மொழியியல் ஆய்வுகளும் தரவகமும்
Language research and corpus

சி.ம.இளந்தமிழ் / C.M. Elanttamil[1]
முனைவர் சாவ் மெங் ஹுவாட் / Dr. Chau Meng Huat[2]

Date of submission: 2022-05-11
Date of acceptance: 2022-06-20
Date of Publication: 2022-07-28
Corresponding author's Name:
C.M. Elanttamil
Email: elanttamil@um.edu.my

Information technology's current development has significantly changed language learning and teaching. Though linguistic research has been continuously undertaken, corpus analysis based on modern Linguistics is very much needed to obtain clear-cut answers with respect to specific research problems and research questions. Corpus Linguistics could identify or discover even the micro-level linguistic features through the development of an exceptionally enormous size corpus and its analysis by objective methods. Corpus Linguistics provides the objective methods for constructing an extensive corpus for a language and its analysis with the help of proper linguistic methods. The current development of Corpus Linguistics and Information Technology enables linguists and language teachers to equip themselves with the necessary skills and opportunities. In a multilingual country such as Malaysia, these types of corpus linguistic research would undoubtedly pave the way for both the language educators and learners of the mother tongue in their teaching and learning tasks. Moreover, it would help to understand the current development in the Lexicon and Grammar of mother tongues.

Key Words: Data, Corpus, Corpus linguistics, Computational linguistics, Analysis

பின்னணி

மொழிகள் பற்றிய மொழியியல் பகுப்பாய்வில், மொழியியலாளர்களால் முன்மொழியப்பட்ட அமைப்புக் கருத்தியல் (Structural School), மாற்றிலக்கணக் கருத்தியல் (Generative School) போன்ற பல கருத்தியல்கள் உள்ளன. ஆராய்ச்சியின் நோக்கத்தைப் பொறுத்து, அந்தந்தக் கருத்தியலை ஆராய்ச்சியாளர்கள் தேர்ந்தெடுக்க வேண்டும். உண்மையான கருத்துப் பரிமாற்றச் செயல்பாட்டில், பயன்படுத்தப்படும் வெளிப்புற மொழியின் (external language) ஆய்வுக்கு, அதாவது மொழிச் செயல்திறனுக்கு (linguistic performance), அமைப்புப் பகுப்பாய்வைத் தெரிந்தெடுக்கலாம். உண்மையான செயல்திறனுக்குப் பின்னால் உள்ள மொழித் திறன் (linguistic competence) பற்றிய ஆய்வுக்கு, மாற்றிலக்கணக் கருத்தியலை ஏற்றுக்கொள்ளலாம்.

மொழியியல் அல்லது மொழிச்செயல்திறன் பகுப்பாய்விலும், மொழியியல் அறிவியலில் இரண்டு வெவ்வேறு முறைகள் உள்ளன. ஒன்று, உண்மையான கருத்துப் பரிமாற்றச் சூழலிலிருந்து தனிமைப்படுத்தப்பட்ட சொற்றொடர்களைப் பற்றிய ஆய்வு. இந்தப் பகுப்பாய்வு முறையானது, பொதுவாக அமைப்பு மொழியியல் (Structural Linguistics) என்று அழைக்கப்படுகிறது. இந்த முறையில், சொற்களும் சொற்றொடர்களும் நுண்மக்

[1] The author is a Lecturer in the Department of Malaysian Language and Applied Linguistics, Universiti of Malaya, Kuala Lumpur, Malaysia. elanttamil@um.edu.my

[2]The author is a Senior Lecturer in the Department of English Language, Universiti of Malaya, Kuala Lumpur, Malaysia. chaumenghuat@um.edu.my

கூறுகளாகப் பகுப்பாய்வு செய்யப்படும். பொதுவாக, இந்தப் பகுப்பாய்வு சொற்றொடர் வரை செல்லும். இந்தக் கருத்தியல் 20 ஆம் நூற்றாண்டின் மொழி கற்பித்தலில் பெரும் தாக்கத்தை ஏற்படுத்தியது. இந்த மொழியியல் பகுப்பாய்வின் வெளியீடானது, மொழி கற்பித்தலின் உள்ளடக்கத்தை உருவாக்கியதோடு வெவ்வேறு விதமான ஆய்வுகளையும் உட்படுத்தியது.

பெர்தியன் குழுவைச் (Firthian school) சேர்ந்த மொழியியலாளர்களால் மற்றொரு முறையானது முன்மொழியப்பட்டு ஏற்றுக்கொள்ளப்பட்டது. இந்தக் கருத்தியல் குழுவானது, சூழலில் காணப்படும் மொழிக்கான பகுப்பாய்விற்கு முக்கியத்துவம் அளிக்கிறது; மொழியை அதன் சூழலிலிருந்து பிரித்துப் பார்ப்பது இல்லை. இந்தக் கருத்தியலைச் சேர்ந்த மொழியியலாளர்கள், ஒரு மொழியின் சூழலில் அல்லது கருத்துப்பரிமாற்றங்களில் பயன்படுத்தப்படும் கூற்றுகளைப் பகுப்பாய்வு செய்தால் மட்டுமே அந்த மொழியின் கூறுகளை அறியலாம் என்று வலியுறுத்துகின்றனர். "பயன்பாட்டில் மொழி" என்பது அவர்களின் கருப்பொருள், அதாவது மொழியின் அமைப்பானது மொழிப் பயன்பாட்டினால் தீர்மானிக்கப்படுகிறது. உண்மையான பயன்பாட்டில் மட்டுமே, சொற்றொடர்கள் அல்லது கூற்றுகள் கருத்துப்பரிமாற்ற மதிப்பைக் கொண்டிருக்கும். எனவே, பெர்த்தியன் கருத்தியலின் கூற்றுப்படி, கூற்றுகளும் பல்வேறு கூற்றுகளைக் கொண்ட முழுப் பனுவலும் / உரைகளும் மொழியியல் பகுப்பாய்வின் பொருளாக இருக்கவேண்டும். மொழிச் சூழலில் உள்ள முழுமையான கூற்றுடன் கூடிய ஒத்திசைவான பனுவல் (coherent text) இந்த மொழிப் பகுப்பாய்வுக் கருத்தியலின் உயர் நிலை ஆகும்.

பிந்தைய காலகட்டத்தில், மேலே குறிப்பிடப்பட்ட இரண்டாம் மொழியியல் கருத்தியல்வாதிகள், இணைப்பனுவலுடன் (co-text), உண்மையான கருத்துப்பரிமாற்றச் சூழலையும் (context) அவர்களின் மொழியியல் பகுப்பாய்வில் சேர்த்தனர். மேலும், அவர்கள் முக அசைவுகள், பிற உடல் அசைவுகள், விளக்கப்படம், வரைபடம் போன்ற மொழிசாராக் (non - verbal) கருத்துப்பரிமாற்றக் கூறுகளைச் சேர்க்க விரும்பினர். அதாவது, உண்மையான கருத்துப்பரிமாற்றத்தின் மொழிசார், மொழிசாராக் கூறுகள் மொழிப் பகுப்பாய்வின் பொருளாக இருக்கவேண்டும். இவ்வாறாக, அவர்கள் "பனுவல்" (Text) என்ற நிலையைக் கடந்து, "கருத்தாடல்" (Discourse) என்ற நிலைக்குச் செல்கிறார்கள். அதாவது, மொழி பகுப்பாய்வின் பொருளானது, கருத்துப்பரிமாற்றமே கருத்துப்பரிமாற்றத்தில் மொழிசார், மொழிசாராக் கூறுகளின் பயன்படுத்தம். பின்னர், மொழிப் பயனரின் சூழல்சார் பொருண்மையியல் (Pragmatics) அறிவை அவர்கள் மொழிப் பகுப்பாய்வில் சேர்த்துக்கொண்டார்கள்.

