சங்க இலக்கியங்களுக்கான மொழித் தொழில் நுட்பக் கருவிகள்

முனைவர். கா. உமாராஜ்
உதவிப் பேராசிரியர்
மொழியியல் துறை
மதுரை காமராஜர் பல்கலைக்கழகம்
www.umarajk.in
umarajk@gmail.com

மொழித்தொழில் நுட்பம் (Language Technology)

கணினியியல் துறையும் மொழியியல் துறையும் இணைந்த ஒரு துறையே கணினி மொழியியல் துறையாகும். விதிகளின் அடிப்படையில் அல்லது புள்ளியியல் அடிப்படையில் இயற்கை மொழியை அல்லது இயற்கை மொழிக்கூறுகளை செயற்கையாக கணினிக்கு தகுந்தவாறு மாற்றி தருவதற்கான கோட்பாடுகளையும் செயல்பாடுகளையும் கொடுப்பதே கணினி மொழியியல் துறையின் முக்கிய பங்காகும்.

மனித மூளையைப் போன்றே கணினிக்கு இயற்கை மொழி அறிவைப் பெறவைத்து ஒரு மொழியிலுள்ள மொழிக்கூறுகளை எளிமையாகப் புரிந்து கொள்ளவும் அம்மொழிக் கூறுகளை செயற்கையாக உருவாக்கவும் செய்ய வைக்கும் முயற்சியே இயற்கை மொழியாய்வு (Natural Language Processing) ஆகும்.

கணினி மொழியியல் (Computational Linguistics), இயற்கை மொழி ஆய்வு (Natural Language Processing-NLP), மொழிப் பொறியியல் (Language Engineering), கணினித் தொழில்நுட்பம் (Computer Technology), மொழியும் உளவியலும் (Language and Psychology), தரவக மொழியியல் (Corpus Linguistics) போன்ற பல துறைகள் உருவாக்கிய கணினி மொழியியல் கோட்பாட்டின் அடிப்படையில் உருவாக்கப்படுகிற தொழில்நுட்பமே மொழித்தொழில்நுட்பமாகும் . ஆகவே மொழித் தொழில்நுட்பம் என்பது பல துறைகளை ஒருங்கிணைத்துக் கூட்டுத் திட்டமாகத் திகழ்கின்றது.

மொழித் தொழில் நுட்பம் இரண்டு வகைப்படும் 1) எழுத்து மொழித் தொழில்நுட்பம் 2) பேச்சு மொழித்தொழில்நுட்பம் . சங்க இலக்கியங்களுக்கு இரண்டு தொழில்நுட்ப ங்களுமே தேவைப்படுகிறது.

சங்க இலக்கியங்களுக்கான மொழித் தொழில் நுட்பம்

சங்க இலக்கியங்களுக்கான மொழித் தொழி நுட்பம் இரண்டு கட்டங்களாக செயல்படுத்தபடுகிறது.

முதல் கட்டமாக இ ணையதளம், வலைப்பூக்கள், விக்கிபீடியா போன்றவற்றில் சங்கத்தமிழ் பனுவல்களை கல்வெட்டுகளை ,ஓலைச்சுவடிகளை இடம்பெறச்செய்வதாகும்.

இரண்டாவது கட்ட மாக கணினிக்குச் சங்கத்தமிழைக் க ற்றுக்கொடுத்து பல்வேறு மொழித் தொழில்நுட்பத் தேவைகளை நிறைவேற்றிக்கொள்வதாகும் . அதற்கு சங்கத் தமிழ் இலக்கணத்தைக் கணினிக்கேற்ற கணிதவாய்பாடுகளாக மாற்றிக் கொடுக்க வேண்டும் . மனித மூளைக்காக உருவாக்கப்பட்டுள்ள தொல்காப்பியம் , நன்னூல் போன்ற இலக்கண நூல்களைக் கணினிக்கேற்ற கணிதவழி இலக்கணமாக மாற்றிக், கணினிக்கு அளிக்கவேண்டும் . மேலும் உலகத்தைப்பற்றிய அறிவும் மிகவும் தேவை . அதன் பயனாக , சொல்லிலோ அல்லது தொடரிலோ பொருள் மயக்கம் ஏற்பட்டால், பேசப்படுகிற பொருள் , பேசும் சூழல் போன்றவை நமக்கு உதவி செய்து , பொருள் மயக்கத்தைத் தவிர்க்கின்றன . கணினிக்கு இதுபோன்ற பின்புல அறிவு இன்று அமையவில்லை. எனவே, அதனால் கணினிக்கு பொருள் மயக்கத்தைத் தவிர்ப்பதற்கு பலவகைகளில் நாம் உதவவேண்டும்.மேலும் சங்கத்தமிழ் சொற்கள் தொடர்கள் சந்தியுடன் காணப்படும் . அவற்றை பிரிப்பதற்கான விதிகளை உருவாக்க வேண்டும்

