கணித்தமிழ் வளர்ச்சிப் பணிக்கான செயல்திட்டங்கள்

தமிழ்த்தரவுமொழியியல் (Tamil Corpus Linguistics) :

1. சங்ககால இலக்கியங்கள் முதற்கொண்டு தமிழ்மொழியிலிருந்து, இன்றைய தமிழ்மொழி வரை , அனைத்து இலக்கியங்களுக்கும் கட்டங்களுக்கும் தரவுத்தளங்களை (Corpora) தரவுமொழியியல் நோக்கில் உருவாக்குவது.
2. ஏற்கனவே உள்ள தரவுத்தளங்களான மதுரைத்திட்டம், இந்திய மொழிகள் நடுவண் நிறுவனத்தின் தரவுத்தளம், பேரா. பா. பாண்டியராஜா உருவாக்கிய தமிழிலக்கியத் தரவுத்தளம், கிரியா தரவுத்தளம் போன்ற தரவுத்தளங்களை உருவாக்கியவர்களிடமிருந்து முறையாகப் பெறுவது.
3. தமிழ்நாடு அரசு பாடநூல் நிறுவனம் இதுவரை உருவாக்கிய பாடநூல்களின் மென்படியைப் (soft copy) பெறுவது.
4. தனியார், நிறுவன வெளியீடுகள், நாளிதழ்கள், வார, மாத இதழ்கள் , மலர்கள் வெளியிடும் ஊடக நிறுவனங்கள் ஆகியோரிடமிருந்து மென்படிகளைப் பெறுவது.
5. தரவுமொழியியல் நோக்கில் தரவுத்தளத்தின் அமைப்பையும் (Structure), பிற கூறுகளையும் (other features such as meta data) தரப்படுத்தி, அதனடிப்படையில் தமிழ்த் தரவுத்தளங்களை உருவாக்குதல். (வேறுபட்ட பல பயன்பாடுகளுக்கும் உதவும் வகையில் உலகத்தரம் வாய்ந்த தரவுத்தளமாக அது அமைய ஏற்பாடு செய்தல்) (Text Encoding Initiative (TEI) என்ற உலக அளவிலான அமைப்பின் பரிந்துரைகளைப் பின்பற்றலாம்)
6. எழுத்துத் தமிழுக்கு மட்டுமல்லாமல், பேச்சுத் தமிழுக்கும் தரவுத்தளத்தை உருவாக்குதல். (இத்தரவுத்தளம் கிளை மொழியியல், சமூக மொழியியல் ஆகியவற்றின் அடிப்படையில் வட்டார, சமூக வழக்குகளின் மொழி நடையையும் வெளிப்படுத்துவதாக அமைய வேண்டும். பேச்சுத் தொழில் நுட்பத்தைத் தமிழுக்குச் செயல்படுத்தி, தேவையான மென்பொருள்களை உருவாக்குவதற்கும், வேறுசில பயன்பாடுகளுக்கும் பேச்சொலியியல் அடிப்படையில் பேச்சுத்தரவும் (பேச்சொலித்தரவு- Speech Corpus) மிகவும் பயன் உள்ளவையாகும் ஆகும்
7. மேற்குறிப்பிட்ட தரவுத்தளங்கள் இரண்டு வகையில் அமைக்கப்பட வேண்டும்; ஒன்று, மொழியியல் குறிப்பு இல்லாத தரவு (Plain Corpus) , மற்றொன்று தேவையான மொழியியல் குறிப்புகளைக் கொண்ட தரவு (Linguistically Annotated Corpus).
8. தரவுத்தளத்தில் பயன்படுத்துகிற மொழியியல் குறிப்புகள் (Linguistic annotations such as Word - class Tagger) தரப்படுத்துதல். அதனடிப்படையில் தரவுகள் உருவாக்குதல்.