எனவே, தற்போது, பெர்தியன் மொழியியலாளர்கள் மொழி / கருத்துப்பரிமாற்றப் பகுப்பாய்வில் கருத்தாடல் (Discourse Analysis) ஆய்வை விரும்புகிறார்கள். இந்த வளர்ச்சியானது, மொழி கற்பித்தலிலும் தனது தாக்கத்தை ஏற்படுத்துகிறது. இந்த மொழியியல் கருத்தினர் கருத்துப் பரிமாற்றத்திற்காகவும் கருத்துப் பரிமாற்றமாகவும் ஒரு மொழியைக் கற்பிக்க விரும்புகின்றனர். இந்த மொழி கற்பித்தல் முறையில் விடோசனும், அவர் சிந்தனை சார்ந்த கல்வியாளர்களும் பங்களிப்பை வழங்கியுள்ளனர். இதன் தொடர்ச்சியாக பல கல்விக் கோட்பாடுகள் உருவாகின அந்த தொடர்ச்சியில் மிக முக்கியமான ஒன்று தரவக மொழியியல் ஆகும்.

தரவக மொழியியல்

தரவக மொழியியல் என்பது:

> ஒரு மின்னணு வடிவத்தில் சேமிக்கப்பட்ட பனுவல்களைக் கொண்ட தரவகத்தை உருவாக்குவது; அதன்மீது தானியங்கு தேடல்களை நடத்துவதற்கும், இயற்கையாக நிகழும் மொழியின் அமைப்பு, ஒழுங்குமுறை பற்றிய நுண்ணறிவைப் பெறுவதற்கும் தொடரடைவி (concordancer) தொடர்பான ஆய்வுகளுக்கும் பல்வகையான மென்பொருள்கள்

ஆராய்ச்சியாளர்களுக்கு உதவுகிறது என வரையறுக்கலாம் (Szudarski, *2017*).

ஒரு மொழிக்கான தரவக உருவாக்கத்தின் முக்கியக் கூறுகள்

தரவகத்தின் அனைத்து தரவுகளும் நம்பகமானவை (authentic) இருக்கவேண்டும் செயற்கையானவையாக (synthetic) இருக்க கூடாது. அதாவது, அவை உண்மையான மொழி பயன்பாட்டிலிருந்து சேகரிக்கப்படுபவை. தரவக மொழியியலாளர்கள் திட்டமிட்டப்பட்ட, கட்டுப்படுத்தப்பட்ட அல்லது புனையப்பட்ட பனுவல்களையும், சோதனை நிலையில் பேசப்பட்ட அல்லது எழுதப்பட்ட பனுவல்களையும் விலக்குகின்றனர்.

இதற்குக் காரணம் என்னவென்றால், தரவக மொழியியலாளர்கள் மொழிப் பயன்பாட்டை விவரிக்க விரும்புகிறார்கள்; உண்மையான மொழிப் பயன்பாட்டில் பொதிந்துள்ள மொழிக் கோட்பாடுகளை முன்மொழிய விரும்புகிறார்கள். செயற்கையாக உருவாக்கப்பட்ட பனுவல்களையோ ஆராய்ச்சியாளரால் திரித்துக் கையாளப்பட்ட பனுவல்களையோ ஆராய்வதில் அவர்கள் எந்தப் பயனையும் இல்லை என்று கூறுகிறார்கள். (Cheng, 2011)

தரவக மொழியியலானது, மொழியியல் ஆய்வினைத்தவிர பிற காரணங்களுக்காக உருவாக்கப்பட்ட மொழியைப்பற்றியும் ஆராயும் தன்மைக் கொண்டது. மொழிப் பற்றிய முடிவுகளையும் உள்ளுணர்வுகளையும் பயன்படுத்துவதற்குப் பதிலாக அல்லது ஒரு குறிப்பிட்ட கருத்தை விளக்கும் மொழியின் எடுத்துக்காட்டுகளை உருவாக்குவதற்குப்பதிலாக, தரவக மொழியியலானது, உண்மையான மொழிப் பயன்பாட்டைப் பிரதிபலிக்கும் மொழியைப் பயன்படுத்துகிறது (Crawford & Csomay, 2015).

தரவுகள் தனிப்பட்ட சொற்களாகவோ அல்லது சொற்றொடர்களாகவோ சேகரிக்கப்படாமல், ஒன்றுக்கு மேற்பட்ட சொற்றொடர்களைக் கொண்டதொரு ஒத்திசைவான பனுவலாகச் (coherent text) சேகரிக்கப்படுகின்றன. ஒரு தரவகத்துக்குப் பனுவல்களைச் சேகரிக்கும்போது கருத்தில் கொள்ள வேண்டிய மற்றொன்று எதுவென்றால், முழுமையான பனுவல்கள் மட்டுமே சேர்க்கப்படவேண்டுமா அல்லது பனுவல்களின் பகுதிகளைச் சேர்ப்பது ஏற்கத்தக்கதா என்பது. எடுத்துக்காட்டாக, தரவகத் தொகுப்பாளர் ஒவ்வொரு பனுவலும் சம அளவினதாக இருக்கவேண்டும் என விரும்பினால், இது ஒரு சிக்கலாக மாறும், அதாவது தரவகத்தில் உள்ள சில பனுவல்கள் முழுமையடையாமல் இருக்கும். பனுவல்களை ஒப்பிடும்போது சில நன்மைகள் இருப்பதாக சிலர் வாதிடுகின்றனர், ஒரு அளவு நிலைப்பாட்டிற்கு ஏற்றவாறு தரவுகளைக் கொள்வது அவற்றின் நம்பகத்தன்மையைப் பாதிக்கிறது; ஒரு குறிப்பிட்ட பனுவல் வகை எவ்வாறு முடிவடைகிறது போன்ற முக்கியக் கூறுகளையும் நீக்குகிறது. எனவே, இதனால் எழுகின்ற ஒருமித்த கருத்து என்னவென்றால், இயற்கையாக இடம்பெறும் பனுவல்களை முழுமையாகச் சேகரிக்க முயற்சிக்கவேண்டும்.

இது தானியங்கு உத்திகளையும், ஊடாட்ட உத்திகளையும் பயன்படுத்தி, பகுப்பாய்விற்கு கணினிகளை விரிவாகப் பயன்படுத்துகிறது. இது அளவறி பகுப்பாய்வு நுட்பங்களையும (quantitative techniques) தரவுசார் பகுப்பாய்வு நுட்பங்களையும் (qualitative techniques) சார்ந்துள்ளது என பைபர் தெளிவாக விவரிக்கின்றார். (Biber et al., 1998)

ஒற்றைப் பனுவலொன்றைப் படிப்பதற்கும், பனுவல் தொகுப்பொன்றை (தரவகம்) ஆய்வு செய்வதற்கான தரவக மொழியியல் கருவிகளைப் பயன்படுத்துவதற்கும் இடையே வேறுபாடுகள் உள்ளன.

