தரவுத்தொகுதி மொழியியல்

(Corpus Linguistics)

Prof. Rajendran Sankaravelayuthan
Amrita Vishwa Vidyapeetham University
Coimbatore 641 112
rajushush@gmail.com

ABSTRACT

The present monography is written in Tamil and the abstract is given in English for wider users. It is an elaborated version of mine on the same topic uploaded in academia.edu and Researchgate. This monograph was lying in my lap since 2000. I could not find time to update it. Now the present one is an updated version. The monograph contains ten chapters: Chapter 1: General Introduction, Chapter 2: Present Scenario of Corpus Linguistics, Chapter 3: Conceptual Classification of Corpora, Chapter 4: Generation of Written Text Corpus, Chapter 5: Text Corpus Processing, Chapter 6: Corpus in Language Technology, Chapter 7: Corpus in Mainstream Linguistics, Chapter 8: Corpus in Machine Translation, Chapter 9: Potential Utilisation of Corpus, Chapter 10: Concluding remarks. This is followed by an exhaustive list of reference.

Chapter 1 General Introduction:

This chapter contains a brief introduction, an explanation about what is corpus, a brief survey of the pre-electronic corpus; the limitations of the corpus, the salient features of a corpus and the resurgence of corpus linguistics.

Chapter 2: Present Scenario of Corpus Linguistics

This chapter contains a brief introduction, a brief introduction of corpus development, the details about the written corpus and the details about speech corpus.

Chapter 3: Conceptual Classification of Corpora

This chapter contains a brief introduction and the details about the conceptual classification of the corpora based on the genere of text, nature of the data, the type of the text and the nature of the application.

Chapter 4: Generation of Written Text Corpus

This chapter contains a brief introduction, the explanation about size of the corpus, the representativeness of texts, the question of nativity, the determination of target users, the selection of time-Spean, the selection of the text types, the method of the data sampling, the method of the data input, the hardware requirement, the management of the corpus files, the method of corpus sanitation and the problem of copyright.

Chapter 5: Text Corpus Processing

This chapter contains a brief introduction, the explanation about the frequency study, the word sorting, the concordance, the lexical collocation, the key word in context, the local word grouping, the word processing, the tagging, the lemmatization, the annotation and the parsing.

Chapter 6: Corpus in Language Technology

This chapter contains a brief introduction, value of corpus in language technology, corpus as a knowledge resource, for designing language technology tools, as a source for translating support system, as a source for human-machine interference systems and corpus in speech technology.

Chapter 7: Corpus in Mainstream Linguistics

This chapter contains a brief introduction followed by descriptions on corpus in lexicography, corpus in lexicology, corpus in formation of technical terms, corpus in grammar writing, corpus in semantic study, corpus in language learning, corpus in dialect study, corpus in sociolinguistics, corpus in psycholinguistics, and corpus in stylistics.

Chapter 8: Corpus in Machine Translation

This chapter contains a brief introduction followed by descriptions on the objective, lessons leant from history, the corpus-based approach, the issues related with corpus based approach, the generation of translation corpora and the alignment of translation corpora.

Chapter 9: Potential Utilization of Corpus

This chapter gives the details about the potential utilization of the corpus.

Chapter 10: Concluding remarks

தரவுத்தொகுதி மொழியியல்

(Corpus Linguistics)

முனைவர் இராசேந்திரன் சங்கரவேலாயுதன்

பணிநிறைவுற்ற பேரசிரியர்

மொழியியல் துறை, தமிழ்பல்கலைக்கழகம், தஞ்சாவூர்

வருகைதரு பேராசிரியர்

கணினி தொழில்நுட்பம் மற்றும் வலைப்பின்னல் மையம்

அமிர்தா விஷ்வ வித்யாபீடம்

கோயம்பத்தூர்

கோயம்புத்தூர்

2020 ஆகஸ்ட்

என்னுரை

"தரவுத்தொகுதி மொழியியல்" தமிழ்ப்பல்கலைக்கழகத்தில் நான் பணியாற்றிய காலகட்டத்தில் பயன்பாட்டுமொழியியல் முதுகலைப் பட்ட மாணவர்களுக்காக நான் எழுதிய எழுத்து வடிவமாகும். நீண்ட நாட்களாக என் மடிக்கணியில் உறங்கிக்கொண்டிருந்த இந்த வரைவை academia.edu, Research Gate என்ற வலைதளங்களில் பதிவேற்றியிருந்தேன். தற்போது அதை விரிவுபடுத்தி இன்றைய அளவில் வேண்டிய மாற்றங்கள் செய்து பேராசிரியர் திருமலை அவர்களின் Language in India என்ற மின் திங்களிதழில் வெளியிடுகின்றேன். பேராசிரியர் திருமலை அவர்களுக்கு எனது இதயபூர்வமான நன்றியைத் தெரிவித்துக்கொள்கின்றேன். தமிழ்ப்பல்கலைக்கழகத்திலிருந்து ஓய்வு பெற்ற பின்னர் கோயம்பத்தூர் அமிர்தா பல்கலைக்கழகத்தில் கணினித் தொழில்நுட்பம் மற்றும் வலையமைப்பு மையம் (Centre for Computational Engineering and Networking (CEN) என்ற துறையில் சேர்ந்து கடந்த 9 ஆண்டுகளாகப் பணியாற்றி வருகையில் பெற்ற அனுபவமும் காலமும் இந்நூலை முழுமையான மற்றும் விரிவான ஒன்றாக மாற்ற உதவின. இவ்வாய்ப்பை எனக்கு நல்கிய பேராசிரியர் கே.பி. சோமன் அவர்களுக்கு எனது நெஞ்சார்ந்த நன்றியைத் தெரிவித்துக்கொள்கின்றேன்.

ஆகஸ்ட், 2020 இராசேந்திரன் சங்கரவேலாயுதன்

பொருளடக்கம்

இயல் 1

அறிமுகம்

1.1. மொழியியலும் தரவுத்தொகுதி மொழியியலும்

மொழியியல் என்பது மனித நாகரீகத்தின் மிகப் பழமையான பாடமாகும். நாம் பல்லாயிரக்கணக்கான ஆண்டுகளாக மனிதமொழியின் வேறுபட்ட நோக்குகளை ஆய்ந்து வருகின்றோம். பல நேரங்களில் மனித மொழி அறிவு, புலனறிவு மற்றும் கருத்துப் பரிமாற்றம் இவை குறித்து ஆய்வதற்கு முக்கியத்துவம் தரப்பட்டது. பல நூற்றாண்டுகளாக இப்பாடம் மனித அறிவின் பிற கிளைகளுடன் அதன் கருத்துருசார் தொடர்பை நிறுவிப் படிப்படியாக அதன் முழு வடிவில் வளர்ந்துள்ளது. இப்போது, இன்றைய நூற்றாண்டின் தொடக்கத்தில் இது எவ்வாறு மொழியின் வேறுபட்ட நோக்குகளைப் பற்றிய கோட்பாடுகள் சீரான மனித மொழி வெளிப்பாடுகளில் உருப்படுத்தம் செய்யப்பட்டுள்ள மொழியின் உண்மையான பயன்பாட்டிற்குச் சான்றாக மெய்படுத்தம் செய்யப்பட்டுள்ளது என்பதை ஆய்வதில் தன் கவனத்தைத் திருப்பியுள்ளது. மொழி ஆய்வின் இப்புதிய திசை மரபு மொழியியலுக்குக் கூடுதல் பரிமாணத்தைத் தந்துள்ளது. மொழியியல் ஆய்வுக்குப் புதிய கருவிகளின் குழுமத்தையும் மொழியின் உண்மையான பயன்பாட்டின் எடுத்துக்காட்டுகளைச் சேகரிக்கும் தொழில் நுட்பங்களையும் நல்கியக் கணிப்பொறித் தொழில் நுட்பத்தின் அறிமுகத்தால் இது சாத்தியமானது. இப்புதிய அணுகுமுறையின் அறிமுகம் மொழியிலுக்கும் மனிதவினத்திற்கும் இரண்டு வழிகளில் நன்மைபுரிந்துள்ளது. முதலாவது இது மொழி மற்றும் மொழிப்பயன்பாடு பற்றிய பழைய கோட்பாடுகள் தொடர்ந்து பயன்படுத்த உகந்ததா என்பதைத் தெரிந்துகொள்ள உதவுகின்றது. இரண்டாவது இது சீரான மனிதச் செயல்பாடுகளில் மொழிச் சான்றின் மற்றும் தகவலின் நேரடியான பயன்பாட்டிற்கு அனுகூலங்கள் தருகின்றது. இது சரியான திசை, மாறுபாடு மற்றும் பயன்பாடு என்பனவற்றின் குறைவால் சிலகாலம் பாதிக்கபட்டுள்ள மொழியலின் மறுமலர்ச்சிக்கும் நிலைநிற்புக்கும் உதவும். சில காலகமாகக் கணிப்பொறித் தொழில் நுட்பத்தின் கண்டுபிடிப்பு மற்றும் முன்னேற்றம் மொழியிலுக்கு ஒரு புதிய பரிமாணத்தைத் தந்துள்ளது.· மொழியைப் புலனறிவுடன் தொடர்புடைய மனிதக் கருத்துப்பரிமாற்றத்தின் கருவியாகப் பார்ப்பதை நோக்கமாகக் கொண்ட செயற்கை அறிவுநுட்பத்தின் (artificial intelligence) பகுதியாகக் கணிப்பொறி மொழியியல் (computational linguistics) என்ற பதிய பாடம் உருவாகியுள்ளது. தரவுத்தொகுதி மொழியியல் கணினிமொழியியலின் முக்கியமான கிளையாகப்

புள்ளியியலில் பயன்படுத்தப்படும் நெறிகள் மற்றும் உத்திகளைப் பின்பற்றி மிக ஒழுங்கான வழியில் சேகரிக்கப்பட்டுள்ள மொழிப் பயன்பாட்டின் அனுபவவாதச் சான்றின் பெரும் அளவைத் தரும் செயல்பாட்டில் முக்கியப் பங்களிப்பு செய்கின்றது. மேலும் இது விரிதரவை ஆயவும் மனித மொழியைப் புரிந்துகொள்ளவும் மனித அறிவியலின் பல களங்களில் பயன்படுத்தவும் கணினி மொழியியலிலும் செயற்கை அறிவுநுட்பத்திலும் தேவையானதாகக் கருதப்படும் தகவல்களைப் பெற மிக நுட்பமான உபாயங்களின் குழுமத்தைத் தருகின்றது. எவ்வாறு மக்கள் மொழியை ஒரு கருவியாகப் பயன்படுதி தங்களுக்குள் கருத்துப்பரிமாற்றம் செய்கின்றனர் என்பதைப் புரிந்துகொள்ள வலுவான புலனறிவு மற்றும் மொழியியல் சார்ந்த நோக்கம் இருக்கின்றது. மனிதவினத்துடன் திறமையாக மொழி ஊடாட்டங்கள் செய்யவியலும் அறிவுநுட்பமுள்ள கணிப்பொறி ஒழுங்குமுறைகளை உருவாக்க தொழிநுட்டபம் அடிப்படையிலான நோக்கம் இருக்கின்றது. இந்த நோக்கங்களுடன் கணிப்பொறி அறிவியலார்களும் மொழியலார்களும் ஒன்றாக இணைந்து இயந்திர மொழிபெயர்ப்புக்கும் தகவல் பிரித்தெடுப்புக்கும் மொழி புரிதலுக்கும் மொழி உருவாக்கத்திற்கும் ஒழுங்குமுறைகளை உருவாக்கியுள்ளனர். இச்செயல்பாடுகளுக்கு அவை சீரான மற்றும் அரிதான பண்புக்கூறுகளைக் கொண்ட இயற்கை மொழியை அனுபவவாத அடிப்படையில் புரிந்துகொள்ள வேண்டியது அவசியமாகும். இங்கு விரிதரவு தவிர்க்க முடியாததாக மாறுகின்றது; ஏனென்றால் தரவுத்தொகுதி அனுபவாதத் தரவுமையங்களின் சேகரிப்புக்குள் அறியவியலும் மொழியின் பெரும்பாலான பண்புக்கூறுகளை வெளிக்கொணரத் திறன்களைக் கொண்டுள்ளது. தற்போது பல எண்ணிக்கையிலான மக்கள் கணிப்பொறியில் பல்வேறு வகைப்பட்ட மொழியியல் தகல்களை நடைமுறைப்படுத்துவதில் ஈடுபட்டுள்ளனர்; ஏனென்றால் கணினி மொழியியலின் மற்றும் இயற்கை மொழியாய்வின் (natural language processing) நோக்கங்கள் மொழியின் பண்புக்கூறுகளை முடிந்த அளவு கணினி அடிப்படையில் பண்பாக்கம் செய்யவேண்டும் என்பதாகும். மேலும் இயற்கை மொழியியல் ஆய்வுக் களத்தில் செய்யப்பட்டுள்ள ஆராய்ச்சிகள் ஒரு இயற்கை மொழியின் மொழியியல் பகுப்பாய்வில் வெளிப்படுத்தப்படும் உள்ளறிவிலிருந்து நன்மையடைந்துள்ளன. எடுத்துக்காட்டாக, ஒரு கணிப்பொறியால் தானாக ஒரு எளிய வாக்கியத்தைப் பொருள்கோள் செய்ய ஒழுங்குமுறையை உருவாக்குபவர் இம்மாதிரி வாக்கியங்களின் முந்தைய மொழியியல் பகுப்பாய்வு பயனுள்ளதாகவும் அறிவுபுகட்டுவதாகவும் இருப்பதைக் காண்பார். எனவே ஒரு விரிதரவுக்குள் சேகரிக்கப்பட்டுள்ள மொழி அலகுகளின் வர்ணனைகளும் தன்மைகளும் மொழித்

தொழில் நுட்பத்திற்கு (language technology) மிக முக்கியமானவைகளாகும். விரிதரவிலிருந்து எடுக்கப்படும் புதிய தகவல்கள் கணினி மொழியியலுக்குப் பயன்படுவதுடன் முக்கிய மொழியியல் களத்திற்குப் பயன்படும் மொழியின் வருணனை மற்றும் புரிந்துகொள்ளல் பற்றிய மதிக்கத்தக்க உள்ளறிவுகளையும் தருகின்றது. மொழியில் மொழி ஊடாட்டம், கருத்துப்பரிமாற்றம் மற்றும் புரிந்துகொள்ளல் என்ற பல்வேறு களங்களில் வெளிப்படுத்தப்பட்டுள்ள ஒரு இயற்கை மொழியின் பயன்பாட்டின் எல்லாத் திசைகளையும் உள்ளடக்கும் விரிந்த பரப்பைக் கொண்ட பல்பரிமாணக் களமாகும். மொழி ஆய்விலும் பயன்பாட்டிலும் தரவுத்தொகுதியின் அறிமுகம், மைய மொழியிலுக்குப் புதிய பரிமாணங்களை உட்படுத்தியுள்ளது. கொள்கை அடிப்படையில் விரிதரவு மொழியியல் (corpus linguistics) இயந்திரத்தால் படிக்கவியலும் உரைகளின் மிகப்பெரிய சேகரிப்புகளின் வழி மொழியியல் நடப்புகளை ஆய்வதை நோக்கமாகக் கொண்ட அணுகுமுறையாகும். இவ்வணுகுமுறை பல ஆய்வுக்களங்களில் பயன்படுத்தப்படுகின்றது: ஒரு மொழியின் வருணனை ஆய்விலிருந்து மொழித் தொழில் நுட்பம் மற்றும் கல்வி வரையிலான எல்லா ஆய்வுகளிலும் பயன்படுத்தப்படுகின்றது. தரவுத்தொகுதி மொழியியலின் உள்ளுறையும் கருத்துரு விரிவாகப் பலவற்றை உள்ளடக்கியதாகும். இது நம்பகமான பேச்சு மற்றும்/அல்லது எழுத்து உரை மாதிரிகளின் எந்தக் கணிசமான பகுதியின் முழுமையான ஆய்வையும் குறிப்பிடும். இது மொழியின் ஒருகால மற்றும் இருகால நோக்குகளை வருவரைவு செய்யும். இலக்கிய உரைகள் மற்றும் செய்தித்தாள் கட்டுரைகள் இவற்றின் சேகரிப்புகளை உள்ளடக்கிய தற்கால மொழியின் பெரிய அளவிலான இயந்திரத்தால் படிக்கவியலும் தரவுத்தொகுதிகளை உட்படுத்தும். மின் விரிதரவு (electronic corpus) என்பது புதிய ஒன்றாகும். இக்களம் அரை நூண்றாண்டு வரலாறு உடையது. எது ஒரு தரவுத்தொகுதியாகக் கருதப்பட வேண்டும், அது எவ்வாறு திட்டமிடப்படவேண்டும், எவ்வாறு உருவாக்கப்படவேண்டும், வகைப்படுத்தப்படவேண்டும், ஆயப்படவேண்டும் மற்றும் பயன்படுத்தப்படவேண்டும் என்பதில் பொதுவான கருத்து இல்லை.

1.2. தரவுத்தொகுதியின் விளக்கம்

அனுபவவாத ஆய்வுகள் எந்த எழுதப்பட்ட அல்லது பேசப்பட்ட உரைகளையும் பயன்படுத்தி அதன் மேல் செய்யப்படுவதாகும். இம்மாதிரியான தனிப்பட்ட உரைகள் பல வகையான இலக்கிய மற்றும் மொழியியல் ஆய்வுகளுக்கு அடிப்படையாக அமைகின்றது. எடுத்துக்காட்டாக ஒரு செய்யுளின் அல்லது ஒரு புதினத்தின் நடையியல் ஆய்வு அல்லது ஒரு தொலைக்காட்சிப் பேச்சின் உரையாடல் ஆய்வு என்பதைக் கூறலாம். ஆனால் அனுபவாத

மொழியியலின் ஒரு வடிவுக்கு அடிப்படையாக அமையும் ஒரு விரிதரவு என்பதன் கருத்துச்சாயல் உரைகளின் குறிப்பிட்ட பகுதியைப் பரிசோதிக்கும் வழிகளிலிருந்து பல அடிப்படையான வழிகளில் வேறுபடும். கொள்கை அடிப்படையில் ஒரு உரைக்குக் கூடுதலான எந்தச் சேகரிப்புத் தொகுதியும் தரவுத்தொகுதி எனப்படும். தரவுத்தொகுதி என்று தமிழில் அழைக்கப்படும் corpus என்பது body (தேகம்/உடல்) என்பதன் இலத்தின் சொல்லாகும்; எனவே தரவுத்தொகுதி என்பதை எந்த ஒரு உரையின் உடல் என்று கூறலாம். ஆனால் விரிதரவு என்ற சொல் தற்கால மொழியியல் என்ற சூழலில் பயன்படுத்தப்படும் போது, எளிய விளக்க உரையைத் தாண்டிக் கூடுதல் சிறப்பு அர்த்தங்களைக் கொண்டிருக்கும் போக்கு காணப்படுகின்றது. இதைப் பின்வரும் நான்கு முக்கியத் தலைப்புகளின் அடிப்படையில் கருதலாம் (Tonny and Wilson 1996:21):

1. மாதிரிப்படுத்தமும் பிரதிநித்துவமும் (Sampling and representativeness)
2. முற்று வடிவம் (Finite Size)
3. இயந்திரம் படிக்க இயலும் வடிவம் (Machine readable form)
4. நிலைபேறான நோக்கீடு (Standard reference)

1.2.1. மாதிரியாக்கமும் பிரதிநித்துவமும்

மொழியியலில் நாம் ஒரு குறிப்பிட்ட உரை அல்லது நூலாளர் என்பதை விட ஒரு மொழியின் முழு வேறுபாட்டில் ஆர்வம் காட்டுவோம். இந்நேர்வுகளில் நமக்குத் தரவு சேகரிப்பில் இரண்டு விருப்புகள் உள்ளன: முதலாவது, நாம் அந்த வகையின் ஒவ்வொரு கூற்றையும் ஆய இயல வேண்டும் அல்லது இரண்டாவது, நாம் அந்த வகையின் ஒரு சிறிய மாதிரியை உருவாக்க இயல வேண்டும். முதலாவது விருப்பு, செயல்பாட்டு அடிப்படையில் இயலாததாகும். உரைகளின் தொகை மிகப்பெரியதாகும். ஆங்கிலம், தமிழ் போன்ற மொழிகளில் இது அதிகரித்துக் கொண்டே செல்கின்றது. இது எல்லை இல்லாதது என்று கூறலாம். எனவே நாம் ஒரு மொழி வகையின் தரவுத்தொகுதியை உருவாக்கச் சம்பந்தப்பட்ட வகையைக் கூடுதலாக உருப்படுத்தம் செய்யும் அதாவது அந்த வகையின் போக்கின் துல்லியமான வெளிப்பாட்டைத் தரும் ஒரு மாதிரியைக் கருத்தில் கொள்வோம். ஒரு வகை நாவல்களிலிருந்து தேர்ந்தெடுக்கப்பட்ட உரை மீது நாம் நமது மாதிரியை அடிப்படையாகக்கொள்ள விரும்புவதில்லை. நாம் பலவகைப்பட்ட எழுத்தாளர்கள் மற்றும் வகைகளின் விரிந்த எல்லைக்கு உட்பட்ட ஒரு மாதிரியை எதிர்பார்ப்போம்; அவற்றை சேர்த்து எடுக்கும்போது அவற்றை சராசரியானது என்று கருத இயலும் மற்றும் நாம்

விருப்பப்படும் முழு மொழி மக்கள்தொகையின் போதுமான அளவு துல்லியமான வெளிப்பாட்டைத் தரும்.

1.2.2. முற்று வடிவம்

மாதிரிபடுத்தத்தைப் போன்று தரவுத்தொகுதி என்பது முற்றுவடிவத்தின் பனுவலின் உடலைக் குறிப்பிடும் போக்கு இருக்கின்றது. எடுத்துக்காட்டாக, ஒரு தரவுத்தொகுதியின் வடிவம் ஒரு மில்லியன் சொற்களில் உருவாக்கப்படலாம். இது உலகப் பொதுமையானது அல்ல. எடுத்துக்காட்டாக, பிர்மிங்காம் பல்கலைக்கழகத்தில் ஜாண் சின்கிளேயரின் COBUILD குழு கண்காணிப்பு தரவுத்தொகுதி (monitor corpus) என்ற உரைகளின் சேகரிப்பின் உருவாக்கத்திலும் ஆய்விலும் ஈடுப்பட்டுள்ளனர். இது திறந்த-முடிவுற்ற தரவுத்தொகுதியாகும். இம்மாதிரியான தரவுத்தொகுதி அகராதி உருவாக்கத்திற்கு முக்கியமானதாகும். லாங்கமானின் COBUILD ஆங்கில அகராதி இந்த விரிதரவால் உருவாக்கப்பட்டதாகும். தரவுத்தொகுதி உருவாக்கத்தின் தொடக்கத்தில் ஆய்வுத்திட்டம் எவ்வாறு மொழி வகை மாதிரிப்படுத்தம் செய்யப்படவேண்டும், முன்வரையறுக்கப்பட்ட ஆக மொத்தம் கிடைக்க எத்தனை எண்ணிக்கையிலான சொற்களின் எத்தனை மாதிரிகள் சேகரிக்கப்படவேண்டும் என்று நிர்ணயிக்கப்படவேண்டும். லங்காஸ்டர்-ஓஸ்லோ/பெர்கன் (லோப்) தரவுத்தொகுதி பிரவுண் தரவுத்தொகுதி 1,000,000 சொற்களைக் கொண்டுள்ளன. பிரிட்டிஷ் தேசிய விரிதரவு 100,000,000 சொற்களைக் கொண்டுள்ளன. கண்காணிப்பு விதரவு போல் அல்லாமல் இம்மாதிரியான தரவுத்தொகுதி சொற்களின் ஆக மொத்தத்தை எட்டியதும் சேகரிப்பு நிறுத்தப்படும் மற்றும் அதன் பிறகு தரவுத்தொகுதியின் வடிவம் அதிகரிப்பதைல்லை.

1.2.3. இயந்திரம் படிக்கவியலும் வடிவம்

இன்றைய காலகட்டத்தில் தரவுத்தொகுதி என்றாலே அது இயந்திரம் படிக்கக் கூடிய வடிவில் இருக்க வேண்டும் என்பதை உட்கோள் செய்யும். பல ஆண்டுகளாக தரவுத்தொகுதி என்பது அச்சிடப்பட்ட உரைகளைக் குறிப்பிடுவதாக இருந்தது. இப்பொழுது இந்நிலை மாறிவிட்டது. சில பெருந்தரவுகள் அச்சு வடிவத்திலும் இருக்கிறது. எடுத்துக்காட்டாக ஆங்கில உரையாடலின் ஒரு விரிதரவு (A Corpus of English Conversation) (Svartvik and Quirk, 1980) என்பது அச்சடிக்கப்பட்ட வடிவில் இருக்கின்றது. இந்த தரவுத்தொகுதி மூல வடிவான லண்டன்-லண்ட் தரவுத்தொகுதி (London-Lund Corpus) என்பதை உருப்படுத்தம் செய்கின்றது. இந்தப் பனுவல்கள் லண்டன்-லண்ட் விரிதரவுக்குள் (London-Lund Corpus) இயந்திரம் படிக்க இயலும்

வடிவத்துடன் இருந்தாலும் இது புத்தக வடிவில் உள்ள மிகக் குறைந்த தரவுத்தொகுதிகளில் ஒன்று என்ற அளவில் குறிப்பிடத்தகுந்தது. லங்காஸ்டர்/ஐ.பி.எம். பேச்சு ஆங்கில தரவுத்தொகுதி (Lancaster/IBM Spoken English Corpus) பேச்சு விரிதரவுக்கு எடுத்துக்காட்டாகும். இயந்திரம் படிக்க இயலும் வடிவம் பல ஆய்வுகளுக்கு பயன் உள்ளதாக அமைகின்றன. இது அச்சு வடிவில் அல்லது பேச்சு வடிவில் உள்ள தரவுத்தொகுதி யைக் காட்டிலும் மிகப் பயனுள்ளதாக இருக்கின்றது.

1.2.4. நிலைபேறான நோக்கீடு

'நிலைபேறான நோக்கீடு' (standard reference) என்பது தரவுத்தொகுதியின் வரையறை விளக்கத்திற்கு முக்கியமான பகுதியாக இல்லாவிட்டாலும் ஒரு தரவுத்தொகுதி அது உருப்படுத்தம் செய்யும் மொழி வகையின் நிலைபேறு நோக்கீடாக இருக்கும் என்பது எதிர்பார்ப்பாகும். எடுத்துக்காட்டாக, அமேரிக்க ஆங்கில மொழியின் பிரவுண் தரவுத்தொகுதி (Brown Corpus), எழுத்து பிரிட்டன் ஆங்கிலத்தின் லோப் தரவுத்தொகுதி (LOB Corpus), பேச்சு பிரிட்டன் ஆங்கிலத்தின் லண்டன்-லண்ட் தரவுத்தொகுதி (London–Lund Corpus) என்பன ஆய்வாளர்களால் பரவலாகப் பயன்படுத்தப்படுகின்றன.

இவ்வாறு இக்கால மொழியியலில் தரவுத்தொகுதி என்பது ஒரு குறிப்பிட்ட பனுவலாக இல்லாமல் கருதப்பட்ட மொழி வகையின் பிரதிநிதியாக இருக்கத்தக்க விதத்தில் மாதிரியாக்கம் செய்யப்பட்டு இயந்திரத்தால் படிக்க இயலும் முற்றுப்பெற்ற வடிவம் என்று வருணிக்கத்தக்க விதத்தில் இருக்கின்றது.

1.3. தரவுத்தொகுதியின் முக்கியமான பண்புக்கூறுகள்

கோட்பாடு அடிப்படையில் corpus என்பததை பின்வருமாறு விளக்கலாம்: Capable Of Representing Potentially Unlimited Selection of Texts. அதாவது தரவுத்தொகுதி ஆற்றல் அடிப்படையில் எல்லையற்ற உரைகளின் தேர்வை உருப்படுத்தம் செய்ய இயலும். It is Compatible to computers, Operational in research application, Representative of source language, Processable by men and machine, Unlimitted in data, and Systematic in formation and representation. அதாவது தரவுத்தொகுதி கணிப்பொறியுடன் இயைபுள்ளது, ஆய்வுப் பயன்பாட்டில் செயன்மைபுரிவது, மூலமொழியின் பிரதிநிதி, மனிதனாலும் பொறியாலும் பகுத்தாயத்தக்கது, தரவில் எல்லையில்லாதது மற்றும் ஆக்கத்திலும் உருப்படுத்தத்திலும்

ஒழுங்குமுறையானது. தரவுத்தொகுதியின் முக்கிய பண்புக்கூறுகளாகப் பின்வருவனவற்றைக் குறிப்பிடலாம்:

அளவு: எப்போதும் எழுப்பப்படும் பொதுவான கேள்வி எவ்வளவு பெரிய விரிதரவை நாம் உருவாக்க வேண்டும் என்பதாகும். இந்தக் கேள்விக்கு விடை தருவது அவ்வளவு எளிதல்ல. இருப்பினும் 'அளவு' என்பது பொதுவாக ஒரு விரிதரவு பேச்சுவடிவிலோ எழுத்து வடிவிலோ மிகப் பெரிய அளவு மொழித் தரவைக் கொண்டிருக்கும் மிகப்பெரிய வடிவைக் குறிப்பிடும். ஒரு விரிதரவின் வடிவ அளவு உண்மையில் அதன் உடலை உருவாக்கும் உறுப்புக்களின் கூட்டுத்தொகையாகும். ஒரு தரவுத்தொகுதியை உருவாக்குவதன் முக்கிய நோக்கம் பெரிய அளவில் மொழித் தரவைச் சேகரிப்பதாகும். 'கண்காணிப்பு விரிதரவின்' கண்டுபிடிப்பு வடிவ அளவின் கணிப்பின் அடிப்படையை 'மொத்த தொகை' என்பதிலிருந்து 'ஒழுக்கின் விகிதம்' என்பதற்கு மாற்றியுள்ளது.

தரம்: தரம் என்பதன் வழுநிலை மதிப்பீடு 'நம்பகத்தன்மை' என்பதாகும். எல்லா விஷயங்களும் பேச்சின் மற்றும் எழுத்தின் மெய்யான இயல்பான பயன்பட்டிலிருந்து பெறப்படவேண்டும். இங்கு மொழியியலாரின் பங்களிப்பு தேவை. அவர் பரிசோதனைக் கட்டுப்பாடுகளோ செயற்கையான சூழல்களோ உள்ள கருத்துப்பரிமாற்றத்திலிருந்து தரவு பெறப்படாமல் சாதாரண கருத்துப்பரிமாற்றத்திலிருந்து பெறப்பட்டதா என்று பரிசோதிக்க வேண்டும். இருப்பினும் இரண்டிற்கும் எல்லைக் கோடு வரைப்பது கடினமானதாகும். எடுத்துக்காட்டாக சில தொலைக்காட்சி நிகழ்வுகள் வேண்டுமென்றே பங்கெடுப்பவர்களைச் செயற்கையான சூழலில் வைத்திருப்பதை உணரலாம். மாறாக சாதாரண உரையாடல் இயற்கையான சூழலில் நிகழ்வதாகக் கருத இயலும்; இருப்பினும் காட்சியளிப்புக்காக ஒன்றோ அதற்கு மேற்பட்ட பங்கெடுப்பாளர்கள் அதை ஒத்திகை பார்த்திருக்கவும்கூடும்.

பிரதிநிதித்துவம்: இது விஷயங்களின் விரிந்த பரப்பெல்லையிலிருந்து பெற்ற மாதிரிகளை உட்படுத்தும். ஒரு மொழியின் கூடுதலான மொழிப் பண்புக்கூறுகளைப் பிரதிநிதித்துவம் செய்யவேண்டி இது எல்லாத் துறைகளிலிருந்தும் களங்களிலிருந்தும் சரிநிகரான விஷங்களைக் கொண்டிருக்கின்றது. இதன் மீது திட்டமிடப்படும் எதிர்கால ஆய்வு ஒரு மொழியைப் பிரதிநிதித்துவம் செய்யும் விரிதரவிலிருந்து பெறப்பட்ட தகவலின் பரிசோதனையையும் வேண்டும்.

எளிமை: தரவுத்தொகுதி எளிய வெளிப்படையான உரைகளைக் கொண்டிருக்கும். அதாவது நாம் பனுவலுக்குள் எந்தக் கூடுதலான மொழித் தகவலும் அடையாளப்படுத்தப்படாத எழுத்துக்களின் தொடர்ச்சியான கோர்வையை எதிர்பார்கின்றோம். ஒரு எளிய வெளிப்படையான உரை மொழிசார்ந்த மற்றும் மொழிசாராத தகவல்களின் பல வகைகளைக் கொண்டிருக்கும் எந்த அடையாளப்படுத்தலுக்கும் எதிரானதாகும்.

சமத்துவம்: தரவுத்தொகுதியில் பயன்படுத்தப்படும் மாதிரிகள் நிகரான அளவில் இருக்கவேண்டும். இருப்பினும் இது விவாதத்திற்குரிய சிக்கல்; பயன்பாட்டின் தேவை அடிப்படையில் சரிவிகித அளவில் மாதிரிகளின் அளவு மாறும் ஆகையால் சரிநிகர் எல்லாவிடத்திலும் ஏற்றுக்கொள்ளப்படமாட்டாது. மாதிரியாக்க மாதிரி (sampling model) விரிதரவைக் கூடுதல் பிரதிநிதித்துவம் செய்யவும் பல்பரிமாணம் உடையதாகச் செய்யவும் வேண்டி போதுமான அளவு மாறும்.

மீளப்பெறும் தன்மை: தகல்கள், எடுத்துக்காட்டுகள் மற்றும் நோக்கீடுகள் என்பன பயன்படுத்துவோர்களால் விரிதரவிலிருந்து எளிதாக மீளப்பெறத்தக்கதாக இருக்கவேண்டும். இது பயன்படுத்துவோர்களுக்கு வேண்டி கணினியில் மின்வடிவில் தரவைப் பாதுகாக்கும் தொழிநுட்பத்தில் கவனம் செலுத்தும். இன்றைய தொழில் நுட்பம் தனிநபர் கணிப்பொறியில் விரிதரவை உருவாக்கவும் நாம் தேவைப்படும் போது எளிதாகத் தரவை மீளப்பெறவும் சாத்தியமானதாகச் செய்துள்ளது.

பரிசோதனைக்கு உட்படுத்தப்படும் தன்மை: பல மூலங்களிலிருந்து சேகரிக்கப்பட்டுள்ள பனுவல் விஷயங்கள் உண்மையானதாகவும் நம்பகத்தன்மையுடையதாயும் இருக்கவேண்டும். விரிதரவு எந்த வகையிலான அனுபவாத ஆய்வுக்கும் பரிசோதனைக்கும் உட்படுத்தப்படவேண்டும். இந்த பண்பு விரிதரவை மொழி வல்லுநர்களால் நம்பத்தகுந்ததாகச் செய்கின்றது; ஏனென்றால் அவர்கள் பிறரால் செய்யப்படும் உற்றுநோக்கல்களைப் பரிசோதிக்கவோ மறுக்கவோ வேண்டி விரிதரவை எந்த அனுபவவாதப் புலனாய்வுக்கும் உட்படுத்த இயலும்.

விரிவுபடுத்தும்தன்மை: விரிவுபடுத்தல் சீர்மையை அதிகரிக்கவேண்டும். இது விரிதரவைக் காலப்போக்கில் மொழிக்குள் காணப்படும் மொழி மாற்றங்களைப் பதிவுசெய்வற்கு இணையானதாகச் செய்யும். காலப்போக்கில் புதிய மொழித் தரவை முறையாக அதிகரிப்பதால் இருகால மொழி ஆய்வுகளுக்கு வாய்ப்பைத் தந்து தரவுத்தொகுதி வரலாற்றுப் பரிமாணத்தைப் பெறும்.

ஆவணப்படுத்தல்: தரவுத்தொகுதிப் பயன்படுத்துவோர் எதிர்கால நோக்கீட்டுக்காகப் பனுவல் மாதிரிகளின் ஆணவப்படுத்தலின் எல்லா தகவல்களையும் ஒரு தனியான இடத்தில் வைத்திருக்கவேண்டும் என்பதைச் சரியான ஆணவப்படுத்தல் உணர்த்தும். தேவை என்றால் ஆணவப்படுத்தலுக்கு எல்லா நோக்கீடுகளும் அடங்கிய ஒரு தலப்பு வழிகாட்டியை உட்படுத்தவேண்டும். விரிதரவின் எளிதான நிர்வாகம், அணுகல் மற்றும் பகுப்பாய்வுக்காக இது சாதாரண பனுவலை அடையாளப்படுத்தலில்.

1.4. தொடக்ககாலத் தரவுத்தொகுதி மொழியியல்

தொடக்க காலத் தரவுத்தொகுதி மொழயியல்' (Early Corpus Linguistics) என்ற கலைச் சொல்லை நாம் சாம்ஸ்கியின் வரவுக்கு முன்னால் உள்ள மொழியியல் என்பதை விளக்கப் பயன்படுத்தலாம். களமொழியியல் (Field Linguistics) மற்றும் அமைப்புமொழியியலார் மரபைச் சார்ந்த பிந்தைய மொழியியலார்கள் பயன்படுத்திய அடிப்படை நெறிமுறையை எந்த ஐயமும் இன்றி விரிதரவு அடிப்படையிலானது என்று கூறலாம். ஆனால் இக்கால கட்டத்திலிருந்தே 'தரவுத்தொகுதி மொழியியல்' என்ற கலைச்சொலைப் பனுவல்களிலும் ஆய்வுகளிலும் காணலாம் என்று அர்த்தம் அல்ல.' ஆரம்ப கால தரவுத்தொகுதி மொழியியல்' என்ற கலைச்சொல் இப்படைப்புகளை வகைப்பாடு செய்யவேண்டி உருவாக்கப்பட்டதாகும். இது சில வழிகளில் சாம்ஸ்கிக்கு முந்தைய மொழியியல் நெறிமுறையைக் குறிப்பதுடன் அது உறவு வைத்துள்ள புதிய தரவுத்தொகுதி மொழியியல் நெறிமுறையுடன் தொடர்புகொண்டுள்ளது. இச்செய்தியை எடுத்துக்காட்டவும் 1950-களுக்கு முன்னர் மொழியியலில் தரவுத்தொகுதியின் பரந்த பயன்பாட்டை நிறுவவும் கீழே அக்காலகட்டத்திய தரவுத்தொகுதி அடிப்படையிலான ஆய்வுகளின் சுருக்கம் தரப்பட்டுள்ளது.

தொடக்ககால தரவுத்தொகுதியை முன் மின்னணுசார் தரவுத்தொகுதி (pre-electronic corpus) என்றும் கூறலாம். எலக்ட்ரானிக் தரவுத்தொகுதியின் (electronic corpus) தலைமுறைக்கு முன்பு, மொழி தரவுத்தளங்கள் அகராதியியல் (lexicography), கிளைமொழியியல் (dialectology), ஒப்பீட்டு மொழியியல் (comparative linguistics), மொழி கற்பித்தல் (language teaching), மொழி ஈட்டம்/பேறு (language acquisition), ஒலியனியல் (phonology) மற்றும் பிற மொழியியல் துறைகளில் பயன்படுத்தப்பட்டன. முந்தைய அறிஞர்கள் தரவுத்தொகுதியை எவ்வாறு உருவாக்கினார்கள், பகுப்பாய்வு செய்தார்கள் மற்றும் அதிலிருந்து தகவல்களை தங்கள் படைப்புகளில் எவ்வாறு பயன்படுத்தினர் என்பதை கீழ்வரும் விளக்கம் காட்டுகிறது.

தரவுத்தொகுதி அடிப்படையிலான மொழி ஆய்வை சரியான கண்ணோட்டத்தில் பார்க்க இது தேவையான தகவலை அளிக்கிறது.

1.4.1. எழுத்துக்கூட்டல் மரபுகள்

எழுத்துக்கூட்டல் மரபுகளை (Spelling Conventions) அறிய வேண்டியும் தரவுகளின் சேகரிப்பு தொடக்க காலகட்டத்தில் மேற்கொள்ளப்பட்டது. காதிங் (Käding 1897) மில்லியன் சொற்கள் கொண்ட ஜெர்மன் மொழியின் மிகப்பெரிய தரவுத்தொகுதியை ஜெர்மன் மொழியின் எழுத்துக்களின் வருகைகளின் மற்றும் எழுத்துக்களின் தொடர்ச்சிகளின் நிகழ்வெண்ணைக் கண்டுப்பிடிக்கப் பயன்படுத்தினார். இவ்விரிதரவு அதன் அளவு காரணமாக அக்காலகட்டத்தில் சிறந்ததாகக் கருதப்பட்டதுடன் சில தற்கால தரவுத்தொகுதிகளுடன் ஒப்பிடுகையில் அளவு அடிப்படையில் மேம்பட்டதாகவும் இருக்கின்றது.

1.4.2.அகராதி தயாரிப்பு

சாமுவேல் ஜான்சன் தமது ஆங்கில மொழியின் அகராதியில் (Dictionary of the English Language 1755) அர்த்தம் மற்றும் பயன்பாட்டை விளக்குவதற்காக கிட்டத்தட்ட 40000 தலைப்புகளுக்கு 150000க்கும் மேற்பட்ட மேற்கோள்களை சேகரித்தார்.

The ஆக்ஸ்போர்டு ஆங்கில அகராதியின் 1ஆவது பதிப்பு (Oxford English Dictionary 1882) (பதிப்பாசிரியர்: ஜேம்ஸ் முர்ரே (James Murray)) சுமார் 414825 பதிவுகளின் பயன்பாட்டை விளக்குவதற்கு பல்வேறு இலக்கியப் பனுவல் மூலங்களிலிருந்து 5 மில்லியன் மேற்கோள்கள் கொண்ட (கிட்டத்தட்ட 50 மில்லியன் சொற்களைக் கொண்ட) ஒரு தரவுத்தொகுதியைப் பயன்படுத்தினார்.

1882இல் தொகுக்கப்பட்ட OEDஇன் முதல் தொகுதி வெளிவந்த பின்னர் அதனை அடுத்து வெளிவந்த துணைத் தொகுதிகள் பல மில்லியன் மேற்கோள்கள் கொண்ட தரவுத்தொகுதியைப் பயன்படுத்தியது. இந்த வேலை 1984 இல் தொடங்கியது மற்றும் OED2 இன் இரண்டாம் பதிப்பு 1989 இல் 20 தொகுதிகளில் வெளிவந்தது.

OED2 (1989) OED1, துணைத்தொகுதிகள் (1972-1986) இவற்றில் பயன்படுத்தப்பட்ட விரிதரவு மற்றும் அறிவியல் மற்றும் தொழில்நுட்பம், இழிவான உரைகள், கெட்டவார்த்தைகள் மற்றும் பிரிட்டனுக்கு வெளியே பயன்படுத்தப்படும் ஆங்கில வகைகள் இவற்றில் இருந்து 50000 புதிய சொற்கள் இவற்றைக் கொண்டிருந்தது. இது 2.4 மில்லியன் மேற்கோள்களிலிருந்து

வரையறுக்கப்பட்ட மற்றும் எடுத்துக்காட்டப்பட்ட 447000 சொற்களுக்கு மேல் உள்ள ஒரு கார்பஸைப் பயன்படுத்தியது.

வெப்ஸ்டரின் புதிய சர்வதேச அகராதியின் (Webster's New International Dictionary) 2ஆவது பதிப்பு (1934) புத்தகங்கள், பத்திரிகைகள், செய்தித்தாள்கள், துண்டுப்பிரசுரங்கள், அட்டவணைகள் மற்றும் கற்றறிந்த பத்திரிகைகள் இவற்றை முறையாக வாசிப்பதில் இருந்து சேகரிக்கப்பட்ட 1665000க்கும் மேற்பட்ட மேற்கோள்களின் கார்பஸைப் பயன்படுத்தியது. மொத்தச் சேகரிப்பில் பதிவுசெய்யப்பட்ட பயன்பாட்டின் 4500000க்கும் மேற்பட்ட புதிய எடுத்துக்காட்டுகள் உள்ளன.

வெப்ஸ்டரின் புதிய சர்வதேச அகராதியின் 3 வது பதிப்பு (1961) கிட்டத்தட்ட அரை மில்லியன் தலைப்புச் சொற்களின் அர்த்தத்தையும் பயன்பாட்டையும் சரிபார்க்கவும் விளக்கவும் 10 மில்லியனுக்கும் அதிகமான மேற்கோள்களின் தரவுத்தொகுதியை அணுகியது.

1.4.3. கிளைமொழி ஆய்வு

ஆங்கிலக் கிளைமொழி சமூகம் (English Dialect Society) (தொடக்கம் 1873) ஆங்கில கிளைமொழி அகராதி (English Dialect Dictionary) (1898-1905) மற்றும் ஆங்கில கிளைமொழி இலக்கணம் (English Dialect Grammar) (1905) ஆகியவற்றை வெளியிட பிராந்திய சொற்களஞ்சியத்தின் கார்பஸை சேகரித்தது.

ஜார்ஜ் வெங்கர் (George Wenker) 1881இல் ஸ்ப்ராகட்லாஸ் டெஸ் டாய்சனை (Sprachatlas des Deutschen) வெளியிட ஜெர்மன் கிளைமொழிகளின் ஒரு பெரிய சொல்சார் தரவுத்தளத்தைச் சேகரித்தார்.

டென்மார்க்கின் (Denmark) மரியஸ் கிறிஸ்டென்சன் (Marius Kristensen) தொடர்ச்சியான பகுப்பாய்வு மற்றும் அடுத்தடுத்த தொகுதிகளில் (1898-1912) டேனிஷ் கிளைமொழி அட்லஸின் (Danish Dialect Atlas வெளியீட்டிற்காக கணிசமான அளவு தரவுத்தொகுதியை சேகரித்தார்.

ஜூல்ஸ் கில்லியரோன் (Jules Gilliéron) என்ற பிரெஞ்சு அறிஞர், ஒரு பயிற்சி பெற்ற களப்பணியாளருடன் அட்லஸ் லிங்குஸ்டிக் டி லா பிரான்ஸின் (Atlas Linguistique de la France) (1902-1910: 13 தொகுதிகள்) பகுப்பாய்வு மற்றும் வெளியீட்டிற்காக பெரிய கிளைமொழி கார்பஸை சேகரித்தார்..

அலெக்சாண்டர் எல்லிஸ் (Alexander Ellis) 20 ஆண்டுகளில் 1145 இடங்களிலிருந்து கைமுறையாக சேகரிக்கப்பட்ட 5 மில்லியன் சொற்களின் பேச்சு கார்பஸைப் பயன்படுத்தி

ஆங்கில கிளைமொழிகளின் தற்போதைய ஒலியனியல் (Existing Phonology of English Dialects) (1889) என்பதை வெளியிட்டார்.

இரண்டாம் உலகப் போருக்குப் பிறகு, யூஜென் டயத் மற்றும் ஹரோல்ட் ஆர்டன் (Eugen Dieth and Harold Orton) ஆகியோர் 5 ஆண்டுகளில் இங்கிலாந்தின் 311 பிராந்தியங்களைச் சேர்ந்த களப்பணியாளர்களால் சேகரிக்கப்பட்ட ஒரு பெரிய பேச்சு கார்பஸுடன் ஆங்கிலக் கிளைமொழிகளின் சர்வே (Survey of English Dialects) (1961) என்பதை வடிவமைத்தனர்.

அமெரிக்காவில், நியூ இங்கிலாந்தின் மொழியியல் அட்லஸ் (Linguistic Atlas of New England) (1943) 2 ஆண்டுகளில் களப்பணியாளர்களால் சேகரிக்கப்பட்ட 213 பிராந்திய வகைகளில் இருந்து சேகரிக்கப்பட்ட பேச்சு தரவுத்தளத்தின் ஒரு கார்பஸைக் கொண்டிருந்தது.

அமெரிக்க பிராந்திய ஆங்கிலத்தின் அகராதி (Dictionary of American Regional English) (1985) அமெரிக்காவில் பயன்படுத்தப்படும் கிளைமொழிகளிலிருந்து தொகுக்கப்பட்ட ஒரு கார்பஸ் தரவுத்தளத்தையும் அணுகியது.

1.4.4. சொல்சார் ஆய்வு

எட்வர்ட் தோர்ன்டைக் 1921இல் ஆசிரியரின் வேர்ட் புத்தகத்தை வெளியிட கிட்டத்தட்ட 4.5 மில்லியன் ஆங்கில சொற்களின் தரவுத்தளத்தை தொகுத்தார். பின்னர், இந்த கார்பஸ் 30000 சொற்களின் ஆசிரியரின் வேர்ட்புக் (The Teacher's Wordbook of 30000 Words) (1940) தயாரிக்க பத்திரிகைகள், கால இதழ்கள் மற்றும் சிறார் இலக்கியங்களிலிருந்து பெறப்பட்ட பனுவல்களுடன் 18 மில்லியன் சொற்களாக விரிவுபடுத்தப்பட்டது,.

எர்னஸ்ட் ஹார்ன் (Ernest Horn) ஒரு அடிப்படை எழுதும் சொல்லகராதி (A Basic Writing Vocabulary) (1926) என்பதை வெளியிட மனித முயற்சிகளின் பல்வேறு மூலங்களிலிருந்து தனிப்பட்ட மற்றும் வெளியிடப்பட்ட கடிதங்களின் பல்வேறு எழுதப்பட்ட பனுவல்களிலிருந்து 5136816-க்கும் மேற்பட்ட இயங்கும் சொற்களின் மற்றொரு தரவுத்தொகுதியைக் கைமுறையாகத் தொகுத்தார்.

மைக்கேல் வெஸ்ட் மற்றும் ஜேம்ஸ் எண்டிகாட் (Michael West and James Endicott) ஆகியோர் 1935ஆம் ஆண்டில் முதல் ஈ.எஃப்.எல் அகராதியை (English as a Foreign Language (EFL) dictionary) (The New Method English Dictionary/தி நியூ மெதட் ஆங்கில அகராதி) வெளியிட்டனர். இது ஒரு பெரிய கார்பஸிலிருந்து சேகரிக்கப்பட்ட 1490 சொற்களின்

சொற்றொகுதியில் கிட்டத்தட்ட 24000 ஐடங்களின் பொருளின் விளக்கங்களைக் கொண்டிருந்தது.

எச். பால்மர், எம். வெஸ்ட், எல். ஃபாசெட், மற்றும் ஏ.எஸ். ஹார்ன்பி (H. Palmer, M. West, L. Faucett, and A.S. Hornby) ஆகியோர் சொற்றொகை தேர்வு குறித்த இடைக்கால அறிக்கையை (The Interim Report on Vocabulary Selection) (1936) வெளியிட ஆங்கிலத்தில் தேர்ந்தெடுக்கப்பட்ட சொற்களின் மற்றொரு பெரிய தரவுத்தொகுதியைப் பயன்படுத்தினார்.

சார்லஸ் ஃப்ரைஸ் (Charles Fries) தி அமெரிக்கன் ஆங்கில இலக்கணம் (The American English Grammar) (1940) என்பதை வெளியிட 2000 கையால் எழுதப்பட்ட தனிப்பட்ட கடிதங்கள் மற்றும் முக்கியமாக புகார்கள் மற்றும் நிதி அல்லது பிற துயரங்களைத் தணிக்க சில அதிகாரத்துவ நடவடிக்கைகளுக்கான கோரிக்கைகள் உள்ளடக்கிய யுஎஸ்ஏ அரசாங்கத்திற்கு எழுதப்பட்ட மற்றொரு 1000 கடிதங்களின் பகுதிகள் குறித்து ஆய்வு செய்தார்.

1.4.5. இலக்கணங்கள் எழுதுதல்

பவுட்ஸ்மா (Poutsma) (1926-29) மற்றும் க்ரூசிங்கா (Kruisinga) (1931-32) ஆகியோர் ஆங்கில செய்தித்தாள்கள் மற்றும் நாவல்களிலிருந்து சேகரிக்கப்பட்ட வாக்கியங்களிலிருந்து பல்வேறு அளவுகளில் உள்ள தரவுத்தொகுதிகளை வடிவமைத்து பயன்படுத்தினர்.

ஓட்டோ ஜெஸ்பர்சன் (1937, 1909-1949) ஆங்கிலத் தொடரியல் மற்றும் இலக்கணத்தைப் பற்றி விவாதிக்க பல்வேறு ஆதாரங்களில் இருந்து ஆங்கிலத்தின் விளக்க எடுத்துக்காட்டுகளின் கார்பஸை தொகுத்தார்.

ராண்டால்ஃப் க்யூர்க் (Randolf Quirk) மற்றும் அவரது குழுவினர் ஆங்கிலத்தின் விரிவான இலக்கணத்தை எழுத அன்றாட ஆங்கிலத்தின் ஒரு தரவுத்தொகுதியைத் (Quirk, et al./க்யூர்க், மற்றும் பலர். 1985) தொகுத்தனர். பிரதிநிதித்துவ எடுத்துக்காட்டுகளைத் தேர்ந்தெடுப்பதில் ஒருதலைப்பட்சத்தின் சிக்கல்களை சமாளிப்பதற்கும், இலக்கண வல்லுநரின் இயல்பானதை விட ஒப்பீட்டளவில் அசாதாரணமான கட்டமைப்புகளைத் தேர்ந்தெடுப்பதை நீக்குவதற்கும் இது உதவும்.

ஹரோல்ட் ஈடன் (Harold Eaton) (1940) ஆங்கில பயன்பாட்டை விவரிக்க பல்வேறு பொருண்மையியல்சார் நிகழ்வெண் பட்டியல்களை (semantic frequency lists) உருவாக்கினார் (லார்ஜ் 1949).

ஃப்ரைஸ் (Fries) எழுதிய ஆங்கிலத்தின் விளக்க இலக்கணம் (A descriptive grammar of English) (1952), வரிவடிவாக்கப்பட்ட தொலைபேசி உரையாடல்களின் ஒரு தரவுத்தொகுதியின் பகுப்பாய்வை அடிப்படையாகக் கொண்டது.

கென்ஹெய்ம் மற்றும் பலர் (Gougenheim et al. 1956) பேச்சு பிரஞ்சு மொழியில் உயர் அதிர்வெண் சொல்சார் மற்றும் இலக்கண விருப்புகளை/தேர்வுகளை விவரிக்கக் கிட்டத்தட்ட 275 தகவலாளிகளிடமிருந்து சேகரிக்கப்பட்ட வரிவடிவமாக்கப்பட்ட பேச்சு பிரெஞ்சு தரவுத்தொகுதியைப் பயன்படுத்தினர்.

பிரஜ் கச்ரு (Braj Kachru 1961) முறையான கருத்தாய்வுகளை ஆராய்வதற்கும், இந்திய ஆங்கிலத்தில் இலக்கணங்களை எழுதுவதற்கு மதிப்புமிக்க உள்ளீடுகளை வழங்குவதற்கும் இந்தியர்களின் 'படைப்பு எழுத்துக்களின்' உரை தரவுத்தொகுதியைப் பயன்படுத்தினார்.

தேசாய் (Desai 1974) மற்றும் நிஹலானி மற்றும் பலர் (Nihalani et al. 1979) இந்தியாவில் ஆங்கிலம் கற்பிப்பதற்கான இலக்கணங்களை எழுதுவதற்கான உண்மையான உள்ளீடுகளை வழங்க எழுதப்பட்ட இந்திய ஆங்கிலத்தின் தேர்ந்தெடுக்கப்பட்ட மாதிரிகளின் தனி தரவுத்தொகுதிகளைப் பயன்படுத்தியது.

1.4.6. பேச்சு ஆய்வு

காட்ஃப்ரே டீவி (Godfrey Dewey) பலவிதமான உரையாடல்களில் இருந்து சேகரிக்கப்பட்ட கிட்டத்தட்ட 100000 சொற்களின் பேச்சு தரவுத்தொகுதியைப் பகுப்பாய்வு செய்த பின்னர் ஆங்கில பேச்சு ஒலிகளின் சார்பு அதிர்வெண் (Relative Frequency of English Speech Sounds) (1923) என்பதை வெளியிட்டார்.

வில்லியம் விட்னி (William Whitney 1874) ஆங்கிலத்தில் பேச்சு அமைப்பொழுங்குகளைப் (speech patterns) படிக்க ஒரு தரவுத்தொகுதியைப் பயன்படுத்தினார்.

பன்சால் (Bansal 1969) இந்தியப் பேச்சு ஆங்கிலத்தின் (Indian Spoken English) பல்வேறு தன்மைகளைக் குறித்து ஆராயவும் பிரிட்டிஷ் மற்றும் இந்திய ஆங்கில மொழிக்கு இடையே இருக்கும் வேறுபாட்டின் தனிக்கூறுகளை அடையாளகாணவும் ஒப்பிடவும் ஒரு பேச்சு தரவுத்தொகுதியைப் பயன்படுத்தினார்.

1.4.7. மொழி கற்பித்தல்

மொழி கற்பித்தல் (Language Pedagogy) குறித்த ஆய்வுகளுக்கும் தொடக்க காலகட்டத்தில் தரவுகள் பயன்படுத்தப்பட்டன. பிரைஸ் மற்றும் டிராவர் (Fries an Traver, 1940)

மற்றும் போங்கர்ஸ் (Bongers, 1947) என்பவர்கள் அயல்மொழி கற்பித்தல் குறித்த ஆய்வுகளுக்கு விரிதரவைப் பயன்படுத்தினர். விரிதரவும் இரண்டாம் மொழி கற்பித்தலும் இருபதாம் நூற்றாண்டின் ஆரம்பகட்ட அரைப்பகுதியில் வலுவான தொடர்பு கொண்டிருந்தன. அயல்மொழி படிப்பவர்களின் சொற்தொகைப் பட்டியல் (Vocabulary list) விரிதரவுகளிலிருந்து எடுக்கப்பட்டன. தான்டைக் (Thorndike, 1921) மற்றும் பாமர் (Palmer, 1933) என்பவர்களின் ஆய்வுகள் அடிப்படையில் எடுக்கப்பட்ட சொல் எண்ணிக்கைகள் இரண்டாம் மொழி கற்பித்தலின் சொற்றொகைக் கட்டுப்பாட்டு இயக்கத்தின் முக்கிய இலக்காக அமைந்தது.

தோர்ன்டைக் (Thorndike,1921), பால்மர் (Palmer 1933), போங்கர்ஸ் (Bongers 1947) மற்றும் பிறர் வெளிநாட்டு மொழி கற்பித்தலில் (foreign language teaching) சொற்றொகைக் கட்டுப்பாட்டு இயக்கத்தின் (vocabulary control movement) வரலாறு மற்றும் கொள்கைகளை வரையறுக்க வெளிநாட்டு மொழி தரவுத்தொகுதிகளைப் (foreign language corpora) பயன்படுத்தினர்.

ஃப்ரைஸ் அண்ட் டிராவர் (Fries and Traver 1940) கற்பிப்பதில் வெளிநாட்டு மொழி கற்பித்தலின் இயல்பு மற்றும் நோக்கதை ஆராய அமெரிக்க கல்வி கவுன்சிலின் (American Council of Education) ஆங்கில தரவுத்தொகுதியைப் (English corpus) பயன்படுத்தினர்.

எச் வி. ஜார்ஜ் மற்றும் பலர் (H.V. George et al. 1950 –1960) ஹைதராபாத்தில் இந்திய மாணவர்களுக்கு ஆங்கிலம் கற்பிப்பதற்கான தொடக்கப்பாடநூல்களை (primers/ப்ரைமர்கள்) வடிவமைக்க நாவல்கள், நாடகங்கள், கதைகள் மற்றும் செய்தித்தாள்கள் மற்றும் பருவ இதழ்கள் உள்ளிட்ட புனைகதை அல்லாத மூலங்களிலிருந்து நூல்களைத் திரட்டுவதன் மூலம் எழுதப்பட்ட பிரிட்டிஷ் ஆங்கிலத்தின் அரை மில்லியன் சொற்களின் தரவுத்தொகுதியை உருவாக்கினர்.

1.4.8. மொழிப் பேறு

18 மற்றும் 19ஆம் நூற்றாண்டுகளில் மொழிப் பேறு குறித்த ஆராய்ச்சி பெரும்பாலும் 'நாட்குறிப்பேடு ஆய்வு' ('diary study') வடிவத்தில் செய்யப்பட்டது. இவை பெரும்பாலும் கவனமாக இயற்றப்பட்ட பெற்றோரின் நாட்குறிப்புகளை அடிப்படையாகக் கொண்டு அவற்றின் குழந்தையின் இருப்பிடங்களைப் பதிவுசெய்கின்றன (Preyer/ப்ரேயர் 1889, Stern/ஸ்டெர்ன் 1924).

20ஆம் நூற்றாண்டில் குழந்தைகளின் மொழி செயல்திறனின் அளவைப் பற்றிய பெரும்பாலான ஆராய்ச்சி மற்றும் மதிப்பீடுகள் குழந்தைகளின் பேச்சு மற்றும் எழுத்தில் இருந்து

உருவாக்கப்பட்ட தரவுத்தொகுதிகளுடன் செய்யப்படுகின்றன. தரவுத்தொகுதிகளின் பகுப்பாய்வுகள் குழந்தைகள் மொழியைச் சரியாகக் கற்றுக் கொண்டதா அல்லது மொழியியல்சார் திறன் குறைவாக உள்ளதா என்பதை அறிய முக்கியமான தடயங்களை வழங்குகின்றன.

'தகவலாளி அளவுகோலில்' 'informant scale' (1927-1957), மொழி பேறின் வளர்ச்சிக்கான பொதுவான விதிமுறைகளை நிறுவுவதற்கான நோக்கத்துடன், பகுப்பாய்வில் பயன்படுத்தப்படும் தரவுத்தொகுதியின் மாதிரி ஏராளமான குழந்தைகளிடமிருந்து (குறுகிய மற்றும் நிலையான நேர அளவிற்குள்) சேகரிக்கப்பட்டது (McCarthy/மெக்கார்த்தி 1954).

'நேர அளவுகோலில்' ('informant scale') (1957இலிருந்து) நீண்ட காலத்திற்குக் குறைந்த எண்ணிக்கையிலான குழந்தைகளிடமிருந்து பெறப்பட்ட கூற்றுகளின் தொகுப்பின் அடிப்படையில் ஒரு வகையான நீளப்பாங்கான ஆய்வு (longitudinal study) செய்யப்பட்டது (Bloom/ப்ளூம் 1970, Brown /பிரவுன் 1973).

1.4.9. ஒப்பீட்டு மொழியியல்

ஒப்பீட்டு மொழியியலும் (Comparative-Linguistics) தரவுத்தொகுதி பயன்படுத்தப்பட்டது. எடுத்துக்காட்டாக, ஈட்டன் (Eaton, 1940) டச்சு, ஜெர்மன், பிரெஞ்சு மற்றும் இத்தாலிய மொழிகளின் சொற்பொருள்களின் நிகழ்வெண்களை ஒப்பீட்டு செய்து ஆய்வு செய்தார். இவ்வாய்வு இன்றைய காலகட்டத்திலும் மிகச் சிறந்ததாகக் கருதப்படுகின்றது. 1990-களின் ஆரம்பத்தில்தான் விரிதரவுகள் மீண்டும் இம்மாதிரியான தகவல்களைப் பெறுவதற்காகப் பயன்படுத்த வேண்டி உருவாக்கப்பட்டன.

1.4.10. தொடரியலும் பொருண்மையியலும்

ஈட்டன்-ஆல் (Eaton 1940) பயன்படுத்தப்பட்ட பொருண்மையியல் நிகழ்வெண் பட்டியல்கள் ஒருமொழிய வர்ணனையில் (monolingual description) ஆர்வம் காட்டிய பிற ஆய்வாளர்களாலும் பயன்படுத்தப்பட்டது. லார்ஜ் (Lorge, 1949) இதற்கு எடுத்துக்காட்டாகும். தொடரியலும் இவ்வாறு பரிசோதிக்கப்பட்டது. ஃப்ரைசின் (Fries, 1952) விரிதரவு அடிப்படையிலான ஆங்கில வருணனை இலக்கணம் இதற்கு ஆரம்பகால எடுத்துக்காட்டாகும். குர்க் மற்றும் பிறரின் (Quirk et al) A Comprehensive Grammar of the English Language முப்பது ஆண்டுகளுக்குப் பிந்தியதாகும். இவ்வகை ஆய்வுகள் ஆங்கிலத்தோடு நின்றுவிடவில்லை. கோகன்ஹைம் மற்றும் பிறர் (Gogenheim et al, 1956) உயர்ந்த நிகழ்வெண் உள்ள சொல் தேர்வுகளையும் இலக்கண தேர்வுகளையும் விளக்க 275 தகவலாளிகளிடமிருந்து

எழுத்துப்பெயர்ப்பு செய்யப்பட்ட பேச்சுப் பிரெஞ்சு மொழியின் தரவுத்தொகுதியை பயன்படுத்தினர். இவ்வெடுத்துக்காட்டுகளிலிருந்து நீண்ட காலமாக மொழியியலில் அடிப்படை விரிதரவு நெறிமுறை விரிவாகப் பயன்படுத்தப்பட்டது என்பது தெரிய வருகிறது.

1.4.11. பிற துறைகள்

ஈட்டன் (1940) ஆங்கிலம், பிரஞ்சு, ஜெர்மன் மற்றும் ஸ்பானிஷ் மொழிகளில் சொல் நிகழ்வு மற்றும் அர்த்தங்களின் அதிர்வெண்ணை ஒப்பிட்டுப் பார்க்க 4 இணை தரவுத்தொகுதிகளைப் (parallel corpora) பயன்படுத்தினார்.

18ஆம் நூற்றாண்டில் விளக்க உரைகள் மற்றும் விமர்சனங்கள் செய்ய ஒரு தரவுத்தொகுதியாகப் பைபிள் பயன்படுத்தப்பட்டது. பல்வேறு பகுதிகளிடையே உண்மை நிலைத்தன்மையை நிரூபிக்க பல்வேறு சொல் பட்டியல்கள் (word lists) மற்றும் சொல்லடைவுகள் (concordances) உருவாக்கப்பட்டன (அலெக்சாண்டர் க்ரூடன் 1769). ஷேக்ஸ்பியர், மில்டன் மற்றும் பிறரின் படைப்புகளில் இதே போன்ற முயற்சிகள் மேற்கொள்ளப்பட்டன (Elliott and Valenza/எலியட் மற்றும் வலென்ஸா 1996).

1897ஆம் ஆண்டில் கோடிங் (Käding) என்ற ஜெர்மன் அறிஞர், கடிதங்களின் நிகழ்வெண் விநியோகம் (frequency distribution), சொற்களில் எழுத்துக்களின் வரிசையை எண்ணுதல் மற்றும் பொது எழுத்துக்கூட்டல் மரபுகளைப் (spelling conventions) படிப்பதற்காக பல்வேறு எழுதப்பட்ட மூலங்களிலிருந்து 11 மில்லியன் ஜெர்மன் சொற்களின் தரவுத்தொகுதியைக் கைமுறையாகத் தொகுத்தார்.

புள்ளியியல்சார் மொழியியலில், ஜே. பி. எஸ்டூப் (J. B. Estoup 1902) ஸ்டெனோகிராஃபிக்கான (stenography) கருவிகளை வடிவமைப்பதற்கான கிராஃபீம் நிகழ்வெண்களைத் தீர்மானிக்க எழுதப்பட்ட பிரெஞ்சு தரவுத்தொகுதியைப் பகுப்பாய்வு செய்தார்.

மார்க்கோவ் (Markov 1913) சில சிறிய ஆங்கில தரவுத்தொகுதிகளைப் பயன்படுத்தி தனது மொழிப் பொறியியல் ஆராய்ச்சியின் (language engineering research) மாதிரியை வடிவமைத்தார்.

ஜிப்ஃப் (Zipf 1936) சில மாதிரி தரவுத்தொகுதிகளுக்கு நெருக்கமான குறிப்புடன் ஆங்கிலத்தில் சொற்களின் விநியோகத்தை புள்ளியியல் அடிப்படையில் ஆய்வு செய்தார்.

1.5. தரவுத்தொகுதியின் வரம்புகள்

(அ) மொழியியல் உற்பத்தி இல்லாமை

சாம்ஸ்கியும் அவரது ஆதரவாளர்களும் மொழியியல் ஆராய்ச்சியில் தரவுத்தொகுதியின் மதிப்பை கடுமையாக விமர்சித்துள்ளனர். 1958இல் டெக்சாஸ் பல்கலைக்கழகத்தில், "எந்தவொரு இயற்கை தரவுத்தொகுதியும் திசை திருப்பப்படும். சில வாக்கியங்கள் வெளிப்படையானவை என்பதால் அவை ஏற்படாது; மற்றவை அவை பொய்யானவை, இன்னும் சிலவற்றில் அவை தவறானவை. தரவுத்தொகுதி, இயற்கையானதாக இருந்தால், அந்த விவரம் [அதன் அடிப்படையில்] வெறும் பட்டியலைத் தவிர வேறொன்றுமில்லை." என அவர் வாதிட்டார்,

மொழியியல் உள்ளார்ந்த தன்மைக்கு தரவுத்தொகுதி ஆதாரங்களை வழங்க முடியாது என்று மரபியலாளர்கள் வாதிடுகின்றனர். அதன் கட்டமைப்பு மற்றும் உள்ளடக்கம் ஆகியவற்றின் அடிப்படையில் இது மொழியியல் 'செயல்திறனை' மட்டுமே குறிக்கிறது, ஆனால் பயனர்களின் மொழியியல் 'திறன்' மற்றும் 'தாராள மனப்பான்மை' ஆகியவற்றைப் பிரதிபலிக்காது. செயல்திறனின் (performance) எடுத்துக்காட்டுகளை மட்டுமே பதிவுசெய்யும் ஒரு தரவுத்தொகுதி, மொழியியலாளர்களுக்கு பயனுள்ளதாக இருக்க இயலாது, அவர்கள் பல்வேறு சூழல்களில் மொழி பயன்பாட்டின் வெளிப்புற ஆதாரங்களை விட, மொழியைப் பற்றிய உள்ளார்ந்த அறிவைப் புரிந்துகொள்ள முற்படுகிறார்கள்.

(ஆ) தொழில்நுட்ப சிக்கல்கள்

தரவுத்தொகுதி உருவாக்குவது என்பது ஒரு பெரிய அளவிலான, பலதரப்பு, ஆர்வமுள்ள வேலை. இது ஒரு சிக்கலானது, நேரத்தை எடுத்துக்கொள்வது, பிழையானது மற்றும் விலை உயர்ந்தது. முழு நிறுவனத்திற்கும் திறமையான தரவு செயலாக்க அமைப்பு தேவைப்படுகிறது, இது அனைவருக்கும் கிடைக்காது, குறிப்பாக இந்தியா போன்ற ஒரு நாட்டில். மொழியியலாளர்கள் கணினி பயன்பாடு மற்றும் தரவு கையாளுதல் ஆகியவற்றில் பயிற்சி பெற வேண்டும். இது ஒரு தொந்தரவான பணி. மற்ற நாடுகளின் மொழியியலாளர்களைப் போலல்லாமல், இந்திய மொழியியலாளர்கள் கணினியை தங்கள் முன்னேற்றத்திற்கு கொண்டு செல்ல ஆர்வமாக இல்லை. கணினி விஞ்ஞானிகள், மறுபுறம், மொழியியலாளர்களுடன் இணைந்து பணியாற்ற ஆர்வமாக இல்லை. இடைவெளி இன்னும் பரந்த அளவில் உள்ளது. எதிர்காலத்தில் இருவருக்கும் இடையிலான பரஸ்பர கூட்டுறவு இடைமுகத்திற்காக நாம் நம்புகிறோம்.

(இ) உரையாடல் ஊடாட்டங்கள்சார் உரைகளின் குறைவு

இன்றைய தரவுத்தொகுதி தினசரி மொழிச் செயல்பாடுகளில் தன்னிச்சையாக நடைபெறுகின்ற, தயாரிக்கப்படாத உரையாடல்களைக் கருத்தில் கொள்ளத் தவறிவிட்டது. உரையாடல் ஊடாட்டம் இல்லாத பனுவல்கள் இல்லாதிருப்பது ஒரு தரவுத்தொகுதியை மனித மொழியின் மதிப்புமிக்க பண்பான தன்னிச்சையின் அம்சம் இல்லாததாக ஆக்குகிறது. தரவுத்தொகுதி, பேசப்படும் அல்லது எழுதப்பட்ட வடிவத்தில், உண்மையில் மொழி பயன்பாட்டின் உண்மையான சூழலில் இருந்து பிரிக்கப்பட்ட தரவுத்தளமாகும். சூழல்களிலிருந்து இல்லாத தன்மை பற்றின்மை ஒரு தரவுத்தொகுதியை உண்மையிலேய தரவுத்தொகுதி (corpus 'சடலம்') அதாவது ஒரு இறந்த தரவுத்தளமாக ஆக்குகின்றது. இத்தகைய தரவுத்தொகுதியில் வாழ்க்கையோடு ஒட்டிய அல்லது தன்னியல்பான உரையாடல் தொடர்புகள்/ஊடாட்டம் (living dialogic interactions), கருத்தாடல் (discourse) மற்றும் பயன்வழியியல் (pragmatics) இவற்றின் பண்புகள் இருக்காது.

இது ஒரு மொழிசார் பேச்சுவார்த்தை (linguistic negotiation) (கடினமான செயல் விளையாட்டு), மொழியில் பயன்பாட்டை (language-in-use) அடையாளம் காணுவது, உரையாடல்களுக்குள் (dialogues) சம்பந்தப்பட்ட வாய்மொழிச் செயல்களைத் தீர்மானிப்பது, உரையாடலில் பங்குபெறுவோரிலிருந்து புலனறிவுசார்ந்த மற்றும் புலனுணர்வு சார்ந்த தகவல்தொடர்பு வழிகளைப் பெறும் பின்னணியை விவரிப்பது என்பனவற்றிற்கு அடிப்படையான உண்மையான நோக்கத்தை வெளிப்படுத்தத் தவறிவிட்டது,

(ஈ) காட்சிக் கூறுகள்சார் தகவலின் பற்றாக்குறை

தரவுத்தொகுதியில் வரைபடங்கள், அட்டவணைகள், படங்கள், வரைபடங்கள், புள்ளிவிவரங்கள், உருவங்கள், சூத்திரங்கள் மற்றும் இதே போன்ற பிற காட்சிக் கூறுகள் இல்லை; அவை சரியான அறிவாற்றல் மற்றும் புரிதலுக்காக உரையின் அமைதியில் பெரும்பாலும் பயன்படுத்தப்படுகின்றன. இத்தகைய காட்சி கூறுகள் இல்லாத ஒரு தரவுத்தொகுதி அதன் பெரும்பாலான தகவல்களை இழக்க நேரிடும்.

இ) பிற வரம்புகள்

தரவுத்தொகுதி உருவாக்கமும் ஆராய்ச்சியும் செயற்கையாக எழுதப்பட்ட பனுவல்களை நோக்கி சாய்ந்தன; இது பேச்சின் முக்கியத்துவத்தை குறைத்தது. இருப்பினும், உண்மையில், பேச்சு எழுதுவதை விட நம்பகமான முறையில் மொழியைப் பிரதிபலிக்கிறது. பேச்சு தரவுத்தொகுதி உருவாக்கத்தின் சிக்கல்கள் அதை ஒரு அரிய பொருளாக ஆக்குகின்றன. உரை

தரவுத்தொகுதியின் எளிதாகக் கிடைக்கும் தன்மை மற்றும் பேச்சுத் தரவுத்தொகுதியின் பற்றாக்குறை உரை தரவுத்தொகுதியை நோக்கி திரும்ப மக்களை தூண்டுகிறது. இருப்பினும், தரவுத்தொகுதி மொழியியல் ஆராய்ச்சியில் பேச்சு தரவுத்தொகுதி முக்கியத்துவதஹை இழந்துவிட்டது என்பதை இது குறிக்கவில்லை. தரவுத்தொகுதியில் சேமிக்கப்பட்டுள்ள மொழி மொழியின் சில சமூக, தூண்டுதல் மற்றும் வரலாற்று அம்சங்களை முன்னிலைப்படுத்தத் தவறிவிட்டது.

ஒரு குறிப்பிட்ட கிளைமொழி ஏன் தரமானதாகப் பயன்படுத்தப்படுகிறது, குழு அடையாளத்தை (group identity) நிறுவுவதற்கும் பராமரிப்பதற்கும் கிளைமொழி வேறுபாடுகள் எவ்வாறு தீர்மானமான பங்களிப்பை வகிக்கின்றன, சமூகத்தில் ஒருவரின் சக்தி, நிலை மற்றும் அந்தஸ்தை எவ்வாறு தனிநபர் வழக்கு தீர்மானிக்கிறது, பொருண்மைக் களங்கள் (domains), நடை வேறுபாடுகள் (registers) போன்றவற்றைப் பொறுத்து மொழி எவ்வாறு வேறுபடுகிறது என்பதைத் தரவுத்தொகுதியால் வரையறுக்க இயலாது.

சில கவிதைகள், பாடல்கள் மற்றும் இலக்கியங்களால் சில உணர்ச்சிகள் எவ்வாறு தூண்டப்படுகின்றன என்பதையும், ஒரு கூற்றின் உத்தேசித்த அர்த்தத்தைத் தீர்மானிக்க உலக அறிவும் சூழலும் எவ்வாறு முக்கிய பங்கு வகிக்கின்றன என்பதையும், நேரம் மற்றும் சமுதாயத்தின் மாற்றத்துடன் மொழி எவ்வாறு உருவாகிறது, பிரிக்கிறது மற்றும் இணைகிறது என்பதையும் வெளிப்படுத்த தரவுத்தொகுதி தவறிவிட்டது.

1.6. தரவுத்தொகுதி மொழியியலின் மறுமலர்ச்சி

தரவுத்தொகுதி மொழியியல் 1950களில் முற்றிலுமாக கைவிடப்பட்டது; பின்னர்1980 களின் முற்பகுதியில் திடீரென மீண்டும் ஒரு முறை ஏற்றுக்கொள்ளப்பட்டது என்பது பொதுவான நம்பிக்கை. இது வெறுமனே பொய்யானது; மேலும் இந்த இடைக்காலத்தின் போது தரவுத்தொகுதி அடிப்படையிலான பணிகளை முன்னோடியாகக் கொண்ட மொழியியலாளர்களுக்கு ஒரு அவதூறு செய்தது.

எடுத்துக்காட்டாக, க்யூர்க் (1960) அவர் 1961இல் தொடங்கிய அவரது லட்சிய சர்வே ஆஃப் ஆங்கில பயன்பாட்டைத் (Survey of English Usage (SEU)) திட்டமிட்டு நிறைவேற்றினார். அதே ஆண்டில், பிரான்சிஸ் & குசெரா இப்போது பிரபலமான பிரவுன் தரவுத்தொகுதிப் பணியைத் தொடங்கினார்; இந்த ஒரு வேலை முடிக்க கிட்டத்தட்ட இரண்டு தசாப்தங்கள் ஆகும். இந்த ஆராய்ச்சியாளர்கள் சிறுபான்மையினராக இருந்தனர்; ஆனால் அவர்கள் உலகளவில்

விசித்திரமாக கருதப்படவில்லை, மற்றவர்கள் அவர்களின் வழியைப் பின்பற்றினர். 1975ஆம் ஆண்டில் ஜான் ஸ்வார்ட்விக் லண்டன்-லண்ட் தரவுத்தொகுதியை நிர்மாணிப்பதற்கான சர்வே ஆஃப் ஆங்கில பயன்பாடு மற்றும் பிரவுன் தரவுத்தொகுதியின் பணிகளை உருவாக்கத் தொடங்கினார்.

இந்த காலகட்டத்தில் கணினி மெதுவாக தரவுத்தொகுதி மொழியியலின் முக்கிய இடமாக மாறத் தொடங்கியது. ஸ்வார்ட்விக் சர்வே ஆஃப் ஆங்கில பயன்பாட்டைக் கணினிமயமாக்கியது, இதன் விளைவாக லீச் (1991) உட்பட சிலர் "இன்றுவரை பேசும் ஆங்கிலம் படிப்பதற்கான ஒப்பிடமுடியாத வளமாக" இருப்பதாக நம்புகிறார்கள்.

கணினிமயமாக்கப்பட்ட தரவுத்தொகுதியின் கிடைக்கும் தன்மை மற்றும் நிறுவன மற்றும் தனியார் கணினி வசதிகள் பரவலாகக் கிடைப்பது தரவுத்தொகுதியின் மறுமலர்ச்சிக்கு ஒரு ஊக்கத்தை அளித்ததாகத் தெரிகிறது.

1.6.1. தொடக்க காலத் தரவுத்தொகுதி குறித்த சாம்ஸ்கியின் கருத்து

சாம்ஸ்கி 1957-65-இல் மொழியியல் ஆய்வின் திசையை அனுபவவாதத்திலிருந்து (empiricism) பகுத்தறிவு வாதத்திற்கு (rationalism) குறைந்த கால கட்டத்திற்குள் மாற்றினார். இங்கு கேள்வி என்னவென்றால் அனுபவாவாதத்திற்கும் பகுத்தறிவு வாதத்திற்கும் வேறுபாடு இருக்கிறதா? இது மொழியியலுக்கு மட்டும் தனித்தன்மையானதா? இயற்கையாக நிகழும் கண்டறிதல் மீதான தீர்மானம் நம்பகமானதா? அல்லது செயற்கையாக ஊக்குவிக்கப்பட்ட உற்றுநோக்கல் நம்பகமானதா? என்ற அடிப்படையான தீர்மானத்திற்கு இது வழிவகுக்கின்றது. பகுத்தறிவாதத்தின் கோட்பாடு செயற்கையான நடத்தையின் தரவு மற்றும் உணர்வு பூர்வமான உள்முகத்தேடல் (Introspective) தீர்மானங்கள் அடிப்படையிலானது. இது ஒரு மொழியின் தாய்மொழி பேசுபவர் அம்மொழியைப் பிரதிபலிப்பதாகவும் அப்பிரதிபலிப்புகளின் அடிப்படையில் கோட்பாட்டு அடிப்படையிலான கூற்றுகளைச் செய்வதாகவும் இருக்கலாம். பகுத்தறிவு வாதத்தாரின் கோட்பாடுகள் மொழியியலின் நேர்வில் மனக் கோட்பாட்டின் வளர்ச்சியின் அடிப்படையில் அமைந்தது. இவை புலனறிவு சாத்தியத்தை அடிப்படை இலக்காகக் கொண்டது. இதன் நோக்கம் மனித மொழியாய்வின் வெளிவிளைவுகளை வெளிப்படுத்து மட்டுமல்லாமல் உண்மையில் எவ்வாறு இவ்வாய்வு செய்யப்படுகிறது என்பதை உருப்படுத்தம் செய்கின்றது என்று கூறுகின்ற ஒரு மொழிக் கோட்பாட்டை உருவாக்குவதாகும். மாறாக, மொழியின் அனுபவாத அணுகுமுறை இயற்கையாகவே நிகழும் தரவின் உற்றுநோக்கலால்

செய்யப்படுகின்றது. இந்நேர்வில் ஒரு மொழியின் ஒரு வாக்கியம் சரியானதா என்று அறிய அம்மொழியின் தரவு ஆயப்பட்டு தீர்மானிக்கப்படும். இந்த அனுபவவாத–பகுத்தறிவாதப் பாகுபாடு கோட்பாட்டைக் கூற பயன்படுத்தப்படும் தரவின் இயல்பின் அடிப்படையிலானது. இந்த இரண்டு அணுகுமுறைகளுக்கும் அனுகூலங்களும் அனுகூலமின்மைகளுக்கும் உண்டு. சாம்ஸ்கி மொழியியல் ஆய்வின் நோக்கத்தை மொழியின் அருவத்தன்மையான விளக்கங்களிலிருந்து மொழியின் உளவியல் உண்மையையும் புலனறிவுசார் சாத்தியமான மாதிரிகளையும் பிரதிபலிக்கும் கோட்பாடுகளுக்கு மாற்றினார். சாம்ஸ்கி இதன்படி மொழியாய்வில் பெருந்தரவைச் சான்றின் ஒரு மூலமாகக் கருதுவதை மதிப்பில்லாமல் ஆக்கினார். சாம்ஸ்கி மொழியியலாரின் மாதிரி செயல்திறனை விட அறிதிறனைதான் அடிப்படையில்தான் தங்கள் மாதிரியை அமைக்க முயல வேண்டும் என்ற காரணத்தால் விரிதரவு மொழியியலாருக்குப் பயனுள்ள கருவியாக இருக்கவியலாது என்று கூறினார். அறிதிறன் (competence)ஒரு மொழியின் உள்ளுறை செய்யப்பட்ட அறிவாகும். ஆனால் செயல்திறன் (performance) மொழி அறிதிறனின் புறச் சான்றாகும். சாம்ஸ்கி மொழியியலார் செயல்திறனைவிட அறிதிறனைத் தான் மாதிரியாக்கம் செய்ய முயன்றனர் என்று வாதிட்டார். செயல்திறன் அறிதிறனைப் பிரதிபலிக்காது என்று வாதிடப்பட்டது. செயல்திறன் நமது அறிதிறனிலிருந்து விலகிய காரணிகளால் பாதிக்கப்டலாம். ஒரு விரிதரவு அதன் இயல்பு அடிப்படையில் வெளிப்படுத்தப்பட்ட கூற்றுகள் ஆகும். அது செயல்திறனின் தரவாகும். இதன்காரணமாக அது மொழி அறிதிறனை மாதிரிப்படுத்த மோசமான வழிகாட்டியாகும். சாம்ஸ்கி அனுபவவாதத்திலிருந்து பகுத்தறிவாதத்திற்கு மாறத் தூண்டினார். மொழியை உற்றுநோக்கல் அடிப்படையில் விளக்காமல் உள்முகத் தேடல் (சிந்தனை) அடிப்படையில் விளக்க வேண்டும் என்றார்.

1.6.2. தரவுத்தொகுதி மொழியியலுக்கான வாதங்கள்

தரவுத்தொகுதி அடிப்படையிலான மொழியியல் ஆய்வுகளுக்கு எதிரான வாதங்கள் எல்லா விரிதரவு அடிப்படையிலான ஆய்வுகளையும் தடைசெய்து விடவில்லை. ஒலியியலில் இயல்பாகப் பெறப்பட்ட தரவுகள் சான்றுகளுக்கான ஆதிக்க மூலமாக இருந்துவந்தது. மொழிப்பேறு ஆய்வுக் களத்திலும் இயல்பாகக் கிடைக்கும் தரவுகள் ஆதிக்கம் செய்தன. குழந்தைகளின் மொழிப்பேறை ஆய்கின்ற மொழியிலார்களுக்கும் உளவியலார்களுக்கும் உள்முகத்தேடல் தீர்மானங்கள் (introspective judgments) கிடைக்கவில்லை. சாம்ஸ்கியே (1964) செயல்திறன் தரவுகளைச் (performance data) சான்றின் மூலமாகக் கருதாது தள்ளி வைத்தல் மொழிப்பேறு ஆய்வுகளுக்குப்

பொருந்தாது என்று கூறினார். பொதுவாகக் கூறினால் 1960களிலும் 1970களிலும் விரிதரவு நெறிமுறை நடைமுறைபடுத்தப்பட்டது. இயல்பாகக் கிடைக்கும் தரவு அதைப் பரிசோதனை செய்ய விரும்புகின்ற எல்லோராலும் உற்றுநோக்கப்படுவதற்கும் சரிபார்க்கப்படுவதற்கும் உட்பட இயலும் என்ற நிலையில் பயனுள்ளது. பேசுபவர் ஒரு உள்முகத்தேடல் தீர்மானத்தைச் செய்கையில் அது சரியானது என்று எவ்வாறு கூற இயலும்? அவர் ஒரு வாக்கியத்தைக் கூறும்போது நாம் அதை உற்றுநோக்கிப் பதிவு செய்யவியலும். ஆனால் அவர் தன் கருத்தை ஒரு சிந்தனைச் செயற்பாங்கால் வெளியிட்டால் நாம் என்ன செய்யவியலும்? அது உற்றுநோக்கவியலாததாக இருக்கும். பதிவு செய்யப்பட்ட வாக்கியத்தால் நாம் மக்களின் கருத்தைப் பெறவியலும். தரவு எல்லோராலும் உற்றுநோக்கத்தக்கதும் விமரிசிக்கத்தக்கதும் ஆகும். மக்கள் கருத்து மற்றும் அதற்கு மாறான தனிநபர் கருத்து என்ற சிக்கல் மொழியியலை மட்டுமற்றி பிற துறைகளையும் பாதித்துள்ளது. தரவுத்தொகுதி அடிப்படையிலான உற்றுநோக்கல்கள் உள்முகத்தேடல் அடிப்படையிலான தீர்மானங்களைவிட மெய்யாகச் சரிபார்க்க உகந்தது. செயற்கையான தரவு செயற்கையானதாகத் தான் இருக்கும். சாம்சன் (Samson, 1992:428) உள்முகத்தேடல் மொழியியலார்களால் ஆயப்பட்ட வாக்கிய வகை தரவுத்தொகுதியில் கிடைக்கும் சான்றுகளின் வகையிலிருந்து மிக விலகி இருக்கின்றது என்று கூறுகின்றார். நிகழ்வெண் மற்றொரு செய்தியை முன்வைக்கின்றது. சில வகை மொழித் தரவுகளை விரிதரவிலிருந்து மட்டுமே சரியாகப் பெறவியலும். மனிதவினம் ஒரு கட்டுமானத்தின் அல்லது சொல்லின் நிகழ்வெண் குறித்த மிகத் தெளிவற்ற கருத்துச்சாயலைத்தான் கொண்டிருக்கின்றது. தரவின் இயல்பான உற்றுநோக்கல் மட்டும் தான் நிகழ்வெண் போன்ற பண்புக்கூறுகளுக்கான சான்றின் நம்பகமான மூலமாக இருக்கின்றது. சாம்ஸ்கி இம்மாதிரியான அளவுசார் தகலைப் (quantitative information) பெறுதல் நன்மை பயப்பதல்ல என்று இச்செய்திக்கு எதிராகக் கூறலாம். லீச் (1992) விரிதரவு விளைவுகளின் புறவயமான சரிபார்ப்புக்கு (objective verification) அனுமதிப்பதால் அறிவியல் நெறிமுறைப் பார்வையின் படி விரிதரவு கூடுதல் வலுவான நெறிமுறையாகும் என்று வாதிடுகின்றார். சாம்ஸ்கி (Chomsky, 1965) மேம்போக்காக 95 விழுக்காடு கூற்றுகள் இலக்கணத்தன்மையற்றது என்று வகைப்படுத்துகின்றார்; எனவே செயல்திறன்சார் தரவு அறிதிறனின் தவறான உருப்படுத்தம் என்கின்றார். இக்கூற்று மிக எளிதாக்கப்பட்ட ஒன்றாகும். லபோவ் (Labov, 1969) எல்லாச் சூழல்களிலும் கூற்றுக்களின் மிகப்பெரும்பான்மையானவை இலக்கணத்தன்மையானவை என்று காட்டுகின்றார். விரிதரவு

அடிப்படையிலான ஆய்வு ஆரம்பத்தில் கருதப்பட்டதுபோல் செல்லாதது அல்ல. விரிதரவு இலக்கணத்தன்மையற்ற வாக்கியங்களின் தொகுதியாக இருக்கத் தேவையில்லை. தரவுத்தொகுதி பொதுவாக இலக்கணத்தன்மையான வாக்கியங்களைக் கொண்டிருக்கும் என்று நம்புவதற்குக் காரணம் இருக்கின்றது என்று தோன்றுகின்றது. நாம் விரிதரவிலுள்ள எல்லா வாக்கியங்களும் இலக்கணத் தன்மையானவை என்று கூறவில்லை. தரவுத்தொகுதிகள் அளவுசார் தரவின் (quantative data) மிகச் சிறப்பான மூலங்களாகும். ஆனால் சாம்ஸ்கி அளவுசார் தரவு மொழியியலுக்குப் பயன்படாது என்று பதில் கூறலாம். ஆனால் அவருடைய கருத்து உண்மை நிகழ்வால் ஆதரிக்கப்படவில்லை என்று நாம் கூறலாம். தரவுத்தொகுதி அல்லது சான்றின் ஒப்பிடவியலும் அளவுசார் தரவைத் தரும் சில பிற இயற்கை மூலமில்லாமல் மொழியியலார்களுக்கோ கணினி மொழியலார்களுக்கோ வலுவான ஆய்வுக் கருவிகள் இருந்திருக்காது. சாம்ஸ்கியால் தரவுத்தொகுதி பற்றி கூறப்பட்ட சில திறனாய்வுகள் பகுதி செல்லத்தக்கதாகும். அவை இன்றைய தரவுத்தொகுதிகள் பற்றி கூடுதல் மெய்யான மனப்பாங்கு வளர உதவுகின்றது. ஆனால் இத்திறனாய்வுகள் (குறைகள்) பகுதி செல்லாதது; அவை தரவுத்தொகுதியின் வலுவின்மைக்கு எதிரான வலுவின் உண்மையான கணிப்பை விலக்குகின்றது. இவ்வுற்றுநோக்கல் ஏன் சிலர் தரவுத்தொகுதி அடிப்படையில் ஆய்வை மேற்கொண்டு வருகின்றனர் என்றும் ஏன் அவர்கள் ஆரம்ப கால விரிதரவு குறித்த முக்கியமான குறைகூறலை நேரிட்டாலும் அதைத் தொடர்கின்றனர் என்றும் குறிப்பிடுகின்றது. இப்போது விரிதரவு என்பது இயந்திரத்தால் படிக்கவியலும் தரவுத்தொகுதி என்ற சொல்லின் ஒருபொருள் பன்மொழியாகப் பயன்படுத்தப்படுகின்றது. இது எதேட்சையானது அல்ல. கணிப்பொறி பயன்படுத்தவேண்டிய போலி உத்தி என்ற வட்டத்திற்குள் இருப்பதாக முன்னர் கருதப்பட்ட உத்திகளை அனுமதிக்கின்து. கணிப்பொறி அதன் தரவைத் தேடும், மீட்கும் மற்றும் கணிக்கும் திறமை காரணமாக இதை அனுமதிக்கின்றது.

சாம்ஸ்கி வாதிட்டபடி அனுபவ மொழித் தரவுத்தளம் (empirical language database) உண்மையில் ஒரு மனிதனின் மொழியியல் திறனின் மோசமான பிரதிபலிப்பா என்ற ஐயம் எழலாம். ஒருவேளை 'இல்லை' ஏனெனில், இயற்கையாகவே நோக்கப்பட்ட தரவு ஒலிப்பியல் தொடர்பான ஆதாரங்களின் மூலத்தை ஆதிக்கம் செலுத்துகிறது; அங்கு அகநோக்குப்பார்வைக்கு (introspection) எந்தப் பங்கும் இல்லை. இயற்கையாக நிகழும் சான்றுகள் பற்றிய நோக்கல் (observation) குழந்தை மொழிப் பேறில் ஆதிக்கம் செலுத்துகிறது; அங்கு உளவியலாளர்கள்

அகநோக்குப்பார்வை அடிப்படையிலான தீர்ப்புகளை நம்பகமான ஆதாரமாகக் கருதுவதில்லை. சாம்ஸ்கி (1964) கூட செயல்திறன் தரவின் (performance data) மதிப்பை மொழி பேறு (language acquisition) மற்றும் ஒலிப்பியல் ஆகியவற்றில் ஆதாரமாக ஒப்புக் கொண்டார்.

தரவுத்தொகுதியின் உருவாக்கம் சில காலம் நிறுத்தப்பட்டாலும், 1950க்கும் 1980க்கும் இடையிலான காலகட்டத்தில் முழுமையாக நிறுத்தப்படவில்லை. ஆர். க்யூர்க் (R. Quirk), என். பிரான்சிஸ் (N. Francis), எச். குசெரா (H. Kucera), ஜே. ஸ்வார்ட்விக் (J. Svartvik), ஜி. லீச் (G. Leech) மற்றும் பலர் தரவுத்தொகுதி உருவாக்கம் மற்றும் பகுப்பாய்வுகளுடன் தொடர்ந்தனர். 1961ஆம் ஆண்டில், ஆர். க்யூர்க் இங்கிலாந்தில் சர்வே ஆஃப் ஆங்கிலப் பயன்பாடு (Survey of English Usage) என்பதை உருவாக்கத் தொடங்கினார். 1961ஆம் ஆண்டில், என். பிரான்சிஸ் மற்றும் எச். குசெரா ஆகியோர் அமெரிக்காவின் பிரவுன் பல்கலைக்கழகத்தில் பிரவுன் தரவுத்தொகுதி (Brown Corpus) என்பதன் உருவாக்கத்தைத் தொடங்கினர். 1975ஆம் ஆண்டில், ஜே. ஸ்வார்ட்விக் லண்டனில் லண்டன்-லண்ட் கார்பஸ் ஆஃப் ஆங்கிலம் (London-Lund Corpus of English) என்பதை உருவாக்கத் தொடங்கினார்.

உயர் கம்ப்யூட்டிங் வசதி கிடைப்பது தரவுத்தொகுதியின் உலகளாவிய வளர்ச்சிக்குப் புதிய வாழ்க்கையை வழங்கியது. நான்கு பதின்ம ஆண்டுகளுக்குள் மின்னணு தரவுத்தொகுதியின் எண்ணிக்கை 1961இல் 1 முதல் 2000இல் 5000க்கும் அதிகமாக வளர்ந்தது. தரவுத்தொகுதியின் அளவும் அதிகரித்தது. பிரவுன் தரவுத்தொகுதி (Brown Corpus), லோப் தரவுத்தொகுதி (Lancaster-Oslo/Bergen Corpus (LOB) Corpus), கே.சி.ஐ.இ. தரவுத்தொகுதி (Kolhapur Corpus of Indian English (KCIE)) என்பன தலா 1 மில்லியன் வார்த்தைகளைக் கொண்டிருந்தன. பர்மிங்காம் ஆங்கில உரையின் தொகுப்பு (Birmingham Collection of English Text) 20 மில்லியன் சொற்களையும், பாங்க் ஆப் ஆங்கிலம் (Bank of English) 200 மில்லியன் சொற்களையும், பிரிட்டிஷ் நேஷனல் தரவுத்தொகுதி (British National Corpus) 400 மில்லியன் சொற்களையும், அமெரிக்கன் நேஷனல் தரவுத்தொகுதி (American National Corpus) 200 மில்லியன் சொற்களையும் கொண்டுள்ளது. தற்போது, உலகெங்கிலும், மொழியியல் மற்றும் மொழி தொழில்நுட்பத்தின் அனைத்து துறைகளிலிருந்தும் மக்கள் தரவுத்தொகுதி உருவாக்கத்தில் ஈடுபட்டுள்ளனர். பண்பாட்டு பண்புகளை விவரிக்க அவர்கள் அதை நம்பகமான வளமாகப் பயன்படுத்துகிறார்கள்; மனித மொழியியல் செயல்திறனைப் பகுப்பாய்வு செய்யுங்கள்; சமூக சூழல்களில் மொழியைக் கவனித்தல்; பகுதிகள், வகைகள் மற்றும் புலங்கள் முழுவதும் மொழியை

ஒப்பிடுக; மொழி தொழில்நுட்பத்திற்கான மென்பொருள், அமைப்புகள் மற்றும் கருவிகளை உருவாக்குதல்; அகராதிகள் மற்றும் குறிப்பு புத்தகங்களை உருவாக்குதல்; மொழி கற்பித்தல் பொருட்கள் மற்றும் அமைப்புகளை வடிவமைத்தல். வெளியீடுகளின் அங்கீகாரம் மற்றும் சரிபார்ப்பு ஆகியவற்றில் ஒப்பீட்டு மதிப்பீட்டைச் செய்தால், உற்பத்தி சட்டத்திலிருந்து அனுபவ மாதிரிக்கான அணுகுமுறையின் மாற்றம் தெளிவாகிறது.

1.6.3. தரவுத்தொகுதி மொழியியலின் பயன்பாடு

தரவுத்தொகுதி மொழியியல் பல ஆராய்ச்சி முறைகளை உருவாக்கியுள்ளது; அவை தரவிலிருந்து கோட்பாட்டிற்கு ஒரு பாதையை கண்டுபிடிக்க முயற்சிக்கின்றன. வாலிஸ் மற்றும் நெல்சன் (Wallis and Nelson 2001) முதலில் அவர்கள் 3A (Annotation, Abstraction and Analysis) கண்ணோட்டம் (perspective) என்று அழைத்ததை அறிமுகப்படுத்தினர்: விவரம் அடையாளப்படுத்தல் (Annotation), சுருக்கம் (Abstraction) மற்றும் பகுப்பாய்வு (Analysis).

விவரம் அடையாளப்படுத்தல் (Annotation) என்பது பனுவல்களுக்கு ஒரு திட்டத்தின் பயன்பாட்டைக் கொண்டுள்ளது. விவரம் அடையாளப்படுத்தும் கட்டமைப்பு, மார்க்அப் (structural markup), சொல்வகைப்பாடு அடையாளப்படுத்தல் (part-of-speech tagging), பாகுபடுத்தல் (parsing) மற்றும் பிற பல உருப்படுத்தங்களைக் (பிரதிநிதித்துவங்களைக்) கொண்டிருக்கலாம்.

சுருக்கம் (Abstraction) என்பது திட்டத்தில் உள்ள சொற்களின் கோட்பாட்டு அடிப்படையில் ஊக்கப்படுத்தப்பட்ட மாதிரி அல்லது தரவுத்தொகுப்பில் உள்ள சொற்களின் மொழிபெயர்ப்பை (mapping மேப்பிங் பொருத்தத்தை) கொண்டுள்ளது. சுருக்கம் பொதுவாக மொழியியலாளர்-இயக்க தேடலை உள்ளடக்கியது; ஆனால் எடுத்துக்காட்டாக, பாகுபடுத்துபவைகளுக்கான விதிமுறை கற்றலும் இதில் அடங்கலாம்.

பகுப்பாய்வு (analysis) தரவுத்தொகுப்பிலிருந்து (dataset) புள்ளிவிவர அடிப்படையில் ஆய்வு செய்தல், கையாளுதல் மற்றும் பொதுமைப்படுத்துதல் ஆகியவற்றைக் கொண்டுள்ளது. பகுப்பாய்வில் புள்ளிவிவர மதிப்பீடுகள், விதி-தளங்களின் தேர்வுமுறை அல்லது அறிவு கண்டுபிடிப்பு முறைகள் ஆகியவை இருக்கலாம்.

இன்று பெரும்பாலான சொல்சார் தரவுத்தொகுதி, சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட (part-of-speech-tagged/POS-tagged) பகுதியாகும். இருப்பினும், 'விவரம் அடையாளப்படுதப்படாத சாதாரண உரையுடன்' ('unannotated plain text') செயல்புரியும்

தரவுத்தொகுதி மொழியியலாளர்கள் கூட முக்கிய சொற்களைத் தனிமைப்படுத்த சில முறைகளைத் தவிர்க்க முடியாமல் பயன்படுத்துகிறார்கள். இத்தகைய சூழ்நிலைகளில் விவரம் அடையளப்படுத்தல் மற்றும் சுருக்கம் ஆகியவை ஒரு சொல்சார் தேடலில் இணைக்கப்படுகின்றன.

விவரம் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியை வெளியிடுவதன் நன்மை என்னவென்றால், பிற பயனர்கள் தரவுத்தொகுதியில் (தரவுத்தொகுதி மேலாளர்கள் மூலம்) சோதனைகளைச் செய்யலாம். தோற்றுவிப்பாளர்களைக் காட்டிலும் பிற ஆர்வங்கள் மற்றும் மாறுபட்ட கண்ணோட்டங்களைக் கொண்ட மொழியியலாளர்கள் இந்த வேலையைப் பயன்படுத்த முடியும். தரவைப் பகிர்வதன் மூலம், தரவுத்தொகுதி மொழியியலாளர்கள் தரவுத்தொகுதியை மொழியியல் விவாதம் மற்றும் மேலதிக ஆய்வின் ஒரு இடமாகக் கருத முடிகிறது.

இயல் 2

தரவுத்தொகுதி மொழியலின் இன்றைய நிலை

2.1. அறிமுகம்

மின் தரவுத்தொகுதியின் அறிமுகம் 1960களில் ஏற்பட்டது. நாற்பது ஆண்டுகளுக்குள் தனிநபரின் ஆர்வத்தாலோ கல்வி மற்றும் ஆய்வு நிறுவனங்களின் ஊக்கத்தாலோ உலகின் பல பாகங்களிலும் பல மொழி விரிதரவுகள் திட்டமிடப்பட்டு உருவாக்கப்பட்டுள்ளன. நாம் அடிப்படையான பண்புக்கூறுகள் அடிப்படையில் தரவுத்தொகுதிகளை இரண்டு வகைகளாகப் பகுக்கலாம்: அவை எழுத்து விரிதரவு மற்றும் பேச்சுத் தரவுத்தொகுதி. தரவுத்தொகுதிகளின் சில எடுத்துக்காட்டுகள் இங்கு தரப்பட்டுள்ளன. ஒவ்வொன்றும் அவற்றின் தகவல் மற்றும் பொருளடக்கம் அடிப்படையில் விளக்கப்பட்டுள்ளன.

2.2. தரவுத்தொகுதி உருவாக்கத்தின் வரலாற்றுச் சுருக்கம்

இலக்கண விளக்கத்தின் ஆரம்பகால முயற்சிகள் சில குறிப்பிட்ட மத அல்லது கலாச்சார முக்கியத்துவம் வாய்ந்த நிறுவனங்களை அடிப்படையாகக் கொண்டவை. எடுத்துக்காட்டாக, ப்ரதிஷிகா (Prātiśākhya) இலக்கியம் வேதங்களில் காணப்படும் சமஸ்கிருதத்தின் ஒலி வடிவங்களை விவரித்தது, மேலும் பைனியின் செம்மொழி சமஸ்கிருதத்தின் (classical Sanskrit) இலக்கணம் குறைந்தபட்சம் அதே கார்பஸின் பகுப்பாய்வை அடிப்படையாகக் கொண்டது. இதேபோல், ஆரம்பகால அரபு இலக்கண வல்லுநர்கள் குர்ஆனின் மொழியில் குறிப்பாக கவனம் செலுத்தினர். மேற்கத்திய ஐரோப்பிய பாரம்பரியத்தில், அறிஞர்கள் பைபிளின் மொழி மற்றும் பிற நியமன நூல்களை விரிவாகப் படிக்க அனுமதிக்க தொடரடைவுகளைத் தயாரித்தனர்.

நவீன தரவுத்தொகுதி மொழியியலில் ஒரு மைல்கல் 1967ஆம் ஆண்டில் ஹென்றி குசெரா (Henry Kučera) மற்றும் டபிள்யூ. நெல்சன் பிரான்சிஸ் (W. Nelson Francis) ஆகியோரால் வெளியிடப்பட்ட தற்போதைய அமெரிக்க ஆங்கிலத்தின் கணக்கீட்டு பகுப்பாய்வின் (Computational Analysis of Present-Day American English) ஆகும்; இது தற்போதைய அமெரிக்க ஆங்கிலத்தைக் கவனமாகத் தேர்வு செய்து தொகுத்த பல்வேறு வகையான மூலங்களிலிருந்து பெறப்பட்ட மொத்தம் ஒரு மில்லியன் வார்த்தைகளைக் கொண்ட பிரவுன் விரிதரவின் பகுப்பாய்வை அடிப்படையாகக் கொண்டது. குசெரா மற்றும் பிரான்சிஸ் இதைப் பலவிதமான கணக்கீட்டு பகுப்பாய்வுகளுக்கு உட்படுத்தினர்; அதில் இருந்து அவர்கள் மொழியியல், மொழி கற்பித்தல், உளவியல், புள்ளிவிவரங்கள் மற்றும் சமூகவியல் ஆகியவற்றின் கூறுகளை இணைத்து

ஒரு செழிப்பான மற்றும் மாறுபட்ட ஓபஸை தொகுத்தனர். மேலும் முக்கிய வெளியீடு ராண்டால்ஃப் க்யூர்க் (Randolph Quirk) ஆங்கிலப் பயன்பாட்டின் மதிப்பாய்வு (The Survey of English Usage) என்பதை அறிமுகப்படுத்திய 'ஆங்கில பயன்பாட்டின் விளக்கத்தை நோக்கி' ('Towards a description of English Usage') ஆகும் (Quirk 1960).

அதன்பிறகு, பாஸ்டன் வெளியீட்டாளர் (Boston publisher) ஹாட்டன்-மிஃப்ளின் (Houghton-Mifflin) அதன் புதிய அமெரிக்க பாரம்பரிய அகராதி-க்கு (American Heritage Dictionary (AHD)) ஒரு மில்லியன் வார்த்தைகள், மூன்று வரி மேற்கோள் தளத்தை வழங்குவதற்காக குசெராவை அணுகினார்; இது தரவுத்தொகுதி மொழியியலைப் பயன்படுத்தி தொகுக்கப்பட்ட முதல் அகராதி ஆகும். பரிந்துரைக்கப்பட்ட கூறுகளை (மொழி எவ்வாறு பயன்படுத்தப்பட வேண்டும்) விளக்கமான தகவலுடன் (அது உண்மையில் எவ்வாறு பயன்படுத்தப்படுகிறது) இணைப்பதற்கான புதுமையான நடவடிக்கையை அமெரிக்க பாரம்பரிய அகராதி எடுத்தது.

பிற வெளியீட்டாளர்கள் இதைப் பின்பற்றினர். பிரிட்டிஷ் வெளியீட்டாளர் காலின்ஸின் கோபில்ட் ஒருமொழி கற்பவர் அகராதி (Collins' COBUILD monolingual learner's dictionary), ஆங்கிலத்தை வெளிநாட்டு மொழியாகக் கற்கும் பயனர்களுக்காக வடிவமைக்கப்பட்டுள்ளது; இது ஆங்கிலத்தின் வங்கியைப் (Bank of English) பயன்படுத்தி தொகுக்கப்பட்டது. ஆங்கிலப் பயன்பாட்டின் மதிப்பாய்வுத் (Survey of English Usage) தரவுத்தொகுதி மிக முக்கியமான தரவுத்தொகுதி அடிப்படையிலான இலக்கணங்களில் ஒன்றான ஆங்கிலத்தின் விரிவான இலக்கணம் (Comprehensive Grammar of English) (க்யூர்க் மற்றும் பலர்/Quirk et al. 1985 1985) என்பதன் உருவாக்கத்தில் பயன்படுத்தப்பட்டது.

பிரவுன் தரவுத்தொகுதி இதேபோன்ற பல கட்டமைக்கப்பட்ட தரவுத்தொகுதிகளையும் உருவாக்கியுள்ளது: லாப் தரவுத்தொகுதி (1960களின் பிரிட்டிஷ் ஆங்கிலம்), கோலாப்பூர் தரவுத்தொகுதி (இந்திய ஆங்கிலம்), வெலிங்டன் தரவுத்தொகுதி (நியூசிலாந்து ஆங்கிலம்), ஆஸ்திரேலிய கார்பஸ் ஆஃப் ஆங்கிலம் (ஆஸ்திரேலிய ஆங்கிலம்), ஃப்ரவுன் தரவுத்தொகுதி (1990களின் முற்பகுதி அமெரிக்கன் ஆங்கிலம்), மற்றும் FLOB தரவுத்தொகுதி (1990களில் பிரிட்டிஷ் ஆங்கிலம்). பிற தரவுத்தொகுதிகள் பல மொழிகள், வகைகள் மற்றும் முறைகளைப் உருப்படுத்தம்செய்கின்றன; மேலும் இவை சர்வதேச ஆங்கில தரவுத்தொகுதி (International Corpus of English) மற்றும் பேச்சு (spoken) மற்றும் எழுத்துப் (written) உரைகளின் (texts) 100

மில்லியன் சொற்களின் சேகரிப்பு கொண்ட வெளியீட்டாளர்கள் (publishers), பல்கலைக்கழகங்கள் (ஆக்ஸ்போர்டு மற்றும் லான்காஸ்டர்) மற்றும் பிரிட்டிஷ் நூலகம் இவற்றில் கூட்டமைப்பால் 1990களில் உருவாக்கப்பட்ட பிரிட்டிஷ் நேஷனல் தரவுத்தொகுதி (British National Corpus) இவற்றையும் உட்படுத்துகின்றன. சமகால அமெரிக்க ஆங்கிலத்தைப் பொறுத்தவரை, அமெரிக்க தேசியத் தரவுத்தொகுதியின் பணிகள் நிறுத்தப்பட்டுள்ளன; ஆனால் 400+ மில்லியன் சொற்கள் கொண்ட தற்கால அமெரிக்க ஆங்கித் தரவுத்தொகுதி (Corpus of Contemporary American English) (1990 - தற்போது வரை) இப்போது ஒரு வலை இடைமுகத்தின் மூலம் கிடைக்கிறது.

எழுத்தாக்கம் செய்யப்பட்ட பேசும் மொழியின் முதல் கணினிமயமாக்கப்பட்ட தரவுத்தொகுதி 1971ஆம் ஆண்டில் மாண்ட்ரீல் பிரஞ்சு திட்டத்தால் (Montreal French Project) கட்டமைக்கப்பட்டது; இதில் ஒரு மில்லியன் சொற்கள் உள்ளன; இது ஒட்டாவா-ஹல் (Ottawa-Hull) பகுதியில் பேசப்படும் பிரெஞ்சு மொழியின் ஷானா போப்லக்கின் (Shana Poplack's) மிகப் பெரிய தரவுத்தொகுதியை ஊக்குவித்தது.

இந்த வாழும் மொழிகளின் தரவுத்தொகுதியைத் தவிர, கணினிமயமாக்கப்பட்ட தரவுத்தொகுதிகளும் பண்டைய மொழிகளில் உரைகளின் தொகுப்பால் உருவாக்கப்பட்டுள்ளன. 1970களில் இருந்து உருவாக்கப்பட்ட ஹீப்ரூ பைபிளின் (Hebrew Bible) ஆண்டர்சன்-ஃபோர்ப்ஸ் தரவுத்தளம் (Andersen-Forbes database) ஒரு எடுத்துக்காட்டு ஆகும்; இதில் ஒவ்வொரு எச்சத்தொடர்களும் தொடரியலின் ஏழு நிலை வரை உருப்படுத்தம்செய்யும் வரைபடங்களைப் பயன்படுத்தி பாகுபடுத்தப்பட்டுள்ளன; மேலும் ஒவ்வொரு கூறும் ஏழு புலத் தகவல்களுடன் அடையாளப்படுத்தப்பட்டுள்ளன. குர்ஆனிய அரபு தரவுத்தொகுதி (The Quranic Arabic Corpus) என்பது குர்ஆனின் செம்மொழி அரபு மொழிக்கான (Classical Arabic language of the Quran) அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி (annotated corpus) ஆகும். இது ஒரு சமீபத்திய திட்டமாகும்; இது உருபனியல் பகுப்பாய்வு, சொல்வகைப்பாடு அடையாளபடுத்தல் (part-of-speech tagging) மற்றும் சார்பு இலக்கணத்தைப் பயன்படுத்தி (dependency grammar) தொடரியல் பகுப்பாய்வு (syntactic analysis) உள்ளிட்ட பல அடுக்கு அடையாளப்படுத்தல்களைக் கொண்டுள்ளது.

தூய மொழியியல் விசாரணையைத் தவிர, தரவுத்தொகுதி தரவு மற்றும் கருவிகளைப் பயன்படுத்தி சட்ட நூல்களைப் புரிந்துகொள்ள முற்படும் சட்டத்தின் வளர்ந்து வரும் துணைத்

துறை மற்றும் தரவுத்தொகுதி மொழியியல் போன்ற பிற கல்வி மற்றும் தொழில்முறைத் துறைகளுக்கு ஆராய்ச்சியாளர்கள் தரவுத்தொகுதி மொழியியலைப் பயன்படுத்தத் தொடங்கினர்

2.3. எழுத்துத் தரவுத்தொகுதி

இணையத்தில் தரவுத்தொகுதிக்கான பெரும்பாலான பொருட்களை இலவசமாகப் பெறலாம். இந்த உரை பொருட்கள் பதிவிறக்கம் செய்ய கிடைக்கின்றன; சிடி-ரோம் அல்லது பிற ஊடகங்களில் உள்ள தரவுத்தொகுதிகள் மற்றும் பிற பொருட்களுக்கு நிறைய செலவாகும். கணினி காப்பகங்களில் பெரிய அளவிலான பிற பொருட்களும் உள்ளன; அவை கல்வியாளர்களால் வணிகரீதியான பயன்பாட்டிற்கு எளிதாகக் கிடைக்கின்றன. எழுத்துத் தரவுத்தொகுதிகளின் சில எடுத்துக்காட்டுகள் கீழே தரப்பட்டுள்ளன.

2.3.1. பிரவுன் தரவுத்தொகுதி

பிரவுன் தரவுத்தொகுதி (Brown Corpus) (இன்றைய அமெரிக்க ஆங்கிலத்தின் நிலையான/தரமான மாதிரி (The standard sample of present day American English))1961–இல் பிரவுன் பல்கலைக் கழகத்தில் நெல்சன் பிரான்சிஸ் ஹென்றி குச்சேரா (W. Nelson Francis and Henry Kuchera) என்பவர்களால் உருவாக்கப்பட்ட முதல் மின் தரவுத்தொகுதி ஆகும். இது 1961-ஆம் ஆண்டு (நாள் காட்டி ஆண்டு) யுஎஸ்ஏ-இல் (USA) அச்சடிக்கப்பட்ட ஆங்கில உரை நடையின் தொடர்ச்சியான உரைகளின் 1014312 சொற்களைக் கொண்டது. இது 2000+ சொற்களைக் கொண்ட 500 மாதிரிகளாகப் (sample) பகுக்கப்பட்டுள்ளது. ஒவ்வொரு மாதிரியும் ஒரு வாக்கியத்தின் தொடக்கத்தில் தொடங்குகிறது ஆனால் ஒரு பத்தியின் அல்லது ஒரு பெரிய பகுதியின் தொடக்கத்தில் தொடங்குகிறது என்பது தேவை இல்லை. ஒவ்வொன்றும் 2000 சொற்களுக்குப் பின் இறுதியுறும் முதல் வாக்கியத்தில் முடிகிறது. ஒவ்வொரு மாதிரியும் உரைநடையின் நடைகள் மற்றும் வகைகளின் விரிந்த பரப்பெல்லையை பிரதிநித்துவம் செய்யும். செய்யுள்களும் நாடகங்களில் உள்ள உரைகளும் எழுத்துக் கருத்தாடல்களுக்கு அப்பாற்பட்ட பேச்சு கருத்தாடலின் கற்பனையான பொழுது போக்காக இருப்பதன் காரணமாக அவைகள் விலக்கி வைக்கப்படுகிறது. கதைகள் உட்படுத்தப்பட்டுள்ளன. ஆனால் 50 விழுக்காடுக்கு மேல் உரையாடல் உள்ள மாதிரிகள் எடுத்துக்கொள்ளப்படமாட்டாது. மாதிரிகள் அவற்றின் பிரதிநித்துவப் பண்பிற்கேயன்றி வேறு எந்த அகவயமாகத் தீர்மானிக்கப்பட்ட சிறப்புக்காகவும் தெரிந்தெடுதக்கப்படவில்லை. தேர்வு செயல் முறைகள் மூன்று நிலைகளைக் கொண்டது:

1. தொடக்க நிலை அகவய வகைப்பாடு

2. ஒவ்வொரு வகைப்பாட்டிலும் எவ்வளவு மாதிரிகள் பயன்படுத்தப்படும் என்பதன் தீர்மானம்

3. ஒவ்வொரு வகைப்பாட்டிற்குள்ளும் மெய்யான மாதிரிகளின் வரைமுறையற்ற தேர்வு

விரிதரவுக்கு இரண்டு முக்கிய பகுதிகள் இருக்கின்றன:

1. தகவல் உரைநடையின் உரைகள்

2. கற்பனை உரைநடையின் உரைகள்

தகவல் உரைகள் பின்வருவனவற்றை உள்ளடக்கும்:

(1) பத்திரிக்கை: அறிக்கை (அரசியல், விளையாட்டு, சமுதாயம், வட்டாரச் செய்திகள், பொருளாதாரம், பண்பாடு)

(2) பத்திரிக்கை: இதழாசிரியர் கருத்து (நிறுவனம்சார்ந்தவை, பதிப்பாசிரியருக்கு தனிநபர் கடிதங்கள்)

(3) பத்திரிக்கை: திறனாய்வு (சினிமா, புத்தகம், இசை, நடனம், போன்றவை)

(4) மதம் (புத்தகங்கள், கால இதழ்கள், சமயக் கட்டுரைகள்)

(5) திறமைகளும் பொழுதுபோக்குகளும் (புத்தகங்கள், கால இதழ்கள்)

(6) புகழ் வாய்ந்த மரபுச் செய்திகள் (புத்தகங்கள்,கால இதழ்கள்)

(7) கடிதங்கள், வாழ்க்கை வரலாறு, நினைவுகள் (புத்தகங்கள், கால இதழ்கள்)

(8) பலதரப்பட்டவை (அரசாங்க ஆவணங்கள், அடித்தள அறிக்கைகள், தொழில் நிறுவன அறிக்கைகள், கல்லூரி பெயர் பட்டியல், தொழில் நிறுவனம், வீடு)

(9) கற்றல் மற்றும் அறிவியல் எழுத்துரைகள் (இயற்கை அறிவியல், மருத்துவம், கணித்தொழில் நுட்பம், பொறியியல்)

(10) சமூக மற்றும் நடத்தை அறிவியல்கள் (அரசியல் அறிவியல், சட்டம், கல்வி)

(11) மனிதவியலிலிருந்து உரைகள்.

கற்பனை உரைநடையின் உரைகள் பின்வருவனவற்றை உள்ளடக்கும்:

(1) பொதுக் கதைகள் (புதினம், சிறுகதைகள்)

(2) அறிவியல் கதைகள் (புதினங்கள், சிறுகதைகள்)

(3) துணிக்கரச் செயல் பற்றிய மற்றும் மேற்கத்திய கதைகள் (புதினங்கள், சிறுகதைகள்)

(4) காதல் கதைகள் (புதினங்கள், சிறுகதைகள்)

(5) நகைச்சுவை (புதினங்கள், சிறுகதைகள்)

2.3.2. லோப் தரவுத்தொகுதி

லங்காஸ்டர் ஓசலோ பெர்கன் தரவுத்தொகுதி (Lancaster Oslo/Bergen (LOB/லோப்) என்பது பிரிட்டிஷ் ஆங்கிலத்தின் ஒரு மில்லியன் சொல் சேகரிப்பாகும். இது லங்காஸ்டர் பல்கலைக் கழகத்தைச் சார்ந்த ஜியாபிரே லீச் (Geofferey Leech) மற்றும் ஓசலோ பல்கலைக் கழகத்தைச் சார்ந்த ஸ்டிக் ஜோகன்சன் (Stig Johansson) என்போர் பெர்கனில் உள்ள மனிதவியலுக்கான நார்வேஜியன் கணிப்பு மையத்துடன் இணைந்து தொகுப்பட்டதாகும். லோப் யுகே–இல் (UK) 1961–ஆம் வருடத்திற்குள் வெளியிடப்பட்ட மூலப்பொருட்களைக் கொண்டது. இது பிரவுண் விரிதரவைப் போல் உரை மாதிரிகளைக் கொண்டது. லோப் தரவுத்தொகுதியின் திட்ட நெறிமுறை யுஎஸ்ஏ-இலும் யுகே-இலும் பயன்படுத்தப்படும் ஆங்கிலத்தின் இரு வகைகளுக்கிடையே உள்ள எதிர்கால ஒப்பீட்டிற்கு வேண்டி பிரவுண் விரிதரவுடன் ஒற்றுமை கொண்டுள்ளது.

2.3.3. ஆஸ்டிரேலிய ஆங்கிலத் தரவுத்தொகுதி

ஆஸ்டிரேலிய ஆங்கில தரவுத்தொகுதி (Australian Corpus of English (ACE)) என்பது ஆஸ்டிரேலியாவிலுள்ள மாக்குயர் பல்கலைக் கழகத்தில் (Macquaire University) மொழியியல் துறையில் 1986–1990 என்ற கால கட்டத்தில் தொகுக்கப்பட்டது. ஆஸ்டிரேலிய ஆங்கில தரவுத்தொகுதி 1986-க்குள் வெளியிடப்பட்ட மூலப்பொருட்களைக் (materials) கொண்டிருக்கின்றது. ஆஸ்டிரேலிய ஆங்கில தரவுத்தொகுதி பல்வேறுபட்ட மொழியியல் ஆய்வுகளில் உதவுவதற்காகத் திட்டமிடப்பட்ட ஆஸ்டிரேலியாவில் முதலில் தொகுக்கப்பட்ட பலபடித்தான தரவுத்தொகுதி ஆகும் (heterogeneous corpus). இது தற்கால ஆஸ்டிரேலிய ஆங்கிலத்தின் திறனான மாதிரியாகும். இது ஆஸ்டிரேலியாவில் கூடுதல் சிறப்புப்பண்புள்ள ஒருபடித்தான தரவுத்தொகுதிகளுடன் (homogeneous corpora) ஒப்பீட்டுப் பார்ப்பதற்கு வேண்டி நோக்கீட்டு தரவுத்தொகுதியாகச் செய்யப்படுகின்றது. இது பிரவுண் தரவுத்தொகுதியிலும் லோப் தரவுத்தொகுதியிலும் பிரதிநிதித்துவம் தரப்பட்டுள்ள இனங்களின் சரிசமத்திற்குப் பொருத்துவதை நோக்கமாகக் கொண்டது. எனவே இது ஏறக்குறைய ஒவ்வொரு வகைப்பாட்டிலும் 2000 சொல் மாதிரிகளின் நிகரான குழுமத்தை உருவாக்குகிறது. சாத்தியமான இடங்களில் துணை இனங்களையும் பொருண்மைப் பரப்புகளையும் (subject areas) பொருத்துவது தேவை என்பதால் ஒவ்வொரு தரவுத்தொகுதி வகைப்பாட்டிற்குள்ளும் மாதிரிச் செயல் முறைகள் பெரும்பாலும் திட்டமிடப்பட்டுள்ளது.