சங்க இலக்கியத்திற்கு இதுவரை உருவாக்கபட்டுள்ள மொழித்தொழில் நுட்பக் கருவிகள்

சங்க தமிழுக்கான தரவகம் மற்றும் சொல்லடைவு கருவி

தரவகம் என்பது மின்னுரைகள் அல்லது மின்நூல்களின் தொகுப்பு எனலாம். “ஒழுங்குமுறையுடன் அதிக அளவில் தேர்வு செய்யப்பட்டு கணினியில் சேமிக்கப்பட்ட இயற்கையான நடைகளை உடைய உரைகள்”. தரவகம் இரு பொரும் பிரிவுகளாகப் பிரிக்கப்படுகிறது 1. தரவகம் 2. குறியீட்டுத் தரவகம் ஆகியன.

தரவகத்தைக் கொண்டு இயற்கை மொழி ஆய்வை ஒரு குறிப்பிட்ட எல்லை வரையே செய்ய இயலும். இந்த இயற்க்கை மொழி ஆய்வு தன்னிறவை அடைய குறியீட்டுத் தரவகத்தின் பங்கு இன்றியமையாதது.
தமிழ் இணையக் கல்விக்கழகம் சங்க இலக்கியம் முதல் தற்கால இலக்கியம் வரையிலான தரவகத்தை உருவாக்கி வருகிறது. http://www.tamilvu.org/

மொழித் தொழில்நுட்பத்தைப் பயன்படுத்தி சங்க இலக்கியத் தரவுகள் செம்மொழித் தமிழாய்வு மத்திய நிறுவனத்தின் மொழித் தொழில்நுட்பத்துறையால் உருவாக்கப்பட்டுவருகிறது (http://www.cict.in).

மேலும் பின்வரும் தளங்களில் சங்க இலக்கியத்திற்கான தரவுகள் கிடைக்கும்

www.projectmadurai.org

www.ta.wikipedia.org

www.tamilvu.org

www.ldcil.org

www.noolagam.org

http://sangamtranslationsbyvaidehi.com

www.karkanirka.org

www.learnsangamtamil.com

www.puram400.blogspot.in

www.ilakkiyam.com

www.sangampoemsinenglish.wordpress.com

www.sangamtamilliterature.wordpress.com

www.classicaltamil.org

சங்க இலக்கிய சொல்லடைவு கருவி

சொல்ல்டைவு கருவியின் மூலம் ஒரு சொல் சங்க இலக்கியங்களில் எந்த எந்த பாடல் வரி எண்ணில் வருகிறது உள்ளது என்பதை அறிந்துக் கொள்ள முடியும். மேலும் அச்சொல் எத்தனை முறை சங்க இலக்கியங்களில் பயின்று வந்துள்ளது என்பதையும் அறியலாம். செம்**மொழித் தமிழாய்வு மத்திய நிறுவனம், தமிழ் இணையக்கல்விக் கழகம் போன்ற நிறுவனங்களும் பண்டியராசன், காமாட்சி போன்ற அறிஞர்களும் சொல்லடைவு கருவியை உருவாக்கியுள்ளர்கள்.**

சங்க இலக்கிய தொடரடைவுக் கருவி

தொடரடைவு என்பது 'ஒரு புத்தகத்தில் அல்லது பதிப்பில் " பயன்படுத்தப்பட்டுள்ள சொற்களின் அகரவரிசைப்படுத்தப்பட்ட பட்டியலாகும் . இப்பட்டியலில் சொற்களுடன்

அவை இடம்பெறும் தொடர்களும் குறிப்பிடப்பட்டிருக்கும் . அதாவது ஒரு புத்தகத்தில் அல்லது பதிப்பிலுள்ள சொற்கள் இடம்பெறும் தொடர்கள் அனைத்தையும் பெறுதல் தொடரடைவு எனக் கூறலாம் . தொடரடைவு என்பதை விக்கிபீடியா பின்வருமாறு பொருள்கொள்கிறது.A Concordance is an alphabetical list of Principal words used in a book or body of work with their immediate contexts. ஆக்ஸ்போர்டு அகராதி தொடரடைவு என்பதைப் பின்வருமாறு பொருள் கொள்கிறது.

