உரையிலிருந்து பேச்சு உருவாக்கம்

சு.சீனிவாசன், சி.தத்தா, பி.சீனிவாசன்
கணிப்பொறிக் கோட்டம், இந்திராகாந்தி அணுவாராய்ச்சி மையம்,
கல்பாக்கம்-603102, காஞ்சிபுரம் மாவட்டம், தமிழ்நாடு

முன்னுரை

எண்ம அடிப்படையில் பேச்சை அலசுதல் (digital speech processing) இன்று முக்கியத்துவம் பெறுகிறது. இவற்றுள் உரையிலிருந்து பேச்சு உருவாக்கம் செய்வது (text-to-speech) அதிகச் சிக்கலற்ற பணியாகத் தோன்றுகிறது. உரைநடையிலுள்ள அனைத்துச் சொற்களையும் செயற்கை முறையில் ஒலிக்கச் செய்வதே பேச்சு உருவாக்கத்தின் (speech synthesis) நோக்கமாகும். இப்பணிக்கு சொற்களைவிட குறுகிய அலகுகளான அசை (syllable), குற்றசை (demisyllable), ஒலியன் (phoneme) ஆகியவை பயன்படுத்தப்படுகின்றன.

மென்பொருள் உருவாக்கம்

இந்திய மொழிகள் அனைத்திற்கும் ஒரு சிறப்பு உண்டு. ஐரோப்பிய மொழிகளைக் காட்டிலும் இந்திய மொழிகளின் வரிவடிவம் (script) சிக்கலானது என்றாலும் ஒலிக்கும்போது எழுதுவதைப் போலவே உச்சரிக்கிறோம். ஆங்கிலத்தின் வரிவடிவம் எளிது என்றாலும் அவற்றிலுள்ள சொற்களை உச்சரிப்பதற்கு நிறைய பட்டறிவு தேவைப்படுகிறது. சொற்களை இடத்திற்கு ஏற்ப- put என்பதை புட் என்றும் cut என்பதை கட் என்றும் வேறுபடுத்தி ஒலிக்கிறோம். ஆனால் இத்தகைய சிக்கல் இந்திய மொழிகளில் கிடையாது. தமிழுக்கு, குறிப்பாக இரு சிறப்புகள் உண்டு. முதல் சிறப்பு தமிழ் நெடுங்கணக்கில் எழுத்துக்கள் குறைவு என்பதும் கூட்டெழுத்துக்கள் இல்லை என்பதாகும். இரண்டாவது சிறப்பு, ஒரே வல்லெழுத்து பல இடநிலைகளில் மாறுபட்ட ஒலிகளை உருவாக்குவதற்கு தமிழில் நியதிகள் வகுக்கப்பட்டிருப்பதாகும்.

தமிழிலுள்ள முதல் எழுத்துக்களுக்கான ஒலியன்கள் முப்பதுதாம் என்றாலும் அவற்றைக்கொண்டு உயிர்மெய் ஒலிகளை உருவாக்க கூடுதல் முயற்சி தேவை. இதற்கு மாற்றாக உயிர்மெய் ஒலிகளைத் தனித்தனியே பதிவுசெய்து பயன்படுத்துவது எளிது. இதன் அடிப்படையில் தமிழோசை (VOICE TAMIL) எனும் மென்பொருள் உருவாக்கப்பட்டுள்ளது. உரையிலிருந்து பேச்சு உருவாக்கம் எவ்வாறு நிகழ்கிறது என்பதைப் படம்-1 விளக்குகிறது.

இந்த மென்பொருளுக்கு உள்ளீடாக தமிழ் உரைப்பகுதி அமைகிறது. மூன்று ஆக்கக் கூறுகளை இம்மென்பொருள் கொண்டுள்ளது. முதல் நிலையில், சொற்களிலுள்ள எழுத்துக்கள் உயிர், மெய், உயிர்மெய் எனப் பிரித்தறியப்படுகிறது. இரண்டாம் நிலையில், இவற்றுக்கான ஒலித் துண்டுகள் (speech segments) தெரிவு செய்யப்படுகின்றன. மூன்றாம் நிலையில், இந்த ஒலித்துண்டுகள் தொகுக்கப்பட்டு பேச்சு உருவாக்கப்படுகிறது. இப்பணிக்கென அனைத்து ஒலிகளுக்குமான நூலகக் கோப்பு ஒன்றும் எழுத்துக் கூட்டலுக்கென சில ஒலிப்பு நியதிகளும் புகுத்தப்பட்டுள்ளன. ஏறக்குறைய 700 கிலோ பைட்டு அளவில் ஒலித் தகவல்கள் வன்தட்டில் (hard disk) கோப்புகளாகப் பதிவு செய்யப்பட்டுள்ளன. அதாவது மொத்தத்தில் ஒரு நிமிட ஒலிப்பதிவு நேரமே இந்த மென்பொருள் செயல்பாட்டிற்கு ஆதாரமாக விளங்குகிறது.

பேச்சு உருவாக்கும் அமைப்பு

இம்மென்பொருள் தனிக் கணிப்பொறியில் (Personal Computer) செயல்படும் வண்ணம் வடிவமைக்கப்பட்டுள்ளது. கணிப்பொறி ஆணைத்தொடர்கள் அனைத்தும் C மொழியில் எழுதப்பட்டுள்ளன. ஏறக்குறைய 350 ஒலித்துண்டுகள் எண்ம வடிவில் கோப்புகளில் சேமிக்கப்பட்டுள்ளன. ஒப்புமை சமிக்கையாகவுள்ள (analog signal) பேச்சொலியை எண்மக் கூறுகளாக மாற்ற 11025 எண்ணிக்கை கொண்ட கூறுவீதத்தை (sampling rate) பயன்படுத்தினோம். இக்கூறுகள் ஒவ்வொன்றும் ஒரு பைட்டு நீளமுடையன. பதிவுசெய்த ஒலி அலைகளை, நேர்த்திசெய்ய வேவ்ஸ் ஸ்டூடியோ என்னும் பயன்பாட்டு மென்பொருளைக் கையாண்டோம்.

குறில் எழுத்துக்களுக்கான சராசரி ஒலிப்பளவு 160 மில்லி செகண்டாகவும், நெடில் எழுத்துக்களுக்கான சராசரி ஒலிப்பளவு 240 மில்லி செகண்டாகவும் இருந்தன. வல்லொற்றுக்களை (plosives) பதிவுசெய்வதில் கூடுதல் கவனம் செலுத்தப்பட்டது. அலைநேர்த்தி (wave editing) செய்யப்படும்போது இவ்வொலிப் பதிவிலுள்ள அமைதி இடைவெளி (silence period) களையப்படாமல் காக்கப்படவேண்டியது அவசியம் என அறியப்பட்டது.

எழுத்துக் கூட்டலும் ஒலித்தொகுப்பும்

எழுத்துக்களின் அணிவகுப்பைச் சொல் என்கிறோம். எழுத்துக்களைக் கூட்டிச் சொல்லை உச்சரிக்க முயலுகிறோம். இதனைக் கணிப்பொறிக்கொண்டு செயல்படுத்த கூடுதல் முயற்சி தேவை. எழுத்துக்களை ஒவ்வொன்றாக உச்சரித்து ஒலித்துண்டுகளை உருவாக்குவதைக் காட்டிலும் அவற்றை அசைச்சொற்களிலிருந்து அகழ்ந்து எடுப்பதே எழுத்துக்கூட்டி வாசிக்க உதவிடும் என்பதைச் சில சோதனைகள் வெளிப்படுத்தின.

எழுத்துக்கூட்டல் எவ்வாறு ஒலித்தொகுப்பாக மாறுகிறது என்பதைப் படம்-2 விளக்குகிறது. இதிலிலுள்ள முதல் அலைப்படம் நேரடியாக ஒலிப்பதிவு செய்யப்பட்ட த,மி,ழ் என்ற மூன்று எழுத்துக்களுக்குரியவை. இரண்டாவது அலைப்படம் அத, அமி, அழ் எனும் மூன்று அசைகளுக்குரியன. மூன்றாவது அலைப்படம் இரண்டாவது அலைப்படத்திலிருந்து பெறப்பட்டது. இதில் மூன்று அசைச் சொற்களிலிருந்தும் தொடக்க ஒலியான அகரம் நீக்கப்பட்டு பின் அவை 'தமிழ்' எனும் ஒலியை உருவாக்குமாறு தொகுக்கப்பட்டுள்ளன. இவ் அலைப்படம் செயற்கை நிலையில் ஒலித்தொகுப்பு (synthesis) செய்திருப்பதைக் காட்டுகிறது. நேரடியாக ஒலிப்பதிவு செய்த 'தமிழ்' எனும் பேச்சொலி நான்காவது அலைப்படத்தில் காட்டப்பட்டுள்ளது.

மூன்றாவது அலைப்படம் ஏறக்குறைய நான்காவது அலைப்படத்தை வடிவில் ஒத்திருக்கிறது. ஒலித்தொகுப்பில் 'மி' எனும் ஈற்றயல் எழுத்துக்குரிய (penultimate demisyllable) ஒலி சற்று வலிமை குன்றி இருக்கிறது. வனம், நவம் என்ற சொற்களை ஒலிக்கும்போது ஈற்றெழுத்துக்கு முன்னுள்ள எழுத்தொலி வலிமை பெறுவதை ஐந்தாவது, ஆறாவது அலைப்படத்தில் காணலாம். இந்த வேறுபாடுகளை அறிந்து ஒலியின் வலிமையைக் கூட்டவும் குறைக்கவும் முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. உரையில் இடம்பெறும் எண்களை இயல்பாக ஒலிக்கும் பொருட்டு பூச்சியம் முதல் ஒன்பது வரை உள்ள எண்கள் நேரடியாக ஒலிப்பதிவு செய்யப்பட்டன.

முடிவுரை

ரோமன் எழுத்துமுறையில் உள்ளீடு செய்யும் எல்லாத் தமிழ் வாக்கியங்களையும் இந்த மென்பொருள் வாசிக்கிறது. சொற்களை அசைகளாகவோ, எழுத்துக்களாகவோ பிரித்து வாசிக்கவும் இதில் வசதி ஏற்படுத்தப்பட்டிருக்கிறது. ஒலிப்பதிவு செய்யும் நபரின் குரலிலேயே சொற்கள் உச்சரிக்கப்படுகின்றன. சொற்களை உணர்வுபூர்வமாக ஒலிப்பதற்கு மேலும் பல உத்திகளை இம் மென்பொருளில் சேர்க்க வேண்டியுள்ளது.

fig 1

fig 2

fig 192 empty

தமிழில் சொல் திருத்தம்

முனைவர் வெ. கிருஷ்ணமூர்த்தி
(முன்னாள் பேராசிரியர், அண்ணா பல்கலைக் கழகம்)
இன்ஃபோரீட், 30(4 c) இரண்டாவது பிரதான சாலை
காந்தி நகர், அடையாறு, சென்னை 600 200

சுருக்கம்
தமிழில் ஒரு சொல் திருத்தி உருவாக்கும்போது எதிர்கொள்ளும் சிக்கல்கள் பற்றி இக்கட்டுரை சுருக்கமாக ஆராய்கிறது. ஒரு சொல்திருத்தியின் நோக்கங்கள் பற்றி முதலில் ஆராயப்படுகிறது. அடுத்து, தமிழில் ஒரு சொல்லை சரியானது என்று சொல்வதற்குச் செய்ய வேண்டிய செயல்பாடுகளும், அதில் வரும் சிக்கல்களும் எடுத்துக்காட்டப்பட்டுள்ளன. சொல்திருத்தி விரைவாகச் செயல்படத் தேவைப்படும் முடிவுகள் குறித்தும் ஆராயப்படுகின்றது.