2.3.4. வெல்லிங்டன் நியூசிலாந்து ஆங்கில எழுத்துத் தரவுத்தொகுதி

வெல்லிங்டன் நியூசிலாந்து ஆங்கில எழுத்துத் தரவுத்தொகுதி (Wellington Corpus of Written Newzealand English (WCWNEZE)) என்பது 1982 கால கட்டத்தில் வெல்லிங்டன் விக்டோரியா பல்கலைக் கழகத்தில் (Wellington Victoria University) மொழியியல் துறையில் உருவாக்கப்பட்டது. நியூசிலாந்து ஆங்கிலத் தரவுத்தொகுதியின் நோக்கம் பிரவுன் தரவுத்தொகுதி, லோப் தரவுத்தொகுதி, ஆஸ்டிரேலிய ஆங்கில தரவுத்தொகுதி என்பதுடன் நேரடியாக ஒப்பீடு செய்வதை அனுமதிக்கும் எழுத்து நீயூசிலாந்து ஆங்கிலத்தின் கணிணியாக்கம் செய்யப்பட்ட மாதிரியைத் தருகிறது. முக்கிய உரை வகைப்பாடுகள் எவ்வளவு இயலுமோ அவ்வளவு லோப் தரவுத்தொகுதியுடன் பொருந்தும் படி வரிசைப்படுத்தப்பட்டுள்ளன. கற்பனை உரை நடையில் குறிப்பாகக் கதைகளில் தேவை கருதிச் சில மாற்றங்கள் செய்யப்பட்டுள்ளன.

2.3.5. கோலாப்பூர் இந்திய ஆங்கிலத் தரவுத்தொகுதி

கோலாப்பூர் இந்திய ஆங்கிலத் தரவுத்தொகுதி (Kolhapur Corpus of Indian English (KCIE) எஸ்.வி சாஸ்த்திரியும் அவரைச் சார்தவர்களும் இணைந்து கோலாப்பூரில் உள்ள சிவாஜி பல்கலைக் கழகத்தில் 1988-ஆம் ஆண்டில் வெளியிடப்பட்ட விசயங்களிலிருந்து எடுக்கப்பட்ட இந்திய ஆங்கிலத்தின் கிட்டத்தட்ட ஒரு மில்லியன் சொற்களைக் கொண்டது. இத்தரவுத்தொகுதி அமேரிக்க, பிரிட்டிஷ் மற்றும் இந்திய ஆங்கிலத்தின் ஒப்பீட்டு ஆய்வுகளுக்கு வேண்டிய மூலப்பொருட்களாகப் பயன்படுகின்றது. இது இந்திய ஆங்கிலத்தின் விரிந்த வர்ணனைக்குக் கொண்டுச் செல்லும் என்று எதிர்பார்க்கப்படுகிறது. லோப் மற்றும் பிரவுன் தரவுத்தொகுதிகளை உருவாக்க பயன்படுத்தப்பட்ட திட்டத்தைப் பின்பற்றி கோலப்பூர் தரவுத்தொகுதிக்கு 15 பொருண்மை வகைப்பாடுகளிலிருந்து (subject areas) உரைகள் தேர்ந்தெடுக்கப்பட்டுள்ளன.

கோலாப்பூர் இந்திய ஆங்கிலத் தரவுத்தொகுதி ஒரு சுதந்திரமான இந்திய மொழியை அதாவது இந்திய ஆங்கிலத்தைச் (Indian English) சிறப்பாகத் திட்டமிடுகிறது. இந்திய ஆங்கிலம் இந்தியச் சொற்றொகை (Indian vocabulary) மற்றும் கலைச்சொற்களால் (terminology) நிறைந்திருக்கிறது; இது தனித்துவமான தொடரியல் முறை மற்றும் பொருண்மையியல்சார் சுமைகளுடன் பிரிட்டிஷ் ஆங்கிலத்திலிருந்து வேறுபட்டது மற்றும் பிரிட்டிஷ் ஆங்கிலத்தின் மிகப்பெரிய நிழலிலிருந்து விடுபட்டது.

தரவை கைமுறையாக கணினியில் எளிய ISCII பயன்முறையில் உள்ளிடுவதால் பயனர்களால் எளிதாக மீட்டெடுக்க முடியும். தரவுத்தொதி உருவாக்கிகள் (corpus generators) மூல உரைகளைச் சேகரிக்க வெவ்வேறு அணுகுமுறைகளைப் பின்பற்றின.

நூல்களைப் பொறுத்தவரை, 1978ஆம் ஆண்டில் இந்தியாவின் பல்வேறு மத்திய மற்றும் தேசிய நூலகங்களிலிருந்து ஆங்கிலத்தில் அச்சிடப்பட்ட நூல்களின் நோக்கீட்டு நூல்களின் தொகுப்பைத் தொகுத்தனர். அனைத்து வகைகளையும் உள்ளடக்கிய நூல்களிலிருந்து கிட்டத்தட்ட 140 நூல்கள் 1200 க்கும் மேற்பட்ட தலைப்புகளிலிருந்து மாதிரிகள் எடுக்கப்பட்டன, இது கோலாப்பூர் இந்திய ஆங்கிலத் தரவுத்தொகுதியின் மொத்த சொல் வலிமையின் 8% ஆகும்.

1978இல் அரசாங்க வெளியீடுகளின் ஆவணங்களைப் பொறுத்தவரையில் பட்டியல் மற்றும் மாதிரியின் அதே நடைமுறை சிறிய எளிமைப்படுத்தலுடன் பின்பற்றப்பட்டது; ஏனெனில் இந்தியாவில் இரண்டு வகையான அரசு ஆவணங்கள் உள்ளன: யூனியன் மற்றும் ஸ்டேட்.

பத்திரிகைப் பொருட்களிலிருந்து வரும் உரைகளுக்கு, செய்தித்தாள்களின் மாதிரி மற்றும் செய்தித்தாள்களின் குறிப்பிட்ட இதழ்கள் முதலில் மேற்கொள்ளப்பட்டன. மொத்தத்தில், 53 ஆங்கிலச் செய்தித்தாள்கள் கருதப்பட்டன; அவற்றில் 6 தேசிய செய்தித்தாள்கள், மீதமுள்ளவை பிராந்திய செய்தித்தாள்கள் ஆகும். ஒவ்வொரு செய்தித்தாளிலிருந்தும், 16 தினசரி இதழ்கள் மற்றும் 4 ஞாயிறு பதிப்புகள் மாதிரி செய்யப்பட்டுள்ளன.

கால இதழ்களுக்கு (periodicals) மத்திய நூலகம், மும்பை மற்றும் தேசிய நூலகம், கொல்கத்தா ஆகியவற்றின் வளங்கள் மாதிரிகளை எடுக்கப் பயன்படுத்தப்பட்டன. கற்றவர் பத்திரிகைகளுக்கு (learned journals) வேறுபட்ட நடைமுறை பின்பற்றப்பட்டது. அறிவியல் மற்றும் தொழில்நுட்ப விசயங்களுக்கு, டாடா அடிப்படை ஆராய்ச்சி நிறுவனம் (Tata Institute of Fundamental Research (TIFR/டிஐஎஃப்ஆர்)) மற்றும் இந்திய தொழில்நுட்ப நிறுவனம்-மும்பையைச் (Indian Institute of Technology-Mumbai (IIT-Mumbai/ஐஐடி-மும்பை) சார்ந்த நூலகங்களிலிருந்து பத்திரிகைகள் எடுக்கப்பட்டன. சமூக அறிவியல் நூல்களுக்கு டிஐஎஸ்எஸ்-மும்பை (Tata Institute od Social Sciences-Mumbai (TISS)-Mumbai)), பம்பாய் பல்கலைக்கழகம், பூனா பல்கலைக்கழகம், சிவாஜி பல்கலைக்கழகம் மற்றும் எம்.எஸ். பல்கலைக்கழகம், பரோடா என்பன பயன்படுத்தப்பட்டன.

இந்தத் தரவுத்தொகுதி இப்போது நோர்வேயின் பெர்கன், மனிதநேயங்களுக்கான நோர்வே கணிப்பு மையத்தில் (Norwegian Computing Centre for the Humanities) நவீன

ஆங்கிலத்தின் சர்வதேச கணினி காப்பகம் (International Computer Archive of Modern English (ICAME)) மூலம் ஆராய்ச்சியாளர்களுக்கு கிடைக்கிறது. இந்த தரவுத்தொகுதி சர்வதேச ஆங்கில தரவுத்தொகுதி-இல் (International Corpus of English (ICE)) இந்திய ஆங்கிலத்தின் பிரதிநிதி மாதிரியாகச் சேர்க்கப்பட வேண்டும்.

2.3.6. ஃப்லோப் தரவுத்தொகுதி (FLOB Corpus)

பிரிட்டிஷ் ஆங்கிலத்திற்கான ஃப்ரீபர்க் லாப் தரவுத்தொகுதி (Freiburg LOB Corpus for British English (FLOB))பிரவுன் மற்றும் லோப் தரவுத்தொகுதிகளுடன் பொருத்தும் ஒரு குழும தரவுத்தொகுதிகளைத் தொகுக்கும் முயற்சியின் விளைவாகும். இது 1990-களின் தொடக்க நிலை மொழியைப் பிரதிநிதித்துவம் செய்யும் என்ற நிலையில் வேறுபடும். இவ்வாய்வுத் திட்டம் 1990 ஏப்பிரலில் ஜெர்மனியில் கிருஸ்டியன் மேயிரின் (Christian Mair) கீழ் தொடங்கப்பட்டது. இத்திட்டத்தின் நோக்கம் மொழியில் தொடர்ந்து நடைபெறும் மாற்றத்தை ஆயவேண்டி அனுபவவாத அடிப்படையை மொழியியலாலருக்குத் தருவதாகும். இது மொழியியலாளர்களுக்குப் பின்வரும் செயல்களில் உதவுகிறது:

(1) இன்றைய ஆங்கிலத்தில் மொழிசார் மாற்றத்தின் மீதான இன்றைய கருதுகோள்களைப் பரிசோதித்தல்

(2) சொல் நிகழ்வெண்களின் முறையான ஒப்பீட்டுகளின் வழி இலக்கியத்தில் இதுவரை கண்டுபிடிக்கப்படாத மாற்றங்களைக் கண்டுப்பிடித்தல்

(3), ஒருகால இடவழக்குகளுக்கும் (British மற்றும் American) நடையியல் வேறுபாட்டுக்கும் இடையிலான சார்பு ஒருபக்கமும் மெய்யான இருகால வளர்ச்சிகள் ஒருபக்கமும் நடந்துகொண்டிருக்கும் மாற்றத்தின் ஆய்வில் ஒரு முக்கிய வழிமுறை சிக்கல்களில் ஒன்றை முறையாகக் கையாளுதல்.

இப்புதிய பிரிட்டிஷ் மற்றும் அமேரிக்க தரவுத்தொகுதிகளின் (British and American corpora) கூடுதல் அனுகூலம் என்னவென்றால் அவை இந்திய, ஆஸ்திரேலிய மற்றும் நியூசிலாந்து தரவுத்தொகுதிகளுடன் (பிந்தைய 1980-களில் உள்ள மொழிப்பயன்பாட்டைப் பிரதிநித்துவம் செய்யும் மாதிரிகள்) ஒப்பிட மூல லோப் மற்றும் பிரவுண் தரவுத்தொகுதிகளைவிடக் கூடுதல் பொருத்தமான தரவு மையங்களைத் தருகின்றன.

லோப்-இன் சிக்கலான குறிய முறையைப் (coding system) பயன்படுத்துவதற்குப் பதிலாக ஃப்லோப்-இன் (FLOB) நிலையான பொதுமைப்படுத்தப்பட்ட மார்க்அப் மொழி (Standard

Generalized Markup Language (SGML/எஸ்.ஜி.எம்.எல்)) அடிப்படையிலான மார்க்-அப் குறியங்களின் எளிமைப்படுத்தப்பட்ட பதிப்பைப் பயன்படுத்தியது; அது சர்வதேச ஆங்கிலத் தரவுத்தொகுதி (International Corpus of English (ICE)) துணைத் தரவுத்தொகுதியைக் (sub-corpora) குறியம் செய் வது வரை நீட்டப்பட்டுள்ளது.

ஃப்ளாப் (FLOB) உரை முடிந்தவரை 'படிக்கக்கூடியதாக' இருப்பதை உறுதி செய்வதற்காக மார்க்-அப் குறியங்களின் பயன்பாடு குறைந்தபட்சமாக வைக்கப்படுகிறது. குறிப்பாக, இரட்டைக் குறியங்களின் (double codes) பயன்பாடு முழுமையாகத் தவிர்க்கப்பட்டுள்ளது. அசல் உரையில் ஆங்கிலம் அல்லாத சொல் சாய்வில் அமைக்கப்பட்டால் அது ஆங்கிலம் அல்லாதாக (<foreign_>word<foreign/>) மட்டுமே குறியம்செய்யப்படும் மற்றும் (<tf_><foreign_> word<foreign/><tf/>) ஆக அல்ல.

2.3.7. பிரிட்டிஷ் தேசியத் தரவுத்தொகுதி

பிரிட்டிஷ் தேசியத் தரவுத்தொகுதி (British National Corpus (BNC)) என்பது வளங்களின் விரிந்த பரப்பெல்லையிலிருந்து பெறப்பட்டு எழுதப்பட்ட மற்றும் பேசப்பட்ட மாதிரிகளின் சேகரிப்பைக் கொண்ட தற்கால ஆங்கிலத்தின் மிகப்பெரிய (100 மில்லியன் சொற்களுக்கு மேல்) தரவுத்தொகுதி ஆகும். இது பேசப்படும் மற்றும் எழுதப்படும் தற்கால பிரிட்டஷ் (current British) ஆங்கிலத்தின் பரந்த பகுதியைப் பிரதிநிதித்துவம் செய்வதற்கு வேண்டி திட்டமிடப்பட்டுள்ளது. தரவுத்தொகுதி ஆக்ஸ்ஃபோர்ட் பல்கலைக்கழ அச்சகம் (Oxford University Press (OUP), அடிசன்-வெஸ்லி லாங்மேன் (Addison-Wesley Longman), லாரூஸ் கிங்ஃபிஷர் சேம்பர்ஸ் (Larousse Kingfisher Chambers), ஆக்ஸ்போர்டு பல்கலைக்கழக கணிப்பு சேவைகள் (Oxford University Computing Services), லான்காஸ்டர் பல்கலைக்கழகத்தின் கணினி ஆராய்ச்சி மையம் (Lancaster University's Centre for Computer Research) மற்றும் பிரிட்டிஷ் நூலக ஆராய்ச்சி மற்றும் கண்டுபிடிப்பு மையம் (British Library's Research and Innovation Centre) ஆகியவற்றால் வழிநடத்தப்பட்ட ஒரு தொழில்துறை / கல்வி கூட்டமைப்பால் உருவாக்கப்பட்டது.

தரவுத்தொகுதியை உருவாக்கும் வேலை 1991-இல் தொடங்கி 1994-இல் முடிக்கத் திட்டமிடப்பட்டது. ஐரோப்பிய ஆய்வாளர்களுக்கு பிளன்சியின் முதல் பொதுவான வெளியீடு பெப்ரவரி 1995-இல் அறிவிக்கப்பட்டது. பிரிட்டிஷ் தேசியத் தரவுத்தொகுதி தற்கால பிரிட்டிஷ் ஆங்கிலத்தின் விரிந்த பரப்பெல்லையைப் உருப்படுத்தம் செய்ய வேண்டி வடிவமைக்கப்பட்டுள்ளது. எழுத்துப் பகுதி (90%) வட்டார மற்றும் தேசீய செய்தித்தாள்கள்,

எல்லா வயதினர்களுக்கும் ஆர்முடைவர்களுக்கும் வேண்டி வெளிவரும் சிறப்பான காலவிதழ்கள் மற்றும் பத்திரிகைகள், கல்விசார் புத்தகங்கள் மற்றும் புகழ்வாய்ந்த கதைப் புத்தகங்கள், வெளியிடப்பட்ட மற்றும் வெளியிடப்படாத கடிதங்கள் மற்றும் அறிக்கைகள், பள்ளி மற்றும் பல்கலைக்கழகக் கட்டுரைகள் மற்றும் பிற பல உரை வகைகளை உள்ளடக்கும். பேச்சுப்பகுதி (10%) பல வயதினர்கள், வட்டாரங்கள், சமூக வகுப்புகள் இவற்றிலிந்து பெறப்பட்ட எழுதப்படாத முறைசாரா உரையாடல்கள் மற்றும் முறையான வணிக மற்றும் அரசாங்க கூட்டங்களிலிருந்து வானொலி மற்றும் தெலைபேசி வரையிலான பரப்பெல்லையைக் கொண்ட சூழல்களின் எல்லா வகைகளிலிருந்தும் சேகரிக்கப்பட்ட பேச்சு மாதிரிகள் என்பனவற்றை உள்ளடக்கும்.

இப்பொதுத்தன்மை பிரிட்டிஷ் தேசியத் தரவுத்தொகுதியை அகராதியியல், பொருண்மையியல், உருபனியல், மொழிப் பகுப்பாய்வு, செயற்கை அறிவுநுட்பம், பேச்சு அறிதல் மற்றும் தொகுப்பு (speech recognition and synthesis), இலக்கிய ஆய்வுகள் மற்றும் முக்கிய மொழியியலின் எல்லாப் பரப்புகளையும் உள்ளடக்கும் ஆய்வு நோக்கங்களின் விரிந்த வகைக்குப் பயனுள்ளதாகச் செய்கின்றது. பிரிட்டிஷ் தேசியத் தரவுத்தொகுதி பேச்சு உரையாடல் (spoken conversation) மற்றும் ஒருவர் உரையாடல் (monologue) இவற்றிலிருந்து ஒலிபெயர்ப்பு செய்யப்பட்ட 863 உரைகளை உள்ளடக்கிய 4124 உரைகளைக் கொண்ட 100106008 சொற்களை உள்ளடக்கியது. ஒவ்வொரு உரையும் மேலும் எழுத்துவடிவ வாக்கிய அலகுகளாகப் பிரிக்கப்பட்டுள்ளது. இதற்குள் ஒவ்வொரு சொல்லுக்கும் சொல் வகுப்பு (சொல்வகைப்பாடு) குறியம் தரப்பட்டுள்ளது. பிரித்தலும் சொல் வகைப்படுத்தலும் CLAWS சொல்வகைப்பாட்டு அடையாளப்படுத்தியால் தானியக்கமாகச் செய்யப்படுகின்றது. பிரிட்டிஷ் தேசியத் தரவுத்தொகுதியில் பயன்படுத்தப்படும் வகைப்பாட்டுத் திட்டம் 65 சொல் வகைப்பாடுகளை வேறுபடுத்துகின்றது.

CLAWS-இலிருந்து வெளியீடு மற்றும் உரைகளின் (text) பல்வேறு கட்டமைப்புப் பண்புகள் (எ.கா. தலைப்புகள், பத்திகள், பட்டியல்கள் போன்றவை) இரண்டையும் உருப்படுத்தம் செய்ய வேண்டி, *உரைக் குறியனாக்க முயற்சியின்* (Text Encoding Initiative (TEI)) வழிகாட்டுதல்களைப் பின்பற்றி (International Organization for Standardization (ISO ஐஎஸ்ஓ)) தரநிலை 8879-ஐப் (தரமான பொதுமையாக்கப்பட்ட மார்க்-அப் மொழி (Standard Generalized Markup Language (SGML/எஸ்ஜிஎம்எல்)) பயன்படுத்தி தரவுத்தொகுதி குறியனாக்கம் (encoded) செய்யப்பட்டுள்ளது. முழு வகைப்பாடு, சூழ்நிலைசார் மற்றும் நோக்கீட்டுநூல்கள்சார் தகவல்கள்

என்பன ஒவ்வொரு உரையுடனும் உரைக் குறியனாக்க முயற்சி-இணக்கமான தலைப்பு வடிவத்தில் சேர்க்கப்பட்டுள்ளன.

2.3.8. அமேரிக்க த் தேசியத் தரவுத்தொகுதி

அமேரிக்க தேசியத் தரவுத்தொகுதி (American National Corpus (ANC)) பிரிட்டிஷ் தேசியத் தரவுத்தொகுதியுடன் ஒப்பிடத்தக்க அமேரிக்க ஆங்கிலத்தை உட்படுத்திய தரவுத்தொகுதியின் உருவாக்கத்திற்கு உதவியது. இதன் நோக்கம் பிரிட்டிஷ் தேசியத் தரவுத்தொகுதி பொதுவினங்களுடன் ஒப்பிடத்தக்கக் குறைந்தது 100 மில்லியன் சொற்களைப் பெறுவதாகும். இது மொழித் தொழில்நுட்ப ஆய்வுக்கு உதவுவதுடன் எல்லா மட்டங்களிலும் மொழிப் பகுப்பாய்வுக்கும் கல்விக்கும் பயன்பட வேண்டி வளமையான தேசிய மூலத்தைத் தருகின்றது. அமேரிக்க தேசியத் தரவுத்தொகுதி அமேரிக்க ஆங்கில அகராதிகளின் வெளியீட்டாளர்கள் மற்றும் மொழி ஆய்வில் ஆர்வமுள்ள நிறுவனங்கள் இவற்றின் கூட்டிணைப்பின் (Consortium) பங்களிப்பு வழி உருவாக்கப்பட்டது. மொழித் தரவுக் கூட்டிணைப்பும் (Linguistics Data Consortium (LDC)) சாதன மற்றும் பொருளாதார ஆதரவு தந்தது. அமேரிக்கத் தேசியத் தரவுத்தொகுதி 10 மில்லியன் சொற்களின் முதல் வெளியீட்டில் உள்ளடக்கப்பட்ட உரைகள் முதலில் பெறப்பட்டவையாகும். எனவே தரவுத்தொகுதி சரிநிகரானதல்ல: பேச்சுப் பகுதி 3224388 சொற்களையும் எழுத்துப்பகுதி 8283828 சொற்களையும் கொண்டிருக்கின்றது.

பேச்சுத் தரவுகளில் ஆங்கிலம் பேசும் அமெரிக்கர்களிடையே 24 எழுதப்படாத தொலைபேசி உரையாடல்களுக்கான எழுத்துப்படிகள் (transcripts) மற்றும் ஆவணக் கோப்புகள் கொண்ட 'கால்-ஹோம்' ('Call-Home') கூறு உள்ளது. 'ஸ்விட்ச்போர்டில்' அமெரிக்க ஆங்கிலத்தின் ஒவ்வொரு முக்கிய கிளைமொழிகளிலிருந்தும் இரு பாலினத்தைச்சார்ந்த 500க்கும் மேற்பட்ட பேசுபவர்கள் பேசும் தன்னிச்சையான உரையாடல்களின் (spontaneous conversations) எழுத்துவடிவமாக்கங்கள் (transcriptions) உட்படும். 'சார்லோட் கதை மற்றும் உரையாடல் சேகரிப்பு' ('Charlotte Narrative and Conversation Collection') வட கரோலினாவில் வசிப்பவர்களின் விவரிப்புகள், உரையாடல்கள் மற்றும் நேர்காணல் பிரதிநித்துவங்களைக் கொண்டுள்ளது. ஒவ்வொரு எழுத்துப்படிக்கான (டிரான்ஸ்கிரிப்டுள்) தலைப்பில் பேசுபவரின் வயது மற்றும் பாலினம் பற்றிய தகவல்கள் சேர்க்கப்பட்டுள்ளன.

எழுதப்பட்ட தரவு நியூயார்க் டைம்ஸின் பனுவல்களை உள்ளடக்கியது; இது ஜூலை 2002இல் ஒற்றைப்படை எண்ணிக்கையிலான ஒவ்வொரு நாட்களுக்கும் நியூயார்க் டைம்ஸ்

நியூஸ்வைரிலிருந்து 4000க்கும் மேற்பட்ட கட்டுரைகளைக் கொண்டுள்ளது. பெர்லிட்ஸ் பயண வழிகாட்டி (The Berlitz Travel Guides) பயணம் மற்றும் தொடர்புடைய பிரச்சினைகள் குறித்து அமெரிக்கர்களால் எழுதப்பட்ட உரைகளைக் கொண்டுள்ளது. ஸ்லேட் இதழ் (Slate Magazine) என்பது ஆன்-லைன் வெளியீடாகும்; இது செய்தி மற்றும் அரசியல், கலை, வணிகம், விளையாட்டு, தொழில்நுட்பம், பயணம், உணவு போன்றவற்றை உள்ளடக்கிய தற்போதைய ஆர்வத்தின் தலைப்புகள் பற்றிய சிறு கட்டுரைகளைக் கொண்டுள்ளது.

அமேரிக்க தேசியத் தரவுத்தொகுதியின் ஆக்ஸ்ஃபோர்ட் பல்கலைக்கழக அச்சக (Oxford University Press (OUP) துணைத் தரவுத்தொகுதி அமெரிக்கர்களால் எழுதப்பட்ட 5 ஆக்ஸ்ஃபோர்ட் பல்கலைக்கழக அச்சக வெளியீடுகளிலிருந்து (ஜவுளித் தொழில், குழந்தை மேம்பாடு, அமெரிக்க அரசியலமைப்பு, பொது உயிரியல் மற்றும் கட்டிடக்கலை) பெறப்பட்ட கால் மில்லியன் புனைகதை அல்லாத சொற்களைக் கொண்டுள்ளது.

சொல்வகைப்பாடு அடையாளப்படுத்தலின் (POS tagging) கை சரிபார்ப்பு எதுவும் இல்லை. 'பொருண்மைக்களம்' (domain), 'துணைப் பொருண்மைக்களம்' (sub-domain), 'பொருள்' (subject), 'பார்வையாளர்கள்' ('audience') மற்றும் 'ஊடகம்' (medium) தொடர்பான முழுமையான தகவல்களைக் கொண்டிருந்தாலும் தலைப்புகள் மிகக் குறைவு.

2.3.9. ஆங்கில வங்கி

ஆங்கில மொழி வங்கி (Bank of English) சொற்கள், இலக்கணம் மற்றும் பயன்பாடு (usage) என்பனவற்றின் பகுப்பாய்விற்கு வேண்டி வடிவமைக்கப்பட்ட தற்கால ஆங்கில மொழியின் மாதிரிகளின் சேகரிப்பாகும். ஆங்கில மொழி வங்கி 1991-இல் COBUILD (Harper Collins Publishers) மற்றும் யுகேயில் உள்ள பிர்மிங்காம் பல்கலைக்கழகம் இவற்றால் 1991-இல் வெளியிடப்பட்டது. 2002 சனவரியில் 450 மில்லியன் சொற்களைக் கொண்ட புதிய தரவுத்தொகுதி வெளியிடப்பட்டது. இது புதிய சாதனங்களின் நிரந்ததர சேர்ப்பினால் தொடர்ந்து வளர்கின்றது. சேகரிப்பு எழுதப்பட்ட மற்றும் பேச்சின் வேறுபட்ட வகைகளின் விரிந்த பரப்பெல்லையைக் கொண்டது. இது நூற்றுக்கணக்கான வேறுபட்ட மூலங்களிலிருந்து ஆங்கில மொழியின் மாதிரிகளைக் கொண்டிருக்கின்றது. எழுதப்பட்ட உரைகள் செய்தித்தாள்கள், இதழ்கள் (magazines), கதைப் புத்தகங்கள் மற்றும் கதையல்லாப் புத்தகங்கள், சிற்றேடுகள் (brochures), துண்டுப்பத்திரிக்கைகள் (leaflets), அறிக்கைகள், கடிதங்கள் மற்றும் பிறவற்றிலிருந்து எடுக்கப்பட்டவைகளாகும். பேச்சுச் சொற்கள் அன்றாட சாதாரண

உரையாடல்கள், வானொலி ஒலிபரப்புகள், கூட்டங்கள், நேர்காணல்கள் மற்றும் விவாதங்கள் போன்றவற்றால் பிரதிநிதித்துவம் செய்யப்பட்டுள்ளது.

பொருளடக்கம் புதுப்பித்த நிலையில் உள்ளது; பெரும்பாலான உரைகள் 1990க்குப் பிறகு தோன்றியவை. உரைகள் மின்னணு வடிவத்தில் பெறப்பட்டு அவற்றை நிலையான/தரமான வடிவத்திற்குக் கொண்டு வர செயலாக்கம் செய்யப்பட்டுள்ளன. சில புத்தகங்கள் ஒளியுணர் (Optical Chracter Recognition (OCR) மென்பொருளைப் பயன்படுத்தி ஸ்கேன் செய்யப்பட்டுள்ளன. பத்திரிகைகள் மற்றும் எபிமெரா ஆகியவை விசைப்பலகை வழி ஏற்றப்பட்டுள்ளன. பேச்சின் கேட்பு பதிவுகள் சிறப்பு பயிற்சி பெற்ற ஊழியர்களால் நேரடியாக கணினியில் எழுத்துப்படிவமாக்கப்பட்டுள்ளன (transcripted).

தரவுத்தொகுதி பிபிசி உலக சேவை வானொலி ஒலிபரப்புகள் (BBC World Service radio broadcasts) மற்றும் அமெரிக்க தேசிய பொது வானொலி (American National Public Radio) இவற்றில் இருந்து மில்லியன் கணக்கான எழுத்துப்படிவமாக்கப்பட்ட பேச்சு வார்த்தைகளைக் கொண்டுள்ளன. பாங்க் ஆப் ஆங்கிலத்தில் பிரதிநித்துவம் செய்யப்பட்டுள்ள உரைகளின் கலவை மற்றும் பல்வேறுவகைகள் தொடர்ந்து மதிப்பாய்வு செய்யப்படுகின்றன; மேலும் உள்ளடக்கத்தின் சமநிலையை பராமரிக்க புதிய வளங்கள்/ஆதாரங்கள் அறிமுகப்படுத்தப்படுகின்றன; இதனால் அது இன்றைய ஆங்கிலத்தின் முக்கிய நீரோட்டத்தை பிரதிபலிக்கிறது.

சொல் சேர்க்கையின் வடிவங்களைத் தேட, சொல் நிகழ்வெண்களைச் சரிபார்க்க, சொற்களின் பயன்பாட்டு மாறுபாடுகளைக் காண மற்றும் அகராதிகள் மற்றும் நோக்கீட்டுப் படைப்புகளில் (reference works) பதிவுசெய்யப்பட்ட தகவல்கள் அதிகாரப்பூர்வமானவையா மற்றும் உறுதியான சான்றுகளால் ஆதரிக்கப்படுகிறதா என்பதைச் சரிபார்க்க, சொல்லியலார்கள் (Lexicologists), அகராதியலார்கள் (lexicographers) மற்றும் கலைச்சொல்லியலார்கள் (terminologists) இந்தத் தரவுத்தொகுதியை அணுகுகிறார்கள்.

பொது மொழியியலாளர்கள், மொழி ஆசிரியர்கள், மொழிபெயர்ப்பாளர்கள் மற்றும் மாணவர்கள் தங்கள் படிப்பு மற்றும் தொழில்முறை நடவடிக்கைகளுக்கு இதை ஒரு அடிப்படை ஆதாரமாகப்/வளமாகப் பயன்படுத்துகின்றனர். நிஜ வாழ்க்கையில் ஆங்கிலம் எவ்வாறு பயன்படுத்தப்படுகிறது என்பதைப் பற்றிய புரிதலை அதிகரிக்கவும், சொற்றொகுதி (vocabulary) மற்றும் இலக்கணத்தின் வகுப்பறை கற்பித்தலை மேம்படுத்தவும் அவர்கள் இதைப் பயன்படுத்துகிறார்கள்.

மொழி ஒரு முக்கிய அங்கமாக இருக்கும் சொல்-செயலி (word-processor), எழுத்துப்பிழை சரிபார்ப்பு (spelling-checker), இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்பு (machine translation system), கணினிமயமாக்கப்பட்ட தகவல் சேவை (computerised information service) போன்ற மொழித் தொழில்நுட்ப அமைப்புகளை மேம்படுத்துவதற்குப் பயனுள்ள ஆங்கில சொற்கள் மற்றும் இலக்கணத்தைப் பற்றிய தரவுகளையும் தகவல்களையும் ஆங்கில வங்கி வழங்குகிறது.

2.3.10. இந்திய மொழிகளின் மின்னணு மற்றும் தகவல் தொழில்நுட்ப அமைச்சக (MIT எம்.ஐ.டி.) தரவுத்தொகுதி

இந்திய மொழிகளுக்கு தரவுத்தொகுதி உருவாக்குவது 1991-இல் தொடங்கப்பட்டது. இது இந்திய அரசின் மின்னணுத் துறை எல்லா இந்திய மொழிகளுக்கும் இயந்திரத்தால் படிக்கவியலும் உரைகளின் தரவுத்தொகுதிகளை உருவாக்க வேண்டி இந்திய மொழிகளுக்குத் தொழில் நுட்ப வளர்ச்சி (Technological Development for Indain Languages) என்ற திட்டத்தைத் தொடங்கியது. இதே நேரத்தில் மொழி ஆய்வுக்கும் (சொல்வகைப்பாடு அடையாளப்படுத்தி, உரை குறியமாக்கி, புள்ளியல் எண்ணி, எழுத்துச் சர்பார்ப்பான், உருபனியல் பகுப்பாய்வி போன்றவற்றை உருவாக்க) மற்றும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு எந்திர உதவிசார் மொழிபெயர்ப்புக்கு (machine-aided translation) வேண்டி கருவிகளை வடிவமைத்தல் என்ற செயல்பாடுகளுக்கு மென்பொருள்களை உருவாக்கவும் ஊக்கம் தரப்பட்டது. டெல்கியில் உள்ள இந்தியத் தொழில் நுட்ப நிறுவனம் (Indian Institute of Technology) இந்திய ஆங்கிலம், இந்தி மற்றும் பஞ்சாபி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் (Central Institute of Indian Languages) தமிழ், தெலுங்கு, கன்னடா மற்றும் மலையாள மொழிகளுக்கு விதரவுகளை உருவாக்கியது. பூனேயில் உள்ள டெக்கான் கல்லூரி (Deccan College) மராத்தி மற்றும் குஜராத்தி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. புபனேஸ்வரிலுள்ள பயன்பாட்டு மொழி அறிவியல்களின் இந்திய நிறுவனம் (Indian Institute of Applied Language Sciences) ஒரியா, வங்காளம், அசாமீஸ் ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. வாரணாசியிலுள்ள சம்பூர்ணநந்தா சமஸ்கிருதப் பல்கலைக்கழகம் சமஸ்கிருத மொழிக்கு விரிதரவு உருவாக்கியது. அலிகார் முஸ்லீம் பல்கலைக்கழகம் (Aligarh Muslim University) உருது, சிந்தி மற்றும் காஷ்மீரி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. கான்பூரிலுள்ள இந்தியத் தொழில் நுட்ப நிறுவனம் மொழி (Indian Institute of technology) பகுப்பாய்விற்கும் இயந்திர மொழிபெயர்ப்புக்கும் கருவிகளையும்

மென்பொருள்களையும் வடிவமைக்கும் பொறுப்பை ஏற்றது. மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் எல்லா மொழிகளின் முழு தரவுத்தொகுதி தரவுமையத்தைச் சேகரித்துப் பாதுகாக்கும் பொறுப்பை ஏற்றது. 1995-இல் இத்திட்டம் முடிவுற்றபோது மின்னணு நிறுவனம் இது நீண்டகால ஆய்வையும் பெரிய முதலீட்டையும் வேண்டும் என்று உணர்ந்து திட்டத்தைத் தொடராமல் நிறுத்தியது. இருப்பினும் தற்போது இந்திய அரசின் எம்.ஐ.டி. (MIT) புதிய வெளிச்சத்திலும் பார்வையிலும் இச்செயல்பாட்டை மீண்டும் நடைமுறைப்படுத்தியுள்ளது (Vishwabharat at http:\\www.mit.gov.in).

2.3.11. பிற உரைத் தரவுத்தொகுதிகள்

தரவுத்தொகுதி டெல் எஸ்பாக்னோ (Corpus del Espagnol) என்ற தரவுத்தொகுதியில் 100 மில்லியன் ஸ்பானிஷ் உரைகள் உள்ளன.

சாண்டியாகோ டி காம்போஸ்டெலா பல்கலைக்கழகத்தில் (University of Santiago de Compostela) உள்ள ஸ்பானிஷ் தொடரியல் ஆராய்ச்சி குழு (Spanish Syntax Research Group) நவீன ஸ்பானிஷ் மொழியில் 1.5 மில்லியன் சொற்களைக் கொண்ட தரவுத்தொகுதியையும் 160000 பகுப்பாய்வு செய்யப்பட்ட உட்பிரிவுகளின் தொடரியல் தரவுத்தளத்தையும் உருவாக்கியது.

மன்ஹைமர் தரவுத்தொகுதி சேகரிப்பு (Mannheimer Corpus Collection) என்பது மொழியியல் ஆராய்ச்சிக்காக ஜெர்மன் ஆன்லைன் கார்ப்பரேஷனின் எப்போதும் வளர்ந்து வரும் தொகுப்பாகும். 1960களின் நடுப்பகுதியில் தொடங்கப்பட்டது; இது நவம்பர் 2002இல் 1850 மில்லியன் சொற்களை எட்டியது.

கோஸ்மாஸ் ஜெர்மன் தரவுத்தொகுதியிலும் (COSMAS German Corpus 1800) மில்லியனுக்கும் அதிகமான இயங்கும் சொற்கள் உள்ளன.

குரோஷிய தேசிய தரவுத்தொகுதியில் (Croatian National Corpus) ஒன்பது மில்லியன் டோக்கன்களுடன் சமகால குரோஷிய மொழியின் முப்பது மில்லியன் சொற்கள் உள்ளன.

பதின்மவயதினர் மொழியின் லண்டன் தரவுத்தொகுதி (Corpus of London Teenage Language (COLT)) ஏறக்குறைய அரை மில்லியன் சொற்களைக் கொண்டுள்ளது; இது எழுத்துக்கூட்டுமுறையில் எழுத்துப்படிவமாக்கம் செய்யப்பட்டு சொல்வகைப்பாட்டிற்காக அடையாளப்பட்டுள்ளது.

1975ஆம் ஆண்டில் ஸ்வீடனின் கெட்ட்போர்க் பல்கலைக்கழகத்தில் (Göteborg University, Sweden) உருவாக்கப்பட்ட ஸ்வீடிஷ் வங்கி (Bank of Swedish (Språkbanken/ஸ்ப்ராக்பாங்கன்)) புனைகதை, சட்டம்சார் உரை, பாராளுமன்ற நடவடிக்கைகளின் அறிக்கை மற்றும் தினசரி செய்தித்தாள்களில் இருந்து நாற்பது மில்லியன் இயங்கும் சொற்களைக் கொண்டுள்ளது.

போர்ச்சுகலின் காம்பினாஸ் பல்கலைக்கழகத்தின் (University of Campinas, Portugal) டைகோ பிரஹே பகுப்பாய்வு செய்யப்பட்ட தரவுத்தொகுதி (Tycho Brahe Parsed Corpus of Portuguese) (1982) வரலாற்றுப் போர்த்துகீசியத்தின் ஐந்து மில்லியன் சொற்களைக் கொண்டுள்ளது.

சுசான் தரவுத்தொகுதி (SUSANNE Corpus) (1992) எழுதப்பட்ட அமெரிக்க ஆங்கிலத்தின் 130000 சொற்களின் விளக்க அடையாளங்களைக் (annotations) கொண்டுள்ளது.

பேச்சு ஆங்கில மாதிரிகளின் அடிப்படையில் சுசானின் ஒரு எதிரிணையை உருவாக்க கிறிஸ்டின் தரவுத்தொகுதி (CHRISTINE corpus) தொடங்கப்பட்டது.

பழைய ஆங்கிலத்தின் புரூக்ளின் தரவுத்தொகுதி (Brooklyn Corpus of Old English) என்பது ஆங்கில உரைகளின் ஹெல்சின்கி தரவுத்தொகுதியின் (Helsinki Corpus of English Texts) பழைய ஆங்கிலப் பகுதியிலிருந்து (Old English Section) உரைகளின் தேர்வு ஆகும். இது இப்போது விளக்க அடையாளம் செய்யப்பட்டுள்ளது (annotated). இது தொகுப்பு (composition), படைப்பாளிகள்/ஆசிரியர்கள் (authors) மற்றும் இனங்களின் (genres) தேதிகளின் ஒரு பரப்பெல்லையைப் பிரதிநித்தவம் செய்யும் பழைய ஆங்கிலத்தின் ஒரு மில்லியனுக்கும் அதிகமான சொற்களைக் கொண்டுள்ளது.

மத்தியகால ஆங்கிலத்தின் பென்-ஹெல்சின்கி தரவுத்தொகுதி (Penn-Helsinki Corpus of Middle English) மத்தியகால ஆங்கிலத்தின் (Middle English) உரைநடை பனுவல்களைக் கொண்டுள்ளது. ஐந்து மில்லியன் சொற்களின் தரவுத்தளம் விளக்க அடையாளம் செய்யப்பட்டுள்ளது (annotated).

ஆங்கிலத்தின் உலகளாவிய ஒப்பீட்டு ஆய்வுகளுக்கான விஷங்களை சேகரிக்கும் நோக்கத்துடன் 1990ஆம் ஆண்டில் சர்வதேச ஆங்கிலத் தரவுத்தொகுதி (International Corpus of English (ICE)) தொடங்கியது. உலகெங்கிலும் உள்ள பதினைந்து ஆராய்ச்சி குழுக்கள் தங்களது சொந்த தேசிய அல்லது பிராந்திய வகை ஆங்கிலத்தின் மின்னணு விரிதரவுகளைத் தயாரிக்கின்றன. ஒவ்வொரு ஐ.சி.இ தரவுத்தொகுதியும் 1989 க்குப் பிறகு தயாரிக்கப்பட்ட ஒரு

மில்லியன் பேச்சு மற்றும் எழுதப்பட்ட ஆங்கிலச் சொற்களைக் கொண்டுள்ளது. பெரும்பாலான பங்கேற்பு நாடுகளுக்கு, சர்வதேச ஆங்கிலத் தரவுத்தொகுதித் திட்டம் தேசிய வகையின் முதல் முறையான விசாரணையைத் தூண்டுகிறது. இந்த திட்டத்தில் ஆஸ்திரேலியா, கனடா, கிழக்கு ஆபிரிக்கா (கென்யா மற்றும் தான்சானியா), கிரேட் பிரிட்டன், ஹாங்காங், இந்தியா, அயர்லாந்து, நியூசிலாந்து, பிலிப்பைன்ஸ், சிங்கப்பூர், இலங்கை மற்றும் அமெரிக்கா ஆகிய நாடுகளின் ஆராய்ச்சி குழுக்கள் உள்ளன. ஒவ்வொரு பிராந்தியக் குழுவும் நாட்டில் காணப்படும் பிராந்திய வகை ஆங்கிலத்தை பிரதிநித்துவம் செய்யும் பேசு மற்றும் எழுதப்பட்ட ஆங்கிலத்தின் ஒப்பிடக்கூடிய மாதிரிகளை சேகரிக்கின்றன. வேறுபட்ட வகைகளின் மாதிரிகளை உறுதிப்படுத்த ஒவ்வொரு குழுவும் வகையும் பல்வேறு வகைகளை பிரத்நுத்துவம் செய்யும் சொல் மாதிரிகளை சேகரிக்கின்றன: தன்னிச்சையான உரையாடல்கள், உரைகள், ஒளிபரப்பு விவாதங்கள், கற்றறிந்த உரைநடை, தனியார் கடிதங்கள், செய்தித்தாள் அறிக்கை, புனைகதை போன்றவை.

தரவுத்தொகுதிகள் தொடரியல், உருபனியல், சொற்றொகை, கருத்தாடல், ஒலியியல் மற்றும் ஒலியனியியல், ஆங்கில மொழி கற்பித்தல், மொழி திட்டமிடல் மற்றும் இயற்கை மொழி செயலாக்கம் குறித்த ஆராய்ச்சிக்கான விசயங்களை வழங்கும்.

ஆங்கிலம் கற்பவரின் உலகளாவியத் தரவுத்தொகுதி (இன்டர்நேஷனல் கார்பஸ் ஆஃப் லர்னர் ஆங்கிலம் (ஐ.சி.எல்.இ)) (International Corpus of Learner English (ICLE) 14 வெவ்வேறு தாய்மொழிப் பின்னணியிலிருந்து ஆங்கிலம் கற்கும் 2+ மில்லியன் சொற்களைக் கொண்டுள்ளது. தரவுத்தொகுதியில் உள்ள எழுத்துக்கள் இரண்டாம் மொழியாக இல்லாமல் வெளிநாட்டு மொழியாக ஆங்கிலம் கற்றவர்களால் பங்களிக்கப்பட்டுள்ளன; மேலும் இது 14 தனித்துவமான துணை தரவுத்தொகுதிகளால் ஆனது; ஒவ்வொன்றும் ஒரு மொழி வகைகளைக் கொண்டுள்ளது (ஆங்கிலம்-பிரஞ்சு, ஆங்கிலம்-ஜெர்மன், ஆங்கிலம்- ஸ்வீடிஷ் போன்றவை).

குழந்தை மொழித் தரவு பரிமாற்ற ஒழுங்குமுறை (Child Language Data Exchange System (CHILDES/சில்ட்ஸ்)) தரவுத்தளம், முதல் மற்றும் இரண்டாம் மொழிகளைக் கற்கிற குழந்தைகள் மற்றும் பெரியவர்களிடமிருந்து சேகரிக்கப்பட்ட எழுத்துப்படிவமாக்கப்பட்ட (டிரான்ஸ்கிரிப்ட்) தரவைக் கொண்டிருக்கின்றது. இது ஆங்கிலம் பேசும் பாடங்களிலிருந்து (English-speaking subjects) பெறப்பட்ட ஆங்கிலத் தரவுத்தொகுதி (English Corpus), உருபனியல்சார் அடையாளங்களைக்/குறிப்புகளைக் (tags) கொண்ட தரவு, (பொருண்மை மயக்கம் நீக்கப்பட்ட சொற்கள்), இருமொழி தரவுத்தொகுதி, மருத்துவ தரவுத்தொகுதி (மருத்துவ பாடங்களிலிருந்து),

தவளைக் கதை தரவுத்தொகுதி (இங்கு விவரிப்புகள் பட புத்தகத்திலிருந்து பெறப்படுகின்றன), கதை தரவுத்தொகுதி (இங்கு விளக்கங்கள் படங்கள் மற்றும் கதைகளிலிருந்து பெறப்படுகின்றன), ஜெர்மானிய மொழி தரவுத்தொகுதிகள் (ஜெர்மானிய மற்றும் நோர்டிக் மொழிகளைக் கற்கும் குழந்தைகளிடமிருந்து பெறப்பட்டது), ரொமான்ஸ் மொழி தரவுத்தொகுதிகள் (ரொமான்ஸ் மொழிகளைக் கற்கும் குழந்தைகளிடமிருந்து பெறப்பட்டது), மற்றும் பிற மொழி தரவுத்தொகுதிகள் (பிற மொழிகளைக் கற்கும் குழந்தைகளிடமிருந்து பெறப்பட்டது) இவற்றைக் கொண்டிருக்கும்.

இரு திசை இணை கொம்பாரா தரவுத்தொகுதி (bi-directional parallel COMPARA corpus) போர்த்துகீசியம்-ஆங்கிலம் மற்றும் ஆங்கிலம்-போர்த்துகீசிய மூல-உரைகள் மற்றும் மொழிபெயர்ப்புகளின் திறந்த-முடிவான தொகுப்பை அடிப்படையாகக் கொண்டது. 62 வெவ்வேறு போர்த்துகீசிய-ஆங்கில உரை இணைகளுக்கு அங்கோலா, பிரேசில், மொசாம்பிக், போர்ச்சுகல், தென்னாப்பிரிக்கா, யுனைடெட் கிங்டம் மற்றும் அமெரிக்காவிலிருந்து சொந்த எழுத்தாளர்கள் மற்றும் மொழிபெயர்ப்பாளர்களின் சமகால மற்றும் சமகால அல்லாத புனைகதைகளின் சாறுகள் இதில் அடங்கும். இந்தத் தரவுத்தொகுதி மொழிபெயர்ப்பைப் படிக்கவும், ஆங்கிலத்தையும் போர்த்துகீசியத்தையும் தானாக ஒப்பிட்டுப் பார்க்கவும் பயன்படுத்தப்படுகிறது.

இத்தாலியின் டஸ்கன் மையம் (Tuscan Centre, Italy) இத்தாலியப் பொருண்மை தரவுத்தொகுதியை (MEANING Italian Corpus (MIC/மிக்) காப்பகப்படுத்தியுள்ளது. இது பொருண்மைக்களம் (டொமைன்) அடிப்படையிலான (domain-based) சொற்பொருண்மை மயக்கநீக்கத்தை (word-sense disambiguation) ஆதரிப்பதற்காக உருவாக்கப்பட்ட சமகால இத்தாலிய மொழியின் 150 மில்லியன் சொற்களின் தரவுத்தொகுதி ஆகும். மிக் செய்தித்தாள் கட்டுரைகள், பத்திரிகை நிறுவன செய்திகள் போன்றவற்றால் ஆனது.

ஆங்கிலம்-நோர்வே இணைத் தரவுத்தொகுதி (English-Norwegian Parallel Corpus (ENPC /ஈ.என்.பி.சி) அசல் உரைகளையும் அவற்றின் மொழிபெயர்ப்புகளையும் கொண்டுள்ளது (ஆங்கிலம்-நோர்வே மற்றும் நோர்வே-ஆங்கிலம்). அசல் உரைகள் நோர்வே மற்றும் ஆங்கிலம் இரண்டிலும் உள்ளன. தரவுத்தொகுதி நான்கு முக்கிய பகுதிகளைக் கொண்டுள்ளது: ஆங்கில மூலங்கள் மற்றும் நோர்வே மொழிபெயர்ப்புகள், நோர்வே மூலங்கள் மற்றும் ஆங்கில மொழிபெயர்ப்புகள். 80 இணை உரைகளின் ஆரம்பத் திட்டம் திட்டத்தின் போது

நீட்டிக்கப்பட்டது மற்றும் தரவுத்தொகுதி தற்போது மொத்தம் 2.6 மில்லியன் சொற்களைக் கொண்ட 100 இணை உரைகளைக் கொண்டுள்ளது. இதில் 60% புனைகதைகள் (குழந்தைகளின் புத்தகங்கள், துப்பறியும் நாவல்கள் மற்றும் பொது இலக்கியங்கள்) மற்றும் 40% உண்மை உரைநடை (எ.கா. பிரபலமான அறிவியல், அரசு வெளியீடுகள், சட்ட நூல்கள், சுற்றுலா தகவல்கள்) உள்ளன. உரை குறியாக்க முன்முயற்சி (Text Encoding Initiative (TEI)) பரிந்துரைகளைத் தொடர்ந்து உரைகள் குறியாக்கம் செய்யப்பட்டன. இது மொழிபெயர்ப்பு ஆய்வுகள் (ஆங்கிலம்-நோர்வே அல்லது அதற்கு நேர்மாறாக) மற்றும் இரண்டு மொழிகளில் ஒப்பிடக்கூடிய அசல் உரைகள் அல்லது ஒரு உரை மற்றும் அதன் மொழிபெயர்ப்பு என்பதை அடிப்படையாகக் கொண்ட ஒப்பீட்டு ஆய்வுகள் ஆகிய இரண்டிற்கும் பயன்படுத்தப்படுகிறது. இரண்டு மொழிகளில் அசல் மற்றும் மொழிபெயர்க்கப்பட்ட உரைகளைப் ஆய்வும் இது உதவுகிறது. இது பயன்பாட்டு மற்றும் கோட்பாட்டு மொழியியல் ஆராய்ச்சிக்குத் தற்போதைய திட்டத்திற்கு அப்பால் கிடைக்கிற ஒரு பொதுவான ஆராய்ச்சி கருவியாகக் கருதப்படுகிறது.

ஜெர்மனியின் செம்னிட்ஸ் பல்கலைக்கழகத்தில் (Chemnitz University, Germany) சேகரிக்கப்பட்ட ஆரம்பகால நவீன ஆங்கில டிராக்ட்களின் லாம்பேட்டர் விரிதரவு (Lampeter Corpus of Early Modern English Tracts (LCEMET/எல்.சி.இ.எம்.இ.டி)) 1640 மற்றும் 1740 க்கு இடையில் வெளியிடப்பட்ட பல்வேறு விசயங்களின்/பொருள்களின் உரைகளை உள்ளடக்கியது. இது பல்வேறு சொல்சார், நடைசார் மற்றும் தொடரியல்சார் ஆராய்ச்சிகளுக்குப் பயன்படுத்தப்படுகிறது. லம்பேட்டர் தரவுத்தொகுதியில் உட்படுத்தபட்ட காலம் ஆங்கில வரலாற்றில் ஒரு முக்கியமான காலத்தையும், ஆங்கிலத்தை ஒரு பல்நோக்கு மொழியாக விரிவுபடுத்துவதையும் குறிக்கிறது. தரவுத்தொகுதிக்கு தேர்ந்தெடுக்கப்பட்ட உரைகள் ஆங்கிலத்தின் தரப்படுத்தல் செயல்முறைகள் மற்றும் உள்நாட்டுப் போர் (Civil War) வெடித்ததற்கும் தொழில்துறை புரட்சியின் (Industrial Revolution) தொடக்கத்திற்கும் இடையிலான வரலாற்று முன்னேற்றங்கள் இரண்டையும் பிரதிபலிக்கின்றன.

எழுதப்பட்ட உரைகளின் எஸ்தோனிய தரவுத்தொகுதி (Estonian Corpus of Written Texts) செய்தித்தாள்கள் (17.5%), ஆவணங்கள் (1.2%), கட்டுரைகள் மற்றும் சுயசரிதைகள்/வாழ்க்கைவரலாறுகள் (9.0%), பொழுதுபோக்குகள் (7.5%), புனைகதை (25.0%) , கலைக்களஞ்சியம் (2.0%), பிரச்சாரம் (6.0%), பிரபலமான அறிவியல் (15.0%), மதம் (0.8%), அறிவியல் (16.0%) போன்றவற்றின் குறிக்கப்படாத/அடையாளப்படுத்தப்படாத மற்றும்

குறியிடப்பட்ட/அடையாளப்படுத்தப்பட்ட உரை மாதிரிகளின் பெரிய தொகுப்பைக் கொண்டுள்ளது.

சிகாகோ பல்கலைக்கழகத்துடன் (University of Chicago) இணைந்து ஒட்டாவா பல்கலைக்கழகத்தில் (University of Ottawa) டெக்ஸ்டெஸ் டி ஃபிராங்காய்ஸ் அன்சியன் (Textes de Français Ancien (TFA/டி.எஃப்.ஏ)) தரவுத்தளம் உருவாக்கப்பட்டுள்ளது. இந்தத் தரவுத்தொகுதி 12 மற்றும் 13ஆம் நூற்றாண்டுகளின் உரைகளைக் கொண்டுள்ளன. பழைய பிரெஞ்சு மொழியின் சொல்லனாக்கப்பட்ட தரவுத்தளத்தை (lemmatised database) தயாரிப்பதற்கு இது டிஜிட்டல் மயமாக்கப்பட்டுள்ளது. பின்னர், மத்திய பிரெஞ்சு உரைகள் (Middle French texts) (14 மற்றும் 15 ஆம் நூற்றாண்டுகள்) இந்தத் தொகுப்பில் சேர்க்கப்பட்டுள்ளன.

நெக்ரா தரவுத்தொகுதி (NEGRA Corpus) ஜெர்மன் செய்தித்தாள்களிலிருந்து எடுக்கப்பட்ட சுமார் ஒரு மில்லியன் வாக்கியங்களைக் கொண்டுள்ளது. தரவுத்தொகுதி சொல்வகைப்பட்டிற்கு அடையாளப்படுத்தப்பட்டுள்ளது மற்றும் வாக்கிய அமைப்புகளுக்கும் அடையாளப்படுத்தப்பட்டுள்ளது. இது ஐரோப்பிய தரவுத்தொகுதி முன்முயற்சியின் பன்மொழி தரவுத்தொகுதிக்குப் பங்களிக்கப்பட்டுள்ளது.

ஒப்பீட்டு இந்தோ-ஐரோப்பிய தரவுத் தரவுத்தொகுதி (Comparative Indo-European Data Corpus) 95 இந்தோ-ஐரோப்பிய பேச்சு வகைகளுக்கான 200-சொற்களின் சொல்புள்ளியியல்சார் (லெக்ஸிகோஸ்டாடிஸ்டிகல்) பட்டியல்கள், இப்பட்டியல்களுக்கிடையிலான புலனறிவுசார் தீர்ப்புகள், சொல்புள்ளியியல்சார் (லெக்ஸிகோஸ்டாடிஸ்டிகல்) சதவீதங்கள், 200 அர்த்தங்களுக்கான தனிப்பட்ட மாற்று விகிதங்கள் போன்றவற்றை உள்ளடக்கியது.

அக்குவைண்ட் ஆங்கில செய்தி உரைத் தரவுத்தொகுதி (ACQUAINT English News Text corpus) மூன்று மூலங்களிலிருந்து பெறப்பட்ட ஆங்கில நியூஸ்வைர் உரைகளைக் கொண்டுள்ளது: சின்ஹுவா செய்தி சேவை (சீனா) (Xinhua News Service (China)), நியூயார்க் டைம்ஸ் செய்தி சேவை (New York Times News Service) மற்றும் அசோசியேட்டட் பிரஸ் வேர்ல்ட் ஸ்ட்ரீம் செய்தி சேவை (Associated Press World Stream News Service). இது சுமார் 375 மில்லியன் சொற்களைக் கொண்டுள்ளது. இது அமெரிக்காவின் தேசிய தரநிலைகள் மற்றும் தொழில்நுட்ப நிறுவனம் (National Institute of Standards and Technology (NIST/ என்ஐஎஸ்டி, USA)) நடத்திய உத்தியோகபூர்வ பெஞ்ச்மார்க் மதிப்பீடுகளில் பயன்படுத்த மொழித் தரவு கூட்டமைப்பால் (Language Data Consortium (LDC/எல்.டி.சி) தயாரிக்கப்பட்டது.

டச்சு சொல்லியல் நிறுவனம் (Institute for Dutch Lexicology) பல்வேறு கல்வி மற்றும் ஆராய்ச்சி நோக்கங்களுக்காக டச்சு மொழியில் பல மில்லியன் வார்த்தைகளை பெரிய தரவுத்தொகுதிகளாக உருவாக்கியுள்ளது.

கேம்பிரிட்ஜ் யுனிவர்சிட்டி பிரஸ்ஸின் (Cambridge University Press) கேன்டர்பரி டேல்ஸ் ப்ராஜெக்ட் (Canterbury Tales Project) சாஸரின் ஆங்கிலத்தின் (Chaucer's English) ஒரு தவுத்தொகுதியைக் கொண்டுள்ளது.

பென்சில்வேனியா பல்கலைக்கழகத்தில் (University of Pennsylvania) உள்ள பென் ட்ரீபேங்க் (Penn Treebank), முதன்மையாக ஆங்கிலத்தில் வோல் ஸ்ட்ரீட் ஜேர்னலின் (Wall Street Journal) கட்டுரைகளையும், செவ்வியல் (கிளாசிக்கல்), வரலாற்று மற்றும் மத உரைகளையும் கொண்டுள்ளது.

2.4. பேச்சுத் தரவுத்தொகுதி

பேச்சு தரவுத்தொகுதி (speech corpus (or spoken corpus) என்பது பேச்சு ஆடியோ கோப்புகள் (database of speech audio files) மற்றும் உரை எழுத்தாக்கங்களின் (text transcriptions) தரவுத்தளமாகும். பேச்சுத் தொழில்நுட்பத்தில் (speech technology), ஒலி மாதிரிகளை உருவாக்கப் பேச்சு தரவுத்தொகுதிகள் பயன்படுத்தப்படுகிறன (பின்னர் அவை பேச்சு அறிதல் அல்லது பேச்சாளர் அடையாள இயந்திரத்துடன் பயன்படுத்தப்படலாம்). மொழியியலில், ஒலிப்பு, உரையாடல் பகுப்பாய்வு, கிளைமொழியியல்ல் மற்றும் பிற துறைகளில் ஆராய்ச்சி செய்ய பேசுத் தரவுத்தொகுதிகள் பயன்படுத்தப்படுகிறன.

ஒரு பேச்சு தரவுத்தொகுதி அத்தகைய தரவுத்தளமாகும். கார்போரா (Corpora) என்பது கார்பஸ் (corpus) என்பதன் பன்மை (அதாவது இது போன்ற பல தரவுத்தளங்கள் கொண்டது).

பேச்சு தரவுத்தொகுதிகளில் இரண்டு வகைகள் உள்ளன:

1. படிக்கப்பட்ட பேச்சு - இதில் அடங்கும்:

- நூல் பகுதிகள் (Book excerpts)
- செய்தி ஒளிபரப்பு (Broadcast news)
- சொற்களின் பட்டியல்கள் (Lists of words)
- எண்களின் வரிசை (Sequences of numbers)

2. தன்னிச்சையான பேச்சு - இதில் அடங்கும்:

- உரையாடல்கள் (Dialogs) - இரண்டு அல்லது அதற்கு மேற்பட்ட நபர்களுக்கு இடையில் (கூட்டங்கள் அடங்கும்);
- விவரிப்புகள் (Narratives) - ஒரு கதையைச் சொல்லும் ஒருவர் (அத்தகைய ஒரு தரவுத்தொகுதி பக்கி தரவுத்தொகுதி (Buckeye Corpus));
- வரைபடப் பணிகள் (Map-tasks) - ஒரு நபர் ஒரு வரைபடத்தில் ஒரு வழியை இன்னொருவருக்கு விளக்குகிறார்;
- நியமனம்-பணிகள் (Appointment-tasks) - தனிப்பட்ட அட்டவணைகளின் அடிப்படையில் ஒரு பொதுவான சந்திப்பு நேரத்தைக் கண்டுபிடிக்க இரண்டு பேர் முயற்சி செய்கிறார்கள்.

ஒரு சிறப்பு வகையான வெளிநாட்டு உச்சரிப்புடன்கூடிய பேச்சுத் தரவுத்தொகுதிகள் (non-native speech databases) பேச்சைக் கொண்டிருக்கும் சொந்தமற்ற பேச்சு தரவுத்தளங்கள் ஆகும்.

2.4.1. லண்டன்-லண்ட் பேச்சு ஆங்கிலத் தரவுத்தொகுதி

லண்டன்-லண்ட் பேச்சு ஆங்கிலத் தரவுத்தொகுதி (London-Lund Corpus of Spoken English (LLC)) இரு திட்டங்களின் விளைவுகளைக் உட்படுத்தியது: (1) 1961-இல் லண்டன் பல்கலைக்கழகக் கல்லூரியில் ராண்டோல்ஃப் குயிர்கால் செயல்படுத்தப்பட்ட ஆங்கில வழக்கின் மதிப்பீட்டாய்வு (Survey of English Usage (SEU)) மற்றும் (2) 1975-இல் சுவீடனிலுள்ள லண்ட் பல்கலைக்கழகத்தில் ஜான் ஸ்வர்த்விக்கால் செயல்படுத்தபப்பட்ட பேச்சு ஆங்கிலத்தின் மதிப்பீட்டாய்வு (Survey of Spoken English (SSE)). இதன் நோக்கம் சேகரிக்கப்பட்டு எழுத்துப்பெயர்ப்பு செய்யப்பட்ட பேச்சுச் சாதனங்களை இயந்திரம் படிக்கவியலும் வடிவில் வைத்திருக்கச் செய்வது ஆகும். உரை மாதிரிகள் கிட்டத்தட்ட 5 மில்லியன் பேச்சுச் சொற்களை உட்படுத்தும். பேச்சுச் சாதனம் இலக்கண ஆய்வின்றி, குறைக்கப்பட்ட எழுத்துப்பெயர்ப்பின் நிலை வழி சென்றுள்ளது. 1980-இன் தொடக்கத்தில் கணினிப்படுத்தப்பட்ட பேச்சு ஆங்கிலத்தின் லண்டன்-லண்ட் தரவுத்தொகுதியின் முதல் நகல் ஆய்விற்கும் புலன்விசாரணைக்கும் வேண்டி உலகம் முழுவதும் உள்ள ஆர்வமுள்ள ஆய்வாளர்களுக்குத் தரப்பட்டது.

2.4.2. பேச்சு ஆங்கிலத்தின் மதிப்பீட்டாய்வு

பேச்சு ஆங்கிலத்தின் மதிப்பீட்டாய்வு (Survey of Spoken English (SSE)) 200 மாதிரி உரைகளைக் கொண்டது. ஒவ்வொரு மாதிரியும் 5000 சொற்களைக் கொண்டிருக்கும். மொத்தத்தில் 1 மில்லியன் சொற்களைக் கொண்டிருக்கும். பேச்சு மூல உரைகள் பின்வருவனவற்றை உள்ளடக்கும்: (1) ஒருவர் உரையாடல்கள் (உடனடியான, சொற்பொழிவு,

விளையாட்டு மற்றும் விளையாட்டல்லா விமர்சனம், தயாரிக்கப்பட்ட ஆனால் எழுதப்படாத சொற்பொழிவு), (2) இருவர் உரையாடல் (நேருக்கு நேரான உரையாடல், இரகசியமாகப் பதிவு செய்யப்பட்டவை மற்றும் வெளிப்டையாகப் பதிவு செய்யப்பட்டவை, தொலைபேசி உரையாடல்) மற்றும் (3) (பேச்சுகள், நாடகங்கள், செய்திகள், எழுதப்பட்ட சொற்பொழிவுகள், கதைகள்). தரவுத்தொகுதியானது பேச்சின் மீக்கூறு மற்றும் மொழிக்கு அப்பாற்பட்ட பண்புக்கூறுகளை அடையாளப்படுத்தும் ஒழுங்குமுறையால் எழுத்துப்பெயர்புச் செய்யப்பட்டுள்ளது. இச்செயற்பாட்டில் ஒவ்வொரு மீக்கூறும் மொழிக்கு அப்பாற்பட்ட பண்புக்கூறுகளும் அவ்வாறே கூர்ந்துகவனிக்கப்பட்டுக் குறிப்பிடப்பட்டுள்ளன. எல்லாப் பேச்சு உரைகளும் இலக்கண அடிப்படையில் ஆயப்பட்டு தரவுத்தொகுதிக்குள் முறையாக உருப்படுத்தம் செய்யப்பட்டுள்ளன.

2.4.3. இயந்திரத்தால் படிக்கவியலும் பேச்சு ஆங்கிலத் தரவுத்தொகுதி

பேச்சு ஆங்கிலத் தரவுத்தொகுதி (Spoken English Corpus (SEC)) தற்காலப் பேச்சு பிரிட்டிஷ் ஆங்கிலத்தின் கிட்டத்தட்ட 100000 சொற்களைக் கொண்ட இயந்திரத்தால் படிக்கவியலும் தரவுத்தொகுதி ஆகும். இது 1984-85-இல் லங்காஸ்டர் பல்கலைக்கழகத்தின் மனிதவியல் ஆய்வு நிதி (Humanities Research Fund) மற்றும் ஐபிம் யுகே லிமிடெட் (IBM UK Ltd) என்பவைகளால் ஆதரவு தரப்பட்டு உருவாக்கப்பட்டது. இதன் நோக்கம் மனிதப் பேச்சில் இசையோட்டம் தருவதற்கான திட்டங்களை ஆய்வதற்கும் பரிசோதனை செய்வதற்கும் வேண்டி பயன்படுத்தவியலும் இயல்பான பேச்சு பிரிட்டிஷ் ஆங்கிலத்தின் மாதிரிகளைச் சேகரிப்பதாகும். பேச்சுக் கூட்டிணைப்புக்கு வேண்டி ஒரு மாதிரியாகப் பொருத்தமுறும் பேச்சு ஆங்கிலத்தின் வகையின் பெருமளவு மாதிரியைச் சேகரிக்க முக்கியத்துவம் தரப்பட்டது. எனவே உயர்ந்த நடையைச் சார்ந்த பேச்சின் சிறிய மாதிரிகள் (அதாவது செய்யுளைப் படித்தல் அல்லது சமயச் சொற்பொழிவு) உட்படுத்தப்பட்டன. பெரும்பாலான உரைகள் பிஎன்சி தரவுத்தொகுதியிலிருந்து எடுக்கப்பட்டுள்ளன. இது விமர்சனங்களிலிருந்து உரைகள், செய்தி ஒலிபரப்பு, பொது கேட்போர் கூட்டத்திற்கான சொற்பொழிவுகள், வரையறுக்கப்பட்ட கேட்போர் கூட்டத்திற்கான சொற்பொழிவுகள், சமய ஒலிபரப்பு, பத்திரிக்கை நடை அறிவிப்பு, கதை படித்தல், செய்யுள் ஒப்பித்தல், இருவர் உரையாடல், பிரச்சாரம் மற்றும் பிற பேச்சுக்கள் என்பனவற்றை உட்படுத்தும். விரிதரவு எழுத்துவடிவ அடிப்படையிலும் மீக்கூறு அடிப்படையிலும் எழுத்துப்பெயர்பு செய்யப்பட்டுள்ளது. மீக்கூறின் மீது தொடரியலின் பாதிப்பின் ஆய்வை ஆனுமதிக்க வேண்டி CLAWS சொல் அடையாளப்படுத்தும் ஒழுங்குமுறையைப் பயன்படுத்தி

இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட வடிவம் உருவாக்கப்பட்டது. இந்த தரவுத்தொகுதி பேச்சுக் கூட்டிணைப்பில் அல்லது பேச்சு அறிதல் களங்களில் ஆய்வுசெய்யும் மக்களுக்குப் பயனுள்ளது என்று நிரூபிக்கப்பட்டுள்ளது. இது இயல்பான பேச்சு ஆங்கிலத்தின் ஒலியியலின் நெருங்கிய ஆய்வுக்கு மாணவர்களுக்குச் சந்தர்ப்பம் தந்து கற்பித்தல் நோக்கங்களுக்கு மதிப்புள்ள மூலவளமாக இருப்பது நிரூபிக்கப்பட்டுள்ளது.

2.4.4. வெலிங்டன் நியூசிலாந்து பேச்சு ஆய்கிலத் தரவுத்தொகுதி

வெலிங்டன் நியூசிலாந்து பேச்சு ஆய்கிலத் தரவுத்தொகுதி (Wellington Corpus of Spoken New Zealand English (WCSNZE)) 1994-இல் ஜானட் ஹோம்சின் வழிகாட்டின் கீழ் வில்லிங்டன் பல்கலைக்கழகத்தில் உருவாக்கப்பட்டது. ஒவ்வொரு பகுதியும் ஒரு மில்லியன் சொற்களைக் கொண்டிருக்கவேண்டும் என்று ஒத்துக்கொள்ளப்பட்டது. நியூசிலாந்து பேச்சு விரிதரவு 1875-க்கும் 1992-க்கும் இடையில் சேகரிக்கப்பட்ட முறையான பேச்சு/ஒருவர்பேச்சு (12%), பகுதி முறையான பேச்சு/பெறப்பட்ட ஒருவர் பேச்சு (13%), முறைசாராப் பேச்சு / உரையாடல் (75%) என்பனவற்றை உள்ளடக்கும். சுருக்கங்கள் பேச்சின் ஒவ்வொரு வகையையும் காணப்படும் சூழல்களின் ஒரு பரப்பெல்லையை உட்படுத்தும் 15 வகுப்புகளாகப் பிரிக்கப்பட்டுள்ளன. இது ஒலிபரப்புப் செய்திகள், ஒலிபரப்பு ஒருவர்பேச்சுகள், வானிலை அறிக்கை ஒலிபரப்பு, விளையாட்டு விமர்சனம், நீதிபதிகளின் சுருக்க உரை, விரிவுரைகள், ஆசிரியர் ஒருவர்பேச்சுகள், நேருக்கு நேரான உரையாடல்கள், தொலைபேசி உரையாடல்கள், வாய்மொழி வரலாற்று நேர்காணல்கள், சமுதாய வழக்கு நேர்காணல்கள், வானொலிப் பேச்சுகள், நேர்காணல்களின் ஒலிபரப்பு, பாராளுமன்ற வாதங்கள், ஒப்பந்த நடவடிக்கைகைகள் மற்றும் கூட்டங்கள் மற்றும் பிறவற்றை உள்ளடக்கும்.

2.4.5. எடின்பர்க் பல்கலைக்கழகப் பேச்சு கால ஆணவமும் ஆங்கில மொழியின் தரவுத்தொகுதியும்

எடின்பர்க் பல்கலைக்கழகப் பேச்சு கால ஆணவமும் ஆங்கில மொழியின் விரிதரவும் (Edinburgh University Speech Timing Archive and Corpus of English (EUSTACE)) என்பது ஒலியியல் ஆய்வாளர்களுக்கும் பேச்சு கூட்டிணைப்பிலும் புரிந்துகொள்ளலிலும் ஈடுபட்டுள்ள பேச்சுத் தொழிநுட்பவியலார்களுக்கும் பயனுள்ள நல்ல மூலவளமாகும். இத்தரவுத்தொகுதி எடின்பரோ பல்கலைக்கழகத்தின் கோட்பாடு மற்றும் பயன்பாட்டு மொழியியல் துறையில் (Department of Theoretical and Applied Linguistics of Edinburgh University) பதிவுசெய்யப்பட்ட

4608 பேச்சு வாக்கியங்களைக் கொண்டது. பிரிட்டிஷ் ஆங்கிலத்தின் 6 பேசுபவர்களால் பேசப்பட்ட இவ்வாக்கியங்கள் பேச்சில் காலத்தின் பல விளைவுகளைப் பரிசோதிப்பதற்கு வேண்டி வடிவமைக்கப்பட்டுள்ளது. இவை நீட்சிக்கும் ஒலிப் பொருளடக்கத்திற்கும் வேண்டி கட்டுப்படுத்தப்பட்டுள்ளன. ஒவ்வொரு வாக்கியத்திலும் முக்கியச் சொற்களின் துணைப் பகுதிகள் கண்டுபிடிக்கப்பட்டுள்ளன. அவை சில வாக்கியங்களின் மீக்கூறு உருப்படுத்தத்தைப் பற்றிய குறிப்புகளை உட்படுத்தும். எடுத்துக்காட்டு வாக்கியங்கள் இணைய தளத்தில் கேட்பதற்காக வைக்கப்பட்டுள்ளது.

2.4.6. கொரியன் பேச்சுத் தரவுத்தொகுதி

கொரியன் தொலைபேசி உரையாடல் தரவுத்தொகுதி (Korean Telephone Conversations speech Corpus (KTCSC)) தொடக்கத்தில் யுஎஸ்ஏ-இல் நண்பர்களை அழை ஆய்வுத்திட்டத்தின் (Korean Telephone Conversations speech Corpus (KTCSC)) பகுதியாகப் பதிவு செய்யப்பட்டது. மொழி அடையாளம் காணல் ஆய்வுத் திட்டத்தின் (Language Identification (LID) Project) ஆதரவில் LDC-ஆல் உரையாடல்கள் சேகரிக்கப்பட்டது. இத்தரவுமையம் கொரியன் தாய்மொழி பேசுபவர்களுக்கு இடையில் 100 தொலைபேசி உரையாடல்களைக் (மொத்தம் 25251 சொற்களைக் கொண்ட) கொண்டிருக்கின்றது. இதில் கிட்டத்தட்ட 44 மணி நேரச் செவிப்புலப் பேச்சு ஆவணங்கள் உள்ளது. ஒவ்வொரு சொல்லும் ஐந்து தகவல் களங்களைக் கொண்டது: (1) ஹங்குலில் இருந்து எழுத்துவடிவம், (2) யேல் ரோமனய்சேசனில் இருந்து எழுத்து வடிவம், (3) உச்சரிப்பு, (4) கொரியன் தொலைப்பேசி உரையாடல் எழுத்துப்பெயர்ப்புகளில் சொற்களின் நிகழ்வெண், (5) சொற்களின் உருபனியல் பகுப்பாய்வு.

2.4.7. பிற பேச்சுத் தரவுத்தொகுதிகள்

ஆஸ்திரிய BADIP (Austrian BADIP) (Banca Dati Dell'italiano Parlato) தரவுத்தளம் 500000 சொற்களின் ஆன்லைன் பதிப்பை கொண்டுள்ளது. பதிப்பு சொல்வகைப்பாட்டு அடையாளங்கள் (POS-tags) மற்றும் சொல்லனாக்கம் (lemmatization) மூலம் செறிவூட்டப்பட்டுள்ளது. கூடுதல் தரவு தொடர்ந்து சேர்க்கப்பட்டு வருகிறது. தரவுத்தளம் கிராஸ் பல்கலைக்கழகத்தின் (ஆஸ்திரியா) (Server of the University of Graz (Austria)) மொழி சேவையகத்தின் ஒரு பகுதியாகும்.

டிரைன்ஸ் பேச்சு உரையாடல் தரவுத்தொகுதி (TRAINS Spoken Dialogue Corpus) 20 வெவ்வேறு பணிகள் மற்றும் 34 வெவ்வேறு பேச்சாளர்களைப் பயன்படுத்தி சேகரிக்கப்பட்ட 98 பணி சார்ந்த பேச்சு உரையாடல்களைக் கொண்டுள்ளது. இந்த உரையாடல்கள், ஒரு இரயில்

பாதை சரக்கு ஒழுங்கமைப்பில் பொருட்களை உற்பத்தி செய்தல் மற்றும் ஏற்றுமதி செய்வது தொடர்பான சில பணிகளை அடைய திட்டங்களை உருவாக்கப் பயனர்களுக்கு உதவ உரையாடலில் திறமையான திட்டமிடல் உதவியாளரை (planning assistant) உருவாக்க சேகரிக்கப்படுகின்றன.

கான்டோனீஸ் பேச்சுத் தரவுத்தளம் (Cantonese Speech Database) சீனா மற்றும் ஹாங்காங்கில் மொபைல் தொலைபேசி வலையமைப்பில் பதிவுசெய்யப்பட்ட 2000 பேச்சாளர்களைக் கொண்டுள்ளது.

பிளெமிஷ் மற்றும் டச்சு பேச்சுத் தரவுத்தளம் (Flemish and Dutch Speech Database) 302 பேசுபவர்கள் (154 ஆண்கள், 148 பெண்கள்) ஒரு காரில் தன்னிச்சையாக உச்சரித்த மற்றும் வாசித்த 120 ஐடங்களின் பதிவுகளைக் கொண்டுளன. அது ஐந்து வெவ்வேறு சேனல்கள் மூலம் பதிவுகள் செய்யப்பட்டுள்ளன; அவற்றில் 4 இன்-கார் மைக்ரோஃபோன்கள் (in-car microphones) வழியாகவும் (1 க்ளோஸ்-டாக் மைக்ரோஃபோன், 3 தொலைதூர மைக்ரோஃபோன்கள்) மற்றும் 1 சேனல் ஓவர் தி ஜிஎஸ்எம் நெட்வொர்க்கிலும் (channel over the GSM network) செய்யப்பட்டன.

உரையாடல் பன்முகத்தன்மை தரவுத்தொகுதி (Dialogue Diversity Corpus (DDC/டி.டி.சி)) உரையாடல்களின் தொகுப்பிலிருந்து தயாரிக்கப்படுகிறது (13 ஆதாரங்கள்/மூலங்கள், ஆங்கிலத்தில் 12 மணி நேரத்திற்கும் மேலான உரையாடல்). இது வாய்வழி வரலாற்று நேர்காணல்கள், வாட்டர்கேட் நாடாக்கள், ஆங்கிலத்தின் பல்வேறு பிராந்திய வகைகள், அமெரிக்கன் ஆங்கிலம், யு.எஸ். உச்ச நீதிமன்ற நடவடிக்கைகள் மற்றும் பிற ஆதாரங்களின் பேச்சு ஆகியவற்றைக் கொண்டுள்ளது. உரையாடல்கள் ஊடாடும் சூழ்நிலைகளின் மிகவும் மாறுபட்ட தொகுப்பில் நிகழ்ந்தன. குறிப்பிட்ட உரையாடல் மாதிரிகளின் கவரேஜின் அகலத்தைப் படிப்பதற்கும், வெவ்வேறு சூழ்நிலைகளில் இருந்து உரையாடல்களை ஒப்பிடும் ஆய்வுகளுக்கும் இது பயன்படுத்தப்படுகிறது.

ஸ்மார்ட்காம் பன்மாதிரித் தரவுத்தொகுதி (SmartKom Multimodal Corpus) 1999க்கும் 2003-க்கும் இடையில் தயாரிக்கப்பட்டது. இது ஸ்மார்ட்காம் பப்ளிக் (SmartKom Public) என்ற தொழில்நுட்ப அமைப்பில் பொது இடங்களில் பதிவுசெய்யப்பட்ட 45 பேசுபவர்களின் உரைகளைக் கொண்டுள்ளது, இது கூடுதல் அறிவார்ந்த தகவல் தொடர்பு சாதனங்களைக் கொண்ட ஒரு பாரம்பரிய பொதுத் தொலைபேசி சாவடியுடன் ஒப்பிடத்தக்கது,.

வெஸ்ட் பாயிண்ட் அரபு பேசுத் தரவுத்தொகுதி (West Point Arabic Speech Corpus) வெஸ்ட் பாயிண்டில் உள்ள யுனைடெட் ஸ்டேட்ஸ் மிலிட்டரி அகாடமியில் (United States Military Academy) வெளிநாட்டு மொழிகள் துறையின் (Department of Foreign Languages) உறுப்பினர்கள் சேகரித்து செயலாக்கிய பேச்சுத் தரவு உள்ளது. இந்தத் தரவுத்தொகுதியின் நோக்கம் அரபு மொழியைக் கற்பிப்பதில் ஒரு உதவியாகப் பயன்படுத்தக்கூடிய தானியங்கி பேச்சு புரிதலுக்கான (automatic speech recognition) ஒலி மாதிரிகளைப் (acoustic models) பயிற்றுவிப்பதாகும். தரவுத்தொகுதி மொத்தம் 11.42 மணிநேர பேச்சு கொண்ட 8516 பேச்சுக் கோப்புகளைக் கொண்டுள்ளது. ஒவ்வொரு பேச்சு கோப்பும் ஒரு நபரை நான்கு உடனடி ஸ்கிரிப்டுகளில்/எழுத்துவடிவங்களில் ஒன்றிலிருந்து ஒரு பிராம்டை வாசிப்பதை பிரதிநுத்துவம்/உருப்படுத்தம் செய்கின்றது.

உணர்ச்சிபூர்வமான மீக்கூறு பேச்சு மற்றும் எழுத்துப்படி விரிதரவு (Emotional Prosody Speech and Transcripts corpus) கேட்பொலி பதிவுகள் மற்றும் உணர்ச்சிபூர்வமான மீக்கூறில் ஆராய்ச்சியை ஆதரிக்க வடிவமைக்கப்பட்ட தொடர்புடைய எழுத்துப்படிகளைக் கொண்டுள்ளது. பதிவுகள் தொழில்முறை நடிகர்கள் பதினான்கு தனித்துவமான உணர்ச்சி வகைகளைக் கொண்ட தொடர்ச்சியான பொருண்மையியல்சார் நடுநிலை கூற்றுகளைப் படிப்பதைக் கொண்டது. இது ஒரு செய்தியின் எழுதப்பட்ட வடிவத்தில் இல்லாத பேச்சின் அம்சங்களைப் (உணர்ச்சி, உள்ளுணர்வு) பெறுவதை நோக்கமாகக் கொண்டுள்ளது. இந்தச் சோதனைகளில் எளிய சொற்றொடர்கள் மாறுபட்ட சூழல்களைப் பிரதிபலிக்கும் வழிகளில் வெளிப்படுத்தப்படுகின்றன. ஒரே சொற்றொடர் வெவ்வேறு கேள்விகளுக்கு பதிலளிக்க, பேச்சாளரிடமிருந்து மாறுபட்ட தொலைவில் உள்ள கேட்பவர்களிடம் உரையாற்ற அல்லது தனித்துவமான உணர்ச்சி நிலைகளை வெளிப்படுத்த பயன்படுத்தப்படலாம்.

பாஸ்க் பேச்சுத் தரவுத்தளம் (Basque Speech Database) நிலையான தொலைபேசி வலையமைப்பில் பதிவுசெய்யப்பட்ட 1060 பாஸ்க் பேசுபவர்களின் பதிவுகளைக் கொண்டுள்ளன. ஒவ்வொரு பேசுபவரும் 43 ஐடங்களை வாசித்தனர் மற்றும் தன்னிச்சையாக உச்சரித்தனர்.

பிஸ்காய்போன் பேச்சுத் தரவுத்தளம் (Bizkaifon Speech Database) ஒலி காப்பகங்கள் மற்றும் பேச்சு பாஸ்கின் கிளைமொழி வகைகளின் தொடர்புடைய தகவல்களைக் கொண்டுள்ளது. இது 21 மணிநேர தன்னிச்சையான மற்றும் வாசிப்புப் பேச்சைக் கொண்டுள்ளது, ஒரு அறையில் ஆர்த்தோகிராஃபிக் டிரான்ஸ்கிரிப்ஷன் உடன் மைக்ரோஃபோனில் பதிவு செய்யப்பட்டுள்ளது.

2.5. சுருக்கவுரை

இந்த இயல் தரவுத்தொகுதி மொழியியலின் இன்றைய நிலை பற்றி விளக்குகின்றது. இயலின் தொடக்கமாக ஒரு சுருக்கமான அறிமுகம் தரப்பட்டுள்ளது. இதைத் தொடர்ந்து, தரவுத்தொகுதி உருவாக்கத்தின் சுருக்கமான அறிமுகம் தரப்பட்டுள்ளது. பின்னர் எழுத்துத் தரவுத்தொகுதிகள் பற்றிய விவரங்கள் ஒருசில எடுத்துக்காட்டுகளுடன் தரப்பட்டுள்ளன. பின்னர் பேச்சுத் தரவுத்தொகுதிகள் பற்றிய விவரங்களும் ஒருசில எடுத்துக்காட்டுகளுடன் தரப்பட்டுள்ளன.

இயல் 3

தரவுத்தொகுதியின் கருத்துருசார் வகைப்படுத்தல்

3.1. அறிமுகம்

மின் தரவுத்தொகுதி (Electronic corpora) புதிய கருத்துருவாகும். இதை எவ்வாறு வகைப்படுத்துவது என்பதில் பொதுவாக ஏற்றுக்கொள்ளத்தகுந்த கருத்து இல்லை. இருப்பினும் அவற்றை வகைப்படுத்த வேண்டுவது அவசியமாகும். தரவுத்தொகுதிகளை வகைப்படுத்த பல காரணிகளைப் பயன்படுத்தலாம். மொழியியல் காரணிகள் அக அடிப்படையிலும் புற அடிப்படையிலும் அமையும். புறக் காரணிகள் பங்கெடுப்பாளர்கள், நேர்வு, சமூக நிலை, மொழியின் கருத்து பரிமாற்றச் செயல்பாடு போன்றவற்றுடன் தொடர்புடைய உரை வகைப்பாட்டியல் (text typology) அடிப்படையில் அமையும். அகக் காரணிகள் மொழியின் சிறுபகுதிகளுக்குள் மொழி அமைப்பொழுங்குகளின் மறுநிகழ்வு அடிப்படையில் அமையும். இவ்வெல்லாச் சிக்கல்களையும் எடுத்துக் கொண்டு தரவுத்தொகுதியைப் பின்வரும் வழியில் பரந்த முறையில் வகைப்படுத்தலாம்:

1. உரையின் இனம் (Genera of text)
2. தரவின் இயல்பு (Nature of Data)
3. உரையின் வகை (Type of Text)
4. திட்ட வரைவின் நோக்கம் (Purpose of design)
5. பயன்பாட்டின் இயல்பு (Nature of application)

இவ்வகைப்படுத்தலைத் தரவுத்தொகுதியின் கருத்துருசார் வகைப்படுத்தல் (Conceptual Classification Of Corpora) எனலாம்.

3.2. உரையின் இனம் அடிப்படையில்

உரையின் இனம் (genera of text) அடிப்படையில் தரவுத்தொகுதியை எழுதப்பட்ட தரவுத்தொகுதி (written corpus) என்றும் பேச்சுத் தரவுத்தொகுதி (speech corpus) என்றும் பேசப்பட்ட தரவுத்தொகுதி (spoken corpus) என்றும் மூன்றாக வகைப்படுத்தலாம்.

3.2.1 எழுத்துத் தரவுத்தொகுதி

எழுத்து விரிதரவு எழுதப்பட்ட, அச்சடிக்ககப்பட்ட, வெளியிடப்பட்ட மற்றும் மின் அணு வடிவில் உள்ள பல மொழித் தரவுகளிலிருந்து சேகரிக்கப்பட்டவையாகும். எடுத்துக்காட்டாக,

இந்திய மொழியின் எம்.ஐ.டி. விரிதரவெள (MIT Corpus of Indian Language) என்பதைக் கூறலாம்.

தரவுத்தொகுதிகளில் உள்ள எழுதப்பட்ட உரைகள் ஸ்கேன் செய்யப்பட்ட அல்லது மின்னணு முறையில் பதிவிறக்கம் செய்யப்பட்ட புத்தகங்கள், செய்தித்தாள்கள் அல்லது பத்திரிகைகளிலிருந்து பெறப்படலாம். பிற எழுதப்பட்ட தரவுத்தொகுதிகளில் இலக்கியப் படைப்புகள் அல்லது ஒரு எழுத்தாளரின் அனைத்து எழுத்துக்களும் இருக்கலாம் (எ.கா. வில்லியம் ஷேக்ஸ்பியர்). சமகால சமுதாயத்தில் மொழி எவ்வாறு பயன்படுத்தப்படுகிறது, காலப்போக்கில் நம் மொழியின் பயன்பாடு எவ்வாறு மாறிவிட்டது, வெவ்வேறு சூழ்நிலைகளில் மொழி எவ்வாறு பயன்படுத்தப்படுகிறது என்பதைப் பார்க்க இதுபோன்ற தரவுத்தொகுதிகள் நமக்கு உதவுகின்றன.

3.2.2 பேச்சுத் தரவுத்தொகுதி

இது முறையான மற்றும் முறைசாராத விவாதங்கள், வாதங்கள், முன்தயாரிக்கப்பட்ட பேச்சுகள், சாதாரணமான மற்றும் இயல்பான பேச்சுகள், இருவர் உரையாடல்கள், ஒருவர் உரையாடல்கள், மேடைப் பேச்சுகள் போன்றவைக் கொண்டிருக்கும். எடுத்துக்காட்டாக, பேச்சு *நியூசிலாண்ட் ஆங்கிலத்தின் வெல்லிங்டன் தரவுத்தொகுதி* (Wellington Corpus of Spoken New Zealand English) என்பது பேச்சுத் தரவுத்தொகுதி ஆகும்.

3.2.3 பேசப்பட்ட தரவுத்தொகுதி

பேசப்பட்ட தரவுத்தொகுதி பேச்சுத் தரவின் நுட்பமான நீட்சியாகும். இதில் பேச்சானது ஒலி பெயர்ப்பிற்காக மட்டுமேயல்லாமல் பிற மாற்றங்கள் செய்யப்படாமல் எழுத்து வடிவில் உருப்படுத்தப்பட்டுள்ளன. எடுத்துக்காட்டாக, நியூசிலாண்ட் பேசப்பட்ட ஆங்கிலத்தின் வெல்லிங்டன் தரவுத்தொகுதி (Wellington Corpus of Spoken New Zealand English) லண்டன்-லண்ட் பேசப்பட்ட ஆங்கிலத்தின் தரவுத்தொகுதி (London-Lund Corpus of Spoken English) என்பனவற்றைக் கூறலாம்.

பேசப்பட்ட தரவுத்தொகுதிகள் பேசும் மொழியின் எழுத்துப்படிளைக் கொண்டுள்ளது. இத்தகைய எழுத்துப்படிகள் மக்களின் வீடுகள் மற்றும் பணியிடங்களில் பதிவு செய்யப்பட்ட சாதாரண உரையாடல்கள் அல்லது தொலைபேசி அழைப்புகள், வணிகக் கூட்டங்கள், வானொலி ஒலிபரப்புகள் அல்லது தொலைக்காட்சி நிகழ்ச்சிகள் போன்றவையாக இருக்கலாம். எழுதப்பட்ட தரவுத்தொகுதிகளைப் போலவே, நிஜ வாழ்க்கையிலும் பல சூழல்களிலும் மொழி எவ்வாறு பயன்படுத்தப்படுகிறது என்பதை பேசும் தரவுத்தொகுதிகள் நமக்குக் காட்டுகிறது.

3.3. தரவின் இயல்பு அடிப்படையில்

தரவின் இயல்பு (nature of data) அடிப்படையில் தரவுத்தொகுதிகளைப் பொதுத் தரவுத்தொகுதி (General Corpus), சிறப்புத் தரவுத்தொகுதி (special corpus), துணைமொழி தரவுத்தொகுதி (sub language corpus), மாதிரித் தரவுத்தொகுதி (sample corpus), இலக்கியத் தரவுத்தொகுதி (literary corpus), கண்காணிப்புத் தரவுத்தொகுதி (monitor corpus) என வகைப்படுத்தலாம்.

3.3.1. பொதுத் தரவுத்தொகுதி

பொதுத் தரவுத்தொகுதி (general corpus) ஒரு வகைத் தரவுத்தொகுதி ஆகும். இதில் பல்வேறு வகையான பொருண்மைகளில் (subjects) எழுதப்பட்ட அல்லது பேசப்படும் பல்வேறு வகையான உரைகள் அடங்கும். சில நேரங்களில் இது மொழி கற்றல், மொழிபெயர்ப்பு போன்றவற்றுக்கான நோக்கீட்டுப் பொருளாக (reference material) அதன் செயல்பாட்டைப் பற்றி "நோக்கீட்டுத் தரவுத்தொகுதி" "reference corpus" என்று அழைக்கப்படுகிறது. பொது தரவுத்தொகுதி (general corpus) வேறுபட்ட துறைகள், இனங்கள், பாடக் களங்கள் மற்றும் நடைகள் இவற்றைச் சார்ந்த பொது உரைகளைக் கொண்டிருக்கும். பிரபலமான பொது தரவுத்தொகுதிகளின் எடுத்துக்காட்டாக 100 மில்லியன் சொற்கள் கொண்ட பிரிட்டிஷ் தேசிய தரவுத்தொகுதி (British National Corpus (BNC/பிஎன்சி)) மற்றும் 400 மில்லியன் சொற்கள் கொண்ட ஆங்கில வங்கி ஆகியவற்றைக் கூறலாம்.

இதன் வடிவம் மற்றும் பயன்பாடு அடிப்படையில் இது உரை சேகரிப்பில் முடிவான எண்ணிக்கையைக் கொண்டது. அதாவது, இதில் உரை வகைகளின் மற்றும் சொற்கள் மற்றும் வாக்கியங்கள் இவற்றின் எண்ணிக்கை ஒரு எல்லைக்குட்பட்டது. இது கால அடிப்படையில் வளரவோ புதிய உரைகளின் இருப்பு அடிப்படையில் மேலும் அதிகரிக்கப்படவோ செய்யாது. ஆனால் இது அளவில் மிகப் பெரிதாகவும் வகைவேறுபாட்டில் வளமானதாகவும் உருப்படுத்தத்தில் விரிந்ததாகவும் பயன்பாட்டில் பரந்தாகவும் இருக்கவேண்டும்.

3.3.2. சிறப்புத் தரவுத்தொகுதி

சிறப்புத் தரவுத்தொகுதி (special corpus) என்பது ஒரு குறிப்பிட்ட வகை உரைகளை உள்ளடக்கிய ஒரு தரவுத்தொகுதி ஆகும். இந்த சிறப்புத்தன்மைக்கு திட்டவட்டமான எல்லைகள் இல்லை, ஆனால் கேள்விக்குரிய உரையின் வகையைக் குறிப்பிடும் சில அளவுகோல்களைக் கருத்தில் கொள்ள வேண்டும். அத்தகைய தரவுத்தொகுதிகளில் ஒரு குறிப்பிட்ட காலக்கெடு (1822

முதல் 1876 வரையிலான உரைகள்) அல்லது ஒரு குறிப்பிட்ட பொருள் (கலை, அரசியல், மருத்துவம்) அல்லது வேறு சில காரணிகளின் அடிப்படையில் சிறப்பு வாய்ந்த சில உரைகள் இருக்கலாம். சில பிரபலமான *சிறப்பு நோக்கங்களுக்கான மொழி* (Language for Special Purpose (LSP)) தரவுத்தொகுதிகள் என்பது 5 மில்லியன் சொற்கள் கொண்ட கேம்பிரிட்ஜ் மற்றும் நாட்டிங்ஹாம் ஆங்கிலக் கருத்தாடல் தரவுத்தொகுதி (Cambridge and Nottingham Corpus of Discourse in English (CANCODE/கான்கோட்) மற்றும் மிச்சிகன் கல்விசார் பேச்சப்பட்ட ஆங்கிலத்தின் தரவுத்தொகுதி (Michigan Corpus of Academic Spoken English (MICASE/மைக்காஸ்) ஆகும்.

சிறப்புத் தரவுத்தொகுதி (special corpus) மொழியின் குறிப்பிட்ட வகை, கிளைமொழி மற்றும் ஆய்வுத் தலைப்பின் சில பண்புகளை வலியுறுத்தும் பாடம் இவற்றிற்கு வேண்டி பொது தரவுத்தொகுதி மாதிரிபடுத்தப்பட்ட உரைகளிலிருந்து திட்டவரைவுசெய்யப்பட்டுள்ளது. எடுத்துக்காட்டாக CHILDES Data Base என்பதைக் கூறலாம். இது நோக்கம் அடிப்படையில் அதன் அளவிலும் உள்ளடக்கத்திலும் வேறுபடும். இது அசாதாரணமான பண்புக்கூறுகளின் கூடுதல் பகுதியைக் கொண்டிருப்பதால் இது மொழியின் வருணனைக்குப் பங்களிப்பு செய்யாது. இது இயல்பான நடத்தையில்லாத மக்களிடமிருந்து தரவுகளைப் பதிவு செய்வதால் இதன் மூலம் நம்பத்தகுந்ததல்ல. சிறப்பு தரவுத்தொகுதி (அதன் தரப்பட்ட நோக்கத்திற்குள் அல்லாமல்) சரிசமமானதல்ல. பிற நோக்கத்திற்குப் பயன்படுத்தினால், மொழிப்பகுதிகளின் பிழையான மற்றும் மாற்றப்பட்ட பார்வையைத் தரும். இது இயல்பான, நம்பகமான மொழியின் ஒன்றோ மற்றொன்றோ வகையின் பண்புக்கூறுகளை வெளிப்படுத்துவதால் இது கொள்கை அடிப்படையில் வேறுபட்டதாகும். மொழியின் பிரதிநித்துவமல்லாத இயல்பை உட்படுத்துவதன் காரணமாக குழந்தைகள், தாய்மொழி அல்லாதவர், புறக் கிளைமொழிகளைப் பயன்படுத்துவோர் என்போரின் மொழி மற்றும் கருத்துப்பரிமாற்றத்தின் மிகச் சிற்பான பரப்புகள் (எ.கா. ஏலம், மருத்துவப் பேச்சு, சூதாட்டம், நீதிமன்ற நிகழ்வு) என்பனவற்றின் தரவுத்தொகுதி சிறப்பு தரவுத்தொகுதிகளாகக் கருதப்படவேண்டும்.

3.3.3. துணை மொழித் தரவுத்தொகுதி

துணை மொழித் தரவுத்தொகுதி (sub language corpus) ஒரு குறிப்பிட்ட மொழியின் ஒரே ஒரு உரை வகையை கொண்டிருக்கும். இது ஒரு நோக்கீட்டுத் தரவுத்தொகுதியிலிருந்து (reference corpus) மொழிக் கற்றையின் மறு நுனியில் இருக்கின்றது. இதன் ஒருபடித்தான அமைப்பும்

(homogeneity of structure) சிறப்பிக்கப்பட்ட சொற்களஞ்சியமும் (specialised lexicon) மாதிரி அடிப்படையில் நல்ல மற்றும் நெருங்கிய தன்மைகளை நிறுவ வேண்டி தரவின் அளவைச் சிறதாக வைக்க அனுமதிக்கின்றது.

3.3.4. மாதிரித் தரவுத்தொகுதி

மாதிரித் தரவுத்தொகுதி (Sample corpus) சிறப்புத் தன்மையான தரவுத்தொகுதியின் ஒரு வகையாகும். இது மிகுந்த கவனத்தோடும் விரிவாகவும் ஆயப்பட்ட உரைகளின் முற்றுநிலைச் சேகரிப்பைக் கொண்டிருக்கும் சிறிய மாதிரிகளால் உருவாக்கப்பட்டதாகும். எடுத்துக்காட்டாக *ஆங்கிலச் செய்தித் தாள்களின் சூரிச் தரவுத்தொகுதி* (Zurich Corpus of English News Papers) என்பதைக் கூறலாம். மாதிரி விரிதரவு ஒருதடவை உருவாக்கப்பட்டுவிட்டால் இதில் ஏதாவது சேர்க்கப்படவோ எந்த வழியிலாவது மாற்றப்படவோ செய்வதில்லை· ஏனென்றால் எந்த மாற்றமும் இதன் உள்ளடக்கத்தை (தரவுமையத்தை) சரிசமமில்லாமல் செய்வதுடன் அதன் ஆய்வுத் தேவையை சிதைக்கும். உரைகளைப் பொறுத்தவரையில் மாதிரிகள் குறைந்த எண்ணிக்கையிலானவை மற்றும் நிரந்தர வடிவ அளவு உள்ளது. எனவே அவை உரைகளாகத் தகுதிபெறாது.

3.3.5. இலக்கியத் தரவுத்தொகுதி

இலக்கியத் தரவுத்தொகுதி (Literary corpus) என்பது மாதிரித் தரவுத்தொகுதியின் ஒரு சிறப்பான வகையாகும். இதில் பல வகைகள் இருக்கின்றன. பைபிள்சார் மற்றும் இலக்கிய ஆய்வுகள் வெகுநாட்களுக்கு முன்பே தரவுத்தொகுதி மொழியியல் என்ற துறையை தொடங்கியது. ஆக்கியோரின் படைப்புகளின் ஒழுங்கை நிறுவுதல் போன்ற இலக்கியக் களத் திறன்கள் இருக்கின்றன. இம்மாதிரியான தரவுத்தொகுதி உருவாக்கத்தின் வகைப்படுத்தும் அளவீடுகள் ஆக்கியவர், இனம் (எ.கா. தூது சிறுகதைகள், கதைகள், போன்றவை), குழு (எ.கா. காதல் புலவர்கள், அகஸ்டன் உரைநடை எழுத்தாளர்கள், விக்டோரியன் கதை ஆசிரியர்கள், போன்றோர்), பொருள் (எ.கா. புரட்சி எழுத்துகள், குடுப்பக் கூற்றுகள், தொழில்மயமாக்கம் போன்றவை) மற்றும் மதிக்கத்தக்க காரணிகளாகப் பிற சிக்கல்கள் என்பனவற்றை உள்ளடக்கும்;. இருப்பினும் அறியப்படாத காரணத்திற்காக நாடகங்களிலிருந்து உருவாக்கப்பட்ட தரவுத்தொகுதி உரைநடையிலிருந்தும் செய்யுளிலிருந்தும் தனியாக வைக்கப்படுகின்றது.

3.3.6. கண்காணிப்புத் தரவுத்தொகுதி

கண்காணிப்புத் தரவுத்தொகுதி (monitor corpus) மொழியின் மாற்றத்தைப் பிரதிபலிக்கும் தரவின் நிரந்தர சேர்க்கைக்கு வழி செய்யும் உரைகளின் வளர்கின்ற, எல்லையற்ற தொகுப்பாகும். தரவுத்தொகுதியின் நிரந்தர வளர்ச்சி மொழியில் மாற்றத்தைப் பிரதிபலிக்கும். ஒரே இணைப்பாக்கத் திட்டவரைவு ஆண்டுக்கு ஆண்டு பின்பற்றப்படுகின்றது. கட்டுப்படுத்தும் விரிதரவின் அடிப்படை ஒரு தனி வருடத்தில் பேசப்பட்ட அல்லது எழுதப்பட்ட உரைகளுக்குக் குறிப்புரை செய்வதாகும். கட்டுப்படுத்தும் தரவுத்தொகுதியிலிருந்து நாம் புதிய சொற்களைக் கண்டுபிடிக்கலாம். வழக்கிலுள்ள வேறுபாட்டைத் தொடர்ந்து அறியலாம்; பொருண்மையிலுள்ள வேறுபாட்டை உற்றுநோக்கி அறியலாம்; நிகழ்வெண்ணின் நீண்ட கால நியமத்தை நிலைநிறுத்தலாம்; சொல் தகவலின் விரிந்த எல்லைப்பரப்பை ஆக்கலாம். தரவின் புதிய மூலங்கள் வருவதாலும் புதிய செயல்முறைகள் அரிதான சாதனங்களை வளமாக இருக்கும் வண்ணம் செய்வதாலும் நாளடைவில் கண்காணிப்புத் தரவுத்தொகுதின் சரிசமநிலை மாறும். இவ்வொழுக்கின் விகிதம் அவ்வவ்போது நேர்செய்யப்படும்.

3.4. உரையின் வகை அடிப்படையில்

உரையின் வகை (type of text) அடிப்படையில் தரவுத்தொகுதி ஒருமொழியத் தரவுத்தொகுதி (monolingual corpus), இரு மொழியத் தரவுத்தொகுதி (bilingual corpus), பன்மொழியத் தரவுத்தொகுதி (multilingual corpus) என்று வகைப்படுத்தலாம்.

3.4.1. ஒருமொழியத் தரவுத்தொகுதி

` ஒருமொழிய தரவுத்தொகுதி (எ.கா. அமிர்தா பல்கலைக்கழகத் தமிழ் தரவுத்தொகுதி) ஒரு காலத்தின் (synchronic) அல்லது பலகாலத்தின் (diachronic) பயன்பாட்டைப் பிரதிநுத்துவம் செய்யும் ஒரு மொழியின் பிரதிநித்துவ உரைகளைக் கொண்டிருக்கும்.

3.4.2. இருமொழியத் தரவுத்தொகுதி

இரு உறவுள்ள அல்லது உறவற்ற மொழிகளின் தரவுத்தொகுதிகள் ஒரே சட்டகத்தில் வைக்கப்படும் போது இருமொழியத் தரவுத்தொகுதி (எ.கா. அமிர்தா பலகலைக்கழக ஆங்கிலம்-தமிழ் தரவுத்தொகுதி) உருவாகும். இம்மொழிகள் இன அடிப்படையில் அல்லது வகைப்பாட்டியல் அடிப்படையில் உறவுள்ளவை என்றால் அவை இணை தரவுத்தொகுதிகளாக மாறும்· உரைகள் சில முன்னரே வரையறுக்கப்பட்ட காரணிகளைப் பின்பற்றி வரிசைப்படுத்தப்படும். இங்கு வடிவ அளவு, பொருளடக்கம் மற்றும் களம் என்பன தரவுத்தொகுதிக்கு தரவுத்தொகுதி மாறும். இணை தரவுத்தொகுதியின் நேர்;வில் இது அனுமதிக்கப்படுவதில்லை.

3.4.3. பன்மொழியத் தரவுத்தொகுதி

பன்மொழியத் தரவுத்தொகுதி (எ.கா. Crater Corpus; இந்திய மொழிகளின் தரவுத்தொகுதிகளின் முயற்சி (Indian Languages Corpora initiative) இரண்டிற்கும் மேற்பட்ட மொழிகளிலிருந்து நல்ல பிரதிநித்துவச் சேகரிப்புகளைக் கொண்டிருக்கும். பொதுவாக இங்கும் இருமொழியத் தரவுத்தொகுதியுலும் உரைகள் வேறுபட்ட மொழிகளைச் சார்ந்திருந்தாலும் ஒற்றுமையுள்ள உரை வகைப்பாடுகள் மற்றும் ஒரே மாதிரியாக்க நெறிமுறைகள் பின்பற்றப்படுகின்றன.

3.5. திட்டவரைவின் நோக்கம் அடிப்படையில்

திட்டவரைவின் நோக்கம் (purpose of design) அடிப்படையில் தரவுத்தொகுதியை விவரம் அடையாளப்படுத்தப்படாத தரவுத்தொகுதி (un-annotated corpus), (விளக்கம்) விவரம் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி (annotated corpus) என்று வகைப்படுத்தலாம்.

3.5.1. விவரம்அடையாளப்படுத்தப்படாத தரவுத்தொகுதி

விவரம் அடையாளப்படுத்தப்படாத தரவுத்தொகுதி (unannotated corpus) (எ.கா. MIT Corpus of Indian Languages) மொழிசார் அல்லது மொழிசாராத் தகவல்களைக் கொண்டிராமல் உரைகளின் எளிய தொடக்க நிலையை உருப்படுத்தம் செய்யும். இது மொழி ஆய்விற்கு ஓரளவுக்கு பயனுள்ளதாகக் கருதப்படுகின்றது. இருப்பினும் தரவுத்தொகுதியின் பயன்பாட்டுக் விளக்க அடையாளங்கள் (annotation) தரப்படும் போது போதிய அளவு அதிகரிக்கும்.

3.5.2. விவரம் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி

விவரம் அடையாளப்படுத்தபட்ட (annotated corpus) தரவுத்தொகுதி (எ.கா. British National Corpus) சில கூடுதலான தகவலகளை (ஆய்வுக் குறிகள், சொல்வகைப்பாட்டுக் குறிகள், இலக்கண வகைப்பாட்டுத் தகவல்கள்) உரையில் பதிவுசெய்ய வேண்டி திட்டவரைவு செய்பவர்களால் வெளியிலிருந்து செருகப்பட்ட அடையாளங்களையும் குறிமங்களையும் கொண்டிருக்கும். அடையாளப்படுத்தப்படாத தரவுத்தொகுதிக்கு மாறாக அடையாளப்படுத்தபட்ட தரவுத்தொகுதி உருபனியல் பகுப்பாய்வு, வாக்கியப் பகுப்பாய்வு, தகவல் மீட்பு, சொற்பொருண்மை மயக்கம் நீக்கல், இயந்திர மொழிபெயர்ப்பு போன்றவற்றை உள்ளடக்கிய மொழித் தொழில் நுட்பத்திற்கு வேண்டிய பல செயல்பாடுகளில் பயன்படும் பொருத்தமான தகவல்களைத் தருவதற்கு கூடுதல் தகுதியானது.

3.6. பயன்பாட்டின் இயல்பு அடிப்படையில்

பயன்பாட்டின் இயல்பு (nature of application) அடிப்படையில் தரவுத்தொகுதியை வரிசைப்படுத்தப்பட்ட தரவுத்தொகுதி (aligned Corpus), இணையான தரவுத்தொகுதி (parallel corpus), நோக்கீட்டு தரவுத்தொகுதி (reference Corpus), ஒப்பீட்டு தரவுத்தொகுதி (comparable corpus), சந்தர்ப்பவாத தரவுத்தொகுதி (opportunistic Corpus) என்று வகைப்படுத்தலாம்.

3.6.1. வரிசைப்படுத்தப்பட்ட தரவுத்தொகுதி

வரிசைப்படுத்தப்பட்ட தரவுத்தொகுதி (aligned corpus) இரு அல்லது பன்மொழி தரவுத்தொகுதியின் ஒரு வகையாகும். இதில் ஒரு மொழியின் மற்றும் அதன் மொழி பெயர்ப்புகளில் உரைகள் வாக்கியங்களுக்கு வாக்கியமாகவும் தொடர்களுக்கு தொடராகவும் சொல்லுக்கு சொல்லாகவும் வரிசைப்படுத்தப் பட்டுள்ளன. எடுத்துக்காட்டாக கனடா ஹேன்ஸர்ட் தரவுத்தொகுதி (The Canadian Hansard Corpus) என்பதைக் கூறலாம்.

3.6.2. இணைத் தரவுத்தொகுதி

இணைத் தரவுத்தொகுதி (parallel corpus) or அல்லது மொழி பெயர்ப்புத் தரவுத்தொகுதி (translation corpus) இரண்டு முறை மொழி பெயர்ப்பு நிகரன்களின் (double-checking translation equivalents) பரிசோதனையை அனுமதிக்கும் மொழிகளின் உரைகளையும் மொழி பெயர்ப்புகளையும் கொண்டிருக்கிறது. இதில் ஒரு மொழியில் உள்ள உரைகள் மற்றும் அவற்றின் மொழி பெயர்ப்புகள் வாக்கியங்களுக்கு வாக்கியம், தொடர்களுக்கு தொடர், சொல்லுக்கு சொல் வரிசைப்படுத்தப் பட்டுள்ளன. எடுத்துக்காட்டாக செம்னிட்ஸ் ஜெர்மன் - ஆங்கிலம்/ஆங்கிலம் - ஜெர்மன் மொழிபெயர்ப்புத் தரவுத்தொகுதி (Chemnitz German – English/ English – German Translation Corpus) என்பதைக் கூறலாம்.

3.6.3. நோக்கீட்டுத் தரவுத்தொகுதி

நோக்கீட்டுத் தரவுத்தொகுதி (reference corpus) ஒரு மொழியின் மொழியைப் பற்றிய விரிவான தரவைத் தரும்படி திட்டமிடப்பட்டுள்ளது. இது ஒரு மொழியின் இலக்கணங்கள், அகராதிகள், சொற்கஞ்சியங்கள் மற்றும் பிற நோக்கீட்டு பொருள்கள் என்பனவற்றை உருவாக்கும் படி தேவையான மொழி வகைகளையும் சிறப்பான சொற்றொகையையும் பிரதிநுத்துவம் செய்வதைக் குறிக்கோளாக கொண்டது. எடுத்துக்காட்டாக ஆங்கிலவங்கி (Bank of English) என்பதைக் கூறலாம்.

3.6.4. ஒப்பீட்டுத் தரவுத்தொகுதி

ஒப்பீட்டுத் தரவுத்தொகுதி (Comparable corpus) ஒன்றுக்கு மேற்பட்ட மொழிகளின் அல்லது மொழி வகைகளின் ஒற்றுமையுள்ள உரைகளின் சேகரிப்பு ஆகும். இது ஒரு வகையான பன்மொழி பெருந்தரவாகும். இதில் பொருளடக்கம், இனம் மற்றும் நடை இவற்றில் வேறுபட்ட மொழிகளின் உரைகளை கொண்டுள்ளது. எடுத்துக்காட்டாக ஐக்கிய ஐரோபியத் தரவுத்தொகுதி (Corpus of European Union) என்று கூறலாம்.

3.6.5. சந்தர்ப்பவாதத் தரவுத்தொகுதி

சந்தர்ப்பவாதத் தரவுத்தொகுதி) (Opportunistic corpus) பல முறைகளிலும் கிடைக்கும் மின் உரைகளின் சேகரிப்பாகும். இது பெரும்பாலும் முடிவுறாமல் இருக்கும். இது பெரும்பாலும் முற்றுப்பெறாமலும் முழுமை பெறாமலும் இருக்கும். கட்டுப்படுத்தும் தரவுத்தொகுதி பெரும்பாலும் சந்தர்ப்பவாதத் தரவுத்தொகுதியாகக் கருதப்படும்.

3.7. சுருக்கவுரை

இவ்வியலில் தரவுத்தொகுதிகளின் கருத்துருசார் வகைப்படுத்தல் பற்றி விளக்குவதை நோக்கமாகக் கொண்டுள்ளது. தொடக்கமாக ஒரு சுருக்கமான அறிமுகம் தரப்பட்டுள்ளது. அறிமுகத்தைத் தொடர்ந்து உரையின் வகையின் அடிப்படையில் தரவுத்தொகுதியை வகைப்பாடு செய்வது குறித்து விளக்கம் தரப்பட்டுள்ளது. அடுத்தபடியாகத் தரவின் தன்மை அடிப்படையில் தரவுத்தொகுதியை வகைப்பாடு செய்வது குறித்து விளக்கம் தரப்பட்டுள்ளது. இதன்பின்னர் உரையின் வகை மற்றும் பயன்பாட்டின் தன்மை ஆகியவற்றின் அடிப்படையில் தரவுத்தொகுதிகளை வகைப்பாடு செய்தல் பற்றிய விவரங்கள் தரப்படுள்ளன.

இயல் 4

எழுத்துத் தரவுத்தொகுதியின் உருவாக்கம்

4.1. அறிமுகம்

தரவுத்தொகுதியை திட்டவரைவு செய்வதிலும் உருவாக்குவதிலும் நிர்வகிப்பதிலும் பல சிக்கல்கள் இருக்கின்றன. தரவுத்தொகுதியை உருவாக்குவதிலும் செயற்பாங்குசெய்வதிலும் உள்ள சிக்கல்கள் விரிதரவின் வகை மற்றும் பயன்பாட்டின் நோக்கம் இவற்றின் அடிப்படையில் வேறுபடும். பேச்சுத் தரவுத்தொகுதி உருவாக்குவதுடன் தொடர்புடைய சிக்கல்கள் உரைத் தரவுத்தொகுதி உருவாக்குவதுடன் தொடர்புடைய சிக்கல்களிலிருந்து வேறுபடும். எனவே தரவுத்தொகுதி உருவாக்கத்தை இரண்டாகப் பகுக்கலாம்:

1. எழுத்து உரைத் தரவுத்தொகுதியை உருவாக்குதல்
2. பேச்சுத் தரவுத்தொகுதியை உருவாக்குதல்

பேச்சுச் தரவுத்தொகுதிகளை உருவாக்க பின்வரும் கருத்துக்கள் கவனத்தில் கொள்ளப்பட வேண்டும்: 1. பயன்பாட்டின் நோக்கம், 2. தகவலாளியின் தேர்வு, 3. தரவுத்தொகுதியின் அளவு, 4. அமைப்புகளின் தேர்வு, 5. தரவு மாதிரியின் முறை, 6. தரவு சேகரித்தலின் முறை, 7. எழுத்துப்பெயர்ப்பின் சிக்கல், 8. தரவு குறியாக்கத்தின் வகை, 9. தரவு கோப்புகளின் நிர்வகிப்பு, 10. உள்ளீட்டுத் தரவின் திருத்தல், 11. உரைகளின் பகுப்பாய்வு, 12. உரைகளின் ஆய்வு.

எழுத்து உரைத் தரவுத்தொகுதி வை உருவாக்குவது பின்வரும் விஷயங்களை உள்ளடக்கும்: 1. விரிதரவின் அளவு, 2. மூல மாதிரி, 3. தாய்மொழித் தன்மை, 4. இலக்கு பயன்படுத்துபவரின் நிர்ணயம், 5. காலகட்டத்தின் தேர்வு, 6. ஆவணங்களின் தேர்வு, 7. உரை ஆவணங்களின் சேகரிப்பு (நூல்கள், செய்திதாள்கள், காலஇதழ்கள் போன்றவை), 8. தரவு மாதிரியின் நெறிமுறை (ஒருவருடைய தேவைக்கு தக்கவாறு சேகரிக்கப்பட்ட விஷயங்களின் வரிசை), 9. தரவு உள்ளீட்டின் முறை (ஒழுங்கற்ற, சீரான, தேர்வு முறை), 10. தரவுத்தொகுதிவைத் திருத்துதல் (தவறுகளைத் திருத்தல், அயற்சொற்கள், மேற்கோள்கள், கிளைமொழிகள் இவற்றை விட்டுவிடுதல்), 11. தரவுத்தொகுதி கோப்புகளின் நிர்வகிப்பு, 12. வெளியீட்டு உரிமையின் சிக்கல்

4.2. தரவுத்தொகுதியின் வடிவ அளவு

தரவுத்தொகுதியின் வடிவ அளவைக் (size of corpus) கருத்தில் கொள்ளவேண்டும். ஒரு தரவுத்தொகுதி எவ்வளவு பெரியதாக இருக்க வேண்டும் என்பது கேள்விக்குறியாகும். விரிதரவு உருவாக்கத்தில் வடிவ அளவு முக்கிய பங்குவகிக்கிறது. இது தரவுத்தொகுதியில் மொத்தம்

எத்தனை சொற்கள் (tokens) இருக்கின்றன, வேறுபட்ட சொற்கள் (types) இருக்கின்றன என்பதுடன் தொடர்புடையதாகும். இது தரவுத்தொகுதியில் எத்தனை வகைபாடுகள் வைக்கப்பட வேண்டும், ஒவ்வொரு வகைப்பாடடிலும் எத்தனை உரைகள் மாதிரிகள் வைக்கப்பட வேண்டும், ஒவ்வொரு மாதிரியிலும் எத்தனைச் சொற்கள் இருக்க வேண்டும் என்ற தீர்மானத்தை உள்ளடக்கும். ஒரு தரவுத்தொகுதியில் எத்தனைச் சொற்கள் அல்லது சொல் வகைகள் இருக்க வேண்டும் என்பது முக்கியமாக இருந்தாலும் தற்போது மிகப் பெரிய தரவுத்தொகுதி தான் மொழியைப் பிரதிபலிப்பதாக அமைகின்றது என்பது நம்பப்படுகிறது. எடுத்துக்காட்டாக, Bank of English, BNC, Cobuild Corpus, Longman/ Lancaster Corpus போன்ற விரிதரவுகள் 100 மில்லியன் சொற்களைக் கொண்டிருக்கின்றன.

4.3. உரையின் பிரதிநிதித்துவத் தன்மை அல்லது மூலமாதிரி

போகப்போக உரைகளின் பிரதிநிதித்துவத் (representativeness of texts) தன்மையின் சூழலில் விரிதரவின் வடிவ அளவு முக்கியத்துவம் இழக்கும். மிகப்பெரிய தரவு ஒரு மொழியின் அல்லது அதன் ஒருவகையை ஒரு சிறிய ஆனால் நன்றாகச் சமநிலையை உடைய தரவுத்தொகுதி வைக் காட்டிலும் கூடுதல் பிரதிநிதித்துவம் செய்யத் தேவையில்லை. உரைகளின் பெரிய சேகரிப்பு நாம் எந்த பொதுமையாக்கத்திலும் பயன்படுத்தவியலும் ஒரு தரவுத்தொகுதி வாக இருக்கத் தேவையிலை. விரிதரவின் ஆய்வின் அடிப்படையிலான கண்டுபிடிப்புகள் ஒரு மொழியின் முழுமைக்கோ அதன் ஒரு குறிப்பிட்ட பகுதிக்கோ பொதுமையாக்கம் செய்யப்பட்டால் தான் நாம் ஒரு தரவுத்தொகுதியைப் பிரதிநிதித்துவத் தன்மையுடையது என்று கூறவியலும். எனவே தான் நாம் தரவுத்தொகுதியில் தரவின் அளவின்மீது கவனக்குவிப்பு செய்யாமல் தரவின் பண்பிற்கு முக்கியத்துவம் தருகின்றோம். அதாவது தரவு ஒரு தரவுத்தொகுதிக்குள் மொழிப்பயன்பாட்டின் எல்லாச் சாத்தியமான பரப்புகளிலிருந்தும் சரியான விகிதத்தில் இருக்கவேண்டும். விரிதரவின் ஒட்டுமொத்த வடிவ அளவு பொருத்தமான பிரதிநித்துவத்தன்மையைப் பெற உதவ வேண்டி மூலப்பொருள்களின் வேறுபட்டத்தன்மையைப் பிரதிபலிக்க வேண்டும். உரை வகைபாடுகளுக்குள் எந்த அளவுக்கு தனிநிலை மாதிரிகளின் எண்ணிக்கை அதிகமாக இருக்கிறதோ அந்த அளவுக்கு மொழியியல் மாறிகளின் ஆய்வின் நம்பகத்தன்மை அதிகமாக இருக்கும். எடுத்துக்காட்டாக, பிரவுண் விரிதரவு (Brown Corpus), லோப் விரிதரவு (LOB Corpus), செவு (SEU) என்பன அமேரிக்காவிலும் யுகேயிலும் பயன்படுத்தப்படும் ஆங்கில மொழியின் நல்ல பிரதிநிதிகளாக நாம் கருதவியலும் அளவுக்கு

கவனமாகத் திட்டவரைவு செய்யப்பட்டுள்ளன. இருப்பினும் மிகக்கூடுதல் வேறுபட்ட அமைப்பையும் பிரதிநிதித்துவச் சட்டகத்தையும் கொண்ட 100 மில்லியன் சொற்களையுடைய பிஎன்சி (BNC) தரவுத்தொகுதியுடன் பிரவுண் (Brown), லோப் (LOB) மற்றும் செவு (SEU) தரவுத்தொகுதிகளை ஒப்பிடும்போது எந்த அளவிற்கு இவ்விரிதரவுகள் பொருளடக்கத்தில் சிறிதாகவும் அமைப்பில் குறைந்த வேறுபாடுடையதாகவும் இருக்கின்றன என்பதைக் காட்டும். இது தரவுத்தொகுதியில் வடிவ அளவு மற்றும் பிரதிநித்துவம் இவற்றிற்கிடையிலான சிக்கல்களை மிக எளிதாக அனுபவவாத அடிப்படையில் தீர்க்கின்றது.

4.4. தாய்மொழித் தன்மை

அடுத்தபடியாகத் தாய்மொழித் தன்மை குறித்த கேள்வி (question of nativity) கருத்தில் கொள்ளப்படவேண்டும். தரவுத்தொகுதி உருவாக்கும் போது தாய்மொழி பேசுபவரின் உரைகளைத் தேர்வு செய்ய வேண்டுமா அல்லது தாய்மொழி பேசாதவரின் உரைகளைத் தேர்வு செய்ய வேண்டுமா என்பது கேள்வி. பொதுவாகக் கூறப் போனால் தாய்மொழி பேசுபவருடைய உரைகள் தாய்மொழி பேசாதவர்களின் உரைகாட்டிலும் நம்பத் தகுந்ததாய் அமையும். இருப்பினும் அது ஒரு கண்காணிப்புத் தரவுத்தொகுதி (monitor corpus) என்றால் தாய்மொழியாகப் பயன்படுத்தாதவர்களைக் காட்டிலும் தாய்மொழியாகப் பயன்படுத்துவோர்களால் உருவாக்கப்படும் உரைகள் முக்கியத்துவம் பெறவேண்டும். ஏனென்றால் மானிட்டர் விரிதரவின் நோக்கம் எல்லா மொழியியல் மற்றும் மொழித் தொழில் நுட்பச் செயல்பாடுகளிலும் முன்மாதிரியாகக் கருதப்படும் மொழியைப் பிரதிநித்துவம் செய்வதாகும். எடுத்துக்காட்டாக, பிரிட்டிஷ் ஆங்கில மொழியின் தரவுத்தொகுதி உருவாக்கத்தின் போது இந்திய ஆங்கில உரைகள் அதில் உள்ளடக்க கூடாது. இது மறுதலையாகவும் அமையும்.

