில் எழுத்துக்களின் பரவல்

(ii) மிகுதியாக முதலெழுத்துக்களாக வரும் எழுத்துக்கள்

தமிழிலக்கண விதிகள் படி அனைத்து எழுத்துக்களும் சொற்களின் முதலெழுத்தாக வராது. சொற்களின் முதலில் வரக்கூடிய எழுத்துக்களில் மிகுதியாக வரும் முதல் 15 எழுத்துக்கள் மற்றும் அவற்றின் நிகழ்வெண்கள் ஆகியவைக் கீழ்காணும் படத்தில் (படம் 3) உள்ளன. இக்கோட்டுரு, கிடை அச்சில் எழுத்துக்களையும், செங்குத்து அச்சில் நிகழ்வெண்களையும் கொண்டுள்ளது.

மார்ச் 21, கட்டுரை எண் 0001



படம் 3: சொற்களின் முதல் எழுத்தாக மிகுதியாக வரும் 15 எழுத்துக்கள்

எழுத்துக்களின் முதலெழுத்து நிகழ்வெண்களைச் சீரற்ற வரம்பெல்லைகளுக்குள் வகைப்படுத்தியிருக்கும் தரவுப்பட்டியலைக் கீழே காண்க.

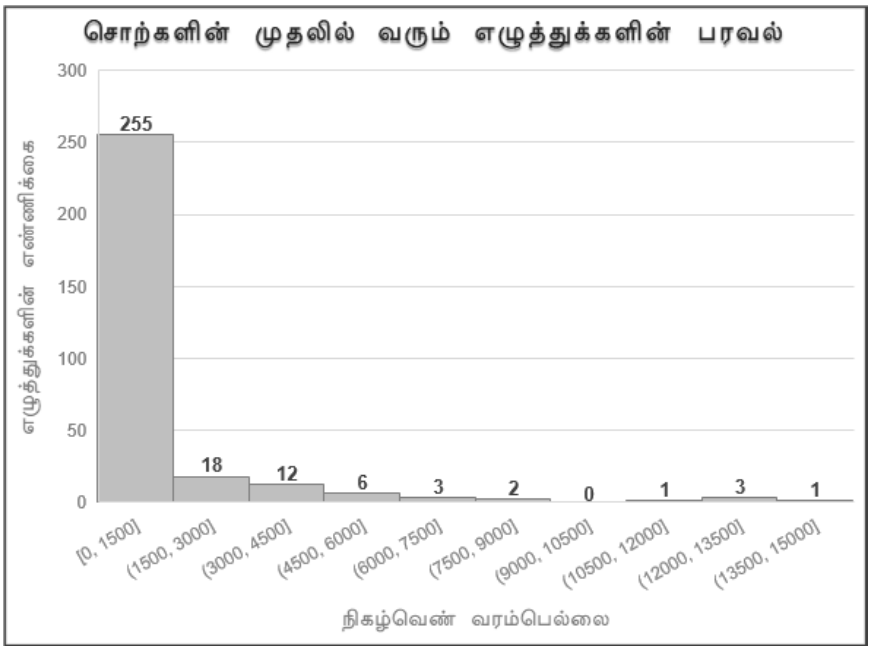
பட்டியல் 2: சொற்களின் முதலெழுத்தாக வரும் எழுத்துக்களின் பரவல்

நிகழ்வெண் வரம்பெல்லை	எழுத்துக்களின் எண்ணிக்கை
0	124
1 to 10	31
11 - 100	47
101- 1000	43
1001 - 2000	5
2001 - 3000	6
3001 - 4000	10
4001 - 5000	5
5001 - 6000	3
6001 - 7000	1
7001 - 8000	4
8001 - 9000	0
9001 - 10000	0
10001 - 11000	0
11001 - 12000	1

12001 - 13000	2
13001 - 14000	1
14001 - 15000	1

தரவுப்பட்டியலின் பரவல் செவ்வகப்படத்தைப் படம் 4இல் காண்க. இப்படத்தின் கிடை அச்சில் சீரற்ற நிகழ்வெண் வரம்பெல்லைகளையும், செங்குத்து அச்சில் குறிப்பிட்ட வரம்பெல்லைகளில் நிகழ்வெண்கள் கொண்ட எழுத்துக்களின் எண்ணிக்கையையும் காணலாம்.

அகராதியிலுள்ள 92% சொற்களின் முதலெழுத்தாக 56 எழுத்துக்கள் இடம்பெறுகின்றன. உயிரெழுத்துக்கள், சொற்களின் முதலெழுத்துக்களாக மட்டுமே இடம்பெற முடியும், மெய்யெழுத்துக்கள் சொற்களின் முதலெழுத்தாக வராது என்ற இலக்கண விதிகளின்படி இவ்வாறு அமைந்திருப்பதாகக் கொள்ளலாம்.

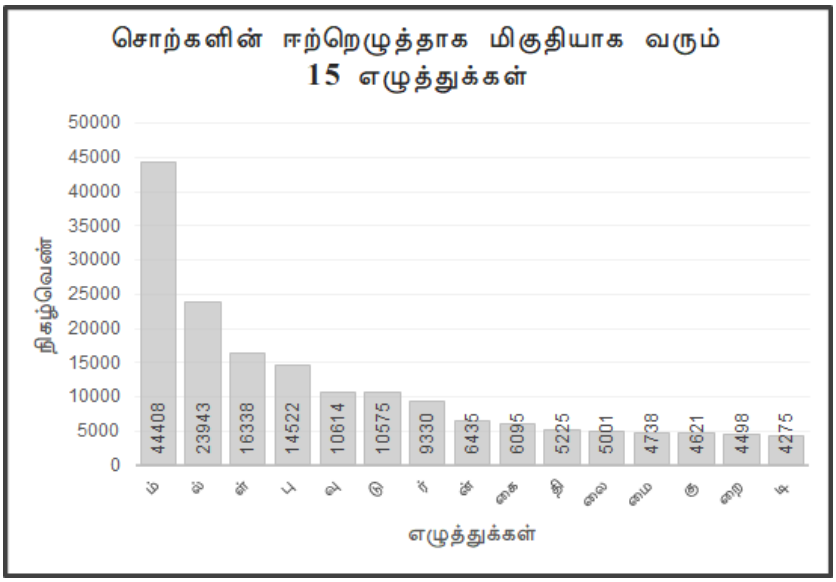


படம் 4: சொற்களின் முதல் எழுத்தாக வரும் எழுத்துக்களின் பரவல்

(iii) மிகுதியாக ஈற்றெழுத்துக்களாக வரும் எழுத்துக்கள்

ஒரு சொல்லின் இறுதி எழுத்தை அச்சொல்லின் ஈற்றெழுத்து என்று அழைப்பர். இப்பகுதியில், சொற்களின் ஈற்றெழுத்துக்களாக வரும் எழுத்துக்களின் புள்ளியியல் விவரங்களைக் காணலாம்.