1. சொல்திருத்தியின் நோக்கம்

சாவிப்பலகை கொண்டு உள்ளீடு செய்த உரையில் எழுத்துப் பிழைகள் இருந்தால் அவற்றை விரைவாகக் களைய உதவுவதுதான் சொல் திருத்தியின் முக்கிய நோக்கம். உள்ளீடு செய்பவரின் நேரத்தை மிச்சப்படுத்துவதே இங்கு முதன்மையான நோக்கம். உரை முற்றிலும் எழுத்துப் பிழை எதுவும் இன்றி வருவது இரண்டாம் நோக்கம். இந்த இரு நோக்கங்களும் பெரிதும் நிறைவேற்றப்பட்டால்தான் ஒரு சொல்திருத்தி உருவாக்கப்பட்டதாக எண்ணலாம். பல சரியான சொற்களை தவறு என்று கூறினாலும், பல தவறான சொற்களை சரியென்று கூறினாலும், இந்த இரு நோக்கங்களும் நிறைவேறுவதில்லை. அதற்கு மாறாக, நேரம்தான் வீணாகும். கணிப்பொறியைப் பயன்படுத்துவது நேரத்தை மிச்சப்படுத்தவே அன்றி நேரத்தை வீணாக்க அல்ல.

ஆங்கிலத்தில் ஒரு சொல்லில் உள்ள பிழைகளைக் கண்டறிவது சற்று எளிதான செயல். ஏனென்றால், ஒரு சொல்லுடன் ing, ed, ly போன்ற சில ஒட்டுக்கள் மட்டுமே சேரும். இவற்றைப் பிரித்துவிட்டு, மீதி உள்ள சொல்லை பட்டியலில் தேடிப் பார்த்தால் போதும். தமிழில் இவ்வாறு இல்லை. தமிழ் சொல் திருத்தியில் அதிக வேலை செய்ய வேண்டியிருக்கிறது. அதனால் நேரம் சற்று அதிகமாக செலவாகும். ஆனாலும் ஒரு பக்கத்தைச் சரிபார்க்க எவரும் இரண்டு, மூன்று நிமிடங்களுக்கு மேல் செலவிட விரும்ப மாட்டார்கள். அதனால் ஒரு சொல்திருத்தி ஒரு பக்கத்தை இரண்டு அல்லது மூன்று நிமிடங்களில் சரிபார்க்க வேண்டும் என்பது ஒரு முக்கிய நோக்கமாக, அளவுகோலாக இருக்க வேண்டும்.

ஒரு பக்கத்தில் எழுதப்பட்ட சொற்கள் எல்லாம் சரியானவையாக இருந்தால், அவற்றை சரியானவை எனக் கூறும் திறன் வேண்டும். பல சரியான சொற்களை தவறு என்று கூறும்போது அதனால் மூன்று விளைவுகள் ஏற்படுகின்றன. முதலில் நேரம் வீணாகிறது. அடுத்து, அதைப் பயன்படுத்துபவருக்கு எரிச்சலூட்டுகிறது. இவற்றை விட முக்கியமாக, பயன்படுத்துபவர் அவர் எழுதியது தவறோ என்ற குழப்பத்தை ஏற்படுத்தும். இதனாலேயே அவர் சரியானதைத் தவறாக்கிவிடும் வாய்ப்பும் உள்ளது.

தமிழில் சொல்திருத்தி உருவாக்குவது கடினம் என்றாலும், இத்தகைய பிழைகள் 5 அல்லது 10 விழுக்காட்டிற்கு மேல் இருந்தால் அதை எவரும் பயன்படுத்தப்போவதில்லை.

அதே போல், தவறான சொல்லை சரி என்று கூறுவதும் நேரத்தை வீணாக்கும். இதுவும் 5 அல்லது 10 விழுப்பாடுகளுக்கு மேலே இருப்பது நல்லதல்ல. இவற்றைச் சோதிக்கும்போது, தரமாக எழுதப்பட்ட கட்டுரைகளைப் பயன்படுத்த வேண்டும். பிழைகளை இரு வகைகளாகப் பிரிக்கலாம். ஒன்று தட்டச்சு செய்மேபோது தானாக வரும் பிழைகள். இவற்றைக் கண்டறிவது சற்று எளிது. அடுத்தது, சொல்திருத்தி சரியாக வேலை செய்கின்றதா எனப் பார்க்க நாமே கொடுக்கும் தவறான சொற்கள். இவற்றைமே கண்டறியும் திறம் இருந்தால்தான் அதனை உண்மையான சொல்திருத்தி எனக் கொள்ளலாம். இத்தகைய பிழைகளைக் கண்டறிவது சற்றுக் கடினமான காரியம். எடுத்துக்காட்டாக, கண்தானே, கண்ணேதான் என்ற இரண்டுமே சரியானவை. ஆனால், தானேகண் என்பது தவறு. கண், தான் மற்றும் ஏ என்ற பகுதிகள் இருப்பதனாலேயே அந்தச் சொல்லை சரியென்று கூறக் கூடாது.

2. சிக்கல்கள்

தமிழில் ஒரு வேர்ச் சொல்லில் இருந்து ஆயிரக் கணக்கில் சொற்களை உருவாக்கலாம். ஒரு சொல்லில் ஏழெட்டுப் பகுதிகள்கூட இருக்கலாம். எடுத்துக்காட்டாக, செய்துகொண்டிருந்தபோதுதானேயப்பா என்பது ஒரே சொல். அது, செய் என்னும் வினைச்சொல்லில் இருந்து உருவாக்கப்பட்டுள்ளது. இதில் உள்ள பகுதிகளைச் சரியானபடி பிரித்து, அவை சரியாகச் சேர்ந்துள்ளனவா என்று பார்த்தால்தான் ஓர் உண்மையான சொல்திருத்தியை உருவாக்க முடியும்.

இப்படி ஒரு சொல்லைப் பிரிக்கும்போது தமிழில் உள்ள புணர்ச்சி விதிகள் அனைத்தையும் கையாண்டு பிரிக்க வேண்டும். ஒரு இடத்தில் பல விதங்களில் பிரிக்க முடியலாம். அவற்றில் பல, சரியான பிரிப்புகள் அல்ல என்பது முழுதும் பிரித்த பிறகே தெரியும். இரு சொற்கள் சேரும்போது, கெடுதல், திரிதல் மற்றும் தோன்றல் என்று மூன்று வகைகளில் எழுத்துக்கள் மாற்றம் பெறுகின்றன. இலக்கண நூல்கள் இவற்றைப் பற்றிக கூறினாலும் சில இடங்களில் 'இது போன்று இன்னும் பல இடங்களிலும் வரும்' என்று கூறுவதால், அந்த இடங்களில் துல்லியமாகச் செயல்படுவது கடினமான காரியமாக உள்ளது.

எடுத்துக்காட்டாக, வந்து + இரு என்பதில் உகரம் கெட்டு, வந்திரு என ஆகிறது. கண் + இல் என்பதில் ண் தோன்றி, கண்ணில் என ஆகிறது. பல் + கள் என்பது பற்கள் எனத் திரிகிறது. ஒரு சொல்லில் உள்ள பகுதிகளைச் சேர்க்கும்போது அந்தப் பகுதிகள் எவை என்று தெரிவதால், தேவையான இலக்கணத்தைப் பயன்படுத்தி அவற்றை இணைப்பது சற்று எளிதான செயல். ஆனால், ஒரு சொல்லை, அதன் பகுதிகளாகப் பிரிக்கும்போது, அவற்றின் பகுதிகளைப் பற்றிய செய்திகள் எதுவும் தெரியாது. அதனால் பிரிக்கும் இடத்தில் ஒரு எழுத்தினைச் சேர்க்க வேண்டுமா, அல்லது ஒரு எழுத்தினை நீக்க வேண்டுமா, அல்லது ஒரு எழுத்தினை மாற்ற வேண்டுமா என்பது தெரியாது. எல்லா வகைகளிலும் செய்துபார்த்தால்தான் சரியான விடை கிடைக்கும்.

3. தேவையான உத்திகள்

தமிழில் ஒரு சொல்லைச் சரி பார்க்கவே பல கணிப்புகளைச் செய்ய வேண்டியிருக்கும். இதற்கு நேரம் அதிகம் செலவாகும். நேரத்தைக் குறைக்க பல உத்திகளைக் கையாள வேண்டும். அவற்றில் சிலவற்றை இங்கு காண்போம்.

தேவையான தகவல்களின் வடிவமைப்பைத் தீர்மானிப்பது ஒரு முக்கிய செயல். இதில், தகவல்களை எந்தக் குறியீட்டில் வைக்கிறோம், எந்தத் தகவல் அமைப்புகளில் வைக்கிறோம் என்பவை அடங்கும்.

தமிழ் எழுத்துக்களை எந்தக் குறியீட்டில் வைத்து கணிப்புகளைச் செய்கிறோம் என்பது ஒரு மிக முக்கியமான முடிவு. இது சொல்திருத்தியின் நேரத்தை அதிக அளவில் பாதிக்கக்கூடியது. அடிப்படைச் செயல்பாடுகள் பல்லாயிரக்கணக்கில் செய்யப்படும்போது, ஒவ்வொரு முறையும் செய்யும் சிறு வீணடிப்பும் பெரு வெள்ளமாக மாறிவிடும். எழுத்துக்களை மெய் + உயிர் என்று வைக்கலாம். உயிர், மெய், உயிர்மெய் என வைக்கலாம். இஸ்கி குறியீட்டில் வைக்கலாம். டேம் அல்லது டேப் குறியீடுகளில் கையாளலாம். கணிப்பில் பயன்படுத்தும் குறியீடு, சொற்களை நினைவகங்களில் தேக்கி வைக்கப்பயன்படுத்தும் குறியீடாக இருக்க வேண்டும் என்பது அவசியமில்லை. இவை இரண்டும் வெவ்வேறாக இருக்கலாம்.

உள்ளே பயன்படுத்தப்படும் குறியீடு, எழுத்துக்களை மெய் + உயிர் என்று பிரிப்பதையும், அவற்றை மீண்டும் சேர்ப்பதையும் மிக விரைவாகச் செய்யும் வகையில் இருக்க வேண்டும். மேலெழுந்தவாரியாகப் பார்க்கும்போது, குறியீட்டின் முக்கியத்துவம் அவ்வளவாகத் தெரியாது. ஆழ்ந்து பார்க்கும்போதுதான் அது புலப்படும். இனி வரும் ஆண்டுகளில் இயல்பு மொழிச் செயலாக்கம் பெருமளவில் பயன்படும். கணிப்பொறியில் கையாளப்படும் மொழிகள் மட்டுமே நிலைத்து நிற்கும். அதில் குறியீடு ஒரு முக்கிய பங்கு வகிக்கிறது.