9. தொடரடைவு (Concordancer) , என்-கிராம்ஆய்வு (N-Gram Analysis) , புள்ளியியல் ஆய்வு (Statistical Analysis) போன்றவற்றிற்குத் தரவுத்தளங்களை உட்படுத்தி, மொழித் தொழில்நுட்பம் (Language Technology), பேச்சுத் தொழில்நுட்பம் (Speech Technology) , மொழி பயிற்றல் (Language Teaching / Learning) போன்றவற்றிற்குப் பயன்படுகிற மொழி விவரங்களையும் கருவிகளையும் அளித்தல்.
10. திருத்தப்படாத, இயற்கையாகக் கிடைக்கின்ற மொழித்தொடர்களை அப்படியே கொண்ட தரவுகளும்(un-corrected Corpus), திருத்தப்பட்ட தரவுகளையும் (corrected, normalized, error-free Corpus) உருவாக்குதல்.
11. தரவுகளை உருவாக்குவதற்குப் பயன்படக்கூடிய அத்தனை மூலாதாரங்களையும் (இணையம் - Web, மேக மூலாதாரங்கள் – Cloud sources போன்றவை) பயன்படுத்துதல்.
12. தமிழ் ஆர்வலர்களைத் தங்களது தரவுகளைத் தமிழ் இணையக் கல்விக்கழகத்தின் சேவைக் கணினிக்கு (server) அளிக்கும் வகையில் வசதிகளை உருவாக்குதல்.

தமிழ் மொழித் தொழில் நுட்பம் (Tamil Language Technlolgy):

1. கணித்தமிழை உள்ளீடு (input) செய்வதற்கான ஒருங்குறி அடிப்படையிலான எழுத்துருக்கள், விசைப்பலகைகள் ஆகியவற்றை தரப்படுத்த, தமிழ் இணையக் கல்விக்கழகம் தேவையான நடவடிக்கைகளை (Standardization) மேற்கொள்ளுதல்.
2. TACE (Tamil All Character Encoding) குறியேற்ற முறையையும் தேவையான ஆய்வுகளுக்குப் பயன்படுத்துதல்.
3. அச்சிலுள்ள நூல்களைக் கணினிக்கு உள்ளீடு செய்வதற்குத் தேவையான மிகச்சிறந்த ஒளிவழி எழுத்தறிவானை (Optical Character Recognizer – OCR) உருவாக்குதல்.
4. ஒளிவழி எழுத்தறிவானுடன் செயல்படும் சொற்பிழைதிருத்தியை உருவாக்குதல்.
5. தமிழ்மொழி அறிவைக் கணினிக்கு அளித்து, அதன் வழி நமக்குத் தேவையான மொழிவழிச் செயற்பாடுகளைக் (Language / linguistic tasks) கணினி வழி மேற்கொள்ளத் தேவையான ஆய்வுகளை மேற்கொள்ளுதல்.
6. தமிழ்மொழியியலின் அடிப்படையில் கணினி ஒலியனியல் (Computational Phonology) , கணினி உருபனியல் (Computational Morphology) , கணினித் தொடரியல் (Computational Syntax) , கணினிப் பொருண்மையியல் (Computational Syntax) கணினிப் பொருண்மைச் சூழலியல் (Computational Pragmatics) , கணினிக் கருத்தாடலியல் (Computational Discourse)போன்ற பல மொழிப்படி நிலைகளுக்கான (linguistic levels) கணினி மாதிரிகள் (Computational Models) – ஆய்வு வடிவங்களை உருவாக்குதல்
7. உருபன் பகுப்பாய்வி (Morphological Parser) , உருபன் உருவாக்கி (Morphological Generator) , தொடர் பகுப்பாய்வி (Syntactic Parser) , பொருண்மை ஆய்வி (Semantic Analyzer) போன்றவற்றை உருவாக்குதல்.
8. தொல்காப்பியர் கோட்பாடுகளைப் பின்பற்றியும் கணினிக்குத் தேவையான மொழி மாதிரிகளை உருவாக்குதல்.
9. மேற்குறிப்பிட்ட கணினிக்கான மொழியியல் ஆய்வின் அடிப்படையில் தமிழுக்கான ஒரு முழுமையான இயற்கைமொழி ஆய்வுப்பொறி (Natural Language Processing Engine – NLP Engine) உருவாக்குதல்.
10. மேற்கூறிய தமிழுக்கான இயற்கைமொழி ஆய்வுப்பொறியின் உதவி கொண்டு, தகவல் பெறுதல் (Information Retrieval / Extraction – IR/IE), உரையுணர்தல் (Text Understanding) , உரைச்சுருக்கம் தருதல் (Text Summarization) , மனிதன் - கணினி உரையாடல் (Human – Machine Interface), தானியங்கு மொழிபெயர்ப்பு (Automatic