தயாரிக்கப்பட்ட எடுத்துக்காட்டுகளின் மேற்கோள் மற்றும் ஒரு கண்காணிப்புத் தரவுத்தொகுதியில் 'ஒழுங்கற்ற' வாக்கியங்களின் பட்டியல் தரவுத்தொகுதியின் மொழியியல் பகுப்பாய்வின் முடிவுகளில் மிகவும் குறிப்பிடத்தக்க விளைவைக் கொண்டுள்ளன. ஏனெனில், அந்த விஷயத்தில், தரவுத்தொகுதியில் சொற்களையும் சொற்றொடர்களையும் 'பயன்படுத்துவதை' (use) விட நிறைய 'குறிப்பீடுகள்' ('mention') நமக்குக் கிடைக்கின்றன. ஒரு தரவுத்தொகுதியை உருவாக்குவதற்கான முக்கிய காரணங்களில் ஒன்று, இயற்கையாக நிகழும் மொழியை பகுப்பாய்வு செய்ய நமக்கு உதவுவதாக இருந்தால், என்ன நடக்கிறது, எது நடக்காது என்பதைப் பார்க்க, பின்னர் தயாரிக்கப்பட்ட பல எடுத்துக்காட்டு வாக்கியங்கள் மற்றும் சொற்றொடர்களை

அனுமதிப்பது குறைவான பொருத்தமாக இருக்கும் முன்மொழியப்பட்ட நோக்கம். இதையும் மற்றும் சிறப்பு தரவுத்தொகுதியில் காணப்படும் பல சாத்தியமான சிக்கல்களையும் தவிர்ப்பதற்கான ஒரு வழி, தரவுத்தொகுதியில் உரைகளைச் சேர்ப்பதற்கான ஒரு அளவுகோலைப் பயன்படுத்துவது; அது இயல்பில் மிகவும் தொழில்நுட்பமாக இருக்கக்கூடாது.

சிறப்புத் தரவுத்தொகுதியைப் பொறுத்தவரை, சொந்தமற்ற பயனர்களால் (non-native users) தயாரிக்கப்படும் உரைகள் கருதப்படுகின்றன; ஏனெனில் ஒரு சிறப்பு தரவுத்தொகுதியின் நோக்கம் சொந்தமற்ற பயனர்களின் பொதுவான தனித்தன்மையை முன்னிலைப்படுத்துவதாகும். இங்கே தரவுத்தொகுதியின் பிரதிநிதித்துவம் பற்றிய கேள்வி ஒட்டுமொத்த மொழியுடன் தொடர்புடையது அல்ல; ஒரு குறிப்பிட்ட வகுப்பு மக்கள் இரண்டாவது மொழியாகப் கற்றுக் கொண்ட மற்றும் அவர்கள் பயன்படுத்திய மொழியுடன் தொடர்புடையது.

இங்கு கருத்தத்தக்கது, மொழியியல் தொடர்புகளின் பல்வேறு முக்கிய நீரோட்டங்களில் ஒரு மொழி பொதுவாக எவ்வாறு பயன்படுத்தப்படுகிறது என்பது பற்றிய தகவல்களைச் சேகரிக்க இயலும் ஒரு தரவுத்தொகுதியை வைத்திருப்பது ஆகும்.

சொல் பயன்பாடு, எழுத்துப்பிழை, தொடரியல் கட்டுமானங்கள், அர்த்தங்கள் போன்றவற்றுக்கான வழிகாட்டுதல்களை வழங்கும் சில உரைகள் மற்றும் நோக்கீடுகளை நாம் உருவாக்க முயற்சிக்கும்போது, பெரும்பாலும் சொந்த பயனர்களின் உரைகளைப் பெற விரும்புகிறோம்.

கொள்கை அடிப்படையில், சொந்த பயனர்களால் எழுதப்பட்ட மற்றும் பேசப்பட்ட இந்த உரைகள் மொழி புரிந்துகொள்ளும் திறனை மேம்படுத்துவதற்கும் மொழி கற்பவர்களுக்குப் பயன்படுத்துவதற்கும் வழிகாட்டியாகவும், பொருத்தமானதாகவும் மற்றும் பிரதிநிதியாகவும் இருக்கும். ஒருவேளை, இது சொந்த மொழி அல்லாத பயனர்களின் விருப்பத்தின் வரிசையில் சரியாகச் செல்கிறது; அவர்கள் இரண்டாவது மொழியைக் கற்கும்போது ஒரு சொந்த மொழி பயனரின் செயல்திறனை அடைவதை நோக்கமாகக் கொண்டுள்ளனர்.

4.5. இலக்குப் பயன்பாட்டாளரின் நிர்ணயம்

இலக்குப் பயன்பாட்டாளரின் நிர்ணயமும் (determination of target users) கருதப்படவேண்டும். ஒரு பொதுவான தரவுத்தொகுதிக்கு குறிப்பட்ட இலக்குப் பயன்பாட்டாளர் இல்லை. யாரும் ஒரு பொது தரவுத்தொகுதியை எதற்கும் பயன்படுத்தலாம். ஒரு சிறப்பு தரவுத்தொகுதிக்கு (specialized corpus) யார் பயன்பாட்டாளர் என்ற கேள்வி முக்கியமானதாகும்.

ஒவ்வொரு விசாரணையாளருக்கும் ஆய்வாளருக்கும் குறிப்பிட்ட தேவை இருப்பதன் காரணமாக தரவுத்தொகுதி அதனடிப்படையில் திட்டவரைவு செய்யப்படவேண்டும். எடுத்துக்காட்டாக மொழிபெயர்ப்பு கருவிகளின் உருவாக்கத்தில் ஈடுபட்டுள்ள ஒரு நபர் பொது தரவுத்தொகுதியைக் காட்டிலும் இணையான தரவுத்தொகுதியை வேண்டுவார். அதுபோல் இரண்டோ அதற்கு மேற்பட்ட மொழிகளின் ஒப்பீட்டு ஆய்வில் ஈடுபட்டுள்ள நபர் மானிட்டர் தரவுத்தொகுதியைக் காட்டிலும் ஒப்பிடவியலும் தரவுத்தொகுதியை வேண்டுவார்.

இலக்கு பயன்பாட்டாளர்	**தரவுத்தொகுதி**
வர்ணனை மொழியியலார்கள்	பொதுத் தரவுத்தொகுதி, எழுத்துத் தரவுத்தொகுதி மற்றும் பேச்சுத் தரவுத்தொகுதி
பேச்சு தொழில் நுட்ப ஆய்வாளர்	பேச்சுத் தரவுத்தொகுதி (உரையிலிருந்து பேச்சு, பேச்சு புரிதல், பேச்சு உருவாக்கம், பேச்சுச் செயலாக்கம், பேச்சு சரிசெய்தல் போன்றவை)
அகராதியியலார்கள் மற்றும் கலைச்சொல் அகராதியியலார்கள்	பொதுத் தரவுத்தொகுதி, கண்காணிப்புத் தரவுத்தொகுதி, சிறப்புத் தரவுத்தொகுதி, நோக்கீட்டு தரவுத்தொகுதி, சந்தர்ப்பவாத தரவுத்தொகுதி போன்றவை
உரையாடல் ஆய்வாளர்கள்	பேச்சுத் தரவுத்தொகுதி, அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி, சிறப்பிக்கப்பட்ட தரவுத்தொகுதி
சமுதாய மொழியிலார்கள்	பொதுத் தரவுத்தொகுதி, பேச்சுத் தரவுத்தொகுதி, எழுத்துத் தரவுத்தொகுதி மற்றும் சிறப்புத் தரவுத்தொகுதி
உளமொழியியலார்கள்	சிறப்புத் தரவுத்தொகுதி, பேச்சுத் தரவுத்தொகுதி, எழுத்துத் தரவுத்தொகுதி
வரலாற்றியலார்கள்	இலக்கியத் தரவுத்தொகுதி, இருகாலத் தரவுத்தொகுதி
சமூக அறிவியலார்கள்	பொதுத் தரவுத்தொகுதி, பேச்சுத் தரவுத்தொகுதி, எழுத்துத் தரவுத்தொகுதி, சிறப்புத் தரவுத்தொகுதி
ஒப்பீட்டு மொழியியலார்கள்	இருமொழியத் தரவுத்தொகுதி, பன்மொழியத் தரவுத்தொகுதி, இணை தரவுத்தொகுதி, ஒப்பிடவியலும் தரவுத்தொகுதி, அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி

மொழிபெயர்ப்பு வல்லுநர்கள்	இருமொழியத் தரவுத்தொகுதி, பன்மொழியத் தரவுத்தொகுதி, இணை தரவுத்தொகுதி, ஒப்பிடவியலும் தரவுத்தொகுதி, அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி
தகவல் மீட்பு வல்லுனர்கள்	பொதுத் தரவுத்தொகுதி, மானிட்டர் தரவுத்தொகுதி மற்றும் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி
அடையாளப்படுத்தும் பகுப்பாயும் பகுத்துக்குறிக்கும் வல்லுநர்கள்	அடையாளப்படுத்தப்பட்ட மானிட்டர், எழுதப்பட்ட தரவுத்தொகுதி, பேச்சுத் தரவுத்தொகுதி, பொதுத் தரவுத்தொகுதி
மையஇலக்கண திட்டவரைவாளர்கள்	ஒப்பிடவியலும் தரவுத்தொகுதி, இருமொழியத் தரவுத்தொகுதி மற்றும் பொதுத் தரவுத்தொகுதி
சொற்பொருண்மைமயக்கநீக்கத்தில் ஈடுபடுபவர்கள்	அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி, மானிட்டர் தரவுத்தொகுதி, எழுதப்பட்ட தரவுத்தொகுதி, பேச்சுத் தரவுத்தொகுதி, பொதுத் தரவுத்தொகுதி
ஆசிரியர்கள், மாணவர்கள்	கற்பவர் தரவுத்தொகுதி, கண்காணிப்புத் தரவுத்தொகுதி, பொதுத் தரவுத்தொகுதி

4.6. காலகட்டத்தின் தேர்வு

காலகட்டத்தின் தேர்வு (selection of time-span) அடுத்த அம்சமாகும். மொழியானது கால அடிப்படையில் மாறும். எனவே ஒரு குறிப்பிட்ட காலகட்டத்திற்கு மொழியின் பண்புக் கூறுகளைக் கண்டுபிடிக்க நாம் ஒரு குறிப்பிட்ட காலகட்டத்தை நிர்ணியக்க வேண்டும். பெருந்தரவு தெளிவான கால அடையாளப்படுத்தலுடன் ஒரு குறிப்பிட்ட காலத்தை உட்படுத்த முயற்சிக்கிறது. எடுத்துக்காட்டாக, இந்திய மொழிக்கான எம்.ஐ.டி. தரவுத்தொகுதி தற்கால மொழியின் நிலையைப் பிரதிநித்துவம் செய்யவேண்டி 1981-இலிருந்து 1995 வரையிலான காலக்கட்டத்தில் வெளியிடப்பட்ட விஷயங்களைக் கொண்டிருக்கின்றது.

4.7. உரை வகையின் தெரிவு

உரைகளின் வகையின் தெரிவு (selection of text type) அடுத்ததாகக் கருதப்பட வேண்டியதாகும். எழுத்துத் தரவுத்தொகுதி திட்டமிடுதலின் முக்கியமான கேள்வி அது எழுத்து உரைகளின் எல்லா வகைகளையும் கொண்டிருக்க வேண்டுமா என்பதாகும். பெரும்பாலான விரிதரவுகள் தரப்படுத்தப்பட்ட எழுத்தாக்கங்களின் உரைகளைக் கொண்டிருப்பதை விரும்புகின்றது. ஒரு பொது தரவுத்தொகுதியின் நோக்கம் எவை ஒரு மொழியின் மைய

(பொதுவான) பண்புக்கூறுகள், எவை தனிப்பட்ட (சிறப்புப்) பண்புக்கூறுகள் என்பதைக் கண்டுபிடிப்பதாகும். எனவே ஒரு தரவுத்தொகுதி தற்காலத்திய எழுத்தாக்கங்களில் மிக நல்லதை மட்டும் உள்ளடக்க வேண்டும் என்பதல்ல. அளவீடு அடிப்படையிலான விகிதம் போதுமானதாகும். இவ்வாறு ஒரு விரிதரவு மனித அறிவின் வேறுபட்ட கிளைகளை எடுக்கப்படும் விஷயங்களின் சேகரிப்பாகும். இங்கு மிகப் புகழ்ப்பெற்ற எழுத்தாளர்களின் எழுத்தாக்கங்களும் அறிமுகம் இல்லாத எழுத்தாளர்களின் எழுத்தாக்கங்களும் ஒன்றாகக் கருதப்படும். தரவுச் சேகரிப்பிற்காக வெளியீடுகளின் ஆணவங்களின் (நூல்கள், செய்தித்தாள்கள், காலவிதழ்கள் போன்றவை) சேகரிப்புக்காக எல்லா விபரநிரல்களும் (catalogues) பட்டியல்களும் பார்க்கப்படவேண்டும். இது குறிப்பிட்ட மூலத்தின் தனித்தன்மை மறைக்கப்படுகின்ற பல்வேறுபட்ட மூலங்களிலிருந்தும் துறைகளிலிருந்தும் விஷயங்களைச் சேகரிப்பதால் அதன் இயல்பில் பெரும்பான்மையும் பலபடித்தானதாக இருக்கின்றது. திரிக்கப்பட்ட பிரதிநிதித்துவத்தின் எந்த வகைக்கும் எதிராக விரிதரவில் வேறுபட்டத்தன்மை பாதுகாக்கப்படுகின்றது. எம்.ஐ.டி.-தமிழ் விரிதரவு இலக்கியத்திலிருந்து 20%ஐயும், நுண்கலையிலிருந்து 5%ஐயும், சமூக அறிவியலிருந்து 15%ஐயும் இயற்கை அறிவியலிருந்து 15%ஐயும் வணிகவியலிருந்து 10%ஐயும் மக்கள் தகவல் தொடர்பிலிருந்துரு 30%ஐயும் மொழி பெயர்ப்பிலிருந்து 5%ஐயும் கொண்டிருக்கிறது. ஒவ்வொரு வகைப்பாடும் சில துணைவகைப்பாடுகளைக் கொண்டிருக்கிறது. எடுத்துக்காட்டாக,

- இலக்கியம் புதினங்கள், சிறுகதைகள், கட்டுரைகள் போன்றவற்றைக் கொண்டிருக்கிறது.
- நுண்கலைகள் ஓவியங்கள், வரைபடங்கள், இசை, சிற்பம் போன்றவற்றைக் கொண்டிருக்கிறது. • சமூக அறிவியல் தத்துவம், வரலாறு, கல்வி போன்றவற்றைக் கொண்டிருக்கிறது.
- இயற்கை அறிவியல் இயற்பியல், வேதியியல், கணிதவியல், புவியியல் போன்றவற்றைக் கொண்டிருக்கிறது.
- வணிகவியல் வரவுசெலவு, வங்கி போன்றவற்றைக் கொண்டிருக்கிறது.
- மக்கள் தகவல் தொடர்பு செய்தித்தாள், காலவிதழ்கள், சுவரொட்டிகள், அறிவிப்புகள், விளம்பரங்கள் போன்றவற்றைக் கொண்டிருக்கிறது.
- மொழிப்பெயர்ப்பு கருதப்பட்ட மொழியில் மொழி பெயர்க்கப்பட்ட எல்லா பொருண்மைகளையும் (subjects) உள்ளடக்கும்.

4.8. தரவு மாதிப்படுத்தலின் நெறிமுறை

அடுத்ததாகக் கருதப்படவேண்டியது தரவு மாதிப்படுத்தலின் நெறிமுறை (method of data sampling) ஆகும். தரவு ஒருவரின் தேவைக்கேற்ப சேகரிக்கப்பட்ட பொருள்களிலிருந்து (விஷயங்களிலிருந்து) பிரித்தெடுக்கப்படவேண்டும். தரவுத்தொகுதியின் கூடுதலான பிரதிநித்துவத்தை நிறுவுவதற்கு பல வழிகள் இருக்கின்றன. நாம் தரவுத்தொகுதி மாதிரிப்படுத்தலின் வழிமுறைகளை வரையறை விளக்கம் செய்வதற்கு முன் நாம் ஆய விரும்புகிற மொழியின் வகையைத் தெளிவாக வரையறை விளக்கம் செய்யவேண்டும். முறையற்ற மாதிரிப்படுத்தும் உபாயம் (random sampling technique) ஒரு விரிதரவைத் திரிக்கப்பட்டதாக மற்றும் பிரதிநித்துவம் இல்லாததாகச் செய்வதிலிருந்து காக்கின்றது. இத்தரப்படுத்தப்பட்ட உபாயம் இயற்கை மற்றும் சமூக அறிவயில்களின் பல பரப்புகளில் பரவலாகப் பயன்படுத்தப்படுகின்றது. மற்றொரு வழி முழு நோக்கீட்டு நூல்களின் அகரவரிசை அட்டவணையைப் பயன்படுத்துவதாகும். பிரிட்டிஷ் தேசிய நூற்பட்டியல் (British National Bibliography) மற்றும் வில்லிங்கின் பத்திரிக்கை வழிகாட்டி (Willing's Press Guide) என்பன லோப் விரிதரவின் உருவாக்கத்திற்குப் பயன்படுத்தப்பட்டன. மற்றொருரு அணுகுமுறை ஒரு மாதிரிபடுத்தல் சட்டகத்தை வரையறை விளக்கம் செய்வதாகும். பிரவுண் விரிதரவின் திட்டவரைவாளர்கள் இதைப் பயன்படுத்தினார்கள். அவர்கள் ஒரு குறிப்பிட்ட ஆண்டில் வெளியிடப்பட்ட நூல்களையும் காலவிதழ்களையும் பயன்படுத்தினார்கள். ஒரு எழுத்து விரிதரவு செய்தித்தாள் அறிக்கை, காதல் கதைகள், சட்ட நியமங்கள், அறிவியல் எழுத்துரைகள், சமூக அறிவியல்கள், தொழில்நுட்ப அறிக்கைகள் போன்ற இனங்களிலிருந்து உருவாக்கப்பட்டலாம்.

4.9. தரவுகளை உள்ளீடு செய்யும் நெறிமுறை

தரவுகளை உள்ளீடு செய்யும் நெறிமுறையும் (method of data input) கருத்தில் கொள்ளப்படவேண்டும். தரவுகள் பின்வரும் வழிகளில் உள்ளீடு செய்யப்படலாம்:

மின் மூலத்திலிருந்து தரவு (data from electronic source): இச்செயற்பாங்கில் மின்வடிவில் காணப்படுகின்ற செய்தித்தாள்கள், கால இதழ்கள், நாளிதழ்கள், பருவ இதழ்கள், நூல்கள் என்பன உட்படுத்தப்படும்.

இணையவலையிலிருந்து தரவு (data from the web): இது இணையப்பக்கம், இணைய தளம் மற்றும் தனிநபர் பக்கங்கள் என்பனவற்றிலிருந்து உரைகளை உட்படுத்தும். உரைகளைக் கணிப்பொறி படித்தறிதல் (machine reading of text): இது ஒளி எழுத்துணரி ஒழுங்குமுறையால்

(OCR) உரைகளைக் கணிப்பொறி படிக்கவியலும் வடிவில் மாற்றும். இம்முறையைப் பயன்படுத்தி அச்சடிக்கப்பட்ட விஷயங்கள் தரவுத்தொகுதியாக விரைவில் மாற்றப்படுகின்றது.

மனித முயற்சியால் தரவை உள்ளீடுசெய்தல் (manual data input): இது கணிப்பொறி விசைப்பலகையால் செய்யப்படுகின்றது. இது கையினால் எழுதப்பட்டவை, பேச்சு மொழியின் எழுத்துப்பெயர்ப்புகள், பழைய கையெழுத்துப்படிகள் என்பனவற்றிலிருந்து தரவைச் சேகரிப்பதற்கான நல்ல வழியாகும்.

தரவை உள்ளீடு செய்யும் செயன்மை மாதிரி நெறிமுறை அடிப்படையிலானது. நாம் ஒரு நூலிலிருந்து ஒவ்வொரு பத்து பக்கங்களுக்குப் பின்னரும் இரண்டு பக்கங்களைப் பயன்படுத்தவிலும். இது தரவுத்தொகுதியை உரைவடிவில் சேமிக்கப்பட்டுள்ள தரவின் நல்ல பிரதிநியாக மாற்றுகின்றது. ஒரு நூலில் வேறுபட்ட எழுத்தாளர்களால் எழுதப்பட்ட வேறுபட்ட விஷங்களை உள்ளடக்கிய பகுதிகள் இருந்தால் எல்லாப் பகுதிகளிலிருந்தும் இச்செயற்பாங்கால் சேகரிக்கப்படும் மாதிரிகள் சரியாகப் பிரதிநிதித்துவம் செய்யப்படும். தலைப்பு கோப்பு (header file) பதிவேடுகளைப் பாதுகாக்கவும் பதிப்புரிமைச் சிக்லைத் தீர்க்கவும் வேண்டி புத்தகத்தின் பெயர், படைப்பாளி/படைப்பாளிகளின் பெயர்கள், வெளியீட்டு ஆண்டு, பதிப்பு எண், வெளியிட்டவரின் பெயர், உள்ளீடாக எடுக்கப்பட்ட பக்கங்களின் எண் போன்ற எல்லா தகல்களையும் கொண்டிருக்கும்.

உள்ளடக்கப்பொருட்களின் விரிவான பதிவுகளை வைத்திருப்பதும் சாதகமானது; இதன் மூலம் ஆவணங்கள், தரவுத்தொகுதியின் வடிவங்களாகத் தேர்ந்தெடுக்கப்படுகின்ற பிறவற்றிலிருந்து அடையாளம் காணப்படுகிறது. உரை புனைகதை அல்லது புனைகதை அல்லாதது, நூல், பத்திரிகை அல்லது செய்தித்தாள், முறையானது அல்லது முறைசாராதது போன்ற போன்ற தகவல்கள் மொழியியல் மற்றும் மொழியியல் அல்லாத ஆய்வுகளுக்குப் பயனுள்ளதாக இருக்கும்.

உள்ளீட்டு நேரத்தில், உரைகளின் இயல்பான வரி திரையில் பராமரிக்கப்படுகிறது. ஒரு பத்தி உள்ளிட்ட பிறகு, ஒரு வெற்று வரி சேர்க்கப்பட்டு, பின்னர் ஒரு புதிய பத்தி தொடங்கப்படுகிறது.

உரைகள் ஒரு சீரற்ற மாதிரி முறையில் (random sampling manner) சேகரிக்கப்பட்டு, ஒரு புதிய மாதிரி உரையின் தொடக்கத்தில் ஒரு தனித்துவமான குறி (unique mark) வைக்கப்படுகிறது.

4.10. வன்பொருளின் தேவை

அடுத்ததாக வன்பொருளின் தேவைக் (hardware requirement) கருத்தில் கொள்ளப்படவேண்டும். தரவுத்தொகுதி உருவாக்கத்திற்கு ஜிஸ்ட் (GIST) அல்லது எழுத்துப்பட்டி அட்டை (Transcript Card (TC)) கொண்ட தனிநபர் கணினி, எழுத்துரு ஆய்வி என்ற (Script Processor (SP)) மென்பொருள், காட்சித்திரை, அச்சு இயந்திரம், விசைப்பலகை, வன்தட்டுகள் போன்றவை தேவை. கோப்புகளை தனிநபர் கருவியில் நிறுவப்பட்டுள்ள எழுத்துபடி அட்டையைப் பயன்படுத்தி உருவாக்கலாம். இது கணினித் திரையில் பல்வேறு இந்திய எழுத்துருக்களைக் காட்சிப்படுத்த அனுமதிக்கின்றது. இந்திய மொழிகளின் எழுத்துக்களில் பயன்படும் பல்வேறு விசைகளின் குறியங்கள் இந்திய தரத்தின் தகவலகத்தால் (Bureau of Indian Standards) தரப்படுத்தப்படுகின்றது. இதை ஒரு ஒரு தனிநபர் கணினியின் உள்ளே நிறுவுவதன் மூலம், கிட்டத்தட்ட முழு அளவிலான உரை சார்ந்த பயன்பாட்டு தொகுப்புகளைப் பயன்படுத்தலாம். இந்திய மொழியில் தரவை உள்ளீடு செய்து மீட்டெடுக்கலாம். மென்பொருள் மானிட்டரில் இரண்டு செயல்பாட்டு காட்சி முறைகளின் தேர்வையும் வழங்குகிறது: ஒன்று வழக்கமான ஆங்கில பயன்முறையில், மற்றொன்று இந்திய பன்மொழி பயன்முறையில்.

4.11. தரவுத்தொகுதி கோப்புகளின் நிர்வாகம்

விரிதரவு நிர்வாகம் (management of Corpus File) மிகக் கடினமான வேலை. இது வைத்திருப்பது, ஆய்வது, திரையிடுவது, பிற தரவுத்தொகுதியிலிருந்து தகவலை மீட்பது போன்ற வேலைகளை உள்ளடக்கும். ஒரு தரவுத்தொகுதி உருவாகிவிட்டால் அது கணிப்பொறியில் சேமிக்கப்பட வேண்டும். அதைப் பாதுகாப்பதற்கும் விரிவாக்குவதற்கும் திட்டங்கள் தேவை; தவறுகள் திருத்தப்பட வேண்டும்; மாற்றங்கள் செய்யப்பட வேண்டும்; முன்னேற்றங்கள் நடைமுறைப்படுத்தப்பட வேண்டும்.

புதிய வன்பொருள் மற்றும் மென்பொருள் தொழில்நுட்பத்தை ஏற்றுக்கொள்ளுதல் மற்றும் பயனர்களின் தேவையில் மாற்றம் ஆகியவையும் கவனிக்கப்படுகின்றன. இது தவிர, மீட்டெடுக்கும் பணி மற்றும் செயலாக்க மற்றும் பகுப்பாய்வுக் கருவிகள் ஆகியவற்றில் தொடர்ந்து கவனம் செலுத்தப்படுகிறது. இந்த வேலைகள் அனைத்தையும் முழு திருப்தியுடன் செயல்படுத்த கணினி தொழில்நுட்பம் மேம்படுத்தப்படவேண்டும். அவ்வாறு நிகழ்ந்தால் மென்பொருள் தொழில்நுட்பம் நம் எல்லா தேவைகளையும் பூர்த்தி செய்யும்.

4.12. தரவுத்தொகுதியின் சீராக்கத்தின் நெறிமுறை

தரவுத்தொகுதியின் சீராக்கத்தின் நெறிமுறை (method of corpus sanitation) தரவுகளை உள்ளீடு செய்த பின் அது திருத்தல்களுக்கும் மாற்றங்களுக்கும் உட்படுத்தப்பட வேண்டும் என்பதைக் கருத்தில் கொண்டதாகும். பொதுவாக நான்கு வகையான தவறுகள் நிகழலாம்:

1. எழுத்தை விட்டுவிடுதல் அல்லது நீக்குதல்
2. எழுத்தைச் சேர்த்தல் அல்லது திரும்பச் செய்தல்
3. எழுத்தைப் பதிலீடு செய்தல்
4. எழுத்தை இடம் மாற்றுதல் அல்லது இடம் பெயர்த்தல்

இவ்வெழுத்துப் பிழைகளை நீக்க நாம் விரிதரவை மனித முயற்சியால் பிரிசோதித்துத் திருத்தல் வேண்டும்.

தரவுத்தொகுதியில் உள்ள சொற்களின் எழுத்துக்கூட்டல் (spelling of words) மூல உரைகளில் பயன்படுத்தப்படும் சொற்களின் எழுத்துக்கூட்டலுக்கு ஒத்ததாக இருக்க வேண்டும் என்பதை உறுதிப்படுத்திக்கொள்ள வேண்டும். சொற்கள் மாற்றப்பட்டதா, மீண்டும் கூறப்பட்டதா அல்லது தவிர்க்கப்பட்டதா, நிறுத்தற்குறிகள் சரியாகப் பயன்படுத்தப்படுகிறதா, கோடுகள் சரியாக பராமரிக்கப்படுகிறதா, ஒவ்வொரு உரைக்கும் தனித்தனி பத்திகள் செய்யப்படுகின்றனவா என்பன்பனவற்றைச் சரிபார்க்க வேண்டும்.

பிழை திருத்தம் தவிர, தரவுத்தொகுதியின் உருவாக்கத்திற்குப் பிறகு அயல் சொற்கள், மேற்கோள்கள், கிளைமொழி வடிவங்கள் ஆகியவற்றைத் தவிர்ப்பதைச் சரிபார்க்க வேண்டும். சொந்தமாக்கப்பட்ட (nativised) அயல் சொற்கள் தரவுத்தொகுதியில் நுழைகின்றன; மற்றவை நீக்கப்படுகின்றன. கிளைமொழி மாறுபாடுகள் சரியாக உள்ளிடப்படவேண்டும். நிறுத்தற்குறிகள் மற்றும் ஒலிபெயர்ப்பு சொற்கள் என்பன அவ்வாறே உருப்படுத்தம் செய்யப்படவேண்டும்.

வழக்கமாக, இயற்கை மற்றும் சமூக அறிவியல் பற்றிய நூல்களில் கதைகள் அல்லது புனைகதை நூல்களை விட அயற் சொற்கள், சொற்றொடர்கள் மற்றும் வாக்கியங்களைக் கொண்டிருக்கும். பிற மொழிகள், கவிதைகள், பாடல்கள் மற்றும் கிளைமொழிகள் இவற்றிலிருந்து மேற்கோள்கள், கணித வெளிப்பாடுகள், இரசாயன சூத்திரங்கள், வடிவியல் வரைபடங்கள், மூல நூல்களின் அட்டவணைகள், படங்கள், புள்ளிவிவரங்கள் மற்றும் பிற குறியீட்டு பிரதிநிதித்துவங்கள் தரவுத்தொகுதியில் உள்ளிடப்படுவதில்லை.

தரவுத்தொகுதி சரியாக திருத்தப்பட்டால் அனைத்து வகையான செயலாக்க வேலைகளும் எளிதாகின்றன.

4.13. பதிப்புரிமைச் சிக்கல்

பதிப்புரிமைச் சிக்கல் (problem of copyright) நாம் எல்லாப் பதிப்புரிமை வைத்திருப்பவர்களிடமிருந்தும் (வெளியிட்டவர் மற்றும்/அல்லது படைப்பாளி, பேச்சு விஷயப்பொருள்களுக்கு எல்லாப் பேசுபவர்கள்) அனுமதி வாங்கிவைத்திருக்கவேண்டும் என்பதை உட்படுத்தும். பதிப்புரிமைச் சட்டங்கள் சிக்சலானவை. எது சரியானது, எது தவறானது, எது சட்டபூர்வமானது எது சட்டத்திற்கு மாறானது என்பதைத் தீர்மானிப்பது கடினம். பதிப்புரிமை சிக்கல்கள் நாட்டிற்கு நாடு வேறுபடும். விஷயங்கள் சொந்தப் பயன்பாட்டிற்கு என்றால் எவ்விதச் சிக்கலுமில்லை. அது வணிக நோக்கத்திற்காக நேரடியாகப் பயன்படுத்தாதது வரை எந்தவிதச் சிக்கலுமில்லை. விஷயங்களைப் பயன்படுத்தி நாம் நாம் வணிகம் செய்ய புதிய கருவிகளையும் ஒழுங்குமுறைகளையும் உருவாக்கலாம். இந்நேர்வில் பதிப்புரிமை மீறப்படவில்லை. ஆனால் நேரடியான வணிகச் செயல்பாட்டிற்குச் சட்டபூர்வமான பதிப்புரிமை உள்ளவர்களிடமிருந்து முன் அனுமதி பெற்றிருக்கவேண்டும்.

4.14. சுருக்கவுரை

பேசுத் தரவுத்தொகுதியின் உருவாக்கம் பற்றி விளக்குவது இவ்வியலின் தலையாக நோக்கம் ஆகும். இவ்வியலின் தொடக்கதில் பேசுத் தரவுத்தொகுதியின் உருவாக்கம் பற்றி ஒரு சுருக்கமான அறிமுகம் தரப்பட்டுள்ளது. இதைத் தொடர்ந்து தரவுத்தொகுதியின் வடிவ அளவு, உரைகளின் பிரதிநிதித்துவம், தாய்மொழித்தன்மை, இலக்குப் பயனர்களைத் தீர்மானித்தல், காலகட்டத்தின் தேர்வு, உரை வகையின் தேர்வு, தரவு மாதிரிப்படுத்தலின் நெறிமுறை, தரவுகளை உள்ளீடு செய்யும் நெறி முறை, வன்பொருளின் தேவை, தரவுத்தொகுதி கோப்புகளின் மேலாண்மை, தரவுத்தொகுதி துப்புரவு முறை மற்றும் பதிப்புரிமை சிக்கல் என்பன பற்றி விளக்கங்கள் தரப்பட்டுள்ளன.

இயல் 5

உரைத் தரவுத்தொகுதி ஆய்வு

5.1. அறிமுகம்

மொழியியலில் ஒரு தரவுத்தொகுதி அல்லது உரைத் தரவுத்தொகுதி (text corpus) என்பது ஒரு பெரிய மற்றும் கட்டமைக்கப்பட்ட உரைகளைக் கொண்ட ஒரு மொழி வளமாகும் (இப்போதெல்லாம் பொதுவாக மின்னணு முறையில் சேமிக்கப்பட்டு செயலாக்கப்படுகிறது). தரவுத்தொகுதி மொழியியலில், புள்ளியியல்சார் பகுப்பாய்வு மற்றும் கருதுகோள் சோதனை, நிகழ்வுகளைச் சரிபார்க்க அல்லது ஒரு குறிப்பிட்ட மொழி எல்லைக்குள் மொழியியல் விதிகளை சரிபார்க்க தரவுத்தொகுதிகள் பயன்படுத்தப்படுகின்றன.

ஒரு தரவுத்தொகுதியில் ஒரு மொழியைச் சார்ந்த உரைகள் அல்லது இருமொழிகளைச்சார்ந்த உரைகள் அல்லது பல மொழிகளைச் சார்ந்த உரைத் தரவுகள் இருக்கலாம். ஒரு மொழியைச் சார்ந்த உரைகள் உள்ள தரவுத்தொகுதி ஒருமொழியத் தரவுத்தொகுதி (monolingual corpus) என்றும் இரு மொழிகளைச் சார்ந்த உரைகள் உள்ள தரவுத்தொகுதி இருமொழியத்தரவுத்தொகுதி (bilingual corpus) என்றும் பல மொழிகளைச் சார்ந்த உரைகள் உள்ள தரவுத்தொகுதி பன்மொழியத் தரவுத்தொகுதி (multilingual corpus) என்றும் அழைக்கப்படும். மொழியியல் ஆராய்ச்சி செய்வதற்கு தரவுத்தொகுதிகளை மிகவும் பயனுள்ளதாக மாற்றுவதற்காக, அவை பெரும்பாலும் விளக்க அடையாளப்படுத்தல் (annotation) எனப்படும் செயல்முறைக்கு உட்படுத்தப்படுகின்றன. தரவுத்தொகுதியை அடையாளப்படுத்தும் ஒரு எடுத்துக்காட்டு, சொல்வகைப்பாடு அடையாளப்படுத்துதல் (ஆங்கிலத்தில் part-of-speech tagging அல்லது POS-tagging) ஆகும்; இதில் ஒவ்வொரு சொல்லின் சொல்வகைப்பாடு (வினை, பெயர்ச்சொல், பெயரடை, போன்றவை அடையாளங்களின் (tags) வடிவத்தில் தரவுத்தொகுதியில் சேர்க்கப்படுகின்றன. மற்றொரு எடுத்துக்காட்டு ஒவ்வொரு சொல்லின் சொல்லன் (லெம்மா/lemma) (அடிப்படை) வடிவத்தைக் அடையாளப்படுத்துவது ஆகும். தரவுத்தொகுதி மொழி அதைப் பயன்படுத்தும் ஆராய்ச்சியாளர்களின் செயல்பாட்டு மொழியாக இல்லாதபோது, இருமொழியாக மாற்றுவதற்கு வரிக்கு இடையிலான பொருள்விளக்கம் (இன்டர்லீனியர் க்ளாசிங்/interlinear glossing) பயன்படுத்தப்படுகிறது.

சில தரவுத்தொகுதிகள் மேலும் கட்டமைக்கப்பட்ட அளவிலான பகுப்பாய்வுகளைப் பயன்படுத்துகின்றன. குறிப்பாக, பல சிறிய தரவுத்தொகுதிகளை முழுமையாக பாகுபடுத்தலாம்.

இத்தகைய தரவுத்தொகுதிகளை வழக்கமாக கிளைவங்கி (ட்ரீபேங்க்ஸ்/Treebank) அல்லது பாகுபடுத்தப்பட்ட தரவுத்தொகுதிகள் என்று அழைக்கிறார்கள். முழுத் தரவுத்தொகுதியும் முழுமையாகவும் தொடர்ச்சியாகவும் அடையாளப்படுத்தப்படுவதை உறுதி செய்வதில் உள்ள சிரமம் காரணமாக இந்தத் தரவுத்தொகுதிகள் பொதுவாகச் சிறியதாக இருக்கும்; ஒன்று முதல் மூன்று மில்லியன் சொற்களைக் கொண்டும் இருக்கும். மொழியியல் கட்டமைக்கப்பட்ட பகுப்பாய்வின் பிற நிலைகள் சாத்தியமாகும்; இதில் உருபனியல், பொருண்மையியல் மற்றும் பயன்வழியியல் அடையாளங்கள் அடங்கும்.

உரை தரவுத்தொகுதியின் மீதான ஆய்வுகள்

பல மொழிகளில் பெரிய தரவுத்தொகுதிகள் சேகரிக்கப்பட்டபின் அவற்றை ஆய்வு உத்திகளுக்கு உள்ளாக்கும் தேவை வரும். தரவுத்தொகுதியிலிருந்து மொழித் தரவை அணுகுவதற்கும் தேவையான தகவல்களை மீளப்பெறுவதற்கும் ஆய்வாளர்கள் ஒழுங்குமுறைகளையும் உபாயங்களையும் உருவாக்கியுள்ளனர். இவற்றை ஆய்வுக்கருவிகள் எனலாம். இவ்வாய்வுக் கருவிகள் மொழி ஆய்வுக்கும் மொழித் தொழில் நுட்ப முன்னேற்றத்திற்கும் பயனுள்ளதாக அமைந்துள்ளன. பலவிதமான தரவுத்தொகுதி ஆய்வு உத்திகள் இருக்கின்றன. எடுத்துக்காட்டாக,

1. புள்ளியியல் ஆய்வி (Statistical Analyzer)
2. தொடரடைவு ஆய்வி (Concordencer)
3. சொல் சேர்ந்து வருகை ஆய்வி (Lexical Collocator)
4. முக்கியசொல் காணும் ஆய்வி (Key-word Finder)
5. சொல்லன் ஆய்வி (Lemmatizer)
6. உருபனியல் ஆய்வியும் உருவாக்கியும் (Morphological analyzer and Generator)
7. சொல் ஆய்வி (Word Processer)
8. சொல் வகைப்பாடு அடையாளப்படுத்தி (Parts-of-Speech Tagger)
9. தரவுத்தொகுதி இடச்சுட்டி/அடையாளப்படுத்தி (Corpus Anotator)
10. பகுப்பான் (Parser)

ஆங்கிலம், பிரஞ்சு, ஜெர்மன் போன்ற மொழிகளுக்கு தரவுத்தொகுதியை ஆயப் பல மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. ஆனால் இந்தி, தமிழ் போன்ற இந்திய மொழிகளுக்கு

அத்தகைய மென்பொருள்கள் சில உருவாக்கப்பட்டுள்ளன. இதற்கான முயற்சிகள் எடுக்கப்பட்டு வருகின்றன.

5.2. நிகழ்வெண் ஆய்வு

ஒரு அதிர்வெண் விநியோகம் என்பது தரவை ஒழுங்கமைக்கவும் சுருக்கமாகவும் பயன்படுத்தப்படும் ஒரு கணக்கெடுப்பு தரவு தொகுப்பின் அட்டவணை பிரதிநிதித்துவம் ஆகும். குறிப்பாக, இது ஒரு தரவுத் தொகுப்பில் ஒரு மாறி எடுக்கும் தரமான அல்லது அளவு மதிப்புகளின் பட்டியல் மற்றும் ஒவ்வொரு மதிப்பும் நிகழும் எத்தனை முறை (அதிர்வெண்கள்).

அதிர்வெண் விநியோகம் என்பது புள்ளிவிவர பகுப்பாய்வு முறைகளின் அடிப்படை கட்டுமானத் தொகுதி மற்றும் கணக்கெடுப்பு தரவை பகுப்பாய்வு செய்வதற்கான முதல் படியாகும். இது ஆய்வாளர்களுக்கு (அ) கணக்கெடுப்பு தரவை அட்டவணை வடிவத்தில் ஒழுங்கமைக்கவும் சுருக்கமாகவும் உதவுகிறது, (ஆ) தரவை விளக்குகிறது, மற்றும் (இ) கணக்கெடுப்பு தரவு தொகுப்பில் வெளிநாட்டவர்களை (தீவிர மதிப்புகள்) கண்டறிய உதவுகிறது.

தரவுத்தொகுதி மொழியியலில் புள்ளிவிவரம்

தரவுத்தொகுதிகள் என்பது மொழியியலாளர்களுக்கான அளவுத் தரவுகளின் இணையற்ற ஆதாரமாகும். எனவே தரவுத்தொகுதி மொழியியலாளர்கள் பெரும்பாலும் புள்ளியியல் விவரங்களின் மூலம் தங்கள் அளவுக் கண்டுபிடிப்புகளை சோதித்துப் பார்க்கிறார்கள் அல்லது சுருக்கமாகக் கூறுகிறார்கள். மொழியியலின் வேறு சில பகுதிகள் புள்ளியியல்சார் கருத்துகள் மற்றும் சோதனைகளைஅடிக்கடி வேண்டுகின்றன; உளவியல் மொழி சோதனைகள், இலக்கண உயர்வு சோதனைகள் மற்றும் கணக்கெடுப்பு அடிப்படையிலான விசாரணைகள்; எடுத்துக்காட்டாக, இவை அனைத்தும் பொதுவாக ஒருவித புள்ளியியல்சார் சோதனைகளை உள்ளடக்குகின்றன. இருப்பினும், தரவுத்தொகுதி பகுப்பாய்வில் நிகழ்வெண் தரவு தொடர்ந்து தயாரிக்கப்படுகிறது; பெரும்பாலான தரவுத்தொகுதி அடிப்படையிலான ஆய்வுகள் சில வகையான புள்ளியியல் விவரப் பகுப்பாய்வுகளை மேற்கொள்கின்றன; இது ஒப்பீட்டளவில் அடிப்படை மற்றும் விளக்கமாக இருந்தாலும் கூட இது செய்யப்படுகின்றது; எ.கா. தரவை ஒருவிதத்தில் விவரிக்க சதவீதங்களைப் பயன்படுத்துதல்.

விளக்கமான புள்ளியியல் தகவல்கள்

தரவுத்தொகுதி மொழியியலில் பெரும்பாலான ஆய்வுகள் வேறு எதுவும் இல்லை என்றால் அடிப்படை விளக்கப் புள்ளியியலைப் (descriptive statistics) பயன்படுத்துகின்றன. விளக்கப்

புள்ளியியல் முக்கியத்துவத்தைச் பரிசோதிக்க முற்படாத புள்ளியியல் ஆகும். மாறாக அவை தரவை ஏதோ ஒரு வகையில் விவரிக்கின்றன. மிக அடிப்படையான புள்ளியியல் அளவீடு அதிர்வெண் எண்ணிக்கை ஆகும்; இது ஒரு தரவுத்தொகுதியில் நிகழும் ஏதோ சில நிகழ்வுகளின் எண்ணிக்கையை எளிமையாகக் கணக்கிடுதல். எடுத்துக்காட்டாக, பி.என்.சியின் எழுத்துத் தரவுப் பகுதியில் லான்காஸ்டர்/Lancaster என்ற வார்த்தையின் 1,103 எடுத்துக்காட்டுகள் உள்ளன. முழு தரவுத்தொகுதியின் சதவீதமாக இதை நாம் வெளிப்படுத்தலாம்; பி.என்.சியின் எழுதுத் தரவுப் பிரிவில் இயங்கும் உரையின் 87,903,571 சொற்கள் உள்ளன; அதாவது லான்காஸ்டர் என்ற சொல் தரவுத்தொகுதியின் எழுத்துத் தரவுப் பிரிவில் உள்ள மொத்தத் தரவுகளில் 0.013% ஐ குறிக்கிறது. 1,103 எண்ணிக்கையை சூழலில் பார்க்கும் மற்றொரு வழி, எழுதப்பட்ட தரவுத்தொகுதியின் முழுமையுடன் ஒப்பிடும்போது அதைப் புரிந்துகொள்ள முயற்சிப்பது. சில நேரங்களில், இங்குள்ளதைப் போலவே, சதவிகிதம் வார்த்தையின் பயன்பாட்டின் அதிர்வெண்ணை அர்த்தமுள்ளதாக தெரிவிக்காமல் போகலாம்; எனவே அதற்குப் பதிலாக ஒரு இயல்பாக்கப்பட்ட அதிர்வெண்ணை (அல்லது ஒப்பீட்டு அளவிலான அதிர்வெண்) உருவாக்கலாம்; இது பின்வரும் கேள்விக்கு பதிலளிக்கிறது: ஒரு ஓடும் உரையில் ஒரு சொல் ஒரு குறிப்பிட்ட எண்ணிக்கையிலான சொல்லளவில் எத்தனை தடவைப் பயன்படுத்தப்படுகின்றது என்று கணக்கிடலாம். இயல்பாக்கப்பட்ட அதிர்வெண்கள் பொதுவாக ஆயிரம் சொற்களுக்குள் அல்லது ஒரு மில்லியன் சொற்களுக்குள் எனக் கணக்கிடப்படுகின்றன.

வகை-டோக்கன் விகிதம் (type-token ratio) எனப்படும் ஒரு சிறப்பு வகை விகிதம் மற்றொரு அடிப்படைத் தரவுத்தொகுதிப் புள்ளிவிவரமாகும். ஒரு டோக்கன் என்பது ஒரு உரையில் ஒரு குறிப்பிட்ட சொல் வடிவத்தின் எந்தவொரு நிகழ்வாகும். உரையில் உள்ள டோக்கன்களின் எண்ணிக்கையை டோக்கன்களின் வகைகளின் எண்ணிக்கையுடன் ஒப்பிட்டுப் பார்க்கும்போது - ஒவ்வொரு வகையும் ஒரு குறிப்பிட்ட, தனித்துவமான சொல் வடிவமாக இருக்கும் - உரையில் எவ்வளவு பெரிய அளவிலான சொற்றொகைப் பயன்படுத்தப்படுகிறது என்பதைக் கூறலாம். ஒரு தரவுத்தொகுதியில் உள்ள வகைகளின் எண்ணிக்கையை டோக்கன்களின் எண்ணிக்கையால் வகுப்பதன் மூலம் வகை-டோக்கன் விகிதத்தை நாங்கள் தீர்மானிக்கிறோம். இதன் விளைவாக சில நேரங்களில் வகை-டோக்கன் விகிதத்தை ஒரு சதவீதமாக வெளிப்படுத்த 100ஆல் பெருக்கப்படுகிறது. இது தரவுத்தொகுதிகளுக்கு இடையிலான சொற்றொகுதி (vocabulary) மாறுபாட்டை அளவிட அனுமதிக்கிறது - விளைவு 1க்கு (அல்லது ஒரு

சதவீதமாக இருந்தால் 100க்கு) நெருக்கமாக இருக்கிறதைப் பொறுத்து அதிகச் சொற்றொகுதி மாறுபாடு இருக்கும்; மேலும் விளைவு 1இலிருந்து கூடுதல் தூரத்தில் இருக்கிறதைப் பொறுத்து, குறைந்த சொற்றொகுதி மாறுபாடு இருக்கும். தரவுத்தொகுதியின் அளவு அதன் வகை-டோக்கன் விகிதத்தைப் பாதிக்கும் என்பதால், ஒத்த அளவிலான தரவுத்தொகுதிகளை மட்டுமே இந்த வழியில் ஒப்பிட முடியும். அளவு வேறுபடும் தரவுத்தொகுதிகளுக்கு, நடைமுறையின் இயல்பாக்குதல் பதிப்பு (தரப்படுத்தப்பட்ட வகை-டோக்கன் விகிதம் (standardised type-token ratio அல்லது STTR/எஸ்.டி.டி.ஆர்) அதற்குப் பதிலாகப் பயன்படுத்தப்படுகிறது.

பலவிதமான புள்ளியியல் அணுகுமுறைகள்

மொழியியல் நீண்ட காலமாகவே புள்ளியியலுடனும் கணக்கியலுடனும் தொடர்புடையது. கணித மொழியியல், கணினி மொழியியல், விரிதரவு மொழியியல், பயன்பாட்டு மொழியியல், குற்றாய்வு மொழியியல், நடை அளவியல் (stylometrics) போன்றவை இயற்கை மொழி விரிதரவிலிருந்து பெறப்படும் வேறுபட்ட புள்ளியியல் மற்றும் அளவியல் தகவல்களை வேண்டுகிறது. மொழியின் வேறுபட்ட பண்புக்கூறுகளைக் குறித்த புள்ளியியல் தகவலின் போதுமாக அறிவு இல்லாமல் நாம் மொழித் தரவைக் கையாளுவதிலும் கண்டறிவதிலும் தவறுகள் செய்வோம். தரவுத்தொகுதியைப் பண்புசார் ஆய்வுகளுக்கும் (qualitative analysis) அளவுசார் ஆய்வுகளுக்கும் (quantitative analysis) உட்படுத்தலாம். பண்பு சார் ஆய்வு முந்தைய அமைப்பு மொழியியல் ஆய்வை குறிக்கும். மாறாக அளவுசார் ஆய்வு தரவுத்தொகுதியின் மேல் அளவு அடிப்படையிலான ஆய்வின் அடிப்படையில் கிடைக்கும் முழுமையான மற்றும் விளக்கமான வர்ணனையைத் தருவதை இலக்காகக் கொள்கிறது. இவ்விலக்கை அடையப் பலவிதமான புள்ளியியல் அணுகுமுறை இருக்கின்றன. அவையான,

1. வர்ணனைப் புள்ளியியல் அணுகுமுறை (descriptive statistical approach): எளிமையான விளக்கப் புள்ளியியல் அணுகுமுறை உற்றுநோக்கப்பட்ட தரவின் மிக முக்கியமான பண்புகளைச் சுருக்கமாகக் கூறுகிறது.

2. அறிதல்/அனுமானப் புள்ளியியல் அணுகுமுறை (inferential statistical approach): அறிதல்/அனுமானப் புள்ளியியல் அணுகுமுறை கேள்விகளுக்குப் பதிலளிக்க அல்லது கருதுகோளை உருவாக்க விளக்கப் புள்ளியியல் அணுகுமுறையிலிருந்து தகவல்களைப் பயன்படுத்துகிறது.

3. மதிப்பீட்டுப் புள்ளியியல் அணுகுமுறை (evaluative statistical approach): மதிப்பீட்டுப் புள்ளியியல் அணுகுமுறை தரவுகளில் உள்ள ஆதாரங்களால் கருதுகோள் ஆதரிக்கப்படுகிறதா என்பதையும், கணித மாதிரி அல்லது தரவின் கோட்பாடு அடிப்படையிலான விநியோகம் யதார்த்தத்துடன் எவ்வாறு தொடர்புடையது என்பதையும் பரிசோதிக்க உதவுகிறது (Oakes 1998: 1).

4. பன்முக புள்ளியியல் அணுகுமுறை (multivariate statistical techniques): ஒப்பீடுகளைச் செய்ய, தரவுத்தொகுதிலிருந்து பெறப்பட்ட மூல அதிர்வெண் தரவுகளிலிருந்து மறைக்கப்பட்ட வடிவங்களை பிரித்தெடுக்க பன்முக புள்ளியியல் நுட்பங்களை (எ.கா., காரணிப் பகுப்பாய்வு Factor Analysis, பல பரிமாண அளவிடுதல் (Multidimensional Scaling), கிளஸ்டர் பகுப்பாய்வு (Cluster Analysis), லாக்-நேரியல் மாதிரிகள் போன்றவை) பயன்படுத்துகிறோம்.

5.3. சொல் வரிசைப்படுத்துதல் (Word Sorting)

அளவுசார் தரவில் நேரடியான அணுகுமுறை எண்ணிக்கை அடிப்படையிலான வரிசைப்படுத்தல் (numerical sorting) ஆகும். எளிய நிகழ்வெண் எண்ணிக்கையிலிருந்து பெறப்படும் தகவலை நெடுங்கணக்கு வரிசையிலும் (alphabetical order) எண்ணிக்கை வரிசையிலும் (numerical order) வரிசைப்படுத்தலாம். இவற்றை மீண்டும் ஏறு வரிசையிலும் இறங்குவரிசையிலும் தரலாம். சொற்களின் நிகழ்வெண்பட்டியல் உரைகளுக்கு ஒரு குழும தடயங்களாகும். உரையை ஆயும் எவரும் ஒவ்வொரு சொல்லும் உரையில் எத்தனை முறை வருகிறது என்பதை அறிவதில் ஆர்வம் காட்டுவர். இப்பட்டியலைப் பரிசோதித்து நாம் ஒரு குறிப்பிட்ட உரையின் அமைப்பைப் பற்றி புரிந்துக்கொள்ளலாம். சொல்லில் மட்டுமின்றி எழுத்துக்கள், மரபுத்தொடர்கள், எச்சத்தொடர்கள் மற்றும் வாக்கியங்கள் இவற்றின் புள்ளி விவரங்களை நாம் தரவுத்தொகுதியிலிருந்து அறிந்துக் கொள்ளலாம். இது ஒரு மொழியின் மொழி அலகுகளையும் பண்புக்கூறுகளையும் புரிந்து கொள்ள பெரிதும் உதவும். எடுத்துக்காட்டாக, நடையியல் ஆய்வுக்கும் இத்தகைய புள்ளியியல் தகவல்கள் பெரிதும் கைகொடுக்கும்.

எளிய பொது நோக்கீட்டுக்கு அகர வரிசைப்படி வரிசைப்படுத்தப்பட பட்டியல் (alphabetical sorted list) பயன்படுத்தப்படுகிறது. அகர வரிசைப்படி ஒரு அதிர்வெண் பட்டியல் இரண்டாம் நிலை பங்கை வகிக்கிறது; ஏனெனில் ஒரு குறிப்பிட்ட பொருளின் அதிர்வெண்ணைச் சரிபார்க்க வேண்டிய அவசியம் இருக்கும்போது மட்டுமே இது பயன்படுத்தப்படுகிறது. இருப்பினும், இது ஒரு ஆய்வின் பொருளாக பயனுள்ளதாக இருக்கும்; ஏனெனில் இது

பெரும்பாலும் சோதிக்கப்பட வேண்டிய கருதுகோள்களை உருவாக்குவதற்கும்; க்ஜெல்மர்-க்கு (Kjellmer 1984) முன் செய்யப்பட்ட அனுமானங்களைச் சரிபார்க்கவும் உதவியாக இருக்கும்.

இந்தியத் தரவுத்தொகுதியில் அதிர்வெண் எண்ணிக்கையைத் தொடங்குவதற்கு முன், தரவுத்தொகுதியில் பயன்படுத்தப்படும் எழுத்துக்கள், சொற்கள், மரபுத்தொடர்கள், சொற்றொடர்கள், உட்பிரிவுகள் மற்றும் வாக்கியங்களைக் கையாள்வது குறித்து நாம் முடிவுகளை எடுக்க வேண்டும். இவை பல்வேறு மொழியியல் பண்புகள் குறித்த தவறான உற்றுநோக்கல் மற்றும் தவறான உய்த்தறிதல் ஆகியவற்றிலிருந்து நம்மைக் காப்பாற்றும்.

5.4. தொடரடைவு ஆய்வு

தொடரடைவுகள் (concordances) சொற்கள் சேர்ந்துவருதலை அடையாளம் காண உதவும். சொற்கள் ஒன்றுக்கொன்று எவ்வளவு நெருக்கமாக வருகின்றன என்பதை அவை காட்டுகின்றன. தொடரடைவுச் செயல்முறை தரவுத்தொகுதியில் பயன்படுத்தப்படும் சொற்களுக்கு ஒரு அட்டவணையை உருவாக்குகிறது. இது சொற்கள் ஒவ்வொன்றும் அதன் சொந்த உரைச் சூழலில் வரும் நிகழ்வுகளின் தொகுப்பாகும். ஒவ்வொரு சொல்லும் உரைகளில் அவற்றின் ஒவ்வொரு நிகழ்வின் இடத்தையும் குறிக்கும் வகையில் அட்டவணைப்படுத்தப் படுகின்றன. இது இன்றியமையாதது, ஏனெனில் இது உரைகளில் பல முக்கியமான மொழி வடிவங்களை அணு உதவும். இது உள்ளுணர்வு வழியாக அணுக முடியாத தகவல்களை வழங்குகிறது.

பி.என்.சி.வெப் (BNCWeb) போன்ற நவீன கணினி அடிப்படையிலான தரவுத்தொகுதிகளைக் கொண்டு, அவற்றை எளிதில் கட்டமைக்க முடியும். ஒரு சொல் இலக்கு சொல்லின் வலது பக்கத்தில் அல்லது அதன் இடது பக்கத்தில் அடிக்கடி வருகிறதா என்பதை பயனர் பார்க்கலாம். இது பேசும் அல்லது எழுதப்பட்ட உரைகளில் சொற்களின் நிலைகளை ஆய்வு செய்ய உதவுகிறது மற்றும் சொல் வகுப்புகளின் நிலைகளை விவரிக்கவும் உதவுகிறது; எடுத்துக்காட்டாக இலக்கணச் சொற்கள் வாக்கியங்களின் தொடக்கத்தில் அல்லது முடிவில் அடிக்கடி வருகிறதா என்று அறிய இயலும். ஒரு சொல்லின் சூழலை ஆராயவும் இது உதவக்கூடும்.

ஹோய் மற்றும் பலர் (Hoey et al) தரவுத்தொகுதியில் உள்ள தொடரடைவுக் கோடுகள் டி சாஸூர் (de Saussure) லாங்கு (langue) என்று அழைக்கப்படுவதை அடையாளம் காண

உதவுகின்றன என்று கூறுகின்றனர். தொடரடைவுகளின் உதவியுடன் ஒரு மொழியின் முறையான வடிவங்கள் அல்லது போக்குகள் காட்டப்படுகின்றன (Hoey et al. 2007: 154).

தொடரடைவு ஆய்வுகளின் ஒரு முக்கிய முடிவு, சொல் அலகுகளின் நீளத்தைக் கணிக்கத்தக்கதாக மாற்றியது. அதற்கு ஒரு எடுத்துக்காட்டு ஹோய் மற்றும் பலர் வழங்கியுள்ளனர் (2007: 154-155). *endure* என்ற வார்த்தையை ஆராய அவர்கள் ஒரு தரவுத்தொகுதியைப் பயன்படுத்தினர்; பின்வரும் உரைகளை உற்றுநோக்கினர்.

(1) that smokers will have to endure 12-hour flights by becoming

(2) remember having had to endure a certain amount of misery

(3) the animals often have to endure hours trapped in the midst

மேற்கண்ட உரைகள் மூலம் *endure* என்பது பலத்தால் வந்த ஒன்றை அல்லது மக்கள் விரும்பத்தகாத ஒன்றை எதிர்கொள்ள வேண்டியிருக்கும் என்பதை விவரிக்க பயன்படுத்தபடுவதை அவர்கள் கண்டுகொண்டனர். *endure* என்பதன் இடது மற்றும் வலது பக்கத்தில் உள்ள காணப்படும் தொரடைவுகள் இதைத் தெளிவாகிறது. இந்த எடுத்துக்காட்டு போன்று வேறு சொற்களுக்கும் உரைகள் பயன்படுத்தப்பட்டால், ஒரு மொழியில் உள்ள சில கட்டமைப்புகள் தொடரடைவுகளின் உதவியுடன் அறியப்படலாம்

சொற்களுக்கிடையே உள்ள தொரடைவை அறிய உதவும் கருவி தொரடைவு ஆய்வி (concordencer) எனப்படும். தொரடைவு ஆய்வி தரவுத்தொகுதியில் வரும் சொற்களின் அகரவரிசை அட்டவணையை உருவாக்கப் பயன்படுகின்றது. இது ஒவ்வொரு சொற்களின் உரைசூழலுடன் கூடிய அவற்றின் நிகழ்வுகளின் சேகரிப்பாகும். ஒவ்வொரு சொல்லும் உரையில் அதன் வருகையிடம் குறிப்பிடப்பட்டு அட்டவணைப்படுத்தப்படும். உரையிலுள்ள பல முக்கியமான மொழி அமைப்பொழுங்குகளை அணுக உதவுவதால் இது மிக முக்கியமானதாகும். இது உள்ளுணர்வால் அறியவியாலாத தகல்களைத் தருகின்றது.

தரவுத்தொகுதியைப் பகுப்பாய்வு செய்வதற்குச் சில தொடரடைவு மென்பொருள்கள் உள்ளன; எடுத்துக்காட்டாக, வரிசைப்படுத்துதல் மற்றும் அதிர்வெண்ணிற்கான மோனோகாங்க் (MonoConc), இணை உரைகளின் செயலாக்கத்திற்கான பராக்கோங்க் (ParaConc), வரிசைப்படுத்துதல் மற்றும் அதிர்வெண் எண்ணுவதற்கான கான்க் (Conc), செயலாக்கத்திற்கான இலவச உரை (Free Text), வரிசைப்படுத்துதல் போன்றவை).

தொடரடைவு ஆய்வி பெரும்பாலும் அகராதிசார் படைப்புகளுக்குப் பயன்படுத்தப்படுகிறது. ஒற்றை மற்றும் பல சொல் சரங்கள், சொற்கள், சொற்றொடர்கள், மரபுத்தொடர்கள் போன்றவற்றைத் தேட இதைப் பயன்படுத்துகிறோம். இது சொல்சார், பொருண்மையியல்சார், தொடரியல்சார் அமைப்பொழுங்குகள் (patterns), உரை அமைப்பொழுங்குகள் (text patterns), இன ஆய்வுகள் (genre studies), இலக்கிய உரைகள் போன்றவற்றைப் படிக்கவும் பயன்படுகிறது (Barlow 1996). மேலும் இது ஒருபொருள்பன்மொழியத் தன்மையுள்ள மற்றும் மொழியில் பல செயல்பாடுகளைக் கொண்டுள்ளன சொற்கள் மற்றும் உருபன்களை ஆய்வதற்கான ஒரு சிறந்த கருவியாகும்.

5.5. உடன்வருகை ஆய்வு

மொழியியலில் உடன்வருகை (co-occurrence) என்பது ஒரு உரைத் தரவுத்தொகுதியில் ஒரு குறிப்பிட்ட வரியில் இரண்டு சொற்கள் (தற்செயல் அல்லது ஒத்திசைவு என்றும் அழைக்கப்படுகிறது) சேர்ந்துவரும் வாய்ப்புக்கான நிகழ்வெண் ஆகும். இந்த மொழியியல் அர்த்தத்தில் உடன்வருகையைச் சொற்பொருள் அருகாமையின் குறிகாட்டியாக அல்லது ஒரு அடையாள வெளிப்பாடாக விளக்கலாம். தரவுத்தொகுதி மொழியியல் மற்றும் அதன் புள்ளியியல் பகுப்பாய்வுகள் ஒரு மொழியில் உடன்வருகைகளின் வடிவங்களை வெளிப்படுத்துகின்றன மற்றும் அச்சொற்களின் பொதுவான சேர்ந்துவருகைகளை அறிய உதவுகின்றன. மொழியியல் கூறுகள் ஒருபோதும் ஒன்றாக நிகழாதபோது ஒரு உடன்வருகைக் கட்டுப்பாடு (co-occurrence restriction) அடையாளம் காணப்படுகிறது. இந்தக் கட்டுப்பாடுகளின் பகுப்பாய்வு ஒரு மொழியின் கட்டமைப்பு மற்றும் வளர்ச்சி பற்றிய கண்டுபிடிப்புகளுக்கு வழிவகுக்கும். உடன்வருகைகள் அதிக பரிமாணங்களில் சொல் எண்ணிக்கையின் நீட்டிப்பைக் காணலாம். தொடர்பு அல்லது பரஸ்பர தகவல் போன்ற நடவடிக்கைகளைப் பயன்படுத்தி இணை நிகழ்வை அளவுரீதியாக விவரிக்க முடியும்.

தரவுத்தொகுதி மொழியியலில், உடன்வருகை என்பது தற்செயலாக எதிர்பார்க்கப்படுவதை விட அடிக்கடி நிகழும் சொற்கள் அல்லது சொற்களின் தொடர்கள். தொடர்நிலையியலில் (phraseology), உடன்வருகை என்பது ஒரு துணை வகைச் சொற்றொடர். மைக்கேல் ஹாலிடே (Michael Halliday) முன்வைத்தபடி, ஒரு தொடர்நிலையியலில்சார் உடன்வருகைக்கான எடுத்துக்காட்டு, *strong tea* என்ற வெளிப்பாடு. அதே அர்த்தத்தை தோராயமாக சமமான *powerful tea* மூலம் தெரிவிக்க முடியும் என்றாலும், இந்த வெளிப்பாடு

ஆங்கிலம் பேசுபவர்களால் அளவுக்கு அதிகமாகவும் மோசமாகவும் கருதப்படுகிறது. மறுதலையாக, தொழில்நுட்பத்தில் தொடர்புடைய வெளிப்பாடு, *powerful computer* என்பது *strong computer* என்பதை விட விரும்பப்படுகிறது. தொடர்நிலையியலில்சார் உடன்வருகை மரபுத்தொடர்களுடன் குழப்பமடையக்கூடாது; அங்கு ஒரு மரபுத்தொடரின் அர்த்தம் அதன் மரபிலிருந்து வேறொன்றிற்கான நிலைப்பாடாக பெறப்படுகிறது, அதே நேரத்தில் உடன்வருகை என்பது வெறும் பிரபலமான கூட்டமைவாகும். ஒரு மொழியைத் திறம்பட பயன்படுத்துவதற்கான திறன், உடன்வருகை எனப்படும் மொழியின் தனித்துவமான அம்சத்தைப் பற்றிய விழிப்புணர்வை உள்ளடக்கியது. பேச்சு அல்லது எழுத்தில் இரண்டு அல்லது அதற்கு மேற்பட்ட சொற்கள் ஒன்றிணைக்கும் மொழியின் நடத்தைதான் உடன்வருகை.

சுமார் ஆறு முக்கிய வகை மோதல்கள் உள்ளன: பெயரடை + பெயர்ச்சொல், பெயர்ச்சொல் + பெயர்ச்சொல் (கூட்டு பெயர்ச்சொற்கள் போன்றவை), வினை + பெயர்ச்சொல், வினையடை + பெயரடை, வினைச்சொற்கள் + முன்னுருபு தொடர் (தொடர் வினைச்சொற்கள்) மற்றும் வினை + வினையுரிச்சொல்.

உடன்வருகைப் பிரித்தெடுத்தல் (collocation extraction) என்பது ஒரு கணினியல்சார் நுட்பமாகும்; இது தரவுத் தோண்டலை (data mining) ஒத்த பல்வேறு கணினி மொழியியல் கூறுகளைப் பயன்படுத்தி ஒரு ஆவணத்தில் அல்லது தரவுத்தொகுதியில் உடன்வருகைகளைக் கண்டறிகிறது.

உடன்வருகை மீண்டும் மீண்டும் நிகழும் சூழல் சார்ந்த பயன்பாட்டின் மூலம் நிறுவப்படுகின்ற பகுதி அல்லது முழுமையான நிலையான வெளிப்பாடுகள் ஆகும். 'குழந்தை உள்ளம்', 'கள்ள நோட்டு', 'கூட்டுக் குடும்பம்' மற்றும் 'கல் மனசு' போன்ற சொற்கள் உடன்வருகை செய்யும் சொற்களின் இணைகளுக்கு எடுத்துக்காட்டுகள் ஆகும்.

உடன்வருகை ஒரு தொடரியல் உறவில் (வினை-செய்ப்படுபொருள்: 'செய்' மற்றும் 'முடிவு' போன்றவை), சொல்சார் உறவு (எதிர்ச்சொல் போன்றவை) அல்லது அவை மொழியியல் ரீதியாக வரையறுக்கப்பட்ட உறவில் இருக்க இயலாது. ஒரு மொழியின் திறமையான பயன்பாட்டிற்கு உடன்வருகை அறிவு மிக முக்கியமானது: உடன்வருகை விருப்பத்தேர்வுகள் மீறப்பட்டால் இலக்கணப்படி சரியான வாக்கியம் ஏற்றுக்கொள்ள இயலாததாக இருக்கும். இது மொழி கற்பிப்பதற்கான ஒரு சுவாரஸ்யமான பகுதியாகும்.

தரவுத்தொகுதி மொழியியலாளர்கள் முக்கியச் சூழலில் சொல் (Key Word in Context (KWIC)) என்பதைக் குறிப்பிடுகின்றனர் மற்றும் அவற்றைச் சுற்றியுள்ள சொற்களை உடனடியாக அடையாளம் காண்கின்றனர். இது சொற்களைப் பயன்படுத்தும் முறையைப் பற்றிய ஒரு கருத்தை அளிக்கிறது.

உடன்வருகையின் செயலாக்கம் பல அளகைளை உள்ளடக்கியது; அவற்றில் மிக முக்கியமானது தொடர்பின் அளவீடு ஆகும்; இது உடன்வருகை முற்றிலும் தற்செயலானதா அல்லது புள்ளிவிவர ரீதியாக முக்கியத்துவம் வாய்ந்ததா என்பதை மதிப்பீடு செய்கிறது. மொழியின் சீரற்ற தன்மை காரணமாக, பெரும்பாலான உடன்வருகைகள் குறிப்பிடத்தக்கவை என வகைப்படுத்தப்படுகின்றன; மேலும் முடிவுகளை மதிப்பீடு செய்ய தொடர்பு மதிப்பெண்கள் பயன்படுத்தப்படுகின்றன. பொதுவாகப் பயன்படுத்தப்படும் தொடர்பு நடவடிக்கைகளில் பரஸ்பர தகவல்கள் (mutual information), டி மதிப்பெண்கள் (t scores) மற்றும் லாக்-வாய்ப்பு (log-likelihood) ஆகியவை அடங்கும்.

சொல்சார் உடன்வருகை ஆய்வி (lexical collocator) சொற்களின் உடன்வருகையை ஆயும். சொற்களின் உடன்வருகை உரைகளில் சொற்களின் பங்களிப்பையும் இடத்தையும் புரிந்து கொள்ளப் பெரிதும் உதவும். இவ்வாய்வி சொற்களில் எந்த இணைகளுக்கிடையில் கூடுதல் உடன்வருகை உறவு இருக்கின்றது என்பதை அறிய உதவும். மேலும் இரண்டு சொற்கள் சேர்ந்து வரும் நிகழ்வுத்தகைமையைத் தற்செயலான விளைவின் அடிப்டையிலான நிகழ்வுத் தகைமையுடன் ஒப்பிடும். ஒவ்வொரு சொல் இணைக்கும் ஒரு மதிப்பெண் தரப்படும்; கூடுதல் மதிப்பெண் கூடுதல் உடன்வருகைக் காட்டும். இது அகராதியிலும் தொழில்நுட்ப மொழிபெயர்ப்பிலும் பயன்படுத்த தரவுத்தொகுதியிலிருந்து பல்சொல் அலகுகளைப் பிரித்தெடுக்க உதவும். இயந்திர மொழி பெயர்ப்புக்கும் பொருண்மை வேறுபாட்டிற்கும் இத்தகைய தகவல்கள் கை கொடுக்கும். இது அர்த்த வேறுபாடுகளை அடையாளம் காண ஒற்றுமையான சொற்களைக் குழும உதவுகின்றது. எடுத்துக்காட்டாக *படி* என்பதன் நான்கு அர்த்தங்களை அதன் உடன்வரும் சொற்கள் மூலம் வேறுபடுத்தி அறியலாம்.

மாடிப்–படி – step

வாசற்-படி – step

புத்தகம் படி – read

சொன்ன-படி – step

ஐந்து லிட்டர் படி – measure

இது போல் மாலை என்ற சொல்லின் பல்பொருள் ஒருமொழியத்தை (polysemy) உடன்வருகையால் வேறுபடுத்தலாம்: *பூ மாலை, ரோஜா மாலை, மாலை நேரம், மாலை வேளை*. உடன்வருகை ஒரே பொருண்மையைக் கொண்ட சொல்களுக்கு இடையேயுள்ள பயன்பாட்டு வேறுபாடுகளையும் வேறுபடுத்த உதவும்.

5.6. சொல்சார் சொல்லடிவகைப்பாடு

சொற்களில் சொல்லடிவகைப்பாடு (lexical collocation) முறை உரைகளில் சொற்களின் பங்கு மற்றும் நிலையைப் புரிந்து கொள்ள உதவுகிறது. மரபுசார் மொழியியல் விளக்கங்கள் மற்றும் கருதுகோள்கள் தரவுத்தொகுதியிலிருந்து சொல்லடிவகைப்பாடு மூலம் திரட்டப்பட்ட புதிய ஆதாரங்களால் சவால் செய்யப்படுகின்றன. அகராதி பற்றிய சமீபத்திய செயல்பாடு (Ooi 1998) பல பொதுவான சொற்களுக்கு, அடிக்கடி வரும் பொருள் நம் மனதில் முதலில் வருவதல்ல, அது அகராதிகளில் நடைபெறுகிறது என்பதைக் காட்டுகிறது. எந்த ஜோடி சொற்களுக்கு இடையில் கணிசமான சேர்ந்துவருகை உறவு உள்ளது என்பதைத் தீர்மானிக்க இது உதவுகிறது.

இது இரண்டு சொற்களின் நிகழ்தகவுகளை ஒரு நிகழ்வாக ஒப்பிடுகிறது, அவை வெறுமனே வாய்ப்பின் விளைவாகும். ஒவ்வொரு ஜோடி சொற்களுக்கும், ஒரு மதிப்பெண் வழங்கப்படுகிறது - அதிக மதிப்பெண் அதிக சொல்லடிவகைப்பாடு ஆகும். இது அகராதியியல் மற்றும் தொழில்நுட்ப மொழிபெயர்ப்பில் பயன்படுத்த தரவுத்தொகுதியிலிருந்து பல்சொல் அலகுகளைப் (multiword units) பிரித்தெடுக்க உதவுகிறது. இது மேலும் அர்த்த மாறுபாடுகளை அடையாளம் காண ஒப்புருமொழிகளை ஒன்றிணைக்க உதவுகிறது (எ.கா. river bank என்பதில் bank என்பது நிலப்பரப்புசார் பயன்பாடு, ஆனால் investment in bank என்பதில் bank என்பது நிதிசார் பயன்பாடு)

இது அவை அர்த்தத்தில் ஒத்தவை சொற்களுக்கு இடையிலான பயன்பாட்டில் உள்ள வேறுபாடுகளைப் பாகுபடுத்த உதவுகிறது. எடுத்துக்காட்டாக *strong* என்பது *motherly, showings, believer, currents, supporter, odour* போன்றவற்றுடன் சொல்லடிவகைப்பாடு செய்யும்; மாறாக *powerful* என்பது *tool, minority, neighbour, symbol, figure, weapon, post* போன்றவற்றுடன் சொல்லடிவகைப்பாடு செய்யும் (Biber at al. 1998: 165). இரண்டு சொற்களுக்கு இடையிலான சொல்லடிவகைப்பாட்டில் உள்ள நுட்பமான வேறுபாடுகள் பற்றிய இத்தகைய தகவல்கள் மாணவர்களுக்கு ஒரு மொழியை சிறந்த முறையில் கற்க உதவுவதில் முக்கிய பங்கு வகிக்கின்றன.

மொழி கற்றல் மற்றும் கற்பித்தல் தவிர, அகராதி எழுதுதல், இயற்கைமொழி ஆய்வு, இயந்திர மொழிபெயர்ப்புக்கு பல்வேறு மொழியியல் பொருட்களின் சொல்லடிவகைப்பாடு பற்றிய தகவல்கள் (எ.கா. சொற்கள், உருவங்கள், மரபுச்சொற்கள் போன்றவை) முக்கியம்.

இருப்பினும், எந்த உடன்வருகை ஒரு குறிப்பிடத்தக்க/முக்கியச் சொல்லடிவகைப்பாடு என்பதைத் தீர்மானிப்பது எளிதானது அல்ல; குறிப்பாக ஒரு மொழி அல்லது மொழி வகையின் சொந்த பேசுபவர் இல்லையென்றால்.

5.7. சூழலில் முக்கிய சொல்

KWIC (க்விக்) என்பது Key Word In Context (சூழலில் முக்கியச் சொல்) என்பதன் சுருக்கமாகும்; இது ஒத்திசைவு வரிகளுக்கான பொதுவான வடிவமாகும். KWIC (க்விக்) என்ற சொல் முதலில் ஹான்ஸ் பீட்டர் லுஹ்னால் (Hans Peter Luhn) உருவாக்கப்பட்டது. இந்த அமைப்பு தலைப்புகளில் முக்கிய சொல் எனப்படும் ஒரு கருத்தை அடிப்படையாகக் கொண்டது, இது மான்செஸ்டர் நூலகங்களுக்கு முதன்முதலில் 1864இல் ஆண்ட்ரியா க்ரெஸ்டாடோரோவால் (Andrea Crestadoro) முன்மொழியப்பட்டது.

தலைப்புகளில் உள்ள ஒவ்வொரு சொல்லையும் (நிறுத்தச் சொற்களைத் (stop words) தவிர) அட்டவணையில் அகர வரிசைப்படி தேட அனுமதிக்க ஒரு கட்டுரை தலைப்புக்குள் சொற்களை வகைப்படுத்தி வரிசைப்படுத்துவதன் மூலம் ஒரு 'சூழலில் முக்கியச் சொல்' அட்டவணை உருவாகிறது. கணினிமயமாக்கப்பட்ட முழு உரை தேடல் (full text search) பொதுவானதாக மாறுவதற்கு முன்பு தொழில்நுட்ப கையேடுகளுக்கு (technical manuals) இது ஒரு பயனுள்ள அட்டவணை முறையாகும்.

சூழலில் முக்கியச் சொல் (Key Word in Context (KWIC)) தரவுத்தொகுதி ஆய்வில் முக்கிய பங்களிப்பு செய்கின்றது. தகவல் ஆய்வில் 'KWIC' (க்விக்) என்ற சொல் அடிக்கடி பயன்படுத்தப்படுகிறது. இது குறிப்பிட்ட சொற்களின் நேர்வுகளை அறிய உதவுகிறது. ஆய்வுக்கு உட்படும் சொல் ஒவ்வொரு வரியின் மத்தியில் கூடுதலான இடைவெளியுடன் தோன்றும். வேறுபட்ட நோக்கங்களுக்கு வேண்டி சூழலின் நீட்சி குறிப்பிடப்பட்டிருக்கும். இது மையத்திலுள்ள சொல்லின் இரண்டு பக்கங்களிலும் இரண்டு, மூன்று அல்லது நான்கு சொற்களின் சூழலைக் காட்டும். இதை என்-கிராம் (N-gram) குழு (set) (bi-gram/இரண்டு, trigram/மூன்று-கிராம், tetra–gram/நன்கு-கிராம்) என்பர். இவ்வமைப்பொழுங்கு ஒருவரின் தேவை அடிப்படையில் மாறும். சொற்கள், தொடர்கள் மற்றும் எச்சத்தொடர்கள் இவற்றின் ஆய்வின்

போது கூடுதலான சூழல் அவற்றைப் புரிந்துக் கொள்ளத் தேவைப்படும். கிவிக் என்பதை ஒரு உரையாக எண்ணுவது நல்லது. சொற்களின் நிகழ்வெண்ணை மையச்சொல்லின் சூழலில் பரிசோதிக்கவேண்டும். எல்லாத் தகவல்களும் எப்போதும் தேவைப்படுவதில்லை; ஆனால் நாம் தேவைப்படும்போது தகவலைப் பயன்படுத்துகின்றோம். க்விக் சொல்லால் பெருந்தரவை அணுகிய பின் நாம் மொழியியல் வர்ணணையில் பல்வேறுபட்ட நோக்கங்களை உருவாக்கலாம் மற்றும் இந்நோக்கங்களை நிறைவேற்ற வழிமுறைகளை வகுக்கலாம். க்விக் சூழலில் முக்கியத்துவம், சேர்ந்து வரும் சொற்களின் பங்களிப்பு, சூழல்களில் சொற்களின் உண்மை நடத்தை, நேர்வின் உண்மைச் சூழ்நிலை மற்றும் சூழல் அடிப்படையிலான கட்டுப்பாடுகள் இவற்றைப் புரிந்துக் கொள்ள க்விக் உதவும்.

தரவுத்தொகுதி ஆய்வில் 'சூழலில் முக்கியச் சொல்' பரவலாகப் பயன்படுத்தப்படுகிறது. (தொடரடைவைப் போன்று) குறிப்பிட்ட சொற்களின் ஒவ்வொரு நிகழ்வையும் பார்க்க இது உதவுகிறது. விசாரணையின் கீழ் உள்ள சொல் ஒவ்வொரு வரியின் மையத்திலும், இருபுறமும் கூடுதல் இடத்துடன் தோன்றும். சூழலின் நீளம் வெவ்வேறு நோக்கங்களுக்காக குறிப்பிடப்படும். இது மையத்தில் சொல்லின் இருபுறமும் இரண்டு, மூன்று அல்லது நான்கு சொற்களின் சூழலைக் காட்டுகிறது. ஒருவரின் தேவைக்கேற்ப இந்த முறை மாறுபடலாம். சொற்கள், சொற்றொடர்கள் மற்றும் உட்பிரிவுகளின் பகுப்பாய்வு நேரத்தில், சிறந்த புரிதலுக்குக் கூடுதல் சூழல் தேவை என்று ஒப்புக் கொள்ளப்படுகிறது.

'சூழலில் முக்கியச் சொல்லை' ஒரு உரையாக நினைப்பது நல்லது, மேலும் மையச் சொல்லின் சூழலில் சொற்களின் அதிர்வெண்ணை ஆராய்வது நல்லது. ஒவ்வொரு முறையும் எல்லா தகவல்களும் தேவைப்படுவது அல்ல, ஆனால் தேவைப்படும் போது தகவல்களைப் பயன்படுத்துகிறோம்.

'சூழலில் முக்கியச் சொல்லால்' ஒரு தரவுத்தொகுதியை அணுகிய பிறகு, மொழியியல் விளக்கத்தில் பல்வேறு நோக்கங்களை வகுக்கலாம் மற்றும் இந்த நோக்கங்களைப் பின்பற்றுவதற்கான நடைமுறைகளை வகுக்கலாம். சூழலின் முக்கியத்துவம், துணை சொற்களின் பங்கு, சூழல்களில் சொற்களின் உண்மையான நடத்தை, நிகழ்வின் உண்மையான சூழல் மற்றும் ஏதேனும் சூழல் கட்டுப்பாடு இருந்தால் புரிந்து கொள்ள 'சூழலில் முக்கியச் சொல்' உதவுகிறது. (ஆங்கில வங்கியில் 'சூழலில் முக்கியச் சொல்' ஒரு தற்சுட்டு வடிவத்துடன் (reflexive form) அடிக்கடி பயன்படுத்தப்படும் வினைச்சொல்லைக் காட்டுகிறது; see, show, present, manifest

and consid - இவை அனைத்தும் பிரதிநிதித்துவம் அல்லது முன்மொழிவை viewing'/'பார்ப்பது' அடக்கும்).

5.8. குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல்

குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் (local word grouping) உரைகளில் சொற்கள் பயன்படுத்தப்படும் அமைப்பொழுங்கைப் பற்றிய தகவல்களை நமக்குத் தரும் மற்றொரு வகையான ஆய்வாகும். இது எங்கு சொல் நிரல் வாக்கியங்களின் பொருண்மைச் சுமையைத் தீர்மானிக்கின்றதோ அங்கும், எங்கு ஒரு உறுப்பின் தனிப்பட்ட பொருண்மைச் சுமை மற்றொரு உறுப்பின் வருகையால் பாதிக்கப்படுகிறதோ அங்கும் முக்கியத்துவம் வாய்ந்ததாகும். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் (தொடர்கள் மற்றும் வாக்கியங்களின்) பகுத்தாய்வதன் போது உறுப்புக்களின் செயல்பாட்டு நடத்தையை நேரிடத் தகவலைத் தருகிறது. இது பயன்பாட்டில் வழக்கமற்ற ஆனால் தற்சுட்டு வடிவுகளுடன் தனிப்பட்ட உறவு கொண்ட வினைவடிவுகளின் (எகா. amuse oneself, please oneself, lend oneself, remind oneself) வகைமுறையைத் தீர்மானிக்கின்றது. இம்மாதிரியான அமைப்பொழுகின் அறிவு, மொழி கற்பவர்களுக்கு இடைப்பட்ட நிலையிலிருந்து உயர்ந்த நிலைக்குச் செல்ல முக்கியமாகும். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதலைப் பயன்படுத்தி நாம் தமிழில் முற்று வினைகள், பெரும்பாலான எச்ச வினைகளைக் கூடுதலாகத் தொடரும் என்றும் பெயர்கள் ஒட்டுகளாலும் பின்னுருபுகளாலும் தொடரப்படும் என்றும் அறிந்து கொள்ளலாம். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதலால் கிடைக்கும் தகவல்கள் சொற்களை வினைக் குழுக்களாகவும் பெயர்க் குழுக்களாகவும் ஆய உதவும். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதலால் கிடைக்கும் தகவல் சொற்களின் குறிப்பிட்ட இடம் சார் சேர்க்கையின் காரணமாக வரும் சொல் மயக்கத்தைத் தீர்க்கின்றது. பொருண்மையின் நுண்மையான வேறுபாடுகள் சூழல்களில் சொற்களின் வருகைமுறையுடன் கூடிய உறுப்புக்களுக்கிடையே உள்ள அக உறவுகளால் பெரும்பாலும் தரப்படுகின்றது. கூட்டுப் பெயர்களுக்கும் கூட்டு வினைகளுக்கும் சொற்களின் குறிப்பிட்ட சேர்க்கையால் உணர்த்தப்படும் பொருண்மையைத் தனிப்பட்ட சொற்களின் பொருண்மைகளிலிருந்து பெறவியலாது.

5.9. சொல் பகுப்பாய்வு

சொல் பகுப்பாய்வு (word processing) தரவுத்தொகுதியில் பயன்படுத்தப்படும் சொற்களைத் தானியக்கமாக ஆய்வதை உள்ளடக்கும். இதன் முக்கிய நோக்கம், ஒரு உரையின்

பகுதிலிருந்து ஒரு சொல்லைக் கண்டுபிடித்து அதை அதன் பயன்பாட்டின் சூழலிருந்து பிரித்து, அதன் உருபொலியனியல் அமைப்பை ஆய்ந்து, அதன் மூலப்பொருண்மையை பெற்று, உரையில் அதன் தொடரியல் பங்களிப்பை விளக்குவது ஆகும். சொல் பகுப்பாய்விலிருந்து கிடைக்கும் தகவல் சொல் அர்த்த மயக்கத்தைத் தீர்ப்பதற்கும் (Word Sense Disambiguation (WSD)) அகராதி உருவாக்குவதற்கும் பகுத்தாய்வதற்கும் மொழி கற்பதற்கும் மற்றும் இது போன்ற பலவற்றிற்கும் பயனுள்ளதாகும். ஆங்கிலம் போன்ற மொழிகளுக்குப் பல சொல் பகுப்பாய்விகள் (word processors) உருவாக்கப்பட்டு பயன்பாட்டில் உள்ளன. ஆங்கிலம் மற்றும் பிற மொழிகளுக்கு பல சொல் செயலிகள் உள்ளன (கிரீன் மற்றும் ரூபின்/Greene and Rubin 1971, கார்டுனென் மற்றும் விட்டன்பர்க்/Karttunen and Wittenburg 1983, கோஸ்கென்னீமி/Koskenniemi 1983, டிஹான்/deHaan 1984, கார்சைடு/Garside 1987, சர்ச் மற்றும் பல/Church et al. 1991, டிரோஸ்/deRose 1991, மெரியால்டோ/Merialdo 1994).

தமிழ் போன்ற இந்திய மொழிகளுக்கும் இத்தகைய பகுப்பாய்விகள் உருவாக்கப்பட்டு பயன்படுத்தப்பட்டு வருகின்றன.

தமிழில் உருபனியல் ஆய்விகள்

தமிழ்மொழி ஒரு திராவிட மொழி. இம்மொழி வினைச்சொல்லை இறுதியாகக் கொண்ட வாக்கியங்களை உடையதாகும். இம்மொழி இலக்கணச் செயற்பாங்குகளை (அதாவது இலக்கணப் பொருண்மைகளை) உருபனியலில் நிலையில் வெளப்படுத்தும் ஒட்டுக்கள் நிறைந்த மொழியாக விளங்குகிறது. இம்மொழியில் உள்ள ஒவ்வொரு வேர் உருபனும் ஆயிரக்கணக்கான சொற்களின் உருவாக்கத்திற்கு (அதாவது திரிபு மற்றும் ஆக்க வடிவுகளின் உருவாக்கத்திற்கு) அடிப்படையாக அமைகிறது. தமிழ் வாக்கியங்கள் எழுவாய்தொடரையும் வினையையும் கொண்ட அமைப்பாகவோ எழுவாய்த்தொடரையும் செயப்படுபொருள் தொடரையும் வினையும் கொண்ட அமைப்பாகவோ எழுவாய்த்தொடரையும் பெயர்த்தொடரையும் (அதாவது பயனிலைத் தொடரையும்) கொண்ட அமைப்பாகவோ வரும். இதன் காரணமாகத் தமிழ் மொழி ஒரு SOV மொழி என்பர். தமிழ் மொழி ஒரு முன் மாதிரி SOV மொழிக்குள்ள எல்லாப் பண்புக்கூறுகளையும் பெற்றுள்ளது. பெயர்கள் வேற்றுமை உருபுகளுக்காகத் திரிபுறுகின்றன. பெயர்த்தொடர்களுக்கும் வினைகளுக்கும் உள்ள உறவுகள் வேற்றுமை உருபுகளாலோ பின்னுருபுகளாலோ வெளிப்படுத்தப்படுகின்றன. வினைகள் கால இடைநிலைகளை ஏற்பதாகவும் எழுவாய்த்தொடருக்கு உடன்பாடாக வினைமுற்று விகுதிகளை ஏற்பதாகவும் வருகின்றன. இதன்

காரணமாக எந்தக் கணிப்பொறி தொடர்பான மொழி ஆய்விற்கும் உருபனியல் பகுப்பாய்வு இன்றியமையாததாகவும் அடிப்படையாகவும் அமைகின்றது. மேலும் எழுத்துப்பிழை திருத்தி உருவாக்கத்திற்கும் இலக்கணப்பிழை திருத்தி உருவாக்கத்திற்கும் உருபனியல் பகுப்பாய்வு அடிப்படையாகும்.

உருபனியல் ஆய்வியின் உருவாக்கம் எந்த இயற்கை மொழி ஆய்வின் செயல்பாட்டிற்கும் மிக முக்கியமான முதற்படியாகும். வேறுபட்ட மொழி வகைகளுக்கு வேறுபட்ட ஆய்வு நெறிமுறைகள் இருக்கின்றன. எளிய இணைப்பு உருபனியல் உள்ள (simple concatinative morphology) மொழிகள் சொல்லும் செயற்பாங்கும் (item and process) மற்றும் சொல்லும் வரிசைமுறையும் (item and arrangement) என்ற மொழி மாதிரிகளால் உருவாக்கப்பட்ட ஆய்விகளால் தமது தேவையைப் பூர்த்தி செய்துகொண்டுள்ளன. பண்புக்கூறுகள் அடிப்படையிலான ஒருமைப்படுத்தும் கோட்பாடுகள் (unification theories) ஸ்பானிஷ் போன்ற மொழிகளுக்குப் பயன்படுத்தப்பட்டது. இக்காலத்தைய மிகப் புகழ்வாய்ந்த கணினிசார் இலக்கணமான தலைமையால் இயக்கப்படும் தொடரமைப்பு இலக்கணம் (Head-Driven Phrase Structure Grammar (HDPSG) பண்புக்கூறு அடிப்படையிலான உருபனியலைப் (Feature Basecd Morphology) பயன்படுத்துகின்றது. இங்கு தமிழ்மொழியின் உருபனியல் ஆய்வுக்குப் பயன்படுத்தப்பட்ட மாதிரிகளும் வேறு முன்னணி மாதிரிகளும் விளக்கப்படும்.

கணிப்பொறி வழி ஆய்வைப் பொறுத்தவரையில் இருமட்ட மாதிரி (two level model) உருபனியலின் முக்கிய கோட்பாடுகளில் ஒன்றாகும். இது பரந்த பரப்பெல்லையில்படும் மொழிகளுக்குச் செயல்படுத்தப்பட்டிருக்கின்றது. முற்றுநிலை தானியங்கி (Finite State Automata (FSA)) மாதிரி சென்னையிலுள்ள AUKBC ஆய்வு மையத்தால் உருவாக்கப்பட்ட உருபனியல் பகுப்பாய்வி முற்றுநிலைத் தானியங்கி அட்டவணையைப் (finite automata state table) பயன்படுத்தி செயல்படுகின்றது. தலைமைத் தொடரில் அமைப்பு இலக்கணம் (Head Phrase Structure Grammar (HPSC)) தமிழுக்கு உமாராணியால் (Umarani, 2001) பயன்படுத்தப்பட்டுள்ளது. ஆம்பிள் என்ற மற்றொரு மாதிரி சொல்லும் வரிசையும் என்ற மாதிரியுடன் தொடர்புடையது. பகுத்துலும் உருவாக்கலும் (Parsing and Generation) இயற்கை மொழிகளின் உருபனியல் ஆய்வின் இரு முக்கியமான நேர்வுகள் ஆகும். பகுப்பாய்வு என்பது பகுப்பாய்வி தரப்பட்ட உள்ளீட்டை புரிந்து கொள்ளவும் ஆயவும் செய்து பெரும்பாலும் அவற்றைப் பொருண்மை மற்றும் இலக்கணப் பண்புக் கூறுகளுடன் வெளியீடாகத் திருப்பித்தரும்

செயல்பாடாகும். பல உரை மீட்பு ஒழுங்குமுறைகளுக்கும் (Text Retrieval Systems (TRS)) இயந்திர மொழிபெயர்ப்புச் (Machine Translation (MT)) செயல்பாடுகளுக்கும் பகுப்பாய்வு முக்கியமானது ஆகும். அதாவது இந்த இரண்டு செயல்பாடுகளுக்கும் தொடக்க இடம் பகுப்பாய்வாகும். ஒரு உணரி (recognizer) உள்ளீட்டுச் சொல்லை வேறுபட்ட உருபன்களாகப் பகுப்பாய்வு செய்கின்றது. ஆங்கிலச் சொல்லான reads என்பது பின்வருமாறு புரிந்து கொள்ளப்படும்: read என்ற வினையில் படர்க்கை இட ஒருமை எண் குறிப்பிடப்பட்ட நிகழ்கால வடிவம். பகுப்பாய்வி உணரியிலிருந்து அது புரிந்து கொள்ளும் சொற்களுக்குப் பொருத்தமான அமைப்பைத் தருகிறது என்ற அளவில் வேறுபடுகின்றது. இவ்வாறு ஒரு உருபனியல் பகுப்பாய்வி உள்ளீடு செய்யப்பட்ட சொல்லை அதன் உறுப்பு உருபன்களாகப் பிரித்து அதனைப் படிநிலை அமைப்பாகப் பகுப்பாய்வு செய்கிறது. இது சொல்லாக்கத்தில் உருபன்களை இணைப்பதற்கு மொழியில் குறிப்பிட்ட இயக்க நெறியின் ஒரு நோக்கை வெளிப்படுத்துகிறது. இவ்விக்கநெறி ஒரு வேருடன் கட்டுண்ட உருபுகளை முன்னொட்டாகவோ உள்ளொட்டாகவோ பின்னொட்டாகவோ வேருடன் சேர்க்கும். தமிழ்மொழியில் இது கூடுதலாகப் பின்னொட்டாக்கம் ஆகும். முன்னொட்டுக்கள் சமஸ்கிருதத்தில் இருந்து வந்த சில கடன் சொற்களில் இருக்கின்றன. மாறாக, உருபனியல் உருவாக்கி (Morphological Generator) உருபன்களை இணைத்துச் சொற்களை உருவாக்கும் நோக்கைப் பயன்படுத்துகிறது. இவ்வாறு உருபனியல் உருவாக்கி பகுப்பாய்வதன் மறுதலையாகும். இதன் உள்ளீடு ஒரு குழும உருபன்களோ அதன் இலக்கண அமைப்புகளோ ஆகும். இது அவற்றை ஒன்றிணைத்து ஒரு மொழியில் நல்வடிவாக்கமுள்ள சொற்களை உருவாக்குகிறது. உருபனியல் பகுப்பாய்வி இச்செயல்பாடுகளில் ஒன்றையோ அதற்கு மேலானவற்றையோ செய்கின்றது. வேறுபட்ட உருபனியல் ஆய்விகள் அவை விளக்கும் மொழிகளின் அடிப்படையில் வேறுபட்ட கோட்பாடு அணுகுமுறைகளைப் பயன்படுத்துகிறது. அதாவது வேறுபட்ட மொழிகளின் இயல்பும் கலவைத்தன்மையும் அவற்றை ஆயப்பொருத்தமான கோட்பாடு அணுகுமுறைகளைத் தீர்மானிக்கிறது. வேறுபட்ட சட்டங்களின் அடிப்படையில் வேறுபட்ட உருபனியல் ஆய்விகள் உருவாக்கப்படுகின்றன. வேறுபட்ட சட்டகங்களின் தோ;வும் அவற்றின் வேறுபட்ட நடைமுறைபடுத்தல்களும் ஒப்பீடு செய்யும் ஆய்வுக்கு அவசியமாகிறது.

இப்பகுதியில் இருமட்ட உருபனியலைப் பயன்படுத்தும் பிசிகிம்மோ (PCKimmo), ஆம்பிள் (Ample) என்பனவும் சொல்லும் வரிசைமுறையும் என்ற மாதிரியின் அடிப்படையில் அமைந்த ஜிஎஸ்மார்ஃபும் (GSMorph) சொல்லும் செயற்பாங்கும் மாதிரியின் அடிப்படையில் அமைந்த

இராமசாமியின் (Ramaswamy, 2000) தமிழ் உருவாக்கியும் (Generator for Tamil) சொல் ஒலியனியல் (Lexical Phonology) அணுகுமுறை அடிப்படையில் அமைந்த இரங்கநாதனின் (Renganathan,1997) டேக் தமிழும் (Tag Tamil) சொல்லும் வரிசைமுறையும் மற்றம் சொல்லும் செயற்பாங்கும் என்ற இரு மாதிரிகளையும் பயன்படுத்தும் வைஷ்ணவியின் தமிழ்; உருபனியல் பகுப்பாய்வியும் உருவாக்கியும் இராசேந்திரனின் உருபனியல் ஆய்வியும் விளக்கப்படுகின்றன. இவை தவிர தமிழில் முயற்சிக்கப்பட்டுள்ள பிற உருபனியல் ஆய்விகளும் பட்டியலிடப்பட்டுள்ளன.

தமிழில் கணிப்பொறி தொடர்பான செயல்பாடுகளுக்கு அடிப்படையாக அமைவது உருபனியல் பகுப்பாய்வும் உருவாக்கமும் ஆகும். பலர் இதில் பங்களிப்பு செய்துள்ளனர். பல உருபனியல் ஆய்விகள் தமிழில் உருவாக்கப்பட்டுள்ளன. இவற்றில் முக்கியமானவைகளைப்பற்றிய சிறு குறிப்பு கீழே தரப்பட்டுள்ளன.

இராஜேந்திரனின் தமிழ் உருபனியல் பகுப்பாய்வி

முதல் முதலில் இந்திய மொழிகளுக்கிடையில் இயந்திர மொழி பெயர்ப்பை மேற்கொள்ள வேண்டி அனுசாராகா (anusaraka ‘language acessor’) என்ற மொழிபெயர்புக் கருவி உருவாக்கத்திற்காகப் பாணினி வடிவவாதம் அடிப்படையில் ஆய்வு மேற்கொண்ட குழுவைச் சார்ந்த ஆய்வாளர்களால் DOE நிதி நல்கையில் தொடங்கிய ஆய்வுதான் இந்தியாவில் இயற்கை மொழி ஆய்விற்கு வித்திட்டது எனலாம். இந்திய மொழிகளுக்கிடையிலான இயந்திர மொழிபெயர்ப்பு ஆய்வுக்காக உருவாக்கப்பட்ட உருபனியல் பகுப்பாய்விகள் அடிப்படையில் தமிழ்-இந்தி மொழி பெயர்ப்புக்காகத் தமிழ் உருபனியல் பகுப்பாய்வியும் இராசேந்திரன் உதவியுன் உருவாக்கப்பட்டது. AUKBC ஆய்வு நிறுவனக் குழுவினரால் உருவாக்கப்பட்ட அபி உருபனியல் பகுப்பாய்வியும் இராசேந்திரனின் மேற்பார்வையில் உருவாக்கப்பட்டதுதான். தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையில் பல்கலைக்கழக மானியக் குழுவின் நிதிநல்கையில் கீழ் மேற்கொள்ளப்பட்ட “Spell and Grammar Checher for Tamil” என்ற உயராய்வுக்காக PROLOG வழியமைப்பு மொழியைப் பயன்படுத்தி ஒரு உருபனியல் பகுப்பாய்வி இராசேந்திரனால் உருவாக்கப்பட்டது. இப்பகுப்பாய்வி இந்திய மொழிகளின் நடுவண் நிறுவன விரிதரவில் பயன்படுத்தப்பட்டு மதிப்பீடு செய்தபோது 95% வெற்றிகரமாக பகுத்துக்குறித்தல் வெளியீடு கிடைத்தது.

முன்னர் குறிப்பிட்டது போன்று இந்த உருபனியல் பகுப்பாய்வி சொல்லடுக்கு அணுகுமுறையை (word and paradigm) அடித்தளமாகக் கொண்ட அடுக்கு அணுகுமுறை (paradigm approach) அடிப்படையில் உருவாக்கப்பட்டுள்ளது.

கணேசனின் உருபனியல் பகுப்பாய்வி

கணேசன் (Ganesan, 1994, 2003) அவர்கள் தமிழுக்காக இந்திய மொழிகளின் மையநிறுவனம் உருவாக்கிய மூன்று மில்லியன் சொற்கள் அடங்கிய விரிதரவை (CIIL Corpus) ஆய வேண்டி ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்கினார். இவர் ஒலியனியல் மற்றும் உருபனியல் விதிகளைக் கையாண்டு இப்பகுப்பாய்வியை உருவாக்கினார். சமீபத்தில் சிறந்த தமிழ் உருபனியல் பகுப்பாய்வியை உருவாக்கியுள்ளார்.

கபிலனின் தமிழ் வினைகளுக்கான உருபனியல் பகுப்பாய்வி

கபிலன் (Kabilan, 1994) அவர்கள் தமிழ் வினைகளுக்கான ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்கினார். இது பேராசிரியர் தெய்வசுந்தரத்தின் (மொழியியல் பகுதி, தமிழ்மொழித் துறை, சென்னைப் பல்கலைக்கழகம்) மேற்பார்வையின் கீழ் நடந்த தமிழ்சார் இயற்கை மொழி ஆய்வின் விளைவாகும்.

தெய்வசுந்தரத்தின் உருபனியல் பகுத்துக் குறிப்பான்

தெய்வசுந்தரம் தமிழுக்கென ஒரு சொல்லாய்வியைத் (Word Processor) தயாரித்துள்ளார். இது மிக நன்றாகச் செயல்பாடு செய்கின்றது. இது இருநிலை மாதிரியை (two-level-model) அடிப்படையாகக் கொண்டது. புணர்ச்சி விதிகள் பகுப்பாய்வுக்குப் பயன்படுத்தப்படுகின்றன. இவர் ஆய்வி (Deivasundaram and Gopal 2003) உருபொலியனியல் மற்றும் உருபனியல் என்ற இருநிலைகளில் செயல்படுகின்றது. இது ஒலியன் சேர்ப்பையும் (phonemic addition), ஒலியன் மாற்றத்தையும் (phonemic change) கணக்கிலெடுக்கின்றது. இவ்வாய்வி உருபனியல் அமைப்பொழுங்கு (mophotactics) அடிப்படையில் ஒட்டுகள் பகுதியுடன் இணைக்கப்படும் நிரலைக் கணக்கிலெடுத்துக் கொண்டு பகுப்பாய்வு செய்யும். இது ஒரு எடுத்துக்காட்டான உருபனியல் பகுப்பாய்வியாகும். முற்றிலும் மொழியல் விதிகளைப் பயன்படுத்தி வெற்றிகரமாக முடிக்கப்பட்டு தமிழுக்கான தலைச் சிறந்த சொல்லாய்வியாக (சொல்லாளரகாக/மென் தமிழாக) வளர்து இது தமிழ் உலகை வலம் வருகின்றது. அரசுப்பணிகளுக்காக இச்சொலாய்வி தமிழக அரசால் பரிந்துரைக்கப்பட்டுப் பரவலாகப் பயன்படுத்தப்பட்டு வருகின்றது.

AUKBC நிறுவனத் தமிழ் உருபனியல் பகுப்பாய்வி

அண்ணா பல்கலைக்கழகம் கே.பி. சந்திரசேகர் ஆராய்ச்சி மையம் (Anna University K.B. Chandrasekhar Research Centre (AUKBC)) ஆய்வு நிறுவனம் (தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையுடன் இணைந்து) உருவாக்கிய API உருபனியல் சொல்லாய்வி தரவுக்கோப்பு, நிலை மாற்றமைவு அட்டவணை (State table.dat) என்பனவற்றில் தந்துள்ள தானியங்கி அட்டவணையைச் சார்ந்து இருக்கிறது. API உருபனியல் சொல்லாய்வி சொற்களை அதன் திரிபுகளிலிருந்து பிரித்தெடுக்கும்படி வடிவமைக்கப்பட்டுள்ளது.

வைஷ்ணவியின் தமிழ் உருபனியல் பகுப்பாய்வியும் உருவாக்கியும்

வைஷ்ணவி (Vaishnavi Ramaswamy, 2003) முனைவர் பட்ட ஆய்வுக்காகத் தமிழ் உருபனியல் பகுப்பாய்வை எடுத்துக்கொண்டு ஆய்வேட்டினைச் செய்துள்ளார். மேலும் வைஷ்ணவி (Vaishnavi Ramaswamy, 2002) தமது ஆய்வியல் நிறைஞர் ஆய்வேட்டிற்காக தமிழுக்கு ஒரு உருபனியல் உருவாக்கியை உருவாக்கி விளக்கமாகத் தந்துள்ளார். அவருடைய உருபனியல் உருவாக்கி சொல்லும் வரிசைமுறையும் (word and arrangement model) மாதிரியையும் சொல்லும் செயற்பாங்கும் மாதிரியையும் (word and process model) சொல்லும் அடுக்கும் (word and paradigm model) ஆகிய மூன்று கோட்பாடுகளையும் பயன்படுத்தி மேற்கொள்ளப்பட்ட ஒரு கலப்பு அணுகுமுறைகாகும். இந்த உருவாக்கி பிசிகிமோவின் (Pckimmo) கூட்டிணைப்பு நெறிமுறையால் செயல்படுகின்றது. வைஷ்ணவி இதைக் கலப்பு மாதிரி (hybrid model) என்று அழைக்கின்றார்.

PERL–இல் உருவாக்கப்பட்ட மற்றும் செயல்படுத்தப்பட்ட கலப்பு மாதிரி (hybrid model) உள்ளுருவாக்கப்பட்ட சொற்களஞ்சியத்தின் சட்டகத்தையும் மொழியின் வேறுபட்ட ஒவ்வொரு சொல் வகுப்பிற்கும் சொல்லும் வரிசைமுறையும் மற்றும் சொல்லும் செயற்பாங்கும் என்பனவற்றின் ஒரு குழும விதிகளையும் கொண்டிருக்கிறது. தரப்பட்ட உருபொலியனியல் விதிகளின் அடிப்படையில் செயல்படும் இணையான துணை நடைமுறைகளை (corresponding subroutines) வேண்டுவதன் மூலம் இதன் ஆய்வு நிரூபிக்கப்படுகிறது. இது ஒவ்வொரு வகைப்பாட்டின் சொல் வடிவுகளைக் கையாளும்படி வடிவமைக்கப் பட்டுள்ளது. இது முதன்மை அமைப்பின் கோப்புகளால் கட்டளையிட்ட படி செயல்படுகிறது. வழியமைப்பு மொழியில் மொழித் தகவலின் மரபுரிமை குறியாக்கம் இருக்கிறது. திறந்த இறுதியுள்ள உருபொலியனியல் மற்றும் மாற்றுருபு விதிகளின் (open ended morphophonimic and allomorphic rules) வழி மொழித் தகவலின் பகுப்பாய்வால் இது நடைப்பெறுகிறது (Vaishnavi Ramaswamy, 2003:101-110).

இம்மாதிரி உருபனியல் சொல்லும்வரிசைமுறையும் மற்றும் சொல்லும் செயற்பாங்கும் மாதிரிகளின் கலவையில் அதன் கோட்பாட்டு அடிப்படையை அமைத்துக் கொள்கிறது. செயற்பாட்டைச் சொற்பகுதியில் (word stem) உருபனியல் இயல்பைப் பொறுத்து ஒரு மாற்றத்தை (transformation) விவரணை செய்வதாக விளக்க இயலும். திரிபு நீக்கத்தின் (inflectional stripping) ஒவ்வவொரு நிலையிலும் குறிப்பிடப்பட்டுள்ள கட்டுப்பாடுகளின் முழுத்தொடர்சிகளின் வழி கடந்த பின்னர் ஒரு சொல் பகுப்பாய்வு செய்யப்படுகிறது. ஒரு பகுப்பாய்வுத் தொகுதி (analysis module) வலமிருந்து இடமாகக் கோர்வை அமைவதை வருடிச் (scanning) செய்கிறது. ஒரு நேரத்தில் ஒரு பின்னொட்டை வருடி அறிந்து பிரித்தெடுக்கிறது. மற்றும் ஒவ்வொரு நேர்விலும் எடுத்துக்காட்டப்பட்ட ஒலியனியல் மற்றும் உருபொலியனியல் விதிகளின் உதவியுடன் மீதிச்சொல்லை மீட்டுருவாக்கம் செய்கிறது. கோர்வை தீர்வது வரை இது நடைபெறும். வெளியீடு கண்டுபிடிக்கப்பட்ட உருபன்களின் வடிவில் இருக்கும். பின்னர் அவை உறுதி செய்யப்படும்.

இவர் தமது உருபனியல் ஆய்வி உருவாக்கத்திற்காக உருபொலியனியல் விதிகள் அடிப்படையில் பெயர்களை 24 குழுக்களாகவும் மாற்றுப்பெயர்களை 15 குழுக்களாகவும் வினைகளை 24 குழுக்களாகவும் பெயரடைகளை இரண்டு குழுக்களாகவும் பகுப்பாய்வு செய்து கணினிக்கு உருபொலியனியல் விதிகளை அறியச்செய்து தமிழ் உருபனியல் பகுப்பாய்வியை உருவக்கியுள்ளார்.

வின்ஸ்டன் குருசின் ஜி.எஸ். மார்ஃப் பகுப்பாய்வியும் உருவாக்கியும்

ஜி.எஸ். மார்ஃப் (GS Morph) என்று சுருக்கமாகக் குறிப்பிடப்படும் எழுத்து அடிப்படையிலான அமைப்பு மொழியியலார் உருபனியல் பகுப்பாய்வி (Grapheme Oriented Structuralist Morphological Processor (GS Morph)) (Sengupta 1999) ஆம்பிள் போன்று சொல்லும் வரிசை முறையும் மாதிரியின் அடிப்படையில் அமைந்தது. இது ஒவ்வொரு உருபனின் விரிவான புற மெய்ப்படுத்தத்தை வேண்டும் அல்லது ஏற்படுத்தும். ஆம்பிள் ஒட்டுக்களை அவை பின்னொட்டுக்களா உள்ளொட்டுக்களா என்பதன் அடிப்படையில் வேறுப்பட்ட கோப்புகளில் சேகரிக்க இயலும். ஆனால் ஜிஎஸ் மார்ஃபின் பகுப்பாய்வு ஒரு வேர்ச் சொற்களஞ்சியத்தையும் (Root Lexicon) பின்னொட்டுக்களின் சொற்களஞ்சியத்தையும் (Suffix Lexicon) உட்படுத்தும்.

ஜிஎஸ் மார்ஃப் சொல்லடைவைப் பொருத்துவதால் செயல்படுகிறது. அதன் உள்ளீடு எழுத்து உருப்படுத்தத்தின் வடிவில் இருக்கிறது. அது வலது எல்லையிலிருந்து இடது எல்லைக்குப்

பகுப்பாய்வைத் தொடங்குகிறது. முதலாவது அது தரப்பட்ட உள்ளீட்டைப் பூஜிய பின்னொட்டால் தொடரப்படும் வேராகக் கருதுகிறது. இதன்படி இது பின்னொட்டுச் சொற்களஞ்சியத்தில் பின்னொட்டுக்காகப் பரிசோதிக்கும். பின்னர் அது வோ;க்கோப்பில் பூஜிய ஒட்டிற்கு நேராக இருக்கும் சொல்லடைவுகளைப் (concordance) பார்க்கும். அது உள்ளீட்டைப் பொருத்துவதில் தோல்வியுற்றால் அது சொல் வடிவத்தைப் பற்றிய அதன் ஆரம்ப அனுமானத்தை மீளப்பெறும். அது உள்ளீட்டின் எழுத்துக்களை வெட்டி எடுக்கத் தொடங்கும். ஒவ்வொரு தடவையும் ஒரு எழுத்தை வெட்டி எடுக்கும்போது அது எல்லா வெட்டி எடுக்கப்பட்ட எழுத்துக்களும் சேர்ந்த ஒன்றைப் பின்னொட்டுச் சொற்களஞ்சித்தில் பொருத்தத்திற்காகப் பார்க்கும். வெட்டி எடுக்கும் செயற்பாங்கு முழுகோர்வையும் தீருவது வரை தொடர்ந்து நடைபெறும். பின்னொட்டுச் சொற்களஞ்சியத்தில் தேடல் வெற்றி பெற்றால் உள்ளீட்டின் மீதிப்பகுதி வேராக முன்மொழியப்பட்டு வேர்க்கோப்பில் பார்க்கப்படும். முன்மொழியப்பட்ட வேர்ச் சொற்களஞ்சியத்தில் காணப்பட்டால் பின்னொட்டு மற்றும் வேர்க் கோப்புகளில் இருந்து அவற்றின் தொடர்புடைய பொருள் தரப்படும். இது வேர் மற்றும் பின்னொட்டுக் கோப்புகளைச் சொல்லடைவு செய்வதன் மூலமும் செய்யப்படும். சொல்லடைவு செய்யப்படும் சமயத்தில் வேர் மற்றும் பின்னொட்டுச் சொற்களஞ்சியங்களுக்குச் சிறு அட்டவணைகள் (hash tables) உருவாக்கப்படும். வழிமுறை வரைவுகள் (algorithms) தொடர்புடைய சிறு அட்டவணைகளில் தொடர்புடைய சொல்லடைவுகளைப் பொருத்திப் பார்க்க வேண்டி பரிசோதிக்கும். முழு உள்ளீடும் தீர்க்கப்படுவது வரை இச்செயற்பாங்கு நிற்பதில்லை என்பதை முக்கியமாகக் கவனிக்க வேண்டும். இப்படி அது ஒரு பொருத்தத்தைக் கண்டுபிடித்து வெளியீட்டைத் தந்தாலும், சாத்தியம் என்றால் பகுப்பாய்வி கூடுதலான பகுப்புகளைக் கண்டு பிடிக்க இயலும் என்பதாகும். இது வேர் மற்றும் மாற்றுருபுகளுக்கு ஒரே ஒலியனியல் வடிவு இருந்தால் பகுப்பாய்வு இரண்டிற்கும் எல்லாச் சாத்தியமான பகுப்பாய்வையும் தரும். தற்போதைய நடைமுறைப்படுத்தலில் விளைவு (effective) வேறுபாட்டைக் காலப் பின்னொட்டில் பார்க்கப்படும் நிலையில் தமிழில் வினைகளுக்கு ஒரு தனி வேர் உள்ளீட்டைத் தருவது, ஜி.எஸ். மார்ஃபில் வேர் மற்றும் பின்னொட்டுக் கோப்புகள் அமைக்கப்படும் முறையின் காரணமாக இது சாத்தியமாகிறது. இது பகுப்பாய்வின் போது ஒரே நடப்பை மிக எளிதாகப் பொருண்மை மயக்கம் தீர்க்க உதவுகிறது.

ஜி.எஸ் மார்ஃப் பகுப்பாய்வி சொல்லின் வரிவடிவ உருபடுத்தத்தை உள்ளீடாக ஏற்கின்றது. சொற்கள் இடது எல்லையிலிருந்து பகுப்பாயப்படுகின்றன. முன்னர் விளக்கப்

பட்டதுபோல் வேர் மற்றும் பின்னொட்டு சொற்களஞ்சியங்களின் பயன்பாட்டால் பகுப்பாய்வு வெற்றிகரமாக நடைபெறும். ஜி.எஸ். மார்ஃபின் உருவாக்கி உட்கூறு பகுப்பாய்வுக்கான அதே சொற்களஞ்சியங்களைப் பயன்படுத்துகின்றது. உள்ளீடு தரப்படும்போது உருவாக்கி அதை வேர்ச் சொற்களஞ்சியத்தில் பரிசோதித்து அதற்குத் தரப்பட்ட சொல்லடைவை எடுக்கும். அது பின்னர் அதே சொல்;லடைவைக் கொண்ட கோர்வைகளைப் பின்னொட்டு அகராதியில் தேடும். ஏதாவது கோர்வை அதே சொல்லடைவுடன் இருந்தால் வோ;மாற்றுரு அப்பின்னொட்டுக் கோர்வையுடன் சேரும். இது பின்னொட்டில் உள்ள உருபன்களின் அர்த்தங்களுடன் திருப்பி அனுப்பப்படும். பகுப்பாய்வுக்குப் பயன்படுத்தப்படும் அதே கோப்புகள் பயன்படுத்தப்படுவதால் உருவாக்குதலும் துல்லியமாக இருக்கின்றது (Winston Cruz, 2002:127)..

வின்ஸ்டன் குருஸ் (Winston Cruz, 2000) தமிழ் வினைச்சொற்களைப் பகுப்பாய்வு செய்வதற்கு ஜி.எஸ். மார்ஃப் முறையைக் கையாண்டார். ஜி.எஸ் மார்ஃப் என்பது உருபன் அமைப்பொழுங்கு குறியீட்டால் ஆனது. வழிமுறை வரைவு (algorithm) இரண்டு கோப்புகளை பொருத்தமானதா இல்லையா என்று மட்டுமே ஆய்வு செய்கிறது. சொல்லாய்வி இரண்டு வடிவங்களை மட்டுமே பயன்படுத்தி இரண்டு கோப்புகள் படைக்கப்பட்டுள்ளது.

துரைபாண்டியின் உருபனியல் பகுப்பாய்வு

தற்கால தமிழில் வினைச்சொல் வடிவத்தினை உருவாக்கும் பகுப்பாய்வு இயந்திரத்தை முழுமையாக உருவாக்கியுள்ளார். இது இவர் உருவாக்கிய சொல்லாய்வியில்/சொல்லாளரில் பயன்படுத்தப் பட்டுள்ளது.

RCILTS-T-இன் தமிழ் உருபனியல் பகுப்பாய்வி

இந்திய மொழி தொழில்நுட்பத் தீர்வுகளுக்கான வள மையம்-தமிழ் (Resource Centre for Indian Language Technological Solutions-Tamil (RCILTS-T)) என்ற அமைப்பின் தமிழ் உருபனியல் பகுப்பாய்வி அட்சரம் (atcharam) என அழைக்கப்படுகின்து. அட்சரம் வேர்ச் சொல்லிலிருந்து ஆக்கச் சொல்லைத் தனியாகப் பகுப்பாய்வு செய்து உருபனுடன் தொடர்புபடுத்துகிறது. இதனைப் பயன்படுத்தி 15 வகைகளை அடிப்படையாகக் கொண்டு இருபதாயிரம் வேர்ச்சொற்கள் தமிழ் அகராதியில் உருவாக்கப்பட்டுள்ளன. இது இரண்டு தொகுதிகளை உடையது. அதாவது பெயர் மற்றும் வினைச் சொல்லை அடிப்படையாகக் கொண்டது. இது ஊக விதிகளுடன் (heuristic rules) பொருண்மை மயக்கம் உடையதாகும். அதற்கு வினை மற்றும் பெயர் திரிபுகளைக் கையாண்டுள்ளனர்.

RCILTS-T-இன் உருபனியல் உருவாக்கி

இந்நிறுவனம் தமிழ்மொழியலும் உருபனியல் உருவாக்கியைத் தயார் செய்துள்ளது. அதற்கு அட்சயம் (Atchayam) எனப் பெயரிடப்பட்டுள்ளது. அட்சயம் என்பது பெயர் மற்றும் வினைச்சொற்களைக் கொண்டு இரண்டு முக்கிய உள்ளீடுகளைப் படைத்துள்ளது. இதில் பெயர்ச்சொல் பிரிவில் பின்னொட்டு இடம்பெறும். எடுத்துக்காட்டாக, பன்மை ஒட்டுக்களையும் திரிபு வடிவங்களையும் வேற்றுமை உருபுகளையும் பின்னுருபுகளையும் உருவாக்கியுள்ளனர். இதே போல் வினைப்பிரிவில் காலங்களையும் இடம்-எண்-பால் உருவாக்கிகளையும் பெயரச்சம், வினையெச்சம், பின்னொட்டு மற்றும் துணை வினைச்சொற்களையும் கொண்டுள்ளது. இது சந்தி விதிகள் மற்றும் 125 உருபனியல் விதிகளையும் பயன்படுத்தி உள்ளது. இது பெயரடை மற்றும் வினையடையையும் கையாண்டுள்ளது. இதில் சொல் மற்றும் வாக்கிய உருவாக்கியைக் கொண்டு ஆய்வு செய்யப்பட்டுள்ளது.

பொருண்மை மயக்கம் அல்லாத தமிழ் உருபனியல்

சொல் வடிவில் பொருண்மை மயக்கம் இருக்கும். இதற்கு மாறாக ஒப்புருச் சொல்லின் (homonymy) வகைப்பாடுகளில் எதிர் பாராத குறுக்கீடுகள் உருவாவதால் உருபனியல் பிறழ்ச்சி, ஒட்டுக்களின் பல்வகைச் செயல்பாடுகள் அல்லது நிச்சயமற்ற பின்னொட்டுச் சொற்கட்டுகளையும் (word bounding) உருவாக்குகிறது. இவ்விதம் வாக்கிய அமைப்புச் சூழல்களை இயல்பாகத் தீர்மானிக்கின்ற பகுப்பாய்வு ஆகும். இதனையே பொருண்மை மயக்கம் அல்லாத உருபனியல் என்கிறோம். இரண்டு அடிப்படை அணுகுமுறைகள் சமவாய்ப்பு (probability) உடையது. விதியை அடிப்படையாகக் கொண்ட அணுகுமுறையில் சில பொருண்மை மயக்கம் தீர்வு காணப்பட முடியாமல் உள்ளது. ஆனால் மிகச் சில தவறுகளைக் கொண்டுள்ளது. பொதுவாகப் புள்ளியியல் அடையாளப்படுத்திகள் முழுவதும் பொருண்மை மயக்கமற்ற வெளியீடுகளைத் தருகின்றன. ஆனால் அவற்றறுள் நிறைய தவறுகள் காணப்படுகின்றன. சமவாய்ப்புள்ள முறையில் 1980-இருந்து உருபனியலில் பொருண்மை மயக்கமற்ற பகுப்பாய்வே அதிகம் இடம் பிடித்துள்ளது. நிலையான புள்ளியியல் முறைகளின் பயன்பாடு முழுவதும் பொருண்மை மயக்கமற்ற வெளியீடுகளாக உள்ளது. 'Influence of Morphology in Word Sense Disambiguation for Tamil' (தமிழ் உருபனியலில் பொருண்மை மயக்கம் அல்லாத சொல் செல்வாக்கு உடையது). பரிசோதனைகளை மேற்கொள்ளும்பொழுது மேற்பார்வை மற்றும் பகுதி மேற்பார்வை அணுகுமுறையை மேற்கொண்டு உருபனியல் திரிபுகளின் பாதிப்புமுறைகளை பற்றியும் திடமான

தீர்வான உருபனியலில் பொருண்மை மயக்கம்மற்ற பகுப்பாய்வை தமிழ்மொழியிலும் பிற திரிபு மொழிகளிலும் பயன்படுத்தலாம் எனத் தீர்வு கண்டுள்ளனர்.