ஒரு தரவகத்தில் உள்ள பனுவல்கள், ஒரு ஒற்றைத் தலையங்கத்தைக் கிடைநிலையில் (horizontal) படிப்பதுபோல தொடக்கத்திலிருந்து இறுதிவரை படிக்கப்படுவதில்லை; மாறாக, பனுவல்கள் யாவும் தொடர்புடைய ஆனால் வெவ்வேறு நிகழ்வுகளின் தொகுப்பாகும், மேலும் அவை முழுவதுமாக ஆராயப்படாமல், பகுதிகளாக ஆராயப்படுகின்றன, ஒரு ஒற்றைக்கூறின் பல எடுத்துக்காட்டுகள்

ஒன்றுக்கொன்று தொடர்புடையவையாகக் காணப்படுகின்றன. இந்த வகையில், தரவகம் கிடைநிலையாகப் படிக்கப்படாமல் செங்குத்தாகப் (vertically) படிக்கப்படுகிறது. ஒரு குறிப்பிட்ட மொழிக் கூறின் பல எடுத்துக்காட்டுகள் ஒரே நேரத்தில் ஆய்வு செய்யப்படுகின்றன (Crawford & Csomay, 2015).

தரவகமொன்றில் உள்ள மாதிரிகள் நம்பகமான, ஒத்திசைவான பனுவலாக இருப்பதால், ஆராயப்படும் மொழியின் கருத்துப்பரிமாற்றச் செயல்பாடுகளைப் புரிந்து கொள்ள இது உதவும். தரவக மொழியியலில் பொதுவாக கருத்துப்பரிமாற்றத்தின் மொழிசார் பகுதி (பனுவல்/உரை) மட்டுமே படிக்கப்பட்டாலும், தரவகத்தில் ஒரு கருத்துப்பரிமாற்ற நிகழ்வின் மொழிசாராப் பகுதியை சேர்க்க எந்தவொரு தடையும் இல்லை.

தரவகங்கள் எந்த முறையில் உருவாக்கப்பட்ட மொழியையும் குறியாக்கம் செய்யலாம் என்று மெகெனரி மற்றும் ஆர்டி கூறுகிறார்கள் (McEnery & Hardie, 2013). எடுத்துக்காட்டாக, பேச்சு வழக்குத் தரவகங்களும், எழுத்து வழக்குத் தரவகங்களும் உள்ளன. கூடுதலாக, சைகை போன்ற துணைமொழியியல் பகுதிகளைச் சில காணொளித் தரவகங்கள் பதிவுசெய்துள்ளன (Knight et al., 2009). மேலும், சைகை மொழியின் தரவகங்களும் உருவாக்கப்பட்டுள்ளன (Crasborn & Zwitserlood, 2008; Schembri & Johnston, 2017).

மேலும், ஒரு குறிப்பிட்ட மொழியின் தரவகத்திற்கான மாதிரிகளைத் தேர்ந்தெடுக்கும்போது, தேவையான மீத்தரவு (metadata) என்று அழைக்கப்படும் ஆசிரியர், பாலினம், வயது, கல்வி நிலை, களம், நடை, பேச்சுச் செயல், கருத்தாடல் வகைகள் போன்ற அனைத்து வெளிப்புறத் தகவல்களும் மாதிரிகளின் பனுவல் பகுதியுடன் சேர்த்துச் சேமிக்கப்படுகிறது. ஆய்வாளர்கள் தங்கள் மொழியியல் பகுப்பாய்வில் சில சூழல்சார்ப் பொருளியல் கூறுகளைச் (pragmatic aspects) சேர்க்க இது உதவும்; இது இப்பகுப்பாய்வைக் கருத்தாடல் ஆய்வாக (discourse analysis) மாற்றும்.

தரவக மொழியியலில் இரண்டு அணுகுமுறைகள்

தரவக மொழியியலில், இரண்டு அணுகுமுறைகள் உள்ளன: ஒன்று தரவகம்சார் அணுகுமுறை (Corpus-based approach), மற்றொன்று தரவக இயக்கு அணுகுமுறை (Corpus-driven approach) என்று அழைக்கப்படுகிறது. டோக்னினிபோனெல்லி (Tognini-Bonelli, 2001), பயனுள்ளதொரு வகையில், தரவகம்சார் அணுகுமுறையையும், தரவக இயக்கு அணுகுமுறையையும் வேறுபடுத்துகிறார்கள். முந்தையதில், தரவக மொழியியல் என்பது ஒரு நெறிமுறையாகக் கருதப்படுகிறது (McEnery & Wilson, 2003). இதில் தரவக மாதிரிகள் பயன்படுத்தப்பட்டுத் தற்போதுள்ள மொழிக் கோட்பாடுகள் சரிபார்க்கப்படுகின்றன.

இதற்கு மாறாக, தரவக இயக்கு அணுகுமுறைகள், தரவக மொழியியலை ஒரு கோட்பாடாகப் பார்க்க முனைகின்றன, இது ஒரு குறுகிய பொருளில், பொருண்மை உருவாக்கத்தையும், பரந்த பொருளில், மொழிப் பயன்பாட்டின் வெவ்வேறு கூறுகளையும் புதியதொரு கோணத்தில் பார்க்க வைக்கிறது (Stubbs, 1993; Teubert, 2005).

தரவகம்சார் அணுகுமுறையை விரும்பும் மொழியியலாளர்கள் தரவக மொழியியலை ஒரு கருவியாகப் (as a tool) பார்க்கிறார்கள்; தரவக இயக்கு அணுகுமுறையை ஏற்பவர்கள் தரவக மொழியியலை மொழிக் கோட்பாடாகக் (as theory of language) கருதுகின்றனர் (Cheng, 2013).

மாற்றிலக்கண மொழியியலாளர்கள் உட்பட, மொழியியலில் வெவ்வேறு கருத்தியல்களைச் சேர்ந்த மொழியியலாளர்கள் தங்கள் மொழியியல் கோட்பாடுகளைச் சரிபார்க்க அல்லது நிறுவவே தரவகத்தைப் பயன்படுத்துகின்றனர். இது தரவகம்சார் அணுகுமுறை என்று அழைக்கப்படுகிறது.

ஆனால், பெர்தியன் கருத்தியலை ஆதரிக்கும் சின்க்ளேர் போன்ற மொழியியலாளர்கள், தரவகம்சார் அணுகுமுறையை ஆதரிப்பவர்கள் அல்ல, மாறாக தரவகஇயக்கு அணுகுமுறையை ஆதரிப்பவர்கள். மொழிகள்குறித்த புதிய உண்மைகளைப் பெற அல்லது முன்மொழிய அவர்கள் தரவகத்தைப் பயன்படுத்த விரும்புகிறார்கள். தரவகப் பகுப்பாய்வின்போது எந்தவொரு முன்கருத்தையும் பயன்படுத்தக்கூடாது என்பது அவர்களின் வாதம்.

இலக்கண வகை வகைப்பாடுகூட தற்போதுள்ள ஒன்றாக இருக்கக்கூடாது: அவை தரவக மாதிரிகள் மீது திணிக்கப்படக்கூடாது என்று சின்க்ளேர் இதற்கு ஒருபடி மேலே செல்கிறார். அதைத் தரவகப் பகுப்பாய்வின் மூலமே பெறவேண்டும் என்கிறார்.

தரவகங்களின் வகைகள்

1. கண்காணிப்புத் தரவக அணுகுமுறை: இங்குத் தரவகம் தொடர்ந்து விரிவடைந்து, காலப்போக்கில் மேலும் மேலும் பனுவல்களைச் சேர்த்துக் கொண்டேயிருக்கும், குறிப்பாக அமெரிக்க COCA தரவகத்தை குறிப்பிடலாம் .

2. சமச்சீரான தரவகம் அல்லது மாதிரித் தரவக அணுகுமுறை: இது, குறிப்பிட்டதொரு காலத்தில் நிலவும் மொழியைப் பிரதிபலிக்கிறதொரு மாதிரித் தரவகம்.