மிகுதியாகச் சொற்களின் ஈற்றில் வரும் முதல் பதினைந்து எழுத்துக்களையும், அவை ஈற்றெழுத்தாக வரும் சொற்களின் எண்ணிக்கையையும் படம் 5இல் உள்ள கோட்டுருவில் காண்க. இக்கோட்டுருவில், ஒவ்வொரு எழுத்தும் எத்தனை சொற்களின் ஈற்றில் இடம்பெறுகின்றது என்ற எண்ணிக்கையைக் காணலாம்.



படம் 5: சொற்களின் ஈற்றெழுத்தாக மிகுதியாக வரும் 15 எழுத்துக்கள்

எழுத்துக்கள், சொற்களின் ஈற்றில் வரும் நிகழ்வெண்களின் சீரற்ற (non-uniform) வரம்பெல்லை வகைப்பாடு, அவ்வெல்லைகளுள் நிகழ்வெண்கள் கொண்ட எழுத்துக்களின் எண்ணிக்கை ஆகியவை கீழுள்ள பட்டியலில் தரப்பட்டிருக்கின்றன. அதன் பரவல் செவ்வகப்படமும் படம் 6இல் தரப்பட்டுள்ளது.

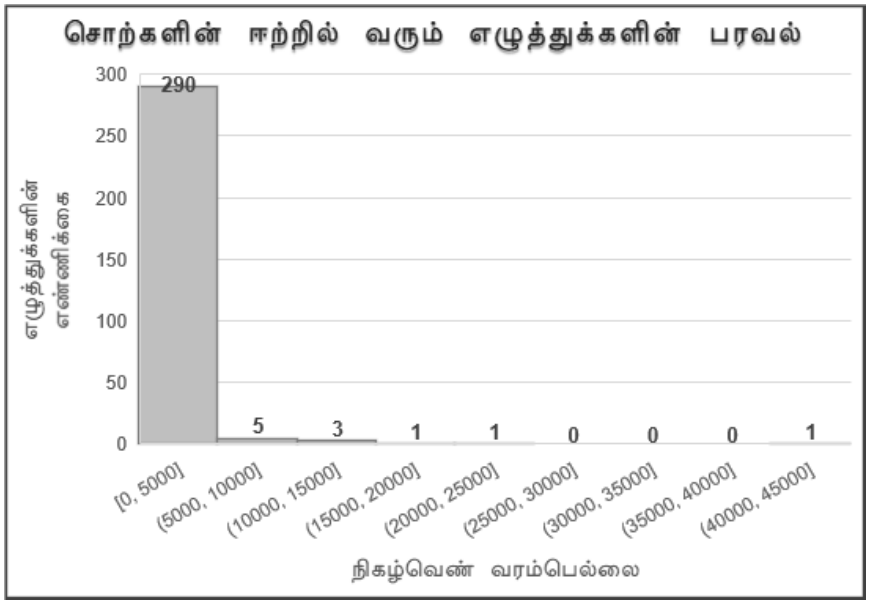
பட்டியல் 3: சொற்களின் ஈற்றெழுத்தாக வரும் எழுத்துக்களின் பரவல்

நிகழ்வெண் வரம்பெல்லை	எழுத்துக்களின் எண்ணிக்கை
0	152
1 to 10	33
11 - 100	46
101- 1000	27
1001 - 2000	14
2001 - 3000	9
3001 - 4000	5
4001 - 5000	4
5001 - 6000	2
6001 - 7000	2
7001 - 8000	0
8001 - 9000	0

9001 - 10000	1
10001 - 20000	4
20000 - 30000	1
30000 - 40000	0
40000 - 50000	1

படம் 6இன் கிடை அச்சில் நிகழ்வெண் வரம்பெல்லைகளும், செங்குத்து அச்சில் அவ்வரம்பெல்லைகளுள் நிகழ்வெண்கள் கொண்ட எழுத்துக்களின் எண்ணிக்கையும் குறிக்கப்பட்டிருக்கின்றன.

152 எழுத்துக்கள் சொற்களின் ஈற்றில் வராமலிருப்பதையும் ஒரு எழுத்தே நாற்பதாயிரம் முறைக்கு மேல் ஈற்றில் வந்திருப்பதையும் காண்க. சொல் அகராதியின் 63% சொற்களுக்கு 20 எழுத்துக்களே ஈற்றெழுத்தாக உள்ளன. சிறிய எண்ணிக்கையிலான எழுத்துக்களே பெரும்பான்மையான சொற்களின் ஈற்றில் வருகின்றன என்பது இத்தரவுகளின் உட்கருத்தாகும்.