அடுத்து, தரவுகளை, சொற்பட்டியல் போன்றவற்றை எவ்வாறு வெளி நினைவகத்தில் எழுதி வைக்கிறோம் என்பது முக்கியம். அத்துடன் எந்தத் தரவுகளை வைக்க வேண்டும் என்பதுவும் முக்கியம்.

அடிப்படையில், பெயர்ச் சொல், வினைச்சொல் என இரு முக்கிய வகைகள் இருந்தாலும், அவற்றின் மாற்றங்களும், இணைப்புக்களும் பலப்பல. அவற்றில் சிலவற்றை மட்டும் இங்கு பார்ப்போம். இவை எடுத்துக்கொண்டுள்ள கணிப்பின் சிக்கலின் தன்மையைக் காட்டும்.

பெயர்ச் சொல்லுடன் வேற்றுமை உருபுகள் சேர்ந்து வரும். இணைப்புக்களின் என்னும் சொல்லில் இணைப்பு, கள், இன் என்று மூன்று பகுதிகள் உள்ளன. இரு பெயர்ச் சொற்கள் இணைந்தும் வரலாம். மாஞ்சோலை ஒரு எடுத்துக்காட்டு.

ஒரு வினைச் சொல் பல வடிவங்களில் வரலாம். வா என்னும் வினை, வந்தான், வருகிறான், வருவான், வந்துகொண்டிருக்கிறான், வந்தானா, வந்த. வரும், வராத, வருகின்ற, வராமல் என்பவை போன்று வரலாம். அத்துடன், வந்தவன், வந்தவர்கள், வராதவர்கள் போன்று பெயர்ச் சொற்களை உருவாக்கலாம். இவை இன்னும் மாற்றம் பெற்று, வந்தவர்களைத்தானேயப்பா என ஆகலாம்.

மரம் என்பது போன்ற சொற்கள் மாறும்போது அத்து என்னும் சாரியை பெற்று, மரத்தை என்று ஆகலாம். மாவும் சோலையும் சேரும்போது நடுவில் ஞ் சேருகிறது. மரமும் வேரும் சேர்ந்து மரவேர் ஆகிறது. பாலும் கடலும் சேரும்போது பாற்கடல் ஆகிறது.

ஒரு சொல்லின் எல்லா மாற்றங்களையும் நினைவகத்தில் வைப்பது என்பது இயலாத காரியம். அடிப்படைச் சொற்களை மட்டுமே வெளி நினைவகத்தில் வைத்திருக்க முடியும். கொடுக்கப்பட்ட சொல் அவற்றில் இருந்து உருவாக்கப்பட்டதா என்பதைக் கண்டறிய

வேண்டும். இதற்கு சொல்லை எந்த அளவு புத்திசாலித்தனத்துடன் பிரிக்கிறோமோ அந்த அளவிற்கு நேரம் குறையும்.

அடுத்து, செயல்பாடுகள் ஒவ்வொன்றினைமே நேரத்தினைக் குறைவாக எடுத்துக் கொள்ளும் வகையில் மிகுந்த கவனத்துடன் எழுத வேண்டும். இதற்கு, தரவுகளை கணிப்பொறியில் எத்தகைய தரவு அமைப்புகளில் (Data Structures) வைக்க வேண்டும் என்பதைக் கவனமாகத் தேர்வு செய்ய வேண்டும். இவற்றைத் திறமையாகப் பயன்படுத்தும் வகையில் இலக்கணக் கருத்துக்களை செயல்முறைகளாக (Algorithms) எழுத வேண்டும். இந்தச் செயல்முறைகள் ஒவ்வொன்றும் நேரத்தைக் குறைவாகச் செலவிடும் வகையில் திறமையாக எழுதப்பட வேண்டும். இந்த வடிவமைப்பு முறைகள்தான் சொல் திருத்தியின் வெற்றி தோல்வியைத் தீர்மானிக்கும். இவற்றை சரியானபடி வடிவமைக்க, தமிழ் இலக்கணத்தை சரியாக அறிந்துகொள்ளும் ஆற்றலும், தரவு அமைப்புகள் மற்றும் செயல்முறைகளைத் திறம்பட வடிவமைப்பதற்குத் தேவையான அனுபவமும் தேவை. இவற்றில் எது குறைந்தாலும் சரியான சொல்திருத்தி உருவாக முடியாது.

4. முடிவுரை

தமிழில் சொல்திருத்தி என்றால், அதில் என்ன எதிர்பார்க்கலாம் என்பதையும், அந்த எதிர்பார்ப்புகளை நிறைவேற்ற வேண்டுமென்றால் எதிர்நோக்க வேண்டிய சிக்கல்கள் பற்றியும் பார்த்தோம்.

தமிழில் சொல்திருத்தி என்னும்போது, அதனுடன், மற்ற இரு செயல்களும் கூட வருகின்றன. தவறான சொல்லைக் கண்டறிந்தவுடன், அந்த இடத்தில் இருக்கலாம் என ஊகிக்கக்கூடிய சில சொற்களின் பட்டியல் ஒன்று கொடுக்கப்படலாம். இதற்கு ஒரு சொல் சரியாக இல்லாத போதும் அது எந்த இலக்கண வகையைச் சேர்ந்தது என்று ஊகிக்க முடிய வேண்டும். அப்போதுதான் இதனைச் செயலாக்க வேண்டும். ஆனால் இது அவ்வளவு எளிதான காரியம் இல்லை.

சொல்லின் இறுதியில் வல்லினம் மிகுமா என்று பார்ப்பது அடுத்த செயல். இதற்கும் ஒவ்வொரு சொல்லின் இலக்கண வகையும் தெரிய வேண்டும். அத்துடன் கூட பல விதிவிலக்குகளையும் கையாள வேண்டும்.

இந்த இரு செயல்பாடுகளையும் தமிழ்ச் சொல்திருத்தியின் மேம்பட்ட நிலை எனக் கருதலாம். தமிழுக்கு சொல்திருத்திகள் இருப்பதாக சில மென்பொருட்கள் கூறுகின்றன. இவற்றின் திறன், அவை எடுக்கும் நேரம், சொற்களை சரியா தவறா என்று சரியாகக் கண்டுணர்தல் போன்றவற்றில் பெரிதும் வேறுபடுகின்றன. எவ்வளவு நேரம் எடுக்கும், எவ்வகைச் சொற்களைச் சரியாகக் கண்டுணரும் என்பதற்கு ஒரு அளவுகோல் இருந்தால் அது பயன்படுத்துவோருக்கு மிகவும் பயனுள்ள தகவலாக இருக்கும். அத்தகைய அளவுகோல் ஒன்றை விரைவாக உருவாக்குவது மக்களுக்குப் பெரிதும் உதவும். ஏனென்றால் அப்போது பல மென்பொருட்களை ஒப்பிட்டுப் பார்க்க ஒரு சரியான வழி கிடைக்கும்.

இயற்கை மொழியாய்வு - விரிதரவு

Mr. K. Subbiah Pillai
Senior Research Fellow/Senior Lectrurer, International Inst. of Tamil Studies
CIT Campus, Tharmani Post, Chennai - 600113, Tamil Nadu, India.

கருத்துப் பரிமாற்றம் (Communication) நிகழ மொழி ஒரு கருவியாகப் பயன்படுகிறது. இக்கருத்துப் பரிமாற்றம் மனிதர்களிடையேயும் விலங்கினங்களிடையேயும் நிகழ்கிறது. தொடக்க நிலையில் காட்சிக் குறியீடுகளின் (Visual Symbols) துணையோடு கருத்துப் பரிமாற்றம் நிகழ்ந்தது. இக் காட்சிக் குறியீடுகளால் தேவைகளை நிறைவு செய்ய இயலாது என்ற நிலை ஏற்பட்டதும் பேச்சொலிகள் (Vocal Sounds) கருத்துப் பரிமாற்றத்திற்குப் பயன்படுத்தப்பட்டன. விலங்கினங்களும் தத்தம் தேவைகளை நிறைவு செய்து கொள்ள மிகக் குறைந்த அளவிலான ஒலிகளை எழுப்பிக் கருத்துப் பரிமாற்றம் நிகழ்த்துகின்றன. கருத்துப் பரிமாற்றம் நிகழப் பயன்படும் மொழிகளை 1. சைகை மொழி (Gesture Language) 2. செயற்கை மொழி (Artificial Language) 3. இயற்கை மொழி (Natural Language) என மூவகைப்படுத்தலாம். மனிதர்களிடையே பொதுநிலையில் கருத்துப்பரிமாற்றம் மனித மொழியாகிய இயற்கை மொழியின் துணையோடு நிகழ்கிறது. இம்மொழியை மனிதன்தான் சார்ந்த சூழலிலிருந்து பெற்று பயன்படுத்தித் தன் அன்றாடத் தேவைகளை நிறைவு செய்கிறான். மொழியியலாளர்கள் (Lingusits) மனித மொழியான இயற்கை மொழியை ஆராய்ச்சி செய்வதையே நோக்கமாகக் கொண்டுள்ளனர். இயற்கை மொழியாய்வுக்காக வடிவமைக்கப்படும் வழியமைப்புகளை (Programs) இயற்கை மொழியாய்வு அமைப்புகள் (Natural Language Systems) என்பர்.

கணினியைப் பயன்படுத்தாத துறைகளே இல்லை எனலாம். வளர்ந்து வரும் அறிவியல் உலகில் எங்கும் எதிலும் கணினியின் பயன்பாடு பல்கிப் பெருகியுள்ளது. இவ்வாறு கணினியைப் பயன்படுத்துவதற்கான காரணங்களை ஆராயுமிடத்து அதன் வேகம் (Speed) கொள்திறன் (Capacity) நேர்த்தி (Accuracy) தானியக்கம் (Auto mation) நம்பகத் தன்மை (Reliability) போன்ற சிறப்பியல்புகள் குறிப்பிடப்பட வேண்டியவைகளாகும். கணிதம் தொடர்பான செயலாக்கங்களுக்காக (Process) வடிவமைக்கப்பட்ட கணினி இன்று மொழியாய்விலும் பரவலாகப் பயன்படுத்தப்பட்டு வருவதைக் காணலாம். இதன் காரணமாக இயற்கை மொழியாய்வு/கணினி மொழியியல் (Natural Language processing-NLP/Computational Linguistics) என்ற மொழியியல் (Linguistics) பிரிவு வளர்ந்து வருகிறது. கணினி மொழியியல், கணினி அறிவியல் (Computer Science) கோட்பாடுகளையும் (Theories) மொழியியல் கோட்பாடுகளையும் (Linguistic Theories) உள்ளடக்கியது.

கணிணியின் செயலாக்கத்திற்குத் (process) தேவையான மூலக்கூறு தரவு (Data) ஆகும். இத்தரவுகளின் அடிப்படையில் செயலாக்கம் செய்வதால் இதனை மின்னணு தரவு செயலாக்க இயந்திரம் (Electronic Data processing Machine-EDP) எனலாம். தரவுகள் எண் தரவுகளாகவோ (Numeric Data) அல்லது எழுத்துத் தரவுகளாகவோ (String data) இருக்கும். இயற்கை மொழியாய்வு மேற்கொள்ள பயன்படும் தரவுகள் பெரும்பாலும் எழுத்துகளாலான உரைத்தரவுகளாக (Textual Data) இருக்கும். உரைத்தரவுகள் எனக்கூறும் போது அவை எழுத்துகளாகவோ (Characters) அல்லது எழுத்துகளாலான சொற்களாகவோ (Words) அல்லது சொற்களாலான சொற்றொடர்களாகவோ (Sentence) அல்லது சொற்றொடர்களாலான பந்திகளாகவோ (Paragraph) அல்லது பந்திகளாலான இயலாகவோ

(Chapter) அல்லது ஒரு படைப்பாளியின் படைப்பிலுள்ள ஒட்டு மொத்த உரையாகவோ (Whole Text) இருக்கலாம். இக்கட்டுரை இயற்கை மொழியாய்வில் விரிதரவு (Corpus) குறித்து விளக்க முற்படுகிறது.