An alphabetical list of words (esp. the important ones) present in a text, usually with citations of the passage concerned (...) Origin from Latin concordara 'agree on'

தொடரின் பொருளை முழுமையாக அறிந்தால் மட்டுமே தொடர்களைத் தெளிவான முறையில் பிரிக்கமுடியும்.

சங்க இலக்கிய இணைய அகராதி

இணைய அகராதி அல்லது மின் அகராதி என்பது பொதுவாக உருவாக்கப்பட்ட ஓர் அகராதியை இணையத்துடன் இணைத்துக் கணினியின் மூலம் தரவுகளைப் பெறுவதாகும்.

இணைய அகராதி என்பது ஆங்கிலத்தில் (Online Dictionary) என்று பொருளாகக் கொள்ளலாம். தமிழாய்வுக்குத் தேவையான கருவி நூற்களான அகராதிகள், பேரகராதிகள், சொற்களஞ்சியங்கள், கலைக்களஞ்சியங்கள் கணினி உதவி கொண்டு உருவாக்குவதில் பல்வேறு முயற்சிகள் மேற்கொள்ளப்பட்டு வெற்றி பெற்றுள்ளன. அவற்றுள் தமிழ் – தமிழ் – ஆங்கில அகராதியாகியசென்னைப் பல்கலைக்கழகத்தின் பேரகராதி (Lexicon), பழனியப்பா பிரதர்ஸ் நிறுவனத்தின் ஆங்கில – தமிழ் அகராதி, பேரா. சண்முகம் பிள்ளை அவர்களின் தமிழ் – தமிழ் அகராதி ஆகியவை குறிப்பிடத்தக்க அகராதிகளாகும். சங்க இலக்கியத்திற்கான இணைய அகராதியை தமிழ் இணையக்கல்விக்கழகம் உருவாக்கியுள்ளது .

கணினி வழி சங்க இலக்கியக்கல்வி

கணினியில் மொழித்தொழில்நுட்பம் பெற்றுள்ள வளர்ச்சியின் பயனாக இன்று மாணவர்களுக்குத் சங்கத்தமிழைக் கற்பித்தலில் கணினிக்குக் குறிப்பிடத்தக்க பங்கு உள்ளது . சங்கத்தமிழ் இலக்கணத்தையும் சொற்களஞ்சியத்தையும் மாணவர்களுக்குக் கற்றுக்கொடுப்பதில் ஆசிரியருக்குத் துணையாகக் கணினி செயல்படமுடியும். மொழி

கற்றலில் இடம்பெறும் பல்வேறு செயல்பாடுகளில் கணினியானது மாணவர்களோடு ஊடாட்டம் (Interaction) புரிந்து, அவர்களுக்கு உதவமுடியும்.

இரண்டுவகைகளில் கணினியைப் பயன்படுத்தலாம் . ஒன்று, தமிழ் இலக்கணத்தை நேரடியாகக் கற்றுக்கொடுக்கவும் பயன்படுத்தலாம் .மற்றொன்று, அவ்வாறு ஆசிரியர் மாணவருக்குக் கற்றுக்கொடுத்தபிறகு, மாணவரின் மொழிப் பயன்படுத்தத்தை வளர்ப்பதற்காகவு ம் பயன்படுத்தலாம்.செம்மொழித் தமிழாய்வு நிறுவனம் , தமிழ் இணையக்கல்விக்கழகம் போன்ற நிறுவனங்கள் இப்பணியை செய்து வருகின்றனர்.

சங்க இலக்கியங்களை மொழித் தொழில் நுட்ப அடிப்படையில் ஆய்வு செய்யும்போது ஏற்படக்கூடிய சிக்கல்கள்.

சங்க இலக்கியத்தில் எது சொல் , எது சொல் ஆகாது என்பதை வரையறுத்துக் கூறுவது மொழிச் சிக்கல் நிறைந்த பகுதியாகும் . சில அறிஞர்கள் தனிச்சொல்லைக் கூட்டுச் சொல்லாகவும் கூட்டுச் சொல்லைத் தனிச் சொல்லாகவும் பிரித்துள்ளனர் . வேற்றுமை உருபுகள், எண்ணும்மைகள், சாரியைகள், அளபெடுத்து வரும் சொற்கள் , அன்மொழித் தொகை, கூட்டு வினைச் சொற்கள் , பகுதிப் பொருள் விகுதி ஆகியவற்றைப் பிரித்துக்கொடுக்க வேண்டுமா ? கொல்யானை, விரிகதிர் போன்ற வினைத் தொகைகளையும், செம்மொழி, செங்கதிர் போன்ற பண்புத் தொகைகளையும் , அவ்விடம், இவ்விடம் போன்ற சுட்டுப்பெயர்களையும் , சால, உறு போன்ற உரிச்சொற்க ளையும் பிரித்துத் தர வேண்டுமா? என்பன போன்ற சிக்கல்கள் எழுகின்றன.