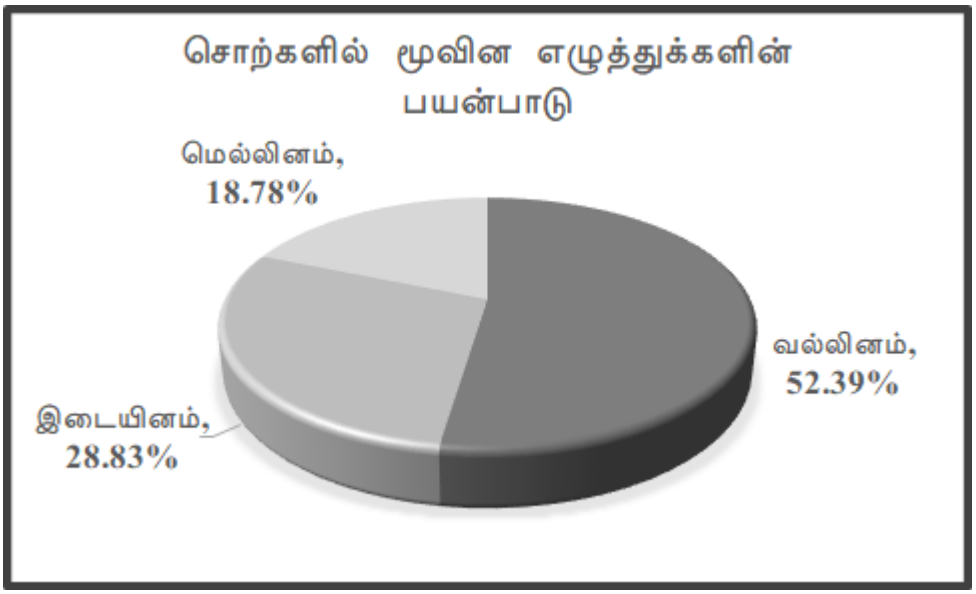
படம் 6: சொற்களின் ஈற்றெழுத்தாக வரும் எழுத்துக்களின் பரவல்

(iv) மூவின எழுத்துக்களின் பயன்பாடு

மெய்யெழுத்துக்களை, அவை குறிக்கும் ஒலிகளின் இயல்புகளுக்கேற்ப வல்லினம், மெல்லினம், இடையினம் என்று பகுக்கலாம். வன்மையான ஒலி கொண்ட எழுத்துக்கள் வல்லினம் என்றும், மென்மையான ஒலி கொண்ட எழுத்துக்கள் மெல்லினம் என்றும்,

வன்மையாகவுமல்லாமல் மென்மையாகவுமல்லாமல் இடைப்பட்ட ஒலி கொண்ட எழுத்துக்கள் இடையினம் என்றும் வகைப்படுத்தப்பட்டுள்ளன.

சொல் அகராதியில் உள்ள சொற்களில் பயன்படுத்தப்பட்டுள்ள எழுத்துக்களில் வல்லின, மெல்லின, இடையின எழுத்துக்களின் விழுக்காடுகளை, கீழ்க்காணும் விளக்கப்படத்தில் காணலாம்.

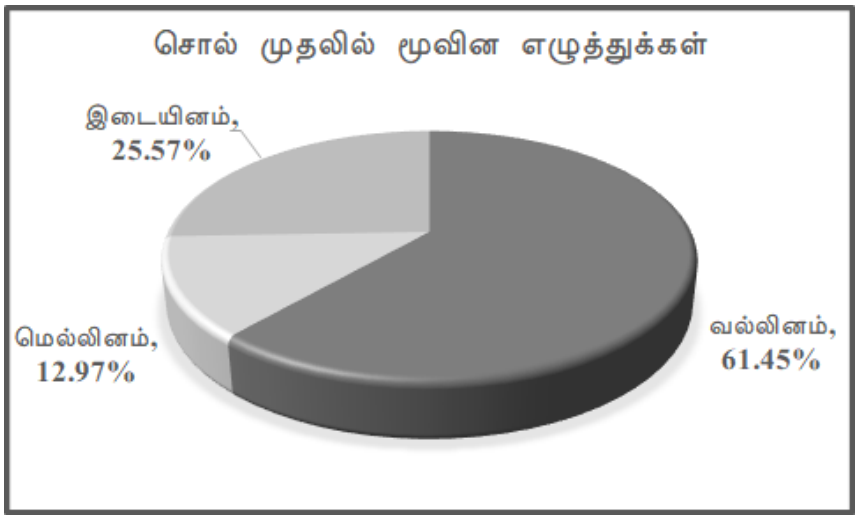


படம் 7: மூவின எழுத்துக்களின் பயன்பாடு

இவ்விளக்கப்படத்தைக் கொண்டு, சொற்களில் பாதிக்கும் மேற்பட்ட விகிதத்தில் வல்லின எழுத்துக்கள் உள்ளதையும், மெல்லின எழுத்துக்கள் 19% மட்டுமே உள்ளதையும் காணலாம்.

(v) முதலெழுத்தாக வரும் மூவின எழுத்துக்களின் பயன்பாடு

சொற்களின் முதலெழுத்துக்களாக வரும் எழுத்துக்களில் வல்லின, மெல்லின, இடையின எழுத்துக்களின் விழுக்காடுகளைக் கீழ்காணும் வட்ட விளக்கப்படத்தில் காண்க.

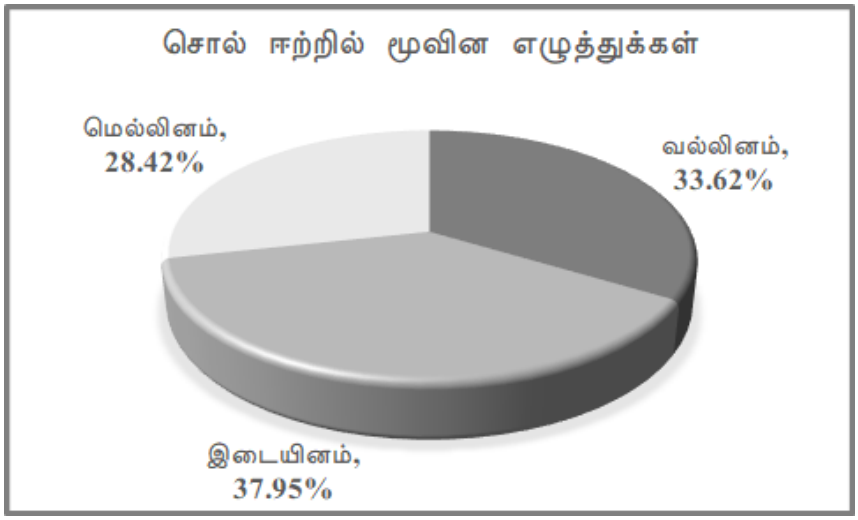


படம் 8: முதலெழுத்தாக வல்லின, மெல்லின, இடையின எழுத்துக்களின் பயன்பாடு

சொற்களின் முதலெழுத்துக்களாகவும் வல்லின எழுத்துக்களே பெரும்பாலும் இடம்பெற்றிருக்கின்றன. மிகக் குறைவான அளவில் (12.9%) மட்டுமே மெல்லின எழுத்துக்கள் சொல் முதலில் இடம்பெறுகின்றன.