ஆய்வாளரின் ஆராய்ச்சிக் கேள்வியைப் பொறுத்து, இரண்டு அணுகுமுறைகள் தோன்றியுள்ளன (Sinclair & Sinclair, 1991). தரவகங்களின் நோக்கங்களையும், இலக்குகளையும், குறிக்கோள்களையும் பொறுத்து, அவை பல்வேறு வகைகளாகப் பகுக்கப்படலாம்.

அட்டவணை 1: தரவகங்களின் வகைகள்.

எண்	தமிழ்	விளக்கம்	ஆங்கிலம்
1	பொது அல்லது நோக்கீட்டுத் தரவகம்	பொதுவாக, ஒரு மொழியின் அனைத்து மொழிக் கூறுகளையும் மொழிசாராக் கூறுகளையும் பகுப்பாய்வு செய்து புரிந்து கொள்வதற்கான தரவகம்.	General or Reference Corpus
2	சிறப்புக் குறிக்கோளுக்கான தரவகம்/ தனித்துவத் தரவகம்	அறிவியல் பனுவல் தரவகம், இலக்கியத் தரவகம், மொழி கற்பித்தல்/கற்றலுக்கான தரவகம் போன்ற குறிப்பிட்ட தொரு துறையின் குறிப்பிட்ட கூறுகளைப் புரிந்து கொளவதற்கான தரவகம்., இங்கே, பொதுவான மொழியியல் கூறுகளிலிருந்து வேறுபட்ட சொல், சொற்றொடர், பிற பனுவல் நிலைகள்போன்ற குறிப்பிட்ட மொழிக் கூறுகளைப் புரிந்து கொள்வதே குறிக்கோளாகும்.	Special purpose or Specialized Corpus
	ஒப்பீட்டுத் தரவகம்	மொழிகளுக்கு இடையே உள்ள ஒற்றுமைகளைப் பகுப்பாய்வு செய்து புரிந்துகொள்ள, இந்த வகைத் தரவகங்கள் பயனுள்ளதாக இருக்கும்.	Comparable Corpus

	இணையொத்த தரவகம்	தானியங்கி இயந்திர மொழிபெயர்ப்பு, லெக்சிகன் மேம்பாடு போன்ற குறிப்பிட்ட பயன்பாடுகளில் பயன்படுத்த, இந்த வகை கார்போரா பயனுள்ளதாக இருக்கும்	Parallel Corpora
	கற்போர் தரவகம்	இது, பனுவல்களின் தொகுப்பு எடுத்துக்காட்டாக ஒரு மொழியைக் கற்போரால் எழுதப்படும் கட்டுரைகள், இந்தத் தரவத்தின் நோக்கமானது, கற்பவர்கள் ஒருவருக்கொருவர் எந்தெந்த வகைகளில் வேறுபடுகிறார்கள் என்பதையும், தாய் மொழி பேசுபவர்களின் மொழியிலிருந்து எந்தெந்த வகைகளில் வேறுபடுகிறார்கள் என்பதையும், கண்டறிவதாகும், இதற்கெனத் தாய்மொழி பேசுவோர் பனுவல்களுடன் ஒப்பிடக்கூடிய தரவகமொன்று தேவைப்படுகிறது.	Learner corpus
	வரலாற்றுத் தரவகம் அல்லது இருகாலத் தரவகம்	இது, வெவ்வேறு காலகட்டங்களில் உள்ள பனுவல்களின் தரவகம். இது, காலப்போக்கில் ஒரு மொழியினது கூறுகளின் வளர்ச்சியைக் கண்டறியப் பயன்படுகிறது	(Hunston, 2008). Historical or diachronic Corpus

தரவக உருவாக்கத்தின் நோக்கத்தைப்பொறுத்து ஒருமொழித் தரவகம் அல்லது பன்மொழித் தரவகம் இருக்கலாம்.

தரவக வடிவமைப்பும் கோட்பாடுகளும்

ஒரு தரவகம் பொதுவாக ஒரு பனுவல் தரவகத்தைக் குறிக்கிறது, இது 'மின்னணு வடிவில் உள்ள மொழிப் பனுவல் பகுதிகளின் தொகுப்பு என வரையறுக்கப்படுகிறது. இது, முடிந்தவரை, மொழியியல் ஆராய்ச்சிக்கான தரவுகளின் ஆதாரமாக ஒரு மொழி அல்லது மொழி வகையை பிரதிநிதித்துவப் படுத்துவதற்கென வெளிப்புற வரன்முறைகளுக்குத் தக்கவாறு தேர்ந்தெடுக்கப்பட வேண்டும் என்கிறார் சின்க்ளேர் (Sinclair, 2005).

இது மொழியியல் பகுப்பாய்விற்குப் பயன்படுத்தப்படும் வகையில், ஆய்வுக்கு எடுத்துக்கொண்ட மொழியைப் பிரதிநிதித்துவப்படுத்துவதாகக் கருதப்படும் பனுவல்களின் தொகுப்பு என்கின்றனர் டோக்னினிபோனெல்லி (Tognini-Bonelli, 2001).

ஹன்ஸ்டன் அவர்கள் (Hunston, 2002) தரவகம் என்பது மின்னணு முறையில் சேமிக்கப்பட்டுப் பயன்படுத்துவதற்கான பனுவல்களின் தொகுப்புகள் அல்லது பனுவலின் பகுதிகள் என்கிறார். ஆயினும் தரவகம் என்பது வெறும் பனுவல்களின் தொகுப்பு அல்ல; அது இயற்கையாக இடம்பெறும் மொழிப் பனுவல்களின் தொகுப்பாகும், இது ஒரு மொழியின் வகை/பண்பை விரித்துக்காட்டுதற்கெனத் தேர்ந்தெடுக்கப்பட்ட முறை என்கின்றனர் சின்க்ளேர் மற்றும் செங் (Cheng, 2011, Sinclair, 1991).

மேலே உள்ள விவாதங்களின் அடிப்படையில், ஒரு மொழியின் தரவகமானது, சில திட்டவட்டமான

வடிவமைப்புக் கொள்கைகளைக் கொண்டு சேகரிக்கப்பட்ட மாதிரிகளைக் கொண்டுள்ளது என்று முடிவு செய்யலாம். இது வெறும் மின்னணுப் பனுவல்களின் தொகுப்பு அல்ல என்பது மிக முக்கியமான கூற்றாகும். மின்னணுப் பனுவல்கள் ஆதாரங்களாகப் பயன்படுத்தப்படலாம். இந்த ஆதாரங்களிலிருந்து மாதிரிகள் சேகரிக்கப்பட வேணடும்.

தரவக மொழியியலின் முன்னோடிகளில் ஒருவர் என அழைக்கப்படும் சின்க்ளேர், தரவக வடிவமைப்பிற்கான முக்கியக் கொள்கைகளைப் பின்வருமாறு வழங்கியுள்ளார்.

1. தரவக உள்ளடக்கங்கள் அவை கொண்டிருக்கும் மொழியைப் பெரிதுபடுத்தாமல் சமூகத்தில் அவற்றின் கருத்துப்பரிமாற்ற நோக்கத்தின் அடிப்படையில் தேர்ந்தெடுக்கப்படுகின்றன.

2. வெளிப்புற வரன்முறைகளுக்கு ஏற்ப தரவகத்தில் உள்ள கருப்பொருளின் கட்டுப்பாடு அமைகிறது, உட்புற வரன்முறைகள் ஏற்ப அமைவது இல்லை.

3. வேறுபடுத்துவதற்காக தனியாக வடிவமைக்கப்பட்டுள்ள தரவகத்தின் கூறுகள் மட்டுமே, வேறுபடுத்தப்படுகின்றன அல்லது அதாவது நெறிப்படுத்தப் படுகின்றன.