உருபனியலும் முற்றுநிலை மாற்றிகளும் (Morphology and Finite State Transducers)

ஒரு சொல்லின் கூறுகளை ஆயும் செயல்பாடுதான் உருபனியல் பகுத்துக்குறித்தல் (Morpholocal Parsing) எனப்படும். எடுத்துக்காட்டு *மரங்கள்* (*மரம்*) N + (*கள்*) PL *மரங்கள்* என்பது மரம் மற்றும் *கள்* என்ற இரு உருபன்களாகப் பிரியும் என்பதைத் தெரிந்து கொள்ளும் செயல்பாடுதான் உருபனியல் பகுத்துக் குறித்தல் எனப்படும். ஒன்றை உள்ளீடாக எடுத்துக் கொண்டு ஒரு வித அமைப்பை அதற்குத் தருவதைப் பொதுவாகப் பகுத்துக் குறித்தல் (parsing) எனலாம். ஆட்சி செய்தி (information retriveal) மீட்புப் பரப்பில் (domain) *மரங்கள்* என்பதிலிருந்து மரம் என்பதைப் பொருத்திப் பார்க்கும் (mapping) சிக்கலைப் பகுதியாக்கம் (stemming) என்பார்கள். உருபனியல் பாகுப்பாய்வு செய்தலோ பகுதியாக்கம் செய்தாலோ பன்மை ஒட்டுகளுக்கு மட்டுமன்றி வேறு பல ஒட்டுகளுக்கும் பயன்படும். எடுத்துக்காட்டாக வந்தான் என்பதை வ + ந்த் + ஆன் என்று பகுக்கலாம் வ என்பது வினைப் பகுதி; ந்த் என்பது இறந்ததால ஒட்டு ஆன் என்பது ஆண்பால் ஒருமை வினைமுற்று விகுதி என்று அடையாளம் காணப்படும். நாம் தமிழ்ப் பெயர்களின் எல்லாம் திரிபு மற்றும் ஆக்க வடிவுகளையும் அகராதியல் பட்டியலிட்டுத் தருவதில்லை. அவ்வாறு செய்வது சிக்கனமான முறையல்ல. ஏனென்றால் கள் என்ற பன்மைப் பின் ஒட்டு விளைவாக்கமுள்ளது (productive). இது எல்லாப் பெயர்களுடனும் சேர்ந்து பன்மைப் பெயர் வடிவங்களை உருவாக்கும் தன்மை உடையது. எனவே வினைகள் மற்றும் பெயர்களின் எல்லாத் திரிபு வடிவுகளையும் பட்டியலிடுவது திறமையற்ற செயல் (அதாவது சிக்கனமானதல்ல); மேலும் விளைவாக்கம் உள்ள பின்னொட்டுகள் புதிய சொற்களுடனும் பயன்படுத்தப்படும். (எடுத்துக்காட்டாக கணினி – கணினிகள்) புதிய சொற்கள் நாள்தோறும் உருவாக்கப்படுகின்றன. இவற்றிற்கெல்லாம் பன்மை வடிவுகளைப் பட்டியலிட்டுக் கொண்டே செல்ல இயலாது. மேலும் துருக்கிய மொழி (Turkish language) போன்ற சிக்கலான உருபனியல் அமைப்புள்ள மொழிகளின் சொற்களின் எல்லா வடிவுகளையும் அகராதியில் சேகரித்து வைக்க இயலாது. உருபனியல் குறைந்த பொருள்தருகின்ற அலகுகளான உருபன்களால் சொற்கள் எவ்வாறு உருவாக்கப்படுகிறது என்ற பாடம் உருபனியலாகும். உருபன் பொருண்மையைக் கொண்டிருக்கின்ற குறைந்த மொழி அலகாகும். எடுத்துக்காட்டாக *கால்* என்ற சொல்லில் ஒரு உருபன் உள்ளது. *கால்கள்* என்ற சொல்லில் *கால்*, *கள்* என்ற இரு உருபன்கள் உள்ளன.

உருபன்களைப் பகுதிகள் (stems) ஒட்டுகள் (affxes) என்ற இரண்டு பெரிய உருபுகளாகப் பிரிக்கலாம். இவற்றிற்கு இடையே உள்ள வேறுபாடு மொழிக்கு மொழி வேறுபாடும் உள் உணர்வின் அடிப்படையில் பகுதி என்பது முக்கியப் பொருள் தரும் சொல்லின் முக்கியமான உருபன் என்றும் ஒட்டுகள் என்பது வேறுபட்ட வகையிலான கூடுதல் பொருள் ஊட்டும் உருபன்கள் என்றும் கூறலாம். ஒட்டுகள் முன்னொட்டுகள் (prefixes) பின்னொட்டுகள் (suffixes) இடையொட்டுகள் (infixes) மற்றும் ஓர ஒட்டுகள் (circumfixes) என்று பகுக்கலாம். முன்னொட்டுகள் பகுதிக்கு முன்வரும் பின்னொட்டுகள் பகுதிக்கு பின்னுமாக வரும். இடையொட்டுகள் பகுதிக்குள் சொருகப்படும். எடுத்துக்காட்டாக தமிழில் *மான்கள்* என்ற சொல்லில் வரும் பன்மை ஒட்டான (உருபன்) *கள்* பின்னொட்டாகும். அநீதி முன்னொட்டாகும். அநியாயம் என்ற சொற்களில் வரும் அ என்பது முன்னொட்டாகும். தமிழிலும் ஆங்கிலத்திலும் ஓர ஒட்டுகள் இல்லை. ஜெர்மன் மொழியில் ஓர ஒட்டுகள் உண்டு என்று பார்த்தோம்.

எடுத்துக்காட்டு:

film- en ‘ to film’ – ge-film-t ‘ filemed’

frag-en ‘to ask’ _ ge-frag-t ‘asked’

lob-en ‘to praise’ _ ge-lob-t ‘Praisd’

Zeig-en to ‘Show’ _ ge-zeig-en ‘Shown’

ge-யும் to-யும் இணைந்து Past Participle (இறந்தகால எச்சம்) வடிவை உருவாக்கும். பிலிப்பைன் மொழியான டாகலாக் (Tagalog) வியட்நாம் மொழியான சரவ் (Charau) போன்ற மொழிகளில் இடை ஒட்டுகள் பயன்படுத்தப்படுகின்றன.

சரவ் மொழியிலிருந்து எடுத்துக்காட்டு:

sulat ‘Write’ – s-um-ulat ‘wrote’

s- in-ulat ‘written’

டாகலாக் மொழியிலிருந்து எடுத்துக்காட்டு:

hingi ‘borrow’ – h-um-ingi

ஒரு சொல் பல உருபன்களின் பிணைக்கப்பட்டுக் உருவாக்கப்படுவதால் முன் ஒட்டுகளும், பின்னொட்டுகளும் பிணைப்பு உருபனியல் (concatenative morphology) எனப்படும். பல வழிகளில் உருபன்கள் மிகச் சிக்கலான வழியின் இணையும் பிணைப்பில்லாத உருபனியல் (non-concutenative morphology) பரவலாக இருக்கும். மேற்சொன்ன டாகலாக் மொழி பிணைப்பு

உருபனியலுக்கு எடுத்துக்காட்டாகும். வேறு ஒருவகையான பிணைப்பில்லாத உருபனியல் வார்ப்புச் சட்ட உருபனியல் (templatic morphology) அல்லது வேர் அமைப்பொழுங்கு உருபனியல் (root-pattern morphology) எனப்படும். அரேபிய மொழி (Arabic Language) ஹீப்ரூ மொழி பிற செமிட்டிக் மொழிகளில் (Semitic Language) சாதாரணமாக வார்ப்புச் சட்ட உருபனியல் காணப்படுகிறது. ஹீப்ரூ மொழியில் வினை இரண்டு உட்கூறுகளால் உருவாக்கப்படுகின்றது. ஒன்று அடிப்படைப் பொருளைக் கொண்டிருக்கும் மூன்று மெய்களால் (CCC) ஆன வேர் (வார்புச் சட்டம் இரண்டு பொதுவாக வினைப்பாட்டுப் (voice) பொருளை வெளிப்படுத்தும் உயிர்கள் எடுத்துக்காட்டாக மூன்று மெய்யால் ஆன 'learn' அல்லது 'study' என்று பொருள் தருகின்ற Lnad என்ற வேர் செய்வினையைக் (active voice) குறிக்கும் CaCaC என்ற வார்புச் சட்டத்துடன் சேர்ந்து lamad 'he studied' 'அவன் படித்தான்' என்ற சொல்லை உருவாக்கும் அல்லது வலியுறுத்தலைக் குறிக்கும். CiCeC என்ற வார்புச் சட்டத்துடன் சேர்ந்து limed 'he taught' 'அவன் கற்பித்தான்' என்ற சொல்லை உருவாக்கும் அல்லது வலியுறுத்தல் செயப்பாட்டு வினையைக் குறிக்கும். CuCaC என்ற வார்புச் சட்டத்துடன் சேர்ந்து lumad 'he was taught' அவன் கற்பிக்கப்பட்டான் என்ற சொல்லை உருவாக்கும். ஒரு சொல்லில் ஒன்றுக்கு மேற்பட்ட ஒட்டுக்கள் வரலாம். எடுத்துக்காட்டாக பானைகளை என்ற சொல்லில் பானை என்ற பகுதியும் கள் என்ற பன்மை விகுதியும் ஐ என்ற 2-ஆம் வேற்றுமை உருபும் உள்ளன. rewrites என்ற சொல்லல் re முன்னொட்டும் write என்ற பகுதியும் s என்ற பின்னொட்டும் உள்ளன. unbelievably என்ற ஆங்கிலச் சொல்லில் un என்ற முன்னொட்டும் believe என்ற பகுதியும் able, ly, என்ற பின்னொட்டுக்களும் உள்ளன. தமிழ் போன்ற மொழிகளில் மூன்றோ நான்கோ ஒட்டுக்களை கொண்டிருக்கும். ஆனால் துருக்கிய மொழி போன்ற மொழிகளில் சொற்கள் ஒன்பதோ பத்தோ ஒட்டுக்களைக் கொண்டிருக்கும். இம்மாதிரியான மொழிகள் ஒட்டுக்கள் நிறைந்த மொழிகள் (agglutinative language) என்று அழைக்கப்படும். உருபன்களிலிருந்து சொற்கள் உருவாக்கப்படுவதை இரண்டு பெரும் வகுப்புகளாகப் பகுக்கலாம். அவைகள் திரிபுகள் (inflection) ஆக்கங்கள் (derivation) என்பன. ஒரு சொல் பகுதி ஒரு இலக்கண உருபனுடன் இணைந்து அதே சொல்வகுப்பைச் சேர்ந்த சொல்வடிவை விளைவிப்பது திரிபு எனப்படும். எடுத்துக்காட்டாக பன்மை ஒட்டான கள் ஒருமைப் பெயர்ப்பகுதியுடன் இணைந்து பன்மை இலக்கண உருபனுடன் இணைந்து பொதுவாக வேறு ஒரு வகுப்பைச் சார்ந்த பெரும்பாலும் பொருளைச் சிக்கலான சொல்வடிவை உருவாக்குவது ஆக்கம் (derivation) எனப்படும்.

முற்றுநிலை உருபனியல் பகுத்துக்குறித்தல் (Finite State Morphological Parsing)

முற்றுநிலைத் தானியங்யின் உருபனியல் பகுத்திக்குறித்தலைப் பின்வரும் எடுத்துக்காட்டு மூலம் விளக்கலாம். நமது நோக்கம் முதல் செங்குத்துப் பத்தியில் கொடுத்துள்ள உள்ளீட்டுச் சொற்களுக்கு இரண்டாவது பத்தியில் தரப்பட்டுள்ள வெளியீட்டைப் பெறுவதாகும்.

உள்ளீடு	**உருபனியல் பகுப்பாய்வு வெளியீடு**
கால்கள்	கால் + பெயர் + பன்மை
கால்	கால் + பெயர் + ஒருமை
கடாக்கள்	கடா + பெயர் + பன்மை
கடா	கடா +பெயர் + ஒருமை
மரங்கள்	மரம் + பெயர் + பன்மை
மரம்	மரம்+பெயர் + ஒருமை
கற்கள்	கல் + பெயர் + பன்மை
கல்	கல்+ பெயர் + ஒருமை
முட்கள்	முள் + பெயர் + பன்மை
முள்	முள்+ பெயர் + ஒருமை

முதல் குத்துநிலை வரிசையில் தரப்பட்ட வடிவங்களுக்கு இரண்டாவது குத்துநிலை வரிசையில் தரப்பட்டவாறு வெளியீட்டைப் பெறுதல் நமது நோக்கமாகும். இரண்டாவது வரிசையில் ஒவ்வொரு சொல்லின் பகுதியும் உருபனியல் பண்பு கூறுகளாகும். எடுத்துக்காட்டாக +பெயர் என்பது அந்த சொல் ஒரு பெயர் என்பதையும் +ஒருமை என்பது ஒருமை எண் என்பதையும் +பன்மை என்பது பன்மை எண் என்பதையும் குறிப்பிடும்.

உருபனியல் பகுப்பான் (Morphological Parser) உருவாக்குவதற்கு பின்வருவன தேவை.

1. அகராதி (Lexicon): பகுதிகள் மற்றும் ஒட்டுகள் இவற்றின் அடிப்படைச் செய்திகளுடன் கூடிய (அதாவது பத்தி பெயரா வினையா போன்ற செய்திகள்) அடங்கிய பட்டியல்

2. உருபன் அமைப்பொழுங்கு (Morphotactics): ஒரு சொல்லுக்குள்ள எந்த வகையைச் சார்ந்த உருபன்கள் பிற வகுப்புகளைத் தொடர்ந்து வரும் என்பதை விளக்குகின்றது. உருபன் வருகை முறையின் மாதிரி எடுத்துக்காட்டாக தமிழில் பன்மை உருபு பகுதியைத் தொடர்ந்து வரும். பகுதிக்கு முன்னர் வராது போன்ற செய்திகளைக் கூறலாம்.

மரம் + கள் +ஐ

* கள் - மரம் - ஐ

* ஐ – கள் - மரம்

* கள் - ஐ – மரம்

*மரம் - ஐ – கள்

[* தவறான உருபன் அமைப்பொழுங்கைக் குறிக்கின்றது]

3. சந்திவிதிகள்/உருபஒலியனியல் விதிகள் அல்லது எழுத்தலகு விதிகள் (morphophonemic Rules அல்லது orthographic rules) தமிழில் சந்தி விதிகள் சாதாரணமாய் இரு உருபன்கள் இணைகையில் ஏற்படும் மாற்றங்களை மாதிரிப்படுத்ததப் பயன்படுகின்றன. ஆங்கிலம் போன்ற மொழிகளில் எழுத்துக்கூட்டல் (spellers) விதிகள் சாதாரணமாய் இரு உருபன்கள் இணைகையில் ஏற்படும் மாற்றங்களை மாதிரிப்படுத்தும்.

மரம் + பன்மை > மரங்கள்

மரம் + ஐ > மரத்தை

city + s > cities

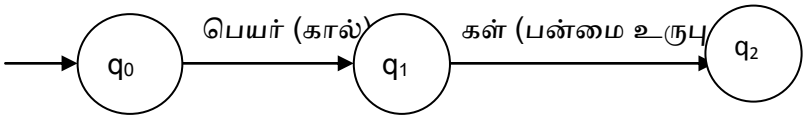
box + s > boxes

உருபன் அமைப்பொழுங்கு பற்றிய அறிவை மாதிரிப்படுத்த எவ்வாறு முற்றுநிலைத் தானியங்கியைப் பயன்படுத்துவது உருபன்களைத் தெரிந்து கொள்ளலின் துணைச் சிக்கல்களை அகராதியில் எவ்வாறு எளிமையாக உருப்படுத்தம் செய்வது என்பது அடுத்து எதிர் கொள்ள வேண்டிய பிரச்சனைகள் ஆகும். இதற்கு நாம் முற்றுநிலை மாற்றிகள் (Finite State Transducers) அகராதியில் உருபனியல் பண்புக் கூறுகளை மாதிரிப்படுத்த அறிமுகப்படுத்த வேண்டி வரும். இறுதியில் எவ்வாறு எழுத்தலகு விதிகளை அல்லது உருபொலியனியல் விதிகளை மாதிரிப்படுத்த முற்றுநிலை மாறிகளை பயன்படுத்தலாம் என்பதையும் தர வேண்டும்.

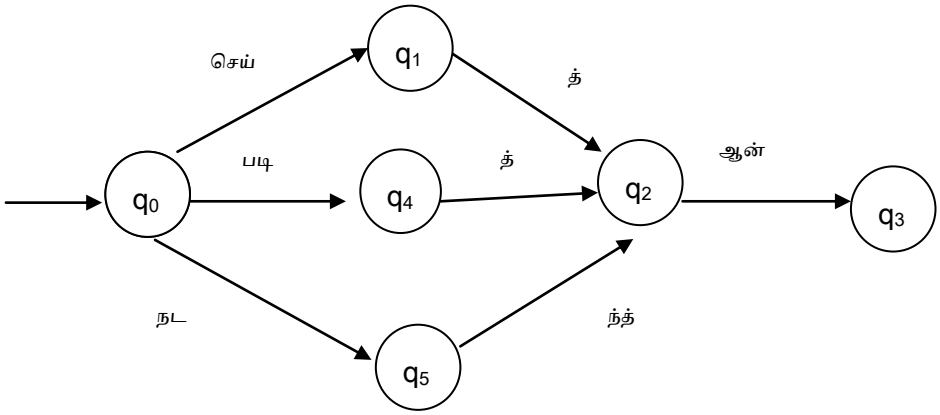
அகராதியும் உருபன் அமைப்பொழுபங்கும் (Lexicon and Morphotactics)

அகராதி சொற்களை வைப்பதற்கான ஒரு களஞ்சியம் எளிய அகராதி மொழியின் எல்லா சொற்களையும் இயற்பெயர்களையும் கொண்டிருக்கும். ஒரு மொழியில் உள்ள எல்லாச் சொற்களையும் பட்டியலிடுவது வசதியானதாகவோ இயலாமலோ இருப்பதால் கணினி அகராதி சாதாரணமாக மொழியின் ஒவ்வொரு பகுதியையும் ஒட்டுகளையும் அவை எவ்வாறு இணையும் என்ற செய்தியைத் தருகின்ற உருபன் அமைப்பொழுங்கின் உருப்படுத்தத்துடன் பட்டியலிட்டு கட்டமைப்பு செய்து தரும். உருபன் அமைப்பொழுங்கை பல வகைகளில் மாதிரிப்படுத்தலாம்.

மிகப் பொதுவான ஒரு மாதிரி முற்றுநிலைத் தானியங்கியாகும். பின் வருவது தமிழ்ப் பன்மை திரிபின் முற்றுநிலை மாதிரியாகும்.



செய்-வகுப்பு வினைகள்	படி-வகுப்பு வினைகள்	நட-வகுப்பு வினைகள்	செய்-வகுப்பு வினைகளின் இறந்தகால ஒட்டு	படி-வகுப்பு வினைகளின் இறந்தகால ஒட்டு	நட-வகுப்பு வினைகளின் இறந்தகால ஒட்டு	படர்க்கை ஆண்பால் ஒட்டு
செய் பெய்	படி பிடி குடி	மற நட அள	த்	த்த்	ந்த்	ஆன்



முற்றுநிலைத் தானியங்கியை பயன்படுத்தி உருபனியல் தெரிந்துகொள்வதின் சிக்கலை தீர்க்கலாம். அதாவது உள்ளீடு செய்கின்ற எழுத்து சரியான தமிழ்ச் சொல்லா இல்லையா என்று உறுதிசெய்து கொள்ளலாம்.

முற்றுநிலை மாறிகளின் உதவியால் உருபனியல் பகுப்பாய்வு

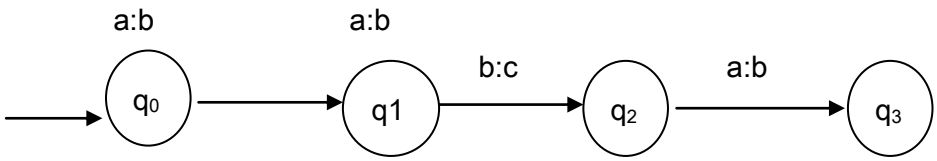
கொஸ்கெனிமி (Koskenniemi) எடுத்துச் சொன்ன இருநிலை உருபனியல் (Two Level Morphological) உருபனியல் பகுப்பாய்வைச் செய்யலாம். இருநிலை உருபனியல் ஒரு சொல்லை உருவாக்குகின்ற உருபன்களின் எளிய இணைப்பை உருப்படுத்தம் செய்யும் சொல்சார் நிலைக்கும் (Lexical Level) (அக அமைப்புக்கும்) இறுதிச் சொல் வடிவின் உண்மையான எழுத்துக் கூட்டலை உருப்படுத்தம் செய்யும் புற அமைப்புக்கும் இடையே உள்ள பொருத்தத்தை உருப்படுத்தம் செய்கின்றது. எழுத்துக் கோர்வைகளால் ஆன மரங்கள் போன்ற புற அமைப்பு நிலையை *மரம்* + பெயர் + பன்மை என்ற உருபன் பண்புக்கூறுகளால் ஆன தொடர்ச்சியாக சொல்சார் நிலையை (அக அமைப்பை) அறிவது உருபனியல் பகுப்பாய்வாகும்.

சொல்நிலை		வ	ண்ண	டு	N	PL	

புறநிலை		வ	ண்ண	டு	க	ள்	

சொன்மை மற்றும் புற நாடாக்களின் எடுத்துக்காட்டு இவ்வாறு இருநிலைக்களுக்கிடையே உருபடுத்தத்தைச் செயல்படுத்த நாம் பயன்படுத்தும் தானியங்கி முற்றுநிலை மாற்றியாகும் (FST) மாற்றிகள் ஒரு குமுமக்குறியீடுகளுக்கும் மற்றொரு குழுமக்குறியீடுகளுக்கும் இடையே உருப்படுத்தம் செய்கின்றது. முற்றுநிலை மாற்றி இதை முற்றுநிலைத் தானியங்கி வழியாகச் செய்கின்றது. இவ்வாறு நாம் ஒரு முற்றுநிலை மாற்றியைக் கோர்வைகளின் இணைகளை தெரிந்துகொள்கின்ற அல்லது உருவாக்குகின்ற இரு நாடா முற்றுநிலை தானியங்கியைக் காணலாம். இவ்வாறு முற்றுநிலை மாற்றிக்கு முற்றுநிலை தானியங்கியை விட கூடுதல் பொதுமைச் செயால்பாடு இருக்கின்றது. முற்றுநிலை தானியங்கி ஒரு முறையான மொழியைக் கோர்வைகளின் ஒரு குழுமமாக விவரணை செய்கின்றது. ஆனால் முற்றுநிலை மாற்றி ஒரு முறையான மொழியைக் கோர்வைகளின் குழுமங்களுக்கு இடையே உள்ள உறவாக விவரணை செய்கின்றது. இது முற்றுநிலை மாறியை ஒரு கோர்வையை உருவாக்கும் ஒரு பொறியாகக் காட்டுகின்றது. ஒரு உள்ளீட்டாலும் வெளிட்டாலும் ஒவ்வொரு வில்லும்

அடையாளப்படுத்தப்பட்ட முற்றுநிலைப்பொறி முற்றுநிலை மாற்றியாகும். ஒரு தொடர்பை முற்றுநிலைமாற்றியாக உருப்படுத்தம் செய்யலாம்.



இணைகளின் ஒரு முற்றுநிலை அல்லது எச்சநிலைக் குழுமம் ஒரு உறவாகும்.

R = (1,1), (2,4), (3,9), (4,16), (5,25)

S = (a, A), (b, B), (c, C), (d, D), (e, E)….

U = (கேள்வி i, பதில் i)

V = (ராமன் வீட்டுக்குப் போய்க்கொண்டிருக்கிறான் Raman is going home)

W = (a,b), (aaa,bbb), (aaaa,bbbb).....

ஒரு உறவை பார்வை அட்டவணையாக (look up table) (உள்ளீடு, வெளியீடு) உருப்படுத்தம் செய்யலாம். அந்த அட்டவணை எல்லையற்ற அளவு பெரிதாய் இருக்கலாம்; பின் தரப்பட்டுள்ளது மாற்றிகளின் நான்கு மடங்கு வழிகளின் சுருக்கமாகும்.

1. அறிவானாக முற்றுநிலைமாற்றி (FST as recognizers)

கோர்வைகளின் ஜோடியை/இணையை உள்ளீடாகவும் வெளியீடாகவும் எடுத்துக்கொண்டு கோர்வை இணைமொழியின் கோர்வை இணையாக இருந்தால் ஏற்றுக்கொள்ளவும் இல்லாவிட்டால் விட்டுவிடவும் செய்யும் மாற்றி முற்றுநிலைமாற்றி அறிவானாகும்.

2. உருவாக்கியாக முற்றுநிலைமாற்றி (FST as generator)

மொழியின் கோர்வைகளின் இணைகளை வெளியீடு செய்கிற பொறி முற்றுநிலை மாற்றி உருவாக்கியாகும் இவ்வாறு வெளியீடு “ஆம்“ அல்லது ”இல்லை” என்பாதம் மற்றும் வெளியீட்டுக் கோர்வைகளின் இணையாகும்.

3. மொழிபெயர்ப்பியாக முற்றுநிலை மாற்றி (FST as translator)

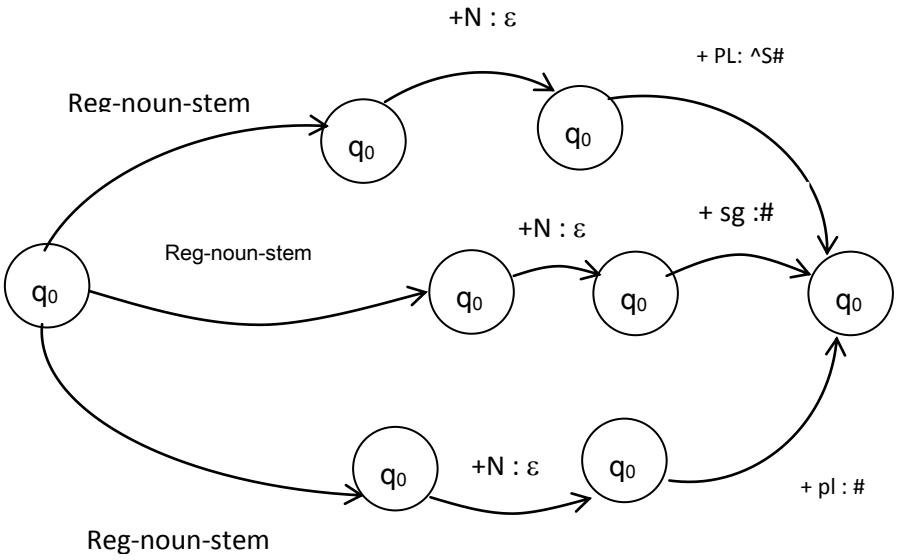
ஒரு கோர்வையைப் படித்து மற்றொரு கோர்வையை வெளியீடு செய்கிற பொறி முற்றுநிலைமாற்றி மொழிபெயர்ப்பியாகும்.

4. உறவுபடுத்தியாக முற்றுநிலைமாற்றி (FST as realtor)

குழுமங்களுக்கு இடையே உறவுகளைக் கணிக்கிற பொறி முற்றுநிலைமாற்றி உறவுப்படுத்தியாகும். எங்கு முற்றுநிலைத்தானியங்கிகள் முறையான மொழிகளுடன் சம உருவம் கொண்டிருக்கிறதோ அங்கு முற்றுநிலைமாற்றிகள் முறையான உறவுகளுடன் சமஉருவம் கொண்டிருக்கும். கோர்வைகளின் குழுமங்களாக இருக்கிற முறையான மொழிகளின் இயற்கையின் நீ;ட்சியான கோர்வைகளின் இணைகளின் குழுமங்கள் முறையான உறவுகளாகும். எ.கா ஆங்கில பன்மைத் திரிபு அகராதியில் கலவை நிலையில் அமையும்.

reg – noun	irreg – pl – noun	irreg –sg -noun
fox	g o:e o:e s e	goose
cat	sheep	sheep
dog	m o:i u:ɛ s:c e	mouse

reg – noun	irreg – pl – noun	irreg –sg -noun
fox	g o:e o:e s e	goose
cat	sheep	sheep
dog	m o:i u:ɛ s:c e	mouse



ஆங்கில பெயர்திரிபின் மாற்றி q1, q2 என்பன ஏற்றுக்கொள்ளும் நிலைகள் ஆகையால், சீரான/முறையான பெயர்கள் பன்மைப் பின்னொட்டு பெறவோ பெறாமலோ இருக்கலாம். ^ என்பது உருபன் - எல்லைக் குறியீடு: · என்பது சொல் எல்லைக் குறியீடு.

எழுத்தலகு விதிகளும் முற்று நிலை மாற்றிகளும்

ஆங்கிலத்திற்கு உருபன் எல்லையில் எழுத்துக் கூட்டல் மாற்றங்கள் தேவை. அதற்காக நாம் எழுத்துக் கூட்டல் விதிகள் அல்லது எழுத்தலகு விதிகளை அறிமுகம் செய்ய வேண்டிவரும். நாம் எழுத்துக் கூட்டல் மாற்றங்களை எளிய உருபன்களின் இணைப்பின் உள்ளீட்டாகவும் கொஞ்சம் மாறுபட்ட உருபன்களின் இணைப்பின் வெளியீடாகவும் எடுத்துக்கொள்ளலாம். பின்வருவது மூன்று நிலைகளைக் காட்டுகின்றது: சொல் நிலை, இடைநிலை, புறநிலை.

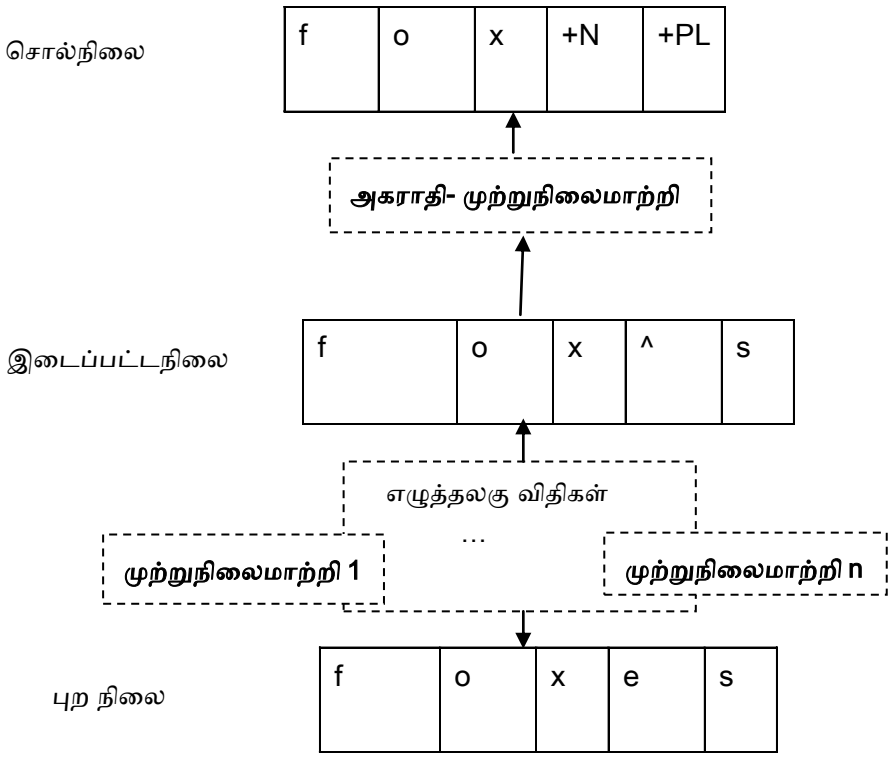
சொல்நிலை	f	o	x	+N	+pl

இடைப்பட்டநிலை	f	o	x	^	s	#

புற நிலை	f	o	x	e	s	#

முற்றுநிலை மாற்றியின் அகராதியையும் விதிகளையும் சேர்த்தல் (Combining FST Lexicon and rules)

நாம் பகுத்துக்குறிக்கவும் உருவாக்கவும் அகராதியையும் விதி மாற்றிகளையும் இணைக்க தயாராகிவிட்டோம். பின்வரும் படம் பகுத்தாயவோ உருவாக்கவோ பயன்படும் இருநிலை உருபன் ஒழுங்குமுறையின் அமைவைக் காட்டுகின்றது.



அகராதி மாற்றி அதன் பகுதிகள் மற்றும் உருபன் பண்புக்கூறுகள் இவற்றைக் கொண்டு சொல் நிலைக்கும் உருபன்களை இணைவதை உருப்படுத்தம் செய்யும் இடைநிலைக்கும் இடையே பொருத்தம் செய்கின்றது. ஒவ்வொரு தனி எழுத்துக் கூட்டல் விதிக்கட்டுப்பாட்டையும் உருப்படுத்தம் செய்கின்ற ஒரு கூட்டம் மாற்றிகள் இடைநிலைக்கும் புறநிலைக்கும் இடையே பொருத்தம் செய்கின்றது.

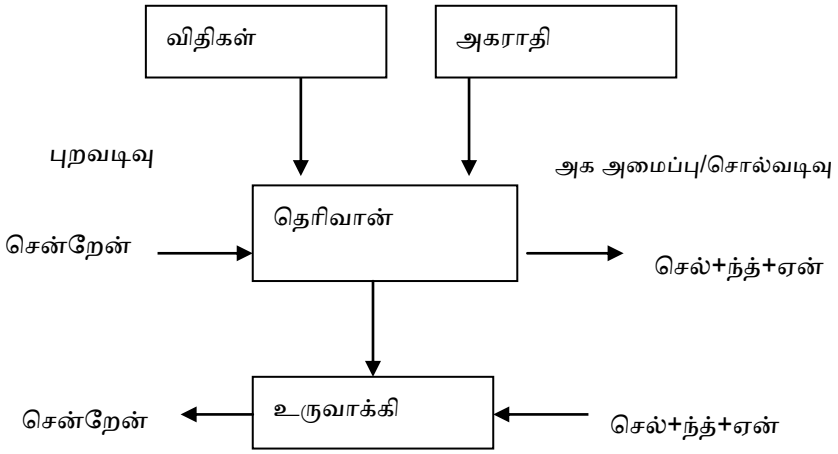
பிசிகிமோ (Pckimmo)

கொசொகென்னிமியின் (Kosekenniemi) இருநிலை உருபனியல் பரவலாகப்; பயன்படுத்தப்படுகின்ற குறிப்பாக ஐரோப்பிய மொழிகளில் பயன்படுத்தப்படுகின்ற ஒன்று. பிசிகிமோ உருபனியல் ஆய்வியாக இரண்டு செயல்களைச் செய்ய இயலும்.

1. தெரிந்துக்கொள்ளுதல் (recognition)
2. உருவாக்குதல் (generation)

பிசிகிமோவின் உருவாக்கும் பகுதி வேரின் சொல்வடிவை உள்ளீடாக ஏற்கும். தெரிவான்/உணரி புற அமைப்பை உள்ளீடாக ஏற்கும் விதிகளை பயன்படுத்தும்: அகரதியைப் பார்க்கும்

தொடர்புள்ள சொல்வடிவையும் பொருள் கோர்வையும் தரும். இதைப் பயன்படுத்துவதில் உள்ள எளிமையும் இம்மாதிரியின் சிக்கனம் அடிப்படையிலான முக்கியத்துவமும் என்னவென்றால் ஒரே குழும விதிகளை தெரிந்துகொள்ளும் செயல்பாட்டிற்கும் உருவாக்கும் செயல்பாட்டிற்கும் பயன்படுத்தலாம். விதிகளை இருதிசைகளிலும் பயன்படுத்தலாம். சில ஒலியன் விதிகளோ உருபன் விதிகளோ எழுத்தலகு விதிகளோ புற அமைப்பிலிருந்து, அக அமைப்பைத் தரும் மாற்றமாகக் கொள்ளப்பட்டால் அவற்றைத் திருப்பிச் செயல்பாட்டிற்காக (சுநஎநசளந pசடிஉநளள) மீண்டும் எழுதத் தேவையில்லை. திரும்புகையில் அது புற அமைப்பிலிருந்து அகஅமைப்பை ஆய்கிறது. இதைப் பின்வரும் படத்தில் காட்டலாம்.



இம்மாதிரியில் பயன்படுத்தப்படும் இருநிலைகள் எல்லாம் மொழியியல் விவரணைகளாகும். இந்த விவரணைகள் மாற்றொலியன் பிறழ்ச்சிகள் (allophonic alternations) உருப ஒலியன் பிறழ்ச்சிகள் (morphophonemic alternations)அல்லது எழுத்தலகு வேறுபாடுகள் (orthographic variations) இவற்றை கையாள இயலும். பிசிகிமோ இருநிலை மொழியியல் விவரணைகளை முற்றுநிலை அட்டவணைகளாக மாற்றும். விரிவாகக்கூறப் புகுந்தால் விதிக்கோப்பில் விதிகள் முற்றுநிலை மாற்றிகளாக (FST) உருபடுத்தம் செய்யப்படும் உருபன் அமைப்பொழுங்குக் கட்டுப்பாடு (Morphotatic Constraints) அகராதிக் கோப்பில் முற்றநிலைத் தானியங்கியாகக் (FST)கூறப்படும். அகராதிப் பகுதி (Lexical Component)

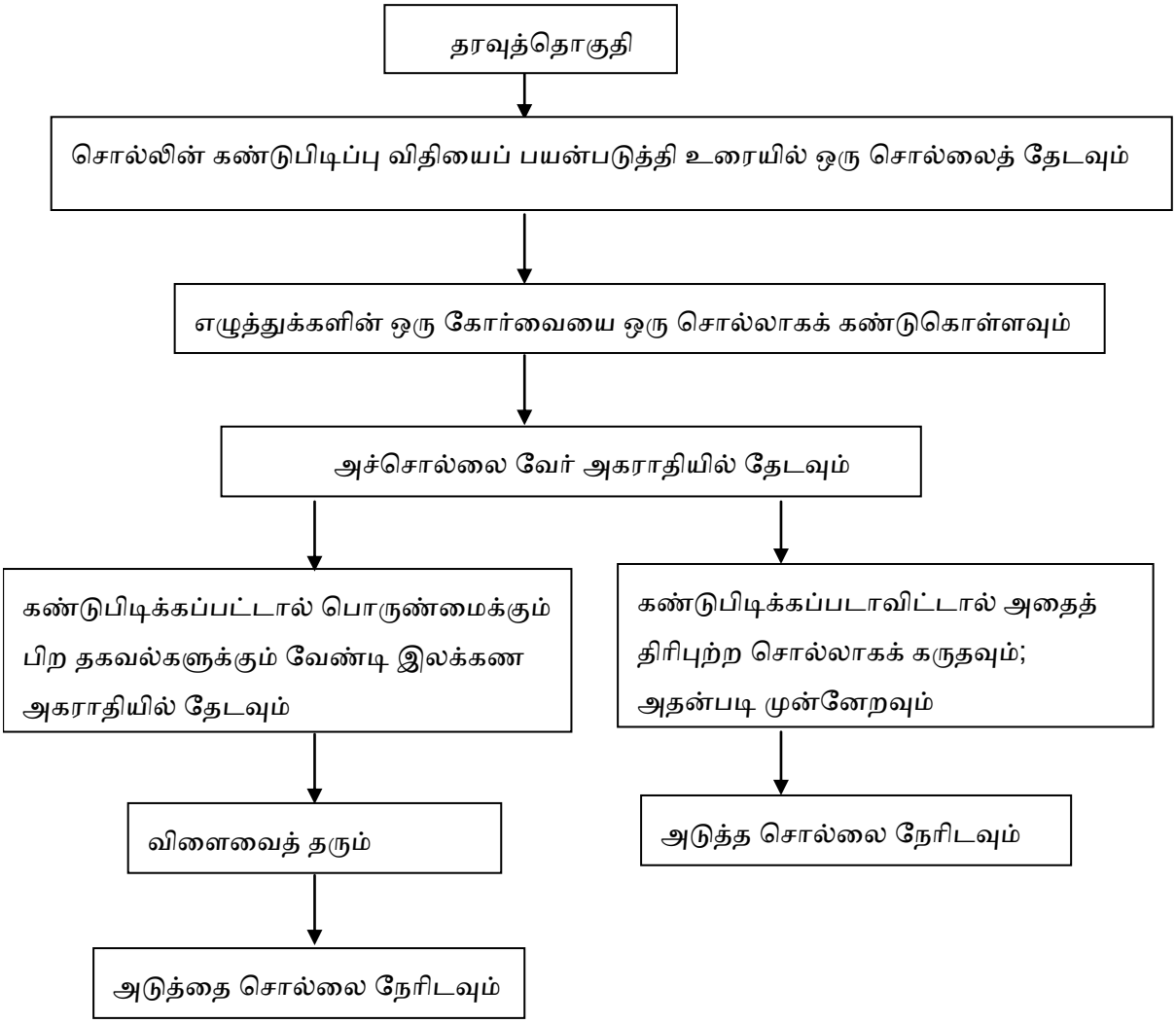
பிசிகிமோவின் அகராதிப் பகுதி தெரிந்துகொள்ளும் செயல்பாட்டின் போது மட்டும்தான் நோக்கப்படும்; உருவாக்கும் செயல்பாட்டில் நோக்கப்படமாட்டாது. அகராதி ஒரு சொல்லின் உருப்படுத்தத்தைக் காட்டும் அதன் உருபன் அமைப்பொழுங்குக் கட்டுப்பாட்டைக் குறிப்பிடும்.

ஒவ்வொரு சொல்லிற்குப் பொருள் தரும் பிசிகிமோவின் ஒவ்வொரு அகராதியும் துணை அகராதிகளாகப் பகுக்கப்பட்டிருக்கும். ஒவ்வொரு துணை அகராதியும் உருபன் அமைப்பொழுங்கில் ஒரே மாதிரி செயலாற்றுகிற ஒரு வகுப்பைச் சார்;ந்த சொற்களை உருப்படுத்தம் செய்யும். ஒவ்வொரு சொற்பதிவும் (lexical entry) ஒரு சொல்லையும் அதன் தொடர்ச்சி வகுப்பையும் அதன் பொருளையும் கொண்டிருக்கும். தொடர்ச்சி வகுப்பு அதைத் தொடர்ந்து வரும் வேறுபட்ட உருபன்களையும் இலக்கணத் தன்மையின் உருபன் அமைப்பொழுங்கு விதிகள் எங்கு கூறப்பட வேண்டும் என்பதைக் காட்டும். ஒரு சொல்லைத் தொடர்ந்து வரும் துணை அகராதிக் குழுவை விவரிக்கும் பிறழ்ச்சியின் பட்டியல் இருக்கும். எடுத்துக்காட்டாக, *கால்* என்ற சொல்லை எடுத்துக் கொள்வோம் அது *ஐ* என்ற இரண்டாம் வேற்றுமை உருபையும் *உக்கு* என்ற நான்காம் வேற்றுமை உருபையும் இல் என்ற இட வேற்றுமை உருருபையும் இலிருந்து என்ற நீக்கல் வேற்றுமை உருபையும் கள் என்ற பன்மை உருபையும் ஆ என்ற வினா பின்னொட்டையும் ஏற்கும். சேய்மை கட்டு அடையில் *அ* என்பது அந்த என்பதன் மாற்று உருபாகும்.

மேற்சொன்ன தமிழ்ச் சொல்லாய்விகள் யாவும் விதி அடிப்படையிலானவை. பின்வருவன தரவுத்தொகுதி அடிப்படையிலானவை.

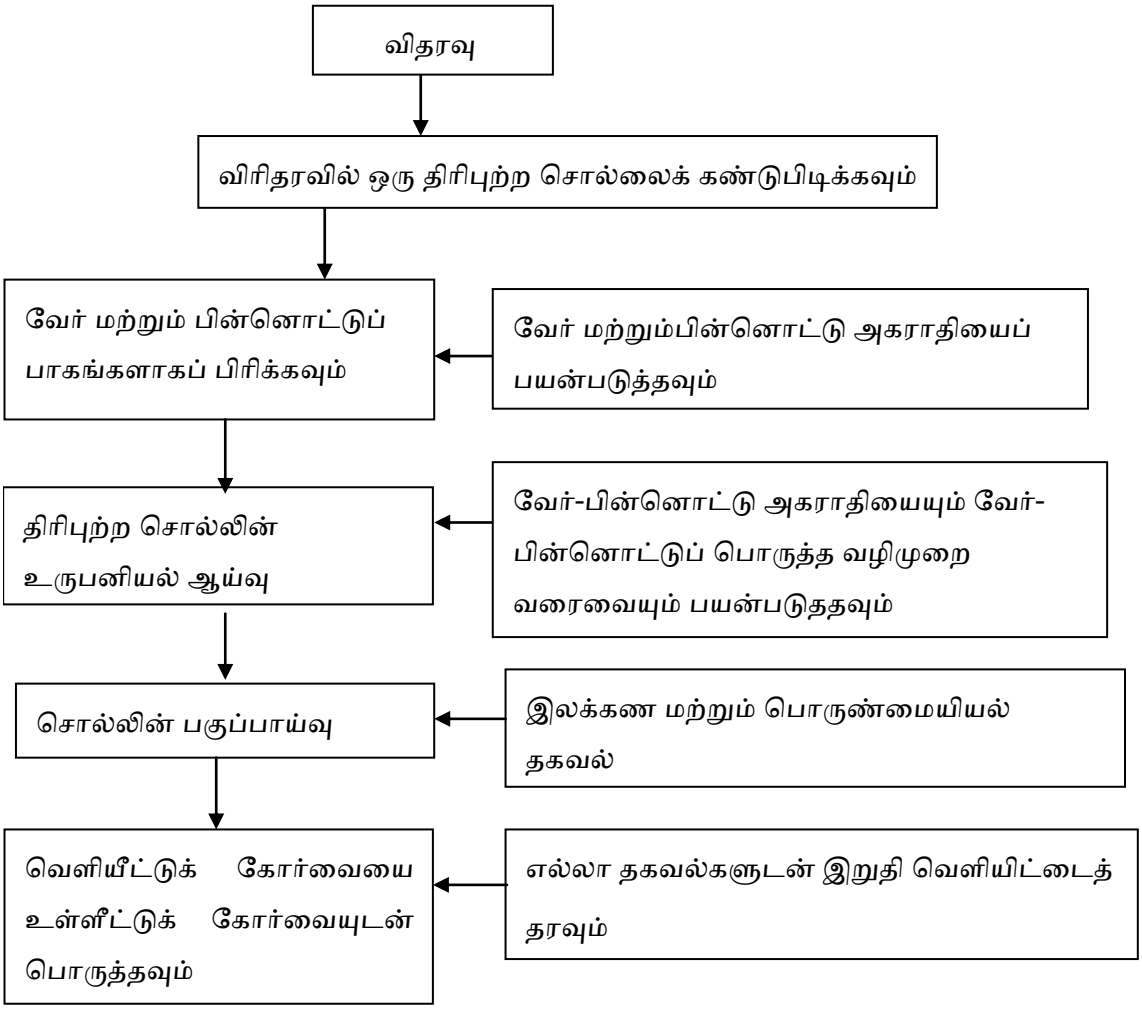
5.9.1. திரிபுறாத சொற்களின் ஆய்வு

பின்வரும் ஒழுக்குப்படம் திரிபுறாத சொற்களின் பகுப்பாய்வை (processing non-inflected forms) விளக்கும்:



5. 9.2. திரிபுற்ற சொற்களைப் பகுத்தாய்தல்

பின்வரும் ஒழுக்குப்படம் திரிபுற்ற சொற்களின் பகுப்பாய்வை (processing inflected words) விளக்கும்:



5. 9.3. இரட்டைச் சொற்களைப் பகுத்தாய்தல் (Processing Double Words)

இரட்டைச் சொற்களின் ஆய்வு உறுப்புகள் அவற்றிற்கு இடையில் ஒரு இடைவெளியால் பிரிக்கப்பட்ட கூட்டுகள், இரட்டுற்ற சொற்கள், வேறுபடுத்தப்பட்ட சொற்கள் என்பனவற்றை உள்ளடக்கும். எல்லா வேறுபடுத்தப்பட்ட சொற்களும் பல்சொல் கோர்வைகளாகும்· அவை மிகத்திறமையுடைய ஆய்வு மற்றும் குறிப்புரை (annotation) வழி கையாளப்படவேண்டும். இரட்டை சொற்களை விரிதரவில் கண்டுபிடித்து ஆய்வது சிக்கலான செயல்முறையாகும். அகராதிகளில் பெரும்பாலும் இத்தகைய இரட்டை சொற்கள் தரப்பட்டிருக்கும். அத்தகைய சொற்களைப் பொருத்தமான அணுகுமுறையால் விரிதரவில் அடையாளம் கண்டு கொள்ளலாம்.

ஆனால் அகராதியில் தரப்படாத இரட்டைச் சொற்களை அடையாளம் காண்பது சிக்கலான செயல் முறையாகும். சேர்ந்து வருகை நிகழ்வெண்கள் இதற்குக் கைகொடுக்கும். அடுத்துவருகின்ற சொற்களின் பொருண்மை இரட்டைச் சொற்களின் பொருண்மையை உறுதிசெய்ய உதவுவதால் இரட்டை சொற்களை ஆய்தல் சொல் மட்டத்தில் பொருண்மை மயக்கம் தீர்க்க உதவும்.

5.10. அடையாயப்படுத்துதல் (Tagging)

குறிப்பிட்ட பண்புக்கூறுகளைக் குறிப்பிட வேண்டி சொற்களுக்குத் தனிப்பட்ட குறியங்களை ஒட்டுவதை உள்ளடக்கும் மொழிக் குறிப்புரையின் ((linguistic annotation) சில வகைகள் குறிப்புரை, குறியங்கள் என்று அறியப்படுவதற்குப் பதிலாக அடையாளப்படுத்துதல் என்று அறியப்படுகின்றது.

5.10.1. சொல் வகைப்பாட்டு அடையாளப்படுத்தல் (Part-of-speech (POS) tagging)

தரவுத்தொகுதி மொழியியலில் சொல்வகைப்பாடு அடையாளப்படுத்தல் (part-of-speech tagging அல்லது POS tagging அல்லது PoS tagging அல்லது POST) இலக்கணம்சார் அடையாளப்படுத்தல் (grammatical tagging) என்றும் அழைக்கப்படுகிறது, இது ஒரு உரையில் (தரவுத்தொகுதி) ஒரு சொல்லை அதன் வரையறை மற்றும் சூழல் இரண்டையும் அடிப்படையாகக் கொண்டு சொல்வகைப்பட்டுக்கு ஒத்ததாகக் அடையாளப்படுத்தும் செயல்முறையாகும். சொற்களைப் பெயர்ச்சொற்கள், வினைச்சொற்கள், பெயரடைகள், வினையடைகள் என அடையாளம் காணும் இதன் எளிமையான வடிவம் பொதுவாகப் பள்ளி வயது குழந்தைகளுக்குக் கற்பிக்கப்படுகிறது.

முன்னர் கையால் செய்யப்பட்ட சொல்வகை அடையாளப்படுத்தல் தற்போது தனித்துவமான சொற்களையும், பேச்சின் மறைக்கப்பட்ட பகுதிகளையும் இணைக்கும் வழிமுறைகளைப் பயன்படுத்தி விளக்கக் சொல்வகை அடையாளங்களின் தொகுப்பால் இப்போது கணினி மொழியியலின் சூழலில் செய்யப்படுகிறது. சொல்வகை அடையாளப்படுத்தலின் வழிமுறைகள் இரண்டு தனித்துவமான குழுக்களாகின்றன: ஒன்று விதி அடிப்படையிலானது மற்றொன்று புள்ளியியல் அடிப்படையானது. முதன்மையானவற்றில் ஒன்றான மற்றும் மிகவும் பரவலாகப் பயன்படுத்தப்படும் ஆங்கில சொல்வகை அடையாளப்பத்திகளில் ஒன்றான ஈ. பிரில்லின் டேக்கர் (E. Brill's tagger) விதி அடிப்படையிலான வழிமுறைகளைப் பயன்படுத்துகிறது.

கொள்கை

சொல்வகை அடையாளப்படுத்தல் சொற்களின் பட்டியலையும் அவற்றின் சொல்வகைப்பாடுகளை வைத்திருப்பதை விட கடினமானது. ஏனெனில் சில சொற்கள் வெவ்வேறு நேரங்களில் ஒன்றுக்கு மேற்பட்ட சொல்வகைப்பாடுகளைக் குறிக்கக்கூடும்; மேலும் சொல்வைப்பாடுகள் சிக்கலானவையாக அல்லது சரியாக வெளிப்படுத்தப்படாதவையாக இருக்கும். இந்நிலை அரிதானது அல்ல; இயற்கை மொழிகளில் (பல செயற்கை மொழிகளுக்கு மாறாக), சொல் வடிவங்களில் பெரும் சதவீதம் தெளிவற்றவை. எடுத்துக்காட்டாக, பொதுவாக ஆங்கிலத்தில் ஒரு பன்மை பெயர்ச்சொல் என்று கருதப்படும் "dogs" என்பது கூட ஒரு வினைச்சொல்லாக இருக்கலாம்:

The sailor dogs the hatch.

சரியான இலக்கண சொல்வகப்பாட்டு அடையாளப்படுத்தல் "dogs" மிகவும் பொதுவான பன்மை பெயர்ச்சொல்லாகப் பயன்படுத்தப்படவில்லை, இங்கே ஒரு வினைச்சொல்லாகப் பயன்படுத்தப்படுகிறது என்பதை பிரதிபலிக்கும். இதைத் தீர்மானிக்க இலக்கண சூழல் ஒரு வழி ஆகும்; "sailor" மற்றும் " hatch" ஆகியவை " dogs" என்பது கடல் சூழலிலைக் குறிக்கின்றது மற்றும் "hatch" என்ற செயப்படுபொருளுக்கு பயன்படுத்தப்படும் ஒரு செயல் எனக் கணிக்க இயலும் (இந்தச் சூழலில், " dogs" என்பது ஒரு கடல்கால அர்த்தத்தில் "(நன்றாக மூடப்பட்ட கதவைப்) பாதுகாப்பாக இறுக்குகிறார்" எனப்பொருள்படும்).

சொல்வகைப்பாடுட்டுக் குழுமம்

ஆங்கிலத்தில் 9 சொல்வகைப்பாடுகள் இருப்பதாகப் பள்ளிகள் பொதுவாக கற்பிக்கின்றன: பெயர்ச்சொல், வினைச்சொல், அடைகொளி அடை, பெயரடை, முன்னுருபு, மாற்றுப்பெயர், வினையடை, இணைப்புக்கிளவி மற்றும் வியப்பிடைச்சொல். இருப்பினும், இன்னும் பல பிரிவுகள் மற்றும் துணை பிரிவுகள் தெளிவாக உள்ளன. பெயர்ச்சொற்களைப் பொறுத்தவரை, பன்மை, உடைமை மற்றும் ஒருமை வடிவங்களை வேறுபடுத்தி அறியலாம். பல மொழிகளில் சொற்கள் அவற்றின் "வேற்றுமை" (எழுவாய், செயப்படுபொருள் போன்றவற்றின் பங்கு), இலக்கணப் பாலினம் மற்றும் பலவற்றிற்கும் குறிக்கப்பட்டுள்ளன; வினைச்சொற்கள் காலம், வினையாற்றுவகை மற்றும் பிற விஷயங்களுக்காகக் குறிக்கப்பட்டுள்ளன. சில குறியீட்டு/அடையாளப்படுத்தும் முறைகளில், ஒரே மூல சொற்களின் வெவ்வேறு திரிபுகள் வெவ்வேறு சொல்வகைப்பாடுகளைப் பெறும்; இதன் விளைவாக அதிக எண்ணிக்கையிலான (விளக்க) அடையாளங்கள் கிடைக்கும். எடுத்துக்காட்டாக, பொது ஒருமைப் பெயர்களுக்கு NN,

பொது பன்மைப் பெயர்களுக்கு NNS, ஒருமை இயற்பெயர்களுக்கு NP (பிரவுண் தரவுத்தொகுதியில் பயன்படுத்தப்படும் சொல்வகைப்பாடு அடையாளங்களைப் பார்க்கவும்). பிற அடையாளப்படுத்தும் அமைப்புகள் குறைந்த எண்ணிக்கையிலான அடையாளங்களைப் பயன்படுத்துகின்றன மற்றும் நுணுக்கமான வேறுபாடுகளைப் புறக்கணிக்கின்றன அல்லது சொல்வகைப்பாட்டிலிருந்து ஓரளவு சுதந்திரமான அம்சங்களாக அவற்றை வடிவமைக்கின்றன.

கணினி மூலம் சொல்வகைப்பாடு அடையாளப்படுத்தலில், ஆங்கிலத்திற்கான சொல்வகைப்பாடு அடையாளங்கள் 50 முதல் 150 வகைப்பாடுகளை வேறுபடுத்துவது பொதுவானது. கோய்ன் கிரேக்கத்தைக் (Koine Greek) (DeRose 1990) குறியிடுவதற்கான புள்ளியியல்சார் முறைகள் குறித்த செயல்பாடு, சொல்வகைப்பாட்டின் 1,000க்கும் மேற்பட்ட சொல்வகைப்பாடுகளைப் பயன்படுத்தியுள்ளது; மேலும் ஆங்கிலத்திற்கு ஒப்பாக அந்த மொழியில் பல சொற்கள் பொருண்மை மயக்கமாக இருப்பதைக் கண்டறிந்துள்ளது. வகைப்பாடு = பெயர்ச்சொல், வகை = பொது, பாலினம் = ஆண்பால், எண் = ஒருமை, வேற்றுமை = செய்ப்படுபொருள்வேற்றுமை, உயிர்மை = இல்லை என்பதற்கான Ncmsan போன்ற மிகக் குறுகிய நினைவுக்குறிப்புகளைப் பயன்படுத்தி உருபனியல் அடிப்படையில் செழிப்பான மொழிகளின் விஷயத்தில் ஒரு உருபனியல்-தொடரியல்சார் விளக்கி (morphosyntactic descriptor) பொதுவாக வெளிப்படுத்தப்படுகிறது.

அமெரிக்க ஆங்கிலத்திற்கான சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கான மிகவும் பிரபலமான "டேக் செட்" என்பது பென் ட்ரீ பேங்க் திட்டத்தில் (Penn Treebank project) உருவாக்கப்பட்ட பென் அடையாளக் குழுமம் (Penn tag set) ஆகும். இது பெரும்பாலும் முந்தைய பிரவுன் தரவுத்தொகுதி (Brown Corpus) மற்றும் லாப் தரவுத்தொகுதி (LOB Corpus) அடையாளக் குழுமங்களுடன் ஒத்திருக்கிறது, இது மிகவும் சிறியதாக இருந்தாலும். ஐரோப்பாவில், ஈகிள்ஸ் வழிகாட்டுதல்களிலிருந்து அடையாளத் தொகுப்புகள் (tag sets) பரந்த பயன்பாட்டைக் காண்கின்றன மற்றும் பல மொழிகளுக்கான பதிப்புகளை உள்ளடக்குகின்றன.

சொல்வகைப்பாட்டு அடையாளப்படுத்தல் பணி பல்வேறு மொழிகளில் செய்யப்பட்டுள்ளது; மேலும் பயன்படுத்தப்படும் சொல்வகைப்பாட்டு அடையாளங்களின் தொகுப்பு மொழியுடன் பெரிதும் மாறுபடும். அடையாளங்கள் பொதுவாக வெளிப்படையான உருபன்சார் வேறுபாடுகளை உள்ளடக்கும் வகையில் வடிவமைக்கப்பட்டுள்ளன; இருப்பினும் இது ஆங்கிலத்தில் பெயர்ச்சொற்களுக்கு அன்றி மாற்றுப்பெயர்களுக்கான வேற்றுமையைக்

குறிப்பது போன்ற முரண்பாடுகளுக்கு வழிவகுக்கிறது; மற்றும் மொழியைக் கடந்து மிகப் பெரிய வேறுபாடுகளுக்கு வழிவகுக்கிறது. கிரேக்க மற்றும் லத்தீன் போன்ற பெரிதும் திரிபுறும் மொழிகளுக்கான அடையாளக் குழுமம் மிகப் பெரியதாக இருக்கும்; இன்யூட் மொழிகள் (Inuit languages) போன்ற ஒட்டுநிலை மொழிகளில் (agglutinative language) சொற்களைக் குறிப்பது கிட்டத்தட்ட சாத்தியமற்றது. மற்றொரு தீவிரத்தில், பெட்ரோவ் மற்றும் பலர் (Petrov et al.) 12 வகைகள் கொண்ட "உலகளாவிய" அடையாளத் தொகுப்பை ("universal" tag set) (எடுத்துக்காட்டாக, பெயர்ச்சொற்கள், வினைச்சொற்கள், நிறுத்தற்குறிகள் போன்றவை; ஆங்கிலத்தில் 'to' வினையெச்சக் குறியீடாகவும் முன்னுருபாகவும் வேறுபாடின்றி இருப்பது போன்றவை) முன்மொழிந்தனர். மிகவிரிந்த அடையாளங்களின் (very broad tags) மிகச் சிறிய அடையாளங்கள் அல்லது மிகத் துல்லியமான குறியீட்டுத் தொகுப்புகளின் சிறிய குறியீட்டு விரும்பத்தக்கதா என்பது கையில் இருக்கும் நோக்கத்தைப் பொறுத்தது. சிறிய அடையாளக் குழுமங்களில் தானியங்கி அடையாளப்படுத்தல் எளிதானது.

வரலாறு

பிரவுன் தரவுத்தொகுதி

சொல்வகைப்பாட்டு அடையாளம் குறித்த ஆராய்ச்சி தரவுத்தொகுதி மொழியியலுடன் நெருக்கமாக பிணைக்கப்பட்டுள்ளது. கணினி பகுப்பாய்விற்கான ஆங்கிலத்தின் முதல் பெரிய கார்பஸ் 1960களின் நடுப்பகுதியில் ஹென்றி குசெரா மற்றும் டபிள்யூ. நெல்சன் பிரான்சிஸ் (Henry Kučera and W. Nelson Francis) ஆகியோரால் பிரவுன் பல்கலைக்கழகத்தில் உருவாக்கப்பட்ட பிரவுன் கார்பஸ் ஆகும். இது தோராயமாக தேர்ந்தெடுக்கப்பட்ட வெளியீடுகளிலிருந்து 500 மாதிரிகளால் ஆன நடப்பு ஆங்கில உரைநடை உரையின் சுமார் 1,000,000 சொற்களைக் கொண்டுள்ளது. ஒவ்வொரு மாதிரியும் 2,000 அல்லது அதற்கு மேற்பட்ட சொற்கள் (2,000 சொற்களுக்குப் பிறகு முதல் வாக்கியத்தின் முடிவில் முடிவடைகிறது, இதனால் கார்பஸில் முழுமையான வாக்கியங்கள் மட்டுமே உள்ளன).

பிரவுன் தரவுத்தொகுதி பல ஆண்டுகளாக சொல்வகைப்பாட்டுக் குறியீடுகளுடன் மிகக் கடினமான உழைப்பால் "குறியிடப்பட்டது". கிரீன் மற்றும் ரூபின் (Greene and Rubin) ஆகியோரால் ஒரு திட்டத்துடன் முதல் தோராயமாக்கல் செய்யப்பட்டது, இதில் என்னென்ன வகைகள் இணைநிகழ்வு செய்யக்கூடும் என்பதற்கான ஒரு பெரிய கையால் செய்யப்பட்ட பட்டியலைக் கொண்டிருந்தது. எடுத்துக்காட்டாக, சார்படை (article) பின்னர் பெயர்ச்சொற்கள்

வரலாம், ஆனால் சார்படைபின்னா வினைச்சொல் (விவாதிக்கக்கூடியது) வரமுடியாது. நிரல் (program) சுமார் 70% சரியைப் பெற்றது. அதன் முடிவுகள் மீண்டும் மீண்டும் மதிப்பாய்வு செய்யப்பட்டு கையால் திருத்தப்பட்டன; பின்னர் பயனர்கள் பிழைப்பட்டியலை அனுப்பினர்; இதனால் 70களின் பிற்பகுதியில் குறியீடுகள் கிட்டத்தட்ட சரியாக இருந்தது (மனித பேச்சாளர்கள் கூட ஒப்புக் கொள்ளாத சில சந்தர்ப்பங்களை அனுமதிக்கிறது).

இந்த தரவுத்தொகுதி சொல்-அதிர்வெண் மற்றும் சொல்வகைப்பாட்டின் எண்ணற்ற ஆய்வுகளுக்குப் பயன்படுத்தப்பட்டது மற்றும் பல மொழிகளில் இதேபோன்ற "குறிக்கப்பட்ட" தரவுத்தொகுதிகளின் வளர்ச்சியை ஊக்குவித்தது. அதை பகுப்பாய்வு செய்வதன் மூலம் பெறப்பட்ட புள்ளிவிவரங்கள் CLAWS (மொழியியல்) மற்றும் வோல்சுங்கா (VOLSUNGA) போன்ற பெரும்பாலான பிந்தைய சொல்வகைப்பாட்டு குறியீடுகளின் அமைப்புகளுக்கு அடிப்படையாக அமைந்தன. இருப்பினும், 2005இல் இது 100 மில்லியன் சொற்களாலான பிரிட்டிஷ் நேஷனல் தரவுத்தொகுதி போன்ற பெரிய தரவுத்தொகுதியால் முறியடிக்கப்பட்டுள்ளது.

சில காலமாக, சொல்வகைப்பாட்டுக் குறியிடுதல் இயற்கையான மொழி செயலாக்கத்தின் பிரிக்க முடியாத பகுதியாகக் கருதப்பட்டது; ஏனென்றால் பொருண்மையிலையோ அல்லது சூழலின் பயன்வழியிலைப் புரிந்து கொள்ளாமல் சொல்வகைப்பாட்டைச் சரியாகத் தீர்மானிக்க முடியாத சில சந்தர்ப்பங்கள் உள்ளன. இது மிகவும் விலை உயர்ந்தது; குறிப்பாக ஒவ்வொரு சொல்லுக்கும் பல சொல்வகைப்பாட்டுச் சாத்தியங்கள் கருதப்படும்போது உயர் நிலைகளைப் பகுப்பாய்வு செய்வது மிகவும் கடினம்.

மறைக்கப்பட்ட மார்கோவ் மாதிரிகளின் பயன்பாடு (Use of hidden Markov models)

1980களின் நடுப்பகுதியில், ஐரோப்பாவில் ஆராய்ச்சியாளர்கள் பிரிட்டிஷ் ஆங்கிலத்தின் லான்காஸ்டர்-ஒஸ்லோ-பெர்கன் தரவுத்தொகுதியை (Lancaster-Oslo-Bergen Corpus) அடையாளப்படுத்தச் செய்யச் செயல்படும் போது, சொல்வகைப்பாடுகளின் மயக்கத்தை நீக்க மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகளைப் (எச்.எம்.எம்./HMMs) பயன்படுத்தத் தொடங்கினர். எச்.எம்.எம்.கள் வேற்றுமைகளை எண்ணுவது (பிரவுன் கார்பஸிலிருந்து போன்றவை) மற்றும் சில வரிசைகளின் நிகழ்தகவுகளின் அட்டவணையை உருவாக்குவது ஆகியவை அடங்கும். எடுத்துக்காட்டாக, ' the' போன்ற ஒரு சார்படையை நீங்கள் பார்த்தவுடன், அடுத்த சொல் ஒரு பெயர்ச்சொல் 40%, ஒரு பெயரடை 40% மற்றும் ஒரு எண் 20% ஆகும். இதை அறிந்தால், ஒரு வினைச்சொல் அல்லது ஒரு மாதிரியை விட "the can" இல் உள்ள " can" என்பது ஒரு வினை

என்பதைவிட ஒரு பெயர்ச்சொல்லாக இருக்க வாய்ப்புள்ளது என்று ஒரு நிரல் தீர்மானிக்க முடியும். பின்வரும் சொற்களைப் பற்றிய அறிவிலிருந்து பயனடைய அதே முறையைப் பயன்படுத்தலாம்.

மிகவும் மேம்பட்ட ("உயர்-வரிசை") எச்.எம்.எம்.-கள் இருமடங்குகளில் மட்டுமல்ல, மும்மடங்குகளிலும் அல்லது பெரிய வரிசைகளிலும் நிகழ்தகவுகளைக் கற்றுக்கொள்கின்றன. எனவே, எடுத்துக்காட்டாக, ஒரு வினைச்சொல்லைத் தொடர்ந்து ஒரு பெயர்ச்சொல்லை நீங்கள் பார்த்திருந்தால், அடுத்தச் சொல் ஒரு முன்னொட்டு, சார்படை அல்லது பெயர்ச்சொல்லாக இருக்கலாம்; ஆனால் மற்றொரு வினைச்சொல் ஆக இருப்பது மிகக் குறைவு.

பல பொருண்மைமயக்கமுள்ள சொற்கள் ஒன்றாக நிகழும்போது, சாத்தியங்கள் பெருகும். எவ்வாறாயினும், ஒவ்வொரு தேர்வையும் நிகழ்தகவுகளை ஒன்றிணைப்பதன் மூலம் ஒவ்வொரு கலவையையும் கணக்கிடுவதும் ஒவ்வொன்றிற்கும் ஒரு தொடர்புடைய நிகழ்தகவை ஒதுக்குவதும் எளிதானது. அதிக நிகழ்தகவு கொண்ட சேர்க்கை பின்னர் தேர்ந்தெடுக்கப்படுகிறது. ஐரோப்பிய குழு CLAWS ஐ உருவாக்கியது, இது ஒரு குறியீட்டுத் திட்டமாகும்; இது சரியாகச் செய்து 93-95% பரப்பெல்லையில் துல்லியத்தை அடைந்தது.

இயற்கையான மொழி பாகுபடுத்தலுக்கான புள்ளிவிவர நுட்பங்களில் (1997), யூஜின் சார்னியாக் சுட்டிக்காட்டியுள்ளபடி, அறியப்பட்ட ஒவ்வொரு வார்த்தைக்கும் மிகவும் பொதுவான குறியீட்டை ஒதுக்குவது மற்றும் அனைத்து அறியப்படாதவைகளுக்கும் "இயற்பெயர்" என்ற குறியீடு 90% துல்லியத்தை அணுகும் ஏனென்றால் பல சொற்கள் பொருண்மை மயக்கமற்றவை; மேலும் பல சொற்கள் மிக அரிதாகவே அவற்றின் குறைவான பொதுவான சொல்வகைப்பாட்டை குறிக்கின்றன.

HMM-அடிப்படையிலான சொல்வகைப்பாட்டுக் குறியீடுதலின் துறையில் CLAWS முன்னோடியாக இருந்தது; ஆனால் இது அனைத்து சாத்தியக்கூறுகளையும் கணக்கிட்டதால் மிகவும் விலை உயர்ந்தது. வெறுமனே பல விருப்பங்கள் இருக்கும்போது இது சில நேரங்களில் காப்பு முறைகளை நாட வேண்டியிருந்தது (பிரவுன் கார்பஸ் ஒரு வரிசையில் 17 தெளிவற்ற சொற்களைக் கொண்ட ஒரு வழக்கைக் கொண்டுள்ளது, மேலும் "still" போன்ற சொற்கள் உள்ளன; அவை 7 தனித்துவமான சொல்வகைப்பாட்டைக் குறிக்கும் (DeRose 1990, பக். 82)).

எச்.எம்.எம்.கள் புள்ளியியல்சார் குறியீடுகளின் செயல்பாட்டை உட்படுத்துகின்றன, மேலும் அவை பல்வேறு வழிமுறைகளில் பயன்படுத்தப்படுகின்றன, அவை இரு-திசை அனுமான அல்காரிதம் ஆகும்.

டைனமிக் நிரலாக்க முறைகள் (Dynamic programming methods)

1987ஆம் ஆண்டில், ஸ்டீவன் டிரோஸ் (Steven DeRose) மற்றும் (கென் சர்ச்) Ken Church ஆகியோர் ஒரே சிக்கலை மிகக் குறைந்த நேரத்தில் தீர்க்க தனித்தனியாக (ஒருவரைஒருவர் சாராமல்) டைனமிக் நிரலாக்க வழிமுறைகளை உருவாக்கினர். அவர்களின் முறைகள் மற்ற துறைகளில் சில காலம் அறியப்பட்ட விட்டர்பி (Viterbi algorithm) வழிமுறையைப் போலவே இருந்தன. டிரோஸ் ஜோடிகளின் அட்டவணையைப் பயன்படுத்தினார்; அதே நேரத்தில் சர்ச் மும்மடங்கு அட்டவணையையும் பிரவுன் தரவுத்தொகுதியில் அரிதான அல்லது இல்லாத மும்மடங்களுக்கான மதிப்புகளை மதிப்பிடும் முறையையும் பயன்படுத்தினார் (மூன்று நிகழ்தகவுகளின் உண்மையான அளவீட்டுக்கு மிகப் பெரிய கார்பஸ் தேவைப்படும்). இரண்டு முறைகளும் 95% க்கும் அதிகமான துல்லியத்தை அடைந்தன. பிரவுன் பல்கலைக்கழகத்தில் டெரோஸின் 1990 ஆம் ஆண்டு ஆய்வுக் கட்டுரையில் குறிப்பிட்ட பிழை வகைகள், நிகழ்தகவுகள் மற்றும் பிற தொடர்புடைய தரவுகளின் பகுப்பாய்வுகளும் அடங்கியிருந்தன; மேலும் கிரேக்க மொழியில் அவரது படைப்பு திரும்பச்செய்யப்பட்டது; அது இதேபோல் பயனுள்ளதாக இருந்தது.

இந்தக் கண்டுபிடிப்புகள் இயற்கை மொழி ஆய்வுத் துறையில் வியக்கத்தக்க வகையில் இடைஞ்சலாக இருந்தன. மொழியியல் பகுப்பாய்வின் பல உயர்ந்த நிலைகளுடன் சொல்வகைப்பட்டுத் தேர்வை ஒருங்கிணைக்கும் மிகவும் அதிநவீன வழிமுறைகளின் வழக்கமான துல்லியத்தை விட அறிக்கையிடப்பட்ட துல்லியம் அதிகமாக இருந்தது: தொடரியல், உருபனியல், பொருண்மையியல் மற்றும் பல. பொருண்மையியல் தேவைப்படும் சில அறியப்பட்ட நிகழ்வுகளுக்கு CLAWS, DeRose மற்றும் Church-இன் முறைகள் தோல்வியடைந்தன; ஆனால் அவை மிக அரிதாகவே நிரூபிக்கப்பட்டன. சொல்வகைப்பாட்டு குறியிடல் பிற செயலாக்கங்களிலிருந்து பிரிக்கப்படலாம் என்பதை இது துறையில் பலருக்கு உணர்த்தியது; இது கணினிமயமாக்கப்பட்ட மொழிப் பகுப்பாய்வின் கோட்பாடு மற்றும் நடைமுறையை எளிதாக்கியது மற்றும் பிற பகுதிகளையும் பிரிப்பதற்கான வழிகளைக் கண்டறிய ஆராய்ச்சியாளர்களை ஊக்குவித்தது. மார்கோவ் மாதிரிகள் இப்போது சொல்வகைப்பாடு ஒதுகீட்டுக்கு ஒரு தரமான முறையாகும்.

மேற்பார்வை செய்யப்படாத குறிச்சொற்கள் (Unsupervised taggers)

ஏற்கனவே விவாதிக்கப்பட்ட முறைகள் குறியீட்டு நிகழ்தகவுகளைக் கற்றுக்கொள்வதற்கு முன்பே இருக்கும் கார்பஸிலிருந்து (pre-existing corpus) செயல்படுவதை உள்ளடக்கியது.

இருப்பினும், "மேற்பார்வை செய்யப்படாத" குறியிடலைப் (Unsupervised taggers) பயன்படுத்தி பூட்ஸ்ட்ராப் செய்ய முடியும். மேற்பார்வை செய்யப்படாத குறியிடல் நுட்பங்கள் அவற்றின் பயிற்சித் தரவுகளுக்கு ஒரு குறிக்கப்படாத தரவுத்தொகுதியைப் பயன்படுத்துகின்றன மற்றும் தூண்டல் மூலம் குறியீட்டுத் தொகுதியை உருவாக்குகின்றன. அதாவது, அவை சொல் பயன்பாட்டில் ஒழுங்கமைபுகளைக் கவனித்து, சொல்வகைப்பாடுகளை ஆக்குகின்றன. எடுத்துக்காட்டாக, புள்ளிவிவரங்கள் "the", "a"மற்றும் "an"ஆகியவை ஒத்த சூழல்களில் நிகழ்கின்றன, அதே நேரத்தில் "eat" என்பது மிகவும் வித்தியாசமானவற்றில் நிகழ்கிறது. போதுமான மறு செய்கையுடன், மனித மொழியியலாளர்கள் எதிர்பார்ப்பதைப் போலவே சொற்களின் ஒற்றுமை வகுப்புகள் (similarity classes) வெளிப்படுகின்றன; மற்றும் வேறுபாடுகள் சில நேரங்களில் மதிப்புமிக்க புதிய நுண்ணறிவுகளைப் பரிந்துரைக்கின்றன.

இந்த இரண்டு வகைகளையும் விதி அடிப்படையிலான, புள்ளியியல் மற்றும் நரம்பியல் அணுகுமுறைகளாக மேலும் பிரிக்கலாம்.

பிற அடையாளங்கள் மற்றும் முறைகள் (Other taggers and methods)

சொல்வகைப்பாடு அடையாளப்படுத்தலின் சில தற்போதைய முக்கிய வழிமுறைகளில் விட்டர்பி வழிமுறை (Viterbi algorithm), பிரில் டேக்கர் (Brill tagger), கட்டுப்பாட்டு இலக்கணம் (Constraint Grammar) மற்றும் பாம்-வெல்ச் வழிமுறை (Baum-Welch algorithm) (முன்னோக்கி-பின்தங்கிய வழிமுறை என்றும் அழைக்கப்படுகிறது) ஆகியவை அடங்கும். மறைக்கப்பட்ட மார்க்கோவ் மாதிரி மற்றும் புலப்படும் மார்கோவ் மாதிரி (visible Markov model) குறியீடுகள் இரண்டையும் விட்டர்பி வழிமுறையைப் பயன்படுத்தி செயல்படுத்தலாம். விதி அடிப்படையிலான பிரில் அடையாளப்படுத்தி அசாதாரணமானது; இது ஒரு விதி அமைப்பொழுங்குகளின் தொகுதியையைக் கற்றுக்கொள்கிறது; பின்னர் புள்ளிவிவர அளவை மேம்படுத்துவதை விட அந்த வடிவங்களைப் பயன்படுத்துகிறது. விதிகள் தொடர்ச்சியாக வரிசைப்படுத்தப்படும் பிரில் டேக்கரைப் போலன்றி, பிஓஎஸ் மற்றும் உருபனியயல் குறியிடுதல் கருவித்தொகுதி (toolkit) ஆர்.டி.ஆர்.பொஸ்டாகர் (RDRPOSTagger) ரிப்பிள்-டவுண் விதிகள் கிளையப்பு (ripple-down rules tree) வடிவத்தில் விதிகளைச் சேகரித்து வைக்கின்றன.

சொல்வகைப்பாடு அடையாளப்படுத்தலின் சிக்கலுக்கு பல இயந்திர கற்றல் முறைகளும் பயன்படுத்தப்பட்டுள்ளன. எஸ்.வி.எம். (SVM Support-vector machine), அதிகபட்ச என்ட்ரோபி வகைப்படுத்தி (maximum entropy classifier), பெர்செப்டிரான் (perceptron) மற்றும் அருகிலுள்ள

அண்டை (nearest-neighbor) போன்ற முறைகள் அனைத்தும் முயற்சிக்கப்பட்டுள்ளன; மேலும் பெரும்பாலானவை 95% க்கு மேல் துல்லியத்தை அடைய முடியும்.

பல முறைகளின் நேரடி ஒப்பீடு ACL விக்கியில் (ACL Wiki) (குறிப்புகளுடன்) தெரிவிக்கப்படுகிறது. இந்த ஒப்பீடு சில பென் ட்ரீபேங்க் தரவுகளில் (Penn Treebank data), அமைக்கப்பட்ட பென் குறியீடுத் தொகுதியைப் (Penn tag set) பயன்படுத்துகிறது, எனவே முடிவுகள் நேரடியாக ஒப்பிடத்தக்கவை. இருப்பினும், பல குறிப்பிடத்தக்க அடையாளப்படுத்திகள் சேர்க்கப்படவில்லை (ஒருவேளை இந்த குறிப்பிட்ட தரவுத்தொகுப்பிற்காக அவற்றை மறுசீரமைப்பதில் ஈடுபட்டுள்ள உழைப்பு காரணமாக இருக்கலாம்). எனவே, இங்கே தெரிவிக்கப்பட்ட முடிவுகள் கொடுக்கப்பட்ட அணுகுமுறையால் அடையக்கூடிய சிறந்தவை என்று கருதக்கூடாது; கொடுக்கப்பட்ட அணுகுமுறையால் அடையப்பட்ட சிறந்தவை கூட இல்லை.

2014ஆம் ஆண்டில், சொல்வகைப்பாட்டு குறியிடலுக்கான கட்டமைப்பு ஒழுங்குமுறை முறையைப் பயன்படுத்தி ஒரு காகித அறிக்கை, நிலையான/தரமான பெஞ்ச்மார்க் தரவுத்தொகுப்பில் 97.36% ஐ அடைகிறது

சொல்வகை அடையாளப்படுத்தும் (Part of Speech Tagging) திட்டம் ஒரு சொல்லை வாக்கியத்தில் அதன் சொல்வகைப்பாட்டுடன் அடையாளப்படுத்தும். இது மூன்று நிலைகளில் செய்யப்படுகின்றது: முன்திருத்தல், தானியங்கு அடையாளம் தருகை, மற்றும் மனிதமுயற்சிசார் பின்திருத்தல்.

முன் திருத்தல் கட்டத்தில், ஒவ்வொரு சொல்லுக்கு அல்லது சொல் சேர்க்கைக்குச் சொல்வகைபாட்டை அடையாளப்படுத்த தரவுத்தொகுதி பொருத்தமான வடிவமாக மாற்றப்படுகிறது. எழுத்து ஒற்றுமை காரணமாக ஒரு சொல்லுக்குப் பல சொல்வகைப்பாடு அடையளாங்கள் வரலாம். சாத்தியமான சொல்வகைப்பாட்டு அடையாளப்படுத்தலின் தொடக்க நிகழ்வுக்குப் பிறகு, உரைகளில் உள்ள பொருண்மை மயக்கத்தை நீக்கச் சொற்களில் சொல்வகைப்பாடுகள் கைமுறையாகச் சரி செய்யப்படுகின்றன.

சொல்வகைப்பாட்டு அடையாளப்படுத்தலின் எடுத்துக்காட்டு (Biber at al. 1998: 258-259). மாதிரி 1 எளிய அடையாளப்படுத்தலைக் காட்டுகிறது. மாதிரி 2 பல பரிமாண அடையாளப்படுத்தலைக் காட்டுகிறது.

அடையாளப்படுத்தாத வாக்கியம் (Untagged sentence)

எ.கா.

A more to stop Mr. Gaitskell from nominating any more labour life peers is to be made at a meeting of labour MPs tomorrow.

அடையாளப்படுத்தப்பட்ட மாதிரிவடிவம் 1

^a_ AT move_ NN to_ TO stop _VB \0Mr_ NPT Gaitskell _NP from _IN nominating _ABG any_DTI. more_AP labour_NN life_NN peers_NNS is_BEZ to-TO be_BE made–VBN at_IN a_AT meeting –NN of_IN labour _NN tomorrow _NR._

அடையாளப்படுத்தப்பட்ட மாதிரிவடிவம் 2

A ^at++++	move ^nn++++
to ^to++++	stop ^vbi++++
Mr ^npt++++	Gaitskell ^np++++
from ^in++++	nominating ^xvbg+++xvbg+
any ^dti++++	more ^ap++++
labour ^nn++++	life ^nn++++
peers ^nns++++	is ^vbz+bez+aux++
to ^to++++	be ^vb+be+aux++
made ^vpsv++agls+xvbnx+	at ^in++++
a ^at++++	meeting ^nn+++xvbg+
of ^in++++	labour ^nn++++
MPs ^npts++++	tomorrow. ^nr+tm+++
.^.+clp+++	

தமிழ் உரை தரவுத்தொகுதியைப் பொறுத்தவரை, பல்வேறு சொல் வகுப்புகளைச் சேர்ந்த சொற்களுக்கு பின்வரும் குறிச்சொல்லைப்/சொல்வகைப்பாட்டு அடையாளங்களைப் பயன்படுத்துகிறோம்:

5. 10.1.1 முக்கியமான சொல் வகுப்புகள் (Major Lexical Classes)

Noun/பெயர்	: [NN]	: எ.கா. பையன்[NN]
Pronoun/மாற்றுப்பெயர்	: [PN]	: எ.கா. நான்[PN]
Adjective/பெயரடை	: [ADJ]	: எ.கா. நல்ல[ADJ]

Adverb/வினையடை	: [ADV]	: எ.கா. வேகமாக [ADV]
Finite verb/வினைமுற்று	: [FV]	: எ.கா. செய்தான்[FV]
Non-finite verb/வினைஎச்சம்	: [NFV]	: எ.கா. செய்ய[NFV]
Postposition/முன்னுருபு	: [PP]	: எ.கா. மேல்[PP]
Indeclinable/திரிபுறாதது	: [IND]	: எ.கா. ஆனால்[IND]
Reduplication/இரட்டுதல்	: [RDP]	: எ.கா. படபட[RDP]

5. 10.1.2 முக்கிய மொழி வகுப்புகள் (Major Language Class)

Tamil	: [TAM]	: எ.கா. மனிதன்[TAM]
English	: [ENG]	: எ.கா. man[ENG]
Foreign	: [FRN]	: எ.கா Admi[FRN]

5.10.1.3 பெயர்ச்சொல்லின் துணை வகுப்புகள்

Compound Noun/கூட்டுப்பெயர்	:[NN_COM]	: எ.கா. பள்ளிக்கூடம்[NN_COM]
Proper Noun/இயற்பெயர்	: [NN_PPN]	: எ.கா. பாரதியார்[NN_PPN]
Proper Noun Tamil/இயற்பெயர் தமிழ்	: [NN_PPN_TAM]	: எ.கா. சென்னை[NN_PPN_TAM]
Proper Noun English/இயற்பெயர் ஆங்கிலம்	: [NN_PPN_ENG]	: எ.கா. லண்டன்[NN_PPN_ENG]
Proper Noun Foreign/இயற்பெயர்அயல்	: [NN_PPN_FRN]	: எ.கா. அரப்[NN_PPN_FRN]
Common Noun/பொதுப்பெயர்	: [NN_CMN]	: E.g. தம்பி[NN_CMN]
Common Noun Tamil/பொதுப்பெயர்	: [NN_CMN_TAM]	: எ.கா. புத்தகம்[NN_CMN_TAM]
Common Noun English/பொதுப்பெயர்ஆங்கிலம்	: [NN_CMN_ENG]	: எ.கா. புக்NN_CMN_ENG]
Common Noun Foreign /பொதுப்பெயர்அயல்	: [NN_CMN_FRN]	: எ.கா. கிதாப்[NN_CMN_FRN]
Collective Noun /தொகைப்பெயர்	: [NN_CLL]	: எ.கா. சங்கம் [NN_CLL]
Collective Noun Tamil/தொகைப்பெயர்தமிழ்	: [NN_CLL_TAM]	: எ.கா. சந்தை[NN_CLL_TAM]
Collective Noun English/தொகைப்பெயர் ஆங்கிலம்	: [NN_CLL_ENG]	: எ.கா. மார்கெட்[NN_CLL_ENG]
Collective Noun Foreign /தொகைப்பெயர்அயல்	: [NN_CLL_FRN]	: எ.கா. பஜார்[NN_CLL_FRN]
Abstract Noun/நுண்மைப்பெயர்	: [NN_ABS]	: எ.கா. கருணை[NN_ABS]
Abstract Noun Tamil/நுண்மைப்பெயர்தமிழ்	: [NN_ABS_TAM]	: எ.கா. பலன்[NN_ABS_TAM]
Abstract Noun English/நுண்மைப்பெயர்ஆங்கிலம்	: [NN_ABS_ENG]	: எ.கா. லாப்[NN_ABS_ENG]
Abstract Noun Foreign/நுண்மைப்பெயர்அயல்	: [NN_ABS_FRN]	: எ.கா. ப்யார்[NN_ABS_FRN]
Material Noun/பருப்பொருள்பெயர்	: [NN_MAT]	: எ.கா. வணி[NN_MAT]
Material Noun Tamil/பருப்பொருள்பெயர்தமிழ்	: [NN_MAT_TAM]	: எ.கா. பால்[NN_MAT_BNG]

Material Noun English/பருப்பொருள்பெயர்ஆங்கிலம்: [NN_MAT_ENG]: எ.கா. ஷார்ட்[NN_MAT_ENG]

Material Noun Foreign/பருப்பொருள்பெயர்அயல் : [NN_MAT_FRN] : எ.கா. ஜாமா[NN_MAT_FRN]

5.10.1.4. மாற்றுப்பெயரின் துணைவகுப்பு(Subclasses of Pronoun)

Personal Pronoun/மூவிடமாற்றுப்பெயர் : [PN_PRS] : எ.கா. நான்[PN_PRS]

Interrogative Pronoun /வினாமாற்றுப்பெயர் : [PN_INT] : எ.கா. யார்[PN_INT]

Relative Pronoun/சார்புமாற்றுப்பெயர் : [PN_RLT] : எ.கா. --- [PN_RLT]

Demonstrative Pronoun/சுட்டுமாற்றுப்பெயர் : [PN_DMS] : எ.கா. இது[PN_DMS]

Indefinite Pronoun/குறிப்பில்லாமாற்றுப்பெயர்: [PN_IDF] : எ.கா. ஒருவன்[PN_IDF]

Definite Pronoun/குறிப்புடைமாற்றுப்பெயர் : [PN_DFN] : எ.கா. அவன்[PN_DFN]

Adjectival Pronoun/பெயரடைமாற்றுப்பெயர் : [PN_AJV] : எ.கா. என்[PN_ADJ]

Emphatic Pronoun/வலியுறுத்துமாற்றுப்பெயர் : [PN_EMP] : எ.கா. நான்தான்[PN_EMP]

Reflexive Pronoun/தற்சுட்டுமாற்றுப்பெயர் : [PN_RFX] : எ.கா. தன்[PN_RFX]

5.9.1.5. பெயரடைகளின் துணை வகுப்புகள் (Subclasses of Adjectives)

Adjectival Compound/கூடுப்பெயரடை : [ADJ_COM] : எ.கா. [ADJ_COM]

General Adjective/பொதுப்பெயரடை : [ADJ_GEN] : எ.கா. கெட்ட[ADJ_GEN]

Qualitative Adjective/பண்புப்பெயரடை : [ADJ_QUL] :எ.கா. அழகான[ADJ_QUL]

Quantitative Adjective/அளவுப்பெயரடை : [ADJ_QUN] : எ.கா. பல[ADJ_QUN]

Qualitative-quantitative Adjective/பண்பு-அளவுபெயரடை: [ADJ_QQN] : எ.கா. எவ்வளவு[ADJ_QQN]

Numeral Adjective/எண்ணுப்பெயரடை : [ADJ_NUM] : எ.கா. இரு[ADJ_NUM]

Cardinal Adjective/அடிப்படைஎண்ணுப்பெயரடை : [ADJ_CRD] : எ.கா. irupatu[ADJ_CRD]

Ordinal Adjective/முறைமைஎண்ணுப்பெயரடை : [ADJ_ORD] : எ.கா. முதல்[ADJ_ORD]

Collective Adjective/தொகைஎண்ணுப்பெயரடை : [ADJ_CLL] : எ.கா. எல்லா[ADJ_CLL]

Pronominal Adjective/மாற்றுப்பெயர்பெயரடை : [ADJ_PN] :எ.கா. என்னுடைய[ADJ_PN]

5.10.1.6. வினையடைகளின் துணை வகுப்புகள் (Subclasses of Adverbs)

General Adverbs/பொதுவினையடைகள் : [ADV_GEN] :எ.கா. விரிவாக[ADV_GEN]

5.10.1.7. முற்று வினைகளின் துணை வகுப்புகள் (Subclasses of Finite Verbs)

Finite verbs in present tense/நிகழ்காலத்தில்முற்றுவினை :[FV_PRT]:எ.கா.செய்கிறான்[FV_PRT]

Finite verbs in past tense/நிகழ்காலத்தில்முற்றுவினை : [FV_PST]: எ.கா. செய்தான்[FV_PST]

Finite verbs in future tense/எதிர்காலத்தில்முற்றுவினை : [FV_FRT]: எ.கா. செய்வான்[FV_FRT]

Finite verbs in habitual tense/வழமைகாலமுற்றுவினை : [FV_HBT]: எ.கா. செய்துவருகிறான்[FV_HBT]

Finite verbs in habitual tense/வழமைகாலமுற்றுவினை :[FV_HBT]:எ.கா. செய்துகொண்டுவருகிறான்[FV_HBT]

Finite verbs in causative sense/காரணமுற்றுவினை : [FV_CUS]: எ.கா. செய்வித்தான்[FV_CUS]

Finite verbs in gerundial sense/தொழிற்பெயர் முற்றுவினை : [FV_GRD]: எ.கா. செய்தது[FV_GRD]

5.10.1.8. திரிபுறாதவைகளின் துணைவகுப்புகள் (Subclasses of Indeclinables)

General Indeclinable/பொதுதிரிபுறாதவை : [IND_GEN]: எ.கா. [IND_GEN]

Emphatic Particle/தேற்ற இடைச்சொல் :[IND_EMP_PRT]: எ.கா. தான்[IND_EMP_PRT]

Negative Particles/எதிர்மறைஇடைச்சொல் : [IND_NEG_PRT] : எ.கா. இல்லை[IND_NEG_PRT]

5.10.1.9. இரட்டுதலின் துணைவகுப்புகள் (Subclasses of Reduplication)

Reduplication of noun/பெயர்இரட்டுதல் : [RDP_NN] : எ.கா. புலிகிலிRDP_NN]

Reduplication of Pronoun/மாற்றுப்பெயர் இரட்டுதல் : [RDP_PN] : எ.கா. அவரவர்[RDP_PN]

Reduplication of Adjective /பெயரடை இரட்டுதல் : [RDP_ADJ] : எ.கா. நல்லநல்ல[RDP_ADJ]

Reduplication of Adverb/வினையடைஇரட்டுதல் : [RDP_ADV] : எ.கா.வேகவேகமாக[RDP_ADV]

Reduplication of Finite Verb/முற்றுவினைஇரட்டுதல் : [RDP_FV] : எ.கா. -----[RDP_FV]

Reduplication of Non-finite verb/எச்சவினைஇரட்டுதல : [RDP_NFV] : எ.கா. செய்யசெய்ய[RDP_NFV]

Reduplication of Postposition/பின்னொட்டின் இரட்டுதல் : [RDP_PP] : எ.கா. மேல்மேலே[RDP_PP]

Reduplication of Indeclinable/திரிபுறாதவை இரட்டுதல் : [RDP_IND] : எ.கா. ------[RDP_IND]

இந்திய அரசின் நிதி நல்கையின் கீழ் நடைபெற்ற இயந்திர மொழி பெயர்பு, தரவுத்தொகுதி உருவாக்கம் என்ற ஆய்வுத்திட்டங்களுக்குத் தேவையான சொல்வகை அடையாளப் படுத்தப்பட்ட உரைகளை உருவாக்க இத்திட்டங்களில் பங்கேற்ற நிறுவனங்களின் கூட்டு முயற்சியால் Bis (Bureau of Indian Standards) tagset for Indian Languages உருவாக்கபட்டு பயன்பாட்டில் உள்ளது. இது ஒரு படிநிலை சொல் அடையாளக் குழுமமாகும். இவ்வடையாளக் குழுமம் (tagset) கீழே பட்டியலிடப்பட்டுள்ளது.

S.No.	English		தமிழ் நிகரன்கள்	எடுத்துக்காட்டுகள்
1	Noun	N	பெயர்	மாம்பழம்
	common	N_NN	பொதுப்பெயர்	மாம்பழம்
	Proper	N_NNP	இயற்பெயர்	கண்ணன்
	Verbal		தொழிபெயர்	படித்தல்
	Nloc	N_NST	இடகாலப்பெயர்	பின்னால், பிறகு
2	Pronoun	*PRP*	மாற்றுப் பெயர்	நான், நீ,அவன்
	Personal	PR_PRP	மூவிடப்பெயர்	நான், நீ, அவன்

	Reflexive	PR_PRF	தற்சிட்டுப்பெயர்	தான்
	Reciprocal	PR_PRC	பரிமாற்றப்பெயர்	ஒருவரையொருவர்
	Relative	PR_PRL	சார்பு மாற்றுப்பெயர்	
	Wh-words	PR_PRQ	வினா மாற்றுப்பெயர்	யார்
3	Demonstrative	DM	சுட்டு	அந்த
	Deictic	DM_DMD	சுட்டு	அந்த
	Relative	DM_DMR	சார்பு	
	Wh-words	DM_DMQ	வினாச்சொல்	எது
4	Verb	V	வினை	பாடு, செய்
	Auxiliary Verb	V_VAUX	துணைவினை	இரு, கொண்டிரு
	Main Verb	V_VM	முதன்மைவினை	நட, ஓடு
	Finite	V_VM_VF	முற்றுவினை	ஓடினான்
	Infinitive	V_VM_VNF	வினை எச்சம்	ஓட, பாட
	Gerund	V_VM_VNG	வினைப் பெயர்	ஓடுதல், பாடுதல்
	Non-Finite	V_VM_VNF	எச்சவினை	வந்து, வந்த
5	Adjective	JJ	பெயரடை	நல்ல, பெரிய
6	Adverb	RB	*வினையடை*	மெல்ல, விரைவாக
7	Post Position	PSP	பின்னுருபு	கொண்டு, விட, காட்டிலும்
8	Conjunction		இணைப்புக்கிளவி	மற்றும்
	Co-ordinator	CCD	சமநிலை இணைப்புக் கிளவி	
	Subordinator	CCS	துணைநிலை இணைப்புக் கிளவி	அனால், ஏனென்றால்
	Quotative		மேற்கோள்	என்று
9	Particles	RP	இடைச்சொல்	உம், ஓ, ஆ
	Default	RP_RPD	வழுநிலை	
	Classifier		பாகுபடுத்தி	
	Interjection	RP_INJ	வியப்பிடைச்சொல்	ஆகா, ஐயோ
	Negation	RP_NEG	எதிர்மறை	
	Intensifier	RP_INTF	மிகப்பான்	மிக
10	Quantifiers	QT	அளவையடை	சிறிது, கொஞ்சம்

	General	QT_QTF	பொது	
	Cardinals	QT_QTC	ஆதார எண்	ஒன்று, இரண்டு
	Ordinals	QT_QTO	முறைமை எண்	ஒன்றாவது, இரண்டாவது, இரண்டாம், மூன்றாம்
11	Residuals	RD	மீதி	
	Foreign word	RD_RDF	அயல் மொழிச்சொல்	புக், மினிட்
	Symbol	RD_SYM	குறியீடு	
	Unknown	RD_UNK	தெரியாதது	
	Punctuation	RD_PUNC	நிறுத்தற்குறி	., ; : ?
	Echowords	RD_ECH	எதிரொலிச்சொல்	(காப்பி) கீப்பீ

5.10.2. இலக்கண அடையாளப்படுத்தல்

சொல்வகை அடையாளப்படுத்தலில் இரண்டாவது மட்டம் இலக்கண அடையாளப்படுத்தல் எனப்படும். இது விரிதரவில் உள்ள ஒவ்வொரு சொல்லுக்கும் இலக்கணத் தகவலைத் தரும் கூடுதல் விரிவான குறிப்புரை ஒழுங்குமுறையாகும். ஆங்கிலத்திற்கு சில இலக்கண அடையாளப்படுத்தல் கருவிகள் உருவாக்கப்பட்டுள்ளன (Leech, Garside, and Bryant 1994). இலக்கண அடையாளப்படுத்தல் இறுதி மனிதமுயற்சிசார் மதிப்பீட்டிற்கும் திருத்தலுக்கும் திட்டவரைவு செய்யப்பட்டுள்ளது.

எலும்புக்கூடு பாகுபடுத்தலுக்குத் தேவையான சொற்களை வேறுபடுத்துவதற்கு இந்த திட்டம் பயன்படுத்தப்படுகிறது.

எடுத்துக்காட்டாக:

புற வடிவம் : சொல்லிவிட்டேன் சொல் வகுப்பு : முற்று வினை

வேர்ப்பகுதி : செல் பின்னொட்டு பகுதி : -விட்டேன்

வினையாற்றுவகை குறியீடு : -விட் எச்சக்குறியீடு : -இ

துணைவினை மார்க்கர்: -விட் காலக்குறியீடு : - ட் (இறந்தகாலம்)

நபர் குறியீடு : -ஏன் (1வது) கௌரவக் குறியீடு : பூஜ்யம்

எண் குறிப்பான் : பூஜ்யம் (ஒருமை.) பொருள் :"நான் சொல்லிவிட்டேன்”

உருபனியல்சார், சொல்சார், பொருண்மையியல்சார் மற்றும் தொடரியல்சார் பகுப்பாய்வு தொடர்பான சிக்கல்களைத் தீர்க்க நமக்கு தரவுத்தொகுதி தேவை; இதில் உரையில் நிகழும் ஒவ்வொரு சொல்லின்ன் இலக்கணத் தகவல்களும் அடங்கும்.

5.10.3. சொல் அர்த்தம் அடையாளப்படுத்தல்

சொல் அர்த்தம் அடையாளப்படுத்தல் (Word Sense Tagging) சொல் அடையாளப்படுத்தப்பட்ட மற்றும் இலக்கணம் அடையாளப்படுத்தப்பட்ட உள்ளீட்டு உரையின் சொற்களை ஏற்கும். தானியங்கு அடையாளப்படுத்தலுக்குப் பின் ஒவ்வொரு சொல்லும் சரியான பொருண்மை வகுப்பாக்கத்தை உறதி செய்ய வேண்டி மனித முயற்சிசார் பின்திருத்தம் செய்யப்படும்.

லான்காஸ்டர் பல்கலைக்கழகத்தில் உள்ள ஒரு திட்டத்திலிருந்து சொல் அர்த்த டேக்கிங்கின் பின்வரும் எடுத்துக்காட்டு எடுக்கப்பட்டுள்ளது.

PPIS1	I	Z8
VV0	like	E2+
AT1	a	Z5
JJ	particular	A4.2+
NN1	shade	O4.3
IO	of	Z5
NN1	lipstick	B4

இந்த அட்டவணையில் உரை கீழ்நோக்கி படிக்கப்படுகிறது, இடதுபுறத்தில் இலக்கண குறியீடுகளும், வலதுபுறத்தில் சொல் அர்த்தக் குறியீடுகளும் உள்ளன.

பொருண்மையியல்சார் குறியீடுகள் பொது கருத்தாடல் புலத்தை குறிக்கும் ஒரு பெரிய எழுத்து (upper case letter), ஒரு புலத்தின் முதல் உட்பிரிவைக் குறிக்கும் ஒரு இலக்கம், ஒரு தசம புள்ளி மற்றும் ஒரு சிறந்த துணைப்பிரிவைக் குறிக்க மேலும் இலக்கத்தைத் தொடர்ந்து, ஒரு பொருண்மையியல்சார் அளவில் நேர்மறை அல்லது எதிர்மறை நிலையைக் குறிக்க ஒன்று அல்லது அதற்கு மேற்பட்ட 'பிளஸ்கள்' அல்லது 'கழித்தல்கள்'.

எடுத்துக்காட்டாக, A4.2 + என்பது 'பொது மற்றும் அருவச் சொற்கள்' (A), துணைப்பிரிவு 'பாகுபாடு' (A4), துணை வகைப்பாடு 'குறிப்பிட்ட மற்றும் பொது' (A4.2) மற்றும் 'பொது' என்பதற்கு எதிரான 'குறிப்பிட்ட' (A4.2 +) என்பதைக் குறிக்கிறது. அதேபோல், E2+ வகை

'உணர்ச்சி நிலைகள், செயல்கள், நிகழ்வுகள் மற்றும் செயல்முறைகள்' (E), துணைப்பிரிவு 'விரும்புவது மற்றும் விரும்பாதது' (E2), மற்றும் 'விரும்பாதது' (E2 +) என்பதை விட 'விரும்புவது' என்பதைக் குறிக்கிறது.

5.10.4 அடையாளப்படுத்தலின் முடிவுகள்

தரவுத்தொகுதியில் குறியிடல் முடிந்ததும், மூன்று வகையான சொற்களைக் காண்கிறோம்: சரியாக குறிக்கப்பட்ட சொற்கள், பொருண்மை மயக்கத்துடன் குறிக்கப்பட்ட சொற்கள் மற்றும் குறிக்கப்படாத சொற்கள்.

5.10.4.1 சரியாக அடையாளப்படுத்தப்பட்ட சொற்கள் (Rightly tagged words)

பெரும்பாலான மாற்றுப்பெயர்கள் (PNs), முற்றுப்பெறாவினைகள் (எச்ச வினைகள்) n-FVs முற்றுப்பெற்ற வினைகள்/வினைமுற்றுகள் (FVs), ndls, திரிபுற்ற பெயர்கள் (NNs), வினையடைகள் (ADVs) மற்றும் பெயரடைகள் (ADJs) சரியாக அடையாளப்படுத்தப்படும். உருத்திரிபுறாதவை (indeclinable) கோர்வைப் பொருத்தத்துடன் அடையாளப்படுத்தப்படும். வேர் அகராதிகள் வேர்களையும் பின்னொட்டு அகராதி (SFX lexicon) பின்னொட்டுகளையும் கொண்டிருப்பதால் பெரும்பாலான மாற்றுப்பெயர்கள் (PNs) மற்றும் பெயர்கள் (NNs) சரியாகக் அடையாளப்படுத்தப்படுகின்றன. மேலும், வேர் மற்றும் பின்னொட்டுக்கு இடையிலான வழிமுறை ஒப்புக்கொள்ளப்படுகிறது. வேர் அகராதி வேர்களைக் கொண்டிருப்பதாலும் பின்னொட்டு அகராதி பின்னொட்டுகளைக் கொண்டிருப்பதாலும் பெரும்பாலான வினைமுற்றுகளும் (FVs) வினையெச்சங்களும் (n-FVs) சரியாக அடையாளப்படுத்தப்படும். மேலும், வேருக்கும் பின்னொட்டுக்கும் இடையிலான வழிமுறை ஒப்புக்கொள்ளப்படுகிறது. சில பெயரடைகளும் (ADJs) வினையடைகளும் (ADVs) அதே வழியில் பின்பற்றப்படும்.

5.10.4.2 அடையாளப்படுத்தப்படாத சொற்கள் (Untagged words)

தரவுத்தொகுதியில் உள்ள சில சொற்கள் அடையாளப்படுத்தப்படாது இருக்கும். வேர் அகராதியில் வேர் வடிவம் இல்லாததால் வினைமுற்றுகளும் (FVs) வினையெச்சங்களும் (n-FVs) அடையாளப்படுத்தப்படாது இருக்கும். பெயர்கள் (NNs), மாற்றுப்பெயர்கள் (PNs) மற்றும் பெயரடைகளுக்குப் (ADJ) பின்னொட்டு பட்டியல் முழுமையானதாக இல்லாதிருக்கலாம் அல்லது வேர்ப் பகுதி வேர் அகராதியில் சேமிக்கப்பட்டுள்ள வேருடுடன் பொருந்தாதுவரலாம். உருத்திரிபுறாதவைகளுக்கு (indeclinables) அகராதியில் பற்றாக்குறை இருக்கலாம். வினையடைகளைப் (ADVs) பொறுத்தவரை, சொற்களுக்கு இடையில் உள்ள இடத்தால் தோல்வி

ஏற்படாம். மேலும், வினையடைகளை வேர் மற்றும் பின்னொட்டு அகராதிகளில் பட்டியலிடுவதில் பற்றாக்குறை இருக்கலாம்.

சில பெயர்சொற்கள் (NNs) உரைகளில் முற்றுவினைகளாகப் (FVs) பயன்படுத்தப்படலாம். இயற்பெயர்கள், ஒலிபெயர்ப்புசெய்யப்பட்ட அயற் சொற்கள், கிளைமொழி வடிவங்கள் வேர் அகராதியில் இல்லாததால் அவை அடையாளப்படுத்தப்படாது வரலாம். இவை பெரும்பாலும் தொழில்நுட்பச் சிக்கல்கள் ஆகும்; இச்சிக்கல்களை அந்தந்த வேர் மற்றும் பின்னொட்டு அகராதிகளைப் பெரிதாக்குவதன் மூலமும் பொருந்தக்கூடிய வழிமுறைகளின் மாற்றத்தினாலும் அகற்றலாம்.

5.10.4.3. பொருண்மை மயக்கத்துடன் அடையாளப்படுத்தப்பட்ட சொற்கள்

அனைத்து இயற்கை மொழிகளிலும் சொல்சார் மட்டத்தில் பொருண்மை மயக்கம் மிகவும் பொதுவானது. ஒற்றைச் சொற்கள் அவற்றின் பயன்பாடுகளின் சூழலைப் பொறுத்து பல சொற்கள், அர்த்தங்கள், நிகழ்வுகள் மற்றும் கருத்துக்களை வெளிப்படுத்த முடியும். சொல் செயலியின் (word processor) செயல்திறன் மற்றும் போதுமான தன்மை அது சொல்சார் பொருண்மை மயக்கங்களைக் (lexical ambiguity) கையாளும் முறையிலிருந்து வருகிறது. சொல்வகைப்பாட்டு அடையாளப்படுத்தலில், பொருண்மை மயக்கம் பொதுவாகச் சொல்சார் மட்டத்தில் நடைபெறுகிறது. பெரும்பாலான சொற்கள் [முற்றுவனைகள் (FVs), பெயர்கள் (NNs), பெயரடைகள் (ADJs) போன்றவை] ஒன்றுக்கும் மேற்பட்ட பொருள் அல்லது அர்த்த வேறுபாட்டை அனுமதிப்பதால் சொல்சார் பொருண்மைமயக்கம் ஏற்படுகிறது. இந்த அர்த்தங்கள் துணை வகைப்படுத்தல் அம்சங்கள், தேர்வு அம்சங்கள், தொடரியல் பண்பு, பொருண்மையியல் பண்பு, மரபுத்தொடர் அர்த்தம், உருவகப் பயன்பாடு மற்றும் பலவற்றில் வேறுபடுகின்றன (Sinclair 1991: 104-105).

சொல்வகைப்பாடு அடையாளப்படுத்தலுக்குப் பிறகு, தரவுத்தொகுதி இரண்டு வகையான பொருண்மை மயக்கத்தைக் கொடுக்கும். கட்டமைப்புசார் பொருண்மை மயக்கம் (Structural ambiguity) பெரும்பாலும் திரிபுறாத சொற்களுக்கு (எ.கா. படி 'step", படி ‘read’) ஏற்படுகிறது, இங்கு வேர் ஒரே மாதிரியான வடிவத்தில் வெவ்வேறு சொல்சார் வகைகளுக்கு உரியதாக இருக்கும். வேர் மற்றும் பின்னொட்டு பகுதி இரண்டிலும் உள்ள ஒற்றுமை காரணமாக கட்டமைப்புப் பொருண்மை மயக்கம் சில திரிபுற்ற சொற்களிலும் குறிப்பிடப்படும் (எ.கா. கடலை – கடலை ‘peanut’ கடல்-ஐ "see-acc"; கண்டது ‘that which is seen’ மற்றும் ‘it saw’). சூழல் சார்ந்த

பாகுபடுத்தல் மட்டத்தில் (context-dependent parsing), இத்தகைய பொருண்மை மயக்கதை நீக்க முடியும்.

உடனடியாக பின்வரும் சொல் (W2) இருப்பதால் தொடர்ச்சிசார் பொருண்மை மயக்கம் (sequential ambiguity) ஏற்படுகிறது; இது முந்தைய சொல்லுடன் (W1) செயலாக்கப்பட்டால், அவையவையின் சுதந்திரமான அர்த்தத்திலிருந்து வேறுபட்ட அர்ததை உருவாக்கும். எடுத்துக்காட்டாக, *கால்* மற்றும் *நடைகள்* ஆகியவை தனிமையில் செயலாக்கப்பட்டால், *கால்* என்பது "leg" என்றும், *நடைகள்* என்றால் "walkings" என்றும் பொருள்படும்; அவை ஒன்றாக *கால்நடைகள்* எனச் செயலாக்கப்பட்டால் "cattle" என்று பொருள்படும்; இது அவற்றின் சுதந்திரமான அர்த்தத்தங்களிலிருந்து வேறுபட்டது. தொடர்ச்சியான பொருண்மை மயக்கதைத் தீர்க்க, சூழலில் முக்கியச்சொல் (KWIC) அல்லது குறிப்பிட்ட இடம்சார் சொல் குழுமத்தின் (Local word grouping (LWG)) அடிப்படையில் தாமதச் செயலாக்க முறையைப் (delayed processing) பயன்படுத்துவது நல்லது.

5.11. தலைச்சொல்லாக்கம்

மொழியியலில் தலைச்சொல்லனாக்கம் (லெமடைசேஷன்/lemmatisation) என்பது ஒரு சொல்லின் திரிபுற்ற வடிவங்களை ஒன்றிணைக்கும் செயல்முறையாகும்; எனவே அவை ஒற்றை சொல்லாகப் பகுப்பாய்வு செய்யப்படலாம்; இது சொல்லின் தலைச்சொல்லனாலால் (லெம்மா/lemma) அல்லது அகராதி வடிவத்தால் (dictionary form) அடையாளம் காணப்படுகிறது.

கணினி மொழியியலில், தலைச்சொல்லனாக்கம் என்பது ஒரு சொல்லின் தலைச்சொல்லனை அதன் நோக்கம் சார்ந்த பொருளின் அடிப்படையில் தீர்மானிக்கும் வழிமுறையாகும் (algorithmic process). பகுதியாக்கம் (stemming) போலல்லாமல், ஒரு வாக்கியத்தில் ஒரு சொல்லின் சொல்வகைப்பாடு மற்றும் பொருளின் நோக்கம் சரியாக அடையாளம் காணப்படுவதையும், அண்டை வாக்கியங்கள் அல்லது ஒரு முழு ஆவணம் போன்ற அந்த வாக்கியத்தைச் சுற்றியுள்ள பெரிய சூழலுக்குள்ளும் தலைச்சொல்லனாக்கம் சார்ந்துள்ளது. இதன் விளைவாக, திறமையான தலைச்சொல்லனாக்க வழிமுறைகளை உருவாக்குவது ஆராய்ச்சியின் திறந்த பகுதி ஆகும் (Mullet et al).

பல மொழிகளில், சொற்கள் பல ஊடுருவிய வடிவங்களில் தோன்றும். எடுத்துக்காட்டாக, தமிழில், 'நட' என்ற வினைச்சொல் 'நடந்தது', 'நடக்கிறது', 'நடக்கும்' அல்லது 'நடக்கிறது' என்று தோன்றலாம். ஒரு அகராதியில் ஒருவர் பார்க்கக்கூடிய 'நடை' என்ற அடிப்படை வடிவம், இந்த

வார்த்தையின் தலைச்சொல்லன் என்று அழைக்கப்படுகிறது. சொல்வகைபாட்டுடன் அடிப்படை வடிவத்தின் தொடர்பு பெரும்பாலும் அச்சொல்லின் சொல்லன் என்று அழைக்கப்படுகிறது.

தலைச்சொல்லனாக்கம் பகுதியாக்கத்துடன் (stemming) நெருக்கமாகத் தொடர்புடையது. வித்தியாசம் என்னவென்றால், ஒரு பகுதியாக்கிகள் (ஸ்டெமர்கள்/stemmer) சூழல் பற்றிய அறிவு இல்லாமல் ஒரு சொல்லில் இயங்குகிறது; எனவே சொல்வகைப்பாட்டைப் பொறுத்து வெவ்வேறு அர்த்தங்களைக் கொண்ட சொற்களுக்கு இடையில் பாகுபாடு காட்ட இயலாது. இருப்பினும், பகுதியாக்கிகள் (ஸ்டெமர்கள்) பொதுவாகச் செயல்படுத்த மற்றும் வேகமாக இயக்க எளிதாக இருக்கும். குறைக்கப்பட்ட "துல்லியம்" சில பயன்பாடுகளுக்கு முக்கியமல்ல. உண்மையில், தகவல் மீட்டெடுப்பு அமைப்புகளுக்குள் பயன்படுத்தப்படும்போது, தலைச்சொல்லனாக்கத்துடன் ஒப்பிடும்போது, பகுதியாக்கம் வினவல் நினைவுகூரும் துல்லியத்தை அல்லது உண்மையான நேர்மறை விகிதத்தை மேம்படுத்துகிறது. ஆயினும்கூட, அத்தகைய அமைப்புகளுக்கு பகுதியாக்கம் துல்லியம் அல்லது உண்மையான எதிர்மறை வீதத்தை குறைக்கிறது. எடுத்துக்காட்டாக பின்வருவனவற்றை உற்றுநோக்கவும்.

- "good" என்ற ஆங்கிலச் சொல்லின் தலைச்சொல்லனாக "good" உள்ளது. ஒரு அகராதி பார்வை தேவைப்படுவதால் இந்த இணைப்பு பகுதியாக்கத்தால் தவறுகிறது.
- "walk" என்ற சொல் "walking" என்ற சொல்லின் அடிப்படை வடிவமாகும், எனவே இது பகுதியாக்கம், சொல்லனாக்கம் இரண்டிலும் பொருந்துகிறது.
- "meeting" என்ற சொல் சூழலைப் பொறுத்து ஒரு பெயர்ச்சொல்லின் அடிப்படை வடிவமாகவோ அல்லது ஒரு வினைச்சொல்லின் ("meet") வடிவமாகவோ இருக்கலாம்; எ.கா., " "in our last meeting" (எங்கள் கடைசி சந்திப்பில்)" அல்லது " We are meeting again tomorrow" (நாங்கள் நாளை மீண்டும் சந்திக்கிறோம்)". பகுதியாக்கம் போலல்லாமல், சொல்லனாக்கம் சூழலைப் பொறுத்து சரியான சொல்லனை/லெம்மாவைத் தேர்ந்தெடுக்க முயற்சிக்கிறது.

லூசீன் (Lucene) போன்ற ஆவண அட்டவணைப்படுத்தல் மென்பொருளானது (Document indexing software) சொல்லின் அடிப்படை வடிவ வடிவத்தைப் அர்த்தம் பற்றிய அறிவு இல்லாமல் சேமிக்க முடியும்; ஆனால் சொல் உருவாக்கம் இலக்கண விதிகளை மட்டுமே கருத்தில் கொள்ளும். பகுதியாக்கப்பட்ட சொல் (stemmed word) ஒரு சரியான சொல்லாக இருக்கவேண்டியதில்லை: கீழேயுள்ள எடுத்துக்காட்டில் காணப்படுவது போல் ' 'lazy' ' என்பது பல பகுதியாக்கிகளால் ' lazi '

என்று பகுதியாக்கப்படுகிறது. ஏனென்றால், பகுதியாக்கப்படுவதன் நோக்கம் பொருத்தமான சொல்லனை/லெம்மாவை உருவாக்குவதல்ல - இது சூழல் பற்றிய அறிவு தேவைப்படும் மிகவும் சவாலான பணியாகும். ஒரு சொல்லின் வெவ்வேறு வடிவங்களை ஒரே வடிவத்திற்கு பொருத்துவதே பகுதியாக்கத்தின் முக்கிய நோக்கம். ஒரு விதி அடிப்படையிலான வழிமுறையாக, ஒரு சொல்லின் எழுத்துக்கூட்டலை மட்டுமே சார்ந்துள்ளது; எடுத்துக்காட்டாக, ' laziness' என்பது 'lazi' என்று பகுதியாக்கப்படும் போது, பகுதியாக்கம் 'lazi' என்ற பகுதி இருப்பதை உறுதிசெய்ய துல்லியத்தை தியாகம் செய்கிறது.

தலைச்சொல்லனாக்கச் (lemmatization) செயல்பாடு தரவுத்தொகுதியில் பயன்படுத்தப்பட்டுள்ள சொற்களின் சொல்வகைப்பாட்டை கண்டுபிடிப்பதுடனும் அவற்றுடன் தொடர்புடைய சொல்லன்களாகக் (lexeme) குறைப்பதுடனும் தொடர்புடையதாகும். எடுத்துக்காட்டாக, படித்தான், படிக்கிறான் என்ற சொல்லிலிருந்து 'படி' எனக் குறைத்துத் தலைச்சொல்லாக்கம் செய்தல். இது சொல்லின் எல்லா மாற்று வடிவங்களையும் உள்ளீடு செய்யாமல் ஒரு தலைச்சொல்லின் மாற்று வடிவங்களைப் பிரித்தெடுக்கவும் பரிசோதிக்கவும் ஆய்வாளர்களை அனுமதிக்கிறது. இது மொழி கற்றலுக்குப் பயன் உள்ளதாக அமைகிறது. இங்கு கற்பவர் ஒரு தலைப்புச் சொல்லின் மொத்த சாத்தியமான எண்ணிக்கையில் பயிற்சி பெறுகிறார்கள். இது எந்தச் சொற்கள் திரிபுறுகின்றன, எத்தனை தடவைத் திரிபுறுகின்றன எவ்வாறு திரிபுறுகின்றன என்பதை அறிவதற்குப் பயன் உள்ளதாக அமையும். எடுத்துக்காட்டாக, பிரவுண் தரவுத்தொகுதியில் (Brown Corpus) ஒரு பகுதி சொல் மற்றும் இலக்கண தகவலுடன் கூடிய சொற்களின் தலைப்பாக்கம் செய்யப்பட்ட வடிவங்களைக் கொண்டிருக்கிறது.

5.12. விவரம் அடையாளப்படுத்தல்

தரவுத்தொகுதி அடையாளப்படுத்தல் (corpus annotation) என்பது ஒரு தரவுத்தொகுதியில் விளக்க மொழியியல் தகவல்களைச் சேர்ப்பது. எடுத்துக்காட்டாக, ஒரு பொதுவான வகை அடையாளப்படுத்தல் என்பது அடையாளங்களை (tags) அல்லது புலக்குறிப்புகளைச் (labels) சேர்ப்பதாகும்; இது ஒரு உரையில் உள்ள சொற்கள் எந்த வகுப்பைச் சேர்ந்தது என்பதைக் குறிக்கும்.

தனியான உரைகளுடன் மட்டுமின்றி தரவுத்தொகுதி அடையாளப்படுத்தல் (corpus annotation) என்று அழைக்கப்படுகின்ற கூடுதல் மொழியியல் தரவுகளைக் கொண்டிருக்கும். இத்தகவல்கள் சொல்வகை அடையாளப்படுத்தல் (parts of speech annotation), மீக்கூறு

அடையாளப்படுத்தல் (prosodic Annotation), பொருண்மை அடையாளப்படுத்தல் (Semantic Annotation), முன் வருகிளவி அடையாளப்படுத்தல் (anaphoric annotation), கருத்தாடல் அடையாளப்படுத்தல் (Discourse Annotation) என வேறுபடும். அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி ஆய்வுக்கு மிகப் பயனுள்ளதாக அமையும். இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி குறிப்புரை செய்யப்பட்ட விரிதரவுகளில் மிகப் பொதுவானதாகும். இதில் சொற்களுக்குச் சொல் வகுப்புகள் தரப்பட்டிருக்கும். பிரவுன் தரவுத்தொகுதி (Brown Corpus), லோப் தரவுத்தொகுதி (LOB Corpus), பிரிட்டிஷ் தேசியத் தரவுத்தொகுதி (British National Corpus (BNC) என்பன இலக்கணத்திற்காகக் குறிப்புரை செய்யப்பட்ட தரவுத்தொகுதி ஆகும். பேச்சு ஆங்கிலத்தின் லண்டன்- லண்ட் தரவுத்தொகுதி (London-Lund Corpus of Spoken English (LLC)) மீக்கூறுகளுக்காகக் குறிப்புரை செய்யப்பட்ட தரவுத்தொகுதி ஆகும். சூசானே தரவுத்தொகுதி (Susanne Corpus) தொடரியலுக்காகக் குறிப்புரைசெய்யப்பட்ட தரவுத்தொகுதி ஆகும். இந்திய மொழிகளின் தரவுத்தொகுதி அடையாளப்படுத்தலின் செயல் இன்னும் சரியாகத் தொடங்கபடவில்லை.

5.12.1. சொல் வகைப்பாட்டு அடையாளப்படுத்தல்

சொல்வகைப்பாட்டு குறிப்புரை செய்வதன் (Part of Speech Annotation) நோக்கம் ஒரு உரையில் உள்ள ஒவ்வொரு சொல் அலகுக்கும் அதன் சொல் வகைப்பாட்டை காட்டும் ஒரு குறியனைத் தருவதாகும். இது தரவுத்தொகுதியிலிருந்து தரவை மீட்டெடுப்பதில் தனித்தன்மையைக் காட்டுகிறது மற்றும் தொடரியல் பகுத்துக்குறித்தலுக்கு உதவுகிறது. இது பொருண்மைக்கள குறிப்புரை செய்தலுக்கும் உதவுகிறது. இது ஒற்றுமையான எழுத்துகளை வேறுபடுத்துவதற்கு நம்மை அனுமதிக்கிறது.

5.12..2. முன்வருகிளவி அடையாளப்படுத்தல்

முன்வருகிளவி அடையாளப்படுத்தல் (anaphoric annotation) திட்டத்தில் எல்லா மாற்றுப்பெயர்களும் பெயர்த்தொடர்களும் பற்றாசு (cohesion) என்பதன் விரிவான சட்டகத்திற்குள் ஒரே குறிபடுபொருளைக் குறிப்பதற்காக குறிப்புரை செய்யப்படுகின்றது. இங்கு முன்வருகிளவி வேறுபட்ட வகைகளுக்காகப் பட்டியலிடப்பட்டு வரிசைப்படுத்தப்படுகின்றன. இம்முன்வருகிளவி குறிப்புரை செய்யும் திட்டம் மாற்றுப்பெயர் தீர்வு போன்ற ஆய்வுகளுக்கும் பரிசோதனை செய்யும் இயங்குமுறைகளுக்கும் பயன்படுத்தப்படுகிறது. இது உரையைப் புரிந்து கொள்வதற்கும் இயந்திர மொழிபெயர்ப்புக்கும் முக்கயமானதாகும்.

5.12.3. மீக்கூறு அடையாளப்படுத்தல்

மீக்கூறு அடையாளப்படுத்தலில் (prosodic annotation) இசையோட்ட ஒழுங்கமைப்புகள், அழுத்தம் மற்றும் விட்டிசை என்பன பேச்சில் அடையாளப்படுத்தப்படும். இது மிகச் சிரமமான குறிப்புரை ஆகும்· ஏனென்றால் மீக்கூறு பிற மொழி மட்டங்களை விட இயல்பில் தன்னுணர்ச்சி அடிப்படையானது. இது பயிற்சி பெற்ற காதால் கவனமாகக் கேட்கப்படுதலைக் வேண்டும். Lancoster / IBM Spoken English Corpus மீக்கூறுக்காகக் குறிப்புரை செய்யப்பட்ட விரிதரவாகும். இவற்றில் அழுத்தப்பட்ட அசைகள் சுதந்திரமான மற்றும் சுதந்திரமற்ற இசைமை இயக்கத்துடன் வேறுபட்ட குறியீடுகளால் அடையாளப்படுத்தப்பட்டுள்ளது. சுற்றுப்புற ஒலியழுத்தமுள்ள அசைகளின் சுரக் குறிப்பிலிருந்து இசையை ஊகிக்க இயலும் அழுத்தமில்லா அசைகள் அடையாளப்படுத்தப்படாமல் விடப்பட்டுள்ளன.