(vi) ஈற்றெழுத்தாக வரும் மூவின எழுத்துக்களின் பயன்பாடு

சொற்களின் ஈற்றில் வரும் மூவின எழுத்துக்களின் விழுக்காடுகள் பத்து விழுகாடு வரம்பெல்லைக்குள் இருப்பதைக் கீழ்க்காணும் வட்ட விளக்கப்படத்தில் காண்க.



படம் 9: ஈற்றெழுத்தாக வரும் மூவின எழுத்துக்களின் விழுக்காடுகள்

சொல் முதலில் பெரும்பான்மையாக வந்த வல்லின எழுத்துக்கள், ஈற்றில் குறைவாகவே வருவதைக் காணலாம். சொல் ஈற்றில் இடையின எழுத்துக்கள் மிகுதியாக வருவதைத் தரவுகள் காட்டுகின்றன.

(vii) குறில், நெடில், மெய், ஆய்த எழுத்துக்களின் பயன்பாடு

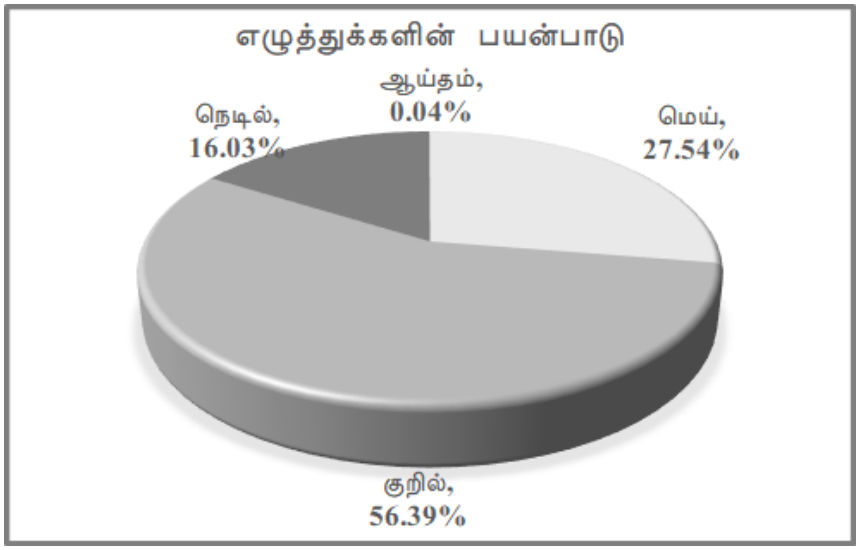
ஒலிக்கப்படும் கால அளவைக் கொண்டு, எழுத்துக்களைக் குறில், நெடில், மெய், என்று பகுப்பர்.

குறில்: குறுகிய ஓசையுடைய எழுத்துக்கள் குறில் அல்லது குற்றெழுத்து என்று அழைக்கப்படும். இவ்வெழுத்துக்கள் ஒரு மாத்திரை நேரம் ஒலிக்கும்.
எ. கா - அ, இ, உ, எ, ஒ

நெடில்: நீண்ட ஓசையுடைய எழுத்துக்கள் நெடில் அல்லது நெட்டெழுத்து என்று அழைக்கப்படும். இவை, இரண்டு மாத்திரை நேரம் ஒலிக்கப்பட வேண்டும்.
எ. கா - ஆ, ஈ, ஊ, ஏ, ஐ, ஓ, ஔ

மெய்: மெய்யெழுத்துக்கள் பதினெட்டும் அரை மாத்திரை அளவு நேரம் ஒலிக்கும்.
ஆய்த எழுத்தும் அரை மாத்திரை நேரம் ஒலிக்கும்.

சொல் அகராதியில் உள்ள சொற்களில், மெய், குறில், நெடில் மற்றும் ஆய்த எழுத்துக்களின் விழுக்காடுகளைக் கீழுள்ள வட்ட விளக்கப்படத்தில் காண்க.

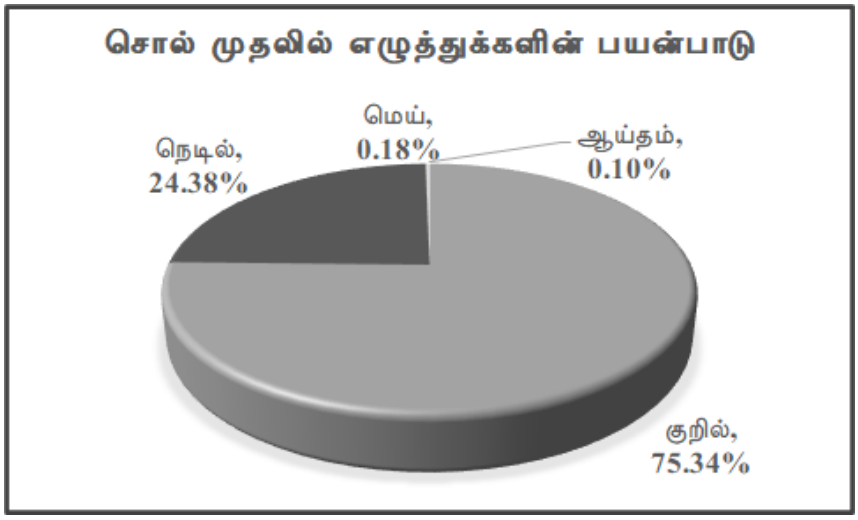


படம் 10: எழுத்துக்களின் பயன்பாடு

குறில் எழுத்துக்கள் பெரும்பான்மையாகவும், ஆய்தம் மிகக் குறைந்த அளவிலும் சொற்களில் இடம்பெறுவதையும், மெய்யெழுத்துக்களின் பயன்பாடு நெடில்களை விட மிகுதியாக இருப்பதையும் காண்க.