தமிழகத்தில் மொழியியல் ஒரு தனித் துறையாகத் தோன்றி வளர்ந்த காலத்தில் குறுந்தொகை இலக்கியங்களுக்கான சொல்லடைவுகள் உருவாக்கப்பட்டன . குறுந்தொகைக்குக் (1974) கலியபெருமாளும், ஐங்குறுநூற்றுக்குக் கிரு ட்டிணம்மாளும் சொல்லடைவுகளை உருவாக்கியுள்ளனர்.

இப்பதிப்புகளில் சொற்பிரிப்பு நெறிமுறைகளைப் பற்றிய விவரங்கள் அதிகம் காணப்படவில்லை. ஆனால் சொற்களிலுள்ள வேற்றுமை உருபுகள் , சாரியைகள், இலக்கணப்பொருள் தரும் உருபுகள் போன்றவை பிரித்துத் தரப்பட்டுள்ளன.

1993 இல் ஆசி யவியல் நிறுவனம் சங்க இலக்கியச் சொற்றொகை (A word index for Sankam Literature) என்னும் நூலை வெளியிட்டது . இது ஜெர்மானிய அறிஞர்கள் தாமஸ்லேமென், தாமஸ் மால்டன் ஆகிய இருவரால் உருவாக்கப்பட்டதாகும் . அதற்குப் பின் 2003-இல் சங்க இலக்கியச் சொல்லடைவைத் தஞ்சை தமிழ்ப் பல்கலைக்கழகம் வெளியிட்டது. இதன் பதிப்பாசிரியர் பேரா. பெ.மாதையன் ஆவார் . இப்பதிப்பும் ஆசியவியல் நிறுவனம் உருவாக்கிய சங்க இலக்கியச் சொற்தொகையும் பெரும்பான்மையும் ஒத்து அமைந்துள்ளன . அதற்கு இரண்டு பதிப்புகளும் கொண்டிருந்த சொற்பிரிப்பு நெறிமுறைகளே காரணமாகும்.

பின்வரும் உருபுகளைப் பிரித்துத் தர வேண்டுமென இரண்டு பதிப்புகளிலுமே குறிப்பிடப்பட்டுள்ளது. (1) வேற்றுமை உருபுகள் (2) சாரியைகள் (3) பின்ஒட்டுக்கள் (4) சுட்டெழுத்துக்கள் (5) பெயரெச்சம் (6) இயல்பு சொற்சேர்க்கை (7) இடைச்சொல். மேலும் பின்வருவனவற்றைத் தனிச்சொல்லாகத்தான் கருத வேண்டுமென இப்பதிப்புகளில் குறிப்பிடப்பட்டுள்ளது . (1) கூட்டு வினையெச்சங்கள் (2) புதுப்பொருள் தரும் சொற்கள் (3) காரணப்பெயர்கள் (4) பகுதிநிலைக் கூட்டுச்சொற்கள்.

சங்க இலக்கியத்திலுள்ள சொல்லடைவை நோ க்குகையில் குறுந்தொகை இலக்கியச் சொற்தொகையில் சில சொற்களின் விடுபாடுகளும் குறைபாடுகளும் இருப்பதைக் காணமுடிகிறது.

சில சொற்களோடு வேற்றுமை உருபுகளைச் சேர்த்தும் குறுந்தொகை இலக்கியச் சொற்தொகை உருவாக்கப்பட்டுள்ளது. சான்றாக

“அகலத்தவனை” என்பதைச் சுட்டலாம் . மேலும் சங்க இலக்கியச் சொற்தொகையில் பின்வரும் தனிச்சொற்கள் பிரித்துக் கொடுக்கப்பட்டுள்ளன . சான்றாக: சேவடி கடுமான் , போன்றவற்றைக் குறிப்பிடலாம் . ஆகவே இதுபோன்ற குறைபாடுகளையும் நீக்கிச் சங்க இலக்கியத்திற்கென மின் அகராதி ,தொடரடைவு உருவாக்கும்போது மொழியியல் அடிப்படையில் மேம்பட்ட சொற்பிரிப்பு முறையைப் பயன்படுத்த வேண்டும்.