Machine Translation) போன்ற பலவகைப் பயன்பாட்டு மென்பொருள்களை தமிழுக்கு உருவாக்குதல்.

11. கணினி, செல்பேசி உட்பட அனைத்து மின்னணு தகவல் தொடர்புச் சாதனங்களிலும் (Electronic Communication Devices) தமிழைப் பயன்படுத்துவதற்குத் தேவையான பலவகை நுகர்வோர் மென்பொருள்களை உருவாக்குதல்.
12. உட்செருகியாகப் (Add-in / Plug-in) பயன்படும் சொற்பிழை திருத்தி, அகராதி, சந்திப்பிழை திருத்திப் போன்ற சிறுசிறு மென்பொருள்களை உருவாக்குதல்.
13. மென்பொருள் நிறுவனங்களின் பெரிய தமிழ்மென்பொருள் உருவாக்கங்களுக்குப் பயன்படும் தமிழ் நிரலாக்கங்களை (DLL, API) உருவாக்குதல்.

தமிழ்ப்பேச்சுத்தொழில்நுட்பம் (Tamil Speech Technology):

1. தமிழ்ப் பேச்சொலிகளின் பௌதீகப்பண்புகளை (Acoustic Physical Properties) ஆய்வதற்குத் தேவையான கணினிவழிப் பேச்சொலி ஆய்வுக்கூடம்(Computerized Speech Lab – CSL) அமைத்தல். (இவ்வாய்வுக்கூடம், மொழி வல்லுநர்கள், பொறியியலாளர்கள், தொழில்நுட்பத்துறையினர் இணைந்து பணியாற்றும் ஒரு ஆய்வுக்கூடமாக அமையும்)
2. பலவகை பேச்சொலித்தரவுகள் (Speech Corpus) உருவாக்கப்பட்டு, பேச்சொலி ஆய்வுக்கருவிகளின் உதவி கொண்டு தமிழ்ப் பேச்சொலிகள் (Phones) அனைத்தையும் ஆய்வு செய்தல்.
3. பேச்சொலிகள் மட்டுமல்லாமல், ஒலியழுத்தம் போன்ற கூறுகளையும் (Amplitude, Stress, Pitch) , ஏற்ற– இறக்கம் (Intonation) போன்ற மேற்கூற்று ஒலியியல் கூறுகளையும் (Suprasegmental features) ஆய்வு செய்தல்.
4. ஒலியன் (Phoneme) , அசை (Syllable) ஆகியவற்றின் அடிப்படையில் தொடர்மொழிக்கான (Continious Speech) கணினிவழி பேச்சுரையை உருவாக்கத் தேவையான அனைத்து ஆய்வுகளையும் மேற்கொள்ளுதல்.
5. மேற்குறிப்பிட்ட ஆய்வுகளின் அடிப்படையில் பேச்சுரை – எழுத்துரை மாற்றி (Automatic Speech Recognizer – ASR) , எழுத்துரை - பேச்சுரை மாற்றி (Text to Speech – TTS) ஆகியவற்றை தமிழுக்கு உருவாக்குதல்.
6. எழுத்துரை – பேச்சுரை மாற்றிக்கான பணிக்கு முன்னுரிமை அளித்தல். (பார்வைத்திறன் இல்லாதவர்களுக்கு மிகவும் இது பயன்படும்)