(viii) முதலெழுத்தாக வரும் குறில், நெடில், மெய் எழுத்துக்களின் பயன்பாடு

சொற்களின் முதலெழுத்தாக அனைத்து எழுத்துக்களும் வர முடியாத நிலையில், குறில்களே பெரும்பான்மையான சொற்களுக்கு முதலெழுத்தாக வருவதை வட்ட விளக்கப்படத்தில் காணலாம்.

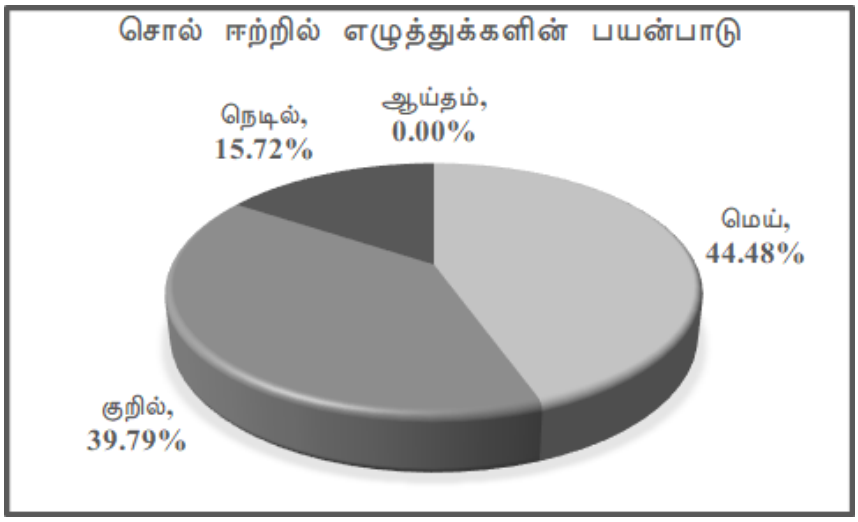


படம் 11: முதலெழுத்தாக வரும் குறில், நெடில், மெய் எழுத்துக்களின் பயன்பாடு

ஏறத்தாழ 25% சொற்களுக்கு நெடில்கள் முதலெழுத்தாக அமைந்துள்ளன. தமிழ் இலக்கண விதிகளின் படி மெய் எழுத்துக்களும், ஆய்தமும் சொல்லின் முதல் எழுத்தாக வராது. "சொல்" அகராதி, இடப் பெயர்கள் போன்ற பிறமொழிப் பெயர்ச் சொற்களின் தமிழ் ஒலிபெயர்ப்புகளையும் உள்ளடக்கியிருக்கின்றது. Fiji, Vladimir என்ற பெயர்களுக்கு ஃபிஜி, வ்லாடிமிர் ஆகிய சொற்கள் ஒலிபெயர்ப்புகள் ஆகும். இவ்வாறான ஒலிபெயர்ப்புகளால், மெய்களும் ஆய்தமும் சொற்களின் முதலெழுத்தாக வருகின்றன. எனவே அவ்வெழுத்துக்களின் முதலெழுத்து விழுக்காடுகள் சுழியமாக இல்லை; ஆயினும் சொல் முதலில் அவற்றின் பயன்பாட்டு விழுக்காடு மிகவும் குறைவாக இருப்பதைக் காண்க.

(ix) ஈற்றெழுத்தாக வரும் குறில், நெடில், மெய் எழுத்துக்களின் பயன்பாடு

சொற்களின் ஈற்றில் மெய்களும் குறில்களும் பெரும்பாலும் வருவதைக் கீழ்க்காணும் வட்ட விளக்கப்படம் காட்டுகிறது.



படம் 12: சொல் ஈற்றில் வரும் குறில், நெடில், மெய் எழுத்துக்களின் பயன்பாடு

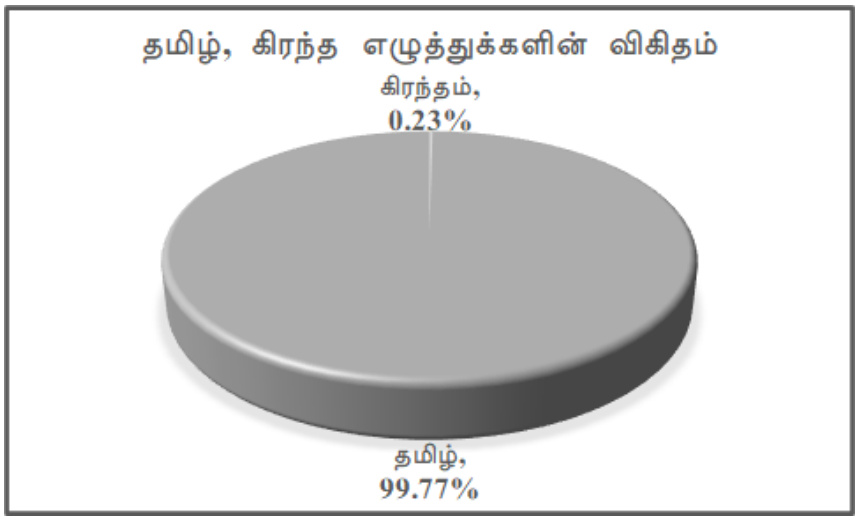
ஒரு தமிழ்ச்சொல்லில் ஆய்த எழுத்து இடம்பெறும் பொழுது, அதன் முன் ஒரு குறிலும், பின் வல்லின உயிர்மெய்யும் வரும்.

குறில்	ஃ	*வல்லின உயிர்மெய்*

பிறமொழிப் பெயர்ச்சொற்களைத் தமிழில் ஒலிபெயர்க்கும் போது, ஆங்கில "f" எழுத்தின் ஒலியை, 'ஃப்' எழுத்துக்களைக் கொண்டு குறிப்பர். இவ்விரண்டு காரணங்களால், ஆய்தம் சொல் ஈற்றில் வராத நிலை உள்ளது. எனவே அதன் பயன்பாட்டு விழுக்காடு சுழியமாக உள்ளது.