இயற்கை மொழியான மனித மொழியின் மாதிரி (Sample) உரைத் தொகுப்பை விரிதரவு என்பர். மொழியாய்வு மேற்கொள்ள பெரிதும் உதவும் இதனை பல்வேறு எழுத்து ஊடகங்களிலிருந்தும் (Written media) பேச்சு ஊடகங்ளிலிருந்து (Spokan media) தொகுப்பர். இந்த ஊடகங்களிலிருந்து திரட்டப்பட்ட உரைத் தொகுப்புகள் கணினியால் படிக்கும் வடிவில் (Machine Readable Form) வன்தகட்டிலோ (Hard Disk) அல்லது மின்காந்த நாடாக்களிலோ (Magnetic Tape) அல்லது அடக்கத் தட்டுகளிலோ (Compact Disk-Read only Memory-CD-ROM) திரட்டி வைக்கப்படுகின்றன. மொழியியலாளர்கள் தாங்கள் மேற்கொள்ள விரும்பும் மொழியாய்வுக்குப் பயன்படுத்தும் நோக்கத்தில் இவ்விரிதரவு திரட்டி வைக்கப்படுகிறது. மேலும் விரிதரவைத் தரவாக ஏற்றுக் கொண்டு செயலாக்கம் செய்ய வடிவமைக்கப்படும் வழியமைப்பின் துணையோடு பல தகவல்களை (Information) உடனுக்குடன் பெறலாம்.

விரிதரவு சொற்களாலான சொற்றொடர்களாலானது இச் சொற்றொடர்களின் அமைப்பை விளக்க முற்படுகையில் மொழியின் இலக்கண அமைப்பையும் எளிதில் அறிந்து கொள்ளலாம். பெருவாரியான விரிதரவைத் திரட்டி வைக்கவும் இத்தரவின் அடிப்படையில் மொழியாய்வு மேற்கொள்ளவும் இன்றைய நிலையில் கணினி ஒன்றே பேருதவியாக இருந்து வருகிறது. இதற்கான காரணம் இதன் சிறப்பியல்புகள் தான் எனத் தெளிவு பெறலாம் . மொழியில் சொற்களின் ஆளுமை, பொருள் வீச்சு (Range of applicability) பயன்பாடு (Function) போன்ற அனைத்துப் பண்புகளையும் அறிய விரிதரவு உதவுகிறது. தொடக்க நிலையில் இவ் விரிதரவு மொழியியலாளர்களின் கவனத்தைப் பெரிதும் ஈர்க்கவில்லை என்றே கூறலாம். பின்னர் தகவல் புரட்சியின் (Information Revolution) காரணமாகப் பெறப்பட்ட இணையம் (Internet) உள்ளிட்ட பல்வேறு தகவல் தொழில்நுட்ப (Information Technology) உத்திகள் மொழியியலாளர்களுக்குப் பல விரிதரவுகளை எளிதில் கிடைக்க வகை செய்தன. இதன் காரணமாக இயற்கை மொழியாய்வு பல நிலைகளில் வளர்ந்து வருகிறது.

தொடக்க நிலையில் பல உள்ளீட்டகங்களின் (Input Units) துணையோடு விரிதரவுகள் கணினியால் படிக்கும் வடிவில் உள்ளீடு (Input) செய்யப்பட்டன. இவ்வுள்ளீட்டகங்களில் குறிப்பிடத்தக்கது விசைப்பலகை (Key Board) இயற்கை மொழியாய்வு மேற்கொள்ள பெருவாரியான உரைத்தரவுகள் தேவைப்படுவதால் இவற்றை பல பணியிடங்களிலிருந்து (Work Station) பல தரவு உள்ளீட்டாளர்களின் (Data Entry operators) துணையோடு உள்ளீடு செய்ய விசைப்பலகை பயன்படுத்தப்பட்டது. தற்சமயம் தொழில்நுட்ப வளர்ச்சியின் காரணமாகப் பெறப்பட்ட ஒளி எழுத்துப் படிப்பான் (Optical Character Recognition/Reader) என்ற கருவி தரவுகளைக் கண்டெடுக்க (Data Capture) உதவுகின்றது. ஆங்கில விரிதரவுகளைத் திரட்ட குருசு வேல் தரவு உள்ளீட்டு இயந்திரம் (Kurzweil Data Entry Machine- KDEM) என்ற உள்ளீட்டுக் கருவி பயன்படுத்தப்பட்டு வருகிறது.

தமிழ் மொழி உள்ளிட்ட பல இந்திய மொழிகளுக்கு விரிதரவுகள் திரட்டப்பட்டு வருகின்றன. இவ்விரிதரவுகளைக் கணினியால் படிக்கும் வடிவில் உள்ளீடு செய்ய தரவு உள்ளீட்டாளர்கள் விசைப்பலகையையே பயன்படுத்துகின்றனர். இந்திய மொழிகளைக் கணினியில் உள்ளீடு செய்ய KDEM போன்ற படிப்பான்கள் பயன்பாட்டில் இல்லை. எனவே, இத்தரவுகளை உள்ளீடு செய்ய அதிக நேரம் செலவாகிறது என்றே கூறலாம். KDEM போன்ற படிப்பான்களை வடிவமைக்கும் முயற்சியில் பல இந்திய நிறுவனங்கள் ஈடுபட்டு வருகின்றன. இம் முயற்சி வெற்றி

பெறும் நிலையில் தமிழ் மொழியிலுள்ள உரைத் தரவுகளை விரைந்து உள்ளீடு செய்து கணினியால் படிக்கும் வடிவில் திரட்டி வைக்கலாம்.

ஆங்கில விரிதரவுகளில் மொழியாசிரியர்களால் பெரிதும் பயன்படுத்தப்படும் விரிதரவுகள் பிரவுண் விரிதரவு (Brown Corpus, 1961-1964) அமெரிக்காவிலுள்ள பிரவுண் பல்கலைக்கழகத்தில் நெல்சன் பிரான்சிஸ் (Nelsen francis) ஹன்றி குச்சரு (Henry Kurcera) என்பவர்களால் தொகுக்கப்பட்ட முதல் அமெரிக்க ஆங்கில விரிதரவு (American English Corpus) திரட்டும் நோக்கத்திற்கேற்ப இந்த விரிதரவைத் தேவைப்படுவோர் பெற்று பயனடயலாம். இது பத்துலட்சம் சொற்களைக் கொண்டது. 500 தலைப்புகளில் மாதிரி உரைகளாகப் பிரிக்கப்பட்டு ஒவ்வொரு தலைப்பிலும் 2000 சொற்களை உள்ளடக்கியது. லங்காஸ்டர் ஓஸ்லோ பெர்கின் விரிதரவு (The Lancaster - oslo / Bergen Corpus) ஏறத்தாள பிரவுண் விரிதரவு போன்று இருந்தாலும் இது பிரிட்டிஷ் ஆங்கில விரிதரவாகும் (British English corpus) .

பேச்சு மொழியில் விரிதரவைத் திரட்டுவது எழுத்துத் தரவுகளை விட கடினமானது. பேச்சு மொழியை ஒலி பெயர்த்த (Translileration) பின்னரே இவற்றை கணினியால் படிக்கும் வடிவில் உள்ளீடு செய்யலாம். பேச்சு விரிதரவுகளில் லண்டன் லண்டு விரிதரவு (London Lund Corpus) 500,000 சொற்களைக் கொண்டது. ரண்டால்ப் குர்க் (Randolph Quirk) என்பவரின் வழிகாட்டலில் மிகநேர்த்தியாக ஒலிபெயர்க்கப்பட்டு (Transliteration) கணினியில் உள்ளீடு செய்யப்பட்டது. இதனைத் தொடர்ந்து லங்காஸ்டர் ஆங்கில பேச்சு விரிதரவு (Lancaster Spoken English Corpus-SEC) மிக நேர்த்தியாக ஒலி பெயர்க்கப்பட்டிருப்பதுடன் சொற்களின் இலக்கண வகைகளும் இடஞ் சுட்டப்பட்டிருக்கின்றன (Annotation)

அமெரிக்க கணினி மொழியியல் கழகம் (Association of Computational Linguistics) விரிதரவுகளை எவ்வாறு திரட்ட வேண்டும் என்றும் அவற்றை எவ்வாறு தேவைப்படுவோர் பெற வேண்டும் என்பதை விளக்கும் வண்ணம் தரவு திரட்டும் வழிமுறைகள் (Data Collection Initiatives) உரை விளக்க வழிமுறைகள் (Text encoding Initiatives) என்ற இரு திட்டப்பணிகளை மேற்கொண்டது. விரிதரவுகளைப் பெற்றுப் பயன்படுத்துவதில் பதிப்பு உரிமை (Copyright) தொடர்பான சட்டச் சிக்கல்களும் இருந்து வந்தன.

உரைப்பகுப்பாய்வு;- (Textual analysis)

இயற்கை மொழியாய்வு உரைத் தரவுகளின் அடிப்படையில் கீழ்க் கண்ட நிலைகளில் மேற்கொள்ளப்படுகிறது .

1. ஒலியமைப்பியல் பகுப்பாய்வு phonological analysis
2. உருபனியல் பகுப்பாய்வு morphological analysis
3. தொடரியல் பகுப்பாய்வு Syntaitic analysis
4. பொருண்மையியல் பகுப்பாய்வு Semantic analysis

ஒலியமைப்பு பகுப்பாய்வு ;-

ஒலியமைப்புப் பகுப்பாய்வு (Phonological analysis) ஒலியியல் (Phonolics) ஒலியனியல் (Phonemics) என்ற இரு பிரிவுகளை உள்ளடக்கியது.

ஒலியியில்;-

உச்சரிப்பு ஒலியியல் (Articalatory Phonetics) ஒலியியக்கவியல் (Acoustics Phonetics) ஒலியுணர்வியல் (Auditory phonetics) என்ற மூன்று ஒலியியல் (phonotics) பிரிவுகளில் ஒலியியக்கவியல், பேச்சு இணைப்பாக்கம் (Speech Synthecis) என்ற நிலையில் ஆராய்ச்சி செய்ய பெரிதும் உதவுகிறது.