5.12.4. பொருண்மை அடையாளப்படுத்தல்

பொருண்மையைக் குறிப்புரை செய்தல் (Semantic Annotation) உரையிலுள்ள சொற்களின் பொருண்மைப் பண்புக்கூறுகளை (குறிப்பாகச் சொல் அர்த்தங்களைக் குறிப்புரை) அல்லது சொற்களுக்கு இடையே உள்ள பொருண்மை உறவுகளைக் (குறிப்பிட்ட செயலின் செயலிகள் பாதிக்கப்படுவர் போன்றவற்றை) குறிப்புரை செய்யும். எடுத்துக்காட்டாக, அவன் பழம் சாப்பிட்டான் என்ற வாக்கியம் பின்வருமாறு பொருண்மைக் குறிப்புரை செய்யப்படும்: அவன் <படர்க்கை, ஆண்பால் ஒருமை, மாற்றுபெயர்> பழம் <தாவரத்தின் பூவிலிருந்து விளையும் இனிப்பான பொருள்> சாப்பிட்டான் <உணவை உட்கொண்டான்> அவன் <எழுவாய்> பழம் <செயப்படுபொருள்> சாப்பிட்டான் <பயனிலை>. எந்தப் பொருண்மை பண்புக்கூறுகள் அடையாளப்படுத்தப்பட வேண்டும் என்பது பற்றி ஒருங்கிணைந்த கருத்து இல்லை. சிலர் Roget's thesaurus-இல் உள்ள பொதுமையான பொருண்மை வகைப்பாடுகளை முன்வைக்கின்றனர். அம்மாதிரியான குறிப்புரை செய்யும் திட்டம் சொற்களின் மூடப்பட்ட வகுப்புகளுக்கும் திறந்த வகுப்புகளுக்கும் நடைமுறைப்படுத்தவேண்டி வடிவமைக்கப்பட்டுள்ளது.

5.12.5. கருத்தாடல் அடையாளப்படுத்தல்

கருத்தாடல் குறிப்புரை செய்தலில் (Discourse Annotation) ஒரு விரிதரவு உரை மற்றும் கருத்தாடல் மட்டங்களில் குறிப்புரை செய்யப்படுகிறது மற்றும் மொழியியல் ஆய்வில் பயன்படுத்தப்படுகிறது. இவ்வகைக் குறிப்புரை கருத்தாடல் ஆய்வில் முக்கிய பங்களிப்பை செய்தாலும் பரவலாகப் பயன்படுத்தப்படவில்லை ஏனென்றால் மொழி வகைப்பாடுகள் சூழல்

சார்ந்தவை மற்றும் அவற்றை உரைகளில் அடையாளம் காணுதல் பிற மொழியியல் நடப்புகளைக் காட்டிலும் விவாதத்திற்குக் கூடுதல் மூல காரணமாக அமையும். வயதுக்குவந்தவர்களின் பேச்சு போக்கைக் கவனிக்க வேண்டி சிலர் London-Lund Spoken Corpus-ஐ 16 கருத்தாடல் அடையாளங்களால் குறிப்புரை செய்துள்ளனர்.

5.13. பகுத்தாய்தல்

பகுத்தாய்தல் (Parsing) என்பது இலக்கண அடிப்படையில் உரைகளின் தானியக்க ஆய்வுடன் தொடர்புடையது (Barnbrook 1998: 170). தொழில் நுட்ப அடிப்படையில் இது ஒரு உரைக்குத் தொடரியல் அமைப்பைத் தரும் செயற்பாங்கை குறிப்பிடுகிறது(McEnery and Wilson 1996: 178). இது பெரும்பாலும் உரையில் அடிப்படை உருபன்-தொடரியல் வகைப்பாடுகள் கண்டுபிடிக்கப்பட்ட பின் நடைமுறைபடுத்தப்படும். வேறுபட்ட இலக்கணங்களின் (சார்பு இலக்கணம், சூழல் கட்டில்லா இலக்கணம், ஒழுங்குமுறையுள்ள செயல்பாட்டு இலக்கணம், நீட்சிசெய்யப்பட்ட ஒட்டு இருக்கணம் போன்றன) அடிப்படையில் பகுத்துக்குறித்தல் இந்த உருபன்-தொடரியல் வகைப்பாடுகளை ஒன்றுடன் ஒன்று உயர்ந்த மட்ட தொடரியல் உறவுகளில் கொண்டுவருகிறது. வாக்கிய நிலை பகுத்துக்குறித்தல் சொல்மட்ட பகுப்பாய்விலிருந்து பெறப்பட்ட தகவலைப் பயன்படுத்தி தானியக்கச் சூழல் அடிப்படையிலான மற்றும் சூழல் சுதந்திரமான தொடரியல் ஆய்வை உட்படுத்தும். ஒரு பகுத்தாயப்பட்ட தரவுத்தொகுதி பகுத்தாய்தலின் கிளைபடங்களைப் பயன்படுத்துவதால் இது கிளைவங்கி (Tree Bank) என்று அழைக்கப்படும். கிளை அமைப்பின் காட்சிப்படம் மிக அரிதாகவே தரவுத்தொகுதி அடையாளப்படுத்தலில் காணப்படும். பொதுவாக ஒத்த தகவல் புலக்குறிப்பு செய்யப்பட்ட அடைப்புக்குறிகளின் குழுமங்களைப் பயன்படுத்தி உருபடுத்தம் செய்யப்படும். எடுத்துக்காட்டு, Pearl sat on a chair என்பது கிளை வங்கியில் பின்வருமாறு தோன்றும்:

[S[NP Pearl_ NP1 NP] [VP sat_VVD [PP on_II [NP a_AT1 chair_ NN1 NP] PP] VP]S]

இதில் உருபனியல்-தொடரியல் தகவல்கள் கீழ்க்கோடுகளால் சொற்களுடன் இணைக்கப்படடுள்ளது. உறுப்புகள் தொடர்களின் தொடக்கத்திலும் இறுதியிலும் முறையே திறக்கும் மற்றும் அடைக்கும் செவ்வக அமைப்புக் குறிகளால் குறிப்புரை செய்யப்பட்டுள்ளது. எல்லாப் பகுத்துக்குறித்தல் ஒழுங்கு முறைகளும் ஒன்றல்ல. முக்கியமான வேறுபாடுகள் பின்வருவன:

1. ஒரு ஒழுங்குமுறை பயன்படுத்தும் உறுப்பு வகைகளின் எண்ணிக்கை

2. இந்த உறுப்பு வகைகள் ஒன்றுடன் ஒன்று இணைவதற்கு அனுமதிக்கப்படும் வழி

இவ்வேறுபாடுகள் இருப்பினும் பெரும்பாலான பகுத்தாய்வுத் திட்டங்கள் சூழல் கட்டுப்பாடு இல்லாத தொடரமைப்பு இலக்கணத்தின் அடிப்படையில் அமைந்தது. இவ்வொழுங்குமுறைக்குள் முழு பகுத்தாய்வுத் (full Parsing) திட்டம் வாக்கிய அமைப்பின் விரிவான ஆய்வைத் தருவதை நோக்கமாக கொண்டது. குறுகிய பகுத்தாய்வுத் திட்டம் (Skeleton Parsing Scheme) தொடரியல் உறுப்பு வகைகளில் குறைந்த நிலையிலான வேறுபடுத்தப்பட்ட குழுமத்தைப் பயன்படுத்துகிறது மற்றும் சில உறுப்பு வகைகளின் அக அமைப்பை விட்டுவிடுகிறது. பகுத்துக்குறித்தல் பெரும்பாலும் மனித ஆய்வால் பின்திருத்தம் செய்யப்படுகிறது; ஏனென்றால் தனியங்கி பகுத்தாய்வு சொல் வகை அடையாளப்படுத்தலை விட வெற்றி விகித்தில் குறைந்தது. முழு மினித இயக்கப் பகுத்தாய்வின் குறைபாடு, தரவுத்தொகுதி பகுத்தாய்வு அல்லது திருத்தல் என்பனவற்றில் ஈடுப்படும் ஆய்வாளரின் ஒழுங்கின்மையாகும். இதை ஈடுகட்ட விரிவான வழிமுறைகள் தரப்படுகின்றன. இருப்பினும் பல்பொருள்கோள் சாத்தியமாகும் போது மயக்கம் ஏற்படலாம். கிளை வங்கிகள் இயற்கை மொழிகளுக்கு அதன் அமைப்பின் சொல் மட்டம், தொடர்மட்டம், வாக்கியமட்டம், செயல்பாடு-பங்கெடுப்பாளர் அமைப்பு மட்டம் என்ற வேறுபட்ட அமைப்புகளின் மட்டங்களில் குறிப்புரைகளைத் தரும் மொழி மூலவளமாக அமைகிறது. கிளைவங்கிகள் இயற்கை மொழி ஆய்வில் தரவு-இயக்க அணுகுமுறைகள், மனித மொழி தொழில் நுட்பங்கள், இலக்கணப் பிரித்தெடுப்பு மற்றும் பொதுவான மொழியியல் ஆய்வுகள் இவற்றின் முன்னேற்றதத்திற்கு முக்கியமானதாகும்.

பயன்பாடுகள்

தரவுத்தொகுதி மொழியியலில் தரவுத்தொகுதகள் முக்கிய அறிவுத் தளமாகும். பயன்பாட்டின் பிற குறிப்பிடத்தக்க பகுதிகள் பின்வருமாறு:

மொழி தொழில்நுட்பம், இயற்கை மொழி ஆய்வு, கணினி மொழியியல்

பல்வேறு வகையான தரவுத்தொகுதிகளின் பகுப்பாய்வு மற்றும் செயலாக்கம் (analysis and processing) கணினி மொழியியல் (computational linguistics), பேச்சு அறிதல் (speech recognition) மற்றும் இயந்திர மொழிபெயர்ப்பு (machine translation) ஆகியவற்றில் அதிக வேலைக்கு உட்பட்டவை; அவை பெரும்பாலும் சொல்வகைப்பட்டு அடையாளப்படுத்தல் மற்றும் பிற நோக்கங்களுக்காக மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகளை (Hidden Markov Model (HMM)) உருவாக்கப் பயன்படுகின்றன. அவற்றிலிருந்து பெறப்பட்ட தரவுத்தொகுதி மற்றும் நிகழ்வெண்

பட்டியல்கள் மொழி கற்பிப்பதற்கு பயனுள்ளதாக இருக்கும். தரவுத்தொகுதியை ஒரு வகை வெளிநாட்டு மொழி எழுதும் உதவியாகக் கருதலாம்; ஏனெனில் தரவுத்தொகுதிகளில் உள்ள உண்மையான பனுவல்களை வெளிப்படுத்துவதன் மூலம் சொந்த மொழி அல்லாத பயனர்களால் பெறப்பட்ட சூழ்நிலைப்படுத்தப்பட்ட இலக்கண அறிவு கற்றவர்களுக்கு இலக்கு மொழியில் வாக்கியத்தை உருவாக்கும் முறையைப் புரிந்துகொள்ள உதவுகிறது, மேலும் பயனுள்ள எழுத்தை செயல்படுத்துகிறது.

இயந்திர மொழிபெயர்ப்பு

பக்கவாட்டு ஒப்பீட்டிற்காக சிறப்பாக (side-by-side comparison) வடிவமைக்கப்பட்ட பன்மொழி தரவுத்தொகுதிகள் சீரமைக்கப்பட்ட இணையான தரவுத்தொகுதிகள் (aligned parallel corpora) என்று அழைக்கப்படுகிறன. இரண்டு மொழிகளில் உரைகளைக் கொண்ட இணையான தரவுத்தொகுதிகளில் இரண்டு முக்கிய வகைகள் உள்ளன. ஒரு மொழிபெயர்ப்பு தரவுத்தொகுதியில் (translation corpus) ஒரு மொழியில் உள்ள பனுவல்கள்/உரைகள் மற்ற மொழியில் உள்ள உரைகளின் மொழிபெயர்ப்புகளாகும். ஒப்பிடக்கூடிய தரவுத்தொகுதியில், பனுவல்கள் ஒரே மாதிரியானவை மற்றும் ஒரே உள்ளடக்கத்தை உள்ளடக்கியது; ஆனால் அவை ஒன்றுக்கொன்றான மொழிபெயர்ப்புகள் அல்ல. ஒரு இணையான உரையை பயன்படுத்த, சமமான உரைக் கூறுகளை (equivalent text segments) (சொற்றொடர்கள் அல்லது வாக்கியங்கள்) அடையாளம் காணும் ஒருவித உரை சீரமைப்பு (text alignment) பகுப்பாய்விற்கு ஒரு முன்நிபந்தனை ஆகும். இரண்டு மொழிகளுக்கு இடையில் மொழிபெயர்ப்பதற்கான இயந்திர மொழிபெயர்ப்பு வழிமுறைகள் பெரும்பாலும் முதல் மொழி தரவுத்தொகுதியையும் இரண்டாவது மொழி தரவுத்தொகுதியையும் உள்ளடக்கிய இணையான துண்டுகளைப் (parallel fragments) பயன்படுத்தி பயிற்சியளிக்கப்படுகின்றன, இது முதல் மொழித் தரவுத்தொகுதியின் உறுப்பு-க்கு-உறுப்பு (element-for-element translation) மொழிபெயர்ப்பாகும்.

மொழிநூல்

வரலாற்று ஆவணங்களின் (historical documents) ஆய்விலும் உரைத் தரவுத்தொகுதிகள் பயன்படுத்தப்படுகின்றன; எடுத்துக்காட்டாக பண்டைய எழுத்துகளைப் புரிந்துகொள்ளும் முயற்சிகளில் அல்லது விவிலியப் புலமையில். குறுகிய காலத்தைசார்ந்த சில தொல்பொருள் தரவுத்தொகுதிகள் அவை சரியான நேரத்தில் ஒரு புகைப்படத்தை/ஸ்னாப்ஷாட்டை வழங்குகின்றன. காலத்தின் அடிப்படையில் மிகக் குறுகிய தரவுத்தொகுதிகளில் ஒன்று, 15-30

ஆண்டு அமர்னா கடிதப் பனுவல்கள் (Amarna letters texts) (கிமு 1350) ஆகும். ஒரு பண்டைய நகரத்தின் தரவுத்தொகுதி, (எடுத்துக்காட்டாக துருக்கியின் "கோல்டெப் உரைகள்" ("Kültepe Texts")) அவை கண்டுபிடிக்கப்பட்ட தளத் தேதிகளால் தீர்மானிக்கப்பட்ட தொடர்ச்சியான தரவுத்தொகுதிகளின் வழியாக செல்லக்கூடும்.

5.14. சுருக்கவுரை

இந்த இயல் உரை ஆய்தல் பற்றி விளக்குவதை நோக்கமாகக் கொண்டுள்ளது. உரை ஆய்தல் பற்றி ஒரு சுருக்கமான அறிமுகம் முதலில் தரப்பட்டுள்ளது. அதைத் தொடர்ந்து, அதிர்வெண் ஆய்வு, சொல் வரிசைப்படுத்தல், தொடரடைவு, சொல்சார் சொல்லடி வகைப்பாடு, சூழலில் முக்கியச் சொல், குறிப்பிட்ட இடம்சாந்த சொற்களைத் தொகுத்தல், சொல் ஆய்தல், அடையாளப்படுத்தல், சொல்லனாக்கம், விவரம் அடையாளப்படுத்தல், பகுப்பாய்வுசெயதல், ஆகியன குறித்து விரிவான விளக்கங்கள் தர்ப்பட்டுள்ளன. தரவுத்தொகுதி உரைகளை பலவித செயற்பாங்குகளுக்கு அல்லது ஆய்வுகளுக்கு உட்படுத்துவது இயற்கை மொழியை ஏதேனும் ஒரு வழியில் பயன்படுத்த முயற்சிக்கும் எந்தவொரு மொழிச் செயலாக்க முறைக்கும் அதிக முக்கியத்துவம் வாய்ந்தது ஆகும். பயனர்களின் மேம்பட்ட தேவைகள் திறமையான மற்றும் பரவலாகப் பொருந்தக்கூடிய அமைப்புகளின் தேவையை உயர்த்துகின்றன. விரிவான செயலாக்கத் திறன்களுக்கான தேவை கோட்பாட்டு, பயன்பாட்டு மற்றும் கணினி மொழியியலில் வலுவான இடைமுகத்தைக் கொண்டுள்ளது. இயற்கை மொழிகளின் சிக்கலான தன்மையைக் கருத்தில் கொண்டு, ஒரு மொழியின் எந்தவொரு பண்புக்கூறு பற்றியும் துல்லியமான முடிவுகளை எடுப்பது எந்திரத்திற்கும் எப்போதும் கடினம். ஆகையால், செயலாக்கத்தில் அவ்வப்போது ஏற்படும் பிழைகள் மொழி செயலாக்கத்தில் ஆராய்ச்சி செய்வதற்கு ஒரு பெரிய தடையாகக் கருதக்கூடாது. பிழைகளைச் சரிபார்க்கக் குறிப்பாக வடிவமைக்கப்பட்ட ஒரு ஊடாடும் கணினி நிரல் இந்த செயல்முறையை மிக வேகமாகவும் நம்பகத்தன்மையுடனும் மாற்றும்.

இயல் 6

மொழித் தொழில் நுட்பத்தில் தரவுத்தொகுதி

6.1 அறிமுகம்

மொழித் தொழில் நுட்பத்தில் தரவுத்தொகுதியின் பங்களிப்பு இன்றைய காலகட்டத்தில் மிக சிறப்பாகவும் பரவலாகவும் பேசப்படும் மற்றும் பயன்படுத்தப்படும் நெறிமுறையாகும். மொழி புரிந்துகொள்தல், பேச்சு புரிந்து கொள்ளுதல், உரை மீட்பு மற்றும் புரிந்துகொள்ளுதல், உரைகளிலிருந்து தகவல்களை மீட்டல், ஒலிவழி எழுத்து புரிந்துகொள்ளுதல் இயந்திரமொழிபெயர்ப்பு போன்றவற்றை உள்ளடக்கும். கணிப்பொறி அடிப்படையிலான மொழியியல் தொழில்நுட்பத்தில் ஆர்வம் கூடி வருகிறது. இருப்பினும் மனித மொழியில் உட்படும் கலவைத்தன்மைகள் மற்றும் சிக்கல்கள் காரணமாக மொழி ஆய்வுக்காகத் திட்டமிடப்படும் கணினி வழியமைப்பு முறைகள் வேறுபட்ட மொழித் தரவுகளின் மிகக் கூடுதலான அளவை வேண்டும். இதன் காரணமாக அளவீட்டு மொழியியல் என்று அழைக்கப்படும் தரவுத்தொகுதி மொழியியல் பயன்பாட்டு மொழியியலின் புதிய கிளையாக உருவாகியுள்ளது. எல்லா மொழிகளிலும் தரவுத்தொகுதி உருவாக்கப்பட்டு அவற்றை மொழியின் பல நிலைகளிலும் அடையாளப்படுத்தும் முயற்சி நடைப்பெற்று வருகிறது. இத்தகைய தரவுத்தொகுதி மொழிக் கருவிகள் தயாரிப்பதற்கு மிகப்பயனுள்ளதாக அமைகிறது. இந்நெறிமுறை மொழித் தொழில் நுட்பமாக வளர்ந்துள்ளது.

நம் சமூகத்தில் மொழி தொழில்நுட்பத்தின் பங்கைப் புரிந்துகொள்வதற்கு பல்வேறு சிக்கல்களைக் கருத்தில் கொள்ள வேண்டும். ஒவ்வொரு தசாப்தத்திலும் தகவல் சேமிப்பு மற்றும் செயலாக்கம் மலிவானதாகி வருகிறது. கணினித் தொழில்நுட்பம் தொடர்ந்து சில காலம் சமூகத்தில் ஊடுருவி மறுவடிவமைக்கும் என்பதை நாம் அறிவோம். நாம் பேசுவதற்கும் கேட்பதற்கும், படிப்பதற்கும் எழுதுவதற்கும் அதிக நேரம் செலவிடுகிறோம். கணினி நம் வாழ்க்கைக்கும் சமூகத்திற்கும் மேலும் மேலும் மையமாகி வருகிறது. இது நாம் பேசு மற்றும் எழுத்துத் தகவல்தொடர்புகளில் - தொலைபேசி மாறுதல் மற்றும் பரிமாற்ற முறைமை, மின்னணு அஞ்சல், சொல் செயலாக்கம் மற்றும் மின்னணு வெளியீடு, முழு உரை தகவல் மீட்டெடுப்பு மற்றும் கணினி புல்லட்டின் பலகைகள் மற்றும் பலவற்றில் ஊடாட்டம் செய்கிறது.

இந்த போக்குகள் இயற்கை மொழி மற்றும் பேச்சு தொழில்நுட்பத்திற்கான மகத்தான பொருளாதார மற்றும் சமூக வாய்ப்பை உருவாக்குகின்றன. பேச்சு மற்றும் உரையை

உருவாக்குதல், பரிமாற்றம் செய்தல், சேமித்தல், தேடுவது அல்லது பேசு மற்றும் உரையை மீட்டுருவாக்கம் செய்வதில் கணினிகள் ஏற்கனவே ஈடுபட்டுள்ளன.

மனித தொடர்பு பேச்சு மற்றும் உரையை அடிப்படையாகக் கொண்டது; மேலும் கணினிகள் பெருகிய முறையில் அதிநவீன வழிகளில் ஈடுபட்டுள்ளன. மொழியியல் தொழில்நுட்பத்தில் எந்தவொரு அடிப்படை மேம்பாடுகளும் முக்கியமான சமூக மற்றும் பொருளாதார பங்குகளைக் காணும் என்பதற்கு இது உத்தரவாதம் அளிக்கிறது. குறைந்த அளவு செலவில், செயல்முறையின் தரத்தை மேம்படுத்த அல்லது சம்பந்தப்பட்ட மனித உழைப்பின் உற்பத்தித்திறனை அதிகரிக்க புதிய அம்சங்களைச் சேர்க்கிறோம்.

இந்த வகையான எளிய எடுத்துக்காட்டுகள் பின்வருவனவற்றை உட்படுத்தும்: சொல் செயலாய்வு (word processing) மூலம் எழுத்துப்பிழைகளைச் சரிசெய்தல்; குரல் அறிதலுடன் (voice recognition) அழைப்புகளைத் திரையிடுவதன் மூலம் அல்லது குரல் தொகுப்பு/உருவாக்கம் மூலம் தகவல்களை வழங்குவதன் மூலம் தொலைபேசி உதவியாளர்களின் பணிச்சுமையைக் குறைக்கப் பேச்சு தொழில்நுட்பத்தைப் பயன்படுத்துதல்; மற்றும் இயந்திர உதவிபெறும் மொழிபெயர்ப்பு (machine-aided translation (MAT) ஒழுங்குமுறைகள் மூலம் எளிதில் திருத்த இயலும் மொழிபெயர்ப்பு வெளியீட்டை வழங்கி மனித மொழிபெயர்ப்பாளர்களை அதிக உற்பத்தி செய்யச்செய்தல்.

மொழி அறிதல் மற்றும் புரிதல், பேச்சு அறிதல் மற்றும் புரிதல், உரை மீட்டெடுப்பு மற்றும் புரிதல், உரைகளிலிருந்து தகவல்களைப் பெறுதல், ஒளிவழி எழுத்து அறிதல் (optical character recognition), இயந்திர மொழிபெயர்ப்பு போன்றவற்றை உள்ளடக்கிய கணினி அடிப்படையிலான மொழியியல் தொழில்நுட்பத்தில் ஆர்வம் அதிகரித்து வருகிறது.

இருப்பினும், மனித மொழியில் உள்ள கலவைத்தன்மை மற்றும் சிக்கல்கள் காரணமாக, அதை செயலாக்க வடிவமைக்கப்பட்ட கணினி நிரல்கள் பலவிதமான மொழியியல் தரவுகளான பேச்சு, உரை, அகராதிகள் மற்றும் இலக்கணங்களுடன் வலுவானதாகவும் பயனுள்ளதாகவும் இருக்க வேண்டும். இத்தகைய தரவுத்தளங்கள் உருவாக்கவும் ஆவணப்படுத்தவும் செலவுமிக்கவை; பராமரிப்பும் விநியோகமும் கூடுதல் செலவுகளைச் சேர்க்கின்றன.

சமீப காலம் வரை, ஆர்வமுள்ள ஆராய்ச்சியாளர்களால் பயன்படுத்த பெரும்பாலான மொழியியல் வளங்கள் எளிதில் கிடைக்கவில்லை. சொத்துரிமை (property right) அல்லது மின்னணு வெளியீட்டின் கூடுதல் சுமைகள் காரணமாக, தனிப்பட்ட ஆராய்ச்சியாளர்களால்

தயாரிக்கப்பட்ட மொழியியல் தரவுத்தளங்கள் பெரும்பாலானவை விநியோகிக்கப் படாமல் ஒரு ஆய்வகத்திலேயே இருக்கின்றன; அல்லது சில ஆராய்ச்சியாளர்களுக்கு வழங்கப்பட்டு ஆனால் மற்றவர்களுக்கு மறுக்கப்பட்டுள்ளன. கடந்த ஆண்டுகளில் குறிப்பிடத்தக்க சில எடுத்துக்காட்டுகள் பகிரப்பட்ட வளங்களின் மதிப்பை நிரூபித்துள்ளன; ஆனால் இவை விதிக்கு மாறாக விதிவிலக்குகள் ஆகும். எடுத்துக்காட்டாக, பிரவுன் தரவுத்தொகுதி பல ஆராய்ச்சியாளர்களால் ஆங்கிலத்தின் மொழி மாதிரிகளை மதிப்பீடு செய்யப் பயன்படுத்தப்படுகிறது.

யு.எஸ். பாதுகாப்புத் துறையின் மேம்பட்ட ஆராய்ச்சி திட்ட முகமை (Advanced Research Projects Agency (ARPA)) மதிப்பீடு மற்றும் பயன்பாட்டிற்காக தொடர்ச்சியான பகிரப்பட்ட பேச்சு தரவுத்தளங்களை உருவாக்கியது. இது பேச்சு அறிதலில் விரைவான முன்னேற்றத்திற்கு வழிவகுத்தது; மேலும் செய்தி புரிதல், ஆவண மீட்டெடுப்பு, பேச்சு புரிதல் மற்றும் இயந்திர மொழிபெயர்ப்பு ஆகியவற்றில் ஆராய்ச்சிக்குப் பயன்படுத்தப்பட்டது.

6.2. மொழித்தொழில் நுட்பத்தில் தரவுத்தொகுதியின் முக்கியத்துவம்

வேறுபட்ட வகையான தரவுத்தொகுதி உருவாக்கத்தால் மொழித்தொழில் நுட்ப ஆய்வும் பயன்பாட்டிலும் தரவுத்தொகுதியின் முக்கியத்துவம் பெருகியுள்ளது. மொழி ஆய்வுக்கான கருவிகள் மற்றும் ஒழுங்கு முறைகள் உருவாக்குவதில் மொழி தரவுத்தொகுதி பெரிதும் பங்களிப்பு செய்கிறது. பயன்பாட்டு அடிப்படையிலான தொழில் நுட்பச் செயலுக்கு தரவுத்தொகுதி உயர்ந்த விளைவைத் தருகிறது. பொதுவாக நாம் தரவுத்தொகுதியின் பயன்பாட்டை இயந்திரத்தால் கட்டுப்படுத்தப்பட்ட கருவிகளை திட்டமிடுவதற்கும் தானியங்கு கருவிகளைப் பரிசோதிப்பதற்கும் பயிற்சி தரவும் சிறப்பான மூலவளமாக காண்கிறோம் தரவுத்தொகுதியிலிருந்து பெறப்பட்ட மொழிப் பண்புக்கூறுகளின் நிகழ்வெண்ணிக்கை மொழி கற்போருக்கு நூல்களைத் திட்டமிடவும், ஒளிவழி எழுத்துக்களைப் புரிந்துகொள்ளும் ஒழுங்குமுறைகளை (OCR) உருவாக்குவதற்கும் தானியங்கு எழுத்துப்பிழைத் திருத்திகளை உருவாக்குவதற்கும் பயனுள்ளதாக இருக்கிறது. குறிப்புரை செய்யப்பட்ட மற்றும் குறிப்புரை செய்யப்படாத விரிதரவுகள் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையை திட்டமிடவும் இயந்திரத்தால் படிக்கப்பெறும் அகராதிகளைத் (Machine Readable Dictionaries) MRDS)) திட்டமிடவும் உருபனியல் ஆய்விகள், சொல் ஆய்விகள், வாக்கியப் பகுத்துக்குறிப்பான்கள் என்பனவற்றை உருவாக்கவும் பயனுள்ளதாக அமைகிறது. விரிதரவைப் பயன்படுத்தி உருவாக்கப்பட்ட பல கணினி கருவிகளும் ஒழுங்குமுறைகளும்

மொழியைப் பயன்படுத்துபவர்களுக்கும் ஆய்வாளர்களுக்கும் எழுத்தாளர்களுக்கும் கல்வியாளர்களுக்கும் ஆசிரியர்களுக்கும் மாணவர்களுக்கும் அறிஞர்களுக்கும் அச்சிட்டு வெளியிடுபவர்களுக்கும் மொழி கற்பவர்களுக்கும் மற்றும் பிறருக்கும் மிகுந்த பயனுள்ளதாக அமைகிறது. பொதுவாக தரவுத்தொகுதியின் முக்கியதுவத்தை மொழித் தொழில்நுட்பத்தின் நான்கு விரிந்த களங்களாகப் பகுக்கலாம். இது அவற்றின் இயல்பு மற்றும் பயன்பாட்டு நோக்கு அடிப்படையில் செய்யப்படும்: அ) அறிவு வளமாகத் தரவுத்தொகுதி, (ஆ) மொழித்தொழில் நுட்பக் கருவிகள் மற்றும் அமைப்புகளை வடிவமைப்பதற்கான ஆதாரமாகத் தரவுத்தொகுதி, (இ) மொழிபெயர்ப்பு ஆதரவு ஒழுங்குமுறைக்கான ஆதாரமாகத் தரவுத்தொகுதி, (ஈ) மனித-இயந்திர இடைமுக ஒழுங்குமுறைகளுக்கான ஆதாரமாகத் தரவுத்தொகுதி.

6.3. அறிவின் மூலவளமாகத் தரவுத்தொகுதி (Corpus as a Knowledge Resource)

தரவுத்தொகுதி மொழியியல் என்பது மின்னணு தரவுத்தளமாக (அதாவது ஒரு தரவுத்தொகுதி) சேமிக்கப்பட்ட இலக்கு மொழியின் பிரதிநிதி மாதிரியை அடிப்படையாகக் கொண்ட மொழி பகுப்பாய்விற்கான அனுபவ அணுகுமுறையாகும் (பைபர், கான்ராட், & ரெப்பன்/ Biber, Conrad, & Reppen, 1998). இந்த துறையில் பெரும்பாலான ஆய்வுகள் ஆயிரக்கணக்கான, மில்லியன் மற்றும் சில நேரங்களில் பில்லியன் கணக்கான சொற்களை உள்ளடக்கிய மிகப் பெரிய தரவுத்தொகுதிகளில் மொழியியல் அம்சங்களின் அளவுசார் பகுப்பாய்விற்கு (quantitative analysis) கணினி மென்பொருளை நம்பியுள்ளன. இருப்பினும், சிறிய அளவிலான ஆய்வுகள் சில நேரங்களில் கணினிகளின் உதவியுடன் அல்லது இல்லாமல் தரவுத்தொகுதி உரைகளின் மனிதப் பகுப்பாய்வு மூலம் மேற்கொள்ளப்படுகின்றன. தரவுத்தொகுதி ஆராய்ச்சியாளர்கள் அளவுசார் முடிவுகளை பண்புசார் பகுப்பாய்வு மற்றும் அந்த முடிவுகளின் விளக்கத்துடன் துணைநிறைவாகக் கண்டுபிடிப்பதும் பொதுவானது. தரவுத்தொகுதி மொழியியல் துறையில் உருவாக்கப்பட்ட வளங்கள், கருவிகள் மற்றும் நுட்பங்கள் அறிவு வளத்தை உருவாக்குவதில் குறிப்பாக முக்கிய பங்கு வகிக்கின்றன.

அறிவு மூலவளமாக தரவுத்தொகுதி பின்வரும் செயல்பாடுகளுக்கு பயன்படுகிறது: 1. பன்மொழி நூலகங்களை உருவாக்குதல், 2. மொழி கற்பவர்களுக்குப் பாடநூல்களைத் திட்டமிடல், 3. (அச்சிட்ட மற்றும் மின்வடிவ) ஒருமொழி அகராதிகளை உருவாக்குதல், 4. (அச்சிட்ட மற்றும் மின் வடிவ) இருமொழி அகராதிகளை உருவாக்குதல், 5. (அச்சிட்ட மற்றும் மின் வடிவ) பன்மொழி அகராதிகளை உருவாக்குதல், 6. (அச்சிட்ட மற்றும் மின்வடிவ) ஒருமொழிச்

சொற்களஞ்சியங்களை (பொருட்புல அகராதிகளை) உருவாக்குதல், 7. பல்வேறு விதமான (அச்சிட்ட மற்றும் மின் வடிவ) நோக்கீட்டுப் பொருள்கள் உருவாக்குதல், 8. இயந்திரத்தால் படிக்கவியலும் அகராதிகளை உருவாக்குதல், 9. பன்மொழிச் சொல் மூலவளங்களை உருவாக்குதல், 10. மின் அகராதிகளை உருவாக்குதல்.

சொல்சார் வளம்

டிஜிட்டல் அகராதி, இயற்கை மொழி செயலாக்கம் மற்றும் டிஜிட்டல் மனிதநேயங்களில், ஒரு சொல்சார் வளமானது ஒன்று அல்லது பல அகராதிகளைக் கொண்ட ஒரு மொழி வளமாகும், எ.கா., ஒரு தரவுத்தளத்தின் வடிவத்தில்.

பண்புகள்

சொல்சார் வளங்களின் இயந்திரத்தால் படிக்கக்கூடிய பதிப்பிற்கான வெவ்வேறு தரநிலைகள் உள்ளன, எ.கா., லெக்சிகல் மார்க்அப் ஃபிரேம்வொர்க் (Lexical Markup Framework (LMF/எல்எம்எஃப்) லெக்சிக்கல் வளங்களை குறியாக்கம் செய்வதற்கான ஒரு ஐஎஸ்ஓ தரநிலை, இதில் ஒரு சுருக்க தரவு மாதிரி மற்றும் எக்ஸ்எம்எல் சீரியலைசேஷன் (XML serialization) ஆகியவை அடங்கும், மற்றும் இணையத்தில் அறிவு வரைபடங்களாக சொல்சார் வளங்களை வெளியிடுவதற்கு ஆர்.டி.எஃப் சொற்றொகையான ஒன்டோலெக்ஸ்-லெமன் (OntoLex-Lemon), எ.கா., மொழியியல் இணைக்கப்பட்ட திறந்த தரவு (Linguistic Linked Open Data).

நேரிடப்படும் மொழிகளின் வகையைப் பொறுத்து, ஒரு சொல்சார் வளத்தை ஒருமொழிய சொல்சார் வளம், இருமொழிய சொல்சார் வளம் அல்லது பன்மொழிய சொல்சார் வளம் என தகுதிபெறலாம். இருமொழி மற்றும் பன்மொழி சொல்சார் வளங்களுக்கு, சொற்கள் இணைக்கப்படலாம் அல்லது ஒரு மொழியிலிருந்து மற்றொரு மொழியுடன் இணைக்கப்படாமல் இருக்கலாம். இணைக்கப்படும்போது, ஒரு மொழியிலிருந்து இன்னொமொழிக்கான நிகரன் இருமொழி இணைப்பு மூலம் அல்லது பன்மொழி குறிப்புகள் மூலம் செய்யப்படுகின்றது .

ஒரே மொழியின் வெவ்வேறு அகராதிகளைக் கொண்ட ஒரு சொல்சார் வளத்தை உருவாக்கவும் நிர்வகிக்கவும் முடியும், எடுத்துக்காட்டாக, பொதுவான சொற்களுக்கு ஒரு அகராதி மற்றும் வெவ்வேறு சிறப்பு களங்களுக்கான ஒன்று அல்லது பல அகராதிகள்.

இயந்திரம் படிக்கக்கூடிய அகராதிக்கு எதிராக இயற்கைமொழி ஆய்வு அகராதி

டிஜிட்டல் அக்ராதியலில் உள்ள சொல்சார் வளங்கள் பெரும்பாலும் இயந்திரத்தால் படிக்கக்கூடிய அகராதி (machine-readable dictionary (MRD/எம்ஆர்டி) என

குறிப்பிடப்படுகின்றன, இது ஒரு அகராதி காகிதத்தில் அச்சிடப்படுவதற்கு பதிலாக இயந்திர (கணினி) தரவுகளாக சேமிக்கப்படுகிறது. இது ஒரு மின்னணு அகராதி மற்றும் சொல்சார் தரவுத்தளமாகும். இயந்திரத்தால் படிக்கக்கூடிய அகராதி என்ற சொல் பெரும்பாலும் இயற்கைமொழி ஆய்வு அகராதியுடன் முரண்படுகிறது. இரண்டுமே திட்டங்களால் பயன்படுத்தப்பட்டாலும், இயந்திரத்தால் படிக்கக்கூடிய அகராதி என்பது காகிதத்தில் அச்சிடப்பட்ட அகராதியின் மின்னணு வடிவமாகும். இதற்கு மாறாக, இயற்கை மொழி ஆய்வை மனதில் கொண்டு அகராதி புதிதாக உருவாக்கப்பட்டபோது இயற்கைமொழி ஆய்வு அகராதி என்ற சொல் விரும்பப்பட்டது.

சொல்சார் தரவுத்தளம் (lexical database)

ஒரு சொல்சார் தரவுத்தளம் என்பது ஒரு சொல்சார் வளமாகும், இது தொடர்புடைய மென்பொருள் சூழல் தரவுத்தளத்தைக் கொண்டுள்ளது, இது அதன் உள்ளடக்கங்களை அணுக அனுமதிக்கிறது. தரவுத்தளமானது சொல்சார் தகவலுக்காக அல்லது சொல்சார் தகவல் உள்ளிடப்பட்ட பொது நோக்கத்திற்கான தரவுத்தளத்திற்காக தனிப்பயன் வடிவமைக்கப்பட்டதாக இருக்கலாம். ஒரு சொல்சார் தரவுத்தளத்தில் பொதுவாகச் சேமிக்கப்படும் தகவல்களில் சொல்வகை மற்றும் சொற்களின் ஒருபொருள்பன்மொழிகள், அத்துடன் வெவ்வேறு சொற்கள் அல்லது சொற்களின் தொகுப்புகளுக்கு இடையிலான பொருண்மையியல் மற்றும் ஒலியனியல் உறவுகள் ஆகியவை அடங்கும்.

அகராதி (Dictionary)

ஒரு அகராதி சொல்சார் அலகால் வரிசைப்படுத்தப்பட்ட பதிவுகளின் பட்டியலைக் கொண்டுள்ளது. ஒவ்வொரு பதிவும் வழக்கமாக ஒரு சொல் அலகு, அதனுடன் தொடர்புடைய வரையறை, சொல்வகைப்பாடு (part-of-speech (POS/பிஓஎஸ்), உச்சரிப்பு, சொற்களின் பயன்பாடுகளைக் காட்டும் எடுத்துக்காட்டுகள் மற்றும் கூடுதல் தகவல்களைக் கொண்டிருக்கும். ஒரு சொல்சார் அலகு, பொதுவாக ஒரு சொல்லாக இருக்கும்; அதேசமயம் அதன் வரையறை ஒரு சொல்லாகவோ கூட்டுச் சொல்லாகவோ சொற்றொடராகவோ பல சொல் வெளிப்பாடாகவோ அல்லது ஒரு சொற்றொடராகவோ இருக்கும். ஒரு ஒருமொழி அகராதிக்கு ஆக்ஸ்போர்டு ஆங்கில அகராதி போன்ற ஒரே ஒரு மொழிய அகராதி எடுத்துக்காட்டாக அமையும். "கிரியாவின் தற்காலத் தமிழ் அகராதி" போன்ற இருமொழிய அகராதி இரண்டு மொழிகளுக்கு இடையிலான சொற்களின் மொழிபெயர்ப்புகளைக் கொண்டிருக்கும். ஒருமொழி அகராதி முக்கியமாக

நூல்களைப் படிப்பதற்கும் புரிந்து கொள்வதற்கும் சொந்த பேசுபவரால் பயன்படுத்தப்படுகிறது. மூல மொழியில் உள்ள சொற்களைப் புரிந்துகொள்ள அல்லது மொழிபெயர்க்க இருமொழி அகராதி பயன்படுத்தப்படுகிறது. இருமொழி அகராதி ஒரு திசை அகராதியாகவோ அல்லது இருதரப்பு/இருதிசை அகராதியாகவோ இருக்கலாம். ஒரு திசை அகராதி மூல மொழியிலிருந்து இலக்கு மொழிக்கான மொழிபெயர்ப்புகளைக் கொண்டிருக்கும்; ஆனால் தலைகீழ் மொழிபெயர்ப்புகள் வழங்கப்பட்டிருக்காது. இதற்கு மாறாக, இருதரப்பு/இருதிசை அகராதி மூல மொழியிலிருந்து இலக்கு மொழிக்கும், இலக்கு மொழியிலிருந்து மூல மொழிக்கும் மொழிபெயர்ப்புகளைக் கொண்டிருக்கும். ஒரு மொழியில் பொதுவாகப் பயன்படுத்தப்படும் எல்லாச் சொற்களையும் உள்ளடக்கிய வெளிப்படையான இருமொழி அகராதிகளைத் தவிர ஒரு அகராதி, குறிப்பிட்ட நோக்கம் அடிப்படையில் ஒருபொருள் பன்மொழிய அகராதியாகவோ (எ.கா. Merriam-Webster's Dictionary of Synonyms), இயற் பெயர்களை மையமாகக் கொண்ட அகராதியாகவோ (எ.கா. A Dictionary of Surnames) அல்லது ஒரு குறுகிய மற்றும் குறிப்பிட்ட பகுதியில் கவனம் செலுத்தும் அகராதியாகவோ (எ.கா. Black's Law Dictionary, and Stedman's Medical Dictionary) இருக்கலாம். புள்ளிவிவரங்கள் 1.1 ஒரு வியட்நாமிய-ஆங்கில காகித இருமொழி அகராதிக்கு ஒரு எடுத்துக்காட்டு.

சொற்களஞ்சியம் (thesaurus)

கில்கரிஃப் (Kilgarriff 2003) ஒரு சொற்களஞ்சியத்தைச் சொற்களை அவற்றின் பொருள் ஒற்றுமைக்கு ஏற்ப தொகுக்கிற ஒரு வளமாக வரையறுக்கிறார். ரோஜெட் (Roget 1911) ஒரு சொற்களஞ்சியத்தில் சொற்கள் அவை அகராதியில் இருப்பதைப் போல அகர வரிசைப்படி அல்லாமல் அவை வெளிப்படுத்தும் கருத்துக்களுக்கு ஏற்ப இருக்க வேண்டும் என்று கூறுகிறார். குறிப்பாக, சோர்கெலின் (Soergel (1974)) கூற்றுப்படி, ஒரு குழும சொற்களஞ்சியம் விளக்கிகள் (descriptors), ஒரு அட்டவணைப்படுத்தும் மொழி (indexing language), ஒரு வகைப்பாடு திட்டம் (classification scheme) அல்லது அமைப்புமுறை சொற்றொகையைக் system (vocabulary) கொண்டிருக்கும். ஒரு சொற்களஞ்சியம் விளக்கிகளுக்கு இடையியான உறவுகளையும் கொண்டிருக்கும். ஒவ்வொரு விளக்கிகளும்/டிஸ்கிரிப்டரும் ஒரு சொல், ஒரு குறிமானம் அல்லது கருத்தை குறிக்கப் பயன்படும் குறியீகளின் மற்றொரு கோர்வை ஆகும். ரோஜெட்டின் இன்டர்நேஷனல் தெசோரஸ் (Roget's International Thesaurus) (Roget 2008), ஓபன் தெசாரஸ்

(Open Thesaurus) (Bhattacharyya 2010) அல்லது thesaurus.com என அழைக்கப்படும் ஒரு பெரிய ஆன்லைன் ஆங்கிலச் சொற்களஞ்சியம்.

சொல்வலை

மில்லர் (Miller 1995) சொல்வலையை/வேர்ட்நெட்டை அறிமுகப்படுத்தினார்; இது ஒரு பெரிய சொல்சார் தரவுத்தளமாகும் (lexical database); இதில் பெயர்ச்சொற்கள், வினைச்சொற்கள், பெயரடைகள் மற்றும் வினையடைகள் என்பன ஒருபொருள் பன்மொழியத் தொகுப்புகள் (synsets) என அழைக்கப்படுகின்ற வரிசைப்படுத்தப்படாத புலனறிவுசார் ஒருபொருள்பன்மொழிகளாக குழுமப்பட்டுள்ளன. ஒவ்வொரு ஒருபொருள் பன்மொழியத் தொகுப்பும் (synset/சின்செட்) ஒரு தனித்துவமான கருத்தை வெளிப்படுத்துகிறது. வேர்ட்நெட் ஒரு செறிவூட்டப்பட்ட அகராதி மற்றும் சொற்களஞ்சியம் ஆகும். ஒரு சொல்சார் அலகு கொடுக்கப்பட்டால், பொது அகராதி மற்றும் சொல்வலை என்பன வரையறைகள், சொல்வகைப்பாடு, மற்றும் எடுத்துக்காட்டுகள் இவற்றைத் தரும். சொல்சார் அலகைப் பொறுத்தவரை, அகராதியில் முக்கியமாக ஒற்றை சொற்கள் உள்ளன; அதே நேரத்தில் வேர்ட்நெட் “tabular array”, “scholarly person” மற்றும் “grape vine” போன்ற குறுகிய சொற்றொடர்களைக் கொண்டிருக்கலாம். ஒரு கருத்தை வழங்கினால், சொல்வலை மற்றும் சொற்களஞ்சியம் கருத்துக்கு பொருந்தக்கூடிய சொற்களைத் தருகின்றன. சொல்வலை ஒருபொருள்பன்மொழியக் குழுமங்களில் (சின்செட்களில்/synsets) உள்ள சொற்கள் புலன்களின் அடிப்படையில் வேறுபடுகின்றன. சொல்வலையில் உள்ள சொற்களுக்கு இடையிலான உறவுகள் (உள்ளடக்குமொழிகள் அல்லது பொதுமைப்படுத்தல், உள்ளடங்குமொழிகள் அல்லது குறிப்பிடுதல், மற்றும் சினைமொழியம் அல்லது பகுதி-முழு உறவுகள் போன்றவை) புலக்குறிப்பு செய்யப்பட்டுள்ளன. தற்போது,மிகப்பெரிய சொல்வலை பிரின்ஸ்டன் சொல்வலை பதிப்பு 3.0 ஆகும்; இதில் 82,115 பெயர்ச்சொற்களின் ஒருபொருள்பனமொழியக் குழுமங்கள், 13,767 வினை ஒருபொருள்பனமொழியக் குழுமங்கள், 18,156 பெயரடை ஒருபொருள்பனமொழியக் குழுமங்கள் மற்றும் 3,621 வினையடை ஒருபொருள்பனமொழியக் குழுமங்கள் உட்பட 117,659 ஒருபொருள்பன்மொழியக் குழுமங்கள் உள்ளன. ஃபின் சொல்வலை, ஜப்பானிய சொல்வலை, யூரோ சொல்வலை ஆகியவை வேறு சில சொல்வலைகளும் உள்ளன. ஆசிய மொழிகளில் (அதாவது, பெங்காலி, இந்தி, இந்தோனேசிய, ஜப்பானிய, கொரிய, லாவோ, மங்கோலியன், பர்மிய, நேபாளி, சிங்கள, சுண்டனீஸ், தாய் மற்றும் வியட்நாமிய மொழிகளில் வேர்ட்நெட்களை

உருவாக்குவதற்கும் பகிர்வதற்கும் ஒரு தளத்தை ஏசியன்வார்ட்நெட் (AsianWordNet (AWN/ஏ.டபிள்யூ.என்)) வழங்குகிறது. துரதிர்ஷ்டவசமாக, ஆசியன் சொல்வலையில் வேர்ட்நெட்டுகளின் முன்னேற்றம் மிகவும் மெதுவாக உள்ளது, மேலும் அவை முடிவடையாமல் உள்ளன.

6.3.1. பன்மொழி நூலகங்களை உருவாக்குதல்

டிஜிட்டல் நூலகம் (digital library), டிஜிட்டல் களஞ்சியம் (digital repository) அல்லது டிஜிட்டல் சேகரிப்பு (digital collectionஎன்பது) டிஜிட்டல் பொருள்களின் ஆன்லைன் தரவுத்தளமாகும்; அவை உரை, நிலையான படங்கள், ஆடியோ, வீடியோ, டிஜிட்டல் ஆவணங்கள் அல்லது பிற டிஜிட்டல் மீடியா வடிவங்களை உள்ளடக்கியது. பொருள்கள் அச்சு அல்லது புகைப்படங்கள் போன்ற டிஜிட்டல் மயமாக்கப்பட்ட உள்ளடக்கத்தையும், முதலில் வேர்ட் செயலி கோப்புகள் அல்லது சமூக ஊடக இடுகைகள் போன்ற டிஜிட்டல் உள்ளடக்கத்தையும் கொண்டிருக்கலாம். உள்ளடக்கத்தை சேமிப்பதைத் தவிர, டிஜிட்டல் நூலகங்கள் சேகரிப்பில் உள்ள உள்ளடக்கத்தை ஒழுங்கமைக்கவும், தேடவும், மீட்டெடுக்கவும் வழிவகை செய்கின்றன.

டிஜிட்டல் நூலகங்கள் அளவு மற்றும் நோக்கத்தில் பெரிதும் மாறுபடும், மேலும் அவை தனிநபர்கள் அல்லது நிறுவனங்களால் பராமரிக்கப்படலாம். டிஜிட்டல் உள்ளடக்கம் உள்நாட்டில் சேமிக்கப்படலாம் அல்லது கணினி நெட்வொர்க்குகள் வழியாக தொலைவிலிருந்து அணுகப்படலாம். இந்த தகவல் மீட்டெடுப்பு அமைப்புகள் ஒருவருக்கொருவர் தகவல்களை இயங்கக்கூடிய தன்மை மற்றும் நிலைத்தன்மை மூலம் பரிமாறிக்கொள்ள முடியும்.

டிஜிட்டல் நூலகங்கள் இணையத்தில் கிடைக்கின்றன. இதன் பொருள் உலகம் முழுவதிலுமிருந்து பயனர்கள் டிஜிட்டல் நூலகங்களை அடைந்து அவற்றின் உள்ளடக்கங்களைப் பயன்படுத்த முயற்சி செய்யலாம். இணையத்தில் ஆங்கிலம் மிகவும் பிரபலமான மொழி என்பதில் எந்த சந்தேகமும் இல்லை, ஆனால் டிஜிட்டல் நூலகங்கள் மையமாக இருக்கும்போது மற்ற மொழிகளையும் கருத்தில் கொள்ள வேண்டும். டிஜிட்டல் நூலகங்கள் மற்றும் பொதுவாக டிஜிட்டல் சேகரிப்புகள் கலாச்சாரம், கல்வி, ஆராய்ச்சி முடிவுகள், அறிவியல், சட்டம் போன்றவற்றின் களஞ்சியங்கள் என்பதை நினைவில் கொள்ளுங்கள்.

டிஜிட்டல் நூலகங்கள் தொலைதூரக் கல்விக்கு ஆதரவாகப் பார்க்கும்போது, வெவ்வேறு மொழிகள் ஒவ்வொரு நாட்டின் கல்விச் செயல்பாட்டின் முக்கியக் கூறுகளாக இருப்பதால் அவற்றைக் கருத்தில் கொள்ள வேண்டும் என்பது தெளிவாகிறது.

ஒரு பாரம்பரிய நூலகத்தில், தலைப்புகள் மற்றும் தொகுதிகளை நாம் குறிப்பிடுகிறோம், ஒரு நூலக சேகரிப்பு ஒரு தலைப்பின் பல தொகுதிகளை வைத்திருக்கலாம், எடுத்துக்காட்டாக உரை புத்தகங்கள். டிஜிட்டல் நூலகத்தில் மூன்று காரணங்களுக்காக தலைப்புகள் (உள்ளடக்கங்கள்) மற்றும் தொகுதிகளை (நிகழ்வுகள், நிகழ்வுகள் அல்லது பொருள்கள்) பிரிப்பது அவசியம். முதலாவது, ஒரு உள்ளடக்கத்திற்கு வெவ்வேறு டிஜிட்டல் வடிவங்களில் நிகழ்வுகள் இருக்கலாம் என்பதுதான் உண்மை. ஒரு உதாரணம் ஹைபர்டெக்ஸ்ட் (HTML) மற்றும் நேரியல் உரையாக (பி.டி.எஃப்) பயன்படுத்தப்படக்கூடிய உரை; ஆன்லைன் காட்சிப்படுத்தலுக்கான முந்தையவை (அனிமேஷன்கள், தொடர்பு, முதலியன) மற்றும் பிந்தையது அச்சிடுதல். இரண்டாவது அலைவரிசை பண்புகள் காரணமாக ஒரு கோப்பை விட ஒரு உள்ளடக்கத்தை பிரிக்க வேண்டும். மூன்றாவது உரிமை நிர்வாகத்திலிருந்து வரலாம். டிஜிட்டல் நூலகம் பதிப்புரிமை செலுத்திய உடன் ஒத்த ஒரு தலைப்பின் 'பிரதிகள்' எண்ணிக்கையுடன் செயல்பட உரிமை உண்டு.

டிஜிட்டல் நூலகம் என்ற சொல் முதன்முதலில் National Science Foundation (NSF/என்எஸ்எஃப்), The Defense Advanced Research Projects Agency (DARPA/தர்பா), The National Aeronautics and Space Administration (NASA/நாசா) டிஜிட்டல் நூலகங்களின் முன்முயற்சியால் 1994இல் பிரபலப்படுத்தப்பட்டது. கணினி நெட்வொர்க்குகள் கிடைப்பதன் மூலம் தகவல் வளங்கள் தேவைக்கேற்ப விநியோகிக்கப்பட்டு அணுகப்படும் என்று எதிர்பார்க்கப்படுகிறது. மெய்நிகர் நூலகம் (virtual library) என்ற சொல் ஆரம்பத்தில் டிஜிட்டல் நூலகத்துடன் மாறி மாறி பயன்படுத்தப்பட்டது, ஆனால் இப்போது முதன்மையாக மற்ற புலன்களில் மெய்நிகர் இருக்கும் நூலகங்களுக்குப் பயன்படுத்தப்படுகிறது (விநியோகிக்கப்பட்ட உள்ளடக்கத்தை ஒருங்கிணைக்கும் நூலகங்கள் போன்றவை). டிஜிட்டல் நூலகங்களின் ஆரம்ப நாட்களில், டிஜிட்டல், மெய்நிகர் மற்றும் மின்னணு சொற்களில் உள்ள ஒற்றுமைகள் மற்றும் வேறுபாடுகள் பற்றி விவாதிக்கப்பட்டது.

பிறப்பு-டிஜிட்டல் (born-digital) என அழைக்கப்படும் டிஜிட்டல் வடிவத்தில் உருவாக்கப்பட்ட உள்ளடக்கம் மற்றும் இயற்பியல் ஊடகத்திலிருந்து மாற்றப்பட்ட தகவல்களுக்கு இடையே பெரும்பாலும் வேறுபாடு காணப்படுகிறது, எ.கா. காகிதம், டிஜிட்டல்மயமாக்கல் மூலம். எல்லா மின்னணு உள்ளடக்கங்களும் டிஜிட்டல் தரவு வடிவத்தில் இல்லை. கலப்பின நூலகம் (hybrid library) என்ற சொல் சில நேரங்களில் பௌதிகச் சேகரிப்புகள் (physical

collections) மற்றும் மின்னணு சேகரிப்புகள் (physical collections) கொண்ட நூலகங்களுக்குப் பயன்படுத்தப்படுகிறது. எடுத்துக்காட்டாக, அமெரிக்கன் மெமரி (American Memory) என்பது காங்கிரஸின் நூலகத்திற்குள் (Library of Congress) உள்ள ஒரு டிஜிட்டல் நூலகமாகும்.

சில முக்கியமான டிஜிட்டல் நூலகங்கள் arXiv மற்றும் இணைய காப்பகம் (Internet Archive) போன்ற நீண்ட கால காப்பகங்களாகவும் செயல்படுகின்றன. அமெரிக்காவின் டிஜிட்டல் பப்ளிக் லைப்ரரி (Digital Public Library of America) போன்ற பிற நூலகங்கள் பல்வேறு நிறுவனங்களிலிருந்து டிஜிட்டல் தகவல்களை ஆன்லைனில் பரவலாக அணுக முயற்சிக்கின்றன.

பன்மொழி டிஜிட்டல் நூலகம்

ஒரு பன்மொழி டிஜிட்டல் நூலகம் என்பது ஒரு டிஜிட்டல் நூலகமாகும், இது அனைத்து செயல்பாடுகளையும் ஒரே நேரத்தில் பல மொழிகளில் செயல்படுத்த வேண்டும் மற்றும் அதன் தேடல் மற்றும் மீட்டெடுக்கும் செயல்பாடுகள் மொழி சுதந்திரமானவை.

அத்தகைய கருத்தை செயல்படுத்துவது பயனருக்கு இடைமுகங்களுக்கான மொழியைத் தேர்வுசெய்யவும், நூலகப் பொருட்களின் மொழியை/மொழிகளைப் பொருட்படுத்தாமல் தேர்ந்தெடுக்கப்பட்ட மொழியில் உள்ள இடைமுகத்திலிருந்து எந்த மொழியிலும் (டிஜிட்டல் நூலகத்தின்) பட்டியலிடும் தகவல்களை அணுகவும் அனுமதிக்கிறது.

அடிப்படையாகக் கொண்ட பன்மொழி டிஜிட்டல் நூலகத்தின் அனைத்து முக்கிய பண்புகளின் வரையறைகளையும், அத்தகைய நூலகத்தின் செயல்பாடுகள், தரவுத்தளத்தின் தரவு மாதிரி மற்றும் மேக்ஸ்வெல் சிஸ்டத்தில் செயல்படுத்தப்படுவதையும் பின்வரும் பிரிவுகள் முன்வைக்கின்றன.

டிஜிட்டல் நூலகங்களின் வகைகள்

நிறுவன களஞ்சியங்கள் (Institutional repositories)

பல கல்வி நூலகங்கள் நிறுவனத்தின் புத்தகங்கள், ஆவணங்கள், ஆய்வறிக்கைகள் மற்றும் பிற படைப்புகளின் நிறுவன களஞ்சியங்களை உருவாக்குவதில் தீவிரமாக ஈடுபட்டுள்ளன, அவை டிஜிட்டல் மயமாக்கப்படலாம் அல்லது 'டிஜிட்டலாகப் பிறந்தவை' (born digital). வர்த்தக களஞ்சியங்களில் ஆராய்ச்சி வெளியிடுவதற்கு மாறாக, வெளியீட்டாளர்கள் பெரும்பாலும் அணுகல் உரிமைகளை மட்டுப்படுத்தும் வகையில், திறந்த அணுகலின் குறிக்கோள்களுக்கு இணங்க, இந்த களஞ்சியங்களில் பல பொது மக்களுக்கு சில கட்டுப்பாடுகளுடன் கிடைக்கின்றன. நிறுவன, உண்மையிலேயே இலவச மற்றும் கார்ப்பரேட் களஞ்சியங்கள் சில

நேரங்களில் டிஜிட்டல் நூலகங்கள் என்று குறிப்பிடப்படுகின்றன. நிறுவன களஞ்சிய மென்பொருள் ஒரு நூலகத்தின் உள்ளடக்கத்தை காப்பகப்படுத்துவதற்கும், ஒழுங்கமைப்பதற்கும், தேடுவதற்கும் வடிவமைக்கப்பட்டுள்ளது. பிரபலமான திறந்த-மூல தீர்வுகளில் டிஸ்பேஸ் DSpace,, ஈபிரிண்ட்ஸ் EPrints, டிஜிட்டல் காமன்ஸ் Digital Commons மற்றும் ஃபெடோரா காமன்ஸ் அடிப்படையிலான அமைப்புகள் ஐலண்டோரா Fedora Commons-based systems Islandora மற்றும் சாம்வெரா Samvera ஆகியவை அடங்கும்.

தேசிய நூலகச் சேகரிப்புகள்

சட்ட வைப்பு பெரும்பாலும் பதிப்புரிமைச் சட்டத்தினாலும், சில சமயங்களில் சட்ட வைப்புக்கான குறிப்பிட்ட சட்டங்களினாலும் மூடப்பட்டிருக்கும்; மேலும் ஒரு நாட்டில் வெளியிடப்பட்ட அனைத்து பொருட்களின் ஒன்று அல்லது அதற்கு மேற்பட்ட பிரதிகள் ஒரு நிறுவனத்தில், பொதுவாக தேசிய நூலகத்தில் பாதுகாக்கச் சமர்ப்பிக்கப்பட வேண்டும். மின்னணு ஆவணங்கள் வந்ததிலிருந்து, ஆஸ்திரேலியாவில் பதிப்புரிமைச் சட்டம் 1968 இல் 2016 திருத்தம் போன்ற புதிய வடிவங்களை உள்ளடக்கும் வகையில் சட்டத்தில் திருத்தம் செய்ய வேண்டியிருந்தது.

அதன் பின்னர் பல்வேறு வகையான மின்னணு வைப்புத்தொகைகள் கட்டப்பட்டுள்ளன. பிரிட்டிஷ் நூலகத்தின் வெளியீட்டாளர் சமர்ப்பிப்பு போர்டல் மற்றும் டாய்ச் நேஷனல் பிப்ளியோதெக்கில் உள்ள ஜெர்மன் மாதிரி ஆகியவை நூலகங்களின் வலைப்பின்னலுக்கு ஒரு வைப்பு புள்ளியைக் கொண்டுள்ளன; ஆனால் பொது அணுகல் நூலகங்களில் உள்ள வாசிப்பு அறைகளில் மட்டுமே கிடைக்கிறது. ஆஸ்திரேலிய தேசிய எடிபோசிட் அமைப்பு அதே அம்சங்களைக் கொண்டுள்ளது; ஆனால் பெரும்பாலான உள்ளடக்கங்களுக்கு பொது மக்களால் தொலைநிலை அணுகலை அனுமதிக்கிறது.

டிஜிட்டல் காப்பகங்கள் (Digital archives)

இயற்பியல் காப்பகங்கள் (Physical archives) இயற்பியல் நூலகங்களிலிருந்து பல வழிகளில் வேறுபடுகின்றன. பாரம்பரியமாக, காப்பகங்கள் பின்வருமாறு வரையறுக்கப்படுகின்றன:

- ஒரு நூலகத்தில் காணப்படும் இரண்டாம்நிலை ஆதாரங்களை விட (புத்தகங்கள், காலக்கோடுகள் போன்றவை) முதன்மை தகவல் ஆதாரங்களை (பொதுவாக ஒரு தனிநபர்

அல்லது நிறுவனத்தால் நேரடியாக தயாரிக்கப்படும் கடிதங்கள் மற்றும் ஆவணங்கள்) கொண்டிருத்தல்.

- அவற்றின் உள்ளடக்கங்களை தனிப்பட்ட ஐடங்களைக் காட்டிலும் குழுக்களாக ஒழுங்கமைத்தல்.
- தனிப்பட்ட உள்ளடக்கங்களைக் கொண்டிருத்தல்.

டிஜிட்டல் நூலகங்களை உருவாக்கப் பயன்படுத்தப்படும் தொழில்நுட்பம் காப்பகங்களுக்கு இன்னும் புரட்சிகரமானது, ஏனெனில் இது இந்த பொது விதிகளில் இரண்டாவது மற்றும் மூன்றை உடைக்கிறது. வேறு வார்த்தைகளில் கூறுவதானால், "டிஜிட்டல் காப்பகங்கள்" அல்லது "ஆன்லைன் காப்பகங்கள்" இன்னும் பொதுவாக முதன்மை ஆதாரங்களைக் கொண்டிருக்கும், ஆனால் அவை குழுக்கள் அல்லது சேகரிப்புகளில் (அல்லது கூடுதலாக) இருப்பதை விட தனித்தனியாக விவரிக்கப்படலாம். மேலும், அவை டிஜிட்டல் என்பதால், அவற்றின் உள்ளடக்கங்கள் எளிதில் மீட்டுருவாக்கம் செய்யக்கூடியவை, உண்மையில் அவை வேறு இடங்களிலிருந்து மீட்டுருவாக்கம் செய்யப்பட்டிருக்கலாம். ஆக்ஸ்போர்டு உரை காப்பகம் பொதுவாக கல்விசார் முதன்மை முதன்மை மூலப்பொருட்களின் பழமையான டிஜிட்டல் காப்பகமாக கருதப்படுகிறது.

காப்பகங்கள் நூலகங்களிலிருந்து வேறுபடுகின்றன. நூலகங்கள் தனித்தனியாக வெளியிடப்பட்ட புத்தகங்கள் மற்றும் சீரியல்கள் அல்லது தனிப்பட்ட பொருட்களின் வரம்புக்குட்பட்ட தொகுப்புகளை சேகரிக்கின்றன. நூலகங்கள் வைத்திருக்கும் புத்தகங்கள் மற்றும் பத்திரிகைகள் தனித்துவமானவை அல்ல; ஏனெனில் பல பிரதிகள் உள்ளன மற்றும் கொடுக்கப்பட்ட எந்த நகலும் பொதுவாக வேறு எந்த நகலையும் போலவே திருப்திகரமாக இருக்கும். காப்பகங்கள் மற்றும் கையெழுத்துப் பிரதி நூலகங்களில் உள்ள பொருள் "கார்ப்பரேட் அமைப்புகளின் தனித்துவமான பதிவுகள் மற்றும் தனிநபர்கள் மற்றும் குடும்பங்களின் ஆவணங்கள்" ஆகும்.

காப்பகங்களின் அடிப்படை பண்பு என்னவென்றால், அவற்றின் தகவல் உள்ளடக்கத்தைப் பாதுகாக்கவும், காலப்போக்கில் புரிந்துகொள்ளக்கூடிய மற்றும் பயனுள்ள தகவல்களை வழங்கவும் அவர்கள் பதிவுகள் உருவாக்கப்பட்ட சூழலையும் அவற்றுக்கிடையேயான உறவுகளின் வலையமைப்பையும் வைத்திருக்க வேண்டும். காப்பகங்களின் அடிப்படை சிறப்பியல்பு காப்பக பிணைப்பின் மூலம் சூழலை வெளிப்படுத்தும் அவர்களின் படிநிலை

அமைப்பில் உள்ளது. காப்பக விளக்கங்கள் காப்பகப் பொருளை விவரிக்க, புரிந்துகொள்ள, மீட்டெடுக்க மற்றும் அணுகுவதற்கான அடிப்படை வழிமுறையாகும். டிஜிட்டல் மட்டத்தில், காப்பக விளக்கங்கள் பொதுவாக குறியாக்கப்பட்ட காப்பக விளக்கம் (Encoded Archival Description (EAD)) எக்ஸ்எம்எல் வடிவமைப்பின் (XML format) மூலம் குறியாக்கம் செய்யப்படுகின்றன. குறியாக்கப்பட்ட காப்பக விளக்கம் (EAD) என்பது காப்பக விளக்கத்தின் தரப்படுத்தப்பட்ட மின்னணு பிரதிநிதித்துவமாகும், இது உலகம் முழுவதும் விநியோகிக்கப்படும் களஞ்சியங்களில் விரிவான காப்பக விளக்கங்கள் மற்றும் வளங்களுக்குத் கூட்டு அணுகலை வழங்குவதைச் சாத்தியமாக்குகிறது.

இந்திய அரசின் டிஜிட்டல் ஆக்கத் திட்டம்

மத்திய மனிதவள மேம்பாட்டு அமைச்சகம் லக்னோவில் ஒரு பன்மொழி அறிவு போர்டல்-பாரத்வானி-ஐ அறிமுகப்படுத்தியது. இதன் குறித்த முக்கியமான குறிப்புகள் கீழே பட்டியலிடப்பட்டுள்ளன.

- பாரத்வானி என்பது இந்தியாவில் அகராதிகளின் மிகப்பெரிய ஆன்லைன் சேமிப்பிடமாகும்.
- இந்தியாவின் பல்வேறு மொழிகளையும் கலாச்சாரங்களையும் ஒரே மேடையில் கொண்டு வருவதை போர்டல் நோக்கமாகக் கொண்டுள்ளது
- போர்ட்டலில் 22 திட்டமிடப்பட்ட மொழிகள் உள்ளன. எண்ணிக்கை 100 இந்திய மொழிகளாக அதிகரிக்கும்
- கல்வியை ஆன்லைனில் உலகமயமாக்குவதன் மூலம் ஊடாடும் அறிவுச் சமூகத்தை உருவாக்குவதை அமைச்சகம் நோக்கமாகக் கொண்டுள்ளது
- பரத்வானி போர்டல் பல மொழிகளில் செயல்படும் இந்தியாவில் உள்ள ஒரே அறிவு போர்டல் ஆகும்.
- பாரத்வானி போர்ட்டல் மூலம் இந்திய மொழிகளின் பன்முகத்தன்மையை மேம்படுத்துவதை அரசாங்கம் நோக்கமாகக் கொண்டுள்ளது.
- போர்ட்டலை குறுக்கு மொழி கற்றல் கருவியாகப் பயன்படுத்தலாம். இது பன்மொழி இலக்கண புத்தகங்கள் மற்றும் படிப்புகளைக் கொண்டுள்ளது, மேலும் ஒலிபெயர்ப்பையும் அனுமதிக்கிறது.

- மைசூரிலுள்ள இந்திய மொழிகளின் மத்திய நிறுவனம் (Central Institute of Indian Languages (CIIL/ சிஐஐஎல்) இந்த திட்டத்தை செயல்படுத்தும் பொறுப்பை ஏற்றுள்ளது.
- போர்ட்டலில் பல கோஷாக்கள் அல்லது பிரிவுகள் உள்ளன, அவை பல்வேறு அதிகாரிகளால் பாடப்புத்தகங்களுக்கான பா தியபுஸ்தகா கோஷ், அனைத்து மொழிகளிலும் கலைக்களஞ்சிய அறிவுத் தளத்திற்கான ஞான கோஷ், அகராதிகள், பொருள்விளக்கச் சொற்கோவைகள், கலைசொல் அகராதிகள் ஆகியவற்றைக் கொண்ட சப்தகோஷ் கோஷ், மொழி கற்றல் புத்தகங்கள் கொண்ட பாஷா கோஷ் மற்றும் மல்டிமீடியா உள்ளடக்கத்திற்கான பஹுமாத்யாம கோஷா
- தளத்தில் ஐ.டி கருவிகளுக்கான சுச்சனா பிரௌத்யோகிகி கோஷாவும் உள்ளது; ஆனால் அது இப்போது இந்திய மொழிகள் திட்டத்திற்கான தொழில்நுட்ப மேம்பாட்டுடன் (டி.டி.எல்.பி) இணைக்கப்பட்டுள்ளது.
- 130 தளத்தில் 130க்கும் மேற்பட்ட அகராதிகள், சொற்களஞ்சியம் மற்றும் சொல் புத்தகங்கள் உள்ளன.
- போர்ட்டலில் உள்ள உள்ளடக்கங்கள் இந்திய பதிப்புரிமைச் சட்டத்தால் பாதுகாக்கப்படுகின்றன.
- உள்ளடக்கங்கள் உரை மற்றும் PDF இல் கிடைக்கின்றன
- பரத்வாணி போர்டல் கூகிள் பிளே ஸ்டோரில் மொபைல் பயன்பாடாகவும் கிடைக்கிறது.

6.3.2. மொழி கற்பவர்களுக்குப் பாடநூல்களைத் திட்டமிடல்

ஒரு மொழி பாடநூல்களை வடிவமைப்பது என்பது பல பரிமாண பணியாகும். இது மொழிப் பாடத்திட்டத்தின் குறிக்கோள்களைப் பற்றிய நுணுக்கமான புரிதல், மொழி கற்றல் கோட்பாடுகள் மற்றும் அவற்றின் கற்பித்தல் தாக்கங்கள் பற்றிய நல்ல அறிவு, தேர்வு வாரியங்களின் தேவை மற்றும் அரசாங்கத்தின் மொழி கொள்கைகள் பற்றிய புரிதல் தேவைப்படுகிறது. ஒரு மொழி பாடப்புத்தகத்தை முடிவு சார்ந்ததாக மாற்றுவதற்கு, மொழியில் பாடத்திட்டப் பொருள் வளர்ச்சியின் கோட்பாடு மற்றும் நடைமுறைக்கு இடையிலான தொடர்பை ஆசிரியர் புரிந்துகொண்டு பாராட்ட வேண்டும். ஒரு மொழிப் பாடநூல் என்பது வெவ்வேறு வகைகளிலிருந்து பெறப்பட்ட நூல்களின் தொகுப்பு மட்டுமல்ல; ஒரு குறிப்பிட்ட சூழலில் பின்பற்றப்பட்ட பாடத்திட்டத்தை தெரிவிக்கும் பரந்த கல்வி நோக்கங்களைக் கணக்கில் எடுத்துக்

கொண்டபின், பாடநூல் வடிவமைப்பாளர்களால் வந்த ஒரு குறிப்பிட்ட கல்விக் கண்ணோட்டத்தை இது பிரதிபலிக்கிறது.

மொழிப் பாடநூல் ஒரு தகவல்தொடர்பு செயல்பாட்டு அணுகுமுறையைப் பின்பற்றுகிறது. இது ஒரு கற்றவருக்குத் தகவல்தொடர்பு மற்றும் இலக்கு மொழியில் ஒரு செயல்பாட்டுத் திறனைப் பெறுவதற்கு உதவுகிறது. புதிய பாடப்புத்தகத்தில் பொருட்களின் தேர்வு மற்றும் செயல்பாடுகளை வழங்குவது செயல்பாட்டு தகவல்தொடர்பு அணுகுமுறையின் கொள்கைகளை அடிப்படையாகக் கொண்டது; இது மொழியை ஒரு கருவியாகக் கருதுகிறது; குறிப்பிட்ட சமூக சூழல்களில் தொடர்பு மற்றும் மொழி செயல்பாடுகளுக்கான வழிமுறையாகும்.

மொழிப் பாடபுத்தகம் சுற்றுச்சூழல், விளையாட்டு, சுகாதாரம், இளமை, பயணம் மற்றும் சுற்றுலா, கலாச்சாரம், உத்வேகம் மற்றும் அறிவியல் மற்றும் தொழில்நுட்பம் போன்ற கருப்பொருள்களைக் கொண்டிருக்கலாம். திறன்களின் ஒருங்கிணைப்பு அனைத்து அலகுகளின் தனித்துவமான அம்சமாகும்; மேலும் இலக்கு மொழியின் ஆக்கபூர்வமான பயன்பாட்டிற்கு கற்பவர்கள் போதுமான வெளிப்பாட்டைப் பெறவேண்டும்: கதை, கவிதை, சுயசரிதை, கட்டுரை, நினைவூட்டல், சிற்றேடு, உரையாடல், தொலைபேசி நேர்காணல், அறிவியல் புனைகதை, பேச்சு போன்றவை. வெவ்வேறு வகைகளிலிருந்து வரையப்பட்ட சுவாரஸ்யமான மற்றும் சவாலான உண்மையான பொருள்களை வெளிப்படுத்துவதன் மூலம் மொழியை எளிதில் புரிந்துகொள்வதற்கும் பயன்படுத்துவதற்கும் கற்பவர்களுக்கு வாய்ப்பளிக்கவேண்டும்.

6.3.3. ஒருமொழி அகராதிகளின் உருவக்கம் (அச்சிடப்பட்ட மற்றும் மின்னணு)

1. அகராதிகளில் உள்ள தகவல்கள் எங்கிருந்து வருகின்றன?

ஒரு அகராதி என்பது ஒரு மொழியின் சொல்லகராதி பற்றிய விளக்கமாகும். இது சொற்களின் அர்த்தத்தை விளக்குகிறது, மேலும் வாக்கியங்களை உருவாக்குவதற்கு அவை எவ்வாறு ஒன்றிணைகின்றன என்பதைக் காட்டுகிறது. ஆனால் அகராதியியலார்கள் அதாவது அகராதிகள் எழுதுபவர்கள் எங்கிருந்து தகவல்களைப் பெறுகிறார்கள்?

சொற்களைப் பற்றிய இரண்டு முக்கிய ஆதாரங்கள் உள்ளன: அகநோக்குப்பார்வை (introspection) மற்றும் உற்றுநோக்குதல் (observation).

- அகநோக்குப்பார்வை என்பது உங்கள் சொந்த மூளையை 'உள்ளே பார்ப்பது' மற்றும் ஒரு சொல்லைப் பற்றி உங்களுக்குத் தெரிந்த அனைத்தையும் நினைவில் வைக்க முயற்சிப்பது

- உற்றுநோக்குதல் என்பது பயன்பாட்டில் உள்ள மொழியின் உண்மையான எடுத்துக்காட்டுகளை ஆராய்வது (செய்தித்தாள்கள், நாவல்கள், வலைப்பதிவுகள், ட்வீட் மற்றும் பலவற்றில்), இதன் மூலம் மக்கள் ஒருவருக்கொருவர் தொடர்பு கொள்ளும்போது சொற்களை எவ்வாறு பயன்படுத்துகிறார்கள் என்பதைப் புரிந்துகொள்ள முடியும்.

ஒரு மொழியின் சரளமாகப் பேசுபவர் அந்த மொழியின் சொற்றொகுதி பற்றி ஏற்கனவே நிறைய அறிந்திருக்க வேண்டும் என்பது வெளிப்படையானது. எனவே அகநோக்குப்பார்வை என்பது சொற்களின் பொருள் மற்றும் அவை எவ்வாறு பயன்படுத்தப்படுகின்றன என்பது பற்றிய நுண்ணறிவுகளின் பயனுள்ள ஆதாரமாக இருக்கும். ஆனால் ஒரு அகராதி ஒரு சொல்லின் நடத்தை குறித்த முழுமையான மற்றும் சீரான கணக்கைக் கொடுக்க வேண்டும்; மேலும் அகநோக்குப்பார்வையால் மட்டுமே இந்த நோக்கத்திற்காகப் போதுமான தகவல்களை வழங்க முடியாது. இதன் விளைவாக, 18ஆம் நூற்றாண்டில் சாமுவேல் ஜான்சனின் காலத்திலிருந்தே அகராதியலார்கள் தங்கள் அகராதிகளை உற்றுநோக்குதலை அடிப்படையாகக் கொண்டுள்ளனர். ஜான்சனின் காலத்தில், மொழியைக் உற்றுநோக்குவது ஒரு உழைப்பு நிறைந்த வணிகமாகும்: இதன் பொருள் நூற்றுக்கணக்கான புத்தகங்களைப் படிப்பது மற்றும் பயன்பாட்டில் உள்ள சொற்களின் நல்ல எடுத்துக்காட்டுகளைப் பெறுவது. ஆனால் இன்றைய கணினி தொழில்நுட்பம் இதையெல்லாம் மிகவும் எளிதாக்குகிறது. இது மொழியின் சொற்றொகையின் நம்பகமான தகவலை இப்போது வழங்கக்கூடிய அளவுக்கு நல்ல மொழி தரவுகளுக்கான அணுகலை நமக்கு வழங்குகிறது.

2. மொழியைக் உற்றுநோக்குவதற்கான வழிகள்: 'மேற்கோள்கள்' மற்றும் தரவுத்தொகுதி

250 ஆண்டுகளுக்கும் மேலாக, மொழியியலாளர்கள் மேற்கோள்களை - பயன்பாட்டில் உள்ள சொற்களின் எடுத்துக்காட்டுகள், புத்தகங்கள் அல்லது பிற மூலங்களிலிருந்து எடுக்கப்பட்டவை - மொழியை விவரிப்பதற்கான அடிப்படையாகப் பயன்படுத்துகின்றனர். மாக்மில்லன் ஆங்கில ஒருமொழிய அகராதிகளின் Buzzword காப்பகத்திலிருந்து எடுக்கப்பட்ட எடுத்துக்காட்டு, ‘‘to green’,’ என்ற வினைச்சொல்லை விளக்கி, இரண்டு அமெரிக்க செய்தித்தாள்களின் மேற்கோள்களை உள்ளடக்கும்:

Green

VERB [TRANSITIVE]

to make something more environmentally-friendly

greening

NOUN [UNCOUTABLE]

The program has *greened* the zoo's primate exhibit and courtyard by installing rain barrels and cisterns to harvest water runoff form rooftops…

SYRACUSE NEW TIMES 20th APRIL 2011

'*of all the sustainability initiatives, day cleaning is "the biggest for the buck; it didn't cost anything," said Marion Coker, manger of strategic business planning and sustainability…The transportation agency said the chage, made on months ago, had contributed, along with other* ***greeing*** *measures to a 12 percent reduction in energy use.'*

COLUMBUS DISPATCH 10th APRIL 2011

மொழியில் ஏற்படும் மாற்றங்களைக் கண்காணிக்கவும், புதிய சொற்கள் மற்றும் சொற்றொடர்கள் பயன்பாட்டுக்கு வரும்போது அவற்றைக் கண்டறியவும் இந்த வகையான தரவு குறிப்பாக பயனுள்ளதாக இருக்கும். ஆதாரங்கள் இப்போது புத்தகங்கள் மற்றும் செய்தித்தாள்களை மட்டுமல்ல, இணையத்திலும் பயன்படுத்தப்படும் மொழியை உள்ளடக்கியதாக விரிவடைந்துள்ளன. எனவே மாக்மில்லனின் வலைப்பதிவு handbags 'கைப்பைகளை' ஒரு வினையெச்சமாகப் பயன்படுத்துவது பற்றி விவாதித்தபோது, பெரும்பாலான மேற்கோள்கள் 'பாரம்பரிய' ஊடகங்களிலிருந்து வந்தவை அல்ல, ஆனால் சமூக வலைப்பின்னல்களில் ட்வீட் மற்றும் பிற இடுகைகளிலிருந்து வந்தன.

மேற்கோள்கள் இன்னும் ஒரு பயனுள்ள பங்கைக் கொண்டுள்ளன, ஆனால் மாக்மில்லன் போன்ற அகராதிகளின் மொழித் தரவின் முக்கிய ஆதாரம் தரவுத்தொகுதி ஆகும். ஒரு தரவுத்தொகுதி என்பது கணினியில் சேமிக்கப்பட்டுள்ள ஆயிரக்கணக்கான வெவ்வேறு 'உரைகளின்' தொகுப்பாகும். இந்த உரைகளில் நாவல்கள், கல்வி புத்தகங்கள் மற்றும் ஆவணங்கள், செய்தித்தாள்கள், பத்திரிகைகள், பதிவு செய்யப்பட்ட உரையாடல்கள் மற்றும் ஒளிபரப்பு நேர்காணல்கள், வலைப்பதிவுகள், ஆன்லைன் பத்திரிகைகள் மற்றும் கலந்துரையாடல் குழுக்கள் மற்றும் பல உள்ளன. ஒரு தரவுத்தொகுதியைப் பயன்படுத்துவதன் முக்கிய அம்சம் என்னவென்றால், உலகெங்கிலும் உள்ள மில்லியன் கணக்கான (அல்லது பில்லியன் கணக்கான) மக்கள் பயன்படுத்தும் அனைத்து ஆங்கிலத்தையும் அதை தொகுப்பவர்களால் கவனிக்க முடியாது; எனவே அதற்குப் பதிலாக ஆங்கில நூல்களின் பிரதிநிதி மாதிரியைப் பார்க்கின்றனர்.

புத்திசாலித்தனமான மென்பொருளைப் பயன்படுத்தி (கீழே காண்க) ஒரு குறிப்பிட்ட சொல், சொற்றொடர், இலக்கண முறை அல்லது சொல்லடி வகைப்பாடு ஆகியவற்றின் ஒவ்வொரு உதாரணத்தையும் நாம் காணலாம். இந்த தகவல் தான் அகராதியில் உள்ள சொற்களைப் பற்றி அகராதி தொகுப்பவர் சொல்லும் அனைத்திற்கும் அடிப்படையாக அமைகிறது.

3. மேக்மில்லனின் தரவுத்தொகுதி வளங்கள்

மேக்மில்லனில் இரண்டு வகையான தரவுத்தொகுதிகள் உள்ளன: பொது மற்றும் சிறப்பு. பொது தரவுத்தொகுதியில் கல்விசார் புத்தகங்கள் மற்றும் பத்திரிகைகள், பிரபலமான மற்றும் இலக்கிய நாவல்கள், தேசிய மற்றும் உள்ளூர் செய்தித்தாள்கள் வரை பலவிதமான தகவல் மற்றும் கற்பனை நூல்கள் உள்ளன. இது இப்போது எழுதப்பட்ட மற்றும் பேசும் ஆங்கிலத்தின் கிட்டத்தட்ட 1.6 பில்லியன் சொற்களைக் கொண்டுள்ளது - அதாவது பத்து ஆண்டுகளுக்கு முன்பு மேக்மில்லன் ஆங்கில அகராதியின் முதல் பதிப்பை உருவாக்கியபோது அவர்கள் பயன்படுத்திய தரவுத்தொகுதியை விட இது எட்டு மடங்கு பெரியது. இதுதான் அவர்கள் பெரும்பாலும் பயன்படுத்தும் தரவுத்தொகுதி. அவர்கள் பயன்படுத்தும் சிறப்புத் தரவுத்தொகுதி பின்வருவனவற்றை உட்படுத்தும்:

- மேக்மில்லன் பாடத்திட்டத் தரவுத்தொக்தி: வேளாண்மை முதல் விலங்கியல் வரையிலான பள்ளி பாடங்களை உள்ளடக்கிய நூற்றுக்கணக்கான பள்ளிப் பாடப்புத்தகங்கள் மற்றும் தேர்வு பாடத்திட்டங்களால் ஆன 20 மில்லியன் சொல் தரவுத்தளம். மேக்மில்லன் பள்ளி அகராதி (Macmillan School Dictionary) மற்றும் மேக்மில்லன் ஆய்வு அகராதி (Macmillan Study Dictionary) தயாரிக்கும் போது இது முதலில் பயன்படுத்தப்பட்டது.
- சுற்றுச்சூழல் அறிவியலின் 60 மில்லியன் சொல் தரவுத்தொகுதி – முதலாவதாக குறிப்பிட்ட களங்களுக்கான திட்டமிடப்பட்ட புதிய தரவுத்தொகுதி
- பெல்ஜியத்தில் உள்ள யுனிவர்சிட்டி கேத்தோலிக் டி லூவின்-லா-நியூவ் (Université catholique de Louvain-la-Neuve) என்ற இடத்தில் ஆங்கில தரவுத்தொகுதி மொழியியல் மையம் (English Corpus Linguistics (CECL/சி.இ.சி.எல்) உருவாக்கிய கற்றல் தரவுத்தொகுதி. சி.இ.சி.எல் உடனான மேக்மில்லனின் ஒத்துழைப்பு கீழே விவரிக்கப்பட்டுள்ளது.

4. ஒரு தரவுத்தொகுதியிலிருந்து தகவல்களை எவ்வாறு பெறுவது?

மொழி தரவுத்தொகுதியிலிருந்து தகவல்களைப் பெறுவதற்கு அகராதியலார்கள் சக்திவாய்ந்த கணினி நிரல்களைப் பயன்படுத்துகின்றனர். ஒரு தரவுத்தொகுதி பகுப்பாய்வு செய்வதற்கான மிகச் சிறந்த மென்பொருள் வகை 'கான்கார்டேன்சர்'/'தொடரடைவி' (concordancer) என்று அழைக்கப்படுகிறது - ஏனெனில் இது தொடரடைவுகளை (concordances) உருவாக்குகிறது, பின்வருவது *remember* என்ற வார்த்தையின் தொடரடைவு:

I don't remember falling asleep. I don't	**remember**	seeing Santa come. Feeling something heavy
apologies. Due to such embarrassment, we're now	**remembered**	well. Exchanging nods of recognition, as
Oh, I hope it doesn't get lost! Does he	**remember**	his way here? Does he know my garden gate
pockets for cold, wet days. This is worth	**remembering**	if you suffer from cold hands. But this
updated URL you are seeking. Finally please	**remember**	that URLs are case sensitive. Ensure you
on his guitar. Tears spilled over as she	**remembered**	it with affection. Across the street was
and a sister, four and six. They barely	**remembered**	Mum, not like me. So when she came in the
turkey, roasted in the oven. Louise could not	**remember**	the last time she had had a hot meal. Steven
the theme of the book at the LSE? I don't	**remember**	wearing a suit at all. But if I did, it
and may lose their interest in reviewing.	**Remember**	to review success as well! See Reviewing
good start and bodes well for her future.	**Remember**	, easy to criticise, hard to create. More
shirt and where was his shoes? He vaguely	**remembered**	a feeling of total happiness and yet now
memories. However, even though they couldn't	**remember**	a certain event they still had feelings
seemed to have stopped at midnight. Then he	**remembered**	long soft red hair touching his face, but
cancer, but no one should take this chance.	**Remember**	most testicular cancers are curable if
late Nineties. stole something? I don't	**remember**	, and that answer isn't just to avoid incriminating
porters. Welfare Life isn't always easy.	**Remember**	that someone else has almost certainly
like to see an answer to this one too... I	**remember**	querying it a few years ago in relation
and masks to protect your face and neck.	**Remember**	, even if it is not hot, you can get severely

ஒரு தொடரடைவி (concordancer) முழு தரவுத்தொகுதியும் பார்த்து, ஒரு குறிப்பிட்ட சொல் அல்லது சொற்றொடரின் ஒவ்வொரு எடுத்துக்காட்டையும் கண்டுபிடித்து, அதன் உடனடி சூழலுடன் - அதன் இருபுறமும் ஏழு அல்லது எட்டு சொற்கள்- காட்டுகிறது. மேலே உள்ள படம் மாக்மில்லன் தரவுத்தொகுதியில் உள்ள அனைத்து வாக்கியங்களின் சிறிய மாதிரியையும் காட்டுகிறது. நாம் அடையாளம் காண மிக முக்கியமான விஷயம் தொடர்ச்சியான வடிவங்கள்: வேறுவிதமாகக் கூறினால், எந்தவொரு அம்சமும் ஒரு முறை மட்டுமல்ல, பல முறையும் நிகழ்கிறது. எடுத்துக்காட்டாக, இந்த தொடரடைவின் முதல் வரி *I don't remember seeing Santa come* என்று கூறுகிறது.

இது இலக்கண முறைக்கு ஒரு எடுத்துக்காட்டு; இங்கு *remember* ஒரு வினைச்சொல்லுடன் *–ing* வடிவத்தில் (அல்லது வினைப்பெயர்/ஜெரண்ட்) பயன்படுத்தப்படுகிறது. மீதமுள்ள ஒத்திசைவை நீங்கள் கவனமாகப் பார்த்தால், அதே கட்டுமானத்தின் மேலும் இரண்டு எடுத்துக்காட்டுகளைக் காணலாம்:

I don't remember wearing a suit at all.

I remember querying it a few years ago.

மேலே உள்ள ஒத்திசைவைப் படிப்பது, *remember* பயன்படுத்தப்படும் வழியின் பொதுவான பல வடிவங்களை அடையாளம் காண்பது எளிது. இவை பின்வருமாறு:

- வினைச்சொல்லைத் தொடர்ந்து *that-clause* வரும் போது: *Finally, please remember that URLs are case sensitive*
- '*…is worth remembering'* என்ற வெளிப்பாடு: *This is worth remembering if you suffer from cold hands*
- வினைச்சொல் வினையெச்சத்துடன் பயன்படுத்தப்படும்போது: *Remember to review success as well!*
- *remember* உடன் பயன்படுத்தப்படும் பொதுவான வினையடைகள்: *He vaguely remembered a feeling of total happiness and yet now it was gone.| They barely remembered Mum, not like me.*

இது போன்ற நூற்றுக்கணக்கான (சில நேரங்களில் ஆயிரக்கணக்கான) எடுத்துக்காட்டுகளை ஸ்கேன் செய்வதன் மூலம், remember போன்ற ஒரு வார்த்தையைப் பற்றிய மிக முக்கியமான உண்மைகளின் படத்தை படிப்படியாக உருவாக்குகிறோம்.

இருப்பினும், இது மிகவும் நேரத்தை எடுத்துக்கொள்ளும். அகராதியலார்கள் முதன்முதலில் தரவுத்தொகுதித் தரவைப் பயன்படுத்தத் தொடங்கியபோது, 1980களில், தரவுத்தொகுதிகள் வெறும் 10 அல்லது 20 மில்லியன் சொற்களைக் கொண்டதாக ஒப்பீட்டளவில் சிறியதாக இருந்தன. இதன் விளைவாக, ஒரு குறிப்பிட்ட சொல்லின் எடுத்துக்காட்டுகளின் எண்ணிக்கையும் (நினைவில் கொள்வது போன்றவை) மிகவும் சிறியதாக இருக்கும் - எனவே அவை அனைத்தையும் பார்க்க முடிந்தது. ஆனால் இன்றைய பில்லியன் சொற்களைக் கொண்ட தரவுத்தொகுதியுடன், இது இனி உண்மை அல்ல. மேக்மில்லனில் பயன்படுத்தப்படும்

தரவுத்தொகுதியில் *remember* என்ற வினைச்சொல்லின் 232,394 எடுத்துக்காட்டுகள் உள்ளன, மேலும் அவை ஒவ்வொன்றையும் படிக்க இயலாது.

தொடரடைவுக்கு அப்பால் (Beyond the concordance)

அதிர்ஷ்டவசமாக, புத்திசாலித்தனமான புதிய மென்பொருள் 'தகவல் மிகைச்சுமை' ('information overload') சிக்கலை தீர்க்கிறது. தொடரடைவுகளுக்கு மேலதிகமாக, இப்போது ஒரு சொல்லைப் பற்றிய அனைத்து முக்கிய உண்மைகளின் திறமையான ஒரு பக்கச் சுருக்கத்தை வழங்கும் 'வேர்ட் ஸ்கெட்சுகள்' ('Word Sketches) என்பதைப் பார்க்கிறோம். *evidence* என்ற பெயர்ச்சொல்லிற்கான ஒரு வேர்ட் ஸ்கெட்சின் ஒரு பகுதி கீழே தரப்பட்டுள்ளது; இந்த மற்றொரு பொதுவான சொல் மாக்மில்லன் தரவுத்தொகுதியில் சுமார் 300,000 வெவ்வேறு எடுத்துக்காட்டுகளைக் கொண்டுள்ளது.

object_of	135665	3.6	subject_of	42624	2.0	adj_subject_of	5993	1.9	a_modifier	93047	2.5	n_modifier	14601	0.6
be	44111	5.35	be	14597	3.76	available	1600	5.62	little	3501	7.94	research	1799	5.94
provide	10909	7.79	suggest	4695	9.46	such	270	3.03	scientific	3121	8.98	expert	1198	7.7
give	9903	7.82	have	2679	3.55	clear	245	4.63	further	3062	7.58	documentary	938	9.42
find	4911	7.17	show	2587	7.06	consistent	158	6.33	clear	2898	7.85	firm	263	5.11
have	4600	4.31	support	1872	6.84	relevant	144	4.39	new	2531	5.35	facie	258	8.91
base	3412	7.92	indicate	1062	7.69	sufficient	138	5.75	good	2368	5.41	material	256	3.53
present	2408	7.78	do	982	4.05	strong	107	3.62	other	2291	5.04	hearsay	244	8.95
see	2297	5.83	relate	664	6.6	overwhelming	99	7.14	oral	2121	9.05	survey	170	4.37
show	2259	6.67	exist	660	6.82	necessary	95	3.72	strong	2048	7.49	V	169	6.41
produce	2122	6.9	come	626	4.87	conclusive	92	8.24	available	1939	5.78	show	166	3.96
gather	1711	8.18	emerge	424	7.03	likely	91	3.32	anecdotal	1773	9.22	quality	165	3.1
support	1582	6.36	concern	391	5.97	admissible	76	8.33	documentary	1684	9.12	DNA	152	6.86

இது எப்படி வேலை செய்கிறது? நிரல் முதலில் விசாரிக்கப்படும் வார்த்தையின் அனைத்து எடுத்துக்காட்டுகளையும் சேகரிக்கிறது - ஒரு தொடரடைவி செய்வது போல. பின்னர் இது இரண்டாம் கட்ட பகுப்பாய்வுக்குப் பொருந்தும். இந்த நேரத்தில், மென்பொருள் குறிப்பிட்ட இலக்கண உறவுகளைப் பார்க்கிறது. *evidence* விஷயத்தில், *evidence* என்பது ஒரு வினைச்சொல்லின் செயப்படுபொருளாக இருக்கும் அனைத்து வாக்கியங்களையும் கண்டுபிடிக்கும்; பின்னர் இந்த அமைப்பொழுங்கில் அடிக்கடி பயன்படுத்தப்படும் வினைச்சொற்களை அடையாளம் காணும். மேலே உள்ள வேர்ட் ஸ்கெட்சின் முதல் நெடுவரிசையில் பட்டியலிடப்பட்ட வினைச்சொற்கள் இவை: மக்கள் *giving evidence, finding evidence, presenting evidence, or gathering evidence* பற்றி பேசுகிறார்கள் (அல்லது

எழுதுகிறார்கள்). இதேபோல், 'a_modifier' என்ற நெடுவரிசை இந்த பெயர்ச்சொல்லை அடிக்கடி மாற்றியமைக்கும் பெயரடைகளின் பட்டியலாகும்: நாம் there is *little* evidence for something, or talk about *clear* evidence, *strong* evidence, or *scientific* evidence என்று கூறலாம். ஒவ்வொரு சொல்லின் அடுத்து வரும் நீல எண் (blue number) தரவுத்தொகுதியில் ஒவ்வொரு சேர்க்கையும் எவ்வளவு அடிக்கடி தோன்றும் என்பதைக் கூறுகிறது: எனவே 'provide + evidence' என்ற சேர்க்கை 10,909 முறை நிகழ்கிறது. இந்த எண்ணைக் கிளிக் செய்வதன் மூலம் *evidence* என்பது *provide* என்பதன் செயப்படுபொருளாய் வரும் அனைத்து வாக்கியங்களையும் காட்டும் ஒரு ஒத்திசைவு கிடைக்கும்.

இந்த மென்பொருள் அகராதியியலார்களின் வாழ்க்கையை எளிதாக்கியுள்ளது, அதே நேரத்தில் மிகவும் துல்லியமான மற்றும் விரிவான தகவல்களை வழங்குகிறது. இது போன்ற நிகழ்ச்சிகள் இப்போது அகராதிக்கான நிலையான கருவிகளாக இருக்கின்றன, ஆனால் வேர்ட் ஸ்கெட்ச் மென்பொருளானது மேக்மில்லனால் முன்னோடியாக இருந்தது மற்றும் மேக்மில்லன் ஆங்கில அகராதியின் முதல் பதிப்பைத் தயாரிக்கப் பயன்படுத்தப்பட்டது.

6. தரவுத்தொகுதி எந்த வகையான தகவல்களை வழங்குகிறது?

சொற்கள் எதைக் குறிக்கின்றன என்பதை அகராதிகள் உங்களுக்குச் சொல்லவில்லை, சொற்கள் எவ்வாறு பயன்படுத்தப்படுகின்றன என்பதையும் விளக்குகின்றன. இந்த இரண்டு செயல்பாடுகளையும் நிறைவேற்றுவதற்கான ஆதாரங்களை தரவுத்தொகுதி நமக்கு வழங்குகிறது.

பொருள்

பல சொற்களுக்கு ஒன்றுக்கு மேற்பட்ட அர்த்தங்கள் உள்ளன, ஆனால் பேச்சாளர் அல்லது எழுத்தாளர் எந்த அர்த்தத்தை விரும்புகிறார் என்பது எப்போதும் தெளிவாகிறது. தரவுத்தொகுதியிலிருந்து இந்த நான்கு வாக்கியங்களில், goal என்ற சொல் அதன் 'கால்பந்து கெலிப்பெண்' அர்த்தத்தில் எப்போது பயன்படுத்தப்படுகிறது, அல்லது அது 'நோக்கம்' அல்லது 'குறிக்கோள்' எப்போது பயன்படுத்தப்படுகிறது எனக் காண்பது எளிது:

But the referee spotted the foul, and disallowed the goal.

African leaders are seeking the support of the international community to achieve these goals.

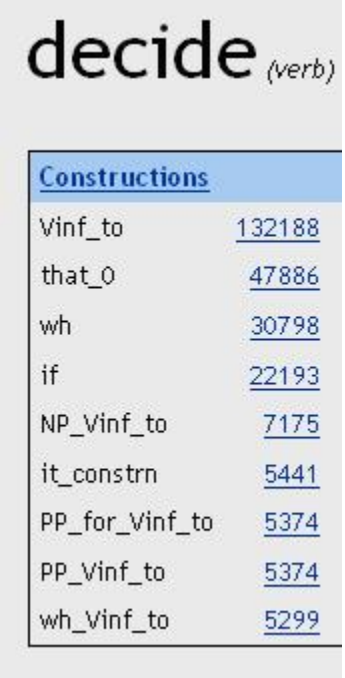
He has made 137 appearances for United and scored 27 goals.

Teachers may use this information to help students set goals for themselves.

உண்மையான உரையாடல்களைப் போலவே, சொல் தோன்றும் சூழலின் மூலம் 'சரியான' பொருளை அடையாளம் காண்கிறோம். சூழலில் சொற்களைப் படிப்பதன் மூலம், அவற்றில் எத்தனை வெவ்வேறு அர்த்தங்கள் உள்ளன என்பதைக் கண்டுபிடிப்போம்.

இலக்கணம்

remember என்பதற்கான தொடரடைவு எவ்வாறு வினைக்கான இலக்கண ஒழுங்மைப்புகளைப் பற்றி நமக்குச் சொல்கிறது என்பதைப் பார்த்தோம்: ஒரு வினைப்பெயர், that-எச்சத்தொடர், வினையெச்சம் மற்றும் பல. இங்கே மீண்டும், வேர்ட் ஸ்கெட்சுகள் மிகவும் அடிக்கடி வரும் 'கட்டுமானங்களை' பட்டியலிடுவதன் மூலம் ஒரு பயனுள்ள குறுக்குவழியை வழங்குகின்றன - எனவே நாம் இனி நூற்றுக்கணக்கான எடுத்துக்காட்டுகளை ஸ்கேன் செய்யத் தேவையில்லை. *decide* என்ற வினைச்சொல்லின் ஒரு சொல் ஸ்கெட்சில் உள்ள இலக்கண வடிவங்களின் பட்டியல் இங்கே:



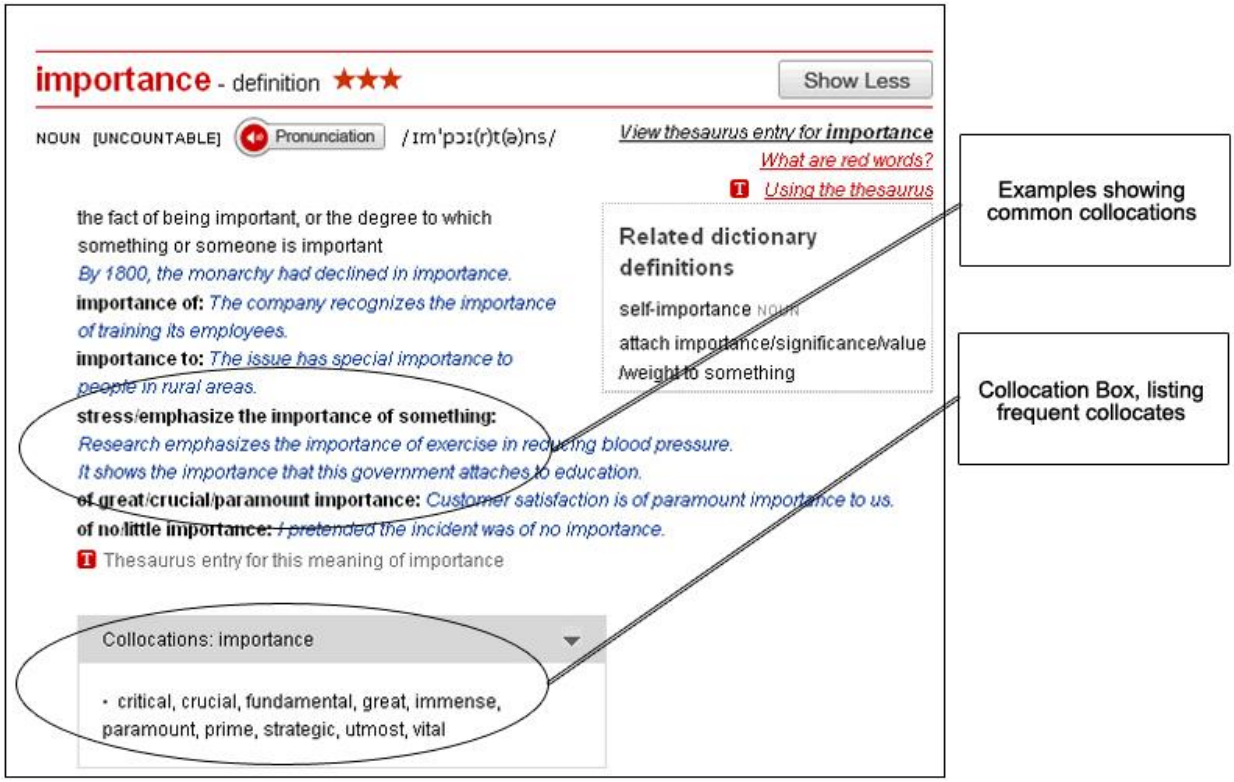
இது *decide* என்பதுடன் அடிக்கடி நிகழும் அமைப்பொழுங்கு எச்சத்தொடர் ('Vinf_to': Three months after that they decided to terminate my employment on health grounds) என்பதை காட்டுகிறது. தரவுத்தொகுதியில் இதற்கு 132,188 எடுத்துக்காட்டுகள் உள்ளன, இது decided பயன்படுத்தப்படும் எல்லா நிகழ்வுகளிலும் கிட்டத்தட்ட பாதி ஆகும். அடுத்த மிகவும் பொதுவான ஒழுங்கமைப்பு that-எச்சத்தொடர் ('that_0': They decided that surrender was the only sensible option), மற்றும் பல.

முடிவெடுப்பதில் அடிக்கடி நிகழும் முறை ஒரு முடிவற்ற பிரிவு ('வின்ஃப்_டோ': மூன்று மாதங்களுக்குப் பிறகு அவர்கள் சுகாதார அடிப்படையில் எனது வேலையை நிறுத்த முடிவு செய்தனர்) என்பதை இது காட்டுகிறது. தரவுத்தொகுதியில் இதற்கு 132,188 எடுத்துக்காட்டுகள் உள்ளன; இது முடிவெடுக்கும் எல்லா நிகழ்வுகளிலும் கிட்டத்தட்ட பாதி. அடுத்த மிகவும் பொதுவான முறை அந்த விதிமுறையுடன் ('that_0': சரணடைதல் மட்டுமே விவேகமான விருப்பம் என்று அவர்கள் முடிவு செய்தனர்), மற்றும் பல.

சொல்லடிவகைப்பாடு (Collocation)

வேர்ட் ஸ்கெட்ச் மென்பொருளானது சொல்லடிவகைப்பாடு அல்லது ஒன்றாகச் வரும்ம் போக்கைக் கொண்ட சொற்களைப் பற்றிய உயர்தர தகவல்களை வழங்குகிறது. ஆதாரங்களுடன் அடிக்கடி பயன்படுத்தப்படும் வினைச்சொற்களின் பட்டியலில் இதை (மேலே) காணலாம், மேலும் இந்த மென்பொருளைப் பயன்படுத்துவது என்பது முதன்முறையாக சொல்லடிவகைப்பாடு பற்றிய ஒரு விரிவான கணக்கைக் கொடுக்க முடியும் என்பதாகும். ஆங்கிலம் இரண்டாவது மொழியாக இருக்கும் எவருக்கும் இது மிகவும் மதிப்பு வாய்ந்தது, ஏனென்றால் உங்கள் கருத்துக்களை இயற்கையாகவும் பொதுவானதாகவும் வெளிப்படுத்தும் வகையில் மோதல் ஒரு முக்கியமாகும். முக்கியத்துவம் வாய்ந்த வார்த்தைக்கான இந்த இடுகையில், அடிக்கடி சேர்ந்துவருகை இரண்டு வழிகளில் காட்டப்படுகின்றன:

- எடுத்துக்காட்டுகளில், *stress*, *emphasize* போன்ற சேர்ந்துவருகைகளை விளக்கும் வாக்கியங்கள் உள்ளன, மேலும் எதையாவது முக்கியத்துவத்தை வலியுறுத்துகின்றன
- ஒரு சொல்லடிவகைப்பாட்டுப் பெட்டியில், இது பெரும்பாலும் முக்கியத்துவத்துடன் செல்லும் சொற்களை ('சொல்லடிவகைப்பாடு) பட்டியலிடுகிறது.



மேக்மில்லன் சொல்லடிவகைப்பாடு அகராதியை (Macmillan Collocations Dictionary) உருவாக்குவதற்கு அதே மென்பொருள் பயன்படுத்தப்பட்டது; இது இயற்கையான ஒலி சேர்க்கைகளை உருவாக்க ஆங்கில சொற்கள் எவ்வாறு ஒன்றிணைகின்றன என்பதற்கான இன்னும் விரிவான விளக்கத்தை வழங்குகிறது.

நடை மற்றும் பிராந்திய வகைகள்

இதுவரை பார்த்த எல்லாச் சொற்களும் (*remember, decide, evidence, importance*) எந்தச் சூழ்நிலையிலும் பயன்படுத்தப்படலாம்: நீங்கள் அவற்றை உரையாடலில் பயன்படுத்தலாம், செய்தித்தாளில் படிக்கலாம் அல்லது கல்வி இதழில் பார்க்கலாம். அவை மொழியியலாளர்களால் 'குறிக்கப்படாதவை' ('unmarked') என்று அழைக்கப்படுகின்றன. ஆனால் சில சொற்கள் மற்றும் வெளிப்பாடுகள் முக்கியமாக ஒரு குறிப்பிட்ட வகை உரையில் காணப்படுகின்றன: எடுத்துக்காட்டாக, பேசும் மொழியில், அல்லது செய்தித்தாள்கள் அல்லது தொழில்நுட்ப எழுத்தில். இதேபோல், பெரும்பாலான ஆங்கில சொற்கள் ஆங்கிலம் பேசும் உலகம் முழுவதும் பயன்படுத்தப்படுகின்றன; ஆனால் சில பிரிட்டிஷ் ஆங்கிலம் அல்லது இந்திய ஆங்கிலம் போன்ற ஒரு குறிப்பிட்ட பிராந்திய வகை ஆங்கிலத்தைச் சேர்ந்தவை.

தரவுத்தொகுதியிலிருந்து இந்த வாக்கியத்தைப் பாருங்கள்:

These two distinct eateries say much about why Charleston has become a mecca for food-lovers.

eatery என்பது 'restaurant'' என்பதற்கான மற்றொரு சொல் - ஆனால் அது 'குறிக்கப்படாதது' ('unmarked').அல்ல. தரவுத்தொகுதியில் உள்ள eatery என்பதன் அனைத்து எடுத்துக்காட்டுகளையும் பார்க்கும்போது, பெரும்பான்மை செய்தித்தாள்கள் மற்றும் பத்திரிகைகளிலிருந்து வந்திருப்பதைக் காண்கிறோம், மேலும் இந்தச் செய்தித்தாள்கள் மற்றும் பத்திரிகைகளில் பெரும்பாலானவை அமெரிக்காவிலிருந்து வந்தவை. எனவே அகராதியில், eatery என்ற சொல்லுக்கு இரண்டு 'புலக்குறிப்புகள் ('லேபிள்கள்'/'labels') உள்ளன: முக்கியமாக அமெரிக்கன் மற்றும் முக்கியமாக பத்திரிகை. தரவுத்தொகுதியின் சான்றுகள் இது போன்ற புலக்குறிப்புகளை/லேபிள்களை நம்பிக்கையுடன் பயன்படுத்த உதவுகிறது.

அதிர்வெண், அது ஏன் முக்கியமானது

மொழியில், அடிக்கடி நிகழும் ஒன்றைக் கற்றுக்கொள்வது மிகவும் பயனுள்ளதாக இருக்கும். *ameliorate* மற்றும் *improve* என்ற சொற்கள் அதிகமாகவோ அல்லது குறைவாகவோ ஒரே மாதிரியாக இருக்கின்றன - ஆனால் *improve* சுமார் 250 மடங்கு பொதுவானது. improve என்பது ஆங்கிலத்தின் சொற்றொகுதியின் 'முக்கியப்' பகுதியாக இருப்பதால், *improve* (அதன் பொருள், இலக்கணம் மற்றும் மோதல்கள்) என்பதைக் கற்பது நலம்பயக்கும்: நீங்கள் அதை அடிக்கடி பார்ப்பீர்கள், கேட்பீர்கள், மேலும் நீங்கள் அதை அடிக்கடி பயன்படுத்த வேண்டியிருக்கும். *ameliorate* இது போன்றதல்ல: நீங்கள் அதைக் காண நேர்ந்தால் (இது சாத்தியமில்லை, ஏனென்றால் இது மிகவும் அரிதானது), நீங்கள் அதை ஒரு அகராதியில் பார்க்கலாம், ஆனால் எந்த சக்தியையும் வீணாக்குவது நல்லது அல்ல.

மிகப் பெரிய தரவுத்தொகுதியுடன், எந்தச் சொற்கள் அடிக்கடி நிகழ்கின்றன என்பதை மட்டுமல்லாமல், எந்த இலக்கண வடிவங்கள் (*decide* + வினையெச்சம் (infinitive) போன்றவை) மற்றும் எந்த சொல்லடிவகைப்பாடு (*crucial* + *importance* போன்றவை) என்பதையும் அடையாளம் காண்பது எளிது. இந்த மேக்மில்லன் அகராதிகளில் நாம் அடிக்கடி விரிவாக விளக்கும் இந்த அடிக்கடி சொற்கள் மற்றும் சேர்க்கைகள் தான், மேலும் 'சிவப்பு' மற்றும் 'கருப்பு' சொற்களுக்கு (('red' and 'black' words (எல்லாவற்றிலும் மிக முக்கியமான அம்சம், மேக்மில்லன் ஆங்கில அகராதி உயர் அதிர்வெண் மைய சொற்களஞ்சியம் மற்றும் முக்கியமாக குறிப்புக்குத் தேவையான குறைவான பொதுவான சொற்களுக்கு இடையில் தெளிவான வேறுபாட்டைக்

கொண்டுள்ளது.)) இடையில் நாம் காணும் வேறுபாடு நமது அகராதிகளின் தனித்துவமான அம்சங்களில் ஒன்றாகும்.

8. எடுத்துக்காட்டுகளைக் கண்டுபிடிக்கத் தரவுத்தொகுதியைப் பயன்படுத்துதல்

அகராதி பயனர்கள் எடுத்துக்காட்டு வாக்கியங்களைப் பாராட்டுகிறார்கள். ஒரு சொல் சூழலில் ஒரு சொல் எவ்வாறு செயல்படுகிறது என்பதைக் காட்டும் ஒரு சிறந்த எடுத்துக்காட்டு, அதன் அர்த்தத்தை விளக்க உதவுகிறது. ஒரு அகராதியில் உள்ள ஒரு வார்த்தையின் எடுத்துக்காட்டு நிஜ வாழ்க்கையில் இந்த வார்த்தையைப் பயன்படுத்தும் விதமாக இருக்க வேண்டும் - எனவே தரவுத்தொகுதியை எடுத்துக்காட்டு வாக்கியங்களின் ஆதாரமாகப் பயன்படுத்துகிறோம்.

தேர்வு செயல்முறை எவ்வாறு செயல்படுகிறது என்பதைப் பார்க்க, மேலே உள்ள முக்கியத்துவத்திற்கான உள்ளீட்டைத் திரும்பிப் பாருங்கள். அகராதியில் (மற்றும் இந்த விஷயத்தில், பல்வேறு பொதுவான சொல்லடிவகைப்பாடுகளை உள்ளடக்கிய) உள்ளிட்ட மிகவும் மதிப்புள்ள வார்த்தையைப் பற்றிய உண்மைகளை அடையாளம் காண வேர்ட் ஸ்கெட்ச் மற்றும் ஒத்திசைவுகளைப் பயன்படுத்துகிறோம். ஆனால் முதல் எடுத்துக்காட்டைக் கவனியுங்கள்:

By 1800, the monarchy had declined in importance.

இந்த எடுத்துக்காட்டு தேர்ந்தெடுக்கப்பட்டது, ஏனென்றால் தரவுத்தொகுதியில் *importance* என்ற வெளிப்பாட்டின் கிட்டத்தட்ட ஆயிரம் எடுத்துக்காட்டுகளை அதன் முன்னால் ஒரு வினைச்சொல்லுடன் காணலாம். இதன் பொருள் முக்கியத்துவம் பயன்படுத்தப்படும் முறையின் பொதுவான அம்சங்களில் ஒன்றாகும். இந்த நிலையில் நிகழும் வினைச்சொற்கள் பொதுவாக *increase, grow, and gain, or decline, decrease* அல்லது *diminish* போன்ற சொற்கள் என்று மேலும் ஆராய்ச்சி காட்டுகிறது.

இந்த அமைப்பொழுங்கின் பல எடுத்துக்காட்டுகளில், இந்த வரிசையை விளக்கும் பலவற்றைக் காண்கிறோம்: [By [date] X had declined in importance.

எடுத்துக்காட்டாக, இவற்றில் ஒன்று பின்வருமாறு கூறுகிறது:

> By the early 12th century, the monasteries, which had been the focal points of religious life, had declined in importance and the way was ready for the introduction of the diocesan system.

ஆனால் இந்த வாக்கியம் அகராதிக்கு மிக நீளமானது, மேலும் இது தேவையற்ற கூடுதல் தகவல்களைக் கொண்டுள்ளது. எனவே இந்த வாக்கியம் சிறிது மாற்றப்பட்டது; அது அகராதியில் பின்வருமாறு சுருக்கப்பட்டது:

By 1800, the monarchy had declined in importance.

இந்த எடுத்துக்காட்டு குறுகிய மற்றும் புரிந்துகொள்ள எளிதானது - ஆனால் இது தரவுத்தொகுதியில் *importance* பயன்படுத்தப்படுவதை உண்மையாகப் பிரதிபலிக்கிறது.

9. 'கற்கும் தரவுத்தொகுதியைப்' பயன்படுத்துதல்: மேக்மில்லன் மற்றும் சி.இ.சி.எல்

மேக்மில்லன் ஆங்கில அகராதியின் (Macmillan English Dictionary) இரண்டாம் பதிப்பை உருவாக்கும் போது, பெல்ஜியத்தில் உள்ள யுனிவர்சிட்டி கத்தோலிக் டி லூவெய்னில் உள்ள ஆங்கில தரவுத்தொகுதி மொழியியல் மையத்துடன் (Centre for English Corpus Linguistics (CECL/சி.இ.சி.எல்) ஆராய்ச்சி ஒத்துழைப்பின் மூலம், வேறுபட்ட வகை தரவுத்தொகுதி - கற்கும் தரவுத்தொகுதி பயன்படுத்தப்பட்டது.

புதுமையான மென்பொருளைப் பயன்படுத்தி, அகராதியலார்கள் மேக்மில்லன் ஆங்கில அகராதியை (Macmillan English Dictionary (MED)) 200 மில்லியனுக்கும் அதிகமான சொற்களைக் கொண்ட ஒரு தனித்துவமான நவீன தரவுத்தொகுதியை - உலக ஆங்கில தரவுத்தொகுதியை (World English Corpus) - அடிப்படையாகக் கொண்டுள்ளனர். MEDஇன் இரண்டாவது பதிப்பு பெல்ஜியத்தில் உள்ள யுனிவர்சிட்டி கத்தோலிக் டி லூவெய்னில் (Université catholique de Louvain) உள்ள ஆங்கில கார்பஸ் மொழியியல் மையத்துடன் (Centre for English Corpus Linguistics) இணைந்து இந்தத் தரவுத்தொகுதியில் சேர்க்கப்பட்டது.

சி.இ.சி.எல் அதன் இயக்குனரான சில்வியன் கிரானெஜரின் கீழ், கற்றல் தரவுத்தொகுதிகளின் வளர்ச்சி மற்றும் பயன்பாட்டில் கவனம் செலுத்துகிறது. ஒரு கற்றல் தரவுவுத்தொகுதியில் உள்ள உரை பேச்சு மற்றும் எழுத்தை உள்ளடக்கியது, இது சொந்த மொழி பேசுபவர்களால் அல்ல, ஆனால் ஒரு மொழியைக் கற்கும் நபர்களால் தயாரிக்கப்படுகிறது. CECLஇன் கற்கும் தரவுத்தொகுதியில் உலகளாவிய ஆங்கிலக் கற்றவர்களின் தரவுகள் அடங்கும், மேலும் இவை பொதுவான கற்பவர்களின் பிரச்சினைகள் குறித்து அகராதியியலார்களுக்கு ஏராளமான தகவல்களை வழங்கின. எடுத்துக்காட்டாக, கற்பவர்களுக்கு உதவி வழங்க இந்தத் தகவல் பயன்படுத்தப்பட்டுள்ளது:

- சரியான சொல்லடிவகைப்பாடுகளை முன்னிலைப்படுத்துதல் மற்றும் எடுத்துக்காட்டுதல்;

- அதிகமாகப் பயன்படுத்தப்படும் சொற்றொகுதி உருப்படிகளுக்குப் பயனுள்ள மாற்றுகளை வழங்குதல்;
- எளிதில் குழப்பம் விளைவிக்கும் சொற்களுக்கு இடையிலான வேறுபாடுகளை சுட்டிக்காட்டும் குறிப்புகளை வழங்குதல்;
- பொதுவான பிழைகளுக்குக் கற்பவர்களை எச்சரிக்க குறிப்பிட்ட எச்சரிக்கைகளை வழங்குதல்.

இது அகராதியின் மையத்தில் உங்கள் எழுதும் திறன்களை மேம்படுத்துதல் பிரிவு (Improve your Writing Skills section), தனிப்பட்ட தலைப்புகளில் சரியான பெட்டிகளைப் பெறுங்கள் (Get it right boxes at individual headwords) மற்றும் சிடிரோமில் பயிற்சிகள் (exercises on the CD-ROM) என்பன கற்பவர்களுக்கு அவர்களின் எழுத்தை மேம்படுத்த உதவும் தனித்துவமான புதிய பொருட்களின் வளர்ச்சிக்கு வழிவகுத்தது, மேக்மில்லன் ஆங்கில அகராதியின் இரண்டாவது பதிப்பு, கற்றவரின் தரவை இந்த முறையான வழியில் பயன்படுத்திய முதல் அகராதி ஆகும்.

6.3.4. இருமொழிய அகராதிகளின் உருவக்கம் (அச்சிடப்பட்ட மற்றும் மின்னணு)

இருமொழி அகராதி அல்லது மொழிபெயர்ப்பு அகராதி (bilingual dictionary or translation dictionary) என்பது ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு சொற்கள் அல்லது சொற்றொடர்களை மொழிபெயர்க்கப் பயன்படும் ஒரு சிறப்பு அகராதி. இருமொழி அகராதிகள் ஒரு திசையில் இருக்கக்கூடும், அதாவது அவை ஒரு மொழியின் சொற்களின் அர்த்தங்களை இன்னொரு மொழியில் பட்டியலிடுகின்றன; அல்லது இரு திசையில் இருக்கக்கூடும்; இது இரு மொழிகளுக்கும் மொழிபெயர்ப்பையும் அனுமதிக்கிறது. இருதிசை இருமொழி அகராதிகள் வழக்கமாக இரண்டு பிரிவுகளைக் கொண்டிருக்கின்றன, ஒவ்வொன்றும் ஒரு மொழியின் சொற்கள் மற்றும் சொற்றொடர்களை அகர வரிசைப்படி அவற்றின் மொழிபெயர்ப்புடன் பட்டியலிடுகின்றன. மொழிபெயர்ப்பிற்கு கூடுதலாக, ஒரு இருமொழி அகராதி பொதுவாக சொல்வகைப்பாடு, பாலினம், வினை வகை, திரிபு மாதிரி மற்றும் பிற இலக்கண தடயங்களைக் குறிக்கிறது. இருமொழி அகராதிகளில் சில நேரங்களில் இருக்கும் பிற அம்சங்கள் சொற்றொடர்கள், பயன்பாடு மற்றும் நடை வழிகாட்டிகள், வினை அட்டவணைகள், வரைபடங்கள் மற்றும் இலக்கண குறிப்புகள். இருமொழி அகராதிக்கு மாறாக, ஒரு மொழியியல் அகராதி சொற்களையும் சொற்றொடர்களையும் மொழிபெயர்ப்பதற்கு பதிலாக வரையறுக்கிறது.

இருமொழி அகராதிகள் பல வடிவங்கள்

இருமொழி அகராதிகள் பல வடிவங்களில் கிடைக்கின்றன, மேலும் அவை பெரும்பாலும் இலக்கண குறிப்பு மற்றும் பயன்பாட்டு எடுத்துக்காட்டுகளையும் உள்ளடக்குகின்றன. (எடுத்துக்காட்டாக யாத்கர் சிந்தி -ஆங்கில அகராதி (Yadgar Sindhi to English Dictionar)).