(x) கிரந்த எழுத்துக்களின் விகிதம்

வடமொழி ஒலிகளை எழுதவும், வடமொழிச் சொற்களை எழுதவும் கிரந்த எழுத்துக்கள் பயன்படுத்தப்படுகின்றன [32]. அவ்வெழுத்துக்களைப் பயன்படுத்தாது, அவற்றுக்கு இணையாக சில குறிப்பிட்ட தமிழெழுத்துக்களும் பயன்படுத்தப்படுகின்றன. சொல் அகராதியின் சொற்களில் பயன்படுத்தப்பட்டிருக்கும் கிரந்த எழுத்துக்களின் விகிதத்தைக் கீழுள்ள வட்ட விளக்கப்படத்தில் காணலாம்.

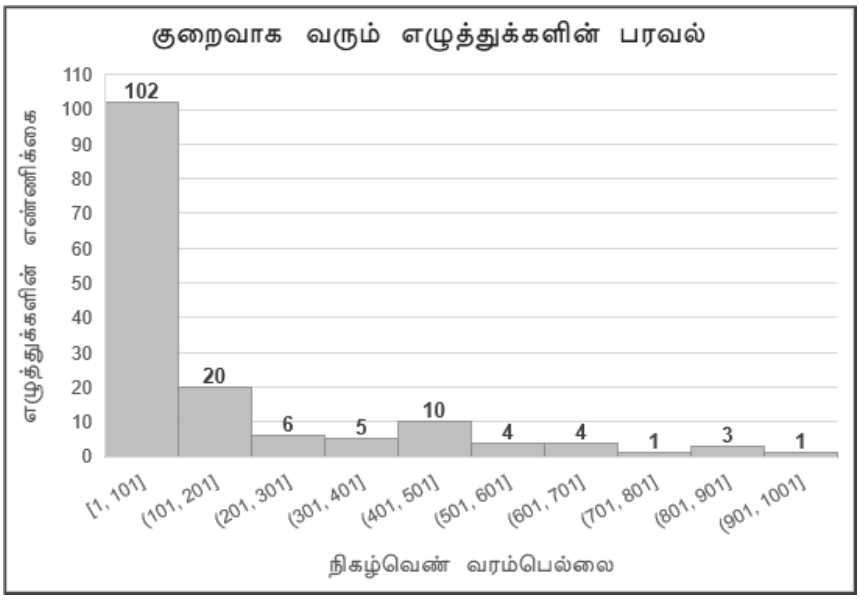


படம் 13: கிரந்த, தமிழ் எழுத்துக்களின் பயன்பாடு

சொல் அகராதிச் சொற்களில் கிரந்தங்களின் பயன்பாடு 0.23% மட்டுமே உள்ளது. ஆகையால் அவை சொல் முதலிலும் ஈற்றிலும் வரும் விகிதங்கள் ஆராயப்படவில்லை.

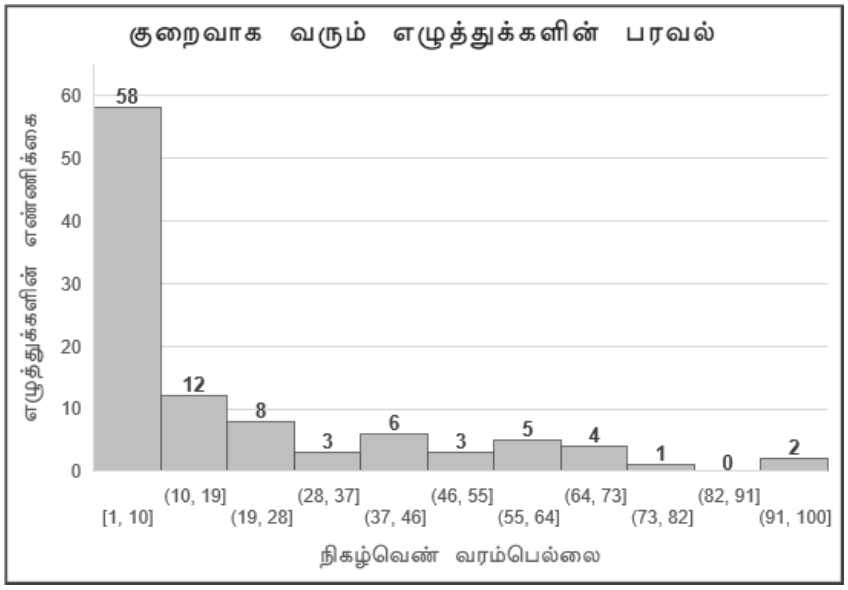
(x) குறைவாக வரும் எழுத்துக்கள்

அகராதியில் உள்ள சொற்களில் குறைவாகப் பயன்படுத்தப்பட்டிருக்கும் எழுத்துக்களில், 1000க்கும் கீழ் நிகழ்வெண்கள் கொண்ட எழுத்துக்களின் பரவலைப் படம் 14இல் காணலாம். சீரற்ற நிகழ்வெண் வரம்பெல்லைகளைப் படத்தின் கிடை அச்சிலும், அவ்வரம்பெல்லைகளுக்குள் நிகழ்வெண்கள் கொண்ட எழுத்துக்களின் எண்ணிக்கையைச் செங்குத்து அச்சிலும் காணலாம்.



படம் 14: சொற்களில் குறைவாக வரும் எழுத்துக்களின் பரவல்

156 எழுத்துக்களின் நிகழ்வெண்கள் 1000க்குக் கீழ் உள்ளது. இவற்றில் 57 எழுத்துக்களின் நிகழ்வெண்கள் 10க்கும் குறைவாக உள்ளது என்பது கண்டறியப்பட்டது. 100க்குக் குறைவாக நிகழ்வெண்கள் கொண்ட எழுத்துக்களின் பரவல் செவ்வகப்படத்தைப் படம் 15இல் காண்க.



படம் 15: சொற்களில் குறைவாக வரும் எழுத்துக்களின் பரவல்

'ங்' என்ற மெய்யெழுத்தும், ங், ஞ் ஆகியவற்றின் உயிர்மெய் வடிவங்களும், 'ஔ' எழுத்தை உயிரெழுத்தாகக் கொண்ட உயிர்மெய்களும் (கௌ, ஙௌ, சௌ...), கிரந்தங்களும், அவற்றின் உயிர்மெய் வடிவங்களும் சொற்களில் குறைவாகப் பயன்படுத்தப்பட்டுள்ளன.