4. தரவகத்தின் அமைப்பை நிர்ணயிக்கும் வரன்முறைகள் குறைவாகவே உள்ளன, ஒன்றுக்கொன்று தனித்தனியாக உள்ளன; இவை பிரதிநிதித்துவத் தரவகத்தை (corpus that is representative) வரையறுப்பதில் சிறப்பு வாய்ந்ததாகவும் இருக்கும்.

5. தரவகத்திற்கான மொழியின் மாதிரிகள், முடிந்தவரை, முழுப் பனுவல்களைக் கொண்டிருக்கும்.

6. சொல்வகைப்பாட்டு இலக்கணக் குறியீடுகள், அச்சுருக்கள் (part-of-speech tags), அச்சிடப்பட்ட ஆவணத்தின் வடிவமைப்பு போன்ற பனுவலொன்றின் எந்தவொரு தகவலும், எளிய பனுவலிலிருந்து (அதாவது, பனுவலின் சொற்கள் மற்றும் நிறுத்தற்குறிகள்) தனியாகச் சேமிக்கப்பட்டு, தேவைப்படும்போது மட்டுமே இணைக்கப்பட வேண்டும்.

7. தரவகத்தின் வடிவமைப்புகளும் உட்கூறுகளும் முழு நியாயத்துடன் (justifications) முழுமையாக ஆவணப் படுத்தப்பட்டுள்ளது.

8. தரவக வடிவமைப்பில் இலக்குக் கருத்துக்களாக (target notions), பிரதிநிதித்துவமும் (representativeness) சமச்சீர்மையும் (balance) அடங்கும்.

9. போதுமான உள்ளடக்கங்களைப் பராமரிக்கையில், தரவகமானது, தன் உட்கூறுகளில் நிலைத்தன்மையை நோக்கமாகக் கொண்டுள்ளது சின்க்ளேர் (Sinclair, 2005).

நன்கு உருவாக்கப்பட்ட தரவகத்தின் முக்கியக் கூறுகள்:

1) தரவகத்தின் அளவு: ஒரு மொழியின் தரவகம் மேலும் பகுப்பாய்வு செய்வதற்குப் பெரியதாக இருக்க வேண்டும். சொற்கள் மில்லியன் அல்லது பில்லியன்களில் இருக்க வேண்டும். இப்போது கணினி தொழில்நுட்ப வளர்ச்சியின் காரணமாகப் மிகப் பெரும் தரவுகளைச் சேமிக்கமுடிகிறது.

2) பிரதிநிதித்துவம்: ஒரு மொழியின் அனைத்து மொழிசார், மொழிசாராக் கூறுகளை மாதிரிகள், பிரதிநிதித்துவப் படுத்தவேண்டும். அனைத்து மொழியியல் கூறுகளும் மொழியின் ஒலியியல், உருபனியல், தொடரியல் கூறுகளும் உருவாக்கப்பட்ட தரவகத்தில் பிரதிநிதித்துவப் படவேண்டும். அதாவது, ஒரு மொழியின் தரவகமானது, அந்த மொழியின் பிரதிநிதியாக இருக்கவேண்டும். தரவகத்தில் எதையும் தவறவிடக்கூடாது.

லீச்சின் (Leech, 1991) கருத்துப்படி, ஒரு தரவகமானது, அதன் உள்ளடக்கங்களை அடிப்படையாகக் கொண்ட கண்டுபிடிப்புகள், அந்த மொழி வகைக்கு பொதுமையாக்கப்பட்டால், அது பிரதிநிதித்துவப்படுத்தவேண்டிய மொழி வகையின் பிரதிநிதியாகக் கருதப்படுகிறது.

பைபர் (Biber, 1993) இந்தத் தரமானது எவ்வாறு எட்டப்படுகிறது என்ற கண்ணோட்டத்திலிருந்து பிரதிநிதித்துவத்தை வரையறுக்கிறார். பிரதிநிதித்துவம் என்பது, இனத்தொகுதியில் உள்ள அனைத்து வகை மாறுபாட்டையும எந்த அளவிற்கு உள்ளடக்கியுள்ளது என்பதைக் குறிக்கிறது. ஒரு தரவகமானது அடிப்படையில் ஒரு மொழி அல்லது மொழி வகையின் மாதிரியாகும் (McEnery et al., 2006).

சமச்சீரான ஒன்று

தரவகத்திற்காகச் சேகரிக்கப்பட்ட மாதிரிகளானவை, சமச்சீரான ஒன்றாக இருக்கவேண்டும். அதாவது, நூல்கள், கட்டுரைகள், இதழ்கள் போன்ற அனைத்து மொழி வளங்களிலிருந்தும் மாதிரிகள் சேகரிக்கப் படவேண்டும் அல்லது தேர்ந்தெடுக்கப்பட வேண்டும். வளங்களின் எண்ணிக்கை அல்லது அளவைக் கணக்கில் கொண்டு, களங்களிலும் (domains) நடைகளிலும், மாதிரிகள் அவற்றின் விகித அளவின் அடிப்படையில் சேகரிக்கப்படவேண்டும்.

மாதிரி முறை

மாதிரிக்காக நாம் கண்டறிவது, பொது இனத்தொகுதிக்கும் பொருந்திவந்தால், ஒரு மாதிரியானது பிரதிநிதியாகக் கருதப்படும். புள்ளிவிவரப்பொருளில்,மாதிரிகள் ஒருபெரிய இனத்தொகையின் சுருங்கிய பதிப்புகளாகும். மாதிரிக் கோட்பாட்டின் நோக்கமானது, அளவின் வரம்புகளுக்கு உட்பட்டு, இனத்தொகையின் சிறப்பியல்புகளை மீட்டுருவாக்கும் மாதிரியைப் பெறுவதாகும். குறிப்பாக, உடனடி ஆர்வமுள்ளவற்றை முடிந்தவரை பெறுவதாகும் (McEnery & Wilson, 1996). மாதிரி முறையானது, புறவயமானதொன்றாக இருக்கவேண்டும். ஆதாரங்களின் எந்தவொரு பகுதிக்கும் பாரபட்சமான அல்லது அகவயமான முன்னுரிமை கொடுக்கக்கூடாது.

தரவக இலக்கணக் குறிப்பீடு

மாதிரிப் பனுவல்களுக்கு வழங்கப்படும் இலக்கணக் குறிப்பீடானது, தரவகத் திட்டங்களில் மற்றொரு முக்கியமான கூறு ஆகும். "குறிப்பாகச் சொல்வதென்றால், ஒரு தரவகம் தானாகவே எதுவும் செய்யமுடியாது; பயன்படுத்தப்பட்ட மொழிச் சேமிப்புக்கிடங்குதானேதவிர வேறொன்றும் இல்லை. இருப்பினும், தரவக அணுக்க மென்பொருளானது, அந்தச் சேமிப்பை மறுசீரமைப்பதன் மூலம் பல்வேறு வகையான ஆய்வுகளை மேற்கொள்ள முடியும்.

ஒரு தரவகமானது, அம்மொழி பேசுவோரின் மொழி அனுபவத்தை, மிகத் தோராயமாகவும், பகுதியளவும் பிரதிநிதித்துவப்படுத்தினால், அணுக்க மென்பொருளானது, அந்த அனுபவத்தை மறுசீரமைக்கிறது, இதைக்கொண்டு பொதுவாகச் சாத்தியமில்லாதவற்றைக்கூட ஆய்வு செய்யலாம். ஒரு தரவகத்தில் மொழிபற்றிய புதிய தகவல்கள் ஏதும் இல்லை; ஆனால் அந்த மென்பொருளானது, நமக்கு ஏற்கனவே தெரிந்ததைப் புதிய கோணத்தில் பார்க்கவைக்கிறது.

