மொழித் தொழில்நுட்ப வழி சங்க இலக்கியத் தரவக உருவாக்கம்

இரா. அகிலன்
நிரலாளர்
செம்மொழித் தமிழாய்வு மத்திய நிறுவனம், சென்னை

தரவகம் என்பது மின்னுரைகள் அல்லது மின்நூல்களின் தொகுப்பு எனலாம். "ஒழுங்குமுறையுடன் அதிக அளவில் தேர்வு செய்யப்பட்டுக் கணினியில் சேமிக்கப்பட்ட இயற்கையான நடைகளை உடைய உரைகள்". தரவகம் என்பது ஒரு குறிப்பிட்ட மொழியின் சொற்கள் அல்லது துறைச் சொற்களை (Language varieties or Register) உள்ளடக்கியதாக இருக்கவேண்டும். தரவகத்தில் இரண்டு முக்கிய காரணிகள் உள்ளன. 1. தரவக வடிவமைப்பு 2. அளவு என்பன, இவ்விரு காரணிகளும் தரவகத்தின் தரத்தை உறுதி செய்கின்றது. தரவகத்தை இரு பெரும் பிரிவுகளாகப் பிரிக்கலாம். 1. உரைத்தரவகம் (Text corpus), 2. பேச்சுத் தரவகம் (Speech corpus).

தமிழ் மொழிக்கான தரவக உருவாக்கத்தை இந்திய மொழிகளின் நடுவண் நிறுவனம், தமிழ்ப் பல்கலைக்கழகம், அமிர்தா பல்கலைக்கழகம் போன்ற ஆய்வு நிறுவனங்கள் மற்றும் பல்கலைக்கழகங்கள் உருவாக்கி வருகின்றன. தமிழ் மொழிக்கான இயற்கை மொழி ஆய்வு தன்னிறைவை அடைய தற்காலத் தமிழ், இடைக்காலத் தமிழ், சங்க காலத் தமிழ் ஆகியவற்றின் அடிப்படையில் தரவுகள் உருவாக்கப்பட வேண்டும். இக்கட்டுரை மொழித் தொழில் நுட்பத்தைப் பயன்படுத்தி சங்க இலக்கியத்திற்கான தரவக உருவாக்கம் பற்றி எடுத்துரைப்பதாக அமைகிறது.

1. **தரவகம்** (Corpus)

கணினியில் பயன்படுத்தும் வகையில் முறையாகச் சேகரித்து வைக்கப்பட்டுள்ள ஒரு மொழியில் அல்லது பல மொழிகளில் அமைந்த பல்வேறு பனுவல்களின் (texts) தொகுப்பு தரவகம். அவ்வாறான தரவகம் ஒருமொழியின் பல்வேறு பரிணாமங்களை எதிரொலிப்பதாகவும் பலதரப்பட்ட புலங்களுக்கு முதன்மை அளிப்பதாகவும் அறிவியல் முறைப்படியும் இருக்க வேண்டும். தரவகத்தின் இருபிரிவுகள் 1. உரைத்தரவகம் 2. பேச்சுத் தரவகம் என்பன, இதில் உரைத்தரவகம் இரு பெரும் பிரிவுகளாகப் பரிக்கப்படுகிறது. 1. தரவகம் (raw corpus), 2. குறியீட்டுத் தரவகம் (tagged corpus) ஆகியன. பொதுவாக தரவகத்தைக் கொண்டு இயற்கை மொழி ஆய்வை ஒரு குறிப்பிட்ட எல்லை வரையே செய்ய இயலும். இந்த ஆய்வுகளை முழுமைப்படுத்த குறியீட்டுத் தரவகம் உருவாக்கப்பட வேண்டும். மேம்பட்ட தேடுதல் (advance search engine), இயந்திர மொழிபெயர்ப்பு (machine translation), சொற்பிழை திருத்தி (spell checker), இலக்கணப்பிழை திருத்தி (grammar checker), சந்தி உருவாக்கம் (santhi developer), உருபனியல் பகுப்பான் (morphological analyzer) போன்ற மொழியியல் கருவிகள் உருவாக்கத்திற்குக் குறியீட்டுத் தரவகம் இன்றியமையாதது.

2. **தரவக மொழியியல்** (Corpus Linguistics)

தரவக மொழியியல் என்பது மொழியின் கூறுகளாகிய சொற்கள், சொல்லுருவாக்கம், இலக்கணம், சொற்றொடர் அமைப்பு, சொற்பொருள் அமைப்பு, உரையாடல் அமைப்புப் போன்றவற்றை மொழிப் பயன்பாட்டின் வழி ஆராய்வதாகும்.