முடிவுரை

சொற்களில் எழுத்துக்களின் பயன்பாட்டு விழுக்காடுகளைச் "சொல்" அகராதியிலுள்ள 240665 சொற்களைத் தரவுகளாகக் கொண்டு இவ்வாராய்ச்சி கண்டறிந்தது. தமிழெழுத்துக்களிலுள்ள பல வகைபாடுகளின் கீழ் வரும் எழுத்துக்கள், சொல்லின் முதலிலும், ஈற்றிலும், பொதுவாகவும் பயன்படுத்தப்பட்டுள்ள விழுக்காடுகள் பட்டியலிடப்பட்டன. குறைவாகப் பயன்படுத்தப்பட்டுள்ள எழுத்துக்களும் ஆராயப்பட்டன.

இவ்வாராய்ச்சி தரும் முடிவுகள், தமிழிலக்கண விதிகளோடு ஒத்திசைவதைக் காண முடிகிறது. சொற்களின் முதலில் வராத எழுத்துக்கள், வல்லினம் மிகுதல், கிரந்தங்களுக்கான தமிழெழுத்துப் பதிலீடுகள், பிறமொழிப் பெயர்ச்சொற்களின் தமிழ் ஒலிபெயர்ப்புகள் போன்றவற்றின் தாக்கத்தை ஆராய்ச்சி முடிவுகளில் காணலாம்.

மூவின எழுத்துக்கள் மற்றும் குறில்-நெடில்-மெய்-ஆய்தம் என்ற இரண்டு எழுத்து வகைப்பாடுகள் மட்டும் இவ்வாராய்ச்சியில் எடுத்துக்கொள்ளப்பட்டிருக்கின்றன. எழுத்து வகைப்பாடுகளான குறுக்கங்கள், அளபெடைகள் போன்றவற்றின் மீதும், எழுத்துக்களின் ஒலிப்பு முறைகள், இடங்கள் மீதும் பின்வரும் காலங்களில் ஆராய்ச்சி மேற்கொள்ளலாம்.

நன்றி

இவ்வுரையில் விவரிக்கப்பட்டிருக்கும் எழுத்துப் புள்ளியியல் தரவுகளைச் "சொல்"அகராதியின் தரவுதளத்திலிருந்து நிரல் கொண்டு தொகுத்தளித்த திரு.இளஞ்செழியன் அவர்களுக்கும், இவ்வுரையைப் படித்திருத்தம் செய்த திருமதி.கன்னியம்மாள் மற்றும் திருமதி.மஞ்சுபாஷினி அவர்களுக்கும் என் நன்றி.

மேற்கோள்

1. பத்ரிராஜூ கிருஷ்ணமூர்த்தி, "Tamil language", https://www.britannica.com/topic/Tamil-language, 2019, அண்மையில் கண்ட நாள் 15/02/2021
2. விக்கிபீடியா, "Letter frequency", https://en.wikipedia.org/wiki/Letter_frequency, 2021, அண்மையில் கண்ட நாள் 15/02/2021
3. லீமான், ஆலிவர், ப்லூம்ஸ்பரி பப்ளிஷிங் வெளியீடு. The Biographical Encyclopedia of Islamic Philosophy. ISBN 9781472569455
4. தாரிக் அல்-தாயெப், "Al-Kindi, Cryptography, Code Breaking and Ciphers", https://muslimheritage.com/al-kindi-cryptography/#note_1, 2003, அண்மையில் கண்ட நாள் 15/02/2021