மேலும் சங்க இலக்கிய தொடரடைவு உருவாக்கும்போது வேற்றுமை உருபுகள் , எண்ணும்மைகள், சாரியைகள், மூவாயிரம் போன்ற எண்ணுப்பெயர்கள் , அளபெடுத்து வரும் சொற்கள், அன்மொழித் தொகை, கூட்டுவினை யெச்சச் சொற்கள், பகுதிப் பொருள் விகுதி ஆகியவற்றைப் பிரித்துக் கொடுக்க வேண்டும்.

ஆங்கில நிகரன்களைத் தேடுவதிலும் பல சிக்கல்கள் உள்ளன . ஒரு சொல்லுக்கு இருவேறு நிகரன்களை கொடுக்க முடியாது. எந்த நிகரன் சரியானவை என்பதை முடிவு செய்ய வேண்டும்.

ஆங்கில நிகரன்களைத் தேடுவதிலும் பல சிக்கல்கள் உள்ளன . 'யாருமில்லை தானே கள்வன்' என்ற தொடரில் கள்வன் என்பதை பன்னீர்செல்வம் 'cheat' என்றும் சண்முகம் பிள்ளை "like a thief" என்றும் பிறமொழிபெயர்ப்பாளர்கள் "thief" என்றும் மொழிபெயர்த்துள்ளனர். ஒரு சொல்லுக்கு இருவேறு நிகரன்களை கொடுக்க முடியாது. எந்த நிகரன் சரியானவை என்பதை முடிவு செய்ய வேண்டும்.

எ.டு

வானின் அகலம் — vaster than the sky / higher than the sea

கடலின் ஆழம் — deeper than the sea / more unfathomable than the water

குறிஞ்சி மரத்தின் மலர்களைக் கொண்டு தேனடை செய்யும் நாடு — hills where the/ honey bees make abundant honey from the black-shammed Kurinji / mountain slopes where bees make rich honey from the flower of the Kurinji that has such

black stalks

பொய்கையிலுள்ள வாளை மீன்கள fresh water sharks in the pools

; valai fish in the wetfield

like a puppet

பாவைபோல

like a doll

lifting his hands and legs

கையும் காலும் தூக்க when others raise their hands and feet he will raise his too

Your mother and my mother do not know each other your father and my father are not you and me

ஞாயும் யாயும் யாரா கியரோ,

எந்தையும் நுந்தையும் எம்முறைக் கேளிர் what would my mother be to yours? what kin in my father to your anyway?

like the mingling of rain water with red soil,

our hearts have mingled

செம்புலப் பெயர் போல

அன்புடை நெஞ்சம் தாங்கலந்தனவே In love our hearts are as red earth pouring rain, mingled beyond parting

உவமைகளுக்குக்கான நிகரன்களைக் கொடுப்பதில் ஏற்படும் சிக்கல்கள்

பின்வரும் பாடல் வரிகளில் சொற்களுக்கான ஆங்கில நிகரன்களை கொடுப்பதில் சிக்கல் இருப்பதை அறிய முடிகிறது.

எ.டு

சிறுமுகை அவிழ்ந்த நறுமலர்; (குறு:220)

ஆம்பல் நாறும் தேம் பொதி துவரவாய் (குறு:300)

your red mouth filled with sweetness is

redolent of the while water lily

பெயல்கண் மறைத்தலின் விசும்புகா ணலையே (குறு:355)

the rain hide everything

and you cannot see the sky

முடிவுரை

சங்க இலக்கியத்திற்கு மொழித் தொழில் நுட்ப அடிப்படையில் இன்னும் பல்வேறு கருவிகள் உருவாக்க வேண்டியுள்ளது. குறிப்பாக இயந்திர மொழிபெயர்ப்பு , எழுத்து பேச்சு மாற்றி , தொடரனியல் குறிப் பான் மற்றும் பகுப்பி , சொல்வலை பேன்றவைகளாகும். தற்போது செய்துவரும் குறியீட்டு தரவகத்தை விரிவுப்படுத்த

வேண்டும். அப்போதுதான் நாம் வரலாற்று முறை இலக்கணம் வரலாற்று முறை அகராதி கிளைமொழி ஆய்வு போன்ற பல பணிகளை செய்ய முடியும்.