3. **மொழித் தொழில்நுட்பம்** (Language Technology)

மொழி காலந்தோறும் பல்வேறு தடங்களில் பயணிக்கிறது. கல்வெட்டுகள், செப்பேடுகள், ஓலைச்சுவடிகள், தாட்சுவடிகள், நூல்கள் என்ற வரிசையில் வெவ்வேறு ஊடகங்களில் பயணித்து வளர்ந்து வந்துள்ளது. தற்பொழுது மொழியானது கணினி, கைபேசி மற்றும் பல்வேறு தொழில்நுட்பக் கருவிகளாலும் பயணிக்கிறது. இந்த நூற்றாண்டில் கணினியும் கைபேசியும் மற்றும் பல்வேறு வகையான தொழில்நுட்பக் கருவிகளும் மக்களுடைய பயன்பாட்டிற்கு வந்ததுள்ளது. இதன் காரணமாக மொழித் தொழில்நுட்பம் வளரத் தொடங்கியது. மொழித் தொழில்நுட்பம் என்பது மொழியை அடிப்படையாகக் கொண்டு மனிதனும் கணினியும் ஊடாடுவதாகும். இவ்வாறு ஊடாடு நிகழ்வதற்குப் பல்வேறு படிநிலைகள் உள்ளன. ஒரு நிலையில் புதிய புதிய கணினி மொழியியல் ஆய்வுகள் நடைபெற வேண்டும். மற்றொரு புறம் பல்வேறு மொழித் தொழில்நுட்பக் கருவிகள் உருவாக்க வேண்டும். மேற்கூரிய மொழித் தொழில்நுட்பத்திற்கு அடிப்படையாக இருப்பது தரவகம்.

மொழித் தொழில்நுட்பம் என்பது துறைகளை ஒருங்கிணைத்துக் கூட்டுத் திட்டமாகத் திகழ்கின்றது. அதாவது தகவல் தொழில்நுட்பம் (Information Technology), 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 இணைந்த கலவையாகக் காணப்படுகிறது. இவை ஆரம்ப காலத்தில் சிறுசிறு ஆய்வுகளாகத் தொடங்கப்பட்டுப் பின்பு துறைகளாக வளரத்தொடங்கின. இந்த கருவிகள் அனைத்தும் பேச்சு மொழியிலும் எழுத்து மொழியிலும் அமைய வேண்டும். இதுவரை மொழித் தொழில் நுட்பத்தைப் பயன்படுத்தியும் மொழித் தொழில்நுட்பத்திற்காகவும் உருவாக்கப்பட்டுள்ள மற்றும் உருவாக்கப்பட வேண்டிய கருவிகளும் தரவுத்தளங்களும் பின்வருமாறு

3.1 மொழித் தொழில்நுட்பக் கருவிகள்

1. சொல்லடைவுக் கருவி (Indexing tool)
2. தொடரடைவுக் கருவி (Concordance tool)
3. பயில்வெண் ஆய்வுக் கருவி (Frequency analysis tool)
4. உரை ஒப்புமைக் கருவி (Text comparison tool)
5. சொல் வகைப்பாட்டுக் கருவி (POS tagger tool)
6. தமிழ் அகர வரிசைக் கருவி (Tamil sorting Tool)
7. உருபனியல் பகுப்பி (Morphological analyzer)
8. உருபனியல் உருவாக்கி (Morphological generator)

9. தொடரனியல் குறிப்பான் (Syntactic tagger)
10. தொடரனியல் பகுப்பி (Syntactic parser)
11. தொடரனியல் குறிப்பான் (Sentence segmenter)
12. சொற்பகுப்பி (Word segmenter)
13. தொடர்ப் பகுப்பி (Phrase predicator)
14. வினா விடை (Question Answering (QA))
15. சொற்பொருள் மயக்க நீக்கி (Word Sense Disambiguation (WSD))
16. சொற்பொருள் மயக்கம் நீக்கி (Knowledge Representation from Text)
17. சொற்பொருள் குறிப்பான் (Semantic tagger Tool)
18. ஒருகால மின்னகராதி (E-dictionary for synchronic Tamil)
19. வரலாற்றுக்கால மின்னகராதி (E-dictionary for Historical Tamil)
20. வரலாற்றுக்கால மின்சொற்கோவை (E-Thesauruses for Historical Tamil)
21. தமிழ்த் தேடுபொறி (Search Engine for Tamil)
22. தமிழ்ச் சொல் வலை (WordNet for Tamil)