5. க்ரிஸ் க்ரிஸ்டென்சன், "Cryptanalysis of the Vigenère Cipher: The Friedman Test", https://www.nku.edu/~christensen/1402%20Friedman%20test%202.pdf, 2015, அண்மையில் கண்ட நாள் 15/02/2021
6. பெர்னார்ட் யிகார்ட், "Letter counting: A stem cell for cryptology, quantitative linguistics and statistics", 2013, https://doi.org/10.1075/hl.40.3.01yc
7. க்ளோயீ உல்ட்ரேஜ், "Morse Code (1836)", medium.com/fgd1-the-archive/morse-code-771534ff98e4#:~:text=The%20design%20of%20Morse%20Code,are%20represented%20by%20a%20symbol. , 2017, அண்மையில் கண்ட நாள் 15/02/2021
8. லியோன் பட்டிஸ்டா ஆல்பெர்டி, மொழிபெயர்ப்பு - கிம் வில்லியம்ஸ், "The mathematical work of Leon Battista Alberti - De componendis cifris", ப்ரிக்ஹாசர் பேசல் வெளியீடு, 2010 - பக்கம் 169- 200
9. ஜாக் க்ரீவ், "Quantitative Authorship Attribution: An Evaluation of Techniques", ஆக்ஸ்போர்டு பல்கலைக்கழக பதிப்பகம் வெளியீடு, 2007
10. ப்ரியேந்திர தேஷ்வால், கல்யாண்மோய் டெப், "Ergonomic Design of an Optimal Hindi Keyboard for Convenient Use", IEEE International Conference on Evolutionary Computation, வான்கூவர், கனடா, 2006, பக்கம் 2187-2194
11. குர்சத் அக்பாக் மற்றும் குழு, "Two-Finger Keyboard Layout Problem: An Application On Turkish Language", https://arxiv.org/ftp/arxiv/papers/1605/1605.05122.pdf, 2016, அண்மையில் கண்ட நாள் 15/02/2021
12. ஜியாவ்ஜுன் பி மற்றும் குழு, "Multilingual Touchscreen Keyboard Design and Optimization", Human–Computer Interaction, 2012, பக்கம் 352-382, DOI: 10.1080/07370024.2012.678241
13. ஜான் நோய்ஸ், "he QWERTY keyboard: a review", International Journal of Man-Machine Studies, 1983, பக்கம் 265-281, ISSN 0020-7373, https://doi.org/10.1016/S0020-7373(83)80010-8.
14. ஃபில் ஃப்லிஷ்மான், "Letter Distributions in Word Games", https://boardgamegeek.com/geeklist/182883/letter-distributions-word-games , 2015, அண்மையில் கண்ட நாள் 15/02/2021
15. விக்கிபீடியா, "Scrabble", https://en.wikipedia.org/wiki/Scrabble , 2020, அண்மையில் கண்ட நாள் 15/02/2021
16. தமிழ் இணைய கல்விக் கழகம், "மாத்திரை", http://www.tamilvu.org/courses/degree/c021/c0211/html/c0211333.htm, அண்மையில் கண்ட நாள் 15/02/2021
17. தமிழ் விக்கிபீடியா, "வல்லினம்", https://ta.wikipedia.org/wiki/%E0%AE%B5%E0%AE%B2%E0%AF%8D%E0%AE%B2%E0%AE%BF%E0%AE%A9%E0%AE%AE%E0%AF%8D, 2011, அண்மையில் கண்ட நாள் 15/02/2021
18. விக்கிபீடியா, "Help:IPA/Tamil", https://en.wikipedia.org/wiki/Help:IPA/Tamil, 2017, அண்மையில் கண்ட நாள் 15/02/2021
19. "Interactive IPA Chart", https://www.ipachart.com/
20. தமிழ் விக்கிபீடியா, "மெல்லினம்", https://ta.wikipedia.org/wiki/%E0%AE%AE%E0%AF%86%E0%AE%B2%E0%AF%8D%E0%AE%B2%E0%AE%BF%E0%AE%A9%E0%AE%AE%E0%AF%8D, 2011, அண்மையில் கண்ட நாள் 15/02/2021
21. தமிழ் விக்கிபீடியா, "இடையினம்", https://ta.wikipedia.org/wiki/%E0%AE%87%E0%AE%9F%E0%AF%88%E0%AE%AF%E0%AE%BF%E0%AE%A9%E0%AE%AE%E0%AF%8D, 2011, அண்மையில் கண்ட நாள் 15/02/2021
22. தமிழ் விக்கிபீடியா, "ஆய்த எழுத்து", https://ta.wikipedia.org/wiki/%E0%AE%86%E0%AE%AF%E0%AF%8D%E0%AE%A4_%E0%AE%8E%E0%AE%B4%E0%AF%81%E0%AE%A4%E0%AF%8D%E0%AE%A4%E0%AF%81, 2011, அண்மையில் கண்ட நாள் 15/02/2021
23. தமிழ் இணைய கல்விக் கழகம், "அசைக்கு உறுப்பாகும் எழுத்துகளின் பெயர்கள்", http://www.tamilvu.org/courses/diploma/c021/c0214/html/c0214311.htm
24. தமிழ் இணைய கல்விக் கழகம், "உயிரளபெடை", http://www.tamilvu.org/courses/diploma/c021/c0211/html/c0211403.htm
25. தமிழ் இணைய கல்விக் கழகம், "ஒற்றளபெடை", http://www.tamilvu.org/courses/degree/c021/c0211/html/c0211404.htm
26. தமிழ் விக்கிபீடியா, "குற்றியலுகரம்", https://ta.wikipedia.org/w/index.php?title=%E0%AE%95%E0%AF%81%E0%AE%B1%E0%AF%8D%E0%AE%B1%E0%AE%BF%E0%AE%AF%E0%AE%B2%E0%AF%81%E0%AE%95%E0%AE%B0%E0%AE%AE%E0%AF%8D, 2013, அண்மையில் கண்ட நாள் 15/02/2021
27. தமிழ் இணைய கல்விக் கழகம், "குற்றியலுகரம்", http://www.tamilvu.org/courses/degree/c021/c0211/html/c0211405.htm
28. தமிழ் விக்கிபீடியா, "குற்றியலிகரம்", https://ta.wikipedia.org/wiki/%E0%AE%95%E0%AF%81%E0%AE%B1%E0%AF%8D%E0%AE%B1%E0%AE%BF%E0%AE%AF%E0%AE%B2%E0%AE%BF%E0%AE%95%E0%AE%B0%E0%AE%AE%E0%AF%8D, 2005, அண்மையில் கண்ட நாள் 15/02/2021
29. தமிழ் இணைய கல்விக் கழகம், "குற்றியலிகரம்", http://www.tamilvu.org/courses/degree/c021/c0211/html/c0211406.htm
30. தமிழ் இணைய கல்விக் கழகம், "குறுக்கங்கள்", http://www.tamilvu.org/courses/degree/c021/c0211/html/c0211407.htm
31. தமிழ் விக்கிபீடியா, "ஐகாரக்குறுக்கம்", https://ta.wikipedia.org/wiki/%E0%AE%90%E0%AE%95%E0%AE%BE%E0%AE%B0%E0%AE%95%E0%AF%8D_%E0%AE%95%E0%AF%81%E0%AE%B1%E0%AF%81%E0%AE%95%E0%AF%8D%E0%AE%95%E0%AE%AE%E0%AF%8D, 2005, அண்மையில் கண்ட நாள் 15/02/2021

32. தமிழ் இணைய கல்விக் கழகம், "கிரந்தம்", http://www.tamilvu.org/tdb/titles_cont/inscription/html/treatise.htm
33. கார்க்கி ஆராய்ச்சி நிறுவனம், "சொல் அகராதி", karky.in/chol, அண்மையில் கண்ட நாள் 15/02/2021
34. How to pronounce, பலுக்கல் அகராதி, https://www.howtopronounce.com/, அண்மையில் கண்ட நாள் 15/02/2021

ஆசிரியர் குறிப்பு

சூர்யா முருகேசன், கணிப்பொறி அறிவியலில் முதுகலைப் பொறியியல் பட்டம் பெற்று, 4.5 ஆண்டுகள் பொறியியல் மாணவர்களைப் பயிற்றுவிக்கும் உதவிப் பேராசிரியராகப் பணிபுரிந்தவர்.
தற்போது, கார்க்கி ஆராய்ச்சி நிறுவனத்தில் "பயில்" வகுப்புகளின் ஒருங்கிணைப்பாளராகவும், கணினி மொழியியல் ஆராய்ச்சியாளராகவும் பணிபுரிகிறார்.