மிகவும் எளிதில் கிடைக்கக்கூடிய மென்பொருள் தொகுப்புகள், "தரவகத்திலுள்ள தரவுகளை மூன்று வழிகளில் செயற்படுத்துகின்றன அவை அதிர்வெண் (frequency), சொற்றொடர் (phraseology), சேகரிப்பு (collection) ஆகியவற்றைக் காட்டுகிறது (Hunston, 2008, 2022). அந்த மொழியைப்பற்றிய தகவலைப் பெறப் பகுப்பி(Parsers), சொல்லடைவி (Concordancer), சொல்லிணைவி (Collocation) போன்ற பொருத்தமான மென்பொருள் கருவிகளைக் கொண்டு, மாதிரிகள் பகுப்பாய்வு செய்யப்படவேண்டும். உருபனியல், தொடரியல், பிற மொழிக் கூறுகள் போன்ற இத்தகைய தகவல்களைக் கொணடு மாதிரிகளுக்கு இலக்கணக்

குறிப்பிட வேண்டும்.

தரவக இலக்கணக் குறிப்பீட்டு வகைகளைப்பற்றி, மெக்கன்ரே - ஹார்டி பின்வருவனவற்றைக் கூறுகிறார்கள்:

> "தரவகமானது, பொதுவாகத் தரவகத்தில் உள்ள தரவுகளின் ஆய்வுக்கு உதவும் மூன்று வகைத் தவல்களைக் கொண்டுள்ளது: மீத்தரவு (metadata), பனுவல் குறிமொழி (textual markup), மொழியியல் இலக்கணக் குறிப்பீடு (linguistic annotation)" (McEnery & Hardie, 2011).

மீத்தரவு என்பது பனுவல் பற்றி உங்களுக்குச் சிலவற்றைத் தரும் தகவலாகும். எடுத்துக்காட்டாக, எழுத்துப் பனுவலில், அதை எழுதியவர் யார், எப்போது வெளியிடப்பட்டது, எந்த மொழியில் எழுதப்பட்டது என்பதைப்பற்றிய தகவல்களையெல்லாம் மீத்தரவு தரும். தரவகப் பனுவலிலோ, தனி ஆவணத்திலோ, தரவுத்தளத்திலோ மீத்தரவைக் குறியாக்கம் செய்யலாம். உண்மையான சொற்களைத் தவிர பிற பனுவலுக்குள் உள்ள தகவல்களைக்கூட பனுவல் குறிமொழி குறியாக்கம் செய்கிறது.

எடுத்துக்காட்டாக, அச்சிடப்பட்ட எழுத்துப் பனுவலில், சாய்வு எழுத்துக்கள் எங்கு தொடங்கி எங்கு முடிவடையும் போன்ற பனுவலின் வடிவமைப்பைக் குறிக்கப் பனுவல் குறிமொழியானது பொதுவாகப் பயன்படுத்தப்படும். ஒலிபெயர்க்கப்பட்ட பேச்சுத் தரவகங்களில், மீத்தரவு மூலமும், பனுவல் குறிமொழி மூலமும் கிடைக்கும் தகவல்கள், ஒலிபெயர்ப்பின் பகுப்பாய்விற்கு மிகவும் முக்கியமானதாக இருக்கலாம். ஒரு தரவகப் பனுவலுள் மொழியியல் தகவலைக் குறியாக்கம் செய்யமுடியும், இதனால் அந்த பகுப்பாய்வைப் பின்னர் முறையாகவும் துல்லியமாகவும் மீட்டெடுக்கமுடியும்.

மீத்தரவு மூலம் நாம் பெறும் தகவலானது, "தரவகப் பயனர்கள் தங்களது கண்டுபிடிப்புகளைப் பொருளுணரவைப்பதற்கும் விளக்குவதற்கும் மிகவும் பயனுள்ளதாக இருக்கும்" (வின்னி செங், 2012,. ப.4).

மேலும், மெக்கன்ரே - ஆண்ட்ரூ ஹார்டி (2012,. ப.31), தரவக ஆய்வாளர்களுக்கு தெளிவாக ஒரு விளக்கத்தை கொடுக்கின்றனர், அதாவது "ஒரு தரவகமானது, மொழியியல் இலக்கணக் குறிப்பீடுகளை உள்ளடக்கும்போது, அந்த இலக்கணக் குறிப்பீடுகளைப்பற்றி என்ன சொல்லமுடியும் என்பதையும், என்ன சொல்லமுடியாது என்பதையும் குறிப்பிடுவது முக்கியம்."

மிக முக்கியமாக, தரவகமானது புதிய தகவல்களைக் கொண்டுள்ளது என்று நாம் கூற முடியாது; அது நிச்சயமாகப் புதுத் தகவல்களைக் கொண்டிருக்காது. இந்த வகையான மொழியியல் பகுப்பாய்வு என்ன செய்கிறது என்றால், அது தரவிலே உறைந்துள்ள உள்ளார்ந்த தகவலை வெளிக்கொணரவைக்கிறது.

ஒரு சொல்லைப் பெயர்ச்சொல் அல்லது வினைச்சொல் என்று அடையாளம் காண்பது என்பது, அவ்வாறு செய்வதன்மூலம் அதை ஒரு பெயர்ச்சொல்லாக மாற்றுகிறோம் என்று பொருளல்ல. தரவக இலக்கணக் குறிப்பீட்டில், நாம் பெயரீட்டுச் செயற்பாட்டில் ஈடுபடுகிறோமே தவிர, எதையும் உருவாக்கவோ, மாற்றவோ இல்லை. அந்த அளவில், ஒரு நிரலின் பார்வையில் அல்லது ஒரு பயனரின் பார்வையில், தரவகம் செறிவூட்டப்பட்டுள்ளது என்று கூறலாம், ஆனால் அந்தத் தரவகத்தில் புதிய தகவல்கள் சேர்க்கப்பட்டுள்ளன என்று கூறமுடியாது.

தரவக இலக்கணக் குறிப்பீட்டில் இரண்டு கருத்துகளுக்கு அதிக முக்கியத்துவம் கொடுக்கப் படவேண்டும். ஒன்று, இலக்கணக் குறிப்பீட்டுத் திட்டம் (annotation scheme); மற்றொன்று, இலக்கணக் குறிப்பீட்டுக்குக் கொடுக்கப்பட்ட வடிவமைப்பு (format given to the annotation).

இலக்கணக் குறிப்பீட்டுத் திட்டம்

முதலாவது கருத்து, எந்த வகையாக இருந்தாலும், எந்த அளவினதாக இருந்தாலும், எந்தவொரு தரவக இலக்கணக் குறிப்பீட்டுத் திட்டத்தினது முக்கிய பகுதியும், இலக்கணக் குறிப்பீட்டுத் திட்டமாகும். முதலாவதும் மிக முக்கியமானதும் எதுவென்றால், இலக்கணக்

குறிப்பீட்டுத் திட்டமானது, இலக்கணக் குறிப்பீட்டுச் செயற்பாட்டின்போது வேறுபடுத்தப்படவேண்டிய மொழியியல் வகைப்பாடுகளைப்பற்றிய வெளிப்படையான, முழுமையான தகவல்களைக் கொண்டிருக்கவேண்டும். இத்தகைய வகைப்பாடுகள், இப்போது உள்ள இலக்கணக் குறிப்பீட்டு வகையைச் சார்ந்தனமட்டுமல்ல; இலக்கணக் குறிப்பீட்டு நோக்கத்தின் அடிப்படையில் விரும்பப்படும் அல்லது தேவைப்படும் குறிப்பிட்ட அளவையும் சார்ந்துள்ளன.