3.2 மொழித் தரவுத்தளம் **(Language Database)**

1. மூல பாடம் (Source Text)
2. சந்திபிரித்த பாடம் (Hyphenated Text)
3. சொல்பிரித்த பாடம் (Segmented Text)
4. குறியீட்டுத் தரவகம் – இலக்கணவகை (Annotated corpus –Pos. TAG)
5. குறியீட்டுத் தரவகம் - தொடரியல் (Annotated corpus –Syntactic TAG)
6. குறியீட்டுத் தரவகம் – பொருண்மையியல் (Annotated corpus –Semantic TAG)
7. குறியீட்டுத் தரவகம் – கருத்தாடவியல் (Annotated corpus - Pragmatic TAG)

4. சங்க இலக்கியத் தரவகம்

தற்காலத் தமிழுக்கான தரவகத்தைப் பல்வேறு ஆய்வு நிறுவனங்கள், பல்கலைக்கழகங்கள் உருவாக்கி வருகிறது. சங்க இலக்கியங்களுக்கான தரவகத்தை மேற்குறிப்பிட்ட மொழியியல் கருவிகளின் துணை கொண்டு உருவாக்கலாம். மொழித் தொழில்நுட்பத்தைப் பயன்படுத்தி சங்க இலக்கியத் தரவுகள் செம்மொழித் தமிழாய்வு மத்திய நிறுவனத்தின் மொழித் தொழில்நுட்பப் துறையால் உருவாக்கப்பட்டுவருகிறது. இதில் முதல் கட்டமாக 20 (தொல்காப்பியம், எட்டுத்தொகை, பத்துப்பாட்டு மற்றும் இறையனார்களவியல்) சங்க இலக்கியங்களுக்கான தரவகம் உருவாக்கப்பட்டு இணையவழித் தொடரடைவு ஒன்று உருவாக்கப்பட்டு இணையத்தில் (http://www.cict.in)அளிக்கப்பட்டுள்ளது. சங்க இலக்கியங்களுக்கான சொல்லடைவை உருவாக்குவதற்குச் சங்க இலக்கிய சொல்லடைவி எனும் ஒரு மென்பொருள் உருவாக்கப்பட்டுத் தரவிரக்கம் செய்து கொள்ளும் வகையில் அளிக்கப்பட்டு வருகிறது. சங்க இலக்கிய ஆய்வுகளை மேம்படுத்தும்நோக்கில் சங்க

இலக்கியங்களுக்கான குறியீட்டுத் தரவகம் உருவாக்கும் பணி நடைபெற்று வருகின்றது. இந்தக் குறியீட்டுத் தரவகங்களை உருவாக்க (POS Annotation for Classical Tamil semi tagger) எனும் மென்பொருள் ஒன்றை உருவாக்கி இதன் வழி சங்க இலக்கியங்ளுக்கான குறியீட்டுத் தரவகம் உருவாக்கபட்டுவருகிறது.

5. மொழியியல் கருவிகள் உருவாக்கும் முறை

சங்க இலக்கிய ஆய்வுகளுக்குப் பயன்படும் மொழியியல் கருவிகள் உருவாக்கம் இரண்டு பிரிவுகளாகப் பிரிக்கலாம். 1.தரவுகள் வழி (Corpus based approach) மொழியியல் கருவிகள் உருவாக்கம் 2. விதிகள் வழி (Rule based approach) மொழியியல் கருவிகள் உருவாக்கம் என்பன.

5.1 தரவுகள் வழி மொழியியல் தொழில்நுட்பக் கருவிகள் உருவாக்கம்

இணையதளங்களிலும், நூல்வடிவிலும் உள்ள சங்க இலக்கியத் தரவுகளைப் பயன்படுத்தி சங்க இலக்கியங்களுக்கான மொழியியல் தொழில்நுட்பக் கருவிகளை உருவாக்கலாம்.