- அச்சிடப்பட்ட அகராதிகள் (Printed dictionaries) - அச்சிடப்பட்ட அகராதிகள் சிறிய பாக்கெட் அளவிலான பதிப்புகள் முதல் பெரிய, விரிவான பல தொகுதி படைப்புகள் வரை இருக்கும்.
- கையடக்க மின்னணு அகராதிகள் (Handheld electronic dictionaries) (மேலும்: பாக்கெட் எலக்ட்ரானிக் அகராதிகள் அல்லது பெட்கள் (Pocket electronic dictionaries or PEDs) - மின்னணு அகராதிகள் ஒரு மினியேச்சர் விசைப்பலகை, பேச்சு புரிந்துகொள்ளுதல் அல்லது அச்சிடப்பட்ட உரையைப் படிக்கும் ஸ்கேனிங் சாதனம் வழியாக உள்ளீட்டைப் பெறும் சிறிய சாதனங்கள், மற்றும் மொழிபெயர்ப்பை ஒரு சிறிய எல்சிடி திரையில் வெளியிடுகிறது அல்லது மொழிபெயர்ப்பைப் கேட்கக்கூடியதாக பேசுகின்றது.
- அகராதி நிரல்கள் (Dictionary programs) - சொற்கள் அல்லது சொற்றொடர்களை உள்ளீடு செய்து கணினிகள் மற்றும் ஸ்மார்ட் போன்களில் மொழிபெயர்க்க அனுமதிக்கும் மென்பொருள்.
- ஆன்லைன் அகராதிகள் (Online dictionaries) - ஆன்லைன் அகராதிகள் அகராதி நிரல்களைப் போன்றவை; இவை பெரும்பாலும் தேட எளிதானவை, ஆனால் எப்போதும் பயன்படுத்த இலவசமல்ல; சில சந்தர்ப்பங்களில் துல்லியம் (குறிப்பாக திறந்த கூட்டு அகராதிகளில்) அல்லது அச்சிடப்பட்ட மற்றும் மின்னணு அகராதிகளின் நோக்கம் இல்லை.
- காட்சி அகராதிகள் **(Visual dictionaries)** - ஒரு காட்சி அகராதி என்பது அச்சிடப்பட்ட அகராதியாகும், இது சரியான மொழிபெயர்ப்பை அடையாளம் காண்பதற்கான நம்பகமான வழியை பயனருக்கு வழங்க முதன்மையாக விளக்கப்படங்களை நம்பியுள்ளது. காட்சி அகராதிகள் பெரும்பாலும் இருமொழியைக் காட்டிலும் பல மொழிகளாக இருக்கின்றன. இரண்டு மொழிகளுக்கு இடையிலான மொழிபெயர்ப்புகளைக் கொண்டிருப்பதற்குப் பதிலாக அவை பெரும்பாலும் நான்கு அல்லது அதற்கு மேற்பட்ட மொழிகளை உள்ளடக்கும்.

தரவுத்தொகுதியைப் பயன்படுத்தி இருமொழிய அகராதி உருவாக்கம்

கடந்த தசாப்தங்களாக அகராதியியல் (லெக்சோகிராஃபி/Lexicography) தரவுத்தொகுதி மொழியியல் முறைகளை இணைத்துள்ளது. மின் அகராதியில் (எலக்ட்ரானிக் டிக்ஸ்னரி/ electronic dictionary) வேலை செய்யத் தொடங்கும் அகராதியலார்கள் (லெக்சோகிராஃபர்கள்/ Lexicographers), புதிதாக கணினி மொழியியலாளர்களாகத் தொடங்கி, அவர்களின் மொழி ஜோடியில் முந்தைய அல்லது குறைவான வேலைகள் எதுவும் செய்யப்படாமல், தரவுத்தொகுதி மொழியியல் முறைகள் தங்கள் திட்டத்திற்கு வழங்கக்கூடிய பங்களிப்புகளை மதிப்பீடு செய்ய வேண்டும், தலைப்புச்சொல்லன் பட்டியல் உருவாக்கம் (லெமா லிஸ்ட் புல்டிங்/lemma list building) மட்டுமல்ல, இருமொழி அகராதி வரைவுகள் மற்றும் நிலையான நுழைவு எடிட்டிங்கில் அவற்றின் ஆவணமாக்கல் செயல்முறை, ஆனால் மாறும் வகையில் உருவாக்கப்பட்ட கார்பஸ் தரவு காட்சிகளைக் கொண்ட அகராதி வெளியீட்டிற்கும். குறைந்த அல்லது நடுத்தர அடர்த்தி கொண்ட மொழி ஜோடியின் சூழலில், எந்த மின்னணு வளங்கள் மற்றும் கருவிகள் தேவை மற்றும் கிடைக்கின்றன என்பதை அவர்கள் கேட்க வேண்டும், மேலும் இருமொழி அகராதி வரைவு என அவற்றின் போதுமான அளவுக்கான கணக்கீட்டு முறைகளுடன் பெறப்பட்ட இருமொழி சொற்களஞ்சியங்களை bilingual glossaries மதிப்பீடு செய்ய வேண்டும்.

இன்று, "தரவுத்தொகுதிப் புரட்சி" (corpus revolution) என்று அழைக்கப்படுபவை (எ.கா. Rundell & Stock/ருண்டெல் & ஸ்டாக், 1992; கிருஷ்ணமூர்த்தி/Krishnamurthy,, 2002; ஹாங்க்ஸ்/Hanks, 2012) அகராதிகள் பயன்பாட்டில் உள்ள மொழியைச் சிறப்பாக பிரதிபலிக்க உதவியது என்பதில் யாருக்கும் சந்தேகமில்லை. எடுத்துக்காட்டாக, தரவுத்தொகுதிக்கு முந்தைய அகராதிகள் (pre-corpus dictionaries) அரிதான சொல் அர்த்தங்களை (மற்றும் மொழிபெயர்ப்பு நிகரன்களைக்) கணக்கில் எடுத்துக்கொள்வதைக் காண முடிந்தது; ஆனால் முக்கியமான பொதுவானவற்றைத் தவறவிட்டன (க்ளோசா/Klosa,, 2007: 111). இப்போது, அதிர்வெண் அளவீடுகள் ஒரு நிலையான அகராதி பணிப்பாய்வுகளின் ஒரு பகுதியாகும்; மேலும் போதுமான பெரிய தரவுத்தொகுதிகள் மற்றும் தொடர்புடைய இயற்கை மொழி ஆய்வுக் (Natural Languge Processing (NLP/என்எல்பி) கருவிகள் இருந்து, ஒருவர் பணிபுரியும் மொழிக்கு அணுகக்கூடியதாக இருந்தால்தான் தொடரியல் உறவுகளால் வரிசைப்படுத்தப்பட்ட சேர்ந்துவருகைகள் (collocates) மற்றும் இணை நிகழ்வுகளைப் (co-occurrences) பற்றிக் கூறும் சொல் ஓவியங்கள் (word sketches) (கில்கரிஃப் & டக்வெல்/Kilgarriff & Tugwell,, 2002) உடனடியாக உருவாக்கப்படலாம்; ஆனால் முக்கியச் சொற்கள் இயைபுகள் (keyword concordances), அதிர்வெண் தரவு மற்றும்

சொல் ஓவியங்கள் (word sketches) என்பன நிலையான அகராதிப் பதிவு தொகுத்தலில் அகராதி ஆவணங்களின் ஆவணப்படுத்தல் செயல்முறைக்கு ஒரு தவிர்க்க முடியாத ஆதாரமாக மாறியதோடு மட்டுமல்லாமல், மேலும் மேலும் அகராதி வலைத்தளங்களில் வரவால் நிலையான "தலையங்கம்" அகராதிப் பதிவு (static "editorial" dictionary entry) தரவுத்தொகுதியால் இயக்கப்படும் உள்ளடக்கத்தால் முழுமை செய்யப்படுகிறது (அல்லது மாற்றப்படுகிறது).

இணைத் தரவுத்தொகுதிகள்

ஒரு இணைத் தரவுத்தொகுதி என்பது இருமொழி அல்லது பன்மொழித் தரவுத்தொகுதி ஆகும், இது இரண்டு அல்லது அதற்கு மேற்பட்ட மொழிகளில் உரைகளைக் கொண்டுள்ளது. அவற்றில் பல விருப்பங்கள் உள்ளன:

- 'அ' மொழியில் எழுதப்பட்ட நூல்களை மட்டுமே கொண்ட இணை தரவுத்தொகுதி மற்றும் அவற்றின் இணையான மொழிபெயர்ப்பு மொழிகளில் ஆ (மற்றும் இ...);
- அ மற்றும் ஆ மொழிகளில் எழுதப்பட்ட சமமான நூல்கள் மற்றும் அவற்றின் மொழிபெயர்ப்புகளைக் கொண்ட இணையான தரவுத்தொகுதி;
- அ, ஆ மற்றும் இ மொழிகளில் உரைகளின் மொழிபெயர்ப்பை மட்டுமே கொண்ட இணை தரவுத்தொகுதி, அதே நேரத்தில் உரைகள் முதலில் இசட் மொழியில் எழுதப்பட்டன. (Teubert 1996:245)

வேறு வார்த்தைகளில் கூறுவதானால், மூலமொழியில் இருந்து இலக்குமொழிக்கு இரு திசை மொழிபெயர்ப்புகள் அல்லது ஒரே ஜோடி மொழிகளுக்கான பரஸ்பர இணையான தரவுத்தொகுதிகள் மற்றும் அசல் பதிப்பு இல்லாமல் இலக்குமொழிகளில் பல மொழிபெயர்ப்புகளை நாம் கொண்டிருக்கலாம்.

மொழி ஆய்வின் பல பயன்பாடுகளில் இணையான தரவுத்தொகுதிகள் சமீபத்தில் மேலும் பிரபலமாகிவிட்டாலும், தப்பெண்ணங்கள் இன்னும் பெரும்பாலும் அவைகளுக்கு எதிராகக் குரல் கொடுக்கின்றன, மேலும் மொழியியல் பகுப்பாய்வில் அவற்றின் பயன்பாட்டினைப் பற்றி ஆட்சேபனைகள் வெளிப்படுத்தப்படுகின்றன. இணையான தரவுத்தொகுதிகளைப் பயன்படுத்துவதற்கு எதிரான வாதங்கள், அதாவது, மொழிபெயர்ப்புகளை அடிப்படையாகக் கொண்ட தரவுத்தொகுதிகள், அ) மொழிபெயர்ப்புகள் இலக்குமொழியைச் சிதைக்கின்றன, ஏனெனில் அவை மூலமொழியின் கண்ணாடிப் படத்தைக் கொடுக்கின்றன, ஆ)

மொழிபெயர்க்கப்பட்ட மொழி அசல் மொழியிலிருந்து வேறுபட்டது மற்றும் இ) மொழிபெயர்ப்பாளர்கள் நம்பமுடியாதவர்கள் மற்றும் தவறுகள் செய்யவர்.

ஆயினும்கூட, மொழிபெயர்ப்பாளர்களின் உள்ளுணர்வு மொழிகளுக்கிடையேயான வேறுபாட்டைக் கொண்டிருப்பதால், இணையான தரவுத்தொகுதி பொதுவாக மொழியியல் மற்றும் குறிப்பாக மாறுபட்ட ஆய்வுகளில் சிறப்பு பங்களிப்பைக் கொண்டுள்ளது. தவிர, மொழிபெயர்ப்பாளர்களைப் பகுப்பாய்வு செய்வதைத் தவிர்த்து மொழிபெயர்ப்பாளரின் மொழிபெயர்ப்பு நிகரன்களைப் பற்றிய அறிவைப் பிடிக்க வேறு வழியில்லை. எனவே அவர்கள் எந்த சந்தேகமும் இல்லாமல் மிகப் பெரிய தாக்கத்தை ஏற்படுத்தக்கூடிய புலம் என்பது அகராதி.

தரவுத்தொகுதி அடிப்படையிலான இருமொழி அல்லது பன்மொழி அகராதி இன்னும் ஆரம்ப நிலையில் உள்ளது. சில சொற்பொழிவாளர்கள் இணையான தரவுத்தொகுதியுடன் பணிபுரிந்துள்ளனர். தவிர, பாரம்பரியமாக, இருமொழி அகராதிகள் ஒரு கெட்ட பெயரைக் கொண்டுள்ளன. சம்பந்தப்பட்ட மொழிகளில் லெக்சிக்கல் அலகுகளுக்கு இடையில் மொழிபெயர்ப்பு நிகரன்களைக் கண்டுபிடிப்பதே பன்மொழி அகராதிகளின் பணி. அகராதிக்கு இணையான தரவுத்தொகுதிகளை வழங்க வேண்டியது உரைகள், வாக்கியத்தால் சீரமைக்கப்பட்ட வாக்கியம்.

உரைகள் வார்த்தைக்கு வார்த்தையாக மொழிபெயர்க்கப்படவில்லை, மாறாக வாக்கியத்திற்கு வாக்கியமாக மொழிபெயர்க்கப்படுகின்றது என்பதை நாம் அனைவரும் அறிவோம். இதனால் பன்மொழி அகராதிக்கு சொல்லடிவகைப்பாடு ஒரு முக்கியமான பிரச்சினையாகிறது.

6.3.5. பன்மொழிய அகராதிகளின் உருவக்கம் (அச்சிடப்பட்ட மற்றும் மின்னணு)

பெரும்பாலும் இருமொழிய அகராதிகளைப் பயன்படுத்தி பன்மொழிய அகராதிகள் உருவாக்கப்படன. தரவுத்தொகுதிகளைப் பயன்படுத்திப் பன்மொழிய அகராதிகள் உருவாக்கம் தொடக்க நிலையிலேயே உள்ளது. விக்கிப்பீடியாவைப் பயன்படுத்தி பன்மொழிய அகராதி உருவக்கம் மேற்கொள்ளப்பட்டுள்ளது.

யூரோசொல்வலை (Euro-wordNet) ஒரு பன்மொழியச் சொல்சார் தரவுத்தளமாகும் (multilingual lexical database). இது அகராதிகளின் (dictionaries) பண்புக்கூறுகளையும் சொற்களஞ்சியங்களின் (thesauri) பண்புக்கூறுகளையும் உள்ளடக்கியது. இதில் சொற்கள் ஒருபொருள் பன்மொழியக் குழுமங்களாகக் குழுமப்பட்டு அவைகள் சொல்சார் உறவுகள்,

பொருண்மை உறவுகளால் [உள்ளடங்குமொழியம்-உள்ளடக்குமொழியம் (hyponymy-hypernymy), சினைமொழியம்-முழுமொழியம் (meronymy-holonymy), எதிர்மொழியம் (antonymy) போன்ற உறவுகளால்] இணைக்கபட்டு ஒரு அடித்தளமான மூலப்பொருண்மையியல் (ontology) மீது சொல்வலைகளாகக் கட்டப்பட்டு ஒருங்கிணைக்கபட்ட அமைப்பாகும். அகராதியை போல் இதில் ஒருபொருள் பன்மொழியக் குழுமங்களுக்கு விளக்கமும் எடுத்துக்காட்டும் தரப்பட்டுள்ளன. இதன் உருவாக்கத்திற்காகத் தரவுத்தொகுதிகளின் தரவுகள் பயன்படுத்தப்பட்டுள்ளன. இந்திய மொழிகளுக்கான சொல்வலை (Indo-wordNet) உருவாக்க முயற்சியும் மேற்கொள்ளப்பட்டு ஒரு நிலையை (30,000 ஒருபொருள் பன்மொழியக் குழுமங்கள் அடங்கிய சொல்வலையாக) அடைந்துள்ளது. தமிழ்சொல்வலையும் இதன் ஒரு பகுதியாக அமையும்.

6.3.6. ஒருமொழியச் சொற்களஞ்சிய உருவாக்கம் (அச்சிடப்பட்ட மற்றும் மின்னணு)

சொற்களின் தொகுப்பு சொற்களஞ்சியம் (thesaurus) எனப்படும். சொற்களஞ்சியம் என்பதை பொருட்புல அகராதி என்றும் குறிப்பிடுவர். சொற்களஞ்சியம் கருத்துருக்களிலிருந்து அல்லது பொருண்மைகளிலிருந்து (அர்த்தங்களிலிருந்து) சொற்களை அடைய உதவும். சொற்களஞ்சியம் என்பதன் விளக்கம் பின்வரும் ஜோன்ஸின் (Jones, 1986:201) மேற்கோள் மூலம் தெளிவாகும்: “ஒரு சொற்களஞ்சியத்தில் ஒருபொருள் பன்மொழிகள் இருக்க வேண்டிய கட்டாயம் இல்லை. மற்றும் ஒருபொருள் பன்மொழி அகராதி சொற்களஞ்சியமாக அமையத் தேவையில்லை. விரிவான பொருளில் சொற்களஞ்சியம் என்பது சொற்களைக் கருத்துருக்கள் (Concepts), தலைப்புகள் (topics) அல்லது பாடப்பொருள்கள் (Subjects) ஆகியவற்றின் அடிப்படையில் பாகுபாடு செய்வதாகும்; ஒரு வகுப்பில் தரப்பட்டுள்ள சில சொற்கள் ஒருபொருள் பன்மொழிகளாகும் என்ற செய்தி நிச்சயமாகத் தேவையற்ற ஒன்றாகும். மாறாக ஒருபொருள் பன்மொழி அகராதி ரொஜெஸ்டின் சொற்களஞ்சியத்தில் (Rogest’s Thesaurus) தரப்பட்டுள்ள கருத்துருப் பாகுபாட்டைத் தராமல் குழுமங்களாக வரும் சொற்களைத்தான் பகுத்தளிக்கும். இது ஒரு பக்குவமற்ற வேற்றுமைப்படுத்தலாகும். சொற்களஞ்சியத்தில் ஒருபொருள் பன்மொழிகள் அடிக்கடி காணப்படும். ஒருபொருள் பன்மொழி அகராதிகளில் உள்ள குழுமங்கள் பெரிதாகவும் ஒருபொருள் பன்மொழியத்தின் பரந்த விளக்கத்திற்கு மாதிரியாகவும் அமைந்துள்ள குழுமங்களை இணைக்கும் குறுக்கு நோக்கீட்டுகளையும் கொண்டிருக்குமானால் அவை சொற்களஞ்சியத் தன்மை பெறலாம்.”

சொற்களஞ்சியம் பற்றிய விரிவான தகவல்கள் இராசேந்திரன் அவர்களால் (இராசேந்திரன், 2001) தரப்பட்டுள்ளன.சொற்களஞ்சியம் என்பது சொற்களை வரிசைப்படுத்தித் தருவது என்ற குறுகிய நோக்கில் அமையாமல் சொற்களின் வலைபோல் பின்னிக்கிடக்கும் பொருண்மை உறவுகளையும் சொல் உறவுகளையும் வெளிப்படுத்த வேண்டும் என்ற நோக்கில் அமைப்பாக்கம் செய்யப்படவேண்டும். சொற்களஞ்சியத்தில் சொற்றொகையைச் சார்ந்த பெயர், வினை, பெயரடை, வினையடை, பின்னுருபுகள், இணைப்பான்கள் மற்றும் பிற செயல்பாட்டுச் சொற்கள் (functional words) தரவுகளாச் சேகரிக்கப்பட்டு ஒரு வகைப்பாட்டியல் (taxonomy) வடிவத்தில் தர முயற்சிக்கப்பட்டுள்ளது.

உரை ஆவணங்களில் தொடர்புடைய சொற்களை அடையாளம் காண்பதற்கான புள்ளிவிவர சொல்லுக்குசொல் இணை நிகழ்வு நடவடிக்கைகளைப் பயன்படுத்துவதன் மூலம், தானியங்கி சொற்களஞ்சியம் தலைமுறை ஆரம்பத்தில் 1970களில் மேற்கொள்ளப்பட்டது. இந்த முறை பல குறைபாடுகளைக் கொண்டிருந்தது: தொடர்பில்லாத பல சொற்கள் அவை அடிக்கடி பயன்படுத்தப்படுவதால் இணைந்திருக்கலாம்; ஒருபொருள் பன்மொழிகள் எப்போதாவது ஒன்றாகப் பயன்படுத்தப்படுகின்றன; ஒற்றைச்சொல் சொற்கள் மட்டுமே கருதப்படுகின்றன, அதேசமயம் பல சிறப்பு களங்களில் பல சொற்கள் இணைந்த கூட்டுசொற்கள் அடிக்கடி பயன்படுத்தப்படுகின்றன; தொடர்புடைய சொற்களின் கொத்து, சொற்களுக்கு இடையிலான உறவுகள் பற்றிய எந்த அறிவும் இல்லாமல் தயாரிக்கப்படுகிறது. ஒரு சொல் அல்லது ஆவண சேகரிப்பில் இணைந்திருப்பதற்கு ஒத்த சொற்களஞ்சியங்கள் ஒத்த இணை நிகழ்வுகளைக் கொண்டிருக்கின்றன என்ற உண்மை, ஒரு சொல்லை அதன் சூழல் சார்ந்த தகவல்களின் அடிப்படையில் ஒரு சொற்றொடருடன் இணைப்பதன் மூலம் கையாளப்பட்டது. SEXTANT அமைப்பு நூல்களில் பலவீனமான தொடரியல் பகுப்பாய்வு முறைகளைப் பயன்படுத்துகிறது, இதேபோன்ற சொற்கள் ஒத்த தொடரியல் உறவுகளில் தோன்றும் என்ற அனுமானத்தின் கீழ் ஆய்வகத்தை உருவாக்குகின்றன. விதிமுறைகள் அவை தோன்றும் இலக்கண சூழலுக்கு ஏற்ப தொகுக்கப்படுகின்றன. மேலே உள்ள இரண்டு முறைகளும் சாத்தியமான அணுகுமுறைகள், ஆனால் சொற்கள், ஒருசொல்போலி அல்லது சினைமொழி ஆகியவற்றின் சொற்பொருள் உறவுகள் போன்ற சொற்களுக்கு இடையில் வரையறுக்கப்படாத உறவுகளின் குறைபாட்டை இன்னும் தீர்க்கவில்லை.

ஏராளமான மென்பொருள் தொகுப்புகள் (மில்ஸ்டெட்/Milstead, 1990) உள்ளன, அவை உரைசார் தகவல் மூலங்களுக்கு இடைமுகங்களை வினவ அல்லது உலாவலாகப் பயன்படுத்த கையால் சொற்களஞ்சியம் உருவாக்க அனுமதிக்கிறது. பொருண்மைக்களம் குறிப்பிட்ட ஆய்வறிக்கையின் இருப்பு, பொருண்மைக்களத்தில் உள்ள முக்கியமான கருத்துகளின் படிநிலை பார்வையை முன்வைப்பதன் இரட்டை நோக்கத்திற்கு உதவுகிறது, அத்துடன் பொருண்மைக்களத்தில் அதே கருத்தை விவரிக்க பயன்படுத்தக்கூடிய மாற்று சொற்கள் மற்றும் சொற்றொடர்களை பரிந்துரைக்கிறது. கோரிக்கையை வெளிப்படுத்துவதற்கான மாற்று வழிகளை அறிவது கணினி அடிப்படையிலான மீட்டெடுப்பில் பொதுவான சிக்கலாகும் (ஃபர்னாஸ் மற்றும் பலர்/Furnas et al., 1987, 1987). ஆகவே, தற்போதுள்ள ஆவண சேகரிப்பில் உள்ள தகவல்களைப் பயன்படுத்துவதற்கான ஒரு மதிப்புமிக்க இணைப்பாக ஒரு சொற்களஞ்சியம் கருதப்படுகிறது. இருப்பினும் பொருண்மைக்களம்-குறிப்பிட்ட மனிதமுயற்சியாலான சொற்களஞ்சிய உருவாக்கத்தில் இரண்டு பெரிய சிக்கல்கள் உள்ளன,

முதல் சிக்கல் சொற்களஞ்சியத்திற்குள் செல்ல சொற்களை அடையாளம் கண்டு அவற்றை ஒழுங்கமைக்கும் மனித முயற்சி. இரண்டாவது சிக்கல் ஆவண சேகரிப்புக்கு கைமுறையாக உருவாக்கப்பட்ட சொற்களஞ்சியத்தின் பொருத்தம் அல்லது பாதுகாப்பு. ஆரம்ப உருவாக்கத்திற்குப் பிறகு, ஆவண சேகரிப்பு மற்றும் பின்னர் சொற்களஞ்சியம் புதுப்பிக்கப்படும் போதெல்லாம் இதே பிரச்சினைகள் மீண்டும் தோன்றும். இந்த சிக்கல்களுக்கு விடையிறுக்கும் வகையில், மனிதமுயற்சியாலான சொற்களஞ்சியம் உருவாக்கத்தை துரிதப்படுத்துவதற்கான சாத்தியக்கூறுகள் அல்லது சொற்களஞ்சியங்களைத் தானாக உருவாக்கி புதுப்பிப்பதற்கான சாத்தியக்கூறுகளை முயற்சிக்கவேண்டும்.

6.3.7. நோக்கீட்டுப் பொருள்களின் உருவாக்கம் (அச்சடிக்கப்பட்டவை மற்றும் மின்னணு வடிவானவை

நோக்கீட்டுப் பொருள்களின் சேகரிப்பு வரையறையின் கண்ணோட்டம் நூல்கள், பெரும்பாலும் அகராதிகள், கையேடுகள் மற்றும் கலைக்களஞ்சியங்கள் போன்ற அதிகாரப்பூர்வமான தகவல்களைக் கொண்ட நோக்கீட்டு நூல்கள் மற்றும் வேறுபட்ட அறிவுகளை உள்ளடக்கிய நூல்கள் நூலகத்தின்/சேகரிப்பு அமைப்பின் பிரிவில் அழைப்பு எண்ணால் அல்லது நோக்கீட்டு எண்ணால் இணைக்கப்பட்டு எளிதில் அணுகும் அல்லது பெறும் படிக்குச் சேகரிக்கப்படவேண்டும் என்பதை வேண்டும். நோக்கிட்டுப் பொருள்களின் இருப்பிடம்

மற்றும் சுழற்சி நிலை பொதுவாக பதிவேடுகளில் பதிவு செய்யப்பட்டிருக்கும். அறிவை யாசிப்பவர்களுக்கு, முக்கியமாக மாணவர்கள் மற்றும் ஆசிரியர்களுக்கு உண்மைகள், புள்ளிவிவரங்கள் மற்றும் பின்னணி தகவல்களை வழங்குவதே நொக்கிட்டுச் சேகரிப்பின் நோக்கம். சேகரிப்பு பாடத்திட்டத்தை மட்டுமல்ல, பயனர்களின் அடிப்படை தேவைகளுக்கான எல்லாத் தகவல்களையும் இது வழங்கும்.

குறிப்பு சேகரிப்புக் கொள்கையின் குறிக்கோள்கள்

- நோக்கீட்டு சேகரிப்புக்கான வழிகாட்டுதல்களை நிறுவுதல்: முதன்மை பயனர்கள், மொழிகள், பொருள் வடிவங்கள் மற்றும் நோக்கீடுச் சேகரிப்பில் காணப்படும் பொருட்களின் வகைகள்.
- புதிய நோக்கீட்டுப் பொருள்களைப் பெறுவதற்கான நடைமுறைகளை நிறுவுதல்: அது விரிவான, புதுப்பித்த, பொருத்தமான மற்றும் வசதியான தகவல்களை வழங்கும்.
- நோக்கீட்டுப் பொருள்களைத் தேர்ந்தெடுப்பதற்கான நடைமுறைகளை நிறுவுதல்.

நோக்கீட்டுச் சேகரிப்பின் பயனர்கள் யாராகவும் இருக்கலாம். மாணவர்கள். மற்றும் ஆசிரியர்கள் ஆகியோரை இச்சேகரிப்பின் முதன்மை பயனர்களாக கருதலாம்; மேலும் கல்வியுடன் அல்லது கற்றலுடன் அல்லது அறிவு பெறுதலில் தொடர்புள்ளவர்கள் யாவரும் பயனர்களாவர்.

பொருட்களின் வடிவம்

சேகரிப்பு அமைப்பு இணையப் பயன்பாட்டை அல்லது வலையமைப்பு பயன்பாட்டை அதிகரிக்க ஆன்லைன் மின்னணு வடிவமைப்பை வலியுறுத்தும்.

நோக்கீடுப் பொருள்கள்

நோக்கீடுப் பொருள்கள் தாள்வடிவ நூல்களாகவோ, மின்வடிவ நூல்களாகவோ, குறுந்தட்டுகளாகவோ, கணினியில் வரிசைப்படுத்தப்பட்டு வகைப்படுத்தப்பட்ட தகவல் பொருள்களாகவோ இருக்கலாம்.

6.3.8. இயந்திரம் படிக்கக்கூடிய அகராதி உருவாக்கம்

இயந்திரம் படிக்கக்கூடிய அகராதி

இயந்திரம் படிக்கக்கூடிய அகராதி (Machine Redable Dictionary (MRD/எம்ஆர்டி) என்பது காகிதத்தில் அச்சிடப்படுவதற்கு பதிலாக இயந்திர (கணினி) தரவுகளாக சேமிக்கப்படும் அகராதி. இது ஒரு மின்னணு அகராதி மற்றும் சொல்சார் தரவுத்தளம் (lexical database) ஆகும். இயந்திரம் படிக்கக்கூடிய அகராதி என்பது மின்னணு வடிவத்தில் உள்ள ஒரு அகராதி, இது ஒரு

தரவுத்தளத்தில் ஏற்றப்படலாம் மற்றும் பயன்பாட்டு மென்பொருள் வழியாக வினவலாம். இது இரண்டு அல்லது அதற்கு மேற்பட்ட மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பை ஆதரிப்பதற்கான ஒற்றை மொழி விளக்க அகராதி அல்லது பல மொழி அகராதியாக இருக்கலாம் அல்லது இரண்டின் கலவையாக இருக்கலாம். பல மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பு மென்பொருள் பொதுவாக இருதரப்பு/இருதிசை அகராதிகளைப் (bidirectional dictionaries) பயன்படுத்துகிறது. ஒரு எம்.ஆர்.டி என்பது தனியுரிம கட்டமைப்பைக் கொண்ட ஒரு அகராதியாக இருக்கலாம், இது பிரத்யேக மென்பொருளால் வினவப்படுகிறது (எடுத்துக்காட்டாக இணையம் வழியாக ஆன்லைனில்) அல்லது இது ஒரு திறந்த கட்டமைப்பைக் கொண்ட ஒரு அகராதியாக இருக்கலாம் மற்றும் கணினி தரவுத்தளங்களில் ஏற்றுவதற்கு கிடைக்கிறது, இதனால் பல்வேறு மென்பொருள் வழியாகப் பயன்படுத்தலாம் பயன்பாடுகள். வழக்கமான அகராதிகளில் பல்வேறு விளக்கங்களுடன் ஒரு தலைச்சொல்லன்/லெம்மா உள்ளது. இயந்திரம் படிக்கக்கூடிய அகராதி கூடுதல் திறன்களைக் கொண்டிருக்கலாம், எனவே சில நேரங்களில் இது ஸ்மார்ட் அகராதி என்று அழைக்கப்படுகிறது. ஸ்மார்ட் அகராதியின் எடுத்துக்காட்டு திறந்த மூல கெல்லிஷ் ஆங்கில அகராதி (Open Source Gellish English dictionary).

அகராதி என்ற சொல் ஒரு மின்னணு சொற்களஞ்சியம் (electronic thesaurus) அல்லது அகராதியைக் குறிக்கப் பயன்படுகிறது, எடுத்துக்காட்டாக எழுத்துச் சரிபார்ப்புகளில் பயன்படுத்தப்படுகிறது. அகராதிகள் (அல்லது சொற்களின்) துணை வகை-சூப்பர் டைப் வரிசைக்கு அகராதிகள் அமைக்கப்பட்டிருந்தால், அது ஒரு வகைபிரித்தல் என்று அழைக்கப்படுகிறது. இது கருத்துக்களுக்கு இடையிலான பிற உறவுகளையும் கொண்டிருந்தால், அது ஒன்டாலஜி என்று அழைக்கப்படுகிறது. தேடல் முடிவுகளை மேம்படுத்த தேடுபொறிகள் ஒரு சொல்லகராதி, வகைபிரித்தல் அல்லது ஆன்டாலஜி/மூலப்பொருண்மையியல் (ontology) ஆகியவற்றைப் பயன்படுத்தலாம். சிறப்பு மின்னணு அகராதிகள் உருவ அகராதிகள் அல்லது தொடரியல் அகராதிகள்.

இயந்திரம் படிக்கக்கூடிய அகராதி என்ற சொல் பெரும்பாலும் இயற்கை மொழி ஆய்வு அகராதியுடன் (NLP dictionary) முரண்படுகிறது; அதாவது இயந்திரம் படிக்கக்கூடிய அகராதி என்பது முன்னர் காகிதத்தில் அச்சிடப்பட்ட ஒரு அகராதியின் மின்னணு வடிவமாகும். இரண்டுமே நிரல்களால் பயன்படுத்தப்பட்டாலும்; இதற்கு முரணாக, இயற்கை மொழி ஆய்வை (என்.எல்.பி.) மனதில் கொண்டு புதிதாக அகராதி கட்டப்பட்டபோது இயற்கை மொழி ஆய்வு அகராதி என்ற

சொல் விரும்பப்பட்டது. இயந்திரம் படிக்கக்கூடிய அகராதி மற்றும் இயற்கை மொழி ஆய்வுக்கான சர்வதேச தரநிர்ணய அமைப்பு (International Organization for Standardization (ISO/ஐஎஸ்ஓ) தரநிலை இரு கட்டமைப்புகளையும் பிரதிநிதித்துவப்படுத்த முடியும்; மேலும் இது சொல்சார் அடையாளப்படுத்தப்பட கட்டமைப்பு (லெக்சிகல் மார்க்அப் ஃபிரேம்வொர்க்/ Lexical Markup Framework) என்று அழைக்கப்படுகிறது.

வரலாறு

முதன்முதலில் பரவலாக விநியோகிக்கப்பட்ட எம்ஆர்டிகள் மெரியம்-வெப்ஸ்டர் ஏழாவது கல்லூரி (Merriam-Webster Seventh Collegiate) (டபிள்யூ 7) மற்றும் மெரியம்-வெப்ஸ்டர் புதிய பாக்கெட் அகராதி (erriam-Webster New Pocket Dictionary (MPD/எம்.பி.டி) ஆகும். ஜான் ஓல்னியின் (John Olney) வழிகாட்டுதலின் கீழ் கணினி மேம்பாட்டுக் கழகத்தில் (System Development Corporation) அரசாங்கத்தால் நிதியளிக்கப்பட்ட திட்டத்தால் இவை இரண்டும் தயாரிக்கப்பட்டன. இரண்டு புத்தகத்தின் தட்டச்சு நாடாக்களும் கிடைக்காததால் அவை கைமுறையாக விசைப்பலகை மூலம் உள்ளீடு செய்யப்பட்டன. முதலில் ஒவ்வொன்றும் காந்த நாடாவின் பல ரீல்களில் அட்டைப் படங்களாக ஒவ்வொரு வரையறையின் ஒவ்வொரு தனி வார்த்தையையும் தனித்தனி பஞ்ச் கார்டில் அச்சிடப்பட்ட அகராதியில் அதன் பயன்பாட்டின் விவரங்களைக் குறிக்கும் பல சிறப்பு குறியீடுகளுடன் விநியோகிக்கப்பட்டன. அகராதியில் உள்ள வரையறைகளை பகுப்பாய்வு செய்வதற்கான ஒரு பெரிய திட்டத்தை ஓல்னி கோடிட்டுக் காட்டினார், ஆனால் பகுப்பாய்வு மேற்கொள்ளப்படுவதற்கு முன்னர் அவரது திட்டம் காலாவதியானது. ஆஸ்டினில் உள்ள டெக்சாஸ் பல்கலைக்கழகத்தில் (University of Texas at Austin) ராபர்ட் ஆம்ஸ்லர் (Robert Amsler) மீண்டும் பகுப்பாய்வைத் தொடங்கினார் மற்றும் தேசிய அறிவியல் அறக்கட்டளையின் நிதியத்தின் கீழ் பாக்கெட் அகராதியின் வகைபிரித்தல் விளக்கத்தை நிறைவு செய்தார், இருப்பினும் வகைபிரித்தல் தரவு விநியோகிக்கப்படுவதற்கு முன்பே அவரது திட்டம் காலாவதியானது. ராய் பைர்ட் மற்றும் பலர் (Roy Byrd et al.) ஐ.பி.எம். யார்க்க்டவுன் ஹைட்ஸ் (IBM Yorktown Heights), ஆம்ஸ்லரின் பணியைத் தொடர்ந்து வெப்ஸ்டரின் ஏழாவது கல்லூரி (Webster's Seventh Collegiate) பற்றிய பகுப்பாய்வை மீண்டும் தொடங்கியது. இறுதியாக, 1980களில் பெல்கூரின் (Bellcore) ஆரம்ப ஆதரவில் தொடங்கி பின்னர் பல்வேறு அமெரிக்க கூட்டாட்சி நிறுவனங்களால் நிதியளிக்கப்பட்டது, இதில் என்எஸ்எஃப் NSF, ஏஆர்டிஏ ARDA, தர்பா DARPA, டிடிஓ DTO மற்றும் ரெஃப்ளெக்ஸ்

(REFLEX), பிரின்ஸ்டன் பல்கலைக்கழகத்தில் (Princeton University) ஜார்ஜ் ஆர்மிட்டேஜ் மில்லர் (George Armitage Miller) மற்றும் கிறிஸ்டியன் ஃபெல்பாம் (Christiane Fellbaum) ஆகியோர் ஒரு அகராதியை உருவாக்கி பரவலாக விநியோகித்தனர் மற்றும் இன்று மிகவும் பரவலாக விநியோகிக்கப்பட்ட கணிசார் அகராதி வளமாக உள்ள வேர்ட்நெட் திட்டத்தில் (WordNet project) அதன் வகைபாட்டியல் (taxonomy) பயன்படுத்தப்பட்டது.

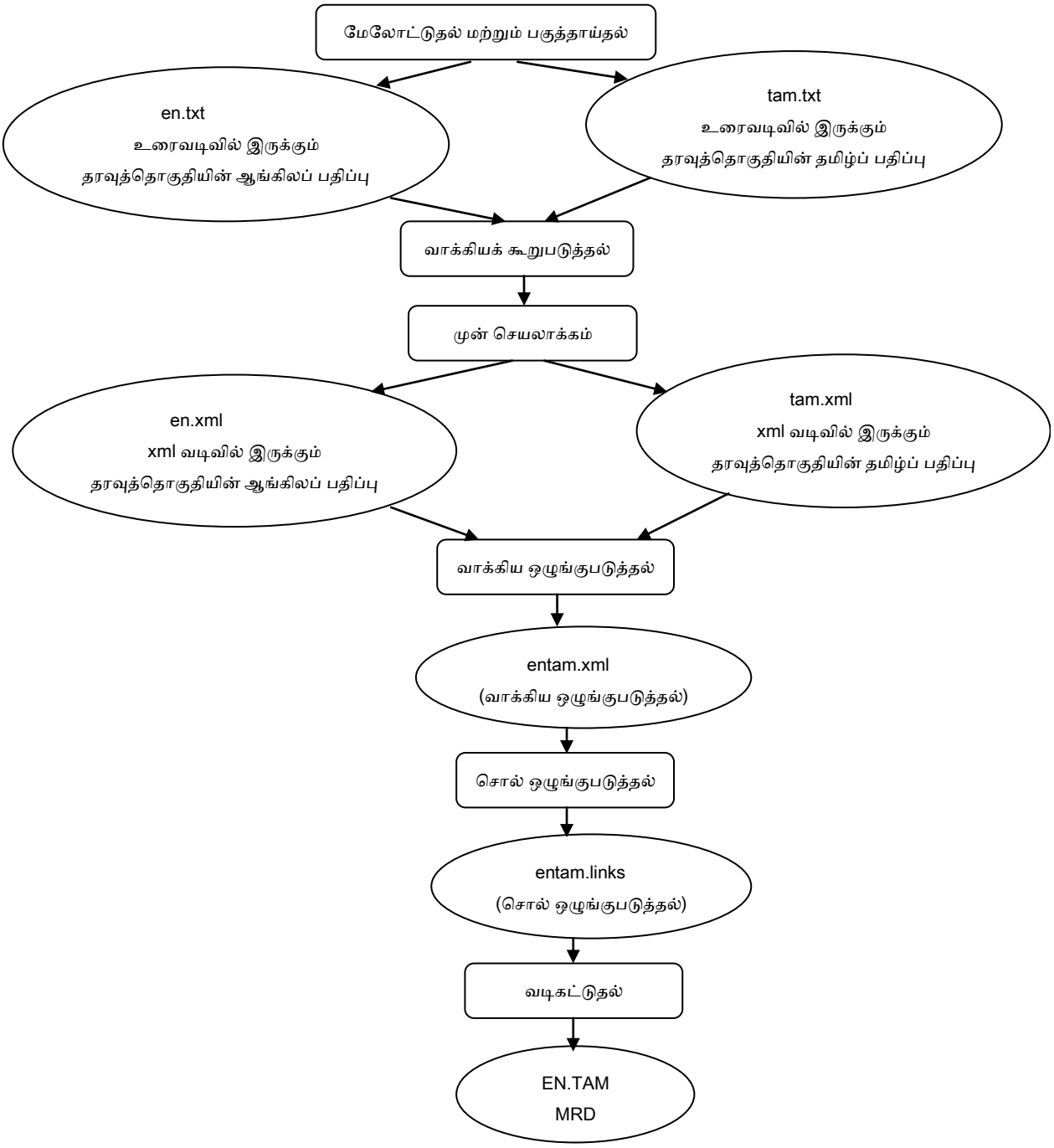
உருவாக்கம்

அன்றாட வாழ்க்கையில் நாம் பயன்படுத்தும் மிக சக்திவாய்ந்த குறிப்பு கருவிகளில் அகராதிகள் ஒன்றாகும். ஒரு மொழியைக் கற்கும் செயல்முறையிலும் அதன் அன்றாட பயன்பாட்டிலும் அவை நன்மை பயக்கும். கடந்த காலத்தில் அனைத்து அகராதிகளும் அச்சிடப்பட்ட வடிவத்தில் இருந்தன. இருப்பினும், தொழில்நுட்பத்தின் விரைவான வளர்ச்சியுடன், டிஜிட்டல் வடிவத்தில் அகராதிகளின் தேவை மிகப்பெரிய அளவில் அதிகரித்துள்ளது. தற்போதுள்ள பாரம்பரிய அகராதிகளை டிஜிட்டல் மயமாக்கும் செயல்முறை நீண்டது, சிக்கலானது, மேலும் நிறைய வளங்கள் தேவைப்படுகின்றன. மேலும், பாரம்பரிய மின்னணு அகராதிகளின் பயன்பாட்டின் சிக்கல் என்னவென்றால், சில சொற்களின் மொழிபெயர்ப்புகள் வெளிப்படையான வடிவத்தில் கொடுக்கப்படவில்லை (வார்த்தைக்கு வார்த்தை அல்லது வார்த்தைக்கு சொற்றொடர்) ஆனால் இலக்கில் உள்ள வாக்கியங்களுக்கு வார்த்தையை உள்ளடக்கிய வாக்கியங்களை நேரடியாக மொழிபெயர்ப்பது மொழி. இந்த வழக்கில், இந்த வார்த்தையின் மொழிபெயர்ப்பை தானாகக் கண்டுபிடிப்பது கடினம். இயந்திரம் படிக்கக்கூடிய அகராதிகள், மறுபுறம், ஒரு வார்த்தைக்கு (சொற்றொடர்) ஒரு சொல் (சொற்றொடர்) கொடுக்கப்பட்ட சரியான மொழிபெயர்ப்பு அல்லது வரைபடத்தைக் கொண்டுள்ளன.

இயற்கை மொழி செயலாக்க சமூகம் இணையான மற்றும் ஒப்பிடக்கூடிய தரவுத்தொகுதி வடிவத்தில் வெவ்வேறு மொழிகளில் வழங்கப்பட்ட பெரிய அளவிலான உரைகள் இருப்பதால் பெரிதும் பயனடைந்துள்ளது. இந்த வகையான உரை சேகரிப்புகள் பலவகையான பயன்பாடுகளுக்கு இருமொழி அகராதிகளைத் தானாக பிரித்தெடுக்க விரிவாக பயன்படுத்தப்படுகின்றன. இயந்திர மொழிபெயர்ப்புத் துறையில் ஆராய்ச்சியாளர்களால் இந்த ஆற்றல் மிகவும் அங்கீகரிக்கப்பட்டுள்ளது, அங்கு புள்ளிவிவர அணுகுமுறைகள் இலக்கண, விதி அடிப்படையிலான நுட்பங்களில் ஆதிக்கம் செலுத்துகின்றன. இந்த போக்கு காரணமாக

இணையான தரவுத்தொகுதிகளைச் செயலாக்குவதற்கான ஏராளமான இலவச மற்றும் திறந்த மூல கருவிகள் உருவாக்கப்பட்டுள்ளன.

தரவுத்தொகுதிகொண்டு உருவாக்கப்படும் ஒரு மாதிரி இருமொழிய இயந்திரம் படிக்கவியலும் அகராதி உருவாக்கம் இங்கு விளக்கப்பட்டுள்ளது. இதன் விளக்கப்படம் கீழே தரப்பட்டுள்ளது.



மேலோடுதல் மற்றும் பகுத்தாய்தல் (Crawling and Parsing)

வலைத்தளம் மூலவளம் என்றால், முதல் படி ஒரு எளிய மேலோட்டி (கிராலர்/crawler) மற்றும் பாகுபடுத்தி (parser) என்பவற்றை உருவாக்குவது. இரு மொழிகளின் இணையாக உள்ள

ஒவ்வொரு உரையின் இணையத்தள முகவரிகளையும் (Uniform Resource Locator (URL)) சேகரித்து, இரு மொழிகளிலும் உரையைப் பதிவிறக்குவதே கிராலரின் நோக்கம். பின்னர், HTML குறியீட்டிலிருந்து கட்டுரையின் உரையைப் பிரித்தெடுக்கவும் மற்றும் தேவையற்ற அனைத்து எழுத்துக்களையும் அகற்றவும் பாகுபடுத்தி பயன்படுத்தப்பட்டது. இறுதியாக, உரைகள் ஒவ்வொரு மொழிக்கும் ஒன்று என இரண்டு உரை கோப்புகளில் சேமிக்கப்படும்; ஒரு வரி ஒரு உரையைக் குறிக்கிறது. இந்த கோப்புகளின் உள்ளடக்கம் கைமுறையாக சரிபார்க்கப்படும்; முதல் கோப்பில் n-ஆவது வரியில் உள்ள கட்டுரை இரண்டாவது மொழியில் மொழிபெயர்க்கப்பட்ட உரைக்கு ஒத்திருக்கிறது என்பதை உறுதிசெய்கிறது. ஒரு மொழியில் காணாமல் போன உரைகள் இரண்டு கோப்புகளிலிருந்தும் அகற்றப்படும்.

வாக்கியக்கூறுபடுத்தல்

அடுத்த கட்டமாக ஒவ்வொரு உரையையும் வாக்கியங்களாகப் கூறுபடுத்த வேண்டும். வாக்கியக் கூறுபடுத்தலுக்கான பல மென்பொருகள் உள்ளன. பொருத்தமான ஒன்றைத் தெரிந்தெடுத்து இதை நிறைவேற்றலாம். அப்லக் (Uplug) என்ற மென்பொருள் வாக்கியங்களைக் கூறுபடுத்தும் தொகுதியை உள்ளடக்கியிருந்தாலும், இந்தத் தொகுதி எளிய விதிகளை நம்பியுள்ளது மற்றும் திருப்திகரமான முடிவுகளைத் தரவில்லை என்று கூறப்படுகின்றது. அதற்கு பதிலாக, பங்க்ட் கருதப்பட்டார். பங்க்ட் (Punkt) என்பது ஒரு கணினி நிரலாகும், இது வாக்கிய எல்லையைக் கண்டறிதலுக்கான மொழி-சுதந்திரமான மேற்பார்வை செய்யப்படாத வழிமுறையை செயல்படுத்துகிறது. உள்ளுணர்வாக புரிந்து கொள்ளப்பட்டால், சுருக்கங்கள் அடையாளம் காணப்பட்டவுடன் தண்டனை எல்லைகளை நிர்ணயிப்பதில் ஏராளமான தெளிவற்ற தன்மைகளை அகற்ற முடியும் என்ற அனுமானத்தின் அடிப்படையில் இது அமைந்துள்ளது. பங்க்ட் திறந்த மூலமாகும், இது பைதான் என்.எல்.டி.கே (இயற்கை மொழி கருவித்தொகுப்பு) மூலம் கிடைக்கிறது, மேலும் சேகரிக்கப்பட்ட நிறுவனத்திற்கு எளிதாகப் பயன்படுத்தலாம். வாக்கியப் பிரிவின் செயல்பாட்டை மேலும் எளிதாக்க, பத்தி HTML குறிச்சொற்களை உள்ளடக்கிய அனைத்து கட்டுரைகளும் முதலில் பத்திகளில் பிரிக்கப்பட்டன, பின்னர் வாக்கியம் பிரிக்கப்பட்டன. இந்த நடவடிக்கைக்குப் பிறகு, முழு கார்பஸிலும் சொல்லத்தக்க எண்ணிக்கையிலான உரைகள் உள்ளன, அதாவது ஒரு மொழிக்கு சொல்லத்தக்க வாக்கியங்கள் உள்ளன.

முன் செயலாக்கம் (Pre-Processing)

தரவுத்தொகுதி வாக்கியக்க கூறாக்கப்பட்டவுடன், தரவுத்தொகுதியை மற்ற அப்லக் (Uplug) தொகுதிகளுடன் மேலும் செயலாக்க அனுமதிக்க அப்லக் முன் செயலாக்க தொகுதி பயன்படுத்தப்படலாம். முன் செயலாக்க தொகுதி உரையை டோக்கனைஸ் செய்கிறது மற்றும் ஒவ்வொரு பத்தி, வாக்கியம் மற்றும் சொல்லுக்கும் அடிப்படை மார்க்அப்பைப் பயன்படுத்தி உரை கோப்புகளை எக்ஸ்எம்எல் வடிவத்தில் மாற்றுகிறது.

வாக்கிய ஒழுங்குபடுத்தல் (Sentence Alignment)

அடுத்து, வாக்கிய ஒழுங்குபடுத்தல் தொகுதி பயன்படுத்தப்பட வேண்டும். இந்த தொகுதியின் நோக்கம் ஒரு கோப்பில் உள்ள அனைத்து வாக்கியங்களையும் மற்ற மொழியில் தொடர்புடைய மொழிபெயர்ப்பு வாக்கியங்களுடன் இணைப்பதாகும். அப்லக்கில் பல வாக்கிய சீரமைப்பு தொகுதிகள் உள்ளன. ஒவ்வொன்றையும் பரிசோதித்தபின், ஹன்அலைன்ஐப் (HunAlign) பயன்படுத்தும் தொகுதி மிகவும் திருப்திகரமான முடிவுகளைக் காட்டும் என்று முடிவு செய்யலாம். ஹன்அலைன் என்பது ஒரு மொழி சுதந்திரமான தொகுதி ஆகும்; இது நீளம் சார்ந்த மற்றும் அகராதி அடிப்படையிலான அணுகுமுறைகளை இணைப்பதன் மூலம் இருமொழி உரைகளில் வாக்கியங்களை ஒருங்கிணைக்கிறது. தரவுத்தொகுதியின் முதல் பாஸில் ஹன்அலைன், வாக்கியங்களின் தோராயமான ஒழுங்குபடுத்தலுக்கும் இந்த ஒழுங்குபடுத்தல் அடிப்படையில் ஒரு அகராதியை உருவாக்கவும் வாக்கியங்களின் நீளம் குறித்த தகவலை பயன்படுத்துகிறது. இரண்டாவது பாஸில், இது தயாரிக்கப்பட்ட அகராதியைப் பயன்படுத்தி வாக்கியங்களை மாற்றியமைக்கிறது. மேலும், ஹன்அலைன் ஒன்று-முதல்-பல மற்றும் பல-முதல்-ஒன்று ஒழுங்குபடுத்தலை உள்ளடக்கியது, இது வாக்கியப் பிரிவு கட்டத்தில் செய்யப்பட்ட பிழைகளை சரியான வாக்கிய ஒழுங்குபடுத்தலுடன் சரிசெய்ய அனுமதிக்கிறது. இந்த கட்டத்தின் விளைவாக வாக்கிய இணைப்புகள் மற்றும் ஒவ்வொரு இணைப்பின் ஒழுங்குபடுத்தல் உறுதியையும் கொண்ட ஒரு எக்ஸ்எம்எல் கோப்பு உருவாகும்.

சொல் ஒழுங்குபடுத்தல் (Word Alignment)

வாக்கியங்கள் ஒழுங்குபடுத்தப்பட்டவுடன், தரவுத்தொகுதி மீது ஒழுங்குபடுத்தல் தொகுதி என்ற சொல் பயன்படுத்தப்பட்டது. சொல் ஒழுங்குபடுத்தல் என்பது ஒழுங்குபடுத்தப்பட்ட வாக்கியங்களில் தொடர்புடைய சொற்களையும் சொற்றொடர்களையும் இணைக்கும் செயல்முறையைக் குறிக்கிறது. இந்த நோக்கத்திற்காக அப்லக்கில் மூன்று வெவ்வேறு தொகுதிகள் உள்ளன: அடிப்படை, குறிச்சொல் மற்றும் மேம்பட்டவை. சிறந்த முடிவுகளை

அடைவதற்கு சம்பந்தப்பட மொழிக்கான சொல்வகைப்பாட்டு அடையாளப்படுத்தி (டேக்கர்) தேவை; இதற்குப் பதிலாக மேம்பட்ட சொல் ஒழுங்குபடுத்தல் தொகுதியைப் பயன்படுத்தினோம்.

வடிகட்டுதல் (Filtering)

தரவுத்தொகுதியைச் செயலாக்குவதற்கான முந்தைய கட்டங்களில் செய்யப்பட்ட பிழைகள் காரணமாக, சொல் ஒழுங்குபடுத்தல் தவறான மொழிபெயர்ப்புகளைக் கொண்டிருக்கும்; அவை நீக்கப்பட வேண்டும். சொல் ஒழுங்குபடுத்தல்களை வடிகட்டுவதற்கான செயல்முறை இரண்டு நிலைகளைக் கொண்டுள்ளது; ஒவ்வொரு கட்டமும் பல விதிகளை உள்ளடக்கியது. மூன்று முறைக்குக் குறைவாக நிகழ்ந்த அனைத்து ஒழுங்குபடுத்தல்களும் சொல் ஒழுங்குபடுத்தல் தொகுதி உருவாக்கிய தவறாகக் கருதப்படும்; மேலும் அவை வடிகட்டுதல் விதிகளைப் பயன்படுத்துவதற்கு முன்பு விலக்கப்பட வேண்டும்.

வடிகட்டுதல் விதிகளைப் பயன்படுத்திய பிறகு, சொல் ஒழுங்குபடுத்தல்களின் பட்டியலை குறிப்பிட்ட எண்ணிக்கையிலான மொழிபெயர்ப்பு ஜோடிகளுக்குப் பட்டியலிட வேண்டும். இது பிரித்தெடுக்கப்பட்ட அகராதியின் அளவு ஆகும்.

6.3.9. பன்மொழிய சொல்சார் வளங்களின் உருவாக்கம்

கையால் உருவாக்கப்பட்ட சொல்சார் வளங்கள் கட்டமைக்க மற்றும் பராமரிக்க விலை உயர்ந்தவை, கவரேஜில் வேறுபடுகின்றன; மேலும் பெரும்பாலும் புதிய மற்றும் களம்/டொமைன் சார்ந்த சொற்கள் இல்லை. ஆகவே, சொல்சார் கையகப்படுத்துதலில் தானியங்கி முறைகள் முக்கியம், குறிப்பாக பன்மொழி அகராதிகள் மற்றும் ஆன்டாலஜிக்களுக்கு, ஒரு மொழிக்கான வளங்களுக்குள்ளான இடைவெளிகளை ஒரு மொழிக்கான ஆதாரங்களுடன் மற்றொரு மொழிக்கான தகவல்களுடன் நிரப்ப ஒரு குறிப்பிட்ட வாய்ப்பு உள்ளது.

இந்தச் சிக்கலுக்கான பல்வேறு அணுகுமுறைகள் உள்ளன; அவற்றில் பல இணையான தரவுத்தொகுதியின் செயல்பாடுகளின் பிற அம்சங்களுடன் தொடர்புடையவை (வெரோனிஸ்/ Veronis 2000). ஒரு இணையான தரவுத்தொகுதி என்பது ஒன்றுக்கு மேற்பட்ட மொழிகளில் மொழிபெயர்க்கப்பட்ட ஆவணங்களின் தொகுப்பாகும்; மேலும் இணையான தரவுத்தொகுதிகள் (parallel corpus) என்பது மொழிபெயர்ப்புச் செயல்முறை குறித்த தகவல்களின் மிகச் சிறந்த ஆதாரங்கள். மூர் (Moore 2001) மொழிபெயர்ப்பு உறவுகளைக் கற்றுக்கொள்வதற்கான புள்ளிவிவர அணுகுமுறையை விவரிக்கிறார்; மேலும் சில பொருத்தமான ஒற்றுமை மதிப்பெண்களைப் பயன்படுத்தி, 'அதிக மதிப்பெண் பெற்ற கூட்டாளரை' ('highestscoring partner') சாத்தியமான மொழிபெயர்ப்பாகத் தேர்ந்தெடுக்கும் பொதுவான முறையைச்

சுருக்கமாகக் கூறுகிறார். இணை அல்லாத தரவுத்தொகுதிகளிலிருந்து சொற்களைப் பிரித்தெடுக்க தொடர்பு மதிப்பெண்களை ஃபங் (2000) ஏற்கிறார்.

மெலமேட்-இன் (Melamed 1996) பைடெக்ஸ்ட் மேப்பிங்ஸ் (bitext mappings) போன்ற இருமொழி உரை-ஒழுங்குபடுத்தல் முறைகளைப் (text-alignment methods) பயன்படுத்தி சிக்கல் பெரும்பாலும் அணுகப்படுகிறது; ஏனென்றால் ஒன்றுக்கொன்று நேரடியாக இணைந்திருக்கும் சொற்கள் ஒரே பொருளைக் கொண்டிருக்க வாய்ப்புள்ளது. காஸ்ஸியர் (Gaussier 1998) நெட்வொர்க் பாய்ச்சல்களின் (networkflows) அடிப்படையில் இணையான வாக்கியங்களுக்கு இடையில் சாத்தியமான ஒழுங்குபடுத்தல்களின் தொகுப்பை விவரிக்கிறார்; மேலும் இந்த பொது அணுகுமுறையானது மிகச் சிறிய தரவுத்தொகுதிகளிலிருந்து கூட சொல்சார் தகவல்களைப் பிரித்தெடுக்கப் பயன்படுகிறது என்பதை நிரூபிக்கிறது. பல ஒழுங்குபடுத்தல் முறைகள் சில நேரங்களில் விதை-அகராதி என குறிப்பிடப்படுகிற அறியப்பட்ட மொழிபெயர்ப்புகளின் மையத்தை நம்பியுள்ளன (அல்லது பயன்படுத்தி மேம்படுத்தலாம்). மெலமேட் (மெலமேட் 1996) சுட்டிக்காட்டியுள்ளபடி, சொல்சார் பிரித்தெடுப்பிற்கான ஒழுங்குபடுத்தல் முறைகளைப் பயன்படுத்துவது ஒரு வட்ட அணுகுமுறையாக இருக்கலாம்.

6.3.10. மின் அகராதி உருவாக்கம்

மின்னணு அகராதி

மின்னணு அகராதி (**Electronic dictionary** (ED) என்ற சொல் நேசியால் பின்வருமாறு வரையறுக்கப்படுகிறது: "மின்னணு அகராதி (அல்லது ED) என்ற சொல் மின்னணு வடிவத்தில் சேமிக்கப்பட்டுள்ள எந்தவொரு நோக்கீட்டுப் பொருளையும் குறிக்க பயன்படுகிறது; இது எழுத்துப்பிழை, பொருள் அல்லது சொற்களின் பயன்பாடு பற்றிய தகவல்களைத் தருகிறது. இவ்வாறு ஒரு சொல் செயலாக்க திட்டத்தில் எழுத்துப்பிழை சரிபார்ப்பு, அச்சிடப்பட்ட சொற்களை ஸ்கேன் செய்து மொழிபெயர்க்கும் சாதனம், ஆன்-லைன் கற்பிக்கும் பொருட்களுக்கான சொற்களஞ்சியம் அல்லது மதிப்பிற்குரிய கடின நகல் அகராதியின் மின்னணு பதிப்பு அனைத்தும் ஒரு வகையான மின்னணு அகராதி ஆகும்; இது சேமிப்பகம் மற்றும் மீட்டெடுக்கும் அதே அமைப்பால் பண்பாக்கம் செய்யப்படும். " (Nesi, 2000 **a**: 819)

எனவே மின்னணு அகராதிகள் அச்சிடப்பட்ட அகராதிகளிலிருந்து முதலில் தரவு சேமிக்கப்படும் முறையினாலும், இரண்டாவதாக இந்த தரவுகளை அணுகும் முறையினாலும் வேறுபடுகின்றன. கூடுதலாக, முல்லர்-ஸ்பிட்சர் மின்னணு அகராதி என்ற வார்த்தையை மனித பயனர்களுக்கு கட்டுப்படுத்துகிறார், ஏனெனில் இது அச்சிடப்பட்ட அகராதியின் அடிப்படை

பண்புகளை மின்னணு அகராதிக்கு மாற்றுவதற்கான முன் நிபந்தனையை தெரிவிக்கிறது (Antje Topel 2014 கட்டுரையிலிருந்து எடுத்தாளப்பட்டுள்ளது.) எனவே மின்னணு அகராதி என்ற சொல், நேசி ஏற்கனவே வாதிட்டபடி, பல்வேறு வகையான மின்னணு அகராதிகளுக்கான பொதுவான சொல். இந்த காரணத்திற்காக, சில கல்வியாளர்கள் மின்னணு அகராதிகளின் வகைபாட்டியலை உருவாக்க முயன்றனர். ஒரு ஆரம்ப முயற்சி வகைப்பாட்டியல் ஸ்டோரர்/ஃப்ரீஸில் காணலாம் (Antje Topel 2014 கட்டுரையிலிருந்து எடுத்தாளப்பட்டுள்ளது.). இதில், ஆசிரியர்கள் ஹவுஸ்மேன் உருவாக்கிய அச்சிடப்பட்ட அகராதிகளின் வகைப்பாட்டியல் மீது தங்கள் படைப்புகளை அடிப்படையாகக் கொண்டுள்ளனர் (Antje Topel 2014 கட்டுரையிலிருந்து எடுத்தாளப்பட்டுள்ளது). அதிலிருந்து, அவர்கள் மொழிகளின் எண்ணிக்கை மற்றும் நிபுணத்துவத்தின் நடுத்தர-சுதந்திர அளவுகோல்களைப் பயன்படுத்துகின்றனர், அதன்படி அவை ஒருமொழி, இருமொழி மற்றும் பன்மொழி அகராதிகள், அத்துடன் பொது மற்றும் சிறப்பு அகராதிகள் (பின்னர் அவை மேலும் துணைப்பிரிவு) இடையே வேறுபடுகின்றன. இவை தவிர, மின்னணு அகராதிகளின் இடைநிலை தனித்துவங்களுக்கு நியாயம் செய்வதற்காக சில நடுத்தர-குறிப்பிட்ட வகைப்பாட்டியல் அம்சங்களை (வெளியீட்டு வடிவம், தனித்துவம், ஹைபர்டெக்ஸ்டுவலைசேஷன், மல்டிமீடியாலிட்டி மற்றும் அணுகல் முறைகள்) சேர்க்கின்றன.

மின் அகராதி என்பது ஒரு அகராதி, அதன் தரவு டிஜிட்டல் வடிவத்தில் உள்ளது மற்றும் பல வேறுபட்ட ஊடகங்கள் மூலம் அணுகலாம். டேப்லெட் அல்லது டெஸ்க்டாப் கணினிகள், மொபைல் பயன்பாடுகள், வலை பயன்பாடுகள் மற்றும் மின்-வாசகர்களின் உள்ளமைக்கப்பட்ட செயல்பாடுகளில் நிறுவப்பட்ட மென்பொருள் உட்பட பல வடிவங்களில் மின்னணு அகராதிகளைக் காணலாம். அவை இலவசமாக இருக்கலாம் அல்லது கட்டணம் தேவைப்படலாம். தகவல்

ஆரம்பகால மின் அகராதிகளில் பெரும்பாலானவை டிஜிட்டல் வடிவத்தில் அச்சு அகராதிகள் கிடைத்தன: உள்ளடக்கம் ஒரே மாதிரியாக இருந்தது, ஆனால் மின்னணுப் பதிப்புகள் பயனர்களுக்கு அதிக சக்திவாய்ந்த தேடல் செயல்பாடுகளை வழங்கின. ஆனால் விரைவில் டிஜிட்டல் மீடியா வழங்கும் வாய்ப்புகள் பயன்படுத்தப்படத் தொடங்கின. இரண்டு வெளிப்படையான நன்மைகள் என்னவென்றால், இடத்தின் வரம்புகள் (மற்றும் அதன் பயன்பாட்டை மேம்படுத்த வேண்டிய அவசியம்) குறைந்த நெருக்கடியாக மாறும், எனவே கூடுதல்

உள்ளடக்கத்தை வழங்க முடியும்; ஆடியோ உச்சரிப்புகள் மற்றும் வீடியோ கிளிப்புகள் போன்ற மல்டிமீடியா உள்ளடக்கத்தை உள்ளடக்குவதற்கான வாய்ப்பு எழுகிறது.

மின் அகராதி தரவுத்தளங்கள், குறிப்பாக மென்பொருள் அகராதிகளுடன் சேர்க்கப்பட்டவை பெரும்பாலும் விரிவானவை மற்றும் அவை 500,000 வரை தலைச் சொற்கள் (headwords) மற்றும் வரையறைகள் (definitions), வினை திரிபாக்கம் அட்டவணைகள் (verb conjugation tables) மற்றும் இலக்கண குறிப்பு பிரிவு (grammar reference section) ஆகியவற்றைக் கொண்டிருக்கலாம். இருமொழி மின்னணு அகராதிகள் மற்றும் திரிபுறும்மொழிகளின் (inflected languages) ஒருமொழி அகராதிகள் பெரும்பாலும் ஒரு ஊடாடும் வினைச்சொல் இணைப்பான் i(nteractive verb conjugat)or உள்ளடக்குகின்றன, மேலும் அவை சொல் பகுதியாக்கம் (word stemming) மற்றும் சொல்லனாக்கம் (லெமடைசேஷன்/lemmatization) திறன் கொண்டவை.

மின் அகராதிகளின் வெளியீட்டாளர்கள் மற்றும் உருவாக்குபவர்கள் (டெவலப்பர்கள்/ developers) தங்கள் சொந்த அகராதியலார்களிடமிருந்து (lexicographers) சொந்த உள்ளடக்கத்தை வழங்கலாம், அச்சு வெளியீடுகளிலிருந்து உரிமம் பெற்ற தரவு அல்லது இரண்டையும், பாபிலோன் மெரியம் வெப்ஸ்டரிடமிருந்து (Merriam Webster) பிரீமியம் உள்ளடக்கத்தை வழங்குவது போலவும், கொலின்ஸ் (Collins), மாஸன் (Masson) மற்றும் சைமன் & ஸ்கஸ்டர் (Simon & Schuster), ஆகியோரிடமிருந்து கூடுதல் பிரீமியம் உள்ளடக்கத்தை வழங்கும் அல்ட்ராலிங்குவாவைப் (Ultralingua) போலவும்., மற்றும் பாராகான் மென்பொருள் டுடன் (Paragon Software Duden,) பிரிட்டானிக்கா (Britannica), ஹர்ராப் (Harrap), மெரியம்-வெப்ஸ்டர் (Merriam Webster), மற்றும் ஆக்ஸ்போர்டிலிருந்து (Oxford) அசல் உள்ளடக்கத்தை வழங்குகின்றன.

கையடக்க அகராதிகள் அல்லது PEDகள்

கையடக்க மின்னணு அகராதிகள் (Handheld dictionaries), "பாக்கெட் எலக்ட்ரானிக் அகராதிகள்" (Pocket Electronic Dictionaries (PED)) அல்லது PEDகள் என்றும் அழைக்கப்படுகின்றன; அவை மினியேச்சர் கிளாம்ஷெல் மடிக்கணினி கணினிகளைப் (miniature clamshell laptop computers) போலவே இருக்கின்றன; அவை முழு விசைப்பலகைகள் மற்றும் எல்சிடி திரைகளுடன் (Liquid Crystal Display (LCD) screens) நிறைவடைகின்றன. அவை முழுமையாக சிறியதாக இருக்க வேண்டும் என்பதால், அகராதிகள் பேட்டரியால் இயங்கும் மற்றும் நீடித்த உறை பொருள்களால் தயாரிக்கப்படுகின்றன. உலகெங்கிலும்

தயாரிக்கப்பட்டாலும், கையடக்க அகராதிகள் குறிப்பாக ஜப்பான், கொரியா, தைவான், சீனா மற்றும் அண்டை நாடுகளில் பிரபலமாக உள்ளன; அங்கு அவை இரண்டாம் மொழியாக ஆங்கிலம் கற்கும் பல பயனர்களுக்கு விருப்பமான அகராதியாகும். ஸ்ட்ரோக் ஆர்டர் அனிமேஷன்கள் stroke order animations,, குரல் வெளியீடு voice output, காஞ்சி (Kanji காஞ்சி என்பது ஜப்பானிய எழுத்து முறைமையில் பயன்படுத்தப்படும் தத்தெடுக்கப்பட்ட லோகோகிராஃபிக் சீன எழுத்துக்கள்). மற்றும் கானாவுக்கான (Kana என்பது ஜப்பானிய எழுத்து அமைப்பின் பகுதிகளை உருவாக்கும் பாடத்திட்டங்களாகும், இது ஜப்பானில் காஞ்சி என அழைக்கப்படும் லோகோகிராஃபிக் சீன எழுத்துக்களுடன் வேறுபடுகிறது.) கையெழுத்துப் புரிதல் (Handwriting recognition (HWR)), மொழி கற்றல் திட்டங்கள், ஒரு கால்குலேட்டர், பிடிஎ போன்ற அமைப்பாளர் செயல்பாடுகள், கலைக்களஞ்சியம், நேர மண்டலம் மற்றும் நாணய மாற்றிகள் மற்றும் குறுக்கெழுத்து புதிர் தீர்வுகள் ஆகியவை கை அகராதிகளின் சில அம்சங்கள். பல மொழிகளுக்கான தரவைக் கொண்ட அகராதிகளில் "ஜம்ப்" அல்லது "ஸ்கிப்-தேடல்" அம்சம் இருக்கலாம், இது பயனர்களை சொற்களைப் பார்க்கும்போது அகராதிகளுக்கு இடையில் செல்ல அனுமதிக்கிறது, மேலும் தலைகீழ் மொழிபெயர்ப்பு நடவடிக்கை முடிவுகளில் காட்டப்படும் சொற்களை மேலும் பார்க்க அனுமதிக்கிறது. பல உற்பத்தியாளர்கள் உரிமம் பெற்ற அகராதி உள்ளடக்கத்தைப் பயன்படுத்தும் கையால் அகராதிகளை உருவாக்குகிறார்கள் அவை மெரியம் வெப்ஸ்டர் அகராதி மற்றும் தெசாரஸ் போன்ற தரவுத்தளத்தைப் பயன்படுத்துகின்றன, மற்றவர்கள் தங்கள் சொந்த அகராதியாலார்களிடமிருந்து தனியுரிம தரவுத்தளத்தைப் பயன்படுத்தலாம். கூடுதல் மெமரி கார்டுகளை வாங்குவதன் மூலம் பல சாதனங்களை பல மொழிகளுக்கு விரிவாக்க முடியும். உற்பத்தியாளர்களில் ஆல்ஃபாலிங்க், அட்ரீ, பெஸ்டா, கேசியோ, கேனான், இன்ஸ்டன்ட் டிக்ட், எக்டாகோ, பிராங்க்ளின், ஜரிவர், லிங்கோ, மாலியாங் சைபர் டெக்னாலஜி, காம்பாக்னியா லிங்குவா லிமிடெட், நூரியன், சீகோ மற்றும் ஷார்ப் ஆகியவை அடங்கும்.

அகராதியியல் செயல்பாட்டில் தரவுத்தொகுதியைப் பயன்படுத்துதல்

தரவுத்தொகுதி அகராதியலர்களுக்கு அளிக்கும் நன்மைகள் சுயமாகத் தெரியும். தரவுத்தொகுதி மேற்கோள்கள் கொடுக்கப்பட்ட சொல்லனின் பொருளை (அல்லது பல அர்த்தங்களை) புனரமைக்கவும், அதன் எழுத்துப்பிழை, ஒலிப்பு அல்லது உருவ மாறுபாடுகளைக் கண்டறியவும், அதன் தொடரியல் மற்றும் சொல்சார் சேர்ந்துவருகைகளை நிறுவவும்

அனுமதிக்கின்றன. தரவுத்தொகுதியில் உள்ள நூல்கள் பற்றிய தகவல்கள், சொல்லனின் பயன்பாடு, அதன் அதிர்வெண், காலவரிசை, புவியியல், முதலியனவும், சொல்லன் பதிவுசெய்யப்பட்ட வகைகள் மற்றும் பெரும்பாலும் அதைப் பயன்படுத்தும் ஆசிரியர்கள் உள்ளிட்ட கூடுதல் தரவை வழங்குகிறது. கார்பஸிலிருந்து மேற்கோள்களும் சொல் அர்த்தங்கள், மரபுத்தொடர்கள் அல்லது சேர்ந்துவருகை ஆகியவற்றின் விளக்கமாகச் செயல்படுகிறது.

மின் அகராதி ஆய்வு (எலக்ட்ரானிக் டிக்ஷனரி ரிசர்ச்/Electronic Dictionary Research (EDR/ஈ.டி.ஆர்) திட்டம்

ஆங்கிலம் மற்றும் ஜப்பானியர்களுக்கிடையேயான மொழிபெயர்ப்பு, இயற்கை மொழி புரிதல் மற்றும் தலைமுறை, பேச்சு செயலாக்கம் போன்ற பல்வேறு இயற்கை மொழி பணிகளுக்கு ஆதரவாகப் பயன்படுத்தக்கூடிய பெரிய அளவிலான, நடைமுறை மின்னணு அகராதி முறையை உருவாக்க 1986ஆம் ஆண்டில் ஐடிஓடியிலிருந்து ஒன்பது ஆண்டு சாசனத்துடன் ஈடிஆர் வெளியேற்றப்பட்டது.

எலக்ட்ரானிக் அகராதியின் ஈ.டி.ஆர் கருத்தாக்கம் இப்போது பொதுவான வழக்கமான ஆன்லைன் அகராதிகளிலிருந்து முற்றிலும் வேறுபட்டது. இந்த பிந்தைய அமைப்புகள் மனிதர்களால் பயன்படுத்த ஆன்லைன் குறிப்பு புத்தகங்களாக வடிவமைக்கப்பட்டுள்ளன. பொதுவாக, அவை காந்த வட்டு அல்லது சிடி-ரோம் இல் சேமிக்கப்பட்டுள்ள வழக்கமான அச்சிடப்பட்ட அகராதியின் சரியான உரை உள்ளடக்கங்களைக் கொண்டுள்ளன. பயன்பாடு ஒரு வழக்கமான அகராதிக்கு ஒத்ததாக இருக்கிறது, தவிர மின்னணு பதிப்பில் மிக விரைவான மற்றும் வசதியான உலாவலுக்கான ஹைபர்டெக்ஸ்ட் போன்ற அம்சங்கள் இருக்கலாம்.

மின் அகராதி ஆய்வின் முதன்மை குறிக்கோள், ஒரு பகுதியாகப் பிடிக்க வேண்டும்: "இயற்கையான மொழியைப் பற்றிய முழுமையான புரிதலுக்கு கணினிக்குத் தேவையான அனைத்து தகவல்களும்" (ஜெட்ரி, 1990). இதில் பின்வருவன அடங்கும்: சொற்களின் அர்த்தங்கள் (கருத்துகள்); கருத்துக்களைப் புரிந்துகொள்ள கணினிக்குத் தேவையான அறிவு; மற்றும் உருபனியல் பகுப்பாய்வு மற்றும் உருவாக்கம், தொடரியல் பகுப்பாய்வு மற்றும் உருவாக்கம் மற்றும் பொருண்மையியல் ஆய்வு ஆகியவற்றை ஆதரிக்க தேவையான தகவல்கள். கூடுதலாக, சொல் இணை நிகழ்வுகள் (word co-occurrences) மற்றும் பிற மொழிகளில் சமமான சொற்களின் பட்டியல்கள் பற்றிய தகவல்கள் அவசியம். சுருக்கமாக, இந்த அமைப்பு சொற்கள் மற்றும்

அவற்றின் அர்த்தங்களைப் பற்றிய மிகப் பெரிய ஆனால் ஆழமற்ற அறிவுத் தளமாக இருக்க வேண்டும்.

மின் அகராதி ஆய்வு ஆராய்ச்சியின் குறிக்கோள்கள் மொழியியல் அறிவின் பரந்த அளவிலான மின்னணு அகராதிகளின் தொகுப்பை உருவாக்குவதாகும். இந்த தகவல் எந்தவொரு குறிப்பிட்ட இயற்கை மொழி செயலாக்கக் கோட்பாடு அல்லது பயன்பாட்டிற்கும் பக்கச்சார்பாக இருக்கக்கூடாது என்பதில் நடுநிலையாக இருக்க வேண்டும்; பொதுவான பொது நோக்கத்திற்கான சொற்கள் மற்றும் தொழில்நுட்ப இலக்கியத்தின் ஒரு தரவுத்தொகுதியிலிருந்து வரும் சொற்களைப் பற்றிய விரிவான தகவல்கள்; மொழியியல் செயலாக்கத்தின் அனைத்து நிலைகளையும் ஆதரிக்கும் திறனை விரிவாக்குவது, அதாவது உருபனியல், தொடரியல் மற்றும் பொருண்மையியல் செயலாக்கம், இயற்கையான சொற்களைத் தேர்ந்தெடுப்பது மற்றும் பிற மொழிகளில் சமமான சொற்களைத் தேர்ந்தெடுப்பது.

6.4. மொழித் தொழில் நுட்பக் கருவிகளைத் வடிவமைக்கத் தரவுத்தொகுதியின் பயன்பாடு (Corpus for designing language technology tools)

மொழித் தொழில் நுட்பக் கருவிகளைத் வடிவமைக்கத் தரவுத்தொகுதியின் பயன்பாடு குறிப்பிடத்தகுந்ததாகும். மொழித் தொழில்நுட்பக் கருவிகளைச் செம்மையாக உருவாகத் தரவுத்தொகுதி இன்றியமையாததாகும். இலக்கண அல்லது மொழியியல் விதிகளைப் பயன்படுத்தி சிறந்த மொழித்தொழில்நுட்பக் கருவிகளை உருவாக்கக் காலவிரையம் ஏற்படும். பல கருவிகளைத் தரவுத்தொகுதியைப் பயன்படுத்தி இயந்திரம் கற்றல் முறையில் எளிதாக உருவாக்க இயலும். மொழித் தொழில்நுட்பக் கருவிகள் பின்வருவனவற்றை உள்ளடக்கும்: சொல்லாய்வு ஒழுங்குமுறை (Word processing system), எழுத்துப்பிழைத்திருத்தும் ஒழுங்குமுறை (Spell checking system), உரை நேர்செய்யும் ஒழுங்குமுறை (Text editing system), உருபனியல் பகுப்பாய்வு ஒழுங்குமுறை (Morphological processing system), வாக்கியப் பகுப்பாய்வு ஒழுங்குமுறை (Sentence parsing system), நிகழ்வெண் கணக்கிடும் ஒழுங்குமுறை (Frequency counting system), சொல் தேடும் இயந்திரம் (Item-search engine), உரை சுருக்கும் ஒழுங்குமுறை (Text summarisation system), உரை அடையாளப்படுத்தும் ஒழுங்குமுறை (Text annotation system), தகவல் மீட்கும் ஒழுங்குமுறை (Information retrieval system), தொடரடைவு ஒழுங்குமுறை (Concordance system), சொற்பொருள் மயக்கம் நீக்கும் ஒழுங்குமுறை Word (Sense Disambiguation system), சொல் வலை திட்டமிடல் (WordNet design design),

பொருண்மையியல்சார் வலைப்பின்னல் அல்லது பொருண்யியல்சார் வலை (Semantic Web or Semantic Net), வகைப்பாடு அடையாளப்படுத்தும் ஒழுங்குமுறை (Parts-of-Speech Tagging system), ஓரிடம்சார் சொற்களை குழுமும் ஒழுங்குமுறை (Local Word Grouping system).

6.4.1. சொல்லாய்வு ஒழுங்குமுறை (Word processing system)

ஒரு சொல்செயலி (Word processer (WP) என்பது ஒரு சாதனம் அல்லது கணினி நிரலாகும்; இது உரையின் உள்ளீடு, திருத்துதல், வடிவமைத்தல் மற்றும் வெளியீட்டை வழங்குகிறது, பெரும்பாலும் சில கூடுதல் அம்சங்களுடன். ஆரம்பகாலச் சொல் செயலிகள் செயல்பாட்டிற்கு அர்ப்பணிக்கப்பட்ட தனித்த சாதனங்களாக இருந்தன; ஆனால் தற்போதைய சொல்செயலிகள் பொது நோக்க கணினிகளில் இயங்கும் சொல் செயலி நிரல்களாகும்.

ஒரு சொல்செயலி நிரலின் செயல்பாடுகள் ஒரு எளிய உரை எடிட்டருக்கும் முழுமையாக செயல்படும் டெஸ்க்டாப் வெளியீட்டு நிரலுக்கும் இடையில் எங்காவது விழும். இருப்பினும் இந்த மூன்றிற்கும் இடையிலான வேறுபாடுகள் காலப்போக்கில் மாறிவிட்டன, அவை 2010 க்குப் பிறகு தெளிவாக இல்லை.

பின்னணி

சொல்செயலி (Word Processor (WP) கணினி தொழில்நுட்பத்திலிருந்து உருவாக்கப்படவில்லை. மாறாக, அவை பொறியியல்சார் இயந்திரங்களிலிருந்து (mechanical machines) உருவாகின; பின்னர் அவை கணினித் துறையுடன் ஒன்றிணைந்தன. சொல் செயலாக்கத்தின் வரலாறு என்பது எழுதுதல் மற்றும் திருத்துவதற்கான இயற்பியல் அம்சங்களின் படிப்படியான தன்னியக்கவாக்கத்தின் கதை; பின்னர் நிறுவனங்களுக்கும் தனிநபர்களுக்கும் கிடைக்கும்படி தொழில்நுட்பத்தின் மெருகேற்றம் செய்யப்பட்டது.

சொல்செயலாக்கம்/சொல்லாய்வு (word processing) என்ற சொல் 1970களின் முற்பகுதியில் அமெரிக்க அலுவலகங்களில் தட்டச்சு செய்பவர்களுக்கு பணியை நெறிப்படுத்தும் யோசனையை மையமாகக் கொண்டிருந்தது, ஆனால் இதன் பொருள் விரைவில் முழு எடிட்டிங்/நேர்செய்தல் சுழற்சியின் தன்னியக்கத்தை நோக்கி நகர்ந்தது.

முதலில், சொல்செயலாக்க/சொல்லாய்வு அமைப்புகளின் வடிவமைப்பாளர்கள் தற்போதுள்ள தொழில்நுட்பங்களை வளர்ந்து வரும் தொழில்நுட்பங்களுடன் ஒன்றிணைத்து தனித்தனி சாதனங்களை உருவாக்கி, தனிப்பட்ட கணினியின் வளர்ந்து வரும் உலகத்திலிருந்து வேறுபட்ட ஒரு புதிய வணிகத்தை உருவாக்கினர். சொல் செயலாக்கத்தின் கருத்து மிகவும்

பொதுவான தரவு செயலாக்கத்திலிருந்து எழுந்தது, இது 1950களில் இருந்து வணிக நிர்வாகத்திற்கு கணினிகளைப் பயன்படுத்துவதாக இருந்தது. வரலாறு மூலம், 3 வகையான சொல் செயலிகள் உள்ளன: இயந்திர, மின்னணு மற்றும் மென்பொருள்.

இயந்திரச் சொல் செயலாக்கம் (Mechanical word processing)

முதல் சொல் செயலாக்கச் சாதனம் (ஒரு தட்டச்சுப்பொறிக்கு ஒத்ததாகத் தோன்றும் கடிதங்களை படியெடுப்பதற்கான இயந்திரம்) ஒரு இயந்திரத்திற்கு ஹென்றி மில்லால் (Henry Mil)l காப்புரிமை பெற்றது; இது மிகவும் தெளிவாகவும் துல்லியமாகவும் எழுதும் திறன் கொண்டது, அதை அச்சகத்திலிருந்து வேறுபடுத்திப் பார்க்க முடியவில்லை.

ஒரு நூற்றாண்டுக்கு மேலாகியும், அச்சுக்கலைஞருக்கு வில்லியம் ஆஸ்டின் பர்ட் (William Austin Burt) பெயரில் மற்றொரு காப்புரிமை தோன்றியது. 19ஆம் நூற்றாண்டின் பிற்பகுதியில், கிறிஸ்டோபர் லாதம் ஷோல்ஸ் (Christopher Latham Sholes) அடையாளம் காணக்கூடிய முதல் தட்டச்சுப்பொறியை உருவாக்கினார், இது ஒரு பெரிய அளவு என்றாலும், அது "இலக்கிய பியானோ" (literary piano) என்று விவரிக்கப்பட்டது.

இந்த இயந்திர அமைப்புகள் வகையின் நிலையை மாற்றுவதற்கும், வெற்று இடங்களை மீண்டும் நிரப்புவதற்கும் அல்லது ஜம்ப் கோடுகளை மீறுவதற்கும் அப்பால் “உரையைச் செயலாக்க” (process text) முடியவில்லை. பல தசாப்தங்கள் கழித்து மின்சாரம் மற்றும் பின்னர் மின்னணுவியல் தட்டச்சுப்பொறிகளில் அறிமுகப்படுத்தப்பட்டது இயந்திர பகுதியுடன் எழுத்தாளர். "சொல் செயலாக்கம்" என்ற சொல் 1950களில் ஜெர்மன் ஐபிஎம் தட்டச்சுப்பொறி விற்பனை நிர்வாகியான உல்ரிச் ஸ்டெய்ன்ஹில்பரால் (Ulrich Steinhilper) உருவாக்கப்பட்டது. இருப்பினும், இது 1960களின் அலுவலக மேலாண்மை அல்லது கணினி இலக்கியங்களில் தோன்றவில்லை, இருப்பினும் இது பின்னர் பயன்படுத்தப்படும் பல யோசனைகள், தயாரிப்புகள் மற்றும் தொழில்நுட்பங்கள் ஏற்கனவே நன்கு அறியப்பட்டவை. ஆனால் 1971 வாக்கில் இந்த வார்த்தையை நியூயார்க் டைம்ஸ் ஒரு வணிக "சலசலப்பு வார்த்தையாக" அங்கீகரித்தது. சொல் செயலாக்கம் மிகவும் பொதுவான "தரவு செயலாக்கம்" அல்லது வணிக நிர்வாகத்திற்கு கணினிகளைப் பயன்படுத்துவதற்கு இணையாக உள்ளது.

1972ஆம் ஆண்டளவில், வணிக அலுவலக மேலாண்மை மற்றும் தொழில்நுட்பத்திற்காக அர்ப்பணிக்கப்பட்ட வெளியீடுகளில் சொல் செயலாக்கம் பற்றிய விவாதம் பொதுவானது, 1970

களின் நடுப்பகுதியில் இந்த சொல் வணிக கால இடைவெளிகளைக் கலந்தாலோசிக்கும் எந்த அலுவலக மேலாளருக்கும் தெரிந்திருக்கும்.

எலக்ட்ரோ மெக்கானிக்கல் மற்றும் எலக்ட்ரானிக் சொல் செயலாக்கம்

1960களின் பிற்பகுதியில், ஐபிஎம் ஐபிஎம் எம்டி/எஸ்டி (காந்த நாடா/ தேர்ந்தெடுக்கப்பட்ட தட்டச்சுப்பொறி)-ஐ உருவாக்கியது. இது இந்த தசாப்தத்தின் முற்பகுதியிலிருந்து ஐபிஎம் செலக்ட்ரிக் தட்டச்சுப்பொறியின் மாதிரியாக இருந்தது, ஆனால் அதன் சொந்த மேசையில் கட்டப்பட்டது, மேலும் காந்த நாடா பதிவு மற்றும் பின்னணி வசதிகளுடன், கட்டுப்பாடுகள் மற்றும் மின் ரிலேக்களின் வங்கியுடன் ஒருங்கிணைக்கப்பட்டது. எம்டி/எஸ்.டி தானியங்கி சொல் மடக்கு, ஆனால் அதற்கு திரை இல்லை. இந்த சாதனம் மற்றொரு டேப்பில் எழுதப்பட்ட உரையை மீண்டும் எழுத அனுமதித்தது, மேலும் நாம் ஒத்துழைக்க முடியும் (டேப்பைத் திருத்த அல்லது நகலெடுக்க மற்றொரு நபருக்கு அனுப்பவும்). சொல் செயலாக்கத் தொழிலுக்கு இது ஒரு புரட்சி. 1969 ஆம் ஆண்டில் நாடாக்கள் காந்த அட்டைகளால் மாற்றப்பட்டன. இந்த மெமரி கார்டுகள் எம்டி / எஸ்.டி உடன் வந்த கூடுதல் சாதனத்தின் பக்கத்தில் அறிமுகப்படுத்தப்பட்டன,

1970களின் முற்பகுதியில், சொல் செயலாக்கம் பின்னர் பல கண்டுபிடிப்புகளின் வளர்ச்சியுடன் கணினி அடிப்படையிலானது (ஒற்றை-நோக்க வன்பொருளுடன் மட்டுமே என்றாலும்). தனிநபர் கணினி (personal computer (PC/பிசி)) வருவதற்கு சற்று முன்பு, ஐபிஎம் நெகிழ் வட்டை உருவாக்கியது. 1970களின் முற்பகுதியில் சிஆர்டி (Cathode-ray tube) காட்சி திரை எடிட்டிங் கொண்ட சொல் செயலாக்க அமைப்புகள் வடிவமைக்கப்பட்டன. இது வேலையைப் படிக்கவும் பதிவு செய்யவும் முடிந்தது.

"வேர்ட் செயலி" என்ற சொற்றொடர் வாங்கை (Wang) ஒத்த சிஆர்டி அடிப்படையிலான இயந்திரங்களைக் (CRT-based machines) குறிக்க விரைவாக வந்தது. இந்த வகையான பல இயந்திரங்கள் வெளிவந்தன, பொதுவாக ஐபிஎம், லானியர் (ஏஇஎஸ் டேட்டா மெஷின்கள் - மறு பேட்ஜ்ட் AES Data machines - re-badged), சிபிடி மற்றும் என்.பி.ஐ போன்ற பாரம்பரிய அலுவலக உபகரண நிறுவனங்களால் விற்பனை செய்யப்படுகின்றன.

சொல் செயலாக்க மென்பொருள்

சொல் செயலாக்கத்தின் இறுதி கட்டம் 1970களின் பிற்பகுதியிலும் 1980களின் பிற்பகுதியிலும் தனிநபர் கணினியின் வருகையுடனும், பின்னர் சொல் செயலாக்க மென்பொருளை உருவாக்கியது. மிகவும் சிக்கலான மற்றும் திறமையான உரையை உருவாக்கும்

சொல் செயலாக்க அமைப்புகள் உருவாக்கப்பட்டு விலைகள் வீழ்ச்சியடையத் தொடங்கின; அவை பொதுமக்களுக்கு மேலும் அணுகக்கூடியதாக அமைந்தன.

தனிப்பட்ட கணினிகளுக்கான (மைக்ரோ கம்ப்யூட்டர்கள்/microcomputers) முதல் சொல் செயலாக்க திட்டம் 1976 டிசம்பரில் விற்பனைக்கு வந்த மைக்கேல் ஷ்ரேயர் மென்பொருளிலிருந்து (Michael Shrayer Software) எலக்ட்ரிக் பென்சில் ஆகும் (Electric Pencil). 1978ஆம் ஆண்டில் வேர்ட்ஸ்டார் (WordStar) தோன்றியது மற்றும் அதன் பல புதிய அம்சங்கள் விரைவில் சந்தையில் ஆதிக்கம் செலுத்தியது. இருப்பினும், வேர்ட்ஸ்டார் ஆரம்பகால சிபி/எம் (கண்ட்ரோல் புரோகிராம்-மைக்ரோ) இயக்க முறைமைக்காக எழுதப்பட்டது, மேலும் இது புதிய எம்எஸ்-டாஸ்-க்கு (மைக்ரோசாஃப்ட் டிஸ்க் ஆப்பரேட்டிங் சிஸ்டம்) மீண்டும் எழுதப்பட்ட நேரத்தில், அது வழக்கற்றுப் போய்விட்டது. XyWrite போன்ற குறைவான வெற்றிகரமான நிரல்கள் இருந்தபோதிலும், MS-DOS சகாப்தத்தில் வேர்ட் பெர்பெக்ட் மற்றும் அதன் போட்டியாளர் மைக்ரோசாப்ட் வேர்ட் இதை முக்கிய சொல் செயலாக்க நிரல்களாக மாற்றின.

1990களில் விண்டோஸ் இயக்க முறைமையின் வளர்ந்து வரும் புகழ் பின்னர் மைக்ரோசாப்ட் வேர்டை எடுத்துக்கொண்டது. முதலில் "மைக்ரோசாஃப்ட் மல்டி-டூல் வேர்ட்" (Microsoft Multi-Tool Word) என்று அழைக்கப்பட்ட இந்த நிரல் விரைவில் "சொல் செயலி" (“word processor”) என்பதற்கு ஒருபொருள்பன்மொழியாக மாறியது.

சொல் செயலி உருவாக்கத்தில் தரவுத்தொகுதியின் பயன்பாடு

சொல் செயலி எழுத்துப்பிழை திருத்தி, இலக்கணத்திருத்தி, சொல்தேர்வான், ஒருபொருள் பன்மொழிய அகராதி போன்ற மொழிக் கருவிகளை உட்படுத்தி வருகின்றது. மேலும் சொல் பகுப்பாய்வு, புணர்ச்சி பிரித்தல், சொல்லனாக்கம் போன்ற செயல்பாடுகளையும் செய்வதற்கு உட்படுதப்பட்டுள்ளது. இக்கருவிகளை உருவாக்கதற்கு தரவுத்தொகுதி பொரும் பயனாற்றுகின்றது. இது குறித்து வேறு இடங்களில் பேசப்படும்.

6.4.2. எழுத்துப்பிழைத்திருத்தும் ஒழுங்குமுறை (Spell checking system)

மென்பொருளில், எழுத்துப்பிழை சரிபார்ப்பு (அல்லது எழுத்துப்பிழை சரிபார்ப்பு) என்பது ஒரு உரையில் எழுத்துப்பிழைகளை சரிபார்க்கும் மென்பொருள் அம்சமாகும். சொல் சரிபார்ப்பு அம்சங்கள் பெரும்பாலும் சொல் செயலி, மின்னஞ்சல் கிளையண்ட், மின்னணு அகராதி அல்லது தேடுபொறி போன்ற மென்பொருள் அல்லது சேவைகளில் உட்பொதிக்கப்படுகின்றன.

வடிவமைப்பு

ஒரு அடிப்படை எழுத்துப்பிழை சரிபார்ப்பு பின்வரும் செயல்முறைகளைச் செய்கிறது:

- இது உரையை ஸ்கேன் செய்து அதில் உள்ள சொற்களைப் பிரித்தெடுக்கிறது.
- இது ஒவ்வொரு வார்த்தையையும் சரியாக உச்சரிக்கப்பட்ட சொற்களின் பட்டியலுடன் ஒப்பிடுகிறது (அதாவது ஒரு அகராதி). இது சொற்களின் பட்டியலைக் கொண்டிருக்கலாம் அல்லது ஹைபனேஷன் புள்ளிகள் அல்லது சொல்சார் மற்றும் இலக்கண பண்புக்கூறுகள் போன்ற கூடுதல் தகவல்களையும் கொண்டிருக்கலாம்.
- கூடுதல் நடைமுறை, உருபனியல் அமைப்பைக் கையாளுவதற்கான மொழி சார்ந்த வழிமுறையாகும். ஆங்கிலம் போன்ற லேசாக திரிபுறும் மொழிக்குக் கூட, எழுத்துப்பிழை சரிபார்ப்பி (spell-checker) பன்மை, வினை வடிவங்கள், சுருக்கங்கள் மற்றும் உடைமைகள் போன்ற ஒரே வார்த்தையின் வெவ்வேறு வடிவங்களைக் கருத்தில் கொள்ள வேண்டும். ஒட்டுநிலை மற்றும் மிகவும் சிக்கலான பெயர்திரிபாக்கம் மற்றும் வினைத்திரிபாக்கம் போன்ற பண்புக்கூறுகளைக் கொண்ட மொழிகளுக்கு, செயல்பாட்டின் இந்தப் பகுதி மிகவும் சிக்கலானது.

ஜெர்மன், ஹங்கேரிய அல்லது துருக்கியம் போன்ற அதிக கூட்டிணைப்பு மொழிகளுக்கான நன்மைகள் தெளிவாக இருந்தாலும், உருபனியல் பகுப்பாய்வு, ஒரு வார்த்தையின் இலக்கணப் பங்களிப்பைப் பொறுத்து பல வடிவங்களை அனுமதிக்கும் ஆங்கிலத்திற்கு ஒரு குறிப்பிடத்தக்க நன்மையை அளிக்கிறதா என்பது தெளிவாகத் தெரியவில்லை.

இந்த கூறுகளுக்கு இணைப்பாக, நிரலின் பயனர் இடைமுகம் பயனர்களை மாற்றுவதை அங்கீகரிக்கவோ நிராகரிக்கவோ நிரலின் செயல்பாட்டை மாற்றவோ அனுமதிக்கும்.

சரியாக உச்சரிக்கப்படும் சொற்களுக்குப் பதிலாகப் பிழைகளை அடையாளம் காண மாற்று வகை எழுத்துப்பிழை சரிபார்ப்பு n-கிராம் போன்ற புள்ளிவிவர தகவல்களை மட்டுமே பயன்படுத்துகிறது. இந்த அணுகுமுறைக்குப் பொதுவாகப் போதுமான புள்ளிவிவர தகவல்களைப் பெற நிறைய முயற்சி தேவைப்படுகிறது. முக்கிய நன்மைகள் குறைந்த இயக்கநேர சேமிப்பிடம் தேவை மற்றும் அகராதியில் சேர்க்கப்படாத சொற்களில் பிழைகளைச் சரிசெய்யும் திறன் ஆகியவை அடங்கும்.

சில சந்தர்ப்பங்களில் எழுத்துப்பிழை சரிபார்ப்பிகள் அந்த எழுத்துப்பிழைகள் மற்றும் பரிந்துரைகளின் நிலையான பட்டியலைப் பயன்படுத்துகின்றன; இந்த குறைந்த நெகிழ்வான

அணுகுமுறை பெரும்பாலும் காகித அடிப்படையிலான திருத்தம் முறைகளில் பயன்படுத்தப்படுகிறது, அதாவது கலைக்களஞ்சியங்களின் உள்ளீடுகளையும் காண்க.

ஒலிப்புத் தகவலுடன் இணைந்து எழுத்துப்பிழை சரிபார்ப்புக்கும் கிளஸ்டரிங் வழிமுறைகள் பயன்படுத்தப்பட்டுள்ளன.

செயல்பாடு

தொடக்ககால எழுத்துத் திருத்திகள் "திருத்துபவைகளுக்கு" ("correctors") பதிலாக "சரிபார்ப்பிகள்" ("verifiers") ஆக இருந்தன. தவறாக எழுதப்பட்ட சொற்களுக்கு அவைகள் எந்த ஆலோசனைகளையும் வழங்கவில்லை. தட்டச்சுப் பிழைகளுக்கு இது உதவியாக இருந்தது, ஆனால் இது தருக்க அல்லது ஒலியியல்சார் பிழைகளுக்கு அவ்வளவு உதவியாக இல்லை. டெவலப்பர்கள்/உருவாக்குநர்கள் எதிர்கொண்ட சவால் தவறாக எழுதப்பட்ட சொற்களுக்கு பயனுள்ள பரிந்துரைகளை வழங்குவதில் உள்ள சிரமம் ஆகும். இதற்குச் சட்டக வடிவத்திற்கு சொற்களைக் குறைத்தல் மற்றும் அமைப்பொழுங்குக்குப் பொருந்தும் வழிமுறைகளைப் பயன்படுத்துதல் இவற்றை வேண்டும்.

எழுத்துப்பிழைச் சரிபார்க்கும் அகராதிகளைப் பொருத்தவரை, "பெரியது, சிறந்தது" என்று தர்க்கரீதியாகத் தோன்றலாம், இதனால் சரியான சொற்கள் தவறானவை எனக் குறிக்கப்படவில்லை. இருப்பினும், நடைமுறையில், ஆங்கிலத்திற்கான உகந்த அளவு 90,000 உள்ளீடுகளாகத் தோன்றுகிறது. இதை விட அதிகமாக இருந்தால், தவறாக எழுதப்படும் சொற்கள் மற்றவைகளை தவறாக நினைப்பதால் தவிர்க்கப்படலாம். எடுத்துக்காட்டாக, தரவுத்தொகுதி மொழியியலின் அடிப்படையில் ஒரு மொழியியலாளர் baht என்ற சொல் தாய் நாணயத்தைக் குறிப்பதைக் காட்டிலும் bath அல்லது bat மட்டையை தவறாக எழுதியுள்ளதாகத் தீர்மானிக்கலாம். எனவே, தாய் நாணயத்தைப் பற்றி எழுதும் ஒரு சிலர் சற்று சிரமத்திற்கு உள்ளானாலும் bath பற்றி விவாதிக்கும் இன்னும் பலரின் எழுத்து பிழைகள் கவனிக்கப்படாமல் இருப்பதை விட அது மிகவும் பயனுள்ளதாக இருக்கும்.

தொடக்கத்தில் MS-DOS எழுத்துப்பிழை சரிபார்ப்புகள் பெரும்பாலும் சொல் செயலாக்கத் தொகுப்புகளிலிருந்து (word processing packages) சரிபார்ப்பு பயன்முறையில் பயன்படுத்தப்பட்டன. ஒரு ஆவணத்தைத் தயாரித்த பிறகு, ஒரு பயனர் எழுத்துப்பிழைகளைத் தேடும் உரையை ஸ்கேன் செய்தனர். இருப்பினும், பின்னர், ஆரக்கிளின் குறுகிய கால CoAuthor போன்ற தொகுப்புகளில் தொகுதி செயலாக்கம் (batch processing) வழங்கப்பட்டது மற்றும் ஒரு

ஆவணம் செயலாக்கப்பட்ட பின்னர் முடிவுகளைப் பார்க்கவும் பயனர்கள் தவறாக அறியப்பட்ட சொற்களை மட்டுமே சரிசெய்யவும் அனுமதித்தது. நினைவகம் மற்றும் செயலாக்கச் சக்தி ஏராளமாக மாறியபோது, 1987ஆம் ஆண்டில் வெளியிடப்பட்ட செக்டர் மென்பொருள் (Sector Software) தயாரிக்கப்பட்ட ஸ்பெல்பவுண்ட் திட்டம் (Spellbound program) மற்றும் வேர்ட் 95 முதல் மைக்ரோசாஃப்ட் வேர்ட் போன்ற ஒரு ஊடாடும் வகையில் பின்னணி எழுத்துப்பிழை சோதனை செய்யப்பட்டது.

சமீபத்திய ஆண்டுகளில், எழுத்துப்பிழைச் சரிபார்ப்பிகள் பெருகிய முறையில் அதி நவீனமாகிவிட்டன; சில இப்போது எளிய இலக்கண பிழைகளை அடையாளம் காணும் திறன் கொண்டவை. இருப்பினும், மிகச் சிறந்த நிலையில் கூட, அவை ஒரு உரையில் உள்ள அனைத்து பிழைகளையும் (ஒப்புருமொழியம்/ஹோமோஃபோன் பிழைகள் போன்றவை) அரிதாகவே கண்டுபிடிக்கின்றன; மேலும் அவை புதுச்சொல்லாக்கங்களையும்/நியோலாஜிசங்களையும் அந்நியச் சொற்களையும் எழுத்துப்பிழைகளாகக் காட்டும். ஆயினும்கூட, எழுத்துப்பிழை சரிபார்ப்பிகள் ஒரு வகை அந்நிய மொழி எழுதும் உதவிக்கருவியாகக் கருதப்படலாம்; இது சொந்தம் அல்லாத மொழி (அந்நிய மொழி) கற்பவர்கள் இலக்கு மொழியில் தங்கள் எழுத்துப்பிழைகளைக் கண்டறிந்து சரிசெய்ய நம்பலாம்.

6.4.3. உரை நேர்செய்யும் ஒழுங்குமுறை (Text editing system)

உரை திருத்தி/சிராக்கி என்பது எளிய உரையைத் திருத்தும் ஒரு வகை கணினி நிரலாகும். மைக்ரோசாஃப்ட் நோட்பேடின் (Microsoft Notepad) பெயரைத் தொடர்ந்து இதுபோன்ற திட்டங்கள் சில நேரங்களில் "நோட்பேட்" மென்பொருள் ("notepad" software) என்று அழைக்கப்படுகின்றன. உரை சீராக்கிகள் இயக்க முறைமைகள் (operating systems) மற்றும் மென்பொருள் மேம்பாட்டு தொகுப்புகளுடன் (software development packages) வழங்கப்படுகிறன; மேலும் அவை கட்டமைப்பு கோப்புகள் (configuration files), ஆவணக் கோப்புகள் (documentation files) மற்றும் நிரலாக்க மொழி மூல குறியீடு (programming language source code) போன்ற கோப்புகளை மாற்றப் பயன்படுகிறது.

எளிய உரைக்கு எதிராக வளமான உரை (Plain text vs. rich text)

எளிய உரைக்கும் (உரை ஆசிரியர்களால் உருவாக்கப்பட்டது மற்றும் திருத்தப்பட்டது) வளமான உரைக்கும் (வேர்ட் செயலிகள் அல்லது டெஸ்க்டாப் வெளியீட்டு மென்பொருளால் உருவாக்கப்பட்டது போன்றவை) இடையே முக்கியமான வேறுபாடுகள் உள்ளன.

எளிய உரை பிரத்தியேகமாக எழுத்து பிரதிநிதித்துவத்தைக் கொண்டுள்ளது. ASCII, ISO / IEC 2022, UTF-8 அல்லது யூனிகோட் போன்ற குறிப்பிட்ட எழுத்துக்குறி குறியீட்டு மரபுகளுக்கு இணங்க, ஒவ்வொரு எழுத்தும் ஒன்று, இரண்டு, அல்லது நான்கு பைட்டுகளின் நிலையான நீள வரிசை அல்லது ஒன்று முதல் நான்கு பைட்டுகளின் மாறி-நீள வரிசை என குறிப்பிடப்படுகிறது.. இந்த மரபுகள் அச்சிடக்கூடிய பல எழுத்துக்களை வரையறுக்கின்றன; ஆனால் இடத்தின் ஓட்டம், வரி முறிவு மற்றும் பக்க முறிவு போன்ற உரையின் ஓட்டத்தை கட்டுப்படுத்தும் அச்சிடப்படாத எழுத்துக்களையும் வரையறுக்கின்றன. எளிய உரையில் உரையைப் பற்றிய வேறு எந்த தகவலும் இல்லை; எழுத்துக்குறி குறியீட்டு மரபு கூட பயன்படுத்தப்படவில்லை. உரை கோப்புகள் வெற்று உரையை பிரத்தியேகமாக சேமிக்கவில்லை என்றாலும், எளிய உரை உரைக்கோப்புகளில் சேமிக்கப்படுகிறது. கணினிகளின் ஆரம்ப நாட்களில், ஒரு மோனோஸ்பேஸ்/ஒற்றையிடைவெளி எழுத்துருவைப் (monospace font) பயன்படுத்தி எளிய உரை காட்டப்பட்டது, அதாவது கிடைமட்ட சீரமைப்பு (horizontal alignment) மற்றும் நெடுவரிசை வடிவமைத்தல் (columnar formatting) சில நேரங்களில் வெள்ளை இடைவெளி எழுத்துக்களைப் (whitespace characters) பயன்படுத்தி செய்யப்பட்டன. பொருந்தக்கூடிய காரணங்களுக்காக, இந்த பாரம்பரியம் மாறவில்லை.

மாறாக வளமான உரை, மெட்டாடேட்டா (metadata), எழுத்து வடிவமைத்தல் தரவு (எ.கா. டைப்ஃபேஸ், அளவு, எடை மற்றும் பாணி), பத்தி வடிவமைத்தல் தரவு (character formatting data) (எ.கா. உள்தள்ளல் indentation, வரிசயமைப்பு/சீரமைப்பு (alignment), எழுத்து மற்றும் சொல் விநியோகம் மற்றும் கோடுகள் அல்லது பிற பத்திகளுக்கு இடையில் இடம்), மற்றும் பக்க விவரக்குறிப்பு தரவு (page specification data) (எ.கா. அளவு, ஓரம் மற்றும் வாசிப்பு திசை). வளமான உரை மிகவும் சிக்கலானதாக இருக்கும். வளமான உரையை பைனரி வடிவத்தில் (எ.கா. டிஓசி/ DOC), மார்க்அப் மொழியுடன் (எ.கா. ஆர்.டி.எஃப்/RTF அல்லது எச்.டி.எம்.எல்./HTML) ஒட்டக்கூடிய உரை கோப்புகள் அல்லது இரண்டின் கலப்பின வடிவத்தில் (எ.கா. அலுவலக திறந்த எக்ஸ்எம்எல்/Office Open XML) சேமிக்க முடியும்.

உரை சீராக்கிகள் எளிய உரை அல்லது எளிய உரையாக விளங்கக்கூடிய எதையும் கொண்ட உரை கோப்புகளை திறந்து சேமிக்க வேண்டும், இதில் வளமான உரைக்கான மார்க்அப் அல்லது வேறு ஏதாவது மார்க்அப் (எ.கா. எஸ்.வி.ஜி/ SVG) அடங்கும்.

உரை சீராக்கிகளின் வகைகள்

சில உரை சீராக்கிகள் சிறிய மற்றும் எளிமையானவைகள்; மற்றவைகள் பரந்த மற்றும் சிக்கலான செயல்பாடுகளை வழங்குகின்றன. எடுத்துக்காட்டாக, யூனிக்ஸ் மற்றும் யூனிக்ஸ் போன்ற இயக்க முறைமைகளில் பைக்கோ எடிட்டர் (pico editor) (அல்லது ஒரு மாறுபாடு) உள்ளது, ஆனால் பலவற்றில் vi மற்றும் Emacs எடிட்டர்களும் அடங்கும். மைக்ரோசாப்ட் விண்டோஸ் அமைப்புகள் எளிய நோட்பேடோடு (Notepad) வருகின்றன, இருப்பினும் பலர்-குறிப்பாக புரோகிராமர்கள்-கூடுதல் அம்சங்களைக் கொண்ட பிற எடிட்டர்களை விரும்புகிறார்கள். ஆப்பிள் மேகிண்டோஷின் (Apple Macintosh கிளாசிக் மேக் ஓஎஸ்ஸின் (Classic Mac OS) (கணினி மென்பொருள்) கீழ் சொந்த சிம்பிள் டெக்ஸ்ட் (native SimpleText) இருந்தது; இது மேக் ஓஎஸ் எக்ஸில் (Mac OS X) டெக்ஸ்ட் எடிட் (TextEdit) மூலம் மாற்றப்பட்டது, இது ஒரு உரை எடிட்டரின் அம்சங்களை வரைகோல்கள், ஓரங்கள் மற்றும் பல எழுத்துரு தேர்வு போன்ற ஒரு சொல் செயலியின் வழக்கமானவற்றுடன் இணைக்கிறது. இந்த அம்சங்கள் ஒரே நேரத்தில் கிடைக்காது, ஆனால் பயனர் கட்டளையால் மாற்றப்பட வேண்டும், அல்லது நிரல் மூலம் கோப்பு வகையைத் தானாக தீர்மானிக்கும்.

பெரும்பாலான சொல் செயலிகள் கோப்புகளை எளிய உரை வடிவத்தில் படிக்கலாம் மற்றும் எழுதலாம், இது உரை எடிட்டர்களிடமிருந்து சேமிக்கப்பட்ட கோப்புகளைத் திறக்க அனுமதிக்கிறது. இருப்பினும், இந்தக் கோப்புகளை ஒரு சொல் செயலியில் இருந்து சேமிக்க, கோப்பு எளிய உரை வடிவத்தில் எழுதப்பட்டிருப்பதை உறுதிசெய்வது அவசியம்; மேலும் எந்த உரை குறியாக்கம் அல்லது BOM (Byte order mark) அமைப்புகளும் கோப்பை அதன் நோக்கத்திற்காக மறைக்காது. வேர்ட்ஸ்டார் போன்ற WYSIWYG (What You See Is What You Get) அல்லாத சொல் செயலிகள் உரை எடிட்டர்களாக சேவையில் எளிதில் அழுத்தப்படுகின்றன, உண்மையில் 1980களில் பொதுவாக இது பயன்படுத்தப்பட்டது. இந்த சொல் செயலிகளின் இயல்புநிலை கோப்பு வடிவம் பெரும்பாலும் மார்க்அப் மொழியை ஒத்திருக்கிறது, அடிப்படை வடிவம் எளிய உரை மற்றும் காட்சி வடிவமைத்தல் அச்சிடப்படாத கட்டுப்பாட்டு எழுத்துக்கள் அல்லது தப்பிக்கும் காட்சிகளைப் பயன்படுத்தி அடையப்படுகிறது. மைக்ரோசாப்ட் வேர்ட் போன்ற பிற்கால சொல் செயலிகள் தங்கள் கோப்புகளை பைனரி வடிவத்தில் சேமித்து வைக்கின்றன, மேலும் அவை எளிய உரைக் கோப்புகளைத் திருத்த ஒருபோதும் பயன்படுத்தப்படுவதில்லை.

சில உரைச் சீராக்கிகள் வழக்கத்திற்கு மாறாக பதிவுக் கோப்புகள் அல்லது ஒரே கோப்பில் வைக்கப்பட்டுள்ள முழு தரவுத்தளம் போன்ற பெரிய கோப்புகளைத் திருத்தலாம். எளிமையான உரை தொகுப்பாளர்கள் கணினியின் பிரதான நினைவகத்தில் கோப்புகளைப் படிக்கலாம். பெரிய கோப்புகளுடன், இது மெதுவான செயல்முறையாக இருக்கலாம், மேலும் முழு கோப்பும் பொருந்தாது. இந்த வாசிப்பு முடிவடையும் வரை சில உரை சீராக்கிகள் பயனரைத் திருத்த அனுமதிக்க மாட்டாது. எடிட்டிங் செயல்திறன் பெரும்பாலும் சிறப்பு இல்லாத எடிட்டர்களிலும் பாதிக்கப்படுகிறது, எடிட்டர் விசை அழுத்தங்கள் அல்லது வழிசெலுத்தல் கட்டளைகளுக்கு பதிலளிக்க வினாடிகள் அல்லது நிமிடங்கள் கூட ஆகும். சிறப்புச் சீராக்கிகள் பெரிய கோப்புகளின் புலப்படும் பகுதியை நினைவகத்தில் மட்டுமே சேமிப்பது, எடிட்டிங் செயல்திறனை மேம்படுத்துதல் போன்ற மேம்படுத்தல்களைக் கொண்டுள்ளன.

சில சீராக்கிகள் நிரல்படுத்தக் கூடியவைகள்; எடுத்துக்காட்டாக குறிப்பிட்ட பயன்பாடுகளுக்கு அவை தனிப்பயனாக்கப்படலாம். புரோகிராம் செய்யக்கூடிய எடிட்டருடன், மீண்டும் மீண்டும் செய்யக்கூடிய பணிகளை தானியக்கமாக்குவது அல்லது புதிய செயல்பாட்டைச் சேர்ப்பது அல்லது எடிட்டரின் கட்டமைப்பிற்குள் ஒரு புதிய பயன்பாட்டைச் செயல்படுத்துவது எளிது. தனிப்பயனாக்குவதற்கான ஒரு பொதுவான நோக்கம், உரை எடிட்டர் பயனர் மிகவும் பரிச்சயமான மற்றொரு உரை எடிட்டரின் கட்டளைகளைப் பயன்படுத்துவது அல்லது பயனர் சார்ந்து காணாமல் போன செயல்பாட்டை நகலெடுப்பது. மென்பொருள் உருவாக்குநர்கள் பெரும்பாலும் அவர்கள் பணிபுரியும் நிரலாக்க மொழி அல்லது மேம்பாட்டு சூழலுக்கு ஏற்றவாறு எடிட்டர் தனிப்பயனாக்கங்களைப் பயன்படுத்துகின்றனர். சில உரை சீராக்கிகளின் நிரல் திறன் நிரலின் முக்கிய எடிட்டிங் செயல்பாட்டை மேம்படுத்துவதற்கு மட்டுப்படுத்தப்பட்டுள்ளது, ஆனால் எமாக்ஸ் உரை கோப்புகளைத் திருத்துவதற்கு அப்பால் நீட்டிக்க முடியும் வலை உலாவுதல் (web browsing), மின்னஞ்சல் வாசித்தல், ஆன்லைன் அரட்டை, கோப்புகளை நிர்வகித்தல் அல்லது கேம்களை விளையாடுவது மற்றும் உரை பயனர் இடைமுகத்துடன் ஒரு லிஸ்ப் செயல்படுத்தும் சூழலாக பெரும்பாலும் கருதப்படுகிறது. யுனிக்ஸ் கலாச்சாரத்தின் பாரம்பரிய ஆசிரியர் போர்களில் அதன் போட்டியாளரான Vi‌ஐப் பின்பற்றவும் Emacs திட்டமிடப்படலாம்.

நிரல்படுத்தக்கூடிய எடிட்டர்களின் முக்கியமான குழு, ஸ்கிரிப்டிங் மொழியாக REXX ஐப் பயன்படுத்துகிறது. இந்த "பழமைவாதச் சீராக்கிகளில்" (orthodox editors) ஒரு "கட்டளை வரி"

உள்ளது, அதில் கட்டளைகள் மற்றும் மேக்ரோக்களை தட்டச்சு செய்யலாம் மற்றும் உரை கோடுகள் எந்த வரி கட்டளைகள் மற்றும் மேக்ரோக்களை தட்டச்சு செய்யலாம். இதுபோன்ற பெரும்பாலான எடிட்டர்கள்/சீராக்கிகள் ISPF/PDF EDIT அல்லது XEDIT-இன் வழித்தோன்றல்கள், ஐபிஎம்மின் விஎம்/ எஸ்பிக்கான தலைமை எடிட்டர் z/VM மூலம். அவற்றில் THE, KEDIT, X2, Uni-edit மற்றும் SEDIT ஆகியவை அடங்கும்.

ஒரு குறிப்பிட்ட பயன்பாட்டிற்காக எழுதப்பட்ட அல்லது தனிப்பயனாக்கப்பட்ட ஒரு உரை திருத்தி பயனர் எதைத் திருத்துகிறார் என்பதைத் தீர்மானிக்கவும் பயனருக்கு உதவவும் முடியும், பெரும்பாலும் நிரலாக்க விதிமுறைகளை பூர்த்திசெய்து தொடர்புடைய ஆவணங்களுடன் உதவிக்குறிப்புகளைக் காண்பிப்பதன் மூலம். மென்பொருள் உருவாக்குநர்களுக்கான பல உரைத் தொகுப்பிகள் மூலக் குறியீடு தொடரியல் சிறப்பம்சமாக மற்றும் நிரல்களை எளிதாகப் படிக்கவும் எழுதவும் தானியங்கி உள்தள்ளல் அடங்கும். புரோகிராமிங் எடிட்டர்கள்/சீராக்கிகள் பெரும்பாலும் ஒரு சேர்க்கப்பட்ட கோப்பு, செயல்பாடு அல்லது மாறியின் பெயரைத் தேர்ந்தெடுக்க பயனரை அனுமதிக்கின்றன; பின்னர் அதன் வரையறைக்கு செல்லவும். ஆரம்ப கர்சர் இருப்பிடத்தை சேமிப்பதன் மூலம் அல்லது கோரப்பட்ட வரையறையை பாப்அப் சாளரத்தில் அல்லது தற்காலிக இடையகத்தில் காண்பிப்பதன் மூலம் குறியீட்டின் அசல் பகுதிக்கு எளிதாக செல்லவும் சில அனுமதிக்கின்றன. சில சீராக்கிகள் இந்த திறனைத் தாங்களே செயல்படுத்துகின்றன; ஆனால் பெரும்பாலும் வரையறைகளைக் கண்டறிய ctags போன்ற துணைப் பயன்பாடு பயன்படுத்தப்படுகிறது.

வழக்கமான அம்சங்கள்

கண்டுபிடித்து மாற்றுதல் - உரைத் தொகுப்பான் கோப்புகளின் குழுக்களில் அல்லது ஊடாடும் வகையில் உரையைத் தேடுவதற்கும் மாற்றுவதற்கும் விரிவான வசதிகளை வழங்குகின்றது. மேம்பட்ட சீராக்கிகள் உரை அல்லது குறியீட்டைத் தேட மற்றும் திருத்த வழக்கமான வெளிப்பாடுகளைப் பயன்படுத்தலாம்.

வெட்டுதல், நகலெடுத்தல், ஒட்டுதல் - பெரும்பாலான உரை சீராக்கிகள் கோப்பிற்குள் அல்லது கோப்புகளுக்கு இடையில் உரையை நகலெடுத்து நகர்த்துவதற்கான முறைகளை வழங்குகின்றன. யுடிஎஃப் -8 குறியிடப்பட்ட உரையை கையாளும் திறன்.

உரை வடிவமைத்தல் - உரைத் தொகுப்பான்கள் பெரும்பாலும் வரி மடக்கு, தானியங்கு உள்தள்ளல், ஆஸ்கி எழுத்துக்களைப் பயன்படுத்தி புல்லட் பட்டியல் வடிவமைப்பு, கருத்து

வடிவமைப்பு, தொடரியல் சிறப்பம்சங்கள் போன்ற அடிப்படை காட்சி வடிவமைப்பு அம்சங்களை வழங்குகின்றன. இவைப் பொதுவாக காட்சிக்கு மட்டுமே மற்றும் வடிவமைத்தல் குறியீடுகளை கோப்பில் செருக வேண்டாது.

செயல்தவிர்த்தல் மற்றும் மீண்டும் செய்தல் - சொல் செயலிகளைப் போலவே, உரை சீராக்கிகளும் கடைசி திருத்தத்தை செயல்தவிர்க்க மற்றும் மீண்டும் செய்வதற்கான வழியை வழங்குகின்றன. பெரும்பாலும்-குறிப்பாக பழைய உரை எடிட்டர்களுடன்-திருத்த வரலாறு ஒரு நிலை மட்டுமே நினைவில் உள்ளது மற்றும் செயல்தவிர் கட்டளையை அடுத்தடுத்து வெளியிடுவது கடைசி மாற்றத்தை "நிலைமாறும்". நவீன அல்லது மிகவும் சிக்கலான சீராக்கிகள் வழக்கமாக பல நிலை வரலாற்றை வழங்குகின்றன; அதாவது செயல்தவிர் கட்டளையை மீண்டும் மீண்டும் வழங்குவது ஆவணத்தை அடுத்தடுத்த பழைய திருத்தங்களுக்கு மாற்றும். ஒரு தனி மீண்டும் செய் கட்டளை மிக சமீபத்திய மாற்றங்களை நோக்கி "முன்னோக்கி" திருத்தங்களை சுழற்றும். நினைவில் வைத்திருக்கும் மாற்றங்களின் எண்ணிக்கை எடிட்டரைப் பொறுத்தது மற்றும் பெரும்பாலும் பயனரால் உள்ளமைக்கப்படுகிறது.

மேம்படுத்தப்பட்ட அம்சங்கள்

மேக்ரோ அல்லது நடைமுறை வரையறை: புதிய கட்டளைகள் அல்லது அம்சங்களை முந்தைய கட்டளைகள் அல்லது பிற மேக்ரோக்களின் சேர்க்கைகளாக வரையறுக்க, ஒருவேளை கடந்து வந்த அளவுருக்கள் அல்லது மேக்ரோக்களின் கூடுடன்.

எடிட்டிங் அமர்வுக்கு இடையில் பயனர் அமைத்த விருப்பங்களைத் தக்கவைத்துக்கொள்ளச் சுயவிவரங்கள்.

குறிப்பிடப்பட்ட பெயர்களைக் கொண்ட சுயவிவர மேக்ரோக்கள், எ.கா., சூழல், சுயவிவரம், திருத்த அமர்வின் தொடக்கத்தில் அல்லது புதிய கோப்பைத் திறக்கும்போது தானாகவே செயல்படுத்தப்படும்.

பல கோப்பு எடிட்டிங்: ஒரு திருத்த-அமர்வின் போது பல கோப்புகளைத் திருத்தும் திறன், ஒவ்வொரு கோப்பின் தற்போதைய-வரி கர்சரை நினைவில் வைத்துக் கொள்ளலாம், ஒவ்வொரு கோப்பிலும் மீண்டும் மீண்டும் உரையைச் செருகவும், கோப்புகளில் உரையை நகலெடுக்கவும் அல்லது நகர்த்தவும், கோப்புகளை அருகருகே ஒப்பிடவும் (ஒருவேளை டைல் செய்யப்பட்ட பல ஆவண இடைமுகத்துடன்), முதலியன.

மல்டி-வியூ எடிட்டர்கள்: ஒரே கோப்பின் பல காட்சிகளைக் காண்பிக்கும் திறன், சுயாதீன கர்சர் கண்காணிப்பு, சாளரங்களில் மாற்றங்களை ஒத்திசைத்தல், ஆனால் சுயாதீன கோப்புகளுக்குக் கிடைக்கும் அதே வசதிகளை வழங்குதல்.

சுருக்குதல் / விரிவாக்குதல், மடித்தல் என்றும் அழைக்கப்படுகிறது: உரையின் பிரிவுகளை பார்வையில் இருந்து தற்காலிகமாக விலக்கும் திறன். இது வரி எண்களின் வரம்பை அடிப்படையாகக் கொண்டதாக இருக்கலாம் அல்லது சில தொடரியல் உறுப்பு அடிப்படையில் இருக்கலாம், எ.கா., ஒரு BEGINக்கு இடையில் உள்ள அனைத்தையும் தவிர்த்து; மற்றும் பொருந்தும் END;.