இலக்கணக் குறிப்பீட்டுத் திட்டத்தின் இரண்டாவது பகுதியானது, பெயரீடுகளுக்கும் வகைப்பாடுகளுக்கும் இடையிலான தொடர்போடு மொழியில் வகைப்பாடுகளைக் குறிக்க வடிவமைக்கப்பட்ட பெயரீடுகளின் (இலக்கணக் குறியீடுகள், குறிமுறைகள், இடுகுறிகள்) தொகுப்பாகும். இந்தப் பெயரீடுகள் சுருக்கமாகவும் உள்ளுணர்வுடன்கூடிய பொருளமைந்ததாகவும் இருக்கவேண்டும். மேலும், ஒரு பொது வகையின் துணைப்பிரிவுகளுக்கிடையில் வேறுபாடுகள் ஏற்படும்போதெல்லாம், அந்தத் துணைப்பிரிவுகளுக்கு இடையிலுள்ள பொதுமைகளைக் கண்டுபிடிப்பதற்கென, இந்தப் பெயரீட்டு அமைப்பு வடிவமைக்கப்படவேண்டும்.

இறுதியாக, இலக்கணக் குறிப்பீட்டுத் திட்டமானது, திட்டத்தில் வரையறுக்கப்பட்டுள்ள மொழியியல் வகைகளுக்கு, தரவகத்தில் உள்ள பல்வேறு மொழியியல் அலகுகள் எவ்வாறு வடிவமைக்கப்படவேண்டும் என்பதை விளக்கும் வழிகாட்டுதல்களின் தொகுப்பையும் கொண்டிருக்கவேண்டும், பின்னர்த் தொடர்ந்து பொருத்தமான பெயரீடுகளுடன் இலக்கணக் குறிப்பீடு இடப்படவேண்டும்.

இலக்கணக் குறிப்பீட்டு வடிவமைப்பு

தரவக இலக்கணக் குறிப்பீட்டுத் திட்டத்தில் இரண்டாவது முக்கியமான கருத்தாகக் கருதப்படுவது, இலக்கணக் குறிப்பீட்டில் பயன்படுத்தப்போகும் வடிவமைவாகும். மூலத் தரவகத்தில் பொருத்தமான மொழியியல் அலகுகளுக்கு எவ்வாறு பெயரீடுகள் பயன்படுத்தப்பட வேண்டும் என்பதை இது மேற்கொள்ளவேண்டும். ஏற்றுக்கொள்ளப்பட்ட எந்தவொரு வடிவமாக இருந்தாலும், இலக்கணக் குறிப்பீடுகள், மூலத் தரவகத்திலிருந்து எளிதில் பிரிக்கக்கூடியதாக இருக்கவேண்டும் என்று பரிந்துரைக்கப்படுகிறது. அதாவது, வேறு வார்த்தைகளில் கூறுவதானால், ஒரே நேரத்தில் மூல நூல்களையும் அவற்றின் இலக்கணக் குறிப்பீடுகளையும் ஆராய்வதுமட்டுமல்லாமல், பனுவல்களை அவற்றின் இலக்கணக் குறிப்பீடுகளிலிருந்து பிரித்துத் தனித்தனியாக ஆராயவும் முடியும் ஜியாபேய் லூ (Lu, *2014*).

தரவகச் சேமிப்பு

தரவகத் திட்டத்தின் முடிவில், மில்லியன்கணக்கான சொற்களைக்கொண்ட மீத்தகவலுடன் கூடிய எளிய பனுவல் கோப்புகளை ஆராய்ச்சியாளர்கள் பெறுவார்கள்.

XML கோப்பு, SQL கோப்பு போன்ற சில பொருத்தமான கோப்பு வடிவங்களில் மற்றொரு மொழியியல் குறியீடு பெற்ற கோப்புகளையும் ஆராய்ச்சியாளர்கள் பெறுவார்கள். இத்தகைய கோப்புகள் பின்னர் கணினி மூலம் செய்யப்படும் ஆய்வுகளுக்கும் மென்பொருள் மேம்படுத்துவதற்கும் பெரும் துணையாக இருக்கும்.

தகவல் மீட்பு

தரவகத் திட்டமானது, சில மாதிரிகளுடன் சேமிக்கப்பட்ட தகவலை மீட்டெடுக்கச் சில பயனர் இடைமுக மென்பொருளைக் (user interface software) கொண்டிருக்கும். சொற்களின் பட்டியல், அவற்றின் அதிர்வெண, தரவரிசை, சொல் வகைப்பாடு, தொடரியல் கூறுகள் (Syntactic features) போன்றவற்றை பயனர் இடைமுகக் கருவிகள் மூலம் மீட்டெடுக்கலாம்.

மேலும், சொல்லடைவு (சூழலின் முக்கியச் சொற்கள்), சொல்லிணைவு (சொற்களுக்கு இடையே உள்ள ஈர்ப்பு) இலக்கணவிணைவு பற்றிய (இலக்கண வகைகளின் அருகாமை

பற்றிய) தகவல்களை ஆராய்ச்சியாளர்கள் பெறுவார்கள்.

மலேசியாவின் தற்கால எழுத்துத் தமிழுக்கான தரவக உருவாக்கமும், பகுப்பாய்வும் தேவை

தற்போது, மலேசியாவின் தற்கால எழுத்துத் தமிழுக்கான தரவகம் எதுவும் கிடைக்கவில்லை. மலேசியத் தமிழுக்குத் தற்சமயம் கிடைக்கும் இலக்கணங்களாவும் பண்டைய இலக்கண நூல்களின் அடிப்படையில் எழுதப்பட்டவை இந்நூல்கள் பல்லாயிரம் ஆண்டுகாளமாக தமிழின் கட்டமைப்பை உறுதி செய்து வருகின்றன. மலேசியாவின் எழுத்து தமிழிலும் பேச்சு தமிழிலும் மாற்றங்கள் ஏற்பட்டுள்ளன. மலேசியாவின் தற்கால எழுத்துத் தமிழுக்குத் மேற்குலகு அறிஞர்கள் செய்கின்ற தரவகம்சார் ஆய்வுகள் அவசியமாகிறது.

ஒரு மொழியின் அமைப்பு, செயல்பாடு (structure and function) ஆகியவற்றின் ஒழுங்குமுறையைப் புரிந்துகொளவதற்குத் தற்கால எழுத்துத் தமிழைப் பற்றிய விரிவான ஆய்வு தேவைப்படுகிறது, இந்த நோக்கத்திற்குத் தரவகம்சார் அணுகுமுறை பொருத்தமானதாக இருக்கும். இதற்காக மலேசியத் தமிழ்ப் பேச்சுச் சமூகத்தின் முழு மக்களையும் பிரதிநிதித்துவப்படுத்தக்கூடிய ஒரு சமச்சீரான தரவகத்தை உருவாக்குவது மிகவும் அவசியம்.