5.2 விதிகள் வழி மொழியியல் தொழில்நுட்பக் கருவிகள் உருவாக்கம்

மொழியியல் நுட்பத்தைப் பயன்படுத்தி கணினி புரிந்துகொள்ளும் வகையிலான மொழியியல் விதிகளை உருவாக்கி இதன் மூலம் தொழில்நுட்பக் கருவிகளை உருவாக்கலாம். இதன் வழி உருவாக்கும் கருவிகள் அதிக சதவிகிதத்திலான முடிவுகளைக் தருகிறது. மேலும் இதற்கு அதிகம் தரவுகள் தேவைப்படுவது இல்லை. மொழியியல் விதிகளைப் பயன்படுத்திச் சங்க இலக்கியத்திற்கான உருபனியல் பகுப்பான் உருவாக்கும் முறை பின்வரும் சங்க இலக்கியத்தில் இடம் பெறும் பன்மைக்குறிட்டிற்கான உருபனியல் பகுப்பான் மொழியியல் விதிகளைப் பயன்படுத்து உருவாக்கும் முறை

1. Check the root word dictionary { if ‘yes’ assign the appropriate tag }
2. Else check the suffix
3. If the suffix is ‘kaḷ’ assign the tag as ‘PL’
4. Check the root word dictionary
5. if yes assign the root word dictionary as ‘NN’ { go to step 22 }
6. Else if the suffix is ‘ṅkaḷ’
7. Remove the first character of the suffix ‘ṅ’
8. Assign the tag as ‘PL’
9. Add (m) in the last character of root word
10. Assign the Tag as ‘NN’ {go to step 22 }
11. Else if the suffix is‘ r̲kaḷ‘
12. Remove the first character of the suffix ‘r̲’
13. Assign the tag as ‘NN’
14. Add (l) in the last character of root word
15. Assign the Tag as ‘NN’ {go to step 22 }
16. Else if the suffix is‘ṯkaḷ‘
17. Remove the first character of the suffix ‘ṯ’
18. Assign the tag as ‘PL’ {go to step 22 }
19. Add (ḷ) in the last character of root word
20. Assign the Tag as ‘NN’
21. Else assign as ‘NN’ {go to step 22 }
22. Stop.

உதா.

மகள் makaḷ (ஐங். 91:3), திங்கள் tiṅkaḷ (பரி. 3:5), மக்கள் makkaḷ (புற.191:3)

கிளைகல் (kiḷaikal) *(நாலெ.191:2)* கலவை+ கள் (kalavai+ kaḷ) *(நாலடி.268)*

இடங்கள் (lṭaṅkaḷ) *(தொல். 1154:2)* சொற்கள் (coṟkaḷ) *(கலி. 81:13)*

முடிவுரை

மொழித் தொழில்நுட்பத்தின் பயனை அடைவதற்கு மிகுதியான உழைப்பும் தமிழ் அறிஞர்கள், மொழியியல் அறிஞர்கள் மற்றும் கணினி வல்லுனர்களின் கூட்டு முயற்சி தேவைபடுகிறது. இந்திய மொழிகள் அனைத்திற்கும் மொழித் தொழில்நுட்பக் கருவிகளை உருவாக்க பல்வேறு ஆய்வுநிறுவனங்கள், பல்கலைக்கழகங்கள் மற்றும் அரசு சாரா நிறுவனங்களும் ஈடுபட்டு வருகின்றன். பிரஞ்சு, ஆங்கிலம், ஜெர்மன் போன்ற மொழிகளில் கணினியும் மனிதர்களும் வேண்டிய தகவல்களைப் பெற்றுக்கொள்ளக்கூடிய இடைமுகம் நன்கு வளர்ச்சியடைந்து உள்ளது. இதுபோல தமிழ் மொழியின் கணினிப் பயன்பாடு சிறக்க புதிய தொழில்நுட்பக் கண்டுபிடிப்புகளை அரசு ஊக்குவிக்க வேண்டும். தொழில்நுட்பங்களை அவரவர் மொழியிலேயே பயன்படுத்த அரசு விழிப்பணர்வை ஏற்படுத்த வேண்டும்.

நூற்பட்டியல்

1. Natural Language Understanding by Allen J. – The Benjamins Publishing Company – 1995
2. கோ.பழனிராஜன், துணைப் பேராசிரியர், மொழியியல் துறை, கேரளா மத்தியப் பல்கலைக் கழகம், கேரளா “மொழித் தொழில்நுட்பம் ஓர் அறிமுகம்” 2013 ‘கணினியியல் தொழில் நுட்பங்களும் சங்க இலக்கிய ஆய்வுகளும்’ தேசியக் கருத்தரங்கு, எஸ் ஆர் எம் பல்கலைக் கழகம், சென்னை.
3. http://en.wikipedia.org/wiki/Corpus_linguistics
4. Speech and Language Processing by Jurafsky, Daniel and James H. Martin, New Delhi-Pearson Education 2002
5. Prof. E.R.Naganathan, R.Akilan "Morphological Analyzer for Classical Tamil text - a Rule based approach " 2012, 12th International Internet conference, Annamalai University, Chidamparam.