தமிழ் மின்னகராதி (Tamil Electronic Dictionary) :

1. தமிழ் மொழி அகராதி, தமிழ் – ஆங்கில அகராதி, ஆங்கிலம் – தமிழ் அகராதி எனப் பலவகை மின்னகராதிகளைத் தமிழ்மொழிக்கு உருவாக்குதல். (கணினிவழி அகராதிகள் (மின்னகராதிகள் – Electronic Dictionaries) இன்றைய மின்னணு உலகத்தில் மிகமுக்கியமான கருவிகளாகப் பலமொழிகளுக்கு உருவாக்கப்பட்டிருக்கின்றன. அச்சில் உள்ள அகராதிகளைவிட (Printed dictionaries) இவ்வகை அகராதிகள் எளிமையாக மாணவர்கள் உட்பட அனைவருக்கும் மிகவும் பயன்படும்.)
2. கணினி, அலைபேசி ஆகியவற்றில் பயன்படும் மென்பொருள்களாக உருவாக்கப்படுகிற மின்னகராதிகளுடன், , கையடக்க மின்னகராதிகளையும் (Hand-held Embedded Dictionaries) உருவாக்குதல்.
3. பொது அகராதி, துறைசார்ந்த அகராதி, சொற்பிறப்பியல் அகராதி , காலமுறை அகராதி , வட்டார வழக்கு அகராதி எனப் பலவகையான மின்னகராதிகளைத் தமிழில் உருவாக்குதல்
4. அகராதியியல் வல்லுநர்கள், கணினி மொழியியல் வல்லுநர்கள், கணினியியலார்கள் இணைந்து மேற்கொள்கிற ஒரு பணியாக இப்பணி அமைதல்.
5. மனித மூளைக்குப் பயன்படுகிற அகராதிகள் (Dictionaries for human brain understanding) மட்டுமல்லாமல், கணினிக்கேற்ற, கணினி புரிந்து கொள்கிற அகராதிகளும் (Dictionaries/ Lexicons understood by Computer) உருவாக்கப்பட்டால் தான் தமிழ் இயற்கை மொழி ஆய்வு வெற்றியடைய முடியும். கணினிவழி மொழிபெயர்ப்பு (Automatic Machine Translation – MT) போன்ற பயனுள்ள தமிழ்மொழிக்கானப் பல பயன்பாட்டு மென்பொருள்களை உருவாக்கமுடியும். சொல்வலை (WordNet), உருவாக்க அகராதி (Generative Lexicon) போன்ற கணினி மொழியியலில் பயன்பட்டு வருகின்ற ஆய்வுமுறைகளைப் பின்பற்றித் தமிழுக்கு அகராதிகள் உருவாக்கப்பட வேண்டும்.

நூலகம்:

தமிழ்மொழி இலக்கியங்கள், இலக்கணங்கள், அகராதிகள் , கணினி மொழியியல், மொழித்தொழில்நுட்பம் , கணினியியல் , புள்ளியியல் போன்ற துறைகளைச் சார்ந்த நூல்கள், இதழ்கள் நூலகத்திற்கு வாங்குதல்.

மென்பொருள் கருவிகள் :

தமிழ் இயற்கைமொழி ஆய்வுக்கும் தமிழ்மென்பொருள் உருவாக்கத்திற்கும் தேவையான மென்பொருள்கருவிகளைக் கொள்முதல் செய்தல்.

மனிதவளம் :

கணினி மொழியியல் துறையில் தேவையான மனிதவளத்தை உருவாக்கத் தேவையான படிப்புகள், பயிற்சிகள், பணிமனைகள் நடத்தப்பட வேண்டும்.

இறுதி அறிக்கை வரைவு தயாரிப்பு - ந. தெய்வ சுந்தரம்