கணினியில் தட்டச்சு செய்து உள்ளீடு செய்யப்பட்ட உரையை ஏற்றுக் கொண்டு அவ்வுரையை உச்சரிக்கும் அமைப்பைப் பேச்சு இணைப்பாக்கம் எனலாம். இதே போன்று மனித பேச்சுறுப்புகளின் அசைவினால் உருவாக்கப்படும் பேச்சைப் புரிந்து கொண்டு உரையாக வெளியிடும் அமைப்பை பேச்சை உணர்தல் (Speech Recognition) எனலாம். ஸ்பெக்டோகிராப் (Spectrograph) என்ற கருவியின் துணையோடு பேச்சொலிப்படங்களை (Spectrogram) எடுத்து ஆராய்ச்சி மேற்கொள்ளவும் மொழியின் ஒலியமைப்பு முறையை அறியவும் விரிதரவு உதவுகிறது. ஒலியனியல் (Phonomics), விரிதரவில் பயின்று வந்த மொத்த எழுத்துகள் (Total Numbers of letters) குறிப்பிட்ட எழுத்தின் நிகழ்வெண்ணிக்கை (frequency of letters) போன்ற தரவுகளைப் பெற்று புள்ளியியல் (Statistics) அடிப்படையில் ஆராய்ச்சி செய்ய உதவுகிறது.

உருபனியல் பகுப்பாய்வு;-

ஒரு மொழியில் பயின்று வரும் ஒலியன்களை மட்டும் அறிந்திருந்தால் போதுமானதன்று. ஒலியன்கள் முறையாக ஒன்றன் பின் ஒன்றாக நிரல்படுத்தப்பட்டு பொருள் உணர்த்தும் நிலையில் சொல்லாக (Word) மாறும் போதுதான் அது கருத்துப்பரிமாற்றத்திற்குப் பயன்படும் நிலையை அடைகிறது. எனவே, ஒன்று அல்லது ஒன்றுக்கு மேற்பட்ட ஒலியன்களின் கூட்டால் பெறப்படும் பொருள் உணர்த்தும் பகுதியை உருபன் (morpheme) என்கிறோம். விரிதரவில் பயின்று வரும் சொற்களைப் பகுத்தெடுத்து அச்சொற்களின் உள் அமைப்பைத் (Internal Structure) தனி உருபன் (free morpheme) என்றும் கட்டுருபன் (Bound morpheme) என்றும் பகுத்து அவற்றின் அமைப்பை அறிந்து கொள்ளலாம்.

தனி உருபன் தனித்துப் பயின்று வரும் . எனவே, இத் தனி உருபன்களைத் தலைப்புச் சொற்களாக (Lexical Head words) கொண்டு அகராதி தொகுக்கலாம். மின்னனு அகராதி தொகுள்ள விரிதரவு பெரிதும் உதவுகிறது. தமிழ் அகராதி வரலாற்றை நிகண்டுகளுக்கு முற்பட்ட காலம், நிகண்டு காலம், அகராதி காலம் என வகைப்படுத்துவர். செய்யுள் வடிவில் எழுதும் வழக்கம் இருந்த காலத்தில் அகராதிகள் நிகண்டுகள் என வழங்கப் பெற்றன. ஐரோப்பியர் வியாபாரத்திற்காக வந்தாலும் தத்தம் மதங்கள் தொடர்பான செய்திகைளப் பரப்ப முற்பட்ட போது நிகண்டு வடிவம் ஏற்றதாக அமையவில்லை. எனவே, சொற்களுக்கு எளிதில் பொருள் விளக்கும் நோக்கத்தில் அகராதிகள் தொகுக்கப்பட்டன.

மின்னணு அகராதி தொகுக்கும் பணி ஜார்ஜ் டவுண் பல்கலைக்கழகத்தில் (George Town University) மொழி பெயர்ப்பைக் கருத்தில் கொண்டு தொகுக்கப்பட்டது. ருசிய ஆங்கில மின்னணு அகராதி இரசாயனம், இயற்பியல், உயிரியல், சமூகவியல் தொடர்பான 50,000 சொற்களுக்கு அகராதி தொகுத்தது. ஐ.பி.எம். நிறுவனம் (I B M Corporation) அமெரிக்க விமானப் படைக்கு உதவ ருசிய- ஆங்கில இருமொழி (Bilingual) அகராதியை 150000 சொற்களால் தொகுத்தது. இதே போன்று பல மின்னணு அகராதிகள் வெளி வரத் தொடங்கின. தமிழ் மொழியில் முதல்முதலில் வெளிவந்த மின்னணு அகராதி கிரியாவின் தற்காலத் தமிழ் அகராதி, மின்னணு அகராதிக்கும் மரபுவழி அகராதிக்கும் உள்ள வேறுபாடு மின்னணு அகராதியில் எப்போது வேண்டுமானாலும் திருத்தங்களை மேற் கொண்டு புதுப்பிக்கலாம்

(Revise) ஆனால் மரபுவழி அச்சிடப்பட்ட அகராதியில் திருத்தங்கள் மேற்கொள்ள மறுபதிப்பு செய்யும் போதுதான் இயலும்.

விரிதரவைப் பயன்படுத்தி சொல்லடைவு (Concordance) தயாரிக்கலாம். இது குறிப்பிட்ட ஒரு சொல் மொழியில் எந்தெந்த சூழலில் பயன்படுத்தப்பட்டுள்ளது என்பதை அறிய உதவும். இவ்வாறு விரிதரவுகளிலிருந்து சொல்லடைவு தயாரிக்க உதவும் வழியமைப்பைச் சொல்லடைவான் (Concordancer) எனலாம். நாம் எந்தச் சொல்லைத் தேடப் பணிக்கிறோமோ, அந்தச் சொல் பயின்று வந்த அனைத்துச் சொற்றொடர்களும் அதிவேகமாக நிரல்படுத்தப்படும். இதனை சூழல் தலைப்புச் சொல் (Key word in Context/KWIC) எனலாம். குறிப்பிட்ட ஒரு சொல் எத்தனை முறை பயன்படுத்தப்பட்டுள்ளது என்பதைப் புள்ளியியல் (Statistics) அடிப்படையில் அறிந்து கொள்ளலாம். இப்புள்ளியியல் தரவுகள் கணினி நடையியல் (Computational Stylistics) ஆராய்ச்சிக்குச் செறிவு சேர்க்கும்.

தொடரியல் பகுப்பாய்வு (Syntactic analysis)

சொற்கள் இலக்கண விதிகளுக்கும் பொருண்மை விதிகளுக்கும் உட்பட்டு நிரல்படுத்தப்படும் போது சொற்றொடர்கள் உருவாக்கப்படுகின்றன. 1950- ஐ ஒட்டிய காலப்பகுதியில் மாற்றிலக்கணம் (Transformational Grammar) உருவாயிற்று. சாமஸ்கி (Noam Chomsky - 1928) எழுதிய ' தொடரியல் அமைப்புகள்' (Synatatic Structires - 1957) தொடரியல் ஆராய்ச்சியில் ஒரு விழிப்புணர்வை ஏற்படுத்தியது. இந்த நூலைத் தொடர்ந்து வெளியான தொடரியல் கோட்பாட்டு நெறிமுறைகள் (Aspects of the theory of syntex) மாற்றிலக்கணக் கோட்பாடுகளுக்கு மேலும் மெருகூட்டின.

தொடரியல் பகுப்பாய்வில் கணினி உள்ளீடாகக் கொடுத்த சொற்றொடர்களை ஏற்றுக்கொண்டு அச்சொற்றொடர்களுக்கு உரித்தான கிளைப்படத்தினைத் திரையில் தோன்றச் செய்ய வேண்டும். உள்ளீடாகக் கொடுத்த சொற்றொடரில் செயலாக்கங்களைச் செய்யும் பகுதியை ஒழுங்குமுறைப் பகுதி எனலாம். தொடரை இனம் கண்டு கொள்வதோடல்லாமல் சொற்கள் எவ்வாறு நிரல்படுத்தப்பட்டுள்ளன என்பதைக் காண இது உதவுகிறது. இவ்வாறு காணுதலைப் பகுத்தல் (Parsing) எனலாம். பகுத்தலைச் செய்யும் ஒழுங்குமுறைப் பகுதியைப் பகுப்பான் (Parser) எனலாம். பகுப்பான் என்பது ஒரு வழியமைப்பு. இவ் வழியமைப்பு உள்ளீடாகக் கொடுத்த சொற்றொடரை ஏற்றுக் கொண்டு அச்சொற்றொடரில் பயின்று வரும் சொற்களுக்கு உரித்தான இலக்கண விளக்கங்களுக்கு ஏற்றாற் போலப் பகுத்துக் கிளைப்படமாகத் தோன்றச் செய்ய வேண்டும்.

விரிதரவு மொழியின் மாதிரிச் சொற்றொடர்களாக இருப்பதால் இச் சொற்றொடர்களில் பயின்று வரும் சொற்களின் இலக்கண வகையைத் தெளிவாகக் குறிப்பிட வேண்டும். இவ்வாறு குறிப்பிடுவதை இடஞ்சுட்டல் (Annotation) எனபர். இடஞ்சுட்டிய பின் பெறப்படும் சொற்றொடர்களின் கிளைப்படங்ளைத் தொகுத்துக் கிளைப்பட வங்கியை (Tree Bank) உருவாக்கலாம். இது இயந்திர மொழி பெயர்ப்பு (Machine Translation) தொடர்பான ஆராய்ச்சி மேற்கொள்ள உதவும்.

பொருண்மையியல் பகுப்பாய்வு (Semantic analysis)

ஒவ்வொரு சொல்லுக்கும் அச்சொல்லுக்கே உரித்தான அமைப்பு (Structure) மற்றும் பயன்பாடு (function) உண்டு. தொடரியல் ஆய்வின் வாயிலாகச் சொற்களைப் பகுக்கச் சொற்களுக்குரித்தான இலக்கண உள்வகைப்பாட்டை (Sub categorization) விளக்கங்களை

வரிசைப்படுத்த வேண்டும். இந்த உள்வகைப்பாட்டு விளக்கம் சொற்கள் சொற்றொடரில் எவ்வாறு பயன்படுத்தப்பட்டுள்ளது என்பதை விளக்குகிறது. எனவே பொருண்மைப் பகுப்பாய்விற்குச் சொற்றொடரில் பயின்று வந்த சொற்களின் சொற்கூறுகள் (Lexical features) தேவைப்படுகின்றன. இச் சொற்கூறுகளைக் கணினிக்குப் புகட்டுவதில் செயற்கைப் புலமைப் (Artificial Intelligence) பிரிவைச் சார்ந்த கணினி வல்லுநர்கள் மிகவும் முனைப்புடன் முயன்று வருகிறார்கள். சொற்றொடரில் பயின்று வந்த சொற்கள் ஒவ்வொன்றிற்கும் சட்டம் (frame) என்ற வடிவமைப்பை ஏற்படுத்துவதுடன் வேற்றுமைச் சட்டங்களை (Case frames) உருவாக்க பொருண்மையில் உத்திகள் துணை செய்கின்றன. விரிதரவுகளின் அடிப்படையில் சொற்களின் பொருண்மைப் பண்புகளை அறிந்து கொள்ளலாம்.