நெடுவரிசை அடிப்படையிலான எடிட்டிங்; ஒரு குறிப்பிட்ட நெடுவரிசையில் தரவை மாற்ற அல்லது செருகும் திறன் அல்லது குறிப்பிட்ட நெடுவரிசைகளுக்கு தரவை மாற்றும் திறன்.

தரவு மாற்றம் - தற்போது திருத்தப்பட்டுள்ள கோப்பில் மற்றொரு உரை கோப்பின் உள்ளடக்கங்களைப் படித்தல் அல்லது இணைத்தல். இயக்க முறைமையின் ஷெல்லுக்கு வழங்கப்பட்ட கட்டளையின் வெளியீட்டைச் செருக சில உரை ஆசிரியர்கள் ஒரு வழியை வழங்குகிறார்கள். மேலும், ஒரு வழக்கு மாற்றும் அம்சம் சிற்றெழுத்து அல்லது பெரிய எழுத்துக்கு மொழிபெயர்க்கலாம்.

வடிகட்டுதல் - சில மேம்பட்ட உரை தொகுப்பாளர்கள் எடிட்டரை கோப்பின் அனைத்து அல்லது பிரிவுகளையும் மற்றொரு பயன்பாட்டிற்கு அனுப்ப அனுமதிக்கிறார்கள் மற்றும் "வடிகட்டப்பட்ட" வரிகளுக்கு பதிலாக கோப்பில் முடிவை மீண்டும் படிக்கலாம். எடுத்துக்காட்டாக, தொடர்ச்சியான வரிகளை அகர வரிசைப்படி அல்லது எண்ணாக வரிசைப்படுத்தவும், கணித கணக்கீடுகளைச் செய்யவும், மூலக் குறியீட்டை உள்தள்ளவும் மற்றும் பலவற்றிற்கும் இது பயனுள்ளதாக இருக்கும். தொடரியல் சிறப்பம்சமாக - ஒழுங்கமைக்கப்பட்ட அல்லது கணிக்கக்கூடிய வடிவத்தில் தோன்றும் மூலக் குறியீடு, மார்க்அப் மொழிகள், கட்டமைப்பு கோப்புகள் மற்றும் பிற உரையை சூழ்நிலைப்படுத்துகிறது. ஒவ்வொரு மொழி உறுப்புக்கும் பயன்படுத்தப்படும் வண்ணங்கள் அல்லது பாணிகளைத் தனிப்பயனாக்க பயனர்கள் பொதுவாக பயனர்களை அனுமதிக்கின்றனர். சில உரை தொகுப்பிகள் எடிட்டரின் முழு பயனர் இடைமுகத்தின் தோற்றத்தையும் உணர்வையும் மாற்ற கருப்பொருள்களை நிறுவவும் பயன்படுத்தவும் பயனர்களை அனுமதிக்கின்றன.

விரிவாக்கம் - புரோகிராமர்களால் பயன்படுத்த விரும்பும் உரை திருத்தி சில சொருகி பொறிமுறையை வழங்க வேண்டும், அல்லது ஸ்கிரிப்ட் செய்யக்கூடியதாக இருக்க வேண்டும்,

எனவே ஒரு புரோகிராமர் தனிப்பட்ட மென்பொருள் திட்டங்களை நிர்வகிக்க தேவையான அம்சங்களுடன் எடிட்டரைத் தனிப்பயனாக்கலாம், குறிப்பிட்ட நிரலாக்க மொழிகள் அல்லது பதிப்பு கட்டுப்பாட்டு அமைப்புகளுக்கான செயல்பாடு அல்லது முக்கிய பிணைப்புகளைத் தனிப்பயனாக்கலாம், அல்லது குறிப்பிட்ட குறியீட்டு பாணிகளுக்கு இணங்க.

சிறப்பு எடிட்டர்கள்

சில எடிட்டர்கள் சிறப்பு அம்சங்கள் மற்றும் கூடுதல் செயல்பாடுகளை உள்ளடக்குகின்றன, எடுத்துக்காட்டாக,

மூலக் குறியத் தொகுப்பாளர்கள் (Source code editors) மூலக் குறியீட்டின் உற்பத்தியை எளிதாக்கக் கூடுதல் செயல்பாட்டுடன் கூடிய உரை எடிட்டர்கள். இவை பெரும்பாலும் பயனர் நிரல்படுத்தக்கூடிய தொடரியல் சிறப்பம்சங்கள் மற்றும் குறியீடு வழிசெலுத்தல் செயல்பாடுகள் மற்றும் குறியீட்டு கருவிகள் அல்லது ஒரு HTML எடிட்டரைப் போன்ற விசைப்பலகை மேக்ரோக்களைக் கொண்டுள்ளன (கீழே காண்க).

மடிப்பு எடிட்டர்கள் (Folding editors). இந்த துணைப்பிரிவில் "பழமைவாத எடிட்டர்கள்" ("orthodox editors") என்று அழைக்கப்படுபவை அடங்கும், அவை செடிட்டின் வழித்தோன்றல்கள். நிரலாக்க-குறிப்பிட்ட அம்சங்கள் இல்லாமல் மடிப்புகளை செயல்படுத்தும் எடிட்டர்கள் பொதுவாக அவுட்லைனர்கள் என்று அழைக்கப்படுகின்றன (கீழே காண்க).

ஒருங்கிணைந்த மேம்பாட்டு சூழல்கள் (integrated development environments (IDE)) பெரிய நிரலாக்க திட்டங்களை நிர்வகிக்கவும் ஒழுங்கமைக்கவும் வடிவமைக்கப்பட்டுள்ளன. எளிமையான உரை எடிட்டிங் தேவையற்ற பல அம்சங்களைக் கொண்டிருப்பதால் அவை வழக்கமாக நிரலாக்கத்திற்கு மட்டுமே பயன்படுத்தப்படுகின்றன.

உலகளாவிய வலை படைப்பாளிகளுக்கு (World Wide Web authors) வலைப்பக்கங்களை உருவாக்கும் பணிக்கு அர்ப்பணிக்கப்பட்ட பல்வேறு HTML எடிட்டர்கள் வழங்கப்படுகின்றன. இவை பின்வருவனவற்றை உள்ளடக்குகின்றன: ட்ரீம்வீவர், கொம்போசர் மற்றும் மின் உரை எடிட்டர். உள்ளமைக்கப்பட்ட HTML ரெண்டரிங் இயந்திரம் அல்லது நிலையான வலை உலாவியில் செயலில் உள்ள ஒரு வேலையைப் பார்க்கும் விருப்பத்தை பலர் வழங்குகிறார்கள்.

மூல வலை எடிட்டர் அல்லது ஐடிஇ பயன்படுத்தி ரூபி அல்லது பி.எச்.பி போன்ற டைனமிக் நிரலாக்க மொழியில் பெரும்பாலான வலை அபிவிருத்தி செய்யப்படுகிறது. எளிமையான நிலையான வலைத்தளங்களைத் தவிர மற்ற அனைவராலும் வழங்கப்பட்ட HTML தளத்தை

கட்டுப்படுத்தும் மென்பொருளால் கூடியிருக்கும் தனி வார்ப்புரு கோப்புகளாக சேமிக்கப்படுகிறது மற்றும் முழுமையான HTML ஆவணத்தை உருவாக்காது.

கணிதவியலாளர்கள், இயற்பியலாளர்கள் மற்றும் கணினி விஞ்ஞானிகள் பெரும்பாலும் எளிய உரை கோப்புகளில் TeX அல்லது LaTeX ஐப் பயன்படுத்தி கட்டுரைகளையும் புத்தகங்களையும் தயாரிக்கிறார்கள். இத்தகைய ஆவணங்கள் பெரும்பாலும் ஒரு நிலையான உரை எடிட்டரால் தயாரிக்கப்படுகின்றன, ஆனால் சிலர் சிறப்பு TeX எடிட்டர்களைப் பயன்படுத்துகிறார்கள்.

அவுட்லைனர்கள் (Outliners) கிளை சார்ந்த எடிட்டர்கள் (tree-based editors) என்றும் அழைக்கப்படுகிறது, ஏனென்றால் அவை ஒரு படிநிலை அவுட்லைன் மரத்தை உரை திருத்தியுடன் இணைக்கின்றன. மடிப்பு (மேலே காண்க) கோடிட்டுக் காட்டும் ஒரு சிறப்பு வடிவமாகக் கருதலாம்.

கூட்டு எடிட்டர்கள் (Collaborative editors) ஒரு நெட்வொர்க்கில் தொலைதூர இடங்களிலிருந்து ஒரே நேரத்தில் ஒரே ஆவணத்தில் பல பயனர்களை வேலை செய்ய அனுமதிக்கின்றன. தனிப்பட்ட பயனர்களால் செய்யப்பட்ட மாற்றங்கள் முரண்பட்ட திருத்தங்களுக்கான சாத்தியத்தை அகற்ற தானாகவே ஆவணத்தில் கண்காணிக்கப்பட்டு ஆவணத்தில் இணைக்கப்படுகின்றன. இந்த எடிட்டர்கள் பொதுவாக எடிட்டர்களிடையே கலந்துரையாடலுக்கான ஆன்லைன் அரட்டை கூறுகளையும் உள்ளடக்குகின்றன.

கவனச்சிதறல் இல்லாத எடிட்டர்கள் (Distraction-free editors) எழுத்தாளரை மீதமுள்ள பயன்பாடுகள் மற்றும் இயக்க முறைமையிலிருந்து தனிமைப்படுத்தும் நோக்கத்துடன் ஒரு குறைந்தபட்ச இடைமுகத்தை வழங்குகின்றன, இதனால் கருவிப்பட்டி அல்லது அறிவிப்பு பகுதி போன்ற இடைமுகக் கூறுகளிலிருந்து கவனச்சிதறல்கள் இல்லாமல் எழுத்தில் கவனம் செலுத்த முடியும்.

நிரல்படுத்தக்கூடிய எடிட்டர்கள் (Programmable editors) பொதுவாக இந்த செயல்பாடுகளை ஏதேனும் அல்லது அனைத்தையும் செய்ய மேம்படுத்தலாம், ஆனால் எளிமையான ஆசிரியர்கள் ஒன்றில் மட்டுமே கவனம் செலுத்துகிறார்கள், அல்லது ஜி.பி.எச்.பெடிட் போன்றவை ஒற்றை நிரலாக்க மொழியை இலக்காகக் கொண்டுள்ளன.

6.4.4. உருபனியல் பகுப்பாய்வு ஒழுங்குமுறை (Morphological processing system)

உருபனியல் பகுப்பாய்வு ஒழுங்குமுறை உருவாக்கத்தில் தரவுத்தொகுதியின் பங்கு மிக முக்கியமாகக்குறிப்பிடத் தகுந்ததாகும். தொடக்ககாலத்தில் உருபனியல் பகுப்பாய்வு

ஒழுங்குமுறை மொழியியல் சார்ந்த விதி அடிப்படையிலேயே உருவாக்கப்பட்டன. உருபன்களும் அவற்றில் வருகை முறைகளும், பகுதிகளும் பகுதி மாற்றங்களும், கட்டுருபன்கள், கட்டில்லா உருபன்கள்/தனி உருபன்கள், வெற்றுருபன்கள், தொடர் உருபன்கள், தொடரிலா உருபன்கள், வேர் உருபன்கள், வேரில்லா உருபன்கள், இரட்டை உருபன்கள், அடுக்கு உருபன்கள், குறைப்பு உருபன்கள் போன்ற பல வகையிலான உருபன்களும், ஒட்டுக்களும் (பின்னொட்டு, முன்னொட்டு, இடையொட்டு போன்றவை) அவற்றின் வருகை முறைகளும், பலவகை இலக்கணச் செயல்பாடுகளும், மாற்றுருபன்களும் அவற்றின் வருகை இடங்களும் உருபொலியயல் விதிகள், சந்தி விதிகள் என்பனவும் கணக்கில் எடுக்கப்பட்டு உருபனியல் பகுப்பாய்வு ஒழுங்குமுறைகள் உருவாக்கப்பட்டன. இவ்வொழுங்குமுறைகளின் குறைபாடுகளை நிவர்த்தி செய்யும் பொருட்டும், காலவிரையத்தைத் தடுக்கும் பொருட்டும் புள்ளியியல் அணுகுமுறைகளும் (stastics based approach) இயந்திரங்கற்றல் அணுகுமுறைகளும் (machine learning approach) பயன்படுத்தப்பட்டன. புள்ளியியல் அணுகுமுறைகளும் இயந்திரங்கற்றல் அணுகுமுறைகளும் தரவுத்தொகுதி அடிப்படையிலானவை. தரவுத்தொகுதி இயக்க (corpus driven) உருபனியல் பகுப்பாய்வு ஒழுங்குமுறைகளும் உருவாக்கப்பட்டன.

டோக்கன்களிலிருந்து இலக்கணத் தகவல்களைப் பெறுவதற்கான செயல்முறையாக உருபனியல் பகுப்பாய்வு வரையறுக்கப்படலாம்; அவற்றின் பகுதி, விகுதி, இடைநிலை தகவல்கள் தரப்படவேண்டும். உருபனியல் பகுப்பாய்வு மூன்று வழிகளில் செய்யப்படலாம்: உருபன் அடிப்படையிலான உருபனிவியல் பகுப்பாய்வு (அல்லது அனிடெம் மற்றும் ஏற்பாடு அணுகுமுறை), சொல்லன் அடிப்படையிலான உருபனியல் (அல்லது ஒரு பொருள் மற்றும் செயல்முறை அணுகுமுறை), மற்றும் சொல் அடிப்படையிலான உருபனியல் பகுப்பாய்வு (அல்லது ஒரு சொல் மற்றும் முன்னுதாரண அணுகுமுறை). கொடுக்கப்பட்ட உள்ளீட்டு டோக்கனின் உருவவியல் பகுப்பாய்விற்கு பொறுப்பான ஒரு நிரலாக ஒரு உருபனியல் பகுப்பாய்வி வரையறுக்கப்படலாம். இது கொடுக்கப்பட்ட டோக்கனை பகுப்பாய்வு செய்கிறது மற்றும் பாலினம், எண், வகுப்பு மற்றும் பல போன்ற உருவ தகவல்களை ஒரு வெளியீடாக உருவாக்குகிறது.

உருபனியல் பகுப்பாய்வின் பல்வேறு முறைகள் பல்வேறு இயற்கை மொழி ஆய்வு (என்.எல்.பி) ஆராய்ச்சி குழுக்கள் உருவவியல் பகுப்பாய்விற்கான வெவ்வேறு முறைகள் மற்றும் வழிமுறைகளை உருவாக்கியுள்ளன. சில வழிமுறைகள் மொழி சார்ந்தவை, அவற்றில் சில மொழி

சுதந்தினமானவை. உருபனியல் பகுப்பாய்வில் ஈடுபட்டுள்ள பல்வேறு முறைகள் பற்றிய சுருக்கமான விளக்கம் பின்வருவனவற்றை உள்ளடக்குகிறது.

• முற்றுநிலைத் தானியங்கி (ஃபைனைட் ஸ்டேட் ஆட்டோமேட்டா/Finite State Automata (FSA)).

• இரண்டு நிலை உருபனியல் (Two Level Morphology)

• முற்றுநிலைத் தானியங்கி மாற்றி (FST).

• பகுதியாக்கி அல்காரிதம் · Stemmer Algorithm.

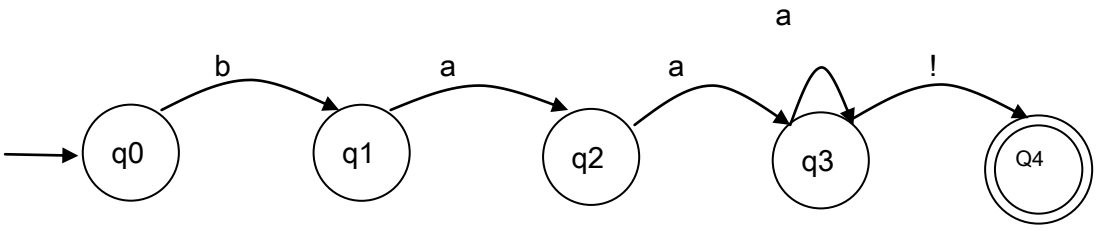
• தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை

• இயக்கிய அக்ரிலிக் சொல் வரைபடம் (Directed Acrylic Word Graph (DAWG))

• சொல்லடுக்கு சார்ந்த அணுகுமுறை (Paradigm Based Approach)

முற்றுநிலை நிலை தானியங்கி (FSA)

ஒரு முற்றுநிலை இயந்திரம் அல்லது முற்றுநிலைத் தானியங்கி என்பது நிலை, மாற்றங்கள் மற்றும் செயல்களால் ஆன நடத்தை மாதிரியாகும். ஒரு முற்றுநிலைத் தானியங்கி என்பது ஒரு வரையறுக்கப்பட்ட எண்ணிகைகளின் நிலைகளில் ஒன்றில் இருக்கக்கூடிய ஒரு கருவி ஆகும். தானியங்கி இறுதி நிலையில் இருந்தால், அது செயல்படுவதை நிறுத்தும்போது, உள்ளீடு குறியீடுகளின் வரிசையாக இருக்கும் இடத்தில் அதன் உள்ளீட்டை ஏற்றுக்கொள்வதாகக் கூறப்படுகிறது. கொடுக்கப்பட்ட மொழியில் ஒரு கோர்வையை ஏற்க அல்லது நிராகரிக்க முற்றுநிலைத் தானியங்கி பயன்படுத்தப்படுகிறது மற்றும் வழக்கமான வெளிப்பாடுகளைப் பயன்படுத்துகிறது. தானியங்கி இயக்கப்படும் போது அது ஆரம்ப கட்டத்தில் இருக்கும் மற்றும் செயல்படத் தொடங்கும். இறுதி நிலையில், அது கொடுக்கப்பட்ட கோர்வையை ஏற்கும் அல்லது நிராகரிக்கும். ஆரம்ப நிலைக்கும் வரையறுக்கப்பட்ட/இறுதி நிலைக்கும் இடையில் மாற்றங்கள் அதாவது ஒரு நிலையிலிருந்து மற்றொரு நிலைக்கு மாறுவதற்கான செயல்முறைகள் உள்ளன. உருபனியல் அகராதி மற்றும் புரிதலைக் குறிக்க முற்றுநிலைத் தானியங்கியைப் பயன்படுத்தலாம்.



மேற்கண்ட முற்றுநிலைத் தானியங்கி பின்வரும் குழுமத்தில் உள்ள ஏதாவது ஒரு கோர்வை ஏற்பதாகக் கருதலாம்.

baa!

baaa!

baaaa!

baaaaa!

இரண்டு நிலை உருவவியல்

கிம்மோ கோஸ்கென்னீமி (Kimmo Koskenniemi) 1983ஆம் ஆண்டில் தனது ஆய்வேட்டில் சொல்-வடிவம் அறிதலுக்கும் (word-form recognition) உருவாக்கத்திற்கும் இரண்டு நிலை உருபனியல் கணினிசார் மாதிரியை வழங்கியுள்ளார். இந்த மாதிரி மொழியியல் வல்லுநர்கள் உருபன் அமைப்பொழுங்குங்கும் (மோர்போடாக்டிக்ஸ்/morphotactics) (எந்த வரிசையில் உருபன்கள் வரக்கூடும் என்பதைக் குறிப்பிடுகிறது) மற்றும் உருபொலியனியல் (மோர்போபோனெமிக்ஸ்/ morphophonemics) (உருபன்கள் வரும் ஒலியியல் சூழலுக்கு ஏற்ப மாற்று வடிவங்கள் அல்லது உருபன்களின் "எழுத்துக்கூட்டல்கள்" இவற்றைக் கணக்கில் எடுக்கின்றது) ஆகியவற்றுக்கு இடையிலான பாரம்பரிய வேறுபாட்டை அடிப்படையாகக் கொண்டது. எடுத்துக்காட்டாக, chased என்ற சொல், பகுதி chase-ஐப் தொடர்ந்து பின்னொட்டு -ed என உருபொலியனியல் அடிப்படையில் பகுப்பாய்வு செய்யப்படுகிறது. இருப்பினும், -ed என்ற பின்னொட்டு சேர்ப்பது chase என்பதன் இறுதி ‘e’ என்பதன் இழப்பை ஏற்படுத்துகிறது; இதனால் chase மற்றும் chas என்பது ஒரே உருபனின் மாற்றுருபு அல்லது மாற்று வடிவங்கள் ஆகும். இவ்வாறு கோஸ்கென்னீமி முற்றுநிலை சொற்களில் ஒலியியல் மாற்றங்களை (phonological alternations) விவரிக்க ஒரு புதிய வழியைக் கண்டுபிடித்தார். இடைநிலை நிலைகளைக் கொண்ட அடுக்கு விதிகளுக்குப் (cascaded rules) பதிலாக, விதிகள் சொல்சார் கோர்வைகளின் புறவடிவ உணர்தலை நேரடியாகக் கட்டுப்படுத்தும் கூற்றுகளாக கருதப்படலாம். விதிகள் தொடர்ச்சியாக பயன்படுத்தப்படாது இணையாக பயன்படுத்தப்படும். கோஸ்கென்னீமி (1983) விதி தொகுப்பி (rule compiler), கூட்டமைவு (composition) அல்லது வேறு எந்த இறுதிநிலை வழிமுறையையும் (finite-state algorithm) சாராத தனது கட்டுப்பாட்டு அடிப்படையிலான மாதிரியை (constraint-based model) செயல்படுத்துவதை உருவாக்கினார்; மேலும் அவர் அதை இரண்டு-நிலை உருபனியல் (two-level morphology) என்று அழைத்தார்.

கோஸ்கென்னீமியின் மாதிரி இரண்டு உருப்படுத்தங்களையும் ஒரு விதியையும் கொண்டுள்ளது:

- ஒரு சொல் வடிவத்தின் புற உருப்படுத்தம்: இது இறுதி செல்லுபடியாகும் வார்த்தையின் உண்மையான எழுத்துவடிவம்.
- ஒரு சொல்-வடிவத்தின் சொல்சார் (உருபொலியனியல் என்றும் அழைக்கப்படுகிறது) உருப்படுத்தம்: இது அடிப்படை வடிவங்களின் எளிமையான ஒருங்கிணைப்பைக் காட்டுகிறது.
- விதி கூறு: இது இரண்டு உருப்படுத்தங்களை ஒன்றையொன்று பொருத்தும் விதிகளைக் கொண்டுள்ளது. ஒவ்வொரு விதியும் ஒரு முற்றுநிலை மாற்றி (finite-state transducer) மூலம் விவரிக்கப்படுகிறது:

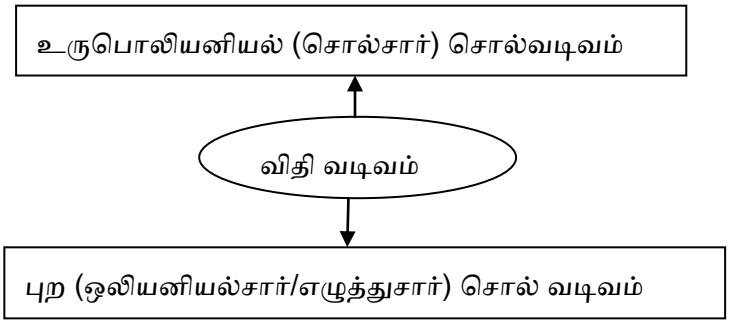
எடுத்துக்காட்டாக, <chased> என்ற சொல்லுக்கு இந்த இரண்டு-நிலை உருப்படுத்தம் வழங்கப்படுகிறது (இங்கு + ஒரு உருபன் எல்லைக் குறியீடு மற்றும் 0 என்பது பூஜ்ய எழுத்து):

சொல்சார் வடிவம் (Lexical form): c h a s e + e d

புற வடிவம் (Surface form): c h a s 0 0 e d

கோஸ்கென்னீமியின் இரு-நிலை உருபனியல் என்பது கணினி மொழியியல் வரலாற்றில் உருபனியல் ரீதியாக சிக்கலான மொழிகளின் பகுப்பாய்விற்கான முதல் நடைமுறை பொது மாதிரியாகும். மொழிச் சார்ந்த கூறுகள், விதிகள் மற்றும் அகராதி என்பன அனைத்து மொழிகளுக்கும் பொருந்தக்கூடிய உலகளாவிய இயக்கநேர இயந்திரத்துடன் இணைக்கப்பட்டன.

படம்: இரு நிலை உருபனியல்



கிம்மோ பாகுபடுத்தி (The KIMMO Parser)

ஆன்ட்வொர்த், 1990 பிசி-கிம்மோ நிரலை உண்மையில் ஷெல் நிரலாக அழைத்தார்; இது பழமையான பிசி-கிம்மோ செயல்பாடுகளுக்கு ஒரு ஊடாடும் பயனர் இடைமுகமாக செயல்பட்டது. இந்தச் செயல்பாடுகள் சி-மொழி மூல குறியீடு நூலகமாக கிடைக்கின்றன, அவை பயனரால் எழுதப்பட்ட நிரலில் சேர்க்கப்படலாம்.

ஒரு மொழியின் பிசி-கிம்மோ விளக்கம் பயனர் வழங்கிய இரண்டு கோப்புகளைக் கொண்டுள்ளது:

(1) விதிமுறைகள் கோப்பு: இது எழுத்துக்கள் மற்றும் ஒலியியல் (அல்லது எழுத்துப்பிழை) விதிகளைக் குறிப்பிடுகிறது, மற்றும்

(2) அகராதி/லெக்சிகன் கோப்பு: இது சொல்சார் அலகுகள் (சொற்கள் மற்றும் உருபன்கள்) மற்றும் அவற்றின் அர்த்தங்களைப் பட்டியலிடுகிறது; மேலும் உருபன் அமைப்பொழுங்குக் கட்டுப்பாடுகளை குறிமாகம் (encodes) செய்கிறது.

பிசி-கிம்மோவில் பொதிந்துள்ள ஒலியியல் கோட்பாட்டு மாதிரி இரண்டு நிலை ஒலியனியல் என (two-level phonology) அழைக்கப்படுகிறது. இரண்டு-நிலை அணுகுமுறையில், சொற்களின் அடிப்படை உருப்படுத்தத்தின் சொல்சார் நிலைக்கும் புற மட்டத்தில் அவை உணரப்படுவதற்கும் இடையிலான தொடர்பாக ஒலியனியல் கருதப்படுகிறது. எடுத்துக்காட்டாக, ஆங்கில எழுத்துக்கூட்டல் விதிகளைக் கணக்கிட, மேற்பரப்பு வடிவ ஒற்றர்கள் அதன் சொற்பொருள் வடிவமான `spy + s பின்வருமாறு தொடர்புடையதாக இருக்க வேண்டும் (இங்கு ` என்ற குறி அழுத்தத்தைக் குறிக்கிறது, + என்பது ஒரு உருபன் எல்லையைக் குறிக்கிறது, மற்றும் 0 பூஜ்ய உறுப்பைக் குறிக்கிறது):

சொல்சார் உருப்படுத்தம்: ` s p y + 0 s

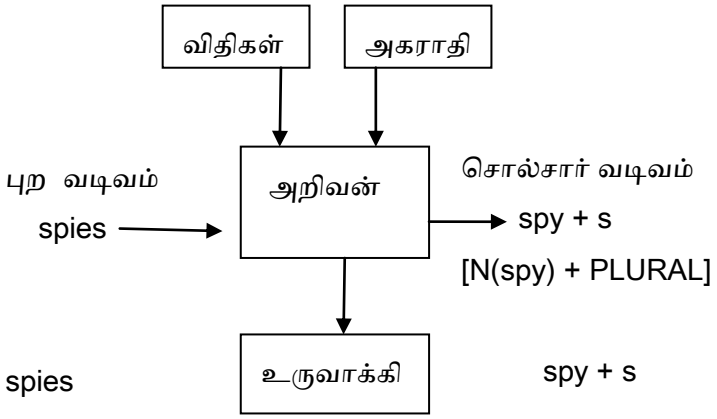
புற உருப்படுத்தம் : 0 s p i 0 e s

சிறப்புத் தொடர்புகளுக்கு விதிகள் எழுதப்பட வேண்டும்: `:0, y:i, +:0, மற்றும் 0:e.

கிமோவின் (PC-KIMMOபிசி) இரண்டு செயல்பாட்டு கூறுகள் உருவாக்கி (generator) மற்றும் அறிவான் (recognizer) என்பனவாகும். உருவாக்கி சொல்சார் வடிவத்தை உள்ளீடாக ஏற்றுக்கொள்கிறது, ஒலியியல் விதிகளைப் பயன்படுத்துகிறது, மேலும் அதனுடன் தொடர்புடைய புற வடிவத்தைத் தருகிறது மற்றும் அகராதியைப் பயன்படுத்தாது. அறிவான் புற வடிவத்தை உள்ளீடாக ஏற்றுக்கொள்கிறது, ஒலியியல் விதிகளைப் பயன்படுத்துகிறது, அகராதியைக் கலந்தாலோசிக்கிறது மற்றும் அதனுடன் தொடர்புடைய அகராதி வடிவத்தை அதன்

அர்த்தத்துடன் வழங்குகிறது. கீழே கொடுக்கப்பட்டுள்ள படம் பிசி-கிம்மோ அமைப்பின் முக்கிய கூறுகளைக் காட்டுகிறது.

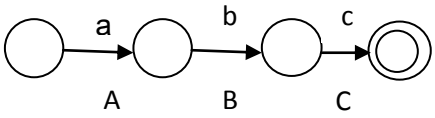
படம்: பிசி-கிம்மோவின் கூறுகள்



முற்று நிலை மாற்றி (Beesley et al., 2003)

முற்று நிலை மாற்றி (Finte State Tranducer (FST/எஃப்எஸ்டி) என்பது -இன் முற்று நிலை தானியங்கி (finite state automata (FSA/எஃப்எஸ்ஏ)) மேம்பட்ட பதிப்பாகும், மேலும் இது கணினியியல் ரீதியாக அகராதியைக் குறிக்கப் பயன்படுகிறது. இரண்டு நிலை உருபனியல் கொள்கையை ஏற்றுக்கொள்வதன் மூலம் இது செய்யப்படுகிறது. இரண்டு நிலை உருபனியல் ஒரு சொல்லை சொல்சார் நிலைக்கும் புற மட்டத்திற்கும் இடையிலான பொருத்தத்தைக் குறிக்கிறது. எஃப்எஸ்டி இரண்டு டேப் தானியங்கியாகக் குறிப்பிடப்படுகிறது. எஃப்எஸ்டி அகராதியில், ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்க எழுத்துக்கூட்டல் விதிகள் மற்றும் எழுத்துகூட்டல் வேறுபாடுகள் இணைக்கப்பட்டுள்ளன. ஒரு எஃப்எஸ்டி வெறுமனே ஒரு மரபு முற்று நிலை தானியங்கி ஆகும், அதன் மாற்றங்கள் ஒற்றைக் குறியீடுகளுடன் இல்லாமல் ஜோடிகளுடன் பெயரிடப்பட்டுள்ளன, எ.கா. Σ = {a: a, b: b, a: c, a: ε, e:, ...}. இது ஒரு முற்று தானியங்கி வழியாக ஒரு குறியீட்டின் குழுமத்தை மற்றொன்றுக்குப் பொருத்துகிறது. கீழேயுள்ள படம் a: A, b: B மற்றும் c: C ஜோடிகளுக்கு மேல் கட்டப்பட்ட ஒரு FSTஐக் காட்டுகிறது.

படம்: ஒரு தன்னிச்சையான முற்று நிலை மாற்றி



கே (Kay 1987) பொதுவாக மொழியியலாளர்கள் மற்றும் குறிப்பாக கணினி மொழியியலாளர்கள், சாத்தியமான இடங்களில் முற்றுநிலை கருவிகளைப் பயன்படுத்துவதன் மூலம் பயனடைவார்கள் என்று கூறுகிறார். அவை கோட்பாட்டளவில் ஈர்க்கக்கூடியவை, ஏனென்றால் அவை கணிதக் கண்ணோட்டத்தில் நன்கு புரிந்து கொள்ளப்படுகின்றன. அவை எளிமையான, நேர்த்தியான மற்றும் மிகவும் திறமையான செயலாக்கங்களை உருவாக்குவதால் அவை கணினியியல்சார் ரீதியாக ஈர்க்கின்றன. பின்வரும் மூன்று காரணங்களால் முற்று-நிலை கருவிகளுடன் கணக்கிடுவது கவர்ச்சிகரமானதாக பீஸ்லி மற்றும் கார்டுனென் (Beesley and Karttunen 2003) வலியுறுத்துகின்றனர்.

முதலாவதாக, முற்று-நிலை இயந்திரங்களின் கணித பண்புகள் நன்கு புரிந்து கொள்ளப்படுகின்றன. இது வரையறுக்கப்பட்ட-நிலை சாதனங்களை மாற்றியமைக்க மற்றும் இணைக்க ஒருவரை அனுமதிக்கிறது

பிற பாரம்பரிய வழிமுறை நிரல்களைப் பயன்படுத்துவது சாத்தியமில்லை. வேறு வார்த்தைகளில் கூறுவதானால் இது குறிப்பாக தலைகீழ், குறுக்குவெட்டு, சேர்க்கை மற்றும் கலவை போன்ற பண்புகள் காரணமாக முற்றுநிலை கருவிகளின் “கணித அழகு” “இணையற்ற நெகிழ்வுத்தன்மை” என்று மொழிபெயர்க்கப்பட்டுள்ளது, (Beesley and Karttunen 2003).

இரண்டாவதாக, முற்று-நிலை கருவிகள் கணினிசார் ரீதியாக திறமையானவை, இதன் விளைவாக சிறந்தவை செயலாக்க வேகம் உள்ளவை.

மூன்றாவதாக, பெரும்பாலான சந்தர்ப்பங்களில், முற்று-நிலை கருவிகள் ஒப்பீட்டளவில் சிறிய நினைவகத்தில் நிறைய தகவல்களைச் சேமிக்க முடியும் (Beesley and Karttunen, 2003).

பகுதியாக்கி (Stemmer)

ஒடுக்களை அகற்ற ஸ்டெம்மர் பயன்படுத்தப்படுகிறது. இது பகுதிகளின் பட்டியல் மற்றும் மாற்று விதிகளைக் கொண்ட விதிகளின் தொகுப்பைப் பயன்படுத்துகிறது.

எடுத்துக்காட்டாக:

writing→ write + ing

ஒரு ஸ்டெமர் நிரலுக்கு, சாத்தியமான அனைத்து ஒட்டுக்களும் மாற்று விதிகளுடன் குறிப்பிடப்பட வேண்டும்.

எ.கா.

ational→ ate relational→ relate

tional→ tion conditional→ condition

மிகவும் பரவலாக பயன்படுத்தப்படும் ஸ்டெமர் வழிமுறை பாட்டர் வழிமுறை (Potter algorithm) ஆகும். இந்திய மொழிகளுக்கும் ஸ்டெம்மரை உருவாக்க சில முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன.

தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை

தரவுத்தொகுதி என்பது ஒரு குறிப்பிட்ட மொழியைச் சேர்ந்த எழுதப்பட்ட உரையின் பெரிய தொகுப்பாகும். மூல தரவுத்தொகுதியை உருபனிய்யல் பகுப்பாய்விற்குப் பயன்படுத்தலாம். இது மூல தரவுத்தொகுதி உள்ளீடாக எடுத்து உரையில் காணப்பட்ட சொல் வடிவங்களின் பிரிவை உருவாக்குகிறது. இத்தகைய பிரிவு உருபனியல் பிரிவை ஒத்திருக்கிறது. ஹெல்சின்கி பல்கலைக்கழகத்தில் உருவாக்கப்பட்ட மோர்ஃபெசர் 1.0 என்பது ஒரு தரவுத்தொகுதி அடிப்படையிலான மொழி சுதந்திரமான உருபனியல் பிரிவு திட்டமாகும். இந்தியமொழிகளிலிருந்து இந்திய மொழிகளுக்கான மொழிபெயர்ப்புத் திட்டம் ஒரு தரவுத்தொகுதி அடிப்படையிலான உருபனியல் பகுப்பாய்வியை வெற்றிகரமாக உருவாக்கி பயன்படுத்தியது, இது சொல்லடுக்கு அடிப்படையிலான அணுகுமுறையையும் தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறையையும் ஒருங்கிணைக்கிறது.

இயக்கிய அக்ரிலிக் வேர்ட் வரைபடம் (Directed Acrylic Word Graph (DAWG))

DAWG என்பது அகராதி பிரதிநிதித்துவம் மற்றும் பலவகையான பயன்பாடுகளுடன் விரைவான கோர்வை பொருத்துதலுக்கான மிகவும் திறமையான தரவு கட்டமைப்பாகும். இந்த முறை கிரேக்க மொழியில் கிரேக்கத்தின் பார்ட்டாஸ் பல்கலைக்கழகத்தால் வெற்றிகரமாக செயல்படுத்தப்பட்டுள்ளது. DAWG தரவு கட்டமைப்பு உருவவியல் பகுப்பாய்வு மற்றும் உருவாக்கம் ஆகிய இரண்டிற்கும் பயன்படுத்தப்படலாம். இந்த அணுகுமுறை மொழி சுதந்திரமானது மற்றும் எந்த உருபனியல் விதிகளையும் அல்லது வேறு எந்த சிறப்பு மொழியியல் தகவலையும் பயன்படுத்தாது (Sagarbas, 2000).

சொல்லடுக்கு அணுகுமுறை (Paradigm approach)

ஒரு சொல்லடுக்கு (paradigm) கொடுக்கப்பட்ட பகுதியின் (stem) அனைத்து சொல் வடிவத்தையும் வரையறுக்கிறது மற்றும் ஒவ்வொரு சொல் வடிவத்துடனும் ஒரு பண்புக்கூறு கட்டமைப்பை வழங்குகிறது. சொல்லடுக்கு அடிப்படையிலான அணுகுமுறை ஊக்கமளிக்கும் திரிபுஅடிப்படையில் வளமையாம்ன் மொழிகளுக்கு திறமையானது. அனுசாரகா ஆராய்ச்சி குழு இந்திய மொழிகளுக்கான மொழி சுதந்திரமான முன்னுதாரண அடிப்படையிலான உருபனியல் தொகுப்பி திட்டத்தை உருவாக்கியது. இது அல்லது இந்த திட்டத்தின் மாறுபட்ட பதிவு இந்திய அரசின் நிதி நல்கையில் நடைபெற்ற மொழிபெயர்ப்புத் திட்டங்களில் பரவலாகப் பயன்படுத்தப்பட்டது.

ஒரு மொழியில் உள்ள சொற்களை உள்ளடக்கிய சொல் வடிவங்களின் வெவ்வேறு அட்டவணைகளை வழங்க மொழியியலாளர் அல்லது மொழி நிபுணர் கேட்கப்படுகிறார். சொல்-வடிவ அட்டவணை ஒரு வேர்களின் தொகுப்பை உள்ளடக்கியது, அதாவது வேர்கள் அவற்றின் சொல் வடிவங்களை உருவாக்குவதற்கான அட்டவணையில் உள்ளார்ந்த வடிவத்தை (அல்லது முன்னுதாரணத்தை) பின்பற்றுகின்றன. இந்த முறையைப் பயன்படுத்தி இந்திய மொழிகளுக்கான கிட்டத்தட்ட அனைத்து உருபனியல் பகுப்பாய்விகளும் உருவாக்கப்பட்டுள்ளன. முன்மாதிரிகளின் அடிப்படையில், நிரல் பகுப்பாய்வு செய்வதற்கான நீக்குதல் சரத்தை உருவாக்குகிறது. சொல்லடுக்கு அடிப்படையிலான அணுகுமுறை பல்வேறு வகையான சொல் அடுக்குகள் அவற்றின் உருபனியல் நடத்தை அடிப்படையில் அமைந்திருப்பதைக் கண்டுபிடிப்பதை நம்பியுள்ளது. இலக்கண அம்சங்கள் மற்றும் சொல் உருவாக்கும் செயல்முறைகளில் உள்ள ஒற்றுமையின் அடிப்படையில் சொற்களை முன்னுதாரணங்களாக வகைப்படுத்தலாம். சொல்லடுக்குக் கட்டுமானத்திற்காக, திரிபுற்ற சொற்களின் மாதிரித் தரவுத்தொகுதி எடுக்கப்படுகிறது. வெவ்வேறு சொல்-வடிவங்களைக் கொண்ட ஒவ்வொரு அட்டவணையும் வேர்களின் தொகுப்பை உள்ளடக்கியது, வேர்கள் அவற்றின் சொல்-வடிவங்களை உருவாக்குவதற்கான வடிவத்தை (அல்லது சொல்லடுக்கை) பின்பற்றுகின்றன என்பதைக் குறிக்கிறது. சொற்கள் பெயர்ச்சொற்கள், வினைச்சொற்கள், பெயரடைகள், வினையடைகள் மற்றும் பின்னொட்டுகள் என வகைப்படுத்தப்படுகின்றன. ஒவ்வொரு வகையும் அவற்றின் உருபொலியனியல்சா (மார்போபோனெமிக்) நடத்தையின் அடிப்படையில் சில வகையான சொல்லடுக்குகளாக வகைப்படுத்தப்படும்.

பின்னொட்டு நீக்குதல் (Suffix stripping)

ஒரு மொழியில் உள்ள சொற்களை பகுப்பாய்வு செய்ய பயன்படுத்தப்படும் மற்றொரு முறை பின்னொட்டு நீக்குதல். அதிக ஒட்டுநிலை மொழிகளில் (agglutinative languages), வேர் அல்லது பகுதியுடன் பின்னொட்டுகளைச் சேர்ப்பதன் மூலம் ஒரு சொல் உருவாகிறது. பின்னொட்டு நீக்குதல் முறை மொழியின் இந்த பண்பைப் பயன்படுத்துகிறது; அதாவது, பகுதியுடன் சிக்கலான பின்னொட்டுகள் இணைக்கப்பட்டுள்ளது. பின்னொட்டு அடையாளம் காணப்பட்டதும், அந்த பின்னொட்டை அகற்றி, சரியான சந்தி விதிகளைப் பயன்படுத்துவதன் மூலம் முழு வார்த்தையின் பகுதியைப் பெறலாம். எடுத்துக்காட்டாக, வந்தான் என்ற தமிழ் வினைச் சொல்லிலிருந்து *ஆன்* என்ற விகுதியும் *த்த்* என்ற இறந்தகால இடைநிலையும் அகற்றப்பட்டு *வ* என்ற பகுதி பெறப்படும்.

6.4.5. வாக்கியப் பகுப்பாய்வு ஒழுங்குமுறை (Sentence parsing system)

பாகுபடுத்தல் அல்லது பகுப்பாய்வு அல்லது தொடரியல் பகுப்பாய்வு என்பது இயற்கையான மொழி, கணினி மொழிகள் அல்லது தரவு கட்டமைப்புகளில், முறையான இலக்கணத்தின் விதிகளுக்கு இணங்க, குறியீடுகளின் கோர்வையைப் பகுப்பாய்வு செய்யும் செயல்முறையாகும். பாகுபடுத்தல் என்ற சொல், ('part of speech)' அதாவது தமிழில் 'சொல்வகைப்பாடு' என்று பொருள்படும் லத்தீன் மொழிச் சொல் *pars* (*orationis*) என்பதிலிருந்து வந்தது.

இந்தச் சொல் மொழியியல் மற்றும் கணினி அறிவியலின் வெவ்வேறு கிளைகளில் சற்று மாறுபட்ட அர்த்தங்களைக் கொண்டுள்ளது. பாரம்பரிய வாக்கியப் பாகுபடுத்தல் பெரும்பாலும் ஒரு வாக்கியம் அல்லது வார்த்தையின் சரியான பொருளைப் புரிந்துகொள்வதற்கான ஒரு முறையாக செய்யப்படுகிறது; சில நேரங்களில் வாக்கிய வரைபடங்கள் போன்ற உபாயங்களின் உதவியுடன். இது பொதுவாக எழுவாய் மற்றும் பயனிலை போன்ற இலக்கணப் பிரிவுகளின் முக்கியத்துவத்தை வலியுறுத்துகிறது.

கணினி மொழியியலில் இந்தச் சொல் கணினியால் ஒரு வாக்கியத்தை அல்லது பிற சொற்களின் கோர்வையை அதன் உறுப்புகளாக/கூறுகளாகப் பிரிக்கும் முறையான பகுப்பாய்வை குறிக்கப் பயன்படுகிறது; இதன் பயனாக உறுப்புகளுக்கிடையில் தொடரியல் உறவைக் காட்டுவதுடன் பொருண்மையியல்சார் மற்றும் பிற தகவல்களையும் கொண்டிருக்கும் பாகுப்பாய்வு கிளையை (parse tree) விளைகிறது. சில பாகுப்பாய்வு வழிமுறைகள் தொடரியல்

அடிப்படையில் பொருண்மை மயக்கமுள்ள உள்ளீட்டுக்கு, ஒரு பகுப்பாய்வுக் காடு (parse forest) அல்லது பகுப்பாய்வுக் கிளைகளின் பட்டியலை உருவாக்கக்கூடும்.

மொழி புரிதலை விவரிக்கும் போது இந்த சொல் உளமொழியிலும் பயன்படுத்தப்படுகிறது. இந்தச் சூழலில், பகுப்பாய்வு என்பது மனிதன் ஒரு வாக்கியத்தை அல்லது சொற்றொடரை (பேச்சு மொழி அல்லது உரையில்) இலக்கணக் கூறுகள்/உறுப்புகள் அடிப்படையில் தொடரியல் உறவுகள், சொல்வகைப்பாடு போன்றவற்றை அடையாளம் காணும் பகுப்பாய்வு செய்யும் முறையைக் குறிப்பிடும். இந்தச் சொல் குறிப்பாக தோட்ட-பாதை வாக்கியங்களை (garden-path sentences) விளக்குவதற்கு மொழி பேசுபவர்களுக்கு எந்த மொழியியல் குறிப்புகள் பயன்படும் என விவாதிக்கப்படும் போது பொதுவானது.

கணினி அறிவியலுக்குள், கணினி மொழிகளின் (computer languages) பகுப்பாய்வில் இந்த சொல் பயன்படுத்தப்படுகிறது; இது தொகுப்பிகள் (compilers) மற்றும் மொழிபெயர்ப்பிகளை (interpreters) எழுதுவதற்கு வசதியாக உள்ளீட்டுக் குறியத்தை அதன் கூறு பகுதிகளாகப் (component parts) பிரிக்கும் தொடரியல் பகுப்பாய்வைக் குறிக்கும். பிளவு (split) அல்லது பிரிவினையை (separation) விவரிக்கவும் இந்த சொல் பயன்படுத்தப்படலாம்.

பாரம்பரிய முறைகள் (Traditional methods)

பகுப்பாய்வின் பாரம்பரிய இலக்கண நடைமுறை, சில நேரங்களில் துணைநிலைத்தொடர் பகுப்பாய்வு (clause analysis) என அழைக்கப்படுகிறது, ஒவ்வொரு பகுதியின் வடிவம், செயல்பாடு மற்றும் தொடரியல் உறவு பற்றிய விளக்கத்துடன் உரையை அதன் சொல்வகைப்பாட்டுக் கூறுகளாக உடைப்பதை உள்ளடக்குகிறது. மொழியின் வினைத்திரிபாக்கத்தையும் பெயர்த்திரிபாக்கத்தையும் ஆய்வதில் இருந்து இது பெருமளவில் தீர்மானிக்கப்படுகிறது, இது பெரிதும் திரிபுற்ற மொழிகளுக்கு மிகவும் சிக்கலானதாக இருக்கும். 'மனிதன் நயைக் கடித்தான்' போன்ற ஒரு சொற்றொடரை ஆய்வது, 'மனிதன்' என்ற ஒற்றை பெயர்ச்சொல் வாக்கியத்தின் எழுவாய் என்பதைக் குறிப்பிடுவதை உள்ளடக்கியது, 'கடித்தான்' என்ற வினைச்சொல் 'கடி' என்ற வினைச்சொல்லின் இறந்தகால ஆண்பால் படர்க்கை இடத்தைக் குறித்து நிற்கும். 'நாய்' என்ற ஒற்றை பெயர்ச்சொல் வாக்கியத்தின் செய்ப்படுபொருள் ஆகும். வாக்கிய வரைபடங்கள் போன்ற நுட்பங்கள் சில நேரங்களில் வாக்கியத்தின் கூறுகளுக்கு இடையிலான உறவைக் குறிக்கப் பயன்படுத்தப்படுகின்றன.

பகுப்பாய்வு என்பது முன்னர் ஆங்கிலம் பேசும் உலகம் முழுவதும் இலக்கணத்தை கற்பிப்பதில் மையமாக இருந்தது, மேலும் எழுதப்பட்ட மொழியின் பயன்பாடு மற்றும் புரிதலுக்கான அடிப்படை என்று பரவலாகக் கருதப்பட்டது. இருப்பினும், அத்தகைய நுட்பங்களின் பொதுவான கற்பித்தல் இப்போது இல்லை.

கணினிசார் முறைகள்

சில இயந்திர மொழிபெயர்ப்பு மற்றும் இயற்கை மொழி செயலாக்க அமைப்புகளில், மனித மொழிகளில் எழுதப்பட்ட நூல்கள் கணினி நிரல்களால் பாகுபடுத்தப்படுகின்றன. மனித வாக்கியங்கள் நிரல்களால் எளிதில் பகுப்பாய்வு செய்யப்படுவதில்லை, ஏனெனில் மனித மொழியின் கட்டமைப்பில் கணிசமான பொருண்மைமயக்கம் உள்ளது, இதன் பயன்பாடு வரம்பற்ற சாத்தியங்களின் பரப்பொல்லைகளுக்கு இடையில் அர்த்தத்தை (அல்லது பொருண்மையியல்) தெரிவிப்பதாகும், ஆனால் அவற்றில் சில மட்டுமே குறிப்பிட்ட வழக்கில் உள்ளன. எனவே "Man bites dog" மற்றும் "Dog bites man" என்ற கூற்றுகள் ஒரு மொழியின் விரம் அடிப்படையில் திட்டவட்டமானது; ஆனால் மற்றொரு மொழியில் "Man bites dog" என்பது அந்த இரண்டு சாத்தியக்கூறுகளையும் வேறுபடுத்துவதற்கு பெரிய சூழலை நம்பியிருக்கும், உண்மையில் அந்த வேறுபாடு இருந்தால் அது கருத்தத்தக்கததாகும். சில விதிகள் பின்பற்றப்படுகின்றன என்பது தெளிவாகத் தெரிந்தாலும் முறைசாரா நடத்தை விவரிக்க முறையான விதிகளைத் தயாரிப்பது கடினம்.

இயற்கையான மொழித் தரவை ஆய்வதற்கு, ஆராய்ச்சியாளர்கள் முதலில் பயன்படுத்த வேண்டிய இலக்கணத்தை ஒப்புக் கொள்ள வேண்டும். தொடரியலின் தேர்வு மொழியியல் மற்றும் கணினிசார் அக்கறைகளால் பாதிக்கப்படுகிறது; எடுத்துக்காட்டாக, சில பகுப்பாய்வு ஒழுங்குமுறைகள் லெக்சிகல் செயல்பாட்டு இலக்கணத்தைப் (Lexical functional grammar (LFG)) பயன்படுத்துகின்றன, ஆனால் பொதுவாக, இந்த வகை இலக்கணங்களை பாகுபடுத்துவது NP-முழுமையானது (NP-complete) என்று அறியப்படுகிறது. தலையால் இயக்கப்படும் சொற்றொடர் தொடரமைப்பு இலக்கணம் (Head-driven phrase structure grammar (HPSG)) என்பது பகுப்பாய்வு சமூகத்தில் பிரபலமாக உள்ள மற்றொரு மொழியியல் வடிவவாதம் ஆகும்; ஆனால் பிற ஆராய்ச்சி முயற்சிகள் பென் ட்ரீபேங்கில் (Penn Treebank) பயன்படுத்தப்படுவது போன்ற குறைவான சிக்கலான வடிவவாதங்களில் கவனம் செலுத்தியுள்ளன. ஆழமில்லாப் பகுப்பாய்வு (Shallow parsing) என்பது பெயர்ச்சொல் சொற்றொடர்கள் போன்ற முக்கிய கூறுகளின்

எல்லைகளை மட்டுமே கண்டுபிடிப்பதை நோக்கமாகக் கொண்டுள்ளது. மொழியியல் சர்ச்சையைத் தவிர்ப்பதற்கான மற்றொரு பிரபலமான உத்தி சார்பு இலக்கண பாகுபகுபாய்வு ஆகும் (dependency grammar parsing).

பெரும்பாலான நவீன பாகுப்பாய்விகள் குறைந்தது ஓரளவு புள்ளிவிவரங்கள் (partly statistical) அடிப்படையிலானதாகும்; அதாவது, அவை ஏற்கனவே அடையாளப்படுத்தப்பட்ட (கையால் பாகுப்பாய்வு செய்யப்பட்ட) பயிற்சி தரவின் தரவுத்தொகுதியை (corpus of training data) நம்பியுள்ளன. இந்த அணுகுமுறை குறிப்பிட்ட சூழல்களில் பல்வேறு கட்டுமானங்கள் நிகழும் அதிர்வெண் பற்றிய தகவல்களை சேகரிக்கக் கணினியை அனுமதிக்கிறது. பயன்படுத்தப்பட்ட அணுகுமுறைகளில் நேரடியான நிகழ்தகவு சூழல் இல்லாத இலக்கணங்கள் (probabilistic context-free grammars PCFGs/பி.சி.எஃப்.ஜிக்கள்), அதிகபட்ச என்ட்ரோபி (maximum entropy) மற்றும் நரம்பியல் வலைகள் (neural networks) ஆகியவை அடங்கும். மிகவும் வெற்றிகரமான ஒழுங்குமுறைகள் பெரும்பாலானவை சொல்சார் புள்ளிவிவரங்களைப் பயன்படுத்துகின்றன (அதாவது, அவை சம்பந்தப்பட்ட சொற்களின் அடையாளங்களையும், அவற்றின் சொல்வகைப்பாட்டையும் கருதுகின்றன). எவ்வாறாயினும், இத்தகைய ஒழுங்குமுறைகள் அதிகப்படியான பொருத்துதலுக்குப் பாதிக்கப்படக்கூடியவையாகும்; மேலும் பயனுள்ளதாக இருக்கச் சில வகையான சீர்படுத்தல் அவசியம்.

நிரலாக்க மொழிகளுக்காகக் கைமுறையாக வடிவமைக்கப்பட்ட இலக்கணங்களைப் போலவே இயற்கை மொழிக்கான பாகுப்பாய்வு வழிமுறைகள் 'நல்ல' பண்புகளைக் கொண்ட இலக்கணத்தை நம்ப இயலாது. முன்னர் குறிப்பிட்டபடி, சில இலக்கண முறைகள் கணினியியல் அடிப்படையில் பகுப்பாய்வது மிகவும் கடினம்; பொதுவாக, விரும்பிய கட்டமைப்பு சூழல் சுதந்திரமானதாக (context-free,) இலாது இருந்தாலும், இலக்கணத்திற்கு ஒருவித சூழல்-சுதந்திர தோராயமானது (context-free approximation) முதல் தேர்ச்சி செய்ய பயன்படுத்தப்படுகிறது. சூழல் சுந்தந்திரமான இலக்கணங்களைப் பயன்படுத்தும் வழிமுறைகள் வழக்கமாக நேரத்தை மிச்சப்படுத்த சாத்தியமில்லாத பகுப்பாய்வுகளைக் குறைக்க சில ஹியூரிஸ்டிக் உடன் பெரும்பாலும் Cocke–Younger–Kasami algorithm (CYK) வழிமுறையின் சில மாறுபாட்டை நம்பியுள்ளன. இருப்பினும் சில ஒழுங்குமுறைகள் துல்லியத்திற்காக வேகத்தை குறைக்கின்றன, எ.கா., ஷிப்ட்-குறைக்கும் வழிமுறையின் (shift-reduce algorithm) நேரியல் நேர பதிப்புகள் (linear-time versions). ஓரளவு சமீபத்திய வளர்ச்சியானது பகுப்பாய்வு மறுசீரமைப்பு ஆகும் (parse

reranking); இதில் பகுப்பாய்விகள் சில பெரிய எண்ணிக்கையிலான பகுப்பாய்வுகளை முன்மொழிகின்றன; மேலும் மிகவும் சிக்கலான அமைப்பு சிறந்த விருப்பத்தைத் தேர்ந்தெடுக்கிறது. பொருண்மையியல் பகுப்பாய்விகள் உரைகளை அவற்றின் அர்த்தங்களின் உருப்படுத்தங்களாக மாற்றுகின்றன.

6.4.6. நிகழ்வெண் கணக்கிடும் ஒழுங்குமுறை (Frequency counting system)

மொழியியலின் பல பயன்பாட்டு அடிப்படையிலான புதிய துணைத் துறைகளின் அறிமுகம் மொழிப் பண்புகளின் புள்ளிவிவரத் தகவல்களைக் வேண்டுகிறது, ஏனெனில் புள்ளிவிவரத் தகவல்கள், மொழி தொழில்நுட்பத்திற்கான கருவிகள் மற்றும் ஒழுங்குமுறைகளை வடிவமைப்பதற்கும், பாடநூல்கள் மற்றும் கற்பிப்பதற்கான உரைப்பொருட்களை உருவாக்குவதற்கும், முன்னர் உருவாக்கப்பட்ட கோட்பாடுகள் மற்றும் உற்றுநோக்கி அறிபவைகளைச் சரிபார்க்கவும் பயனுள்ளதாய் அமைகின்றது. உரைகள் அல்லது தரவுத்தொகுதிகளின் அதிர்வெண் எண்ணிக்கையை ஒப்பிடுவது பல பயன்பாடுகள் மற்றும் அறிவியல் துறைகளில் ஒரு முக்கியமான பணியாகும். கொடுக்கப்பட்டுள்ள ஒரு குழும உரைகளில், “எக்ஸ் என்ற சொல் அடிக்கடி நிகழ்கின்றது”, அல்லது “காலப்போக்கில் எக்ஸ் அடிக்கடி நிகழ்கிறது”அல்லது “எக்ஸ் என்ற சொல் பெண்களின் பேச்சை விட ஆண்களின் பேச்சில் அடிக்கடி நிகழ்கிறது” போன்ற ஒரு கருதுகோளை சோதிக்க விரும்புப்போது புள்ளியியல் சார் நிகழ்வெண் ஆய்வு கைகொடுக்கும்.

மொழியியலில், சொற்களின் அதிர்வெண்கள் மற்றும் பழமொழிகள் போன்ற பிற நிகழ்வுகள், பொருண்மையியல் அடையாளங்கள்கள், என்-கிராம் போன்றவை மக்கள் எவ்வாறு தொடர்பு கொள்கின்றன என்பதை அறியப் பரவலாகப் பயன்படுத்தப்படுகின்றன. சொற்களின் சூழ்நிலை நடத்தை மொழியில் மாறுபடும் மற்றும் இனம், தலைப்பு, படைப்பாளி (பாலினம், வயது, சமூக வகுப்பு) போன்ற காரணிகள் பலவற்றால் பாதிக்கப்படுகிறது. எடுத்துக்காட்டாக, எழுதப்பட்ட மொழியில், குறிப்பாக செய்தித்தாள் உரைகளில், நடையியல் குறிக்கோள்களின் காரணமாக ஒரு சொல்லை மீண்டும் மீண்டும் பயன்படுத்துவது தவிர்க்கப்படுகிறது. மாறாக உரையாடலில், சொற்களின் ஆரம்பம் மற்றும் தொடரியல் கட்டமைப்புகள் முக்கிய பங்கு வகிக்கின்றன. இயற்கையான மொழி ஒரேவிதமானதல்ல. எனவே குறிப்பிட்ட வார்த்தையைப் பொறுத்தது சொற்களின் அதிர்வெண்களில் பெரிய மாறுபாடு உள்ளது.

எலக்ட்ரானித் தரவுத்தொகுதியை அறிமுகப்படுத்தப்படுவதற்கு நீண்ட காலத்திற்கு முன்பே மொழி ஆய்வில் அளவு தகவல்களைப் பயன்படுத்த ஒப்புதல் அளிக்கப்பட்டது. அரை நூற்றாண்டுக்கு முன்னர், Flesch (1946) ஆங்கிலச் சொற்களைப் பற்றிய சில சுவாரஸ்யமான உற்றுநோக்கல்களை முன்வைத்தார். ஆங்கிலத்தில், 1.12 எழுத்து நீளம் கொண்ட சொற்கள் புரிந்துகொள்வது 'மிகவும் எளிதானது', 1.23 அசை நீளம் கொண்ட சொற்கள் 'எளிதானது', 1.39 அசை நீளம் கொண்ட சொற்கள் 'மிகவும் எளிதானது', 1.47 அசை நீளம் கொண்ட சொற்கள் 'தரமானவை', 1.55 அசை நீளமுள்ள சொற்கள் 'மிகவும் கடினம்', 1.67 எழுத்து அசை கொண்ட சொற்கள் 'கடினம்', மற்றும் 1.92 எழுத்துக்கள் அல்லது அதற்கு மேற்பட்ட சொற்கள் பொதுவான மொழி பயனர்களுக்குப் புரிந்துகொள்வது 'மிகவும் கடினம்'. சில ஆங்கில உரைநடை நூல்களில் சொற்களின் சராசரி நீளத்தைக் கண்டறிய டீவி (Dewey 1923) ஒரு தரவுத்தொகுதியையும் ஆய்வு செய்தார். Gibson (1962) ஷேக்ஸ்பியரின் எழுத்துக்களிலும், பைபிளின் அங்கீகரிக்கப்பட்ட பதிப்பிலும் சொற்களின் சராசரி நீளத்தைக் கணக்கிட இலக்கிய நூல்களின் தரவுத்தொகுதியைப் பயன்படுத்தினார். எல்டர்டன் (Elderton 1949), ஹெர்டன் (Herden 1956), குட் (Good 1957), மில்லர், நியூமன் மற்றும் ப்ரீட்மேன் (Miller, Newman, and Friedman 1958), எட்வர்ட்ஸ் மற்றும் சேம்பர்ஸ் (Edwards and Chambers 1964) மற்றும் பிறவற்றிற்கும் இதே போன்ற படைப்புகள் வரவு வைக்கப்படலாம்.

எலக்ட்ரானிக் வடிவத்தில் மொழித் தரவுத்தொகுதியை அறிமுகப்படுத்திய பின்னர், கணினி மூலம் பல்வேறு வகையான தரவுத்தொகுதிகளை எளிதில் அணுகுவதன் காரணமாக சொல்-நிலை புள்ளிவிவர ஆய்வுகள் கூடுதல் ஆர்வத்துடன் முயற்சிக்கப்படுகின்றன. இதுபோன்ற சில படைப்புகளை நாம் குறிப்பிடலாம், அவை ஆங்கிலத்தில் பல்வேறு வகையான தரவுத்தொகுதிகளை அடிப்படையாகக் கொண்டவை. ஒரு சுவாரஸ்யமான ஆய்வில், லீச், பிரான்சிஸ் மற்றும் சூ (Leech, Francis, and Xu 1994) ஆகியோர் ஆங்கிலத்தில் சொல் அர்த்தத்தில் தனித்தனி அல்லாத வகைகளின் இருப்பை ஆராய்ந்தனர்; அதே நேரத்தில் கில்கரிஃப் (Kilgarriff 1996) ஆங்கிலத்தின் வெவ்வேறு உரை தரவுத்தொகுதிகளின் சொல்சார் பங்குகளில் இருக்கும் ஒற்றுமைகள் மற்றும் வேறுபாடுகளை ஆராய்கிறது. மெக்னெரி மற்றும் வில்சன் (Mcenery & Wilson 1996) பிரவுன் தரவுத்தொகுதி மற்றும் லோப் தரவுத்தொகுதி (LOB Corpus) ஆகியவற்றுக்கு இடையேயான அடிப்படை வேறுபாடுகளின் சிறந்த இடங்களைக் கண்டுபிடிக்க பல புள்ளிவிவரக் கணக்கீட்டு முறைகளைப் பயன்படுத்துகின்றனர். ஜு (Xu, 1996) பல்வேறு

பாடப்பிரிவுகளைச் சேர்ந்த பல ஆங்கில தரவுத்தொகுதிகளில் பயன்படுத்தப்படும் ஆங்கில சொற்களின் சராசரி நீளத்தை அளவிட புள்ளிவிவரங்களையும் பயன்படுத்துகிறது. பைபர், கான்ராட் மற்றும் ரெப்பன் (Biber, Conrad & Reppen 1998) வெவ்வேறு மொழியியல் அலகுகள் நிகழும் அதிர்வெண்ணைக் கணக்கிட வெவ்வேறு புள்ளிவிவர முறைகளைப் பயன்படுத்தினர், இதில் ஆங்கிலத்தின் தரவுத்தொகுதியின் பல்வேறு வகையான உருபன்கள் மற்றும் சொற்கள் அடங்கும். இதேபோன்ற முயற்சியின் நன்மதிப்பு ஓக்ஸ்-க்கும் (Oakes 1998) செல்கிறது; அவர் ஆங்கிலம் மற்றும் ஜெர்மன் சொற்களின் இயல்பான நீளத்தைக் வார்த்தைகளில் பயன்படுத்தப்படும் எழுத்துக்களின் எண்ணிக்கை அடிப்படையில் கண்டுபிடிக்க ஆங்கிலம் மற்றும் ஜெர்மன் மொழிகளின் பல தரவுத்தொகுதிகளைப் பயன்படுத்துகிறார்.

சொல் அதிர்வெண் அடிப்படையில் பட்டியல்கள்

அதிர்வெண் அடிப்படையில் சொல் பட்டியல்கள் (Word lists by frequency) என்பது ஒரு மொழியின் சொற்களின் பட்டியல்கள், கொடுக்கப்பட்ட சில உரை தரவுத்தொகுதிக்குள் நிகழும் அதிர்வெண் மூலம், நிலைகள் அல்லது தரவரிசைப் பட்டியலாகச் சொற்றொகையை ஈட்டும் நோக்கத்திற்காக சேவை செய்யப்படுகின்றன. அதிர்வெண் அடிப்படையில் ஒரு சொல் பட்டியல் "கற்பவர்கள் தங்கள் சொற்றொகைக் கற்றல் முயற்சிக்குச் சிறந்த ஈட்டத்தைப் பெறுவதை உறுதி செய்வதற்கான பகுத்தறிவு அடிப்படையை வழங்குகிறது" (Nation 1997), ஆனால் இது முக்கியமாகப் பாட எழுத்தாளர்களுக்காக மட்டுமே கருதப்படுகிறது, இது நேரடியாக கற்பவர்களுக்கு அல்ல. அதிர்வெண் பட்டியல்கள் அகராதி நோக்கங்களுக்காகவும் உருவாக்கப்படுகின்றன; பொதுவான சொற்கள் விடபடாமல் இருப்பதை உறுதிசெய்ய ஒரு வகையான சரிபார்ப்பு பட்டியலாகச் செயல்படுகின்றன. தரவுத்தொகுதி உள்ளடக்கம், தரவுத்தொகுதி பதிவு மற்றும் "சொல்" என்பதன் வரையறை ஆகியவை சில முக்கிய ஆபத்துகள் ஆகும். சொல் எண்ணுதல் ஆயிரம் ஆண்டுகள் பழமையானது, 20ஆம் நூற்றாண்டின் நடுப்பகுதியில் இன்னும் பிரம்மாண்டமான பகுப்பாய்வு செய்யப்பட்டு, திரைப்பட வசன வரிகள் (SUBTLEX megastudy) போன்ற பெரிய தரவுத்தொகுதிகளின் இயற்கை மொழி மின்னணு செயலாக்கம் (natural language electronic processing) ஆராய்ச்சித் துறையைத் துரிதப்படுத்தியுள்ளது.

கணினி மொழியியலில், ஒரு அதிர்வெண் பட்டியல் என்பது அவற்றின் அதிர்வெண்ணுடன் ஒரு வரிசைப்படுத்தப்பட்ட சொற்களின் பட்டியல் (சொல் வகைகள்), இங்கு அதிர்வெண் என்பது

வழக்கமாக கொடுக்கப்பட்ட தரவுத்தொகுதியில் நிகழ்வுகளின் எண்ணிக்கையைக் குறிக்கிறது, இதிலிருந்து தரவரிசை பட்டியலில் உள்ள இடமாகத் தகுதிநிலையை ஆக்க இயலும்.

6.4.7. சொல் தேடு பொறி (Item-search engine)

ஒரு தேடுபொறி என்பது கணினி கணினியில் சேமிக்கப்பட்ட தகவல்களைக் கண்டறிய உதவும் வகையில் வடிவமைக்கப்பட்ட தகவல் மீட்டெடுப்பு ஒழுங்குமுறை (information retrieval system) ஆகும். தேடல் முடிவுகள் பொதுவாக ஒரு பட்டியலில் வழங்கப்படுகின்றன, அவை பொதுவாக வெற்றிகள் என்று அழைக்கப்படுகின்றன. தகவல் சுமைகளை நிர்வகிப்பதற்கான பிற நுட்பங்களைப் போலவே, தகவலைக் கண்டுபிடிப்பதற்குத் தேவையான நேரத்தையும், கலந்தாலோசிக்க வேண்டிய தகவல்களின் அளவையும் குறைக்க தேடுபொறிகள் உதவுகின்றன. ஒரு தேடுபொறியின் மிகவும் பொது, புலப்படும் வடிவம் ஒரு வலை தேடுபொறி (Web search engine) ஆகும்; இது உலகளாவிய வலையில் (World Wide Web) தகவல்களைத் தேடுகிறது.

தேடுபொறிகள் எவ்வாறு செயல்படுகின்றன

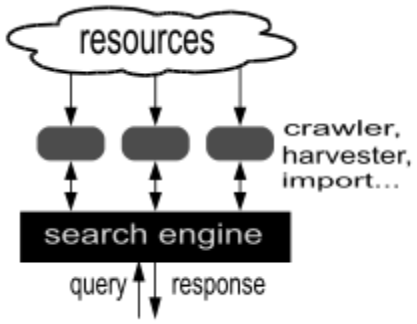
தேடுபொறிகள் ஒரு குழுவிற்கு ஒரு இடைமுகத்தை வழங்குகின்றன; இது பயனர்களுக்கு ஆர்வமுள்ள ஒரு ஐட்டங்களைப்/பொருளைப் பற்றிய அளவுகோல்களைக் குறிப்பிட உதவுகிறது மற்றும் பொருந்தக்கூடிய பொருள்களை இயந்திரம் கண்டறியும். அளவுகோல்கள் தேடல் வினவலாக (search query) குறிப்பிடப்படுகின்றன. உரைத் தேடுபொறிகளின் (text search engines) விஷயத்தில், தேடல் வினவல் பொதுவாக ஒன்று அல்லது அதற்கு மேற்பட்ட ஆவணங்களைக் கொண்டிருக்கக்கூடிய விரும்பிய கருத்தை அடையாளம் காணும் சொற்களின் தொகுப்பாக வெளிப்படுத்தப்படுகிறது. கண்டிப்பாக வேறுபடுகின்ற தேடல் வினவல் தொடரியலின் பல நடைகள்/பாணிகள் உள்ளன. இது முந்தைய தளங்களிலிருந்து தேடுபொறிகளில் பெயர்களை மாற்றலாம். சில உரை தேடுபொறிகள் பயனர்கள் வெள்ளை இடத்தால் பிரிக்கப்பட்ட இரண்டு அல்லது மூன்று சொற்களை உள்ளிட வேண்டும் என்றாலும், பிற தேடுபொறிகள் பயனர்கள் முழு ஆவணங்கள், படங்கள், ஒலிகள் மற்றும் பல்வேறு வகையான இயற்கை மொழிகளைக் குறிப்பிட உதவும். வினவல் விரிவாக்கம் (Query expansion (QE)) எனப்படும் ஒரு செயல்முறையின் மூலம் தரமான பொருட்களின் தொகுப்பை வழங்குவதற்கான வாய்ப்பை அதிகரிக்கச் சில தேடுபொறிகள் தேடல் வினவல்களுக்கு மேம்பாடுகளைப் பயன்படுத்துகின்றன. வினவல் புரிந்துகொள்ளும் (Query understanding) முறைகளைத் தரப்படுத்தப்பட்ட வினவல் மொழியாகப் பயன்படுத்தலாம்.

வினவலால் குறிப்பிடப்பட்ட அளவுகோல்களை பூர்த்தி செய்யும் பொருட்களின் பட்டியல் (list of items) பொதுவாக வரிசைப்படுத்தப்படுகிறது அல்லது தரப்படுத்தப்படுகிறது. பொருள்களைத் தரவரிசைப்படுத்துதல் (மிக உயர்ந்தது முதல் மிகக் குறைவானது) விரும்பிய தகவலைக் கண்டுபிடிக்க தேவையான நேரத்தைக் குறைக்கிறது. நிகழ்தகவு தேடுபொறிகள் (probabilistic search engines) ஒற்றுமையின் அளவீடுகளின் அடிப்படையில் பொருள்களை வரிசைப்படுத்துகின்றன (ஒவ்வொரு ஐட்டத்திக்கும் வினவலுக்கும் இடையில், பொதுவாக 1 முதல் 0 வரையிலான அளவில், 1 மிகவும் ஒத்ததாக இருக்கும்) மற்றும் சில நேரங்களில் புகழ் அல்லது அதிகாரம் (பிப்லியோமெட்ரிக்ஸைப் (Bibliometrics) பார்க்கவும்) அல்லது பொருத்தமான பின்னூட்டத்தைப் பயன்படுத்தும். பூலியன் தேடுபொறிகள் பொதுவாக ஒழுங்கைப் பொருட்படுத்தாமல் பொருந்தக்கூடிய ஐடங்களை மட்டுமே திருப்பித் தருகின்றன, இருப்பினும் பூலியன் தேடுபொறி என்ற சொல் பூலியன் பாணி தொடரியல் (ஆபரேட்டர்களின் பயன்பாடு AND, OR, NOT, XOR) ஒரு நிகழ்தகவு சூழலில் குறிப்பிடலாம்.

சில அளவுகோல்களின்படி விரைவாக வரிசைப்படுத்தப்பட்ட பொருந்தக்கூடிய ஐடங்களின் தொகுப்பை வழங்க, ஒரு தேடுபொறி பொதுவாக அட்டவணையிடல் என குறிப்பிடப்படும் ஒரு செயல்முறையின் மூலம் பரிசீலிக்கப்படும் பொருட்களின் குழு பற்றிய மெட்டாடேட்டாவை சேகரிக்கும். குறியீட்டுக்கு பொதுவாக சிறிய அளவிலான கணினி சேமிப்பிடம் தேவைப்படுகிறது, அதனால்தான் சில தேடுபொறிகள் குறியிடப்பட்ட தகவல்களை மட்டுமே சேமித்து வைக்கின்றன, ஒவ்வொரு பொருளின் முழு உள்ளடக்கத்தையும் அல்ல, அதற்குப் பதிலாக தேடுபொறி முடிவு பக்கத்தில் உள்ள ஐட்டங்களுக்கு செல்ல ஒரு முறையை வழங்குகின்றன. மாற்றாக, தேடுபொறி ஒவ்வொரு ஐட்டத்தின் நகலையும் ஒரு தற்காலிக சேமிப்பில் சேமிக்கக்கூடும், இதன் மூலம் பயனர்கள் ஐட்டத்தின் குறியீட்டு நேரத்தில் அல்லது காப்பக நோக்கங்களுக்காக அல்லது மீண்டும் மீண்டும் வரும் செயல்முறைகளை மிகவும் திறமையாகவும் விரைவாகவும் செயல்படுத்த முடியும்.

பிற வகை தேடுபொறிகள் ஒரு குறியீட்டை சேமிக்காது. கிராலர் (Crawler), அல்லது சிலந்தி வகை தேடுபொறிகள் (spider type search engines) (a.k.a. நிகழ்நேர தேடுபொறிகள்) தேடல் வினவலின் போது ஐட்டங்களைச் சேகரித்து மதிப்பிடலாம், தொடக்க உருப்படியின் உள்ளடக்கங்களை அடிப்படையாகக் கொண்ட கூடுதல் ஐடங்களை மாறும் வகையில் கருத்தில் கொள்ளலாம் (ஒரு விதை அல்லது விதை URL என அழைக்கப்படுகிறது இணைய கிராலரின்

நேர்வில்). மெட்டா தேடுபொறிகள் (Meta search engines) ஒரு குறியீட்டையோ அல்லது தற்காலிக சேமிப்பையோ சேமிக்கவில்லை, அதற்குப் பதிலாக ஒன்று அல்லது அதற்கு மேற்பட்ட தேடுபொறிகளின் குறியீட்டை அல்லது முடிவுகளை மறுபயன்படுத்தி, ஒருங்கிணைந்த, இறுதி முடிவுகளை வழங்குகின்றன. (கீழ்வரும் படமும் இத்தலைப்பில் கீழ் தந்துள்ள செய்திகளும் விக்கிபீடியாவிலிருது எடுத்தாளப் படுகின்றது.)



தேடுபொறிகளின் வகைகள்

மூலமாக

- டெஸ்க்டாப் தேடல் (Desktop search)
- கூட்டாட்சி தேடல் (Federated search)
- மனித தேடுபொறி (Human search engine)
- மெட்டாசர்ச் இயந்திரம் (Metasearch engine)
- பன்முகத் தேடல் (Multisearch)
- வலை தேடுபொறி Web search engine

உள்ளடக்க வகை மூலம்

- ஆடியோ தேடுபொறி (Audio search engine)
- முழு உரை தேடல் (Full text search)
- படத் தேடல் (Image search)
- வீடியோ தேடுபொறி (Video search engine)

இடைமுகத்தால்

- அதிகரிக்கும் தேடல் (Incremental search)
- உடனடி பதில் (Instant answer)
- சொற்பொருள் தேடல் (Semantic search)

- தேர்வு அடிப்படையிலான தேடல் (Selection-based search)
- குரல் தேடல் (Selection-based search)

தலைப்பு மூலம்

- நூலியல் தரவுத்தளம் (Bibliographic database)
- நிறுவன தேடல் (Enterprise search)
- மருத்துவ இலக்கியம் மீட்டெடுப்பு (Medical literature retrieval)
- செங்குத்து தேடல் (Vertical search)

6.4.8. உரை சுருக்கும் ஒழுங்குமுறை (Text summarisation system)

தானியங்கி உரைச் சுருக்கம் என்றால் என்ன?

தானியங்கி உரைச் சுருக்கம் அல்லது உரைச் சுருக்கம் என்பது ஒரு நீண்ட ஆவணத்தின் குறுகிய மற்றும் ஒத்திசைவான பதிப்பை உருவாக்கும் செயல்முறையாகும். உரைச் சுருக்கம் என்பது ஒரு குறிப்பிட்ட பயனர் (அல்லது பயனர்கள்) மற்றும் பணி (அல்லது பணிகள்) ஆகியவற்றுக்கான சுருக்கப்பட்ட பதிப்பை உருவாக்க ஒரு மூலத்திலிருந்து (அல்லது மூலங்களிலிருந்து) மிக முக்கியமான தகவல்களை வடிகட்டுவதற்கான செயல்முறையாகும். நாம் (மனிதர்கள்) பொதுவாக இந்த வகை பணியில் சிறந்தவர்கள், ஏனெனில் இது முதலில் மூல ஆவணத்தின் பொருளைப் புரிந்துகொள்வதும் பின்னர் பொருளை வடிகட்டுவதும் புதிய விவரத்தில் முக்கிய விவரங்களைக் கைப்பற்றுவதும் அடங்கும். எனவே, உரையின் சுருக்கங்களைத் தானாக உருவாக்குவதன் குறிக்கோள், இதன் விளைவாக வரும் சுருக்கங்களை மனிதர்களால் எழுதப்பட்டதைப் போலவே சிறந்ததாகப் பெறுவது ஆகும். மனிதர்களால் உருவாக்கப்பட்ட சுருக்கங்களை வெற்றிகரமாகப் பின்பற்றும் சுருக்கத்தை ஒரு இயந்திரம் உருவாக்கக்கூடிய நுட்பங்களை உருவாக்குவதே தானியங்கி சுருக்கமயமாக்கல் பணியின் சிறந்தது. மூல ஆவணத்தின் சாராம்சத்தைக் கைப்பற்றும் சொற்களையும் சொற்றொடர்களையும் உருவாக்குவது மட்டும் போதாது சுருக்கம் துல்லியமாக இருக்க வேண்டும் மற்றும் ஒரு புதிய முழுமையான ஆவணமாகச் சரளமாகப் படிக்கப்பட வேண்டும். தானியங்கு உரைச் சுருக்கம் என்பது முக்கிய தகவல் உள்ளடக்கம் மற்றும் ஒட்டுமொத்த பொருளைப் பாதுகாக்கும் போது சுருக்கமான மற்றும் சரளமான சுருக்கத்தை உருவாக்கும் பணியாகும்.

உரை சுருக்கம் என்பது ஒரு நீண்ட உரை ஆவணத்தின் குறுகிய, துல்லியமான மற்றும் சரளமான சுருக்கத்தை உருவாக்குவதில் உள்ள சிக்கலாகும். ஆன்லைனில் கிடைக்கக்கூடிய

உரைத் தரவின் எப்போதும் வளந்துகொண்டிருக்கும் அளவைத் நேரிட தானியங்கி உரைச் சுருக்கம் முறைகள் பெரிதும் தேவைப்படுகின்றன, அவை தொடர்புடைய தகவல்களைக் கண்டுபிடிப்பதற்கும் தொடர்புடைய தகவல்களை விரைவாகப் பயன்படுத்துவதற்கும் உதவுகின்றன. உரை சுருக்கம் ஏன் முக்கியமானது, குறிப்பாக இணையத்தில் கிடைக்கும் உரையின் செல்வத்தைப் பெற கைகொடுக்கும்.

ஏராளமான உரை பொருள் உள்ளது, அது ஒவ்வொரு நாளும் மட்டுமே வளர்ந்து வருகிறது. வலைப்பக்கங்கள், செய்தி கட்டுரைகள், நிலை புதுப்பிப்புகள், வலைப்பதிவுகள் மற்றும் பலவற்றை உள்ளடக்கிய இணையத்தைப் பற்றி சிந்தியுங்கள். தரவு கட்டமைக்கப்படாதது மற்றும் அதைத் தொடர நாம் செய்யக்கூடியது தேடலைப் பயன்படுத்துவதும் முடிவுகளைத் தவிர்ப்பதும் ஆகும். இந்த உரைத் தரவின் பெரும்பகுதியை குறுகிய, கவனம் செலுத்திய சுருக்கங்களாகக் குறைக்க வேண்டிய அவசியம் உள்ளது, இவை இரண்டும் முக்கிய விவரங்களைக் பெறுகின்றன; எனவே இரண்டையும் நாம் மிகவும் திறம்பட வழிநடத்தலாம்; மேலும் பெரிய ஆவணங்களில் நாம் தேடும் தகவல்கள் உள்ளதா என்பதை சரிபார்க்கலாம். டிஜிட்டல் ஆவணங்களின் வடிவத்தில் உள்ள உரை தகவல்கள் விரைவாக பெரிய அளவிலான தரவைக் குவிக்கின்றன. இந்த பெரிய அளவிலான ஆவணங்கள் கட்டமைக்கப்படாதவை: இது கட்டுப்பாடற்றது மற்றும் பாரம்பரிய தரவுத்தளங்களாக ஒழுங்கமைக்கப்படவில்லை. ஆகவே ஆவணங்களைச் செயலாக்குவது ஒரு செயலற்ற பணியாகும், பெரும்பாலும் தரநிலைகள் இல்லாததால். எல்லா உரையின் சுருக்கங்களையும் கைமுறையாக உருவாக்க முடியாது; தானியங்கி முறைகளுக்கு ஒரு பெரிய தேவை உள்ளது.

தானியங்கி உரை சுருக்க கருவிகள் தேவைப்படுவதற்கான காரணங்கள்

ஜுவான்-மானுவல் டோரஸ்-மோரேனோ என்போர் “தானியங்கி உரை சுருக்கம்” (Automatic Text Summarization) (Juan-Manuel Torres-Moreno, 2014) என்ற தமது நூலில் நமக்கு தானியங்கி உரை சுருக்க கருவிகள் தேவைப்படுவதற்கான 6 காரணங்களை வழங்குகிறார்கள்.

1. சுருக்கங்கள் வாசிப்பு நேரத்தைக் குறைக்கின்றன.
2. ஆவணங்களை ஆராய்ச்சி செய்யும் போது, சுருக்கங்கள் தேர்வு செயல்முறையை எளிதாக்குகின்றன.
3. தானியங்கிச் சுருக்கம் குறியீட்டு செயல்திறனை மேம்படுத்துகிறது.

4. தானியங்குச் சுருக்கமயமாக்கல் வழிமுறைகள் மனிதச் சுருக்கங்களை விட குறைவான சார்புடையவை.
5. தனிப்பயனாக்கப்பட்ட தகவல்களை வழங்குவதால் கேள்விக்கு பதிலளிக்கும் அமைப்புகளில் தனிப்பயனாக்கப்பட்ட சுருக்கங்கள் பயனுள்ளதாக இருக்கும்.
6. தானியங்கி அல்லது பகுதி தானியங்கிச் சுருக்க ஒழுங்குமுறைகளைப் பயன்படுத்துவது வணிகச் சுருக்க சேவைகளை அவர்கள் வாங்கக்கூடிய நூல்களின் எண்ணிக்கையை அதிகரிக்க உதவுகிறது.

உரை சுருக்கங்களின் எடுத்துக்காட்டுகள்

ஒரு பெரிய ஆவணத்தின் சுருக்கத்திற்கு பல காரணங்களும் பயன்பாடுகளும் உள்ளன. மனதில் உடனடியாக வரக்கூடிய ஒரு எடுத்துக்காட்டு, ஒரு நீண்ட செய்தி கட்டுரையின் சுருக்கமான சுருக்கத்தை உருவாக்குவது, ஆனால் இன்னும் பல உரை சுருக்கங்கள் ஒவ்வொரு நாளும் நாம் வரக்கூடும். 1999 ஆம் ஆண்டின் "தானியங்கி உரை சுருக்கத்தில் முன்னேற்றங்கள்" என்ற தலைப்பில் அவர்களின் புத்தகத்தில், ஆசிரியர்கள் உரைச் சுருக்கத்தின் ஒவ்வொரு நாளும் எடுத்துக்காட்டுகளின் பயனுள்ள பட்டியலை வழங்குகிறார்கள்.

- தலைப்புச் செய்திகள் (headlines) (உலகம் முழுவதிலுமிருந்து)
- சுருக்க உரை (outlines) (மாணவர்களுக்கான குறிப்புகள்)
- நிகழ்ச்சிக்குறிப்புகள் (minutes) (கூட்டத்தின்)
- முன்னோட்டங்கள் (previews) (திரைப்படங்களின்)
- விஷயச்சுருக்கங்கள் (synopses) (சோப் ஓபரா பட்டியல்கள்)
- மதிப்புரைகள் reviews (ஒரு புத்தகம், குறுவட்டு, திரைப்படம் போன்றவை)
- சுருக்கத் தொகுப்புகள் digests (டிவி வழிகாட்டி)
- சுயசரிதை (biography) (விண்ணப்பங்கள், இரங்கல்கள்)
- சுருக்கங்கள் (abridgments) (குழந்தைகளுக்கான ஷேக்ஸ்பியர்)
- புல்லட்டின் bulletins (வானிலை முன்னறிவிப்புகள் / பங்குச் சந்தை அறிக்கைகள்)
- சாராம்சம் (sound bites) (தற்போதைய பிரச்சினையில் அரசியல்வாதிகள்)
- வரலாறுகள் (histories) (முக்கிய நிகழ்வுகளின் காலவரிசை)

உரை சுருக்கத்திற்கான முக்கிய படிகள்:

ஆவணங்களை சுருக்கமாக மூன்று முக்கிய படிகள் உள்ளன. இவை தலைப்பு அடையாளம் காணல், விளக்கம் மற்றும் சுருக்க உருவாக்கம்.

1. தலைப்பு அடையாளம் காணல்: உரையில் மிக முக்கியமான தகவல்கள் அடையாளம் காணப்படுகின்றன. தொகுப்பு அடையாளம் காண வெவ்வேறு நுட்பங்கள் பயன்படுத்தப்படுகின்றன; அவை நிலை, குறிபுச் சொற்றொடர்கள், சொல் அதிர்வெண். சொற்றொடர்களின் நிலையை அடிப்படையாகக் கொண்ட முறைகள் தலைப்பு அடையாளம் காண மிகவும் பயனுள்ள முறைகள்.

2. விளக்கம்: சுருக்கம் சுருக்கங்கள் விளக்க படி வழியாக செல்ல வேண்டும். இந்த கட்டத்தில், பொதுவான உள்ளடக்கத்தை உருவாக்குவதற்காக வெவ்வேறு பாடங்கள்/பொருண்மைகள் இணைக்கப்படுகின்றன.

3.3. சுருக்கம் உருவாக்கம்: இந்த கட்டத்தில், கணினி உரை உருவாக்கும் முறையைப் பயன்படுத்துகிறது.

அணுகுமுறைகள் (Approaches)

. தானியங்கி உரை சுருக்கம் ஒரு பழைய சவால்; ஆனால் தற்போதைய ஆராய்ச்சி திசையானது பயோமெடிசின், தயாரிப்பு மதிப்பாய்வு, கல்வி களங்கள், மின்னஞ்சல்கள் மற்றும் வலைப்பதிவுகள் ஆகியவற்றில் வளர்ந்து வரும் போக்குகளை நோக்கித் திசை திருப்புகிறது. இந்த பகுதிகளில், குறிப்பாக உலகளாவிய வலையில் தகவல் சுமை இருப்பதே இதற்குக் காரணம். இயற்கை மொழி ஆய்வு (என்.எல்.பி) ஆராய்ச்சியில் தன்னியக்க சுருக்கம் ஒரு முக்கியமான பகுதியாகும். இது ஒன்று அல்லது அதற்கு மேற்பட்ட உரைகளின் சுருக்கத்தை தானாக உருவாக்குவதைக் கொண்டுள்ளது. பிரித்தெடுக்கும் ஆவணச் சுருக்கத்தின் நோக்கம் அசல் ஆவணத்திலிருந்து பல கூற்று வாக்கியங்கள், பத்திகளை அல்லது பத்திகளை தானாகவே தேர்ந்தெடுப்பதாகும் .நியூரல் நெட்வொர்க் (Neural Network), வரைபடக் கோட்பாடு (Graph Theoretic), தெளிவில்லாத (Fuzzy) மற்றும் கிளஸ்டரை (Cluster) அடிப்படையாகக் கொண்ட உரை சுருக்கம் அணுகுமுறைகள் ஒரு அளவிற்கு திறம்பட செயல்படுவதில் வெற்றி பெற்றுள்ளன. ஒரு ஆவணத்தின் சுருக்கம். பிரித்தெடுக்கும் மற்றும் சுருக்க முறைகள் (extractive and abstractive methods) இரண்டும் ஆராய்ச்சி செய்யப்பட்டுள்ளன. பெரும்பாலான சுருக்க நுட்பங்கள் பிரித்தெடுக்கும் முறைகளை அடிப்படையாகக் கொண்டவை. சுருக்க முறை என்பது மனிதர்களால் செய்யப்பட்ட சுருக்கங்களுக்கு ஒத்ததாகும். இப்போது சுருக்கமான சுருக்கத்திற்கு மொழி

உருவாக்கத்திற்கு கனரக இயந்திரங்கள் தேவைப்படுகின்றன, மேலும் டொமைன் குறிப்பிட்ட பகுதிகளுக்குள் நகலெடுப்பது கடினம்.

தானியங்கி சுருக்கத்திற்கு இரண்டு பொதுவான அணுகுமுறைகள் உள்ளன: பிரித்தெடுத்தல் மற்றும் சுருக்கம்.

பிரித்தெடுத்தல் அடிப்படையிலான சுருக்கம் (Extraction-based summarization)

இங்கே, அசல் தரவிலிருந்து உள்ளடக்கம் பிரித்தெடுக்கப்படுகிறது, ஆனால் பிரித்தெடுக்கப்பட்ட உள்ளடக்கம் எந்த வகையிலும் மாற்றப்படவில்லை. பிரித்தெடுக்கப்பட்ட உள்ளடக்கத்தின் எடுத்துக்காட்டுகள் ஒரு உரை ஆவணத்தை "அடையாளம்" செய்ய அல்லது அகரவரிசை அட்டவணைப்படுத்தப் பயன்படுத்தக்கூடிய முக்கியச் சொற்றொடர்கள் அல்லது மேலே கூறப்பட்டுள்ளபடி ஒரு சுருக்கத்தை கூட்டாக உள்ளடக்கும் முக்கிய வாக்கியங்கள் (தலைப்புகள் உட்பட) மற்றும் பிரதிநிதி படங்கள் அல்லது வீடியோ பிரிவுகள் ஆகியவற்றை உள்ளடக்கும். உரையைப் பொறுத்தவரை, பிரித்தெடுத்தல் சறுக்குதல் செயல்முறைக்கு ஒத்ததாக இருக்கிறது, இதில் சுருக்கம் (கிடைத்தால்), தலைப்புகள் மற்றும் துணை தலைப்புகள், புள்ளிவிவரங்கள், ஒரு பிரிவின் முதல் மற்றும் கடைசி பத்திகள் மற்றும் விருப்பமாக ஒரு பத்தியில் முதல் மற்றும் கடைசி வாக்கியங்கள் ஒருவர் முழு ஆவணத்தையும் விரிவாக படிக்க தேர்ந்தெடுப்பதற்கு முன்பு படிக்கப்படுகின்றன.

சுருக்கம் சார்ந்த சுருக்கம் (Abstraction-based summarization)

இது முக்கியமாக உரைக்கு பயன்படுத்தப்பட்டது. சுருக்க முறைகள் (Abstractive methods) அசல் உள்ளடக்கத்தின் உள் பொருண்மையியல்சார் பிரதிநிதித்துவத்தை உருவாக்குகின்றன, பின்னர் இந்த பிரதிநிதித்துவத்தைப் பயன்படுத்தி ஒரு மனிதன் வெளிப்படுத்தக் கூடியவற்றுடன் நெருக்கமான ஒரு சுருக்கத்தை உருவாக்கும். ஒரு உரையைப் பிரித்தெடுப்பதை விட வலுவாகச் சுருக்க, சுருக்கம் பிரித்தெடுக்கப்பட்ட உள்ளடக்கத்தை மூல ஆவணத்தின் பகுதிகளை பொழிப்புரை செய்வதன் மூலம் மாற்றலாம். இருப்பினும் இத்தகைய மாற்றம், பிரித்தெடுப்பதை விட கணினிசார் அடிப்படையில் மிகவும் சவாலானது; இது இயற்கையான மொழி ஆய்வு (natural language processing) மற்றும் அசல் ஆவணம் ஒரு சிறப்பு அறிவுத் துறையுடன் தொடர்புடைய சந்தர்ப்பங்களில் அசல் உரையின் களத்தைப் பற்றிய ஆழமான புரிதல் ஆகிய இரண்டையும் உள்ளடக்கியது. "பொழிப்புரை/பெயர்ப்புரை"-ஐப் படம் மற்றும் வீடியோவுக்குப் பயன்படுத்துவது

இன்னும் கடினம்; அதனால்தான் பெரும்பாலான சுருக்கமயமாக்கல் அமைப்புகள் பிரித்தெடுத்தலாக (extractive) அமைகின்றன.

உதவிச் சுருக்கம் (Aided summarization)

அதிக சுருக்க தரத்தை நோக்கமாகக் கொண்ட அணுகுமுறைகள் ஒருங்கிணைந்த மென்பொருள் மற்றும் மனித முயற்சியை நம்பியுள்ளன. இயந்திர உதவி மனித சுருக்கத்தில் (Machine Aided Human Summarization), பிரித்தெடுக்கும் நுட்பங்கள் உட்படுத்துவதற்காக முக்கிய பத்திகளை முன்னிலைப்படுத்துகின்றன (இதில் மனிதன் உரையைச் சேர்க்க்வோ நீக்கவோ செய்வான்). மனித உதவி இயந்திர சுருக்கத்தில் (Human Aided Machine Summarization) ஒரு மனிதப் பின் செயலாக்க மென்பொருள் (human post-processes software) கூகிள் மொழிபெயர்ப்பின் தானியங்கி மொழிபெயர்ப்பின் வெளியீட்டை ஒருவர் திருத்துவதைப் போல வெளியீடு செய்கிறது.

தரவுத்தொகுதி அடிப்படியிலான உரைச்சுருக்க ஒழுங்குமுறை

தரவுத்தொகுதி அடிப்படியிலான உரைச்சுருக்க ஒழுங்குமுறை பல உருவாக்கப்பட்டுள்ளன.

பார்த்தா லால் உருவாக்கப்பட்ட உரை சுருக்க ஒழுங்குமுறை

பார்த்தா லால் (Partha Lal, 2002) தரவுதொகுதியைப் பயன்படுத்தி உருவக்கக்கப்பட உரை சுருக்க ஒழுங்குமுறை குறிப்பிடத் தகுந்ததாகும். இந்த திட்டத்தில், உரை சுருக்கம் ஒழுங்குமுறை ஒன்று உருவாக்கப்பட்டுள்ளது. சுருக்கமாக ஆவணத்தில் உள்ள வாக்கியங்களுக்கு மதிப்பெண்களை ஒதுக்குவதன் மூலமும், சுருக்கத்தில் அதிக மதிப்பெண் வாக்கியங்களைப் பயன்படுத்துவதன் மூலமும் கணினி செயல்படுகிறது. மதிப்பெண் மதிப்புகள் வாக்கியத்திலிருந்து பிரித்தெடுக்கப்பட்ட அம்சங்களை அடிப்படையாகக் கொண்டவை. அம்ச மதிப்பெண்களின் நேரியல் கலவை (linear combination) பயன்படுத்தப்படுகிறது. அம்சத்திலிருந்து மதிப்பெண் வரையிலான கிட்டத்தட்ட அனைத்து மேப்பிங் மற்றும் நேரியல் கலவையில் உள்ள குணக மதிப்புகள் (coecient values) ஒரு பயிற்சித் தரவுத்தொகுதியிலிருந்து பெறப்படுகின்றன. சில முற்சுட்டு தீர்மானம் (anaphor resolution) செய்யப்படுகிறது. இந்த ஒழுங்குமுறையை மதிப்பீடு செய்ய ஆவண புரிந்துணர்வு மாநாட்டில் (Document Understanding Conference) சமர்ப்பிக்கப்பட்டது. அடிப்படை சுருக்கத்திற்குக் கூடுதலாக, பயனரை உரையைக் குறிவைக்கும் சிக்கலை தீர்க்க சில முயற்சிகள் மேற்கொள்ளப்படுகின்றன. நோக்கம் கொண்ட பயனருக்கு பின்னணி அறிவு அல்லது வாசிப்பு திறன் குறைவாக இருப்பதாகக் கருதப்படுகிறது. சுருக்கத்தில்

பயன்படுத்தப்படும் தனிப்பட்ட சொற்களை எளிதாக்குவதன் மூலமும், வலையில் (web) இருந்து தேவையான பின்னணித் தகவல்களை பெறுவதன் மூலமும் கணினி உதவுகிறது.

6.4.9. உரை அடையாளப்படுத்தும் ஒழுங்குமுறை (Text annotation system)

தரவுத்தொகுதி அடையாளப்படுத்தல்/சிறுகுறிப்புசெய்தல் என்பது ஒரு தரவுத்தொகுதியில் விளக்க மொழியியல் தகவல்களைச் சேர்ப்பது. எடுத்துக்காட்டாக, ஒரு பொதுவான வகை அடையாளப்படுத்தல்/சிறுகுறிப்புசெய்தல் என்பது அடையாளங்களை/குறிச்சொற்களை அல்லது லேபிள்களைச் சேர்ப்பதாகும்; இது ஒரு உரையில் உள்ள சொற்கள் எந்த வகைப்பாட்டைச்/வகுப்பைச் சேர்ந்தது என்பதைக் குறிக்கிறது. இது சொல்வகைப்பாடு அடையாளப்படுத்தல் (part-of-speech tagging (அல்லது POS tagging/ பிஓஎஸ் டேக்கிங்) என்று அழைக்கப்படுகிறது; எடுத்துக்காட்டாக, ஒரே எழுத்துக்களைக் ஆனால் வெவ்வேறு அர்த்தங்கள் அல்லது உச்சரிப்பு கொண்ட சொற்களை வேறுபடுத்துவதில் பயனுள்ளதாக இருக்கும். ஒரு ஆங்கில உரையில் ஒரு சொல் present உச்சரிக்கப்பட்டால், அது ஒரு பெயர்ச்சொல்லாகவோ (= 'gift'), ஒரு வினைச்சொல்லாகவோ (=give someone a present) அல்லது பெயரடையாகவோ (='not absent') இருக்க இயலும். ஒரே மாதிரியாகத் தோன்றும் இந்த சொற்களின் அர்த்தங்கள் மிகவும் வேறுபட்டவை; மேலும் உச்சரிப்பின் வித்தியாசமும் உள்ளது; ஏனெனில் present என்ற வ்வினைச்சொல் இறுதி எழுத்துக்களில் அழுத்தத்தைக் கொண்டுள்ளது. சொல்வகைப்பாடுகளை அடையாளப்படுத்தும்/குறிப்பிடும் ஒரு எளிய முறையைப் பயன்படுத்துதல் - அடிக்கோடிட்டுக் குறியீட்டால் சொற்களை குறிச்சொற்களை இணைத்தல் - இந்த மூன்று சொற்களையும் பின்வருமாறு அடையாளப்படுத்தும்/குறிப்பிடும்:

present_NN1 (ஒற்றை பொதுவான பெயர்ச்சொல்)

present_VVB (ஒரு சொற்பொருள் வினைச்சொல்லின் அடிப்படை வடிவம்)

present_JJ (பொது பெயரடை)

உரையை அடையாளப்படுத்துவதன் முக்கியத்துவம்

எதையும் பற்றி கற்பிக்கக்கூடிய அனைத்து வகையான தகவல்களுக்கும் இணையம் ஒரு அற்புதமான ஆதாரம் என்பது அனைவருக்கும் தெரியும்: வித்தைகாட்டல், நிரலாக்கம், ஒரு கருவியை வாசித்தல் மற்றும் பல. இருப்பினும், இணையத்தில் உள்ள மற்றொரு அடுக்கு தகவல் உள்ளது; எனவேதான் அந்த விஷயங்கள் அனைத்தும் (மற்றும் வலைப்பதிவுகள், மன்றங்கள், ட்வீட்டுகள் போன்றவை) தொடர்பு கொள்ளப்படுகின்றன. உரைகள், படங்கள், திரைப்படங்கள் மற்றும் ஒலிகள் உட்பட அனைத்து வகையான ஊடகங்களிலும் இணையம் தகவல்களைக்

கொண்டுள்ளது மற்றும் மொழி என்பது தகவல்தொடர்பு ஊடகம்; இது உள்ளடக்கத்தைப் புரிந்துகொள்ளவும் உள்ளடக்கத்தை பிற ஊடகங்களுடன் இணைக்கவும் அனுமதிக்கிறது. இருப்பினும், ஆர்வமுள்ள பயனர்களுக்கு இந்த தகவலை வழங்குவதில் கணினிகள் சிறந்தவை என்றாலும், அவை மொழியைப் புரிந்து கொள்வதில் மிகவும் திறமையானவை.

கோட்பாட்டு மற்றும் கணினி மொழியியல் மொழியின் ஆழமான தன்மையை அவிழ்ப்பதில் கவனம் செலுத்துகிறது மற்றும் மொழியியல் கட்டமைப்புகளின் கணினிசார் பண்புகளை (computational properties) தனதாக்கிக்கொள்கிறது. மனித மொழி தொழில்நுட்பங்கள் (Human language technologies (HLTs/எச்.எல்.டி) இந்த நுண்ணறிவுகளையும் வழிமுறைகளையும் பின்பற்றி அவற்றைச் செயல்பாடாக, மொழியைப் பயன்படுத்தி கணினிகளுடன் நாம் தொடர்பு கொள்ளும் வழிகளை பாதிக்கும் உயர் செயல்திறன் கொண்ட திட்டங்களாக மாற்ற முயற்சிக்கின்றன. ஒவ்வொரு நாளும் அதிகமான மக்கள் இணையத்தைப் பயன்படுத்துவதால், ஆராய்ச்சியாளர்களுக்குக் கிடைக்கும் மொழியியல் தரவுகளின் அளவு கணிசமாக அதிகரித்துள்ளது; இது மொழியியல் மாடலிங் சிக்கல்களை மனிதர்கள் தாமாகவே செயலாக்கக்கூடிய சிறிய அளவிலான தரவுகளுக்கு மட்டுப்படுத்தப்படுத்தாமல் மாறாக இயந்திரம் கற்றல் பணிகளாகப் பார்க்க அனுமதிக்கிறது.

இருப்பினும், ஒரு கணினியை ஒரு பெரிய அளவிலான தரவை வழங்குவதும், பேசக் கற்றுக் கொள்ள வேண்டும் என்று எதிர்பார்ப்பதும் போதாது - கணினி வடிவங்கள் மற்றும் அனுமானங்களை எளிதில் கண்டுபிடிக்கும் வகையில் தரவைத் தயாரிக்க வேண்டும். தரவுத்தொகுப்பில் தொடர்புடைய மெட்டாடேட்டாவைச் சேர்ப்பதன் மூலம் இது வழக்கமாக செய்யப்படுகிறது. தரவுத்தொகுப்பின் கூறுகளைக் குறிக்கப் பயன்படுத்தப்படும் எந்த மெட்டாடேட்டா குறிச்சொல்லும்/அடையாளங்களும் உள்ளீட்டின் மீது சிறுகுறிப்பு/அடையாளப்படுத்தல் என அழைக்கப்படுகிறது. இருப்பினும், வழிமுறைகள் திறமையாகவும் திறம்படவும் கற்றுக்கொள்ள, தரவில் செய்யப்படும் சிறுகுறிப்பு/அடையாளப்படுத்தல் துல்லியமாக இருக்க வேண்டும், மேலும் இயந்திரம் செய்யும்படி கேட்கப்படும் பணிக்கு பொருத்தமானதாக இருக்க வேண்டும். இந்தக் காரணத்திற்காக, மொழி சிறுகுறிப்பின்/அடையாளப்படுத்தலின் ஒழுக்கம் அறிவார்ந்த மனித மொழி தொழில்நுட்பங்களை வளர்ப்பதில் ஒரு முக்கியமான இணைப்பாகும்.

மொழியியல் விளக்கத்தின் அடுக்குகள்

அடையாளப்படுத்தப்பட்ட (சிறுகுறிப்புசெய்யப்பட்ட (annotated)) தரவுத்தொகுதியை உருவாக்குவதற்கு முறையான மொழியியல் பயிற்சி அவசியம் இல்லை என்றாலும், பல வகையான சிறுகுறிப்புசெய்யும்/அடையாளப்படுத்தும் பணிகளின் எடுத்துக்காட்டுகளை அறிந்துகொள்ளவேண்டும்; மேலும் பயன்படுத்தப்படும் மொழிபற்றிய பல்வேறு அம்சங்களைப் பற்றிய அடிப்படை புரிதல் மொழிக்கூறுகளை அடையாளப்படுத்துவதற்கு மிகவும் உதவியாக இருக்கும். இலக்கணமானது மொழியில் நன்கு உருவாக்கப்பட்ட கட்டமைப்புகளை உருவாக்குவதற்குப் பொறுப்பான வழிமுறைகளுக்கு பொதுவாக வழங்கப்படும் பெயர். அறிவாற்றல் வடிவமைப்பு அல்லது விளக்க வசதிக்காக, பெரும்பாலான மொழியியலாளர்கள் இலக்கணத்தை தனித்துவமான தொகுதிகள் அல்லது அமைப்புகளைக் கொண்டதாகவே கருதுகின்றனர். இந்த பகுதிகளில் பொதுவாக தொடரியல், பொருண்மையியல், உருபனியல், ஒலியனியல் (மற்றும் ஒலியியல்) மற்றும் அகராதி ஆகியவை அடங்கும். மனித செயல்பாட்டில் மொழி எவ்வாறு உட்பொதிந்துள்ளது என்பதோடு தொடர்புடைய இலக்கணத்திற்கு அப்பாற்பட்ட பகுதிகள் கருத்தாடல், நடைமுறைவாதம்/பயன்வழியியல் மற்றும் உரைக் கோட்பாடு ஆகியவற்றில் அடங்கும். பின்வரும் பட்டியல் இந்த பகுதிகளின் விரிவான விளக்கங்களை வழங்குகிறது:

தொடரியல்

வாக்கியங்கள் எவ்வாறு சொற்களை இணைக்கின்றன என்ற ஆய்வு. சொல்வகைப்பாடுகளை ஆராய்வதும், அவை எவ்வாறு பெரிய கட்டுமானங்களை உருவாக்குகின்றன என்பதும் இதில் அடங்கும்.

பொருண்மையியல்

மொழியில் பொருள் பற்றிய ஆய்வு. பொருண்மையியல் சொற்களுக்கு இடையிலான உறவுகளையும் அவை பிரதிநிதித்துவப்படுத்தப் பயன்படுவதையும் ஆராய்கிறது.

உருபனியல்

ஒரு மொழியில் உள்ள மிகச்சிறிய பொருள் அலகுகளின் ஆய்வு. ஒரு உருபன் என்பது பொருள் அல்லது செயல்பாட்டைக் கொண்ட மொழியின் மிகச்சிறிய அலகு ஆகும், இது சொற்கள், முன்னொட்டுகள், இணைப்புகள் மற்றும் பொருளை வழங்கும் பிற சொல் கட்டமைப்புகளை உள்ளடக்கிய ஒரு வரையறை ஆகும்.

ஒலியனியல்

ஒரு குறிப்பிட்ட மொழியின் ஒலி அமைப்பொழுங்குகளின் ஆய்வு. எந்தெந்த ஒலிகள் குறிப்பிடத்தக்கவை மற்றும் பொருளைக் கொண்டிருக்கின்றன என்பதை தீர்மானிப்பது ஆய்வின் அம்சங்களில் அடங்கும் (அதாவது, ஒலியன்கள்): அசைகள் எவ்வாறு கட்டமைக்கப்பட்டு இணைக்கப்படுகின்றன; மற்றும் மொழியில் உள்ள தனித்தனி அலகுகளை (கூறுகளை) விவரிக்க என்ன அம்சங்கள் தேவை, அவை எவ்வாறு விளக்கப்படுகின்றன.

ஒலியியல்

மனித பேச்சின் ஒலிகளைப் பற்றிய ஆய்வு, அவை எவ்வாறு உருவாக்கப்படுகின்றன மற்றும் உணரப்படுகின்றன. ஒரு ஒலியன் என்பது ஒரு தனிப்பட்ட ஒலியைக் குறிப்பிடும் சொல்; இது அடிப்படையில் மனித பேச்சின் மிகச்சிறிய அலகு ஆகும்.

அகராதி/லெக்சிகன்

ஒரு மொழியில் பயன்படுத்தப்படும் சொற்கள் மற்றும் சொற்றொடர்களின் ஆய்வு, அதாவது ஒரு மொழியின் சொல்லகராதி.

கருத்தாடல் பகுப்பாய்வு

தகவல் பரிமாற்றங்களைப் பற்றிய ஆய்வு; வழக்கமாக உரையாடல்களின் வடிவத்தில், குறிப்பாக வாக்கிய எல்லைகளில் தகவல்களின் ஒழுக்கு.

பயன்வழியியல்/நடைமுறைவாதம்

உரையின் சூழல் ஒரு வெளிப்பாட்டின் பொருளை எவ்வாறு பாதிக்கிறது என்பதையும், மறைக்கப்பட்ட அல்லது முன்னறிவிக்கப்பட்ட பொருளை ஊகிக்க என்ன தகவல் அவசியம் என்பதையும் ஆய்வு செய்கிறது.

உரைக் கட்டமைப்புப் பகுப்பாய்வு

பெரிய உரை அமைப்புகளை உருவாக்க விவரிப்புகள் மற்றும் பிற உரை பாணிகள்/நடைகள் எவ்வாறு கட்டமைக்கப்படுகின்றன என்பது பற்றிய ஆய்வு.

வேறுபட்ட வகையான அடையாளப்படுத்தல்கள் (Different kinds of annotation)

முன்னர் பார்த்த்த மொழியியல் விளக்கத்தின் அடுக்குகளான/நிலைகளான ஒலியியல், ஒலியனியல், உருபனியல், தொடரியல், பொருண்மையியல், கடுத்தாடல் ஆகியவற்றிற்கு இணங்க வேறுபட்ட அடையாளப்படுத்தல்கள் உரையில் அல்லது தரவுத்தொகுதியில் செய்யப்படும். சொல்வகப்பாடு அடையாளப்படுத்தல் தவிர, பிற வகை அடையாளப்படுத்தல்கள் உள்ளன; ஒரு

தரவுத்தொகுதி அல்லது உரையின் மொழியியல் பகுப்பாய்வு அடிப்படையில் வெவ்வேறு நிலைகள் உள்ளன. எடுத்துக்காட்டாக பின்வருவன்வற்றைக் கூறலாம்:

ஒலியியல்சார் அடையாளப்படுத்தல் (phonetic annotation): எடுத்துக்காட்டாக பேச்சுத் தரவுத்தொகுதியில் உள்ள வார்த்தை எப்படி உச்சரிக்கப்படுகிறது என்பது பற்றிய தகவல்களைச் சேர்ப்பது ஒலியியல்சார் அடையாளப்படுத்தல் ஆகும்.

மீக்கூறு அடையாளப்படுத்தல் (prosodic annotation): அசையழுத்தம், இசையோட்டம் மற்றும் விட்டிசை போன்ற மீக்கூறு பண்புக்கூறுகள ஒரு பேச்சுத் தரவுத்தொகுதியில் சேர்பது மீக்கூறு அடையாளப்படுத்தல் ஆகும்.

தொடரியல் அடையாளப்படுத்தல் (syntactic annotation): கொடுக்கப்பட்ட வாக்கியம், தொடரியல் பகுப்பாய்வு அடிப்படையில் சொற்றொடர்கள், எச்சத்தொடர்கள் போன்ற அலகுகளாக எவ்வாறு பகுப்பாய்வு செய்யப்படுகிறது என்பது பற்றிய தகவல்களைச் சேர்ப்பது தொடரியல் அடையாளப்படுத்தல் ஆகும்.

பொருண்மையியல்சார் அடையாளப்படுத்தல் (semantic annotation): சொற்களின் பொருண்மையியல்சார் வகைப்பாடு பற்றிய தகவல்களைச் சேர்ப்பது பொருண்மையியல்சார் அடையாளப்படுத்தல் ஆகும். எடுத்துக்காட்டாக, *cricket* என்ற சொல் எந்த எழுத்துக்கூட்டல் வித்தியாசமும் உச்சரிப்பு வித்தியாசமும் இல்லாமல் ஒரு விளையாட்டைக் குறிப்பிடும் சொல்லாகவும் ஒரு பூச்சியைக் குறிப்பிடும் சொல்லாகவும் வெவ்வேறு பொருண்மையியல்சார் வகைப்பாடுகளைச் சார்வதை உணர்த்தும்.

பயன்வழியியல் அடையாளப்படுத்தல் (pragmatic annotation): பேசும் உரையாடலில் (spoken dialogue) நிகழும் பேச்சுச் செயல் (speech act) (அல்லது உரையாடல் செயல் (or dialogue act)) பற்றிய தகவல்களைச் சேர்ப்பது பயன்பாட்டியல் அடையாளப்படுத்தல் ஆகும். எடுத்துக்காட்டாக ஆங்கிலத்தில் பேசும்போது வெவ்வேறு சந்தர்ப்பங்களில் okay எனச் சொல்வது ஒரு ஒப்புதல் (acknowledgement), பின்னூட்டத்திற்கான கோரிக்கை (a request for feedback), ஏற்றுக்கொள்ளுதல் (acceptance) அல்லது விவாதத்தின் ஒரு புதிய கட்டத்தைத் தொடங்கும் ஒரு பயன்வழியியல்சார் குறிப்பானாக (pragmatic marker) இருக்கலாம்.

கருத்தாடல் அடையாளப்படுத்தல் (discourse annotation): ஒரு உரையில் முற்சுட்டுசார் இணைப்புகள் பற்றிய தகவல்களைச் சேர்ப்பது கருத்தாடல் அடையாளப்படுத்தல் ஆகும். எடுத்துக்காட்டாக, I'll saddle the horses and bring them round (நான் குதிரைகளை சேணம்

போட்டு அவற்றை வட்டமாகக் கொண்டு வருவேன்) என்ற வாக்கியத்தில் வரும். [பிரவுன் கார்பஸிலிருந்து ஒரு எடுத்துக்காட்டு] them என்பதையும் (தமிழில் *அவற்றை*) அதன் முற்சுட்டு the horses (தமிழில் *குதிரைகளை*) என்பதையும் இணைக்கும்படி அடையாளப்படுத்தல்.

நடையியல் அடையாளப்படுத்தல் (stylistic/ஸ்டைலிஸ்டிக் அடையாளப்படுத்தல்): எடுத்துக்காட்டாக, உரையில் அல்லது தரவுத்தொகுதியில் பேச்சு மற்றும் சிந்தனை விளக்கக்காட்சி பற்றிய தகவல்களைச் சேர்ப்பது (நேரடிப் பேச்சு, மறைமுகப் பேச்சு, சுதந்திர மறைமுகச் சிந்தனை போன்றவை).

சொல்சார் அடையாளப்படுத்தல்: ஒவ்வொரு சொல் வடிவத்தின் சொல்லனின்/லெம்மாவின் அடையாளத்தை ஒரு உரையில் சேர்ப்பது - அதாவது ஒரு அகராதியில் அதன் தலைப்பாக (எ.கா. lying என்பதன் சொல்லனாக/லெம்மாவாக LIE வருவது) போன்ற வார்த்தையின் அடிப்படை வடிவம்.

இயந்திரம்கற்றல் அணுகுமுறை

சதுரங்கம் அல்லது ஜியோபார்டியை எவ்வாறு வெல்வது என்பதிலிருந்து ஓட்டுநர் குறுகிய பாதை திசைகளை தீர்மானிப்பது வரை, ஒவ்வொரு நாளும் புதிய மற்றும் அற்புதமான சிக்கல்களைத் தீர்க்க மக்கள் கணினிகளைக் கற்றுக்கொடுத்தது போல் தெரிகிறது. ஆனால் கணினிகளால் செய்ய முடியாத பல பணிகள் இன்னும் உள்ளன, குறிப்பாக மனித மொழியைப் புரிந்துகொள்ளும் உலகில். புள்ளிவிவர முறைகள் இந்த சிக்கல்களை அணுகுவதற்கான ஒரு சிறந்த வழியாக நிரூபிக்கப்பட்டுள்ளன; ஆனால் இயந்திரக் கற்றல் (machine learning (ML/ எம்.எல்) நுட்பங்கள் பெரும்பாலும் ஒரு பெரிய அளவிலான தரவைக் காட்டிலும், ஒரு தரவுத்தொகுப்புக்குப் (dataset) பொருத்தமான சுட்டிகள் மூலம் வழிமுறைகள் வழங்கப்படும்போது சிறப்பாக செயல்படுகின்றன. இயற்கையான மொழியைப் பற்றி விவாதிக்கும்போது, இந்தச் சுட்டிகள் பெரும்பாலும் அடையாளங்கள்/சிறுகுறிப்புகள் (annotations) - மெட்டாடேட்டா வடிவத்தில் வருகின்றன, அவை உரையைப் பற்றிய கூடுதல் தகவல்களை வழங்கும். இருப்பினும், ஒரு கணினியைத் திறம்பட கற்பிப்பதற்காக, அதற்கு சரியான தரவை வழங்குவது முக்கியம், மேலும் அதில் இருந்து கற்றுக்கொள்ள போதுமான தரவு இருக்க வேண்டும்.

6.4.10. தகவல் மீட்கும் ஒழுங்குமுறை (Information retrieval system)

தகவல் மீட்டெடுப்பு (Information retrieval IR/ஐஆர்) என்றால் என்ன?

தகவல் மீட்டெடுப்பு (IR/ஐஆர்) என்ற வார்த்தையின் பொருள் மிகவும் பரந்ததாக இருக்கும். 1951ஆம் ஆண்டில், கால்வின் மூர்ஸ் (Calvin Mooers) " “Information retrieval” (“தகவல் மீட்டெடுப்பு") என்ற வார்த்தையை உருவாக்கினார், இதன் மூலம் ஒரு வருங்கால தகவல் பயனாளர் தகவலுக்கான கோரிக்கையை பயனுள்ள குறிப்புகளின் தொகுப்பாக மாற்ற முடியும். கால்வின் மூயர்ஸின் கூற்றுப்படி: "தகவல் மீட்டெடுப்பு என்பது தகவலின் விளக்கத்தின் அறிவுசார் அம்சங்களையும், தேடலுக்கான அதன் விவரக்குறிப்பையும், மேலும் செயல்பாட்டைச் செய்வதற்கு எந்த அமைப்புகள், நுட்பங்கள் அல்லது இயந்திரங்கள் பயன்படுத்தப்படுகின்றன என்பதையும் உட்படுத்துகிறது." இருப்பினும், ஒரு கல்வித் துறையாக, தகவல் மீட்டெடுப்பு பின்வருமாறு வரையறுக்கப்படலாம்: “தகவல் மீட்டெடுப்பு (IR/ஐஆர்) என்பது பெரிய சேகரிப்பிலிருந்து (பொதுவாக கணினிகள் சேமிக்கப்படும்) தேவையான தகவலைத் திருப்திசெய்யும் கட்டமைக்கப்படாத இயல்பைக்கொண்ட (பொதுவாக உரை) மூலப்பொருளைக் (பொதுவாக ஆவணங்கள்) கண்டுபிடிப்பது,

தகவல் மீட்டெடுப்பு (IR/ஐஆர்) மாதிரியாக்கம் (மாடலிங்), வடிவமைத்தல் மற்றும் பெரிய அளவிலான தகவல்களுக்கு விரைவான மற்றும் பயனுள்ள உள்ளடக்க அடிப்படையிலான அணுகலை வழங்கக்கூடிய அமைப்புகளை செயல்படுத்துவதை நோக்கமாகக் கொண்டுள்ளது. ஐஆர் அமைப்பின் நோக்கம், பயனரின் தகவல் தேவைக்கு உரை ஆவணங்கள், படங்கள் மற்றும் வீடியோ போன்ற தகவல் பொருட்களின் பொருத்தத்தை மதிப்பிடுவதாகும். இத்தகைய தகவல் தேவை வினவலின் வடிவத்தில் குறிப்பிடப்படுகிறது, இது வழக்கமாக சொற்களின் பையுடன் ஒத்திருக்கும். பயனர்கள் தங்கள் தகவல் தேவைக்கு பொருத்தமான தகவல் பொருட்களில் மட்டுமே ஆர்வமாக உள்ளனர். தகவல் பொருட்களின் பிரதிநிதித்துவம் மற்றும் அமைப்பு பயனருக்கு அவர் விரும்பும் தகவல்களை எளிதாக அணுக வேண்டும். ஒரு ஐஆர் அமைப்பின் முதன்மை குறிக்கோள், பயனர் வினவலுடன் தொடர்புடைய அனைத்து தகவல் பொருட்களையும் மீட்டெடுப்பது, அதே சமயம் மிகக்குறைவாகவே பொருந்தாத பொருட்களை (ஐடம்களை) மீட்டெடுப்பது. மேலும், மீட்டெடுக்கப்பட்ட தகவல் பொருள்கள் (ஐடங்கள்) மிகவும் பொருத்தமானவையிலிருந்து குறைந்த தொடர்புடையவையாக தரப்படுத்தப்பட வேண்டும்.

தகவல் மீட்டெடுப்பு (Information Retrieval) தரவு மீட்டெடுப்பிலிருந்து (data retrieval) வேறுபட்டது. தரவு மீட்டெடுப்பு முக்கியமாக ஒரு தொகுப்பின் எந்த ஆவணங்களில் பயனர் வினவலில் உள்ள முக்கிய வார்த்தைகளைக் கொண்டுள்ளது என்பதை தீர்மானிப்பதைக்

கொண்டுள்ளது, இது பயனர் தகவல் தேவையை பூர்த்தி செய்ய போதுமானதாக இல்லை. உண்மையில், ஒரு ஐஆர் அமைப்பின் பயனர் கொடுக்கப்பட்ட வினவலை திருப்திப்படுத்தும் தரவை மீட்டெடுப்பதை விட ஒரு விஷயத்தைப் பற்றிய தகவல்களை மீட்டெடுப்பதில் அதிக அக்கறை கொண்டுள்ளார். ஒரு தகவல் மீட்டெடுப்பு முறைமைக்கு, மீட்டெடுக்கப்பட்ட பொருள் துல்லியமாக இருக்கலாம் மற்றும் சிறிய பிழைகள் கவனிக்கப்படாமல் போகக்கூடும், ஆனால் தரவு மீட்டெடுக்கும் முறைக்கு, இருப்பினும் மீட்டெடுக்கப்பட்ட ஆயிரம் பொருள்களில் ஒரு தவறான பொருள் மொத்த தோல்வி என்று பொருள். ஐஆர் அமைப்பு மிகவும் பயனுள்ளதாக இருக்க, ஒரு தொகுப்பில் உள்ள தகவல் பொருட்களின் (ஆவணங்கள்) உள்ளடக்கங்களை எப்படியாவது "விளக்கம்" செய்து பயனர் வினவலுக்குப் பொருந்தக்கூடிய அளவிற்கு ஏற்ப அவற்றை வரிசைப்படுத்த வேண்டும். ஆவண உள்ளடக்கத்தின் இந்த "விளக்கம்" என்பது ஆவண உரையிலிருந்து வாக்கிய மற்றும் சொற்பொருள் தகவல்களைப் பிரித்தெடுப்பதும் பயனர் தகவல் தேவைக்கு பொருந்த இந்த தகவலைப் பயன்படுத்துவதும் அடங்கும். சம்பந்தப்பட்ட கருத்து தகவல் மீட்டெடுப்பின் மையத்தில் உள்ளது. மீட்டெடுப்பு செயல்முறையின் மிகவும் கடினமான பகுதி எந்த ஆவணங்களுடன் தொடர்புடையது என்பதை தீர்மானிப்பது அல்லது ஒரு குறிப்பிட்ட வினவலை பூர்த்தி செய்வது. பொருந்தக்கூடிய வரிசையை குறைப்பதில் ஆவணங்கள் முன்னுரிமை அளிக்கப்பட வேண்டும். வினவலுடன் தொடர்புடைய ஆவணங்கள் உயர்ந்த இடத்தில் இருக்கும்போது அதேசமயம் பொருத்தமற்றது குறைந்த தரத்தில் இருக்கும் போது ஒரு ஐஆர் அமைப்பு அதன் அதிகபட்ச மீட்டெடுப்பு செயல்திறனை அடைகிறது.