தமிழ்ச் சொற்களின் அமைப்பையும் செயல்பாட்டையும் முழுமையாகப் புரிந்துகொள்வதே இந்த ஆராய்ச்சியின் பின்னால் இருக்கும் ஆராய்ச்சிக் கேள்வியாக இருக்கும். தமிழுக்கு உள்ள தேவையான மென்பொருள் கருவிகளைக்கொண்டு பகுத்தல் (parsing), சொல்வகைப்பாட்டுக் குறியீடு (part-of-speech tagging) உள்ளிட்ட தேவையான உருபனியல் பகுப்பாய்வுகள் (morphological analyses) மேற்கொள்ளப்படும். சொல் வகைகளின் அதிர்வெண் (frequency), சொல் வகை (word types), சொல் வில்லை விகிதம் (type - token ratio) போன்ற சில புள்ளியியல் ஆய்வுகளும் statistical studies மேற்கொள்ளப்படும். சொல்லடைவு (Concordance), சொல்லிணைவு (Collocation), இலக்கணவிணைவு ஆய்வுகள் மேற்கொள்ளப்படும். எனவே மலேசியத் தமிழுக்கான தரவக உருவாக்கமும் பகுப்பாய்வும் காலத்தின் தேவையாக அமைகிறது.

மலேசியத் தமிழ்த் தரவகத்தின் பங்களிப்பு

மேற்கூறிய தரவகம்சார்த் தமிழ் ஆய்வானது, கீழ்க்காணும் வகைகளில் பங்காற்றும்

1. கோட்பாட்டு ஆய்வுகள் (Theoretical studies): மலேசியாவின் தற்கால எழுத்துத் தமிழின் தரவகம்சார் ஆய்வு, தமிழின் தற்போதைய நிலையைப் புரிந்துகொள்ளப் பெரிதும் உதவும். மேலும், தமிழின் புதிய வளர்ச்சிகளைப் புரிந்துகொள்ளவும் உதவும்.
2. கற்பித்தல் நோக்கங்கள் (Pedagogical purposes):

அ. முன்மொழியப்பட்ட ஆய்வானது, தமிழ்ப்பள்ளி மாணவர்களுக்குத் தேவையான கற்பித்தல் / கற்றல் கருவிகளை உருவாக்க உதவும்.

ஆ. தற்கால எழுத்துத் தமிழுக்குக் கற்பித்தல் இலக்கணங்களை எழுத இது உதவும்.

இ. தற்கால எழுத்துத் தமிழுக்கு அகராதிகளை உருவாக்க இது மிகவும் பயனுள்ளதாக இருக்கும்.

ஈ. புதிய சொற்களஞ்சியங்களை உருவாக்குவதற்கும் இது மிகவும் பயனுள்ளதாக இருக்கும்.

3. இயற்கை மொழி ஆய்வு (Natural Language Processing): எழுத்துப்பிழை திருத்திகள், சந்திப்பிழை திருத்திகள்போன்ற பலவேறு மொழித் தொழில்நுட்பப் பயன்பாடுகளுக்கு மிகவும் பயன்படும் தற்காலத் தமிழுக்கான உருபன் பகுப்பிகளை உருவாக்குதல்.

அப்படி ஒரு முழுமைப்பெற்ற தரவகம் உருவாக்கப்பட்டால் நம் மொழி வளர்ச்சிக்கும் மொழி திட்டமிடலுக்கும் பெருந்துணையாக அமையும்.

References

Agasthialingom, S. (2002). *Structure of Tamil Language* (in Tamil). Chidambaram: Meyyappan Publishing, .

Biber, D. (1993). *Representativeness in corpus design. Literary and Linguistic Computing,* 8(4), 243-257. https://doi.org/10.1093/llc/8.4.243

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Cheng, W. (2011). *Exploring corpus linguistics: Language in action*. Routledge

Cheng, Winnie. (2012). *Exploring Corpus Linguistics- Language in Action*. London: Routledge (Taylor and Francis Group),

Crawford, William J & Eniko Csomay. (2016). *Doing Corpus Linguistics*. London: Routledge (Taylor and Francis Group),

Crasborn, O. A., & Zwitserlood, I. E. P. (2008). *The Corpus NGT: an online corpus for professionals and laymen*.

Crawford, W., & Csomay, E. (2015). *Doing corpus linguistics*. Routledge.

De Silva, M.W.S. Diglossia and Literacy. Mysore: Central Institute of Indian Languages, 1976. "Some Consequences of Diglossia." *York Papers in Linguisics* (1974): 71-90.

Deiva Sundaram, N. Diglossic . (1980). *Situation in Tamil (Sociolinguistic Approach)*. Unpublished Ph.D. dissertation. Chennai: University of Madras.

Hunston, Susan. (2002). *Corpora in Applied Linguistics*. London: Cambridge University Press,

Hunston, S. (2002). *Pattern grammar, language teaching, and linguistic variation. Using corpora to explore linguistic variation*.

Hunston, S. (2008). *Starting with the small words: Patterns, lexis and semantic sequences*. International journal of corpus linguistics.

Kothandaraman, Pon. (1997). *A Grammar of Contemporary Literary Tamil*. Chennai: International Institute of Tamil Studies.

Leech, G. (1991). *New resources, or just better old ones? The Holy Grail of representativeness*. In Corpus linguistics and the web

Lehmann, Thomas. (1998*). "Old Tamil." Steever, Sanford. The Dravidian Languages*. London: Routledge,.

Lu, Xiaofei. (2014). *Computational Methods for Corpus Annotation and Analysis*. London: Springer,

McEnery, T., & Wilson, A. (1997). *Teaching and language corpora (TALC*). ReCALL, 9(1), 5-14.

McEnery, A., & Xiao, R. (2003). *The Lancaster Corpus of Mandarin Chinese.*

McEnery, T., & Wilson, A. (2003). *Corpus linguistics. The Oxford handbook of computational linguistics*, pp.448-463.

McEnery, Tony & Richard Xiao & Yukio Tono. (2006). *Corpus-Based Language Studies-an advanced resource book*. London: Routledge (Taylor & Francis Group),

McEnery, T., & Hardie, A. (2011). *Corpus linguistic: Method, theory and practice*. Cambridge University Press.

McEnery, Tony and Andrew Hardie. (2012). *Corpus Linguistics*. London: Cambridge University Press.

McEnery, T., & Hardie, A. (2013). *The history of corpus linguistics. The Oxford handbook of the history of linguistics*, 727-745.

Nuhman, M.A. (2007). *Fundamental Tamil Grammar (in Tamil)* . Puththanathhtam: Adaiyalam.

O'Keeffee, Anne & Michael McCarthy & Ronald Carter. (2007). *From Corpus to Classroom (Language Use and Language Teaching).* London: Cambridge University Press,

Paramasivam Muthusamy, Atieh Farashaiyan. (2016). "Language Change and Maintenance of Tamil language in the." *International Journal of Humanities and Social Science Invention* 55-60.

Rajendiran, N. Balakrishnan Muniappan, Manickam Govindaraju, G. (2012). "Identity and Language of Tamil Community in Malaysia: Issues and Challenges." DOI: 10.7763/IPEDR. . V48.: 17.

Schembri, A., & Johnston, T. (2017). Usage-based grammars and sign languages: Evidence from Auslan, BSL and NZSL. *Julkaisematon käsikirjoitus. Saatavilla Internetissä*: https://www. academia. edu/5244167.

Seeni Naina Mohammad, C. (2013). *Good Tamil Grammar (in Tamil).* Puththanaththam: Adaiyalam.

Siddharththan,(2003). *Singapore Tamil Grammar through easy Tamil*. Chennai: Narmatha Publishing,

Sinclair, J. (2005). *Language as a String of Beads. Strategies in academic discourse*, 19, 163.

Stubbs, M. (1997). *Language and the mediation of experience: Linguistic representation and cognitive orientation. The handbook of sociolinguistics*, 358-373.

Szudarski, P. (2017). *Corpus linguistics for vocabulary: A guide for research*. Routledge.

Teubert, W. (2005). *My version of corpus linguistics. International journal of corpus linguistics*, 10(1), 1-13.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work. Corpus Linguistics at Work*, 1-236.