கணினி, தகவல் மீட்டல் (Information Reterival) என்ற நிலையில் மிகவும் பயனுள்ளதாக இருக்கிறது. விரிதரவுகளைத் தொகுத்து முறைப்படி ஆவணப்படுத்தும் (Documentation) நிலையில் இத்தரவுகள் தொடர்பான தகவலைத் தேவைப்படும் போதெல்லாம் விரைந்து பெற்று பயனடையலாம். சான்றாக பெயரெச்சத் தொடரின் (Relative Clause) அமைப்பு குறித்து ஆராய்ச்சி செய்யும் ஆய்வாளர்கள் இவ்விலக்கண அமைப்பைக் கொண்ட சொற்றொடர்களை விரைந்து திரட்டி எடுத்துக் கொள்ளலாம். இவ்வாறு திரட்டி எடுக்க வேண்டுமாயின் விரிதரவு, இடஞ்சுட்டும் (Annotation) சொற்களின் இலக்கண விளக்கத்தையும் சொற்றொடரின் அமைப்பையும் தெளிவாக உள்ளடக்கி இருக்க வேண்டும்.

விரிதரவு குறிப்பிட்ட மொழி தொடர்பான அனைத்து அமைப்பு விதிகளையும் விளக்கும் நோக்கத்தில் திரட்டப்படுகின்றது. இதனை மீண்டும் மீண்டும் தேவைப்படுவோர் பெற்று பயன்படுத்தும் (Reusability of Resource) வண்ணம் கிடைக்கச் செய்வதால் இயற்கை மொழியாய்வு - தமிழ் பல நிலைகளில், இணையத்தின் துணையோடு விரிவடையும் என்பது திண்ணம்.

துணைநூல் பட்டியல்

Butler, C. S. 1985 Computers in Linguistics, Basil Black well Ltd., New York

_____ 1992 Computers and Written Texts[ed.], Basil Black well Ltd., New York

Davis, G. B. 1969 Computer data processing, MC Goraw Hill, INC., Sydney, [Reprinted 1986]

Wishman, R 1986 Computational Linguistics An Introduction Cambridge University Press, Cambridge

Harvis, M. D. 1985 Introduction to Natural Language Processing, Reston publishing company, INC, Virgina

Hill, A. A 1969 Linguistics [ed.] voice of America Forum Lectures, U.S.

King, M 1983 Parssing Natural Language [ed.] Academic press, New York

Landau, S. I. 1984 Dictionaries the Art and craft of Lexicography, Cambridge University

Leech, G.et.al, 1995 Spoken English [ed.] Longman, New York

Subbiah pillai, K. 1990 Computer Analysis of Aryhadin Wiippu, unpublished project work Department of Electronics, CIT , Chennai

Subbiah pillai, K1992 Computer Analysis of case system in modern Tamil [unpublished project work] Department of Electronics CIT Chennai

_____ 1998 இயற்கை மொழியாய்வு - தமிழ் உ.த. நிறுவனம் சென்னை

Winston, P.H. 1992 Artificial Intelligence, Addition wesley publishing company California [Reprinted 1993]

204 empty

Teaching of Tamil Scripts and Their Impact Through Keyboard of Computers

Dr. N. Nadaraja Pillai,
Central Institute of Indian Languages
Manasagangotri, Mysore - 570 006, India

Introduction

The main concern of this paper is to explore the possibility of integrating the teaching of Tamil scripts and its input through the keyboard of the computer. It is obvious that keyboard operation should be same one as the writing/hand movement of the letters. Only this will help the user understand the input method.

Tamil as everybody knows follows a syllabic system and hence the keyboard operation should also follow that.

Teaching of Tamil Scripts

The teaching of Tamil scripts, now-a-days is based on the strategy of pattern perception, contrastive observation and similarity in hand movements. Following this principle, a new method of teaching the Tamil scripts was evolved, in which the letters are grouped into 11 groups. They are as follows:

1. Ta, pa, ya, ma, za
2. ii, ra, ca, ka, ta
3. a, aa, I
4. na, nga
5. e, ee, nja
6. la, va
7. Ra, ai
8. La, na, Na
9. o, oo, au
10. u, uu, ahu
11. sa, sha, ja, ha, ksha, sri

The secondary symbols of the vowels will be introduced as and when the vowel is taught.

Even if, this shape similarity method is followed, the teaching script has to end in the traditional arrangement of letters also as in,

a, aa, i, ii, u, uu, e, ee, ai, o, oo, au - the vowels and
k, ng, c, nj, T, N, t, n, p, m, y, r, l, v, z, L, R, n - the consonants and

finally the grantha letters, namely,

s, sha, ja, ha, ksha, and sri.

The learning of this arrangement helps the learners to refer to the Tamil dictionary, since the dictionary entries are arranged in the traditional system only. Any arrangement based on the manner or point of articulation will definitely hamper the learning . Take for example; there is a move to arrange the consonants in the following way.

k, ng, c, nj, T, N, t, n, p, m, R, n, y, r, l, v, L, z.

Yet another arrangement suggested is as follows:

k, c, T, t, p, R, vallinam 'hard sounds'
ng, nj, N,n, m, n mellinam 'soft/nasal sounds'
y, r, l, v, L, z, iDaiyinam 'in between sounds'

This is the tradional way of classifying the consonants into three.

There is yet another arrangement suggested based on the point of articulation. The following will suggest it.

k, ng velar sounds
c, nj, y Palatal sounds
T, N, L, Z Retroflex sounds
r, n, l, r Alveolar sounds
t, n Dental sounds
v Labio-dental sounds
P, m Bilabial sounds

None of the arrangement would help the learners except the shape similarity method and final arrangement of these on the traditional model. While the shape similarity method is based on similarity in the shape and hand movement, and helps in learning words, word formation, etc. after each group of letters, the traditional arrangement is mostly done on the point of articulation with minor variations. Hence, data input has a direct connection with the manner in which the learning the script is done.

Secondary Symbol and Writing System

There are some problems in the learning/ teaching of vowel-consonant letters. As far as vowels are concerned, there is no problem in learning, but the writing of vowel-consonant combination letters pose lot of problems.

There are six ways of writing them.

1. The secondary symbol follows the main letter.

ik + aa > kaa

2. The secondary symbols precede the main letter

ik + e > ke
ik + ee > kee
ik + ai > kai

3. The secondary symbols are written on both sides of the letter

ik + o > ko
ik + oo > koo
ik + au > kau

4. The secondary symbols is written on the letter

ik + i > ki
ik + ii > kii

5. The secondary symbols are written in three different ways depending on where the end point of hand movement stops while writing the main letter. The end points become the beginning of the secondary symbols. This is the case with u and up.

There are three ways of writing them.

(a) k + u > ku
T + u > Tu
m + u > mu
r + u > ru
n + u > nu
Z + u > ZU

(b) ng + u >ngu
c + u > cu
p + u > pu
y + u > yu
v + u >vu

(c) nj + u >.nju
N + u > Nu
t + u > tu
n + u > nu
l + u > lu
R + u > Ru
n + u >nu

The long counterparts are written in the same manner with minor modifications denoting the length. Except ( )all other letters behave regularly.

(a) ku Tu mu ru Lu Zu
kuu Tuu muu ruu Luu Zuu

(b) ngu cu pu yu vu
nguu cuu puu yuu vuu

(c) nju Nu tu nu lu Ru nu
njuu Nuu tuu nuu luu Ruu nuu

6. The secondary symbols of the grantha, as far as / u / and / uu / are concerned, are different as in the case of the following:
ju juu shu shuu

The above classification is the basic step for teaching the vowel- consonant letters. This has to be followed in the typing through the keyboard of the computer also. In case, a system of a consonant and a vowel leading to vowel-consonant formation is adopted, this will affect the learning of the scripts adversely. The secondary symbols discussed under 2,3 and 5 will not only disturb the learning but lead to errors also.

Automation of Typing

It is well known fact that the pure consonants in Tamil are to be written with over the letter. Though there are certain combinations fixed rules, which presuppose the occurrence of pure consonants, which facilitates typing through computers, it would be appropriate to type the pure consonants with dot over it since most of the errors are already due to not putting the dot. If a programme, which facilitates automatic dotting, were made, it would affect the learning and writing too. Hence it is advised to use the pure consonants in the keyboards and the combination of the pure consonants and the vowel will give the vowel-consonant letters. The computer will take care of the rest, which is not necessary for the user.

ik + a > ka
ik + aa > kaa
ik + I > ki, etc. and not

ka + aa > kaa
ka + I > ki
ka + u > ku, etc.

The following are some of the doubling of consonants.

Pakkam this will be typed as ip+a, ik, ik+a, m

Accam, caTTam, aNNan, cattam, appaa, ammaa, ayyan, vallam, kaLLan, maaRRam, annam

Normally only the nasal counterpart of the stop consonant only precedes them, as in
Tangkam, panjcam, paNTam, tantam, rampam, kunRam

There are a few exceptions also, as in:
anpu, maaNpu, kaNkaL, maankaL, kaaTci, muyaRci, paartteen, paayccineen, etc.

These can not be however, automatic. Hence the automatic dotting though look simple for programming actually hampers the learning as well as goes away from the tradition too.

In case we process these in such an automatic way, that is, the keeping a dot would be automatically done over the first one in these combinations, which may look a simplified one, would on the contrary will lead to errors. The hand movement while writing these words presupposes the dotting first before writing the second consonant of the combination.

Conclusion

An attempt has been made through this paper to give justification for a correlation between hand movements, writing secondary symbols, etc. and the key board input in the computer. Modifications based on traditional way should be followed at the end of teaching, if it is to facilitate the learners. Further more, Tamil scripts are not only to write the Tamil language, but the other minority as well as the tribal languages of Tamilnadu. This responsibility of Tamil should also be taken into consideration before finalizing any change.

210 empty

வட்டு இயக்க அமைப்பு சார்ந்த ஒலியியல் தமிழ் விசைப் பலகை இயக்கி வடிவமைத்தல்

Dr.மு.பொன்னவைக்கோ (தலைமைப் பேராசிரியர்),
பெ.இராமமூர்த்தி & ஜெ.பி.பிரசன்னா (B.E. இறுதியாண்டு மாணவர்கள),
கணிப்பொறித்துறை,
கிரசன்ட் பொறியியல் கல்லூரி, சென்னை 600 048.

முன்னுரை

கணிப்பொறியின் வளர்ச்சியும் பயன்பாடும் நாளுக்கு நாள் பெருகிக் கொண்டே வருகின்றது. கணிப்பொறி எல்லா துறையினருக்கும் ஒரு இன்றியமையாத கருவியாக வளர்ந்துள்ளதோடு மனிதனின் அன்றாட வாழ்க்கைக்கும் தேவையான சாதனமாக வளர்ந்து வருகின்றது. நகர மக்களோடு நின்றுவிடாமல் கிராம மக்களின் பயன்பாட்டிலும் கணிப்பொறி இடம் பெற்று வருகின்றது. கணிப்பொறியை பயன்படுத்த மிக முக்கியமான சாதனம் உள்ளீட்டு கருவியாகிய விசைப்பலகையாகும்.

இதுவரை வடிவமைக்கப்பட்டுள்ள கணிப்பொறி விசைப்பலகைகளை மூன்று வகைகளாக பிரிக்களாம். அவையாவன

1. ரோமன் எழுத்து விசைப்பலகைகள்.
2. தட்டச்சு வகை விசைப்பலகைகள்.
3. ஒலியியல் வகை விசைப்பலகைகள்.

தற்போது பயன்படுத்தும் மேக்கின்டாஸ் தமிழ் விசைப்பலகைகள் அனைத்தும் வின்டோஸ் (windows) தளத்தை மட்டுமே சார்ந்துள்ளது. ஆனால் பெரும்பாலான தொகுப்பிகள் DOSல் வடிவமைக்கப்பட்டுள்ளது. எனவே, தமிழ் தெரிந்தவர்கள் எளிதாக (Compilers) DOSயை உபயோகப்படுத்த இவ்விசைப்பலகையை உருவாக்கியுள்ளோம்.