தகவல் மீட்டெடுப்பு (ஐஆர்) என்பது தொடர்புள்ள வளங்களின் தொகுப்பிலிருந்து ஒரு தகவல் தேவைக்கு பொருத்தமான தகவல் அமைப்பு வளங்களைப் பெறுவதற்கான செயல்பாடு ஆகும். தேடல்கள் முழு உரை அல்லது பிற உள்ளடக்கம் அடிப்படையிலான சொல்லடைவாக்கம் (indexing) முறையின் அடிப்படையில் இருக்கலாம். தகவல் மீட்டெடுப்பு என்பது ஒரு ஆவணத்தில் தகவல்களைத் தேடுவது, ஆவணங்களைத் தேடுவது, தரவை விவரிக்கும் மெட்டாடேட்டாவைத் தேடுவது மற்றும் உரைகள், படங்கள் அல்லது ஒலிகளின் தரவுத்தளங்களைத் தேடுவது.

தகவல் சுமை (information overload) எனப்படுவதைக் குறைக்கத் தானியங்குத் தகவல் மீட்டெடுப்பு ஒழுங்குமுறைகள் (Automated information retrieval system) பயன்படுத்தப்படுகின்றன. ஐஆர் அமைப்பு என்பது புத்தகங்கள், பத்திரிகைகள் மற்றும் பிற ஆவணங்களுக்கான அணுகலை வழங்கும் ஒரு மென்பொருள் அமைப்பு; அந்த ஆவணங்களை

சேமித்து நிர்வகிக்கிறது. வலை தேடுபொறிகள் (Web search engines) மிகவும் புலப்படும் ஐஆர் பயன்பாடுகள் ஆகும்.

கண்ணோட்டம் (Overview)

ஒரு பயனர் கணினியில் வினவலுக்குள் (query) நுழையும்போது தகவல் மீட்டெடுப்புச் செயல்முறை தொடங்குகிறது. வினவல்கள் தகவல் தேவைகளின் முறையான மொழிவுகள் (statements), எடுத்துக்காட்டாக வலை தேடுபொறிகளில் தேடல் கோர்வைகள் (search strings). தகவல் மீட்டெடுப்பில் ஒரு வினவல் சேகரிப்பில் ஒரு பொருளை (object) தனித்துவமாக அடையாளம் காணவில்லை. அதற்கு பதிலாக, பல பொருள்கள் வினவலுடன் பொருந்தக்கூடும், ஒருவேளை வெவ்வேறு அளவு பொருத்தத்துடன் (different degrees of relevancy).

ஒரு பொருள் (object என்பது உள்ளடக்க சேகரிப்பு அல்லது தரவுத்தளத்தில் உள்ள தகவல்களால் குறிப்பிடப்படும் ஒரு உருப்பொருள் (entity). பயனர் வினவல்கள் தரவுத்தளத் தகவலுடன் பொருந்துகின்றன. இருப்பினும், ஒரு தரவுத்தளத்தின் கிளாசிக்கல் SQL வினவல்களுக்கு மாறாக, தகவல்களை மீட்டெடுப்பதில், பெறப்பட்ட முடிவுகள் வினவலுடன் பொருந்தாமல் இருக்கலாம் அல்லது பொருந்தாது, எனவே முடிவுகள் பொதுவாக தரவரிசைப்படுத்தப்படுகின்றன. முடிவுகளின் தரவரிசை தரவுத்தள தேடலுடன் ஒப்பிடும்போது தகவல் மீட்டெடுப்பு தேடலின் முக்கிய வேறுபாடாகும்.

பயன்பாட்டைப் பொறுத்து தரவு பொருள்கள் data objects, எடுத்துக்காட்டாக, உரை ஆவணங்கள், படங்கள், ஆடியோ, மன வரைபடங்கள் mind maps அல்லது வீடியோக்களாக இருக்கலாம். பெரும்பாலும் ஆவணங்கள் ஐஆர் அமைப்பில் நேரடியாக வைக்கப்படுவதில்லை அல்லது சேமிக்கப்படுவதில்லை, மாறாக அதற்கு பதிலாக ஆவண பதிலிகள் அல்லது மெட்டாடேட்டா மூலம் கணினியில் குறிப்பிடப்படுகின்றன.

பெரும்பாலான ஐஆர் அமைப்புகள் தரவுத்தளத்தில் உள்ள ஒவ்வொரு பொருளும் வினவலுடன் எவ்வளவு பொருந்துகின்றன என்பதற்கான எண் மதிப்பெண்ணைக் கணக்கிடுகின்றன, மேலும் இந்த மதிப்புக்கு ஏற்ப பொருட்களை வரிசைப்படுத்துகின்றன. மேல் தரவரிசைப் பொருள்கள் பின்னர் பயனருக்குக் காண்பிக்கப்படும். பயனர் வினவலைச் செம்மைப்படுத்த விரும்பினால் செயல்முறை மீண்டும் செய்யப்படலாம்.

மாதிரிவகைகள்

முதல்பரிமாணம்: கணித அடிப்படை

கணக்-கோட்பாட்டு மாதிரிகள் (Set-theoretic models) ஆவணங்களை சொற்கள் அல்லது சொற்றொடர்களாகக் உருப்படுத்தம்செய்கின்றன. ஒற்றுமைகள் பொதுவாக அந்தத் கணங்களில் கணக்-கோட்பாட்டு செயல்பாடுகளிலிருந்து பெறப்படுகின்றன. பொதுவான மாதிரிகள்:

- நிலையான/தரமான பூலியன் மாதிரி Standard Boolean model
- விரிவாக்கப்பட்ட பூலியன் மாதிரி Extended Boolean model
- தெளிவற்ற மீட்டெடுப்பு Fuzzy retrieval

இயற்கணித மாதிரிகள் (Algebraic models) ஆவணங்கள் மற்றும் வினவல்களைப் பொதுவாக திசையன்கள் vectors, மெட்ரிக்குகள் matrices அல்லது டுப்பிள்களாக tuples குறிக்கின்றன. வினவல் திசையன் query vector மற்றும் ஆவண திசையனின் (document vector) ஒற்றுமை அளவிடக்கூடிய மதிப்பாகக் குறிப்பிடப்படுகிறது.

- திசையன் விண்வெளி மாதிரி (Vector space model)
- பொதுவான திசையன் விண்வெளி மாதிரி (Generalized vector space model)
- (மேம்படுத்தப்பட்ட) தலைப்பு சார்ந்த திசையன் விண்வெளி மாதிரி (Enhanced) Topic-based Vector Space Model)
- விரிவாக்கப்பட்ட பூலியன் மாதிரி (Extended Boolean model)+
- உள்ளுறை
- த சொற்பொருள் அகரவரிசை அட்டவணைப்படுத்தல் a.k.a. மறைந்த சொற்பொருள் பகுப்பாய்வு (Latent semantic indexing a.k.a. latent semantic analysis)

நிகழ்தகவு மாதிரிகள் (Probabilistic models) ஆவணத்தை மீட்டெடுக்கும் செயல்முறையை ஒரு நிகழ்தகவு அனுமானமாகக் (probabilistic inference) கருதுகின்றன. கொடுக்கப்பட்ட வினவலுக்கு ஒரு ஆவணம் பொருந்தக்கூடிய நிகழ்தகவுகளாக ஒற்றுமைகள் கணக்கிடப்படுகின்றன. பேயஸின் தேற்றம் (Bayes' theorem) போன்ற நிகழ்தகவுக் கோட்பாடுகள் பெரும்பாலும் இந்த மாதிரிகளில் பயன்படுத்தப்படுகின்றன.

- இரும சுதந்திர மாதிரி (Binary Independence Model)
- ஒகாபி (பிஎம் 25) (okapi (BM25)) தொடர்புடைய செயல்பாட்டை அடிப்படையாகக் கொண்ட நிகழ்தகவு பொருத்தமான மாதிரி (Probabilistic relevance model)
- நிச்சயமற்ற அனுமானம் (Uncertain inference)
- மொழி மாதிரிகள் (Language models)

- திசைதிருப்பல்-இருந்து-சீரற்ற தன்மை மாதிரி (Divergence-from-randomness model)
- உள்ளுறைந்த டிரிச்லெட் ஒதுக்கீடு (Latent Dirichlet allocation)

பண்புக்கூறுகளை அடிப்படையாகக் கொண்ட மீட்டெடுப்பு மாதிரிகள் (Feature-based retrieval models) ஆவணங்களை பண்புக்கூறு செயல்பாடுகளின் (அல்லது பண்புக்கூறுகளின்) மதிப்புகளின் திசையன்களாகப் பார்க்கின்றன; மேலும் இந்தப் பண்புக்கூறுகளை ஒற்றை பொருத்த மதிப்பெண்ணாக (single relevance score) இணைப்பதற்கான சிறந்த வழியைத் தேடுகின்றன, பொதுவாகத் தரவரிசை முறைகளைக் கற்றுக்கொள்வதன் மூலம். பண்புக்கூறுச் செயல்பாடுகள் ஆவணம் மற்றும் வினவலின் தன்னிச்சையான செயல்பாடுகளாகும், மேலும் இது வேறு எந்த மீட்டெடுப்பு மாதிரியையும் மற்றொரு பண்புக்கூறாக எளிதாக உள்ளடக்க/இணைக்க இயலும்.

இரண்டாவது பரிமாணம்: மாதிரியின் பண்புகள்

சொற்கூறு-இடைச்சாருமைகள் அற்ற மாதிரிகள் (Models without term-interdependencies) வெவ்வேறு சொற்கூறுகளை/சொற்களை சுதந்திரமாகக் கருதுகின்றன. இந்த உண்மை வழக்கமாக திசையன் இடைவெளி மாதிரிகளில் சொற்கூறு திசையன்களின் செங்குத்தான அனுமானத்தால் அல்லது நிகழ்தகவு மாதிரிகளில் சொற்கூறு மாறிகளுக்கான சுதந்திர அனுமானத்தால் பிரதிநித்துவம்செய்யப்படுகிறது.

இயல்பாயுள்ள சொற்கூறு இடைச்சாருமைகளைக் கொண்ட மாதிரிகள் (Models with immanent term interdependencies), சொற்கூறுகளுக்கு இடையிலான இடைச்சாருமையின் உருப்படுத்தத்தை/பிரதிநித்துவத்தை அனுமதிக்கின்றன. எவ்வாறாயினும், இரண்டு சொற்கூறுகளுக்கு இடையிலான இடைச்சாருமைகளின் அளவு, மாதிரியால் வரையறுக்கப்படுகிறது. இது முழு ஆவணங்களின் கணத்திலும் அந்தச் சொற்கூறுகளின் இணை நிகழ்விலிருந்து (co-occurrence) நேரடியாகவோ அல்லது மறைமுகமாகவோ (எ.கா. பரிமாணம்சார் குறைப்பு மூலம்) ஆக்கப்படுகிறது.

ஆழ்நிலைச் சொற்கூறு இடைச்சாருமைகளைக் கொண்ட மாதிரிகள் (Models with transcendent term interdependencies) சொற்களுக்கு இடையில் இடைச்சாருமைகளை உருப்படுத்தம் செய்ய அனுமதிக்கின்றன; ஆனால் அவை இரண்டு சொற்கூறுகளுக்கு இடையிலான இடைச்சாருமை எவ்வாறு வரையறுக்கப்படுகிறது என்று குறிப்பிடவில்லை. அவை

இரண்டு சொற்கூறுகளுக்கு இடையில் இடைச்சாருமையின் அளபுக்காக வெளிப்புற வளத்தை நம்பியுள்ளன. (எடுத்துக்காட்டாக, ஒரு மனிதன் அல்லது அதிநவீன வழிமுறைகள்.)

செயல்திறன் மற்றும் சரியான நடவடிக்கைகள்

ஒரு தகவல் மீட்டெடுப்பு ஒழுங்குமுறையின் மதிப்பீடு (evaluation of an information retrieval system) என்பது ஒரு ஒழுங்குமுறை அதன் பயனர்களின் தகவல் தேவைகளை எவ்வளவு சிறப்பாகப் பூர்த்தி செய்கிறது என்பதை மதிப்பிடுவதற்கான செயல்முறையாகும். பொதுவாக, அளவீடு தேடப்பட வேண்டிய ஆவணங்களின் தொகுப்பையும்/சேகரிப்பையும் தேடல் வினவலையும் கருதுகிறது. பூலியன் மீட்டெடுப்பு அல்லது மேல்-கே மீட்டெடுப்பிற்காக (top-k retrieval) வடிவமைக்கப்பட்ட பாரம்பரிய மதிப்பீட்டு அளவீடுகள், துல்லியம் மற்றும் நினைவுகூரல் ஆகியவை அடங்கும். எல்லா நடவடிக்கைகளும் ஒரு அடிப்படை உண்மைக் கருத்தைக் கருதுகின்றன: ஒவ்வொரு ஆவணமும் ஒரு குறிப்பிட்ட வினவலுக்கு பொருத்தமானவை அல்லது பொருந்தாதவை என்று அறியப்படுகிறது. நடைமுறையில், வினவல்கள் தவறாக முன்வைக்கப்படலாம் மற்றும் பொருத்தத்தின் வேறுபட்ட குறைகள் இருக்கலாம்.

தகவல் மீட்டெடுப்புகளின் வடிவமைப்பு மற்றும் செயல்பாடு (IRs design and operation)

எந்தவொரு தகவல் மீட்டெடுப்பு ஒழுங்குமுறையின் அடிப்படைக் கொள்கை என்னவென்றால், அது சேவை செய்யும் ஹோஸ்ட் அமைப்பின் சட்டங்களால் நிர்வகிக்கப்படுகிறது. ஐஆர்எஸ் வடிவமைப்பில் கருத்தில் கொள்ள வேண்டிய சில முக்கியமான காரணிகள்: i) சேவை செய்ய வேண்டிய பயனர் வகுப்பைக் குறிப்பிடவும். நேரடி பயனர்களுக்கும் பயனாளிகளுக்கும் இடையில் வேறுபாடு காணுங்கள். ஒரே தகவல் மீட்டெடுப்பு அமைப்புகளுக்கு ஒன்றுக்கு மேற்பட்ட பயனர் அல்லது பயனாளி வகுப்பு இருந்தால், அமைப்பின் சிறப்பியல்புகளைப் பயனர் வகுப்பினருடன் பொருத்துவதன் அடிப்படையில் தெளிவான வேறுபாட்டு முன்னுரிமைகளை ஒதுக்கி அறிவிக்கவும். ii) ஒழுங்குமுறை/கணினி நோக்கம் கொண்ட பயன்கள் மற்றும் சிக்கல்-வகுப்பைக் குறிப்பிடவும், அவற்றில் முன்னுரிமைகளைச் சரிசெய்யவும், இருப்பினும் வரையறுக்கப்பட்ட சிக்கலுடன் தகவல் மீட்டெடுப்பு அமைப்பை வடிவமைப்பது கடினமான பணியாகும். iii) ஒழுங்குமுறையின்/கணினியுப் தரவுத்தளத்தில் பெறப்பட வேண்டிய, ஒழுங்கமைக்கப்பட்ட மற்றும் சேமிக்கப்பட வேண்டிய தகவல் பொருட்களின் வரம்பைக் குறிப்பிடவும். iv) தரவுத்தளத்தில் உள்ளீடுகளை குறுக்கு உறுதிசெய்தல் மற்றும் சரிபார்த்தல். v) முக்கியத்துவம், பொருத்தப்பாடு, ஏற்புடைமை மற்றும் மிகைமை

இவற்றிற்காக (அவ்வப்போது அல்லது குறைபாடு ஏற்படும்போது) தரவுத்தளத்தைப் பரிசோதனை செய்ய, மறுசீரமைக்க மற்றும் திருத்த இந்த அமைப்பு செலவு குறைந்ததாக இருக்க வேண்டும் vi) வன்பொருள் மென்பொருள் மற்றும் பிற கூறுகள்: ஒழுங்குமுறையின் வன்பொருள் மற்றும் மென்பொருள் தேவைகளைப் பகுப்பாய்வு செய்வது அவசியம்.

வடிவமைப்பு (Design)

ஒரு தகவல் மீட்டெடுப்பு ஒழுங்குமுறையின் வடிவமைப்பு, பகுப்பாய்வின் சிக்கல்கள் மற்றும் இணைப்பாக்கத்தின் சிக்கல்களை உள்ளடக்கியது. ஒரு தகவல் மீட்டெடுப்பு ஒழுங்குமுறையைப் பகுப்பாய்வு செய்வதன் நோக்கம் ஒரு பகுதி வரிசையில் சாத்தியமான தகவல் மீட்டெடுப்பு ஒழுங்குமுறைகளின் ஒரு தொகுப்பை ஏற்பாடு செய்வதாகும். தகவல் மீட்டெடுப்பு ஒழுங்குமுறை அதன் செயல்பாட்டில் செயலூக்கமுடையதாகவும் திறமையாகவும் இருக்க வேண்டும். ஒரு ஐஆர் ஒழுங்குமுறை, சாத்தியமானதாக இருக்க, ஹோஸ்ட் நிறுவனத்தில் தகவல்களைச் செயலாக்குவதை நிர்வகிக்கும் அடிப்படைக் கொள்கைகளுடன் இணங்க வேண்டும்.

6.4.11. தொடரடைவு ஒழுங்குமுறை (Concordance system)

ஒரு தொடரடைவு என்பது ஒரு நூல்/உரையில் அல்லது பணி அமைப்பில் பயன்படுத்தப்படும் முக்கிய சொற்களின் அகர வரிசையாகும், ஒவ்வொரு வார்த்தையின் ஒவ்வொரு நிகழ்வையும் அதன் உடனடி சூழலுடன் பட்டியலிடுகிறது. நேரம், சிரமம் மற்றும் கணினிக்கு முந்தைய சகாப்தத்தில் ஒரு தொடரடைவை உருவாக்குவதற்கான செலவு காரணமாக வேதங்கள், பைபிள், குர்ஆன் அல்லது ஷேக்ஸ்பியர், ஜேம்ஸ் ஜாய்ஸ் அல்லது கிளாசிக்கல் லத்தீன் மற்றும் கிரேக்க ஆசிரியர்களின் படைப்புகள் போன்ற சிறப்பு முக்கியத்துவம் வாய்ந்த படைப்புகளுக்கு மட்டுமே தொடரடைவுகள் தொகுக்கப்பட்டுள்ளன.

தொடரடைவு ஒரு அகரவரிசை சொல்லடவை (index) விடக் கூடுதலானதாகும்; அவை வர்ணனை, வரையறைகள் மற்றும் தலைப்புசார் குறுக்குச் சொல்லடடைவுப்படுத்தல் (topical cross-indexing) போன்ற கூடுதல் பொருள்களை (material) உள்ளடக்குகின்றன; இது கணினிகளால் உதவப்படும்போது கூட உழைப்பு மிகுந்த செயல்முறையாகத்தான் உருவாக்குகிறது.

முன் கணினி/கணக்கீட்டு சகாப்தத்தில், தேடல் தொழில்நுட்பம் கிடைக்கவில்லை; மேலும் பைபிள் போன்ற நீண்ட படைப்புகளைப் படிப்பவர்களுக்கு ஒரு தொடரடைவு வழங்கியது,

அவர்கள் தேடிக்கொண்டிருக்கும் ஒவ்வொரு வார்த்தையையும் தேடல் முடிவுகளுடன் ஒப்பிடக்கூடிய ஒன்று; இன்று, பல சொற்களைப் பற்றிய வினவல்களின் முடிவை (பிற சொற்களுக்கு அருகில் சொற்களைத் தேடுவது போன்றவை) இணைக்கும் திறன் தொடரடைவு வெளியீட்டில் ஆர்வத்தை குறைத்துள்ளது. கூடுதலாக, சொல் சூழலின் அடிப்படையில் மொழியியல் தகவல்களைத் தானாக அடையாளம் காணும் வழிமுறையாக உட்கீடையான பொருண்மைசார் சொல்லடைவுப்படுத்தல் (அட்டவணைப்படுத்தல்) போன்ற கணித நுட்பங்கள் முன்மொழியப்பட்டுள்ளன.

இருமொழி தொடரடைவு (bilingual concordance) என்பது சீரமைக்கப்பட்ட இணை உரையை அடிப்படையாகக் கொண்ட ஒரு தொடரடைவு ஆகும்.

ஒரு தலைப்புசார் தொடரடைவு (topical concordance) என்பது ஒரு நூல் உள்ளடக்கிய பாடங்களின் பட்டியல் (வழக்கமாக பைபிள்), அந்த பாடங்களை உள்ளடக்கிய உடனடி சூழலுடன். ஒரு பாரம்பரிய தொடரடைவைப் போலன்றி, சொல்லடைவு செய்யப்பட்ட சொல் வசனத்தில் தோன்ற வேண்டியதில்லை. நேவின் தலைப்புசார் பைபிள் (Nave's Topical Bible) என்பது மிகவும் பிரபலமான தலைப்புசார் தொடரடைவு ஆகும்., and the

முதல் பைபிள் தொடரடைவு வல்கேட் பைபிளுக்கு (Vulgate Bible) ஹக் ஆஃப் செயிண்ட் செர் Hugh of St Cher (d.1262) தொகுத்தது; அவருக்கு உதவ 500 துறவிகளைப் பயன்படுத்தினார். 1448ஆம் ஆண்டில், ரப்பி மொர்தெகாய் நாதன் (Rabbi Mordecai Nathan) எபிரேய பைபிளுக்கு தொடரடைவை நிறைவு செய்தார். அவருக்குப் பத்து ஆண்டுகள் பிடித்தன. கிரேக்க புதிய ஏற்பாட்டிற்கான ஒரு தொடரடைவு 1599இல் ஹென்றி ஸ்டீபன்ஸ் (Henry Stephens) அவர்களால் வெளியிடப்பட்டது, மேலும் செப்டுவஜின்ட் (Septuagint) இரண்டு ஆண்டுகளுக்குப் பிறகு 1602இல் கான்ராட் கிர்ச்சரால் (Conrad Kircher) செய்யப்பட்டது. ஆங்கில பைபிளின் முதல் தொடரடைவு 1550இல் திரு மார்பெக் (Mr Marbeck) அவர்களால் வெளியிடப்பட்டது. க்ரூடனின் (Cruden) கூற்றுப்படி, இது 1545ஆம் ஆண்டில் ராபர்ட் ஸ்டீபன்ஸ் வகுத்த வசன எண்களைப் பயன்படுத்தவில்லை, ஆனால் திரு காட்டனின் "அழகான பெரிய தொடரடைவு" இதைச் செய்தது. பின்னர் க்ரூடனின் கான்கார்டன்ஸ் Cruden's Concordance மற்றும் ஸ்ட்ராங்கின் கான்கார்டன்ஸ் (Strong's Concordance) ஆகியவற்றைப் பின்பற்றினார்.

ஒரு உரையைப் படிக்கும்போது, மொழியியலில் தொடரடைவுகள் அடிக்கடி பயன்படுத்தப்படுகின்றன. உதாரணத்திற்கு:

- ஒரே வார்த்தையின் வெவ்வேறு பயன்பாடுகளை ஒப்பிடுதல்
- முக்கிய வார்த்தைகளைப் பகுப்பாய்வு செய்தல்
- சொல் அதிர்வெண்களை பகுப்பாய்வு செய்தல்
- சொற்றொடர்கள் மற்றும் மரபுத்தொடர்களைக் கண்டுபிடித்து பகுப்பாய்வு செய்தல்
- துணை உறுப்புகளின் மொழிபெயர்ப்புகளைக் கண்டறிதல், எ.கா. பிடெக்ஸ்டுகள் மற்றும் மொழிபெயர்ப்பு நினைவுகளில் கலைசொல்.
- சொல்லடைவுகள் மற்றும் சொல் பட்டியல்களை உருவாக்குதல் (வெளியிடுவதற்கும் பயனுள்ளதாக இருக்கும்)

அமெரிக்கன் தேசியத் தரவுத்தொகுதி (American National Corpus), பிரிட்டிஷ் தேசியத் தரவுத்தொகுதி (British National Corpus), தற்கால அமெரிக்கன் ஆங்கிலத் தரவுத்தொகுதி (Corpus of Contemporary American English) போன்ற தேசிய உரை தரவுத்தொகுதிகளில் தொடரடைவு நுட்பங்கள் பரவலாகப் பயன்படுத்தப்படுகின்றன. தொடரடைவு நுட்பங்களைப் பயன்படுத்தும் தனித்த பயன்பாடுகள் தொடரடைவிகள் (concordancers) அல்லது அதற்கு மேற்பட்ட மேம்பட்ட தரவுத்தொகுதி மேலாளர்கள் (advanced corpus managers) என அழைக்கப்படுகின்றன. அவைகளில் சில சொல்வகைப்பாட்டு அடையாளங்களை ஒருங்கிணைத்து, தரவுத்தொகுதி மொழியியலில் ஏற்றுக்கொள்ளப்பட்ட பல்வேறு வகையான தேடல்களை நடத்துவதற்கு பயனருக்கு தனது சொந்த சொல்வகைப்பாட்டு அடையாளப்படுத்தப்பட தரவுத்தொகுதிகளை உருவாக்க உதவுகின்றன.

6.4.12. வாக்கியப் பகுப்பாய்வு ஒழுங்குமுறை (Sentence parsing system)

வாக்கிய பாகுபடுத்தல்

கணினி தொடரியல் (computational syntax) ஆய்வு வாக்கியங்களைத் தொடர்களாகவும் சொற்களாகவும் பகுத்து அவற்றின் தொடரியல் மற்றும் சொல்வகைப்பாட்டுப் பண்புக்கூறுகளை அடையாளப்படுத்தித் தரும் செயல்பாடாகும். வாக்கியங்கள் தரப்படுகையில் அவற்றை தொடரியல் கூறுகளாகப் பகுப்பது மற்றும் தொடரியல் கூறுகள் தரப்படும் போது வாக்கியங்களை உருவாக்குவது என்ற இரு செயல்பாடுகள் இதில் அடங்கும்.

6.4.12.1.கணிப்பொறி வழி மொழி ஆய்வுக்கான இலக்கண வடிவமைப்புகள்

ஒரு கூற்றின் பொருத்தமான பொருளைத் தருமாறு சொற்களுக்கு இடையே உள்ள உறவுகளைப் பிரித்தெடுக்கும் நுட்பத்தைத் தருவது இந்த இலக்கண வடிவமைப்பின் முக்கியமான செயல்பாடாகும். வாக்கியங்களை பகுப்பாய்வு செய்வதற்குப் பயன்படுத்துகின்ற செயல்முறையிலிருந்து தனியாக ஒரு குழும விதிகளால் நாம் மொழியை பண்பாக்கம் செய்ய

விரும்புகிறோம். அம்மாதிரியான குழும விதிகள் முறையான இலக்கணம் எனப்படும். முறையான இலக்கணம் ஒரு குழும இலக்கண வாக்கியங்களை வரையறுக்கிறது. அவ்வாக்கியங்களுக்கு அமைப்பைத் தருகின்றது. மொழியைப் பேசுகின்றவர்களின் இலக்கணம் குறித்த உள்ளுணர்வுடன் பொருத்துகின்ற வாக்கியத்தின் பொருளை வரையறுக்கின்ற முறையான இலக்கணத்தை உருவாக்குகிறதுதான் நமது செயல். தொடக்ககால கட்டத்தில் சூழல்வரையறை இல்லாத இலக்கண வடிவவாதம் பயன்படுத்தப்பட்டது. தற்போது கட்டுப்பாடு அடிப்படையிலான இலக்கண முன்மாதிரிகள் பயன்படுத்தப்படுகின்றன. அவை கீழே பட்டியலிடப்பட்டுள்ளன:

- செயல்பாட்டு ஒருங்கிணைக்கப்பட்ட இலக்கணம் (Functional Unification Grammar)
- தலை இயக்கத் தொடரமைப்பு இலக்கணம் (Head-driven phrase Structure Grmmar)
- செயல்பாட்டு இலக்கணம் (Fuctional Grammar)
- சொல் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar)
- வகைப்பாட்டு ஒன்றிணைக்கப்பட்ட இலக்கணம் (Categorical Unification Grammar)
- கிளை சேர்க்கப்படும் இலக்கணம் (Tree adjoining Grammar)
- சார்பு இலக்கணம் (Dependency grammar)

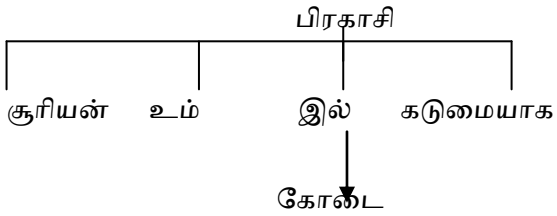
வடிவம்சார் இலக்கணம் (Formal Grammar)

வாக்கியங்களைப் பகுப்பாய்வு செய்வதற்குப் பயன்படுத்துகின்ற செயல்முறையிலிருந்து தனியாக ஒரு குழும விதிகளால் நாம் மொழியை பண்பாக்கம் செய்ய விரும்புகிறோம். அம்மாதிரியான குழும விதிகள் முறையான இலக்கணம் ஒரு குழும இலக்கண வாக்கியங்களை வரையறுக்கிறது; அந்த வாக்கியங்களுக்கு அமைப்பைத் தருகின்றது. மொழியைப் பேசுகின்றவர்களின் இலக்கணம் குறித்த உள்ளுணர்வுடன் பொருத்துகின்ற வாக்கியத்தின் பொருளை வரையறுக்கின்ற முறையான இலக்கணத்தை உருவாக்குகிறதுதான் நமது செயல்.

சூழல் வரையறை இல்லாத இலக்கணம் (Context Free Grammar)

சூழல் வரையறை இல்லாத இலக்கணம் மரபு மொழியியலரால் அண்மை உறுப்பு இலக்கணங்கள் என்றும் வழியமைப்பு மொழித் திட்டமிடுபவர்களால் பேக்ஸ் இயல்பான வடிவு (Bgackes Normal Form) என்றும் சில கணினி பயன்பாட்டில் மறுதரவு அமைப்பொழுங்கு (Recursive Pattern) என்றும் அழைக்கப்படுகின்றது. இவைகளைப் பொதுவாகத் தொடரமைப்பு இலக்கணம் (Pharse Structare Grammar) என்பதன் குறிப்பிட்ட வகை எனலாம்; இவைதான் ஆக்கமுறை இலக்கணத்திற்கு (Generative Grammar) அடிப்படையாக அமைந்தது. அமைப்புக்கான மூன்று அணுகுமுறைகள்: புடைப்பெயர்வு பின்னல் (Transition Network), தட்டையான அமைப்பு (flat structure) உடையவை, கிளையமைப்பு (hierarchical structure) உடையவை. நம்முடைய உள்ளுணர்வு உறுப்பமைவை எதிர்பார்க்கின்றது. இதை வாக்கியம் என்பது பின்னலால் விற்களின் தொடர்ச்சியாகப் பொருந்துவதாக கொள்ள இயலாது. வாக்கியம்

துண்டுகளால் அல்லது தொடர்களால் ஆனது. தொடர்களின் அமைப்பு, பொருள் எவ்வாறு பெறப்படுகின்றது என்று கூறுவதால் இவ்வகையிலான அமைப்பாக்கம் முக்கியமானதாகும். மொழி அமைப்பை விளக்க பல வகைப்பட்ட விதிகளின் வடிவமைப்புகள் (formulations) இருக்கின்றன. எவ்வாறு மொழி அமைப்பை நன்றாக விளக்கலாம் என்ற அணுகுமுறையின் அடிப்படையில் அவற்றின் விளக்கங்கள் வேறுபடுகின்றன. தலை மற்றும் அடை அணுகுமுறை (head and modifier approach), இலக்கணத்தின் மரபுத்தன்மையிலான பல அறிமுகங்கங்கள் அமைப்புகளை அடிப்படை அமைப்பொழுங்கிலிருந்து நீட்சி செய்யப்பட்ட அடை செய்யப்பட்ட அமைப்பாக விவரணை செய்கின்றன. எடுத்துக்காட்டாக, வாக்கியத்தின் அக அமைப்புப் பெயர்த் தொடரைத் தொடர்ந்து வரும் வினை என்று கொள்ளலாம். பெயர் அடைகள், சுட்டடைகள் (அடைகொளி அடைகள்) என்பன பெயரை அடை செய்வதாகவும் வினையடை போன்றவை வினையை அடை செய்வதாகவும் கருதலாம்; அடைகளையும் அடைகளால் விவரிக்கலாம். ஒவ்வொறு நிலையிலும் தனிச்சொல்லாக ஒரு தலை இருக்கும். அதற்கு ஒன்றோ அதற்கு மேலோ அடைகள் வரலாம். எடுத்துக்காட்டாக, *சூரியன் கோடையில் கடுமையாகப் பிரகாசிக்கும்* என்ற வாக்கியத்தைப் பின்வருமாறு ஆயலாம்.



பொருளையும் தொடரியலையும் உள்ளடக்கிய காரணத்தால் தலை அடை அணுகுமுறை மொழியியலில் முக்கிய இடம் வகிக்கின்றது. ஒரு பொருள் (சாதனம்) எப்படி இருக்கிறது என்று கூறும் பெயரடை, பெயருக்குக் கூடுதலான விவரணையைத்/விளக்கத்தைத் தரும். வினையை விவரிக்கும் வினையடை முறை, காலம், இடம் மற்றும் பிற செய்திகளைத் தரும். இவ்வகையிலான சொல் அமைப்பிற்கும் கருத்துரு அமைப்பிற்கும் இடையிலான பொருத்தம் மொழியைப் பொருளை வெளிப்படுத்துகிற ஒருமித்த ஒழுங்குமுறையாகப் பார்ப்பதால் பயனுள்ளதாய் இருக்கின்றது. தலை-அடை இலக்கணம் சார்பு இலக்கணமாக வடிவமைக்கப்பட்டு பயன்படுத்தப்படுகின்றது. சார்பு இலக்கணக் கொள்கை எவ்வாறு அமைப்பொழுங்கின் நிலையான அறிவு அமைப்பாக்கம் செய்யப்படுகின்றது. அல்லது வாக்கியம் ஆய்வு செய்யப்படுகின்றது என்று கூறாமல் அமைப்புகளை விளக்குவதின் வழிகளை வலியுறுத்துகின்றன.

அண்மை உறுப்பு அணுகுமுறை தொடரமைப்புச் சட்டகம் (Phrase Structure Formwork)

தொடரமைப்புச் சட்டகம் அண்மை உறுப்பு ஆய்வு அடிப்படையிலானது. அண்மை உறுப்பாய்வு ஒரு வாக்கியத்தை அண்மை உறுப்புகளாகப் பகுக்கும்.

அண்மை உறுப்பாய்வின் மாதிரிகள் (Models of IC analysis)

அண்மை உறுப்பாய்வு தொடரியல், உருபனியல், ஒலியனியல் என்ற எல்லா நிலைகளிலும் செய்யப்படுகின்றது. அண்மை உறுப்பாய்வில் பல மாதிரிகள் உள்ளன. அவற்றில் மூன்று வகை மாதிரிகள் கீழே தரப்பட்டுள்ளன.

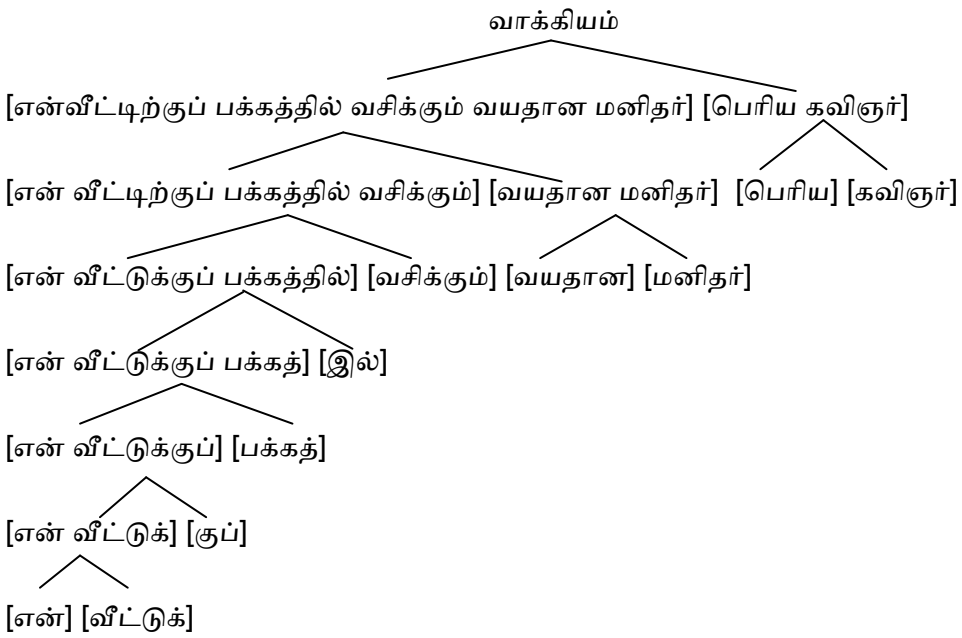
1. அடையாளப்படுத்தாத பெட்டிப்படம்

பின்வரும் எடுத்துக்காட்டில் தரப்பட்ட வாக்கியத்தின் அண்மை உறுப்புகள் பகுக்கப்பட்டு அடையாளப் படுத்தாமல் பெட்டிப்படமாகத் தரப்பட்டுள்ளது:

<table>
<tr><td>என்</td><td>வீட்டுக்</td><td>குப்</td><td>பக்கத்</td><td>இல்</td><td>வசிக்கும்</td><td>வயதான</td><td>மனிதர்</td><td>பெரிய</td><td>கவிஞர்</td></tr>
<tr><td colspan="3">என் வீட்டுக்</td><td rowspan="4"></td><td rowspan="5"></td><td colspan="2" rowspan="4">வயதான மனிதர்</td><td colspan="2" rowspan="5">பெரிய கவிஞர்</td></tr>
<tr><td colspan="4">என் வீட்டுக்குப்</td></tr>
<tr><td colspan="5">என் வீட்டுக்குப் பக்கத்</td></tr>
<tr><td colspan="6">என் வீட்டுக்குப் பக்கத்தில்</td></tr>
<tr><td colspan="8">என் வீட்டுக்குப் பக்கத்தில் வசிக்கும்</td></tr>
<tr><td colspan="8">என் வீட்டுக்குப் பக்கத்தில் வசிக்கும் வயதான மனிதர்</td></tr>
<tr><td colspan="10">என் வீட்டுக்குப் பக்கத்தில் வசிக்கும் வயதான மனிதர் பெரிய கவிஞர்</td></tr>
</table>

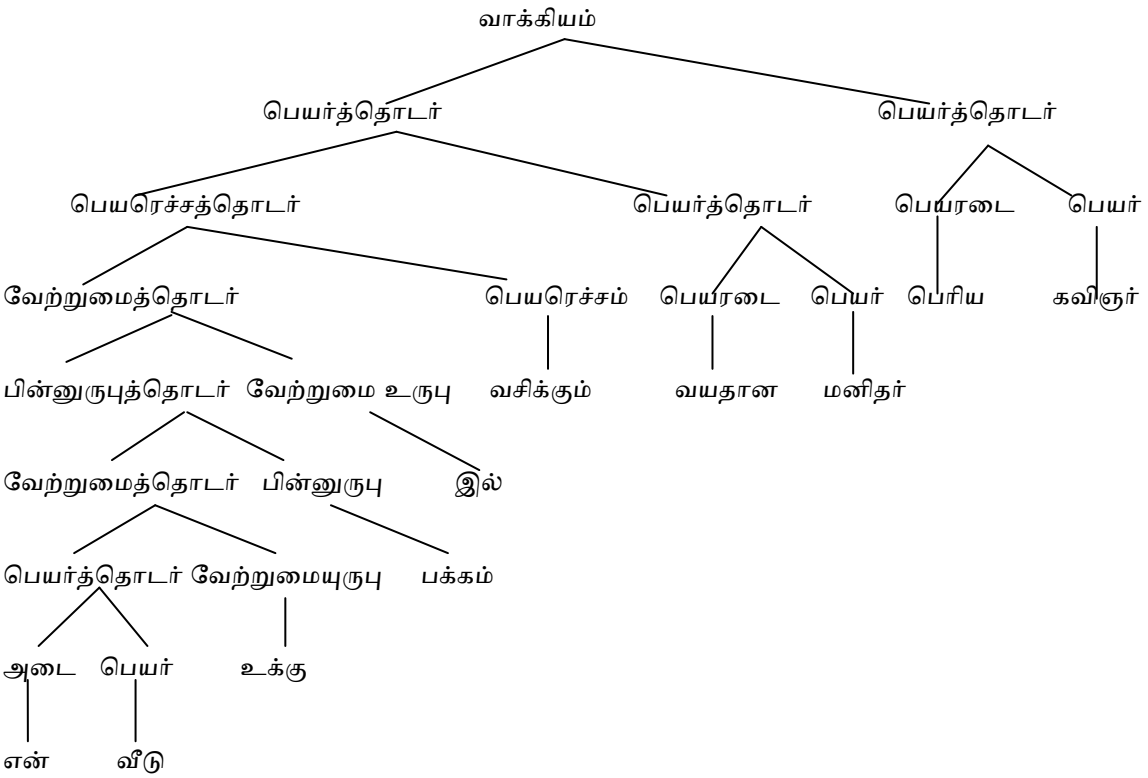
1. அடைப்புக்குறிப் படம்
2. [[[[என்] [வீட்டிற்]] [குப்]] [பக்கத்]] [இல்]] [வசிக்கும்]] [[வயதான] [மனிதர்]]] [[பெரிய] [கவிஞர்]]
3. கிளைப்படம்

வாக்கியப் பகுத்தாய்வைப் பின்வருமாறு படிநிலை அமைப்பில் அண்மை உறுப்புகள் கட்டுக்களை ஒன்றுக்குள் ஒன்றாக அமைவதைப் பின்வருமாறு கிளைப்படத்தில் காட்டலாடம்.



4. அடையாளப்படுத்தப்பட்ட படம்

இறுதி நிலையில் வரும் அண்மை உறுப்புகளைச் சொல்வகைப்பாட்டிற்காகவும் அவற்றால் விளையும்; கட்டுகளைத் தொடர்களாகக் கொண்டும் குறிப்பீடு செய்தால் பின்வருமாறு அடையாளப்படுத்தப்பட்ட படம் விளையும்.



அண்மை உறுப்புப் பகுப்பாய்வு குறித்த சில செய்திகள் கீழே பட்டியலிடப்பட்டுள்ளன.

1. அண்மை உறுப்பு பகுநிலை (analytic) அடிப்படையிலானது.

2. அடிப்படை மொழி அலகுகளைக் காண்பதற்காக அறிவியல் கண்டுபிடிப்புச் செயல்முறை (discovery procedures) செயல்பாட்டால் உருவானது.

3. வாக்கியங்களை இரண்டு அண்மை உறுப்புப் பாகங்களாகவும் ஒவ்வொரு சிறு பாகத்தையும் மேலும் சிறு அலகுகளாகவும் (உருபன் வகை) பகுப்பது இக்கோட்பாட்டின் கொள்கையாகும்.

4. சில மொழியியலார் வாக்கியத்தை பெரிய அலகுகளாகக் கொண்டு மேலிருந்து கீழ் நோக்கி (top-down) ஆய்வாளர்கள் சிலர் கீழிருந்து மேல்நோக்கி (bottom-up)ஆய்வார்கள்.

5. அண்மை உறுப்பு ஆய்வு அடையாளப்படுத்தப்பட்டால் மரபிலக்கணப் பகுப்பாய்வு (Parsing) போலவே அமையும். அதாவது வாக்கியத்தை இலக்கண வகைப்படுகளாக பகுத்தல் நிகழும்.

6. மரபு அடிப்படையிலான வாக்கியப் பகுப்பாயவும் தொடரமைப்பு இலக்கணம் அடிப்படையிலானதாகும். இவ்விலக்கணம் பின்வரும் கேள்விகளுக்கு விடை தருகின்றது.

1. ஒரு வாக்கியத்;தின் உருப்புகள் யாவை?

2. அவை எவ்வாறு ஒழுங்குபடுத்தப்பட்டுள்ளன

7. மரபிலக்கணங்கள் வாக்கியங்களையும் தொடர்களையும் விளக்கப் பின்வரும் கூற்றுகளைத் தருகின்றன.

1. ஒரு வாக்கியத்தில் பெயர்த்தொடரும் (Noun Phrase) வினைத்தொடரும் அடங்கும்.

2. ஒரு பெயர்த்தொடரில் ஒரு பெயர் மட்டுமோ அடைகொளி அடை + பெயரோ அடைகொளி அடை + பெயரடை + பெயரோ வரும்.

எடுத்துக்காட்டாக,

1. அந்தப் பையன்

2. அந்த நல்லப் பையன்.

3. ஒரு வினைத்தொடரில் பெயர் மட்டுமோ வினை மட்டுமோ வினை + துணை வினையோ பின்வருபுத் தொடர் + வேற்றுமைத் தொடர் + வினையோ வரும். எடுத்துக்காட்டு,

வந்தான்

வந்துக்கொண்டிருந்தான்

வீட்டிலிருந்து வந்துக்கொண்டிருந்தான்

பழத்தைச் சாப்பிட்டான்

4. பின்னுருபுத் தொடரில் வேற்றுமைத் தொடரும் பின்னுருபும் வரும்.

எடுத்துக்காட்டு, வீட்டிலிருந்து

5. வேற்றுமைத் தொடரில் பெயர்த்தொடரும் வேற்றுமை உருபும் வரும்.

எடுத்துக்காட்டு,

வீட்டை, வீட்டின்

திட்டவட்டமற்ற சிக்கனமில்லாத இம்மாதிரியான விளக்கக்கூறுகள் தருவதற்குப்பதிலாக சில மரபு மற்றும் கருத்துச் சாயல் அடிப்படையில் ஒரு குழு ஆக்கவிதிகள் (Generation rules) அல்லது விரித்தெழுது விதிகள் (rewrite rules) தரலாம். அம்மாதிரியான விதிகளைத் தொடரமைப்பு விதிகள் (phrase structure rules) எனலாம். தொடரமைப்பு விதிகள் பின் மாதிரியான பொது விதிச் சுருக்கத்தைக் (rules-schema) கொண்டிருக்கும்.

அ → ஆ

அதாவது 'அ' என்ற குறியீட்டைத் தொடக்ககுறியீடு என்ற குறியீடாக விரித்தெழுதலாம். விரித்தெழுதும் செயல்பாட்டைத் தொடங்கிவைப்பதற்குப் பயன்படுத்தும் குறியீட்டைத் தொடக்ககுறியீடு எனலாம்.

எ.கா

வா → பெ.தொ + வி.தொ.

என்ற விரித்தெழுது விதியில் 'வா' என்பது தொடக்கக்குறியீடு. மேற்சொன்ன விதி ஒரு வாக்கியத்தில் பெயர்த்தொடரும் வினைத்தொடரும் வரும் என்று தெரிவிக்கின்றது. '+' என்ற குறி

பிணைப்பைக் (concatenation) குறிப்பிடப் பயன்படுகின்றது. பெ.தொ, வி.தொ என்ற புலக்குறிகளை (labels) விரித்தெழுதலாம்.

எ.கா.

பெ.தொ → அடைகொளி அடை + பெயர்

என்ற விதி பெ.தொ. கணுவை விரித்தெழுதுகின்றது. (அடைகொளி அடையில் *அந்த, இந்த, என், உன், அவனது, அவளது, அதன், அவற்றின்* என்பன அடங்கும்). *பையன், சிறுமி* போன்றன பெயர்ச்சொல்லில் அடங்கும். கணு என்பது கிளைப்படத்தில் ஒன்றோ அல்லது அதற்கு மேற்பட்டகிளைகளோ வெளிவரும் தொடக்கவிடமாகும். தமிழில் வி.தொ என்பதை பி.உ.தொ + வே.தொ + வி.தொ + து.வி என விரித்து எழுதலாம்.

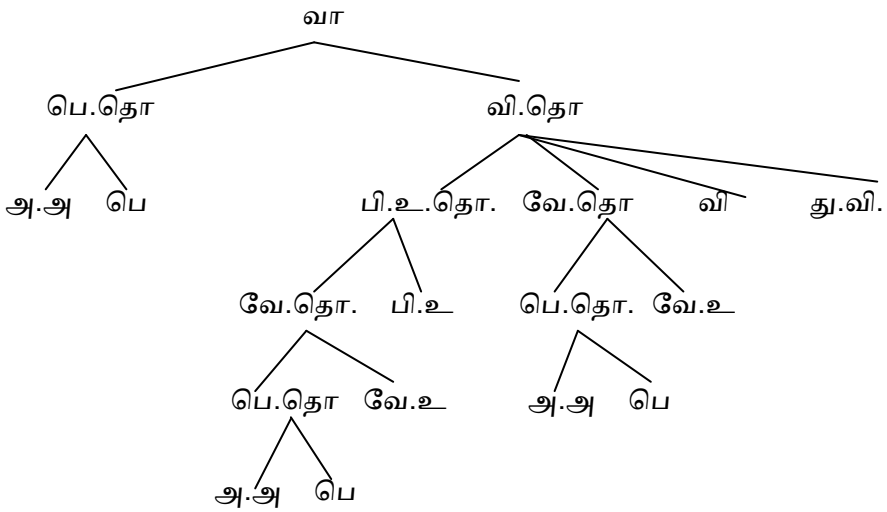
வி.தொ → பி.உ.தொ + வே.தொ + வி + து.வி.

பி.உ.தொ என்ற கணுவைப் பின்வருமாறு விரித்து எழுதலாம்.

பி.உ.தொ → பெ.தொ + பி.உ.

(வா = வாக்கியம், பெதொ = பெயர்த்தொடர், விதொ = வினைத் தொடர், பிஉதொ = பின்னுருபுத்தொடர், வேதொ = வேற்றுமைத் தொடர்; துவி = துணைவினை, வேஉ = வேற்றுமை உருபு, பிஉ = பின்னுருபு, அ அஅ = அடைகொளி அடை, விஅ =வினை அடை, பெஅ = பெயரடை, எஅ = எண்ணடை, மிஅ =மிகை அடை)

இதுவரை தந்த விதிகளைப் பின்வருமாறு கிளைபடமாகத் தரலாம்.



மேற்கண்ட கருத்துச் சாயல்களைப் பயன்படுத்தி தமிழ் இலக்கணத்தில் ஒரு பகுதிக்கு தொடரமைப்பு விதிகள் உருவாக்கலாம்.

வா →பெ.தொ + வி.தொ.

வி.தொ → வே.தொ + (பி.உ.தொ + வி +து.வி

து.வி → கொண்டிரு, இரு, வேண்டு, பார், போ

வி → கொல், வெட்டு, நட, பார், ஒடி

பெ.தொ → அ.அ + (பெ.அ) + பெ

பெ → பையன், பாம்பு, மரம், மேஜை, சிறுமி, பெண்.

அ.அ → அந்த, ஒரு, ஓர், தன்

பெ.அ → வயதான, இளமையான, அழகான

பி.உ.தொ → வே.தொ + பி.உ

வே.தொ → பெ.தொ + வே.உ

பி.உ → பக்கத்தில், மேலே, கீழே, நோக்கி, அருகில், அடியில்

வே.உ → ஐ, இல், .ஆல், உக்கு, ஓடு

மேற்கண்ட தொடரமைப்பு இலக்கணம் (Phrase sturcture grammar) பின் வரும் தமிழ் வாக்கியங்களை உருவாக்கும்.

அந்தப் பையன் ஒரு பாம்பைப் கொல்லப் பார்த்தான் (பெ.தொ + வே.தொ + வி + து.வி) அந்த அழகான பெண் தன் வீட்டை நோக்கி நடந்தாள் (பெ.தொ + பி.உ.தொ + வி)

அந்த வயதான மனிதர் மரத்தை வெட்டப்போனார். (பெ.தொ + வே.தொ + வி + து.வி)

ஆனால் மேற்சொன்ன விதிகள் பின்வரும் பொருளற்ற வாக்கியங்களையும் உருவாக்கும்.

* மேஜை அந்தப் பெண்ணை மரத்திற்கு அருகில் கொன்றது.

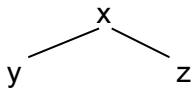
* மரம் அந்தப் பையனைப் போகும்

* அந்தச் சிறுவன் அந்தச் சிறுமியை ஒடிக்க முடியும்

* மேஜை அந்த வயதான மனிதரைப் பார்க்கும்

மேற்சொன்ன விதிகள் சுழல் வரையறையில்லாத விதிகளாகும். அவற்றைச் சுழல் கட்டுண்ட விதிகளாக (Contex sensitive rules) மாற்றினால் மேற்சொன்ன வாக்கியங்களின் உருவாக்கத்தைத் தடை செய்யலாம். சூழல் வலையரை இல்லாத இலக்கணத்தைப் பயன்படுத்தி வாக்கியத்தைப் பகுத்தல், பகுப்பாய்வு/பகுத்தாய்தல் (parsing) எனப்படும். உள்ளீடு செய்யப்படும் கோர்வையைத் (input string) தெரிந்து கொள்வதும் அதற்கு அமைப்பைத் தருகிறதும் ஆன இரண்டும் இணைந்த செயல்பாடுதான் பகுத்தாய்தலாகும். வாக்கியத்தை தெரிந்து கொள்வதும் அதற்கு அமைப்பைத் தருகிறதும் ஆன செயல்பாடு தொடரியல் பகுத்தாய்தல் (syntactic parsing) எனப்படும். பகுத்தாயும் கருவி பகுத்தாய்வான் (parser) எனப்படும். இவ்வாறு பகுத்தாய்வான் இரு செயல்களைச் செய்ய வேண்டும். ஒன்று ஒரு கோர்வைத் தரப்படும் போது பகுத்தாய்வான் முதலில் அதைப் பகுத்தாய வேண்டிய மொழியின் வாக்கியம் என்பதைத் தெரிந்து கொள்ள வேண்டும். இந்நிலையில் பகுத்தாய்வானுக்கு உள்ளடங்கிய தெரிந்துகொள்வான்கள் (built-in recognizers)

இருக்க வேண்டும். இரண்டாவது அது வெளியீடு செய்கிற அமைப்பை வாக்கியத்திற்கு தர வேண்டும். இலக்கணத்தில் இருக்கிற மொழியியல் செய்திகளை பகுத்தாய்வான்கள் சார்ந்திருக்க வேண்டும் என்பது இதன் மூலம் தெரிகின்றது. வெளியீட்டு வசதிகளுடன் கூடிய தெரிந்துகொள்வான்களைப் (recognizers) பொறுத்த வரையில் இலக்கணங்களும் பகுத்தாய்வான்களும் பகுத்தாயும் பண்பு வாய்ந்தவை. x→y என்ற விதி ஒரு உருவாக்கமாகத் தரப்பட்ட கலவைநிலை கட்டளையாகக் கருதப்பட வேண்டும்.



பகுத்தாய்வான் பெரும்பாலும் குறியீடுகளின் கோர்வையை ஏற்கும்: அதற்கு ஒரு விதிகையைப் பயன்படுத்தும் அது ஒரு கோர்வையின் விரித்தெழுது விதி போல் வரும். எடுத்துக்காட்டாக b → d என்ற விதியைப் பயன்படுத்தி abc என்பதை adc என்று விரித்தெழுதலாம்; adc என்பதை adc என்றும் விரித்தெழுதலாம். 'ABC' 'ADC', மற்றும் 'Adc' என்பன வாக்கிய வடிவுகள் (Sententical forms) அல்லது ஆக்கங்கள் எனப்படும். 'ADC' என்ற கோர்வை 'ABC'-யில் இருந்து ஒரு விதியைப் பயன்படுத்தி பெறப்படுவதால் ADC-ஆனது ABC-யிலிருந்து நேரடியாகப் பெறப்பட்டது ஆகும். 'ADC' என்ற கோர்வை 'ABC'-யில் இருந்து ஒரு விதியைப் பயன்படுத்தி பெறப்படுவதால் ADC-ஆனது ABC-யிலிருந்து மறைமுகமாகப் பெறப்பட்டது ஆகும். ஒவ்வொரு படியிலும் பகுத்தாய்வான்கள் ஏதோ ஒரு அமைப்பை வெளியிடும். ஒரு குழும விதிகளின்படி வரக்கூடிய எல்லா அமைப்புகளையும் நாம் தெரிந்தால் ஒருவாக்கியத்தைப் பகுத்தாயலாம்.

6.4.12.2. பகுத்தாயும் நடைமுறைத்திறன் (parsing strategies)

மொழியியல் அறிவைப் பிரதிபலிக்கின்ற விதிகளின் அடிப்படையில் பகுத்தாய்வான் செயல்படும். வாக்கியங்களுக்கு அமைப்புத் தருகையில் ஒரு குழும விதிகளைப் பல நிரல்களில் நிறைவேற்றலாம். ஒவ்வொரு வேறுபட்ட நிரலும் வேறுபட்ட பகுத்தாய்தல் நடைமுறைத் திறத்துடன் தொடர்புடையது. பகுத்தாய்வான்களை அவற்றோடு தொடர்புடைய நடைமுறைத்திறன் அடிப்படையில் வகைப்பாடு செய்யலாம். பகுத்தாயும் நடைமுறைத் திறனாக இரண்டு அளவீடுகள் அடிக்கடி பயன்படுத்தப்படுகின்றனவும் நிலைபேறு பெற்றவனவும் ஆகும். முதலாவது அது பகுத்தாயும் கோர்வைக்கு வெளியீடுகளின் மொழியியல் அமைப்பைக் கவனக் குவிப்பு செய்கிறது. உள்ளீட்டுக் கோர்வையிலிருந்து தொடங்கி அமைப்பு கட்டப்படுவதனால் அந்த பகுத்தாய்வான் கீழிருந்து மேல் (bottom-up) செயல்படுவதாகக் கூறலாம். அல்லது தொடக்கக் குறியீட்டிலிருந்து (கிளை அமைப்பின் வேர்) தொடங்குவதனால் பகுத்தாயவான் மேலிருந்து கீழ் (top-town) செயல்படுவதாகக் கூறலாம். பகுத்தாயும் நடைமுறைத் திறன்களை வகைப்படுத்தும் அளவீடுகளை ஒரு எடுத்துக்காட்டு மூலம் விளக்கலாம். பின்வரும் ஒரு குழும விரித்தெழுது விதிகளை எடுத்துக்கொள்வோம்.

1a. s→ AB

1b. s→ cd

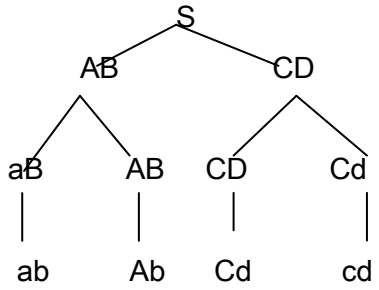
1c. A→ a

1d. B→ b

1e. C→ C

1f. D→d

தரப்பட்டுள்ள ஒரு குழும விதிகளிலிருந்து அவற்றால் பெறப்படும் எல்லா ஆக்கங்களின் திட்டத்தைக் கட்ட முடியும். தொடக்க குறியீடான S-இல் இருந்து இரண்டு விதிகளைச் செயல்படுத்தலாம். விதி 1a AB-இன் ஆக்கத்தையும் விதி 1b CD-இன் ஆக்கத்தையும் விளைவிக்கும். AB-யில் விதி 1c -ஐயும் விதி 1d-ஐயும் பயன்படுத்தி முறையே aB மற்றும் Ab என்ற கோர்வைகளைப் பெறலாம். இந்த எண்ணத்தைப் பின்பற்றி தரப்பட்ட ஒரு குழும விரித்தெழுது விதிகள் அனுமதிக்கும் ஆக்கங்கங்களின் எல்லா கோர்வைகளையும் கண்டுகொள்ளலாம். விளைவு பெரும்பாலும் கிளை வடிவாக உருப்படுத்தம் செய்யப்படும்.



மேலே தரப்பட்ட கிளை அமைப்பு 1 இல் தரப்பட்ட விதிகளால் விளையும் ஆக்கங்களின் எல்லா கோர்வைகளையும் காட்டுகின்றது. கிளையில் வரும் ஒவ்வொரு கணுவும் (AB,CD, aB, ab, Ab, CD, Cd, cd) வாக்கிய வடிவில் நடைமுறைப்படுத்தக் கூடிய வேறுபட்ட விதிகளின் தேர்வை உருப்படுத்தம் செய்யும். கிளை அமைப்பின் கிளைகள் (ab, cd) அதாவது அடியில் வரும் கணுக்கள் மேலும் விதிகளின் பயன்படுத்த இயலாத வாக்கிய வடிவுகளை உருப்படுத்தம் செய்யும். இந்த கிளை அமைப்பை மொழியியலார் எவ்வாறு ஒரு வாக்கியத்தின் பாகங்கள் இணைகின்றன என்பதைக் காட்டுகின்ற மொழியியல் அமைப்பை உருப்படுத்தம் செய்யப் பயன்படுத்துகின்ற கிளையுடன் ஒப்பிட்டு குழப்பிக்கொள்ளக்கூடாது. மேல் சொன்ன விதிகள், கிளை அமைப்பைக் காட்ட இயலும் ஆக்கங்களின் எல்லாக் கோர்வைகளையும் தருகின்றன. இந்த வகையான கிளை அமைப்பின் அடிப்படையில் ஆழம் முதல் (depth-first) அல்லது அகலம் முதல் (breadth- first) என்ற பகுத்தாயும் நடைமுறைத்திறன்களின் பாகுபாடு அமையும்.

மேலிருந்து கீழ் மற்றும் கீழிருந்து மேல் பகுத்தாய்தல் (Top- Down) Versus bottom-up Parsing)

பின்வரும் இரு எடுத்துக்காட்டுகள் எவ்வாறு மேலிருந்து கீழ் மற்றும் கீழிருந்து மேல் பகுத்தாய்வான்கள் பின்வரும் வாக்கியத்திற்கு அமைப்பு தருகின்றது என்பதை விளக்கும்.

1. ஒரு பூனை எலியைப் பிடித்தது.

இரண்டும் பின்வரும் விரித்தெழுதும் விதிகளைப் பயன் படுத்தப்படுவதாக வைத்துக்கொள்வோம்.

2.a. வா →பெ.தொ + வி.தொ

2. b. பெ.தொ →அ.அ + பெ

2. c. வி.தொ →வே.தொ + வி

2. d. வி.தொ →வி

2. e. வே.தொ →பெ.தொ + வே உ

2. f. வி →பிடித்தது

2. g. பெ→ பூனை

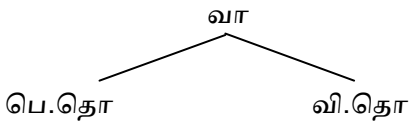
2. h. பெ →எலி

2. i. அ.அ →ஒரு

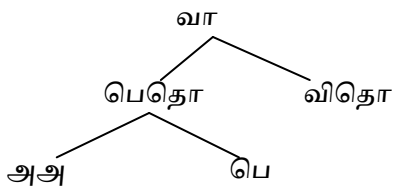
2. j. வே.உ→ ஐ

மேலிருந்து கீழ் பகுத்தாய்தல் (மேல்-கீழ் பகுத்தாய்த்தல்) (Top- Down parsing)

மேலிருந்து கீழ் பகுத்தாய்வான்கள் எப்பெழுதும் தொடக்கக் குறியீடான 'வா' என்பதிலிருந்து தொடங்கும் அதில் பயன்படுத்த விதிகளைக் கண்டு கொள்ளும் அதை விரிக்கும் எடுத்துக்காட்டில் 2.a என்ற விதிதான் இருக்கிறது. அதைப்பயன்படுத்தி விளையும் விளைவு பின்வரும் அமைப்பைத் தரும்.



இரு புதிய கணுக்கள் தோன்றும். இந்த கணுக்கள் இறுதிக் குறியீடுகளா (terminal symbols) என்று முதலில் பகுத்தாய்வான் பார்க்கும். அப்படியானால் அந்தக் குறியீடுகள் பகுத்தாயப்படுகின்ற கோர்வையுடன் பொருத்திப்பார்க்கப்படும். இல்லாவிடில் பகுத்தாய்வான் முதல் இறுதியுறா குறியீட்டுக் கணுவை விரிக்கும். மேற்கண்ட எடுத்துக்காட்டில் பெதொ வில் 2.b என்ற விதி பயன்படுத்தப்பட்டு பின்வரும் அமைப்பு விளையும்.



புதிதாக உருவாக்கப்பட்ட இந்தக் கணுக்களும் இறுதிக் கணுக்கள் அல்ல.

கீழிருந்து மேல் பகுத்தாய்தல் (Bottam-up parsing)

கீழ்-மேல் பகுத்தாய்தல் உள்ளீடு செய்யப்பட்ட கோர்வையைத் தொடக்கமாகக் கொண்டு செயல்படத் தொடங்கி அதை வேரான 'வா' என்று சுருக்கும். அது ஒரு வாக்கியத்தை உள்ளீடாக ஏற்கும் சொற்களை (இறுதியுற்ற குறியிடுகளை) அவற்றின் வகைப்பாட்டால் இடம்பெயர்க்கும். இதற்காக அது விரித்தெழுது விதிகளின் வலப்பக்கம் பார்த்து அவற்றை இடப்பக்கம் உள்ள வகைப்பாடாகச் சுருக்க வேண்டும். மீண்டும் *ஒரு பூனை எலியைப் பிடித்தது* என்ற வாக்கியம் 2a-j என்ற விதிகளால் கூறப்பட்டுள்ள இலக்கணத்தால் அமைப்பு பெறும். எடுத்துக்காட்டாக முதலில் 2f. என்ற விதியைப் பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்.

வி
↓
ஒரு பூனை ஒரு எலி ஐ பிடித்தது

2.j.-ஐப் பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்

வே.உ வி
↓ ↓
ஒரு பூனை ஒரு எலி - ஐ பிடித்தது

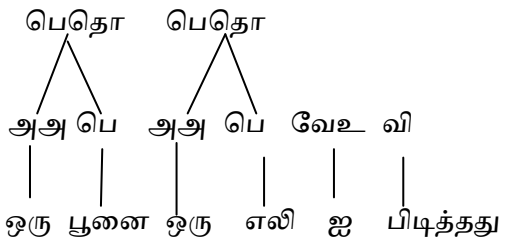
2.h.-ஐ பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்

பெ பெ வே.உ வி
↓ ↓ ↓ ↓
ஒரு பூனை ஒரு எலி ஐ பிடித்தது

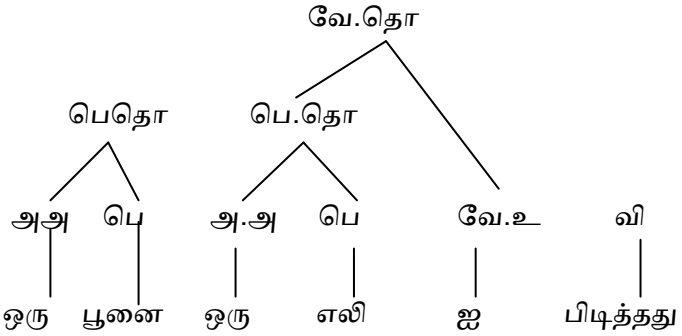
2 .i.-ஐப் பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்

அஅ பெ அஅ பெ வே.உ வி
↓ ↓ ↓ ↓ ↓ ↓
ஒரு பூனை ஒரு எலி ஐ பிடித்தது

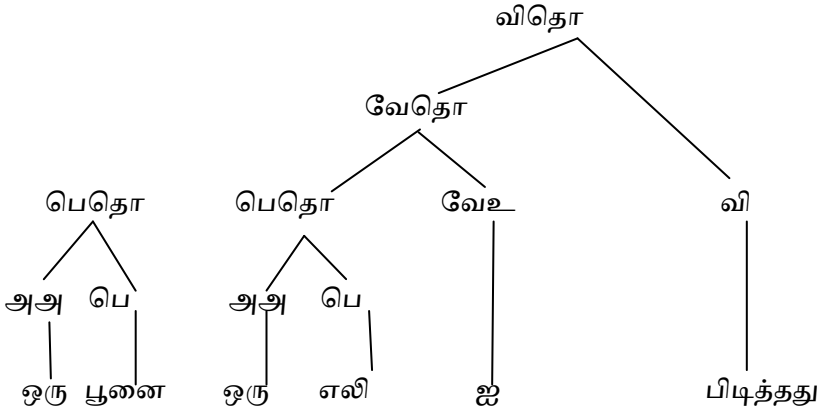
இந்த இடத்தில் 2b என்ற விதி 'அ.அ' என்பதையும் 'பெ' என்பதையும் இணைந்து அவற்றை பெ.தொ எனக் குறிக்கும்.



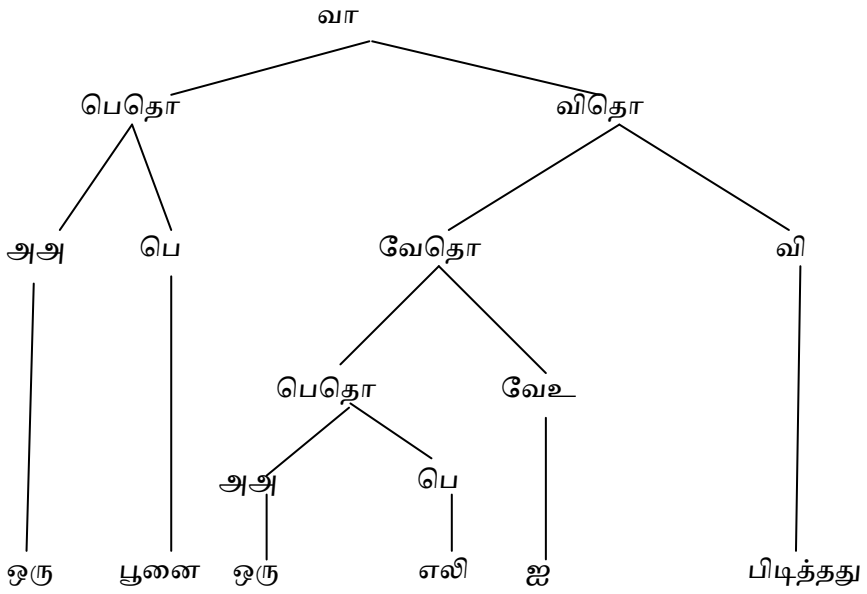
இந்த இடத்தில் 2e என்ற விதி வேஉ என்பதையும் பெதொ என்பதையும் இணைத்து அவற்றை வேதொ என சுருக்கும்.



விதி 2c ஐ-வேதொ ஐயும் வி-ஐயும் இணைத்து வி.தொ. எனச் சுருக்கும்.



விதி 2a பெ.தொ ஐயும் வி.தொ ஐயும் இணைத்து வா ஆகச் சுருக்கும்.



பகுத்தாய்வான் வாக்கியத்தின் வலப்பக்கத்திலிருந்து தொடங்கி கோவையின் முன்னோக்கிச் செயல்படுவதாகக் கொள்வது அவ்வளவு சரியானதல்ல. தெளிவாகக் கூறினால் மொழி இம்மாதிரி செயல்படுவதில்லை. இதன் அடிப்படையில் கீழ்-மேல் பகுத்தாய்வான் உளவியல் அடிப்படையில் சரியானது அல்ல என்று வாதிடலாம். வேறுபட்ட ஆக்கங்களின் தொடர்ச்சியைப் பின்பற்றினாலும் ஒரே இலக்கணத்தைப் பின்பற்றுவதனால் இடது பக்கத்திலிருந்து தொடங்கும் பகுத்தாய்வானும் வலது பக்கத்திலிருந்து தொடங்கும் பகுத்தாய்வானும் பகுத்தாய்தல் அடிப்படையிலும் பெற்ற விளைவின் அடிப்படையிலும் ஒன்றாகும். இந்த வலமிருந்து இடம் என்ற வேறுபடலுக்கும் வலது பகுத்துத்தாய்தல் அல்லது இடது பகுத்தாய்தல் என்பதற்கும் தொடர்பில்லை. வலது பகுத்தாய்தலானது கீழ்-மேல் பகுத்தாய்தலின் விளைவாகும். இது விதிகளின் வலப்பக்கத்தில் காணப்படும் குறியீடுகளைப் பார்த்து வாக்கிய வடிவுகளைச் சுருக்கும். ஒரு மேல்-கீழ் பகுத்தாய்வான் இடது பகுத்தாய்தலை நடைமுறைப்படுத்தும். இது விதிகளின் இடப்பக்கத்தில் காணப்படும் குறியீடுகளை விரித்து வாக்கிய வடிவுகளை ஆக்கும். இதுபோல் ஒரே மொழியியல் செய்திகளின் படி ஒரே வாக்கியங்களுக்கு ஒரே அமைப்புகளைத் தருவதால் ஒரே இலக்கணத்தை நோக்கீடு செய்யும் மேல்-கீழ் பகுப்பானும் கீழ்-மேல் பகுப்பானும் ஒன்றாகும். அவைகள் வேறுபட்ட ஆக்கங்களின் தொடர்ச்சியைப் பின்பற்றுகிறது என்ற உண்மையுடன் எவ்வளவு நினைவகம் தேவை எவ்வளவு கணிப்பு நேரத்தை உட்படுத்துகிறது என்பதாலும் வேறுபடும்.

ஆழம் முதல் மற்றும் அகலம் முதல் பகுத்தாய்தல் (Depth first Versus breath first Parsing)

பின்வரும் ஒரு குழும விரித்தெழுது விதிகளை எடுத்துக் கொள்வோம்.

3 a. S ⟶ AB

3 b. S ⟶ CD

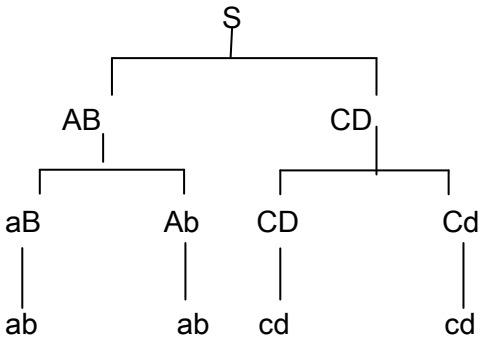
3 c. S → a

3 d. B → b

3e. C → c

3 f. D → d

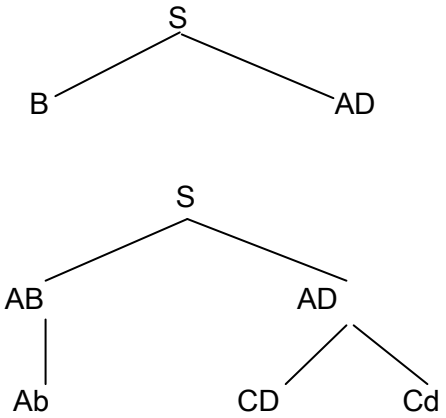
இவற்றோடு தொடர்புடைய பின்வரும் ஆக்கக் கிளையமைப்பையும் எடுத்துக் கொள்வோம்.

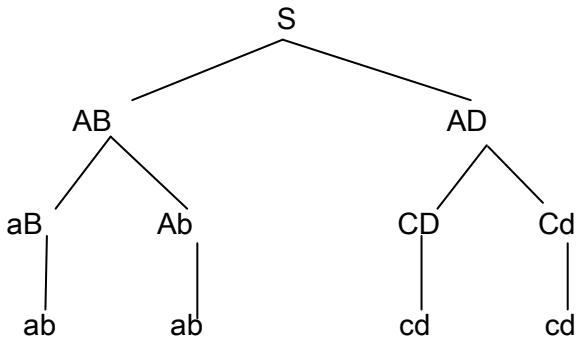


இந்த கிளை அமைப்பை இருவிதமாக அணுகலாம்: ஒன்று அதன் செங்குத்து நிலையில் கவனம் செலுத்திப் பெறலாம்; இரண்டு அதன் இடை நிலையில் கவனம் செலுத்திப் பெறலாம். இந்த இரு வேறுபட்ட நோக்கும் பகுத்தாயும் நடைமுறைத்திறனின் வேறுபாட்டை விளைவிக்கும். ஆழம் முதல் பகுத்துக்குறித்தல் நாம் மேலே தரப்பட்டுள்ள கிளை அமைப்பின் செங்குத்து வழியை எடுத்துக் கொள்வோம். வேரை வாக்கியத்துடன் இணைக்கும் ஒரு தனி செங்குத்து விதியைத் தேர்ந்தெடுப்போம். எடுத்துக்காட்டாக கிளைக் கணுவில் அடங்கும் ab என்ற வாக்கியத்திற்கு பின்வருமாறு வழியைத் தேர்ந்தெடுக்கலாம்.

இந்த வழி கிளை அமைப்பில் வரும் எந்தச் செங்குத்தான வழியைப் போல வாக்கிய வடிவுகளின் ஒரு தொடர்ச்சியைத் தரும். ஒவ்வொரு வாக்கிய வடிவும் வேருடனோ அல்லது வேறு தனி விதியால் விளைந்த வாக்கிய வடிவுடனோ ஒரு தனி விதியைப் பயன்படுத்துவதால் விளைவதாகும். ஒரு குறிப்பிட்ட ஆக்கத்திற்கு பல விதிகளைப் பயன்படுத்தலாம் என்றாலும் ஒரே ஒரு விரித்தெழுது விதிதான் தக்கவைக்கப்படும் பிற விருப்புகள் கிளையின் பிற வழிகளில்

வெளிப்படுத்தப்படும். ஒரு தனிச் செங்குத்தான வழியாகப் பின்பற்ற இயலாது. ஒரு கிளை அமைப்பில் ஒரு தனி செங்குத்தான வழியிலே உருப்படுத்தம் செய்யத்தக்க வாக்கிய வடிவுகளைப் பின்பற்றும் பகுத்தாய்வான் ஆழம்-முதல் பகுத்தாய்வான் என்று அழைக்கப்பெறும். மேல்-கீழ் மற்றும் கீழ்-மேல் பகுத்தாய்வான்கள் இவ்வகையைச் சாரும். மேற்கண்ட கிளைப்படம் ஒரே வாக்கியத்தை ஒன்றுக்கும் மேற்பட்ட வழிகளில் தொடக்கக் குறியீடான 'வா' என்பதுடன் இணைக்கலாம் எனக் காட்டும். ஆழம் முதல் பகுத்தாய்வான் ஒரு சமயத்தில் ஒரு வழியை ஆய்வதால் தொடக்கத்தில் தேர்ந்தெடுக்கப்பட்ட பாதை தவறாகப் போகலாம். இந்த இடங்களில் பகுத்தாய்வான் சரிசெய்து தவறிலிருந்து விடுபட வேண்டியது அவசியம் இதன் காரணமாக ஆழம் முதல் பகுப்பான்கள் பின்னழுவல் வசதியுடன் நிறைவேற்றப்பட வேண்டும். ஆழம் முதல் பகுத்தாய்வான் நாம் ஒரு ஆக்கக்கிளை அமைப்பை இடைநிலைப் பரிமாணத்திற்கு முக்கியத்துவம் கொடுத்து கிளை அமைப்பின் ஒரு நிலையில் எல்லாக் கணுக்களையும் எடுத்துக் கொண்டு பார்க்கலாம். எடுத்துக்காட்டாக ஆக்கக்கிளை அமைப்பின் வேர் 'S' என்ற வாக்கிய வடிவைத் தாங்கும் கணுவைக் கொண்டிருக்கும். AB மற்றும் AD என்ற வாக்கிய வடிவுகளைக் கொண்ட இரண்டு மகள் கணுக்கள் 'S' என்பதிலிருந்து வேறுபட்ட விதிகளால் விளைந்தனவாகும். இம்மாதிரியான நிகழ்வில் பகுத்தாய்வான் ஒரே நேரத்தில் வேறுபட்ட ஆக்கங்களை உருவாக்கும். அடுத்த அடியில் இரண்டின் விளைவுகளின் மேல் இயலும் எல்லா விதிகளும் செயல்படுத்தப்படும். இவ்வகையிலான பகுத்தாய்வான் அகலம் முதல் பகுத்தாய்வான் எனப்படும்.





வீதி முதல் பகுத்தாய்வான்கள் உருவாக்கப்பட்ட எல்லா வாக்கிய வடிவிலும் இயலும் எல்லா விதிகளைச் செயல்படுத்தும் அவை ஆக்கக் கிளை அமைப்பின் இடைநிலைப் பரிமாணத்தை ஆயும். ஒரே நேரத்தில் வரும் எல்லாத் தேர்வுகளையும் தீர்த்து விட்டு தோல்வி அல்லது வெற்றி என்ற முடிவுக்கு எடுத்துச் செல்லும். இவ்வகையிலான செயல்முறை தோல்வியின் போது பின்னழுவலை அதிக பட்சமாகும். தவறான தேர்வால் விளையும் ஆக்கத் தொடர்ச்சி இல்லாது போனாலும் ஒரே சமயத்தில் உருவான எல்லா வெற்றிப் பிறழ்ச்சிகளும் எஞ்சும்.

வகைப்பாடு (Classification)

	ஆழம் முதல்	அகலம் முதல்
மேல்-கீழ்	மேல்-கீழ் ஆழம்-முதல்;	மேல்-கீழ் அகலம்-முதல்
கீழ்-மேல்	கீழ்-மேல் ஆழம்-முதல்	கீழ்-மேல் அகலம்-முதல்

6.3.13. சொற்பொருள் மயக்கம் நீக்கும் ஒழுங்குமுறை (Word Sense Disambiguation system)

கணினி மொழியியலில், சொல்-அர்த்த மயக்கநீக்கம் (WSD) என்பது ஒரு வாக்கியத்தில் ஒரு சொல்லின் எந்த அர்த்தம் பயன்படுத்தப்படுகிறது என்பதை அடையாளம் காண்பது தொடர்பான ஒரு திறந்த சிக்கலாகும். இந்த சிக்கலுக்கான தீர்வு கணினி தொடர்பான பிற எழுத்து உரைகளைப் பாதிக்கிறது; அதாவது கருத்தாடல், தேடுபொறிகளின் பொருத்தத்தை மேம்படுத்துதல், முற்சுட்டுத் தீர்மானம், ஒத்திசைவு மற்றும் அனுமானம்.

மனித மூளை சொல் அர்த்த மயக்கநீக்கத்தில் மிகவும் திறமையானது. இயல்பான மொழி இவ்வளவு தேவைப்படும் வகையில் உருவாகிறது என்பது நரம்பியல் யதார்த்தத்தின் (neurologic reality) பிரதிபலிப்பாகும். வேறு வார்த்தைகளில் கூறுவதானால், மனித மொழி மூளையின் நரம்பியல் நெட்வொர்க்குகள் வழங்கிய உள்ளார்ந்த திறனைப் பிரதிபலிக்கும் வகையில் (வடிவமைக்க உதவியது) வளர்ந்தது. கணினி அறிவியல் மற்றும் அது இயக்கும் தகவல் தொழில்நுட்பத்தில், இயற்கை மொழி செயலாக்கம் (natural language processing) மற்றும் இயந்திரம் கற்றல் ஆகியவற்றைச் செய்வதற்கான கணினிகளில் திறனை வளர்ப்பது நீண்டகால சவாலாக உள்ளது.

சொல்சார் வளங்களில் குறியாக்கம் செய்யப்பட்ட அறிவைப் பயன்படுத்தும் அகராதி அடிப்படையிலான முறைகள் முதல், கைமுறையாக அர்த்தம் அடையாளப்படுத்தப்பட்ட எடுத்துக்காட்டுகளின் ஒரு தரவுத்தொகுதியில் ஒவ்வொரு தனித்துவமான சொல்லுக்கும் ஒரு வகைப்படுத்தி பயிற்சியளிக்கப்படும் மேற்பார்வையிடப்பட்ட இயந்திர கற்றல் முறைகள் (supervised machine learning methods) வரை, அர்த்தங்களைத் தூண்டும் சொற்களின் நிகழ்வுகளைக் கொத்தாக்கம் செய்யும் முழுமையாக மேற்பார்வை செய்யப்படாத முறைகள் (completely unsupervised methods) வரை பலவிதமான நுட்பங்கள் ஆராய்ச்சி செய்யப்பட்டுள்ளன. இவற்றில், மேற்பார்வையிடப்பட்ட கற்றல் அணுகுமுறைகள் இன்றுவரை மிகவும் வெற்றிகரமான வழிமுறைகளாக உள்ளன.

தற்போதைய வழிமுறைகளின் துல்லியத்தை பல எச்சரிக்கைகள் இல்லாமல் குறிப்பிடுவது கடினம். ஆங்கிலத்தில், கரடுமுரடாகப் பதம்செய்யப்பட்ட (ஒப்புருமொழி/ஹோமோகிராஃப்) மட்டத்தில் துல்லியம் வழக்கமாக 90%ஐ விட அதிகமாக உள்ளது; குறிப்பிட்ட ஒப்புருமொழிகளில் (ஹோமோகிராஃப்களில்) சில முறைகள் 96%க்கும் அதிகமான துல்லியமானவை. நேர்த்தியாகப் பதம்செய்யப்பட்ட அர்த்த வித்தியாசங்கள் மீதான மதிப்பீடுப் பயிற்சிகளில் (செம்இவல்-200/ SemEval-2007, சென்செவல்-2/Senseval-2) 59.1% முதல் 69.0% வரையிலான சிறந்த துல்லியங்கள் பதிவாகியுள்ளன; இங்கு எப்போதும் அடிக்கடி அர்த்தத்தைத் தேர்ந்தெடுக்கும் எளிய சாத்தியமான வழிமுறையின் அடிப்படைத் துல்லியம் முறையே 51.4% ஆகும் மற்றும் 57%.

சொற்பொருள் மயக்கநீக்கம் பற்றி

சொற்பொருள் மயக்கநீக்கத்திற்கு இரண்டு கடுமையான உள்ளீடுகள் தேவைப்படுகின்றன: மயக்கநீக்கம் செய்ய வேண்டிய அர்த்தங்களைக் குறிப்பிடுவதற்கான ஒரு அகராதி மற்றும் பொருள்மயக்கம் நீக்கப்படவேண்டிய மொழித் தரவுகளின் தரவுத்தொகுதி (சில முறைகளில், மொழி எடுத்துக்காட்டுகளின் பயிற்சித் தரவுத்தொகுதி தேவைப்படுகிறது). சொற்பொருள் மயக்கநீக்கச் செயல்பாடு இரண்டு வகைகளைக் கொண்டுள்ளது: "சொல்சார் மாதிரி" ("lexical sample") மற்றும் "அனைத்து சொற்களும்" செயல்பாடு ("all words" task). முன்னர் தேர்ந்தெடுக்கப்பட்ட இலக்குச் சொற்களின் ஒரு சிறிய மாதிரியின் நிகழ்வுகளை வேறுபடுத்துவதை முந்தையது உள்ளடக்கும்; பிந்தையதில் பொருள்மயக்கநீக்கம் செய்யவேண்டிய இயங்கும் உரையின் அனைத்து சொற்களும் இருக்க வேண்டும். பிந்தையது மிகவும் யதார்த்தமான மதிப்பீட்டு வடிவமாகக் கருதப்படுகிறது, ஆனால் கார்பஸ் உற்பத்தி

செய்வதற்கு மிகவும் விலை உயர்ந்தது, ஏனென்றால் மனித சிறுகுறிப்பாளர்கள் ஒவ்வொரு வார்த்தையின் வரையறைகளையும் ஒவ்வொரு முறையும் ஒரு குறிச்சொல் தீர்ப்பை எடுக்க வேண்டியிருக்கும் போது, ஒரு தொகுதிக்கு ஒரு முறை அல்ல ஒரே இலக்கு வார்த்தைக்கான நிகழ்வுகள்.

இவை அனைத்தும் எவ்வாறு செயல்படுகின்றன என்பதைக் குறிக்க, "மாலை" என்ற (எழுதப்பட்ட) சொல்லின் தனித்துவமான அர்த்தங்களின் மூன்று எடுத்துக்காட்டுகளைக் கவனியுங்கள்:

1. கழுத்தில் அணியும் ஒன்று
2. அந்தி நேரம்

மற்றும் வாக்கியங்கள்:

1. அவள் கழுத்தில் மாலை அணிந்திருக்கிறாள்.
2. அவள் மாலை நேரத்தில் உலாவப் போனாள்.

தமிழைப் புரிந்துகொள்ளும் நபர்களுள், முதல் வாக்கியம் மேலே தந்துள்ள முதல் அர்த்தத்தில் "மாலை (கழுத்தில் அணிவது)" என்ற சொல்லைப் பயன்படுத்துகிறது என்றும் இரண்டாவது வாக்கியத்தில் "மாலை (நேரம்)" என்ற இரண்டாவது அர்த்தத்தில் சொல் பயன்படுத்தப்படுகிறது என்றும் அறிந்துகொள்வர். இந்த மனிதத் திறனை பிரதிபலிக்க வழிமுறைகளை உருவாக்குவது பெரும்பாலும் கடினமான பணியாக இருக்கலாம்.

வரலாறு

1940களில் இயந்திர மொழிபெயர்ப்பின் ஆரம்ப நாட்களில் சொற்பொருள் மயக்கநீக்கம் முதன்முதலில் ஒரு தனித்துவமான கணக்கீட்டு/கணினிசார் செயல்பாடாக வடிவமைக்கப்பட்டது, இது கணினி மொழியியலில் மிகப் பழமையான சிக்கல்களில் ஒன்றாகும். வாரன் வீவர் (Warren Weaver) தனது புகழ்பெற்ற 1949ஆம் ஆண்டு மொழிபெயர்ப்பு பதிவுக்குறிப்பில் (memorandum on translation) முதலில் ஒரு கணினிசார் சூழலில் சிக்கலை அறிமுகப்படுத்தினார். ஆரம்பகால ஆராய்ச்சியாளர்கள் சொற்பொருள் மயக்கநீக்கத்தின் முக்கியத்துவத்தையும் சிரமத்தையும் நன்கு புரிந்து கொண்டனர். பார்-ஹில்லெல் (Bar-Hillel 1960) கீழ்கண்ட உதாரணத்தைப் பயன்படுத்திச் சொற்பொருள் மயக்கத்தை "எலக்ட்ரானிக் கம்ப்யூட்டர்" மூலம் தீர்க்க இயலாது, ஏனெனில் இது பொதுவாக அனைத்து உலக அறிவையும் மாதிரிப்படுத்துவதை வேண்டும் என்றார்.

”bass" என்றசொல் கீழ்காணும் வாக்கியங்கள் அடிப்படையில் பின்வரும் மூன்று அர்தங்களில் பயன்படுத்தப்படும்:

I went fishing for some sea bass.

The bass line of thesog is too weak.

1.ஒரு வகை மீன்

2.குறைந்த அதிர்வெண் சுரங்கள்

3.ஒரு வகை கருவி

1970களில், சொற்பொருள் மயக்கநீக்கம் என்பது வில்க்ஸின் விருப்பத்தேர்வு பொருண்மையியலில் (Wilks' preference semantics) தொடங்கி செயற்கை நுண்ணறிவுத் துறையில் உருவாக்கப்பட்ட பொருண்மையியல்சார் பொருள்விளக்க ஒழுங்குமுறைகளின் துணைச் செயல்பாடாகும். இருப்பினும், சொற்பொருள் மயக்கநீக்கம் ஒழுங்குமுறைகள் அந்த நேரத்தில் பெரும்பாலும் விதி அடிப்படையிலானவை மற்றும் கையால் குறியிடப்பட்டவை என்பதால் அவை அறிவு ஈட்டச் சிக்கலுக்கு ஆளாயின.

1980களில், ஆக்ஸ்போர்டு மேம்பட்ட கற்பவரின் தற்கால ஆங்கிலம் அகராதி (Oxford Advanced Learner's Dictionary of Current English (OALD) போன்ற பெரிய அளவிலான சொற்சார் வளங்கள் கிடைத்தன: கை-குறியீட்டு முறை இந்த வளங்களிலிருந்து தானாகவே பிரித்தெடுக்கப்பட்ட அறிவால் மாற்றப்பட்டது, ஆனால் சொற்பொருண்மை மயக்கநீக்கம் இன்னும் அறிவு சார்ந்த அல்லது அகராதி அடிப்படையாக இருந்தது .

1990களில், புள்ளிவிவரப் புரட்சி கணினியல் மொழியியல் மூலம் பரவியது, மற்றும் சொற்பொருள் மயக்கநீக்கம் மேற்பார்வையிடப்பட்ட இயந்திரம் கற்றல் நுட்பங்களைப் (supervised machine learning techniques) பயன்படுத்துவதற்கான ஒரு முன்னுதாரண சிக்கலாக மாறியது.

2000களில் மேற்பார்வையிடப்பட்ட நுட்பங்கள் துல்லியத்தின் மேன்மையை அடைந்தது, எனவே கவனம் கரடுமுரடான அர்த்தங்கள், டொமைன் தழுவல், பகுதி மேற்பார்வை மற்றும் மேற்பார்வை செய்யப்படாத தரவுத்தொகுதி அடிப்படையிலான ஒழுங்குமுறைகள், வெவ்வேறு முறைகளின் சேர்க்கைகள் மற்றும் வரைபடம் அடிப்படையிலான முறைகள் மூலம் அறிவு சார்ந்த ஒழுங்குமுறைகளின் திரும்பிவரவு ஆகியவற்றுக்குக் கவனம் திருப்பப்பட்டுள்ளது. இன்னும், மேற்பார்வையிடப்பட்ட அமைப்புகள் தொடர்ந்து சிறப்பாகச் செயல்படுகின்றன.

அணுகுமுறைகள் மற்றும் முறைகள்

அனைத்து இயற்கை மொழி செயலாக்கத்தையும் போலவே, சொற்பொருள் மயக்கநீக்கத்திற்கு இரண்டு முக்கிய அணுகுமுறைகள் உள்ளன - ஆழமான அணுகுமுறைகள் deep approaches) மற்றும் ஆழமற்ற அணுகுமுறைகள் (shallow approaches).

ஆழ்ந்த அணுகுமுறைகள் உலக அறிவின் விரிவான திண்மைக்கான அணுகலைக் கருதுகின்றன. " you can go fishing for a type of fish, but not for low frequency sounds" மற்றும் " songs have low frequency sounds as parts, but not types of fish " போன்ற அறிவு பின்னர் bass என்ற வார்த்தையை எந்த அர்த்தத்தில் பயன்படுகிறது என்பதைத் தீர்மானிக்கப் பயன்படுத்தப்பட்டது. இந்த அணுகுமுறைகள் நடைமுறையில் மிகவும் வெற்றிகரமாக இல்லை, முக்கியமாக இதுபோன்ற அறிவுத் திண்மை கணினி-படிக்கக்கூடிய வடிவத்தில், மிகக் குறைந்த களங்களுக்கு வெளியே இல்லை. இருப்பினும், அத்தகைய அறிவு இருந்திருந்தால், ஆழமான அணுகுமுறைகள் ஆழமற்ற/மேலோட்டமான அணுகுமுறைகளை விட மிகவும் துல்லியமாக இருக்கும். மேலும், கணினி மொழியியலில் ஒரு நீண்ட பாரம்பரியம் உள்ளது, குறியீட்டு அறிவின் அடிப்படையில் மற்றும் சில சந்தர்ப்பங்களில் இத்தகைய அணுகுமுறைகளை முயற்சிப்பது, சம்பந்தப்பட்ட அறிவு மொழியியல் அல்லது உலக அறிவு என்பதை தெளிவாகக் கூறுவது கடினம். முதல் முயற்சி, 1950களில் இங்கிலாந்தில் உள்ள கேம்பிரிட்ஜ் மொழி ஆராய்ச்சி பிரிவில் மார்கரெட் மாஸ்டர்மேன் (Margaret Masterman) மற்றும் அவரது சகாக்களின் முயற்சியாகும். இந்த முயற்சி தலைப்புகளின் குறிகாட்டியாக ரோஜெட்டின் தெசோரஸின் பஞ்ச்-கார்டு பதிப்பையும் அதன் எண்ணிடப்பட்ட "தலைகளையும்" பயன்படுத்தியது மற்றும் ஒரு தொகுப்பு குறுக்குவெட்டு வழிமுறையைப் பயன்படுத்தி உரையில் மறுநிகழ்வுகளைத் தேடியது. இது மிகவும் வெற்றிகரமாக இல்லை, ஆனால் பிற்கால செயல்பாடுகளுக்கு வலுவான உறவுகளைக் கொண்டிருந்தது, குறிப்பாக 1990களில் ஒரு ஆய்வக முறையின் யாரோவ்ஸ்கியின் இயந்திர கற்றல் தேர்வுமுறை.

ஆழமில்லா/மேலோட்டமான அணுகுமுறைகள் உரையைப் புரிந்துகொள்ள முயற்சிக்கவில்லை. " if bass has words sea or fishing nearby, it probably is in the fish sense; if bass has the words music or song nearby, it is probably in the music sense" போன்ற தகவல்களைப் பயன்படுத்தி அவர்கள் சுற்றியுள்ள சொற்களைக் கருதுகின்றனர். சொல் அர்த்தங்களுடன் அடையாளப்படுத்தப்பட்ட சொற்களின் பயிற்சி தரவுத்தொகுதியைப்

பயன்படுத்தி இந்த விதிகள் கணினியால் தானாகவே பெறப்படலாம். இந்த அணுகுமுறை, கோட்பாட்டளவில் ஆழமான அணுகுமுறைகளைப் போல சக்திவாய்ந்ததாக இல்லை என்றாலும், கணினியின் வரையறுக்கப்பட்ட உலக அறிவின் காரணமாக, நடைமுறையில் சிறந்த முடிவுகளை அளிக்கிறது. இருப்பினும், tree மற்றும் dogs இரண்டிற்கும் அருகில் bark என்ற வார்த்தையைக் கொண்டிருக்கும் The dogs bark at the tree போன்ற வாக்கியங்களால் குழப்பமடையலாம்.

சொற்பொருள் மயக்கநீக்கத்திற்கு நான்கு வழக்கமான அணுகுமுறைகள் உள்ளன:

1. அகராதி மற்றும் அறிவுச் சார்ந்த முறைகள் (Dictionary- and knowledge-based methods): இவை எந்தவொரு தரவுத்தொகுதி ஆதாரங்களையும் பயன்படுத்தாமல், முதன்மையாக அகராதிகள், சொற்களஞ்சியம் மற்றும் சொல்சார் அறிவுத் தளங்களை நம்பியுள்ளன.
2. அரை மேற்பார்வையிடப்பட்ட அல்லது குறைந்த மேற்பார்வையிடப்பட்ட முறைகள் (Semi-supervised or minimally supervised methods): இவை ஒரு சிறிய அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி போன்ற பூட்ஸ்ட்ராப்பிங் செயல்பாட்டில் விதை தரவுகளாக அல்லது சொல்-ஒழுங்கமைக்கப்பட்ட இருமொழித் தரவுத்தொகுதி போன்ற அறிவின் இரண்டாம் நிலை மூலத்தைப் பயன்படுத்துகின்றன.
3. மேற்பார்வையிடப்பட்ட முறைகள் (Supervised methods): இவை பயிற்சியளிக்க உணர்வு-சிறுகுறிப்பு நிறுவனத்தைப் பயன்படுத்துகின்றன.
4. மேற்பார்வை செய்யப்படாத முறைகள் (Unsupervised methods): இவை முற்றிலும் வெளிப்புற தகவல்களைத் தவிர்த்து, அடையாளப்படுத்தப்படாத கச்சாத் தரவுத்தொகுதியிலிருந்து நேரடியாகச் செயல்படுகின்றன. இந்த முறைகள் சொல் அர்த்தப் பாகுபாடு (word sense discrimination) என்ற பெயரிலும் அறியப்படுகின்றன.

இந்த எல்லா அணுகுமுறைகளும் பொதுவாக ஒவ்வொரு சொல்லையும் சுற்றியுள்ள n உள்ளடக்க சொற்களின் சாளரத்தைத் தரவுத்தொகுதியில் தெளிவுபடுத்துவதன் மூலமும் மேலும் அந்த n சுற்றியுள்ள சொற்களை புள்ளிவிவர ரீதியாக பகுப்பாய்வு செய்தும் செயல்படுகின்றன. பயிற்சியளிப்பதற்கும் பின்னர் தெளிவுபடுத்துவதற்கும் பயன்படுத்தப்படும் இரண்டு ஆழமில்லா/மேலோட்டமான அணுகுமுறைகள் நேவ் பேய்ஸ் வகைப்படுத்திகள் (Naïve Bayes classifiers) மற்றும் முடிவுக்/தீர்மானக் கிளைகள் (decision trees). சமீபத்திய ஆராய்ச்சியில், ஆதரவு திசையன் இயந்திரங்கள் (support vector machines) போன்ற கர்னல் அடிப்படையிலான முறைகள் (kernel-based methods) மேற்பார்வையிடப்பட்ட கற்றலில் சிறந்த செயல்திறனைக்

காட்டியுள்ளன. வரைபட அடிப்படையிலான அணுகுமுறைகள் (Graph-based approaches) ஆராய்ச்சி சமூகத்திலிருந்தும் அதிக கவனத்தை ஈர்த்துள்ளன; மேலும் தற்போது இன்றைய நிலைக்கு நெருக்கமான செயல்திறனை அடைந்துள்ளன.

அகராதி- மற்றும் அறிவு சார்ந்த முறைகள் (Dictionary- and knowledge-based methods)

லெஸ்க் வழிமுறை (Lesk algorithm) (Lesk 1986) என்பது வளருகிற அகராதி அடிப்படையிலான முறை (seminal dictionary-based method) ஆகும். உரையில் ஒன்றாகப் பயன்படுத்தப்படும் சொற்கள் ஒன்றுக்கொன்று தொடர்புடையவை என்பதையும், சொற்களின் வரையறைகளிலும் அவற்றின் அர்த்தங்களிலும் அந்தத் தொடர்பைக் காணலாம் என்ற கருதுகோளின் அடிப்படையில் இது அமைந்துள்ளது. இரண்டு (அல்லது அதற்கு மேற்பட்ட) சொற்கள் அவற்றின் அகராதி வரையறைகளில் மிகப் பெரிய சொல் மேலுறல் கொண்ட அகராதி அர்த்தங்களைக் கண்டுபிடிப்பதன் மூலம் பொருள்மயக்கநீக்கம் செய்யப்படுகின்றன. எடுத்துக்காட்டாக, "pine cone" இல் உள்ள சொற்களைத் பொருள்மயக்கநீக்கம் செய்யும் போது, பொருத்தமான அர்த்தங்களின் வரையறைகள் *evergreen* மற்றும் *tree* (குறைந்தது ஒரு அகராதியிலாவது) ஆகிய சொற்களை அடங்கும். இதேபோன்ற அணுகுமுறை (Diamantini et al 2015) இரண்டு சொற்களுக்கு இடையேயான குறுகிய பாதையைத் தேடுகிறது: இரண்டாவது சொல் முதல் சொல்லின் ஒவ்வொரு பொருண்மைசார் மாறுபாட்டின் வரையறைகளுக்கிடையில் மீண்டும் தேடப்படுகிறது, பின்னர் முந்தைய வரையறைகளில் ஒவ்வொரு சொல்லின் ஒவ்வொரு பொருண்மைசார் மாறுபாட்டின் வரையறைகளுக்கிடையில் மீண்டும் தேடப்படுகிறது; இவ்வாறு தொடரப்படும். இறுதியாக, முதல் சொல்லிலிருந்து இரண்டாவது சொல்லுக்கான தூரத்தைக் குறைக்கும் பொருண்மைசார் மாறுபாட்டைத் தேர்ந்தெடுப்பதன் மூலம் முதல் சொல் பொருண்மை மயக்கம் நீக்கப்படுகிறது.

வரையறைகளைப் பயன்படுத்துவதற்கு மாற்றாக, பொதுவான சொல்-அர்த்தத் தொடர்பைக் கருத்தில் கொள்வதும், சொல்வலை/வேர்ட்நெட் போன்ற கொடுக்கப்பட்ட சொல்சார்/லெக்சிகல் அறிவுத் தளத்தின் அடிப்படையில் ஒவ்வொரு இணை/ஜோடி சொல் அர்த்தங்களின் பொருண்மைசார் ஒற்றுமையைக் கணக்கிடுவதும் ஆகும். செயற்கை அறிவு ஆராய்ச்சியின் ஆரம்ப நாட்களின் பரவல் செயல்படுத்தும் ஆராய்ச்சியை (spreading activation research) நினைவூட்டும் வரைபட அடிப்படையிலான முறைகள் (Graph-based methods) சில வெற்றிகளுடன் பயன்படுத்தப்பட்டுள்ளன. மிகவும் சிக்கலான வரைபட அடிப்படையிலான

அணுகுமுறைகள் கிட்டத்தட்ட அதேபோல் மேற்பார்வையிடப்பட்ட முறைகளையும் (Navigli & Velardi 2005: 1063-1074) அல்லது குறிப்பிட்ட களங்களில் அவற்றை விஞ்சுவதைக் காட்டுகின்றன. (Navigli, Litkowski & Hargraves 2007: 30-35; Agirre, Lopez de Lacalle & Soroa 2009: 10). சமீபத்தில், degree போன்ற எளிய வரைபட இணைப்பு நடவடிக்கைகள், போதுமான வளமையான சொல்சார் அறிவுத் தளத்தின் முன்னிலையில் அதிநவீன சொற்பொருள் மயக்கநீக்கத்தைச் செய்கின்றன என்று தெரிவிக்கப்பட்டுள்ளது. Navilgli & Lapata (2010: 678-692). மேலும், விக்கிபீடியாவிலிருந்து வேர்ட்நெட்டுக்கு பொருண்மைசார் உறவுகளின் வடிவத்தில் அறிவை தானாக மாற்றுவது எளிய அறிவு அடிப்படையிலான முறைகளை மேம்படுத்துவதாகவும், சிறந்த மேற்பார்வை செய்யப்பட்ட அமைப்புகளுக்கு போட்டியாகவும், டொமைன்-குறிப்பிட்ட அமைப்பில் அவற்றை விஞ்சவும் உதவுகிறது. (Ponzetto & Navigli 2010: 1522-1531).

தேர்வு விருப்பத்தேர்வுகள் (அல்லது தேர்வுக் கட்டுப்பாடுகள்) பயன்படுத்துவதும் பயனுள்ளதாக இருக்கும். எடுத்துக்காட்டாக, ஒருவர் பொதுவாக உணவை சமைக்கிறார் என்பதை அறிந்து, ஒருவர் "I am cooking basses"-இல் bass என்ற சொல்லை வேறுபடுத்தலாம் (அதாவது, இது ஒரு இசைக்கருவி அல்ல).

மேற்பார்வை/கண்காணிக்கப்பட்ட முறைகள் (Supervised methods)

மேற்பார்வை/கண்காணிக்கப்பட்ட செய்யப்பட்ட முறைகள் சூழல் சொற்பொருண்மை மயக்கம் நீக்குவதற்குப் போதுமான ஆதாரங்களை வழங்க முடியும் என்ற அனுமானத்தின் அடிப்படையில் அமைந்துள்ளது (எனவே, பொது அறிவு மற்றும் பகுத்தறிவு தேவையற்றதாகக் கருதப்படுகிறது). பண்புக்கூறுத் தேர்வு feature selection, அளவுரு மிகையாக்கம் parameter optimization மற்றும் குழுமக் கற்றல் (ensemble learning) போன்ற தொடர்புடைய நுட்பங்கள் உட்பட, ஒவ்வொரு இயந்திரக் கற்றல் வழிமுறையும் சொற்பொருண்மை மயக்கம் நீக்குவதற்குப் பயன்படுத்தப்படலாம். ஆதரவு திசையன் இயந்திரங்கள் (Support Vector Machines) மற்றும் நினைவக அடிப்படையிலான கற்றல் (memory-based learning) ஆகியவை இன்றுவரை மிகவும் வெற்றிகரமான அணுகுமுறைகளாகக் காட்டப்பட்டுள்ளன, ஏனெனில் அவை பண்புக்கூறு இடத்தின் உயர் பரிமாணத்தை சமாளிக்கக்கூடும். எவ்வாறாயினும், இந்த மேற்பார்வையிடப்பட்ட முறைகள் ஒரு புதிய அறிவு ஈட்டச் சிக்கலுக்கு உட்பட்டுள்ளன; ஏனெனில் அவை பயிற்சிக்காக, உருவாக்க அதிக உழைப்பு மற்றும் விலை தேவைப்படும்

அர்த்தம் கைமுறையாக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகளின் (sense-tagged corpora) கணிசமான அளவை நம்பியுள்ளன.

அரை மேற்பார்வை முறைகள்

பயிற்சி தரவு இல்லாததால், பல சொல் அர்த்தமயக்கநீக்க வழிமுறைகள் அரை மேற்பார்வையிடப்பட்ட கற்றலைப் பயன்படுத்துகின்றன, இது புலக்குறிப்பு செய்யப்பட்ட மற்றும் புலக்குறிப்பு செய்யப்படாத தரவை அனுமதிக்கிறது. யாரோவ்ஸ்கி வழிமுறை அத்தகைய வழிமுறையின் ஆரம்ப எடுத்துக்காட்டு (Yarowsky 1995, pp. 189–196). இது சொல் அர்த்த மயக்கநீக்கத்திற்காக மனித மொழிகளின் பண்புகளான 'ஒரு இணைவமைதிக்கு (collocation) ஒரு அர்த்தம்' மற்றும் 'ஒரு கருத்தாடலுக்கு ஒரு அர்த்தம்' என்பனவற்றைப் பயன்படுத்துகிறது. உற்றுக்கவனித்தலிருந்து, சொற்கள் கொடுக்கப்பட்ட கருத்தாடலிலும் கொடுக்கப்பட்ட இணவமைதியிலும் ஒரே ஒரு அர்த்தத்தை மட்டுமே வெளிப்படுத்துகின்றன.

பூட்ஸ்ட்ராப்பிங் அணுகுமுறை (bootstrapping approach) ஒவ்வொரு வார்த்தையின் சிறிய அளவிலான விதைத் தரவுகளிலிருந்து (seed data) தொடங்குகிறது: கைமுறையாக அடையாளப்படுத்தப்பட்ட பயிற்சி எடுத்துக்காட்டுகள் அல்லது குறைந்த எண்ணிக்கையிலான உறுதியான முடிவு விதிகள் (எ.கா., ' bass' சூழலில் ' 'play' ' எப்போதும் இசைக்கருவியைக் குறிக்கிறது). எந்தவொரு மேற்பார்வையிடப்பட்ட முறையையும் பயன்படுத்தி, ஆரம்ப வகைப்படுத்தியைப் (initial classifier) பயிற்றுவிக்க விதைகள் பயன்படுத்தப்படுகின்றன. இந்த வகைப்படுத்தியானது தரவுத்தொகுதியின் அடையாளப்படுத்தப்படாத பகுதியில் மிகவும் நம்பிக்கையான வகைப்பாடுகள் மட்டுமே உட்படுத்தப்பட்டுள்ளன ஒரு பெரிய பயிற்சி தொகுப்பைப் பிரித்தெடுக்கப் பயன்படுத்தப்படுகிறது. முழு தரவுத்தொகுதியும் நுகரப்படும் வரை அல்லது கொடுக்கப்பட்ட அதிகபட்ச எண்ணிக்கையிலான மறு செய்கைகளை அடையும் வரை ஒவ்வொரு புதிய வகைப்படுத்தியும் அடுத்தடுத்து பெரிய பயிற்சி தரவுத்தொகுதியில் பயிற்சியளிக்கப்படுகிற செயல்முறை மீண்டும் நிகழ்கிறது.

அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகளைத் துணைநிறைவு செய்யும் இணைநிகழ்வுத் தகவல்களை வழங்க பிற அரை-மேற்பார்வையிடப்பட்ட நுட்பங்கள் பெரிய அளவிலான அடையாளப்படுத்தப்படாத தரவுத்தொகுதிகளைப் பயன்படுத்துகின்றன. இந்த நுட்பங்கள், வெவ்வேறு களங்களுக்கு மேற்பார்வை செய்யப்பட்ட மாதிரிகளின் ஏற்றுக்கொள்கைக்கு உதவுகின்றன.

மேலும், ஒரு மொழியில் பொருண்மைமயக்கம் உள்ள ஒரு சொல் பெரும்பாலும் சொல்லின் அர்த்ததைப் பொறுத்து இரண்டாவது மொழியில் வெவ்வேறு சொற்களில் மொழிபெயர்க்கப்படும். மொழியைக் கடந்த அர்த்த வேறுபாடுகளை ஊகிக்க சொல்-ஒழுங்கமைக்கப்பட்ட இருமொழி தரவுத்தொகுதி (Word-aligned bilingual corpora) பயன்படுத்தபடுகிறது; இது ஒரு வகையான அரை மேற்பார்வை ஒழுங்குமுறை ஆகும்.

மேற்பார்வை செய்யப்படாத முறைகள் (Unsupervised methods)

மேற்பார்வை செய்யப்படாத கற்றல் சொற்பொருண்மை மயக்கநீக்கம் ஆராய்ச்சியாளர்களுக்கு மிகப்பெரிய சவாலாக உள்ளது. ஒத்த அர்த்தங்கள் ஒத்த சூழல்களில் நிகழ்கின்றன என்பதன் உள்ளார்ந்த அனுமானம், இதனால் சூழல் ஒற்றுமையின் சில அளவைப் (Schütze 1998: 97–123) பயன்படுத்தி சொல் நிகழ்வுகளை கிளஸ்டரிங்/கொத்தாக்கம் செய்வதன் மூலம் அர்த்தங்களை உரையிலிருந்து தூண்டலாம்; இந்தச் செயல்பாடு சொல் அர்த்தத் தூண்டல் (word sense induction) அல்லது பாகுபாடு (discrimination) எனக் குறிப்பிடப்படுகிறது. பின்னர், சொல்லின் புதிய நிகழ்வுகளை நெருங்கிய தூண்டப்பட்ட கொத்துகளாக அல்லது அர்த்தங்களாக வகைப்படுத்தலாம். இம்முறை மேலே விவரிக்கப்பட்ட பிற முறைகளை விட செயல்திறன் குறைவாக உள்ளது; ஆனால் ஒப்பீடுகள் கடினம், ஏனெனில் தூண்டப்பட்ட அர்த்தங்கள் அறியப்பட்ட சொல் அர்த்தங்களின் அகராதிக்கு மாற்றப்பட வேண்டும். அகராதி அர்த்தங்களின் தொகுப்பிற்கு/கணத்திற்கு மேப்பிங்/பொருத்துதல் விரும்பப்படவில்லை என்றால், கொத்து/கிளஸ்டர் அடிப்படையிலான மதிப்பீடுகள் (என்ட்ரோபி மற்றும் தூய்மை நடவடிக்கைகள் உட்பட) செய்யப்படலாம். மாற்றாக, சொல் அர்த்தத் தூண்டல் முறைகள் ஒரு பயன்பாட்டிற்குள் சோதிக்கப்பட்டு ஒப்பிடப்படலாம். உதாரணமாக, முடிவுக் கொத்துகளின் தரம் மற்றும் முடிவு பட்டியல்களின் அளகைப் பன்மயமாக்கத்தை அதிகரிப்பதன் மூலம் சொல் அர்த்தத் தூண்டல் வலை தேடல் முடிவு கொத்தாக்கத்தை/கிளஸ்டரிங்கை (Web search result clustering) மேம்படுத்துகிறது என்று காட்டப்பட்டுள்ளது (Navigli & Crisfulli 2010; DiMaroc & Navigli 2013). மேற்பார்வை செய்யப்படாத கற்றல், கை முயற்சியைச் சார்ந்து இல்லை ஆதலால் அறிவு ஈட்டத் தடையை சமாளிக்கும் என்று நம்பப்படுகிறது.

நிலையான அளவு அடர்த்தியான திசையன்கள் (fixed size dense vectors) (சொல் உட்பொதிப்புகள் (Word embedding)) மூலம் அவற்றின் சூழலைக் கருத்தில் கொண்டு சொற்களை உருப்படுத்தம் செய்வது பல இயற்கைமொழி ஆய்வு ஒழுங்குமுறைகளில் மிக அடிப்படையான

தொகுதிகளில் ஒன்றாக மாறியுள்ளது. பெரும்பாலான பாரம்பரிய சொல் உட்பொதித்தல் நுட்பங்கள் பல அர்த்தங்களுடன் சொற்களை ஒற்றை திசையன் பிரதிநிதித்துவத்துடன் இணைக்கின்றன என்றாலும், அவை இன்னும் சொல் அர்த்த மயக்க நீக்கத்தை மேம்படுத்த பயன்படுத்தப்படலாம். சொல் உட்பொதித்தல் நுட்பங்களுக்கு மேலதிகமாக, சொல்சார்/லெக்சிகல் தரவுத்தளங்கள் (எ.கா., வேர்ட்நெட்/WordNet, கான்செப்ட்நெட்/ConceptNet, பாபெல்நெட் /BabelNet) சொற்களையும் அவற்றின் அர்த்தங்களையும் அகராதிகளாக மேப்பிங்/பொருத்தம் செய்வதில் மேற்பார்வை செய்யப்படாத அமைப்புகளுக்கும் உதவலாம். சொல்சார்/லெக்சிக்கல் தரவுத்தளங்கள் மற்றும் சொல் உட்பொதிப்புகளை இணைக்கும் சில நுட்பங்கள் ஆட்டோ எக்ஸ்டென்ட் (AutoExtend) மற்றும் மிகவும் பொருத்தமான அர்த்த அடையாளப்படுத்தல் (Most Suitable Sense Annotation (MSSA/எம்.எஸ்.எஸ்.ஏ) ஆகியவற்றில் வழங்கப்படுகின்றன. ஆட்டோஎக்ஸ்டெண்டில், அவை சொற்கள் மற்றும் அவற்றின் சொல் அர்த்தங்கள் போன்ற அதன் பண்புகளாக ஒரு பொருள் உள்ளீட்டு பிரதிநிதித்துவத்தை (object input representation) துண்டிக்கும் முறையை அறிமுகப்படுத்துகின்றன. ஆட்டோ எக்ஸ்டெண்ட் சொற்களை (எ.கா. உரை) மற்றும் சொல் அல்லாத (எ.கா. வேர்ட்நெட்டில் உள்ள ஒத்திசைவுகள்) பொருள்களை (objects) கணுக்ககளாகவும், கணுக்களுக்கிடையிலான உறவை விளிம்புகளாகவும் வரைய/பொருத்த வரைபடக் கட்டமைப்பைப் (graph structure) பயன்படுத்துகிறது. ஆட்டோ எக்ஸ்டெண்டில் உள்ள உறவுகள் (விளிம்புகள்) அதன் கணுக்களுக்கு இடையிலான சேர்க்கையை அல்லது ஒற்றுமையை வெளிப்படுத்தலாம். முந்தையது ஆஃப்செட் கால்குலஸின் (offset calculus) பின்னால் உள்ளுணர்வைக் கைப்பற்றுகிறது, பிந்தையது இரண்டு கணுக்களுக்கு இடையிலான ஒற்றுமையை வரையறுக்கிறது. மிகவும் பொருத்தமான அர்த்த அடையாளப்படுத்தலில் (MSSA), மேற்பார்வை செய்யப்படாத ஒரு பொருண்மைமயக்கநீக்க ஒழுங்குமுறை ஒரு நிலையான சூழல் சாளரத்தில் முன் பயிற்சி பெற்ற சொல் உட்பொதித்தல் மாதிரி மற்றும் வேர்ட்நெட்டைப் பயன்படுத்தி மிகவும் பொருத்தமான சொல் அர்த்தத்தைத் தேர்ந்தெடுக்கச் சொல் அர்த்தங்களுக்கு இடையிலான ஒற்றுமையைப் பயன்படுத்துகிறது. ஒவ்வொரு சூழல் சாளரத்திற்கும், மிகவும் பொருத்தமான அர்த்த அடையாளப்படுத்தல் (MSSA) ஒவ்வொரு சொல் அர்த்தத்தின் வரையறையின் சென்ட்ராய்டை வேர்ட்நெட்டின் அர்த்தங்களில் (அதாவது, குறுகிய வரையறுக்கும் அர்த்தம் மற்றும் ஒன்று அல்லது அதற்கு மேற்பட்ட பயன்பாட்டு எடுத்துக்காட்டு) முன் பயிற்சி பெற்ற சொல் உட்பொதித்தல் மாதிரியைப் பயன்படுத்தி அதன் சொற்களின் சொல்

திசையன்களைச் சராசரியால் கணக்கிடுகின்றது. இந்த சென்ட்ராய்டுகள் ஒரு குறிக்கோள் சொல்லின் உடனடி அருகிலுள்ள அண்டை வருபவைகளுடன் (அதாவது, முன்வரும் மற்றும் பின்வரும் சொற்கள்) மிக உயர்ந்த ஒற்றுமையுடன் சொல் அர்த்ததைத் தேர்ந்தெடுக்கப் பின்னர் பயன்படுகின்றன. எல்லாச் சொற்களும் அடையாளப்படுத்தப்பட்டு, பொருண்மைமயக்கம் நீக்கப்பட்ட பிறகு, அவை எந்தவொரு நிலையான சொல் உட்பொதித்தல் நுட்பத்திலும் ஒரு பயிற்சித் தரவுத்தொகுதியாகப் பயன்படுத்தப்படலாம். அதன் மேம்பட்ட பதிப்பில், மிகவும் பொருத்தமான அர்த்த அடையாளப்படுத்தல் (எம்.எஸ்.எஸ்.ஏ.) அதன் பொருண்மைமயக்கநீக்கச் செயல்முறையை மீண்டும் மீண்டும் செய்ய சொல் அர்த்த உட்பொதிப்புகளைப் பயன்படுத்தலாம்.

ற அணுகுமுறைகள்

பிற அணுகுமுறைகள் அவற்றின் முறைகளில் வித்தியாசமாக வேறுபடலாம்:

- இயல்புநிலை தர்க்கத்தின் செயல்பாட்டுப் பொருண்மையியல் அடிப்படையில் சொற்பொருள் மயக்கநீக்கம் (Disambiguation based on operational semantics of default logic) (Galitsky, 2005).
- பொருட்புல உந்தல் சொற்பொருள் மயக்கநீக்கம் (Domain-driven disambiguation) (Gliozzo, Magnini & Strapparava 2004: 380–387. Buitelaar et al. 2006: 275–298)
- மேலாதிக்க சொல் அர்த்தங்களை அடையாளம்காணல் (Identification of dominant word senses) (McCarthy et al. 2007: 553–590, Mohammad & Hirst 2006: 121–128, Lapata & Keller 2007: 348–355)
- குறுக்கு மொழிச் சான்றுகளைப் பயன்படுத்திச் சொற்பொருள் மயக்கநீக்கம் (WSD using Cross-Lingual Evidence) (Ide, Erjavec & Tufis 2002: 54–60; Chan & Ng 2005: 1037–1042).
- பாட்டம் கோட்பாடையும் பங்கு மற்றும் குறிப்பு இலக்கணம் (Role and Reference Grammarதையும் (RRG)) இணைக்கும் ஜான் பாலின் மொழி சுதந்திர இயற்கை மொழி புரிதலில் சொற்பொருள் மயக்கநீக்கத் தீர்வு (WSD solution in John Ball's language independent NLU combining Patom Theory and RRG (Role and Reference Grammar)) (Weaver 1949).
- கட்டுப்பாட்டு அடிப்படையிலான இலக்கணங்களில் வகை அனுமானம் (Type inference in constraint-based grammars) (Stuart M. Shieber, 1992)

உள்ளூர் தடைகள் மற்றும் சுருக்கம்

அறிவு ஈட்டத் தடை/சிக்கல் சொற்பொருண்மை மயக்கநீக்கச் சிக்கலை தீர்க்க ஒரு பெரிய தடையாக இருக்கலாம். மேற்பார்வை செய்யப்படாத முறைகள், அகராதிகள் மற்றும் சொல்சார் தரவுத்தளங்களில் மட்டுமே அரிதாகவே வடிவமைக்கப்பட்டுள்ள சொல் அர்த்தங்களைப் பற்றிய அறிவை நம்பியுள்ளன, இது. மேற்பார்வையிடப்பட்ட முறைகள் ஒவ்வொரு சொல் அர்த்ததிற்கும் கைமுறையாக அடையாளப்படுத்தப்பட்ட எடுத்துக்காட்டுகளின் இருப்பைச் சாந்திருக்கின்றன; இந்தத் தேவை சோதனை நோக்கங்களுக்காக சென்செவல் பயிற்சிகளில் செய்யப்படும் ஒரு சில சொற்களுக்கு மட்டுமே உள்ளது.

சொற்பொருண்மை மயக்கநீக்க ஆராய்ச்சியின் மிகவும் நம்பிக்கைக்குரிய போக்குகளில் ஒன்று, இதுவரை அணுகக்கூடிய மிகப்பெரிய தரவுத்தொகுதியான உலகளாவிய வலையை சொல்சார் தகவல்களை தானியக்கமாகப் பெறுவதற்குப் பயன்படுத்துவது ஆகும் (Kilgarrif & Grefenstette 2003: 333–347). சொற்பொருண்மை மயக்கநீக்கம் பாரம்பரியமாக தகவல் மீட்டெடுப்பு (Information Retrieval (IR)) போன்ற பயன்பாடுகளை மேம்படுத்த இயலும் ஒரு இடைநிலை மொழி பொறியியல் தொழில்நுட்பமாக (intermediate language engineering technology) புரிந்து கொள்ளப்படுகிறது. இருப்பினும், இந்த விஷயத்தில் தலைகீழும் உண்மை ஆகும்: வலை தேடுபொறிகள் எளிய மற்றும் வலுவான தகவல் மீட்டெடுப்பு நுட்பங்களை செயல்படுத்துகின்றன, அவை சொற்பொருண்மை மயக்கநீக்கத்தில் பயன்படுத்த தகவல்களுக்கு வலையை வெற்றிகரமாக சுரங்கப்படுத்தலாம். பயிற்சி தரவின் வரலாற்றுப் பற்றாக்குறை சில புதிய வழிமுறைகள் மற்றும் நுட்பங்களின் தோற்றத்தைத் தூண்டியுள்ளது, இது சொல் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகளைத் தானியக்கமாக ஈட்டுவதில் விவரிக்கப்பட்டுள்ளது.

வெளி அறிவு வளங்கள்

அறிவு என்பது சொற்பொருள் மயக்