தமிழ் தொகுப்பு (Compilers in Tamil) வடிவமைக்கத் ....ல எழுத்துக்களைக் கொண்டு வடிவமைக்கப்பட்டுள்ள கணிப்பொறி மொன்களைஸ் பயன்படுத்த தேவையான 'முற்செயலாக்கி' (Preprocessors) போன்றவற்றை வடிவமைக்க DOS - ஐ அடிப்படையாகக் கொண்ட விசைப்பலகைத் தேவைப்படுகிறது.

எனவே, இப்படிப்பட்ட பயன்பாடுகளுக்கும், DOS இயக்க அமைப்H(tm) தIN(tm) உருக்களை கணிப்பொறிக்கு உள்ளிட, DOS இயக்க அமைப்பு செயலும் விசைப்பலகை தேவைப்படுவதால், இந்த முயற்சியில், B.E. இறுதியாண்டு அடிப்படையாக, DOS-ஐ ஜூலையில் தமிழ் உருவடிவ அமைப்பில் விசைப்பலகை இயக்கியும் (Keyboard Driver) வடிவமைக்கப்பட்டது. அந்த விசைப்பலகையைப் பற்றி இக்கட்டுரை விவரிக்கிறது. இந்த விசைப்பலகை இட அமைப்பு தமிழ் இணையம்99 விசைப்பலகையைத் தழுவி அமைக்கப்பட்டுள்ளது.

Design:

Tamil Font Generation:

Fonts are very vital for any software, since a well-developed font has a good feel while working. Calling Interrupts and overwriting the previously stored character by means of the newly designed Tamil characters does the font generation. For accessing english and Tamil keys, the Tamil characters are installed between 162 to 255 keeping TABxxx - Bilingual coding scheme for Tamil as reference.

Tamil characters are generated in text mode in DOS platform. This is done by calling VDU interrupt number 10h along with the sub functions. The appearance of a character is designed in an 8 x 16 matrix. The matrix size cannot be increased or decreased, because, the subfunction of the above interrupt supports only 8x16 character size. If a pixel is to be lighted, '1' is entered in the matrix otherwise'0' is entered. In the Bitap according to the shape of the character the pixel is lighted or not lighted.

The mentioned interrupt requires some subfuntions namely, service number 11h, subfunction number 0h, number of bytes per character defined by the table and first character in the table.

In the character cell according to the shape of the character the pixel is lighted or not lighted.

For example, 'ம' is drawn as in Fig.1.

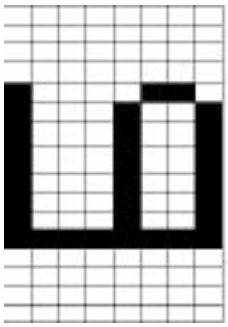

Fig.1 Character cell for a single character ம.

This is now converted to a hexa-decimal number in a row wise manner. Since the matrix has 16 rows, 16bytes of hexa-decimal number will be available for the generation of a single character. The hexa equivalent of the Tamil character 'ம' is

0x00, 0x00, 0x00, 0x00, 0x86, 0x89, 0x89, 0x89
0x89, 0x89, 0x89, 0xff, 0x00, 0x00, 0x00, 0x00

The hexa values of the ASCII character codes are converted into actual character shapes on the screen using translation table or character generator.

Mapping of Tamil fonts:

Tamil keyboard driver helps to input data into computer in Tamil, this requires Tamil fonts and its mapping to the keys of the keyboard. The aim is to map the Tamil keys(overwritten instead of extended ASCII characters) to the keyboard.

The keyboard contain a microprocessor which senses the key pressed and sends the key scan code value over the keyboard serial link keyboard controller on the motherboard receives the kscan scan byte. Then it translates that kscancode into the system scan code and places the value in its output buffer. The motherboard controller then issues an interrupt request indication that data is available. The interrupt request calls the interrupt 9 handler, the keyboard BIOS. The keyboard BIOS reads the scan code from the motherboard controller and translates the scancode into an ASCII byte. The keyboard handler puts both the scan code and the ASCII byte into the next available spot in the 16 byte FIFO keyboard buffer. Lastly the keyboard interrupt is cleared, and the keyboard bios exists, returning control to the task running at the time of the keyboard interrupt.

The OS or application program uses 16h, the intermediate keyboard BIOS services, to access the keyboard buffer. Interrupt 16h functions are used to find if a key is available and to determine the value of a key.

The above working is to be followed while creating our own function routine. This function routine is designed so that all the Tamil characters are accessed in the place of English keys.

When a key is pressed, each key generates its own scan code, using that scancode set the English or Tamil mode for any combination of keys. System should check whether it is in Tamil mode or English mode. If it is in English mode it returns to the original ROM BIOS routine. If it is in Tamil mode, it checks which key is pressed, then traps that key,and passes the character into the place of the original English character, in the keyboard buffer. Then display routine is called to display that particular character on the screen.

Conclusion

The keyboard driver, developed in DOS environment was tested and found very effective. The fonts generated in DOS platform are not that good looking for want of space. However, after the data in Tamil is input into the system, it is viewed through windows through an interface for good shape of the Tamil fonts.

214 empty

The enabling technology for Tamil

N. Anbarasan
APPLESOFT, Bangalore - 560 010, India
<e-mail : aplesoft@vsnl.com>

Abstract

Innumerable softwares are available off-the shelf to meet the various requirements. These softwares are ranging from word processing to authoring. There are attempts to develop equivalent softwares for these softwares to meet the vernacular requirements. But, they are not successful and are not able to cope up with developments of technologies. In to-day's technological enhancement, the method of having Tamil on computers have resulted in reality for Tamil computing.

Introduction

Computers are handy tools for automation with appropriate programming. The computerisation process is to automate the mannual process to enable quality information management and benefit the user. Language implementation has the same objective.

Even though PCs were introduced during 1974, they became popular only when good softwares like Word star, Lotus 1-2-3 etc were made available. Unless good softwares to meet today's requirement are made available, there cannot be any quantum jump in the use of Language software. Good softwares not only enhances the usage of the computers but also increases their penetration. It is unfortunate that no good software is available to cater to the needs of the vernacular users. As far as administration is concerned, there are hundreds of general purpose softwares available off-the shelf. These softwares range from simple word processing to complex database management through fascinating Desk Top Publishing. These softwares are constantly revised and upgraded to keep pace with the rapidly advancing technologies in hardware and operating systems. With the ever increasing demand, the softwares are becoming more sophisticated.

Today, the world is witnessing a shift in the usage of computers with umultilingual capabilities. Thus the softwares and web contents being developed are multilingual in nature.

Computerisation

The computerisation in the home land of Tamil community (Tamilnadu) is vastly and variedly implemented, The computerisation caters to various requirements at various levels of the government departments ranging from land records to destitute pensionery schemes, Government schools and at various service sectors with an ever green hopes of achieving IT

revolution for anything and everything. This aims to guarantee to the majority population, the right to access to information thereby ensuring transparency in the governance.

Requirement of software

Well. Just switching over to computerisation and having lust for creating computer awareness will not result in proper implementation. To see and enjoy, the computerisation needs to be through the Language of the people. Hence, the existing hardwares need a good software for Tamil and in Tamil to achieve the goal of IT revolution and ensure proper implementation.

In the wake of computerisation process, almost every computer needs a software - Language specific for transacting official business in Tamil. As is seen, the MS Office Suite is widely used to derive the results in English. Apart from Office suite, based on the requirement, a Department may require a tailor-made software to meet its own requirment. With the majority of staff members having insufficient working knowledge of English and conveniently placed with Tamil and the local population have no or little knowledge of English as also coupled with the policy on usage of Tamil for administrative convenience, an inevitable requirement arises for an "Enabling" Tamil software to offer whatever fascinating features offered by MS Office suite.

Types of Tamil softwares

To meet the requirement of our Language for various requirements (applications) two methods are being followed.

1. To write new software to meet the requirement of our Language. So far most of the softwares cater to the needs of word-processing. There are certain scattered efforts to develop database applications, spread sheet application, programming Language etc. None of these were successful except word processing with minimal features. The few wordprocessors, which are available in the market are

    Bharathi
    Surabhi Inscript Processor
    Valluvan
    Kamban
    Padhami
    Leap

2. There is enormous range of software, available for English. For any user, the natural choice could be to use the same English software for vernacular usage as well. As the user is already familiar with the English software and its operational details, it is convenient to have the same, rather than re-learning a new set of command. Based on this approach, certain software is being developed and it is such software, which is successfully used for the obvious reasons. This type of development could be classified as "Interface software". The Interface softwares available in the market are:

ISM
Inscrypt
Ilango

The interface software merely allows inputting of Language text into the application softwares on popular Operating Systems. The formal approach could be to enable the input and display of Language text at all levels. Such software can be conveniently termed as "Enabling" software. SURABHI 2000 and Kanian 2000 are based on this approach. These are feature packed softwares and setting new benchmarks amongst vernacular softwares and set a new standard. The ultimate aim of the "Enabling" software is to obtain the maximum out of the English softwares.

The off-the shelf software is developed for English having simple script, where the letters are placed only side by side. The software developed for English. cannot be used for Indian languages, which have complex scripts. The software is developed with features specific to English- such as "Find" and "Replace", "Spell check", "Dictionery", "Autocorrect", "Mail merge" etc. The softwares available in the market to meet the requirment of vernacular demand provides no support to these features either directly or indirectly.

Required Features of Tamil software

The desirable features of a Tamil software, which could enhance the usage of Tamil on computers could be:

Co-exists with other Windows based applications
Configurable options
Find and Replace in Tamil
Find - Files or Folder
Intelligent keyboard manager
Choice of keyboard layouts
Shortcut in Tamil
Multitasking
Sorting
Spellcheck
Dictionery
User interfaces in Tamil

Tamil software development fund

As an outcome of Tamilnet 99 and in the series of ever green forethoughts of the Government of Tamilnadu, a "Tamil software development fund" has been created and a meagre amount is being offered to the Tamil software developers for development of softwares with specific requirement. Though, the offer in no way affords the developers financially viable for such development, it enthuses their interest to do their mettle for the enthronement of Tamil on computers.

Localisation of Windows 95 and 98

The Tamil software development fund enables the developers to aim for new technogies. A few developers have already been able to get some funds from this caretaking fund. APPLESOFT, is one amongst such Companies to get the funds. APPLESOFT undertakes the project for localisation of Windows 95 and 98 and the aim of the project is to take Tamil at OS (operating system) level with user interfaces in Tamil. The user interaces convered under this project is as under :-

Window title
Standard Menu
Standard pulldown menus
Standard common dialog boxes such as Open, save, Print etc
Start menu
Start menu sub menu
Desktop icons
Status bar text
Control panel
Tooltip texts
Status bar texts
Rebar menus, buttons

Conclusion

What-ever fascinating and fabulous softwares are available for English, they are not able to provide the Government or the user to have the same features for Tamil. The market statistics claim to have earned turnouts running into thousands of crores of ruppees from exporting softwares by a few thousand Companies engaged on development of such softwares. However, the achievements out of Language softwares, are obviously negligible or not noticeable. An intersting and contrasting truth to note here, is that only an handful of companies with lust for Language and guts to survive, are engaged in research, development and marketing the beloved Language softwares. Only their efforts are making Tamil to servive in the digital world. While the software developers other than the vernacular specific, are making giant leaps of their track, these exalted with Language specific softwares are becoming underprivileged and uneconomic.

It is unfortunate that, the Government often compare these small time Tamil Software developers are compared on par with application software developers. The Language being the carrier of Tamil culture and its recognised identity, the concerned Govt are yet to recognise the role of the Tamil software developers.

Unless the Govts comes with a policy to support these Tamil software developers at least on par with the professionals engaged in culture, arts such as cinema, folk arts etc, Tamil cannot survive in the digital revolution.

Compilation Of Electronic Dictionary For Tamil

Dr. M. Ganesan
Centre of Advanced Study in Linguistics, Annamalai University
Annamalainagar - 608002, Tamilnadu, India

Introduction

In the computer era language development and technology development are having impact on each other. There is a need to develop a language interms of grammar and lexical studies in such a way that it suit the modern technology. Similarly technology has to be developed to cope with the intricacies of languages such as scripts, writing system, etc. The long term goals of NLP (Natural Language Processing) research to develop.

i. Machine Aided Translation (MAT) systems for various natural languages.
ii. Systems for man-machine communication through natural languages.
iii. Text-to-speech and speech-to-text systems, and
iv. Computer Aided learning/Teaching (CALT) materials.

These goals can be achieved in stages through several subsystems which comprise of linguistic tools / information at the background and software tools at the foreground. The linguistic tools for the use of machine can be either in the form of rules (mostly grammatical information) or in the form of databases (mostly lexical information). Grammar which describes the structure of a language is mainly written for human beings, especially for language experts. Such grammars as such may not be adequate for a machine to understand the language as it does not have any common sense and other world knowledge which are necessary for the proper interpretation of the grammar. Similarly conventional dictionaries and lexicons prepared for human users provide authentic reference to meanings and grammatical information. Those information are also limited mainly because of the constraint of space. Addition of more information would make it voluminous in size and that would be inconvenient for users to handle it. Thus, there are different types of specialized dictionaries like historical, etymological, professional (law, medicine, etc.) pedagogical, etc., depending upon the requirement of the variety of users. All the information available in those dictionaries are grossly inadequate for the use of machines. It is, therefore, necessary to prepare computational grammar and lexicons for natural languages in such a way that they can be used by machines and also that the benefits of technology can be made available to the human users to acquire more information with less effort and cost. In this direction, this paper describes the limitation of information available in the printed dictionaries, advantages of Electronic Dictionary (ED) over a printed dictionary, designing and compilation of an ED, uses of computer corpora to the lexicographers, various software tools needed for corpus analysis, etc.

Limitation of Information in Printed Dictionary

Dictionary is a tool mainly used to acquire lexical knowledge, and to some extent, grammatical information of a language. For a lexeme the type of information normally available in a dictionary are parts of speeches, pronunciation, meanings, citations, and special uses, etc. Sometimes etymology, synonyms and antonyms, register, etc., are also provided in some dictionaries. For the most of the Indian languages such a wide variety of dictionaries are not available. It may be mostly because of the limited users for the Indian language dictionaries, when comparing to English dictionary. If one analyses the reasons for not using the dictionary for Indian languages, he may attribute that the type of information available in the dictionary are limited and not meeting the requirement of the users. For example, a learner of Tamil wants to know the meaning for the word Vanta:n. The word as such is not attested as an entry in any Tamil dictionary. To get the meaning of the word the learner has to know that the root of the word is va:. So a considerable amount of knowledge on Tamil morphology is necessary from the learner side to find the meaning. Otherwise dictionary should have all the inflected and derived forms as a separate entry, which is practically not possible, because a verb in Tamil can be conjugated to around 1600 forms (which include particles, post positions, etc. suffixed to a verb). Further in the print medium the size of the dictionary will be unmanageably voluminous. Secondly, if one wants to check the spelling of an inflected word like collikkoLLa, the dictionaries are of no use to him. Such limitations of information are basically due to the structural constitution of a language. Languages like Tamil are highly agglutinative by nature and there is, therefore, a need to overcome the limitations with the help of technology.

Electronic Dictionary

Computers, as we know, have a lot of storage capacity and computation capability. The features can be made use of to overcome the limitations of space and information in a printed dictionary. Electronic Dictionary, in general, means that having dictionary information in electronic medium. But on the basis of the purpose for which it is used, and the type of infomation incorporated in it, it can be classified into different types. Dictionaries for human use, Dictionaries for on-line references to both human and machine, dictionaries with more grammatical information for language processing by machine, dictionaries / lexicon for MT (Machine Translation) systems, etc., are some of the different types of electronic dictionaries. An ED must aim to provide more lexical and grammatical information, instead of reproducing the printed one in the electronic medium.

Advantages of Electronic Dictionary

The medium itself is the greatest advantage. In print whatever information stored could only be retrieved / referred to in the same order. Whereas in computer medium the information stored can be processed using programs so that the exact information which are required can be retrieved easily. Besides this, the followings are some of the order major advantages of E.D.

i. Provides more grammatical information like sub-categorization, collocation, selectional restriction, etc., than the one available in print medium.

ii. Various types of specialized dictionaries (professional, pedagogical, etc.) can be extracted from an ED.
iii. allows to extract lists of nouns, verbs, etc.
iv. can provide paradigms for nouns and verbs.
v. gives pronunciation through voice.
vi. displays animated pictures.
vii. is available in machine readable from so that any modification or updation can be done easily.
viii. readily available for on-line references to both human users and machine.
ix. machine can make use of the information selectively from the dictionary for different applications like Machine Translation, language processing, CALT, speech recognition, etc.
x. a bi/multilingual dictionary can be compiled from a monolingual ED and vice-versa, and
xi. if properly designed, ED can be reversible one. i,e. a Tamil- English bilingual dictionary can be used as an English - Tamil dictionary.

A learner who wants to get the meanings of a word which is in inflected or derived form can give the word as such, the ED, using a morphological analyser finds out the root form and displays the meanings. Even if one is interested to see all the inflected forms of the word, they can be generated and listed with grammatical labeling. It also helps to find out the spelling of an inflected form which is not possible in other means.

Compilation of Electronic Dictionary

The discipline of lexicography, atleast in the Western countries, has changed almost beyond recognition. In dictionary- making , whether it is for print or computer, the technology is maximum utilised. Lexicography involves both mental and mechanical works almost equally. The entire mechanical works can be easily carried out by computers using suitable programs. The machine can also provide various processed information which actually helps the lexicographers to accomplish the most of the mental tasks with ease. Computers can be involved in all the four stages of dictionary- making.

1) data-collection,
2) entry-selection,
3) entry construction and
4) entry arrangement.

In the case of compilation of an ED one has to decide a number of factors such as the type and quantum of information to be provided in the ED, the structure of databases, the method of retrieval of information, etc, will be advance.

An ED can be designed with three major sub-systems, viz.

1. system for data collection,
2. system for data storage and

3. system for information retrieval

At the time of developing these systems, the features of computers such as colour, graphics, animation, voice, memory, speed, etc., the information requirement of different users, presentations of basic information and rarely retrieved information, etc., should be kept in mind.

Language corpora and its use in Dictionary making

"Corpora are essentially, bodies of natural language materials (whole texts, samples from texts or sometimes just unconnected sentences) which are stored in machine readable form" (Leech, 1992: 115).Basically, corpora provide authentic data of contemporary use of languages. The major advantages of corpora are that any specific information can be retrieved selectively and through computer programs data can be manipulated for various purposes, as they are stored in an organized way and are in machine readable form. The use of computerized corpus data on a massive scale helps lexicographic in a number of ways :

1) to select the head word
2) to give authentic real-life material as examples
3) helps lexicographer to decide on sense distinction
4) to provide grammatical information
5) to give the statistical information like frequency of occurrence of a word in the corpus, etc.,
6) to provide information about the sub-categorization, collocation and selectional restriction of a lexical item.

A number of dictionaries (some are entirely in new types) have been published in English using large corpus data. In the case of Tamil, computer corpora to a size of 3.5 million words have been created by the Central Institute of Indian Languages (CIIL), Mysore. It is a primary corpus; data are collected from the books, journals, News papers, Government documents, etc. published during the year 1981 to 1990 to represent the language use of contemporary Tamil. They are classified into 6 major categories and 76 sub-categories. The CIIL has also designed a trilingual (Tamil-Hindi-English) electronic dictionary with various features discussed in this paper.

Tools for lexicographers

Corpora can be viewed as large sources of information comprising of textual narratives and can be augmented with additional information like labeling for grammatical categories at different levels. The primary motive for arranging corpora in machine readable form is to introduce an element of automation, which cannot be realized unless an efficient retrieval system is available. The software tools for lexicographers in general and for electronic dictionary in particular are listed below:

1) Corpus Manager : It is a software which allows to organize corpus in a classified order. Any corpus text can be inserted or deleted from language corpora. It also allows to retrieve any text selectively from the corpus.
2) Word Tagger:Grammatical information in addition to the text can be provided in the corpus data. They can be labeled at end of an item (morpheme, word, phrase, etc.) so that based on the tagged information data can be retrieved . For example, if a corpus is tagged for parts - of - speeches, one can easily retrieve all the verbs used in the corpus.
3) Frequency count : It is an effective tool to get different statistical information like the frequency of occurrence of an item (say a word or a phrase) in a given text.
4) KWIC (Key Word In Context) Concordance : This tool scans the corpus for occurrence of specific word (even morpheme, phrase, sentence, etc.) and present all those occurrences with linguistic content both all left and right of the word. It is very useful to suggest different shades of meaning, the collocational behaviour and selectoral restriction of a word. Citations, suitable to the senses of the word can be extracted from the corpus instead of creating intuitive sentences.
5) Morphological Analyser: It analyses a word for its morphemic components. In an ED, it is useful to find out the root of a word form in order to search the head word.
6) Paradigm Genarator : A software which can generate the entire inflected and derived forms of lexical items. For languages like Tamil when incorporated in the ED, learners can get any specific inflected form and can also use for verifying the spelling of a word.

Most of the above mentioned tools have been developed in CIIL for Tamil.

Conclusion

The technology should be fully exploited for the development of Tamil. Electronic dictionary as mentioned in this paper, must be different from the printed version of dictionaries by incorporating additional features and by providing more information. Creating an electronic dictionary for Tamil with all these features will be an asset to Tamil language and its community.

References:

1. Ekka. Francis, BD Jayaram and M. Ganesan., Final Report : Development of corpora of Text of Indian languages in machine readable form, Part II (Tamil, Telugu, Kannada, Malayalam) Mysore: CIIL, 1995
2. M. Ganesan. "A Scheme for Grammatical Tagging of Corpora in Indian Languages" in Technology and Languages, (Ed) B.B Rajapurohit, Mysore : CIIL, 1994.
3. Leech Geoffery, "Corpora Annotation Schemes" in Literary and Linguistic Computing, Vol. 8 No 4, 1993.
4. Meijs, Williem, "Linguistic Corpora and Lexicography" in Annual Review of Applied Linguistics, Vol 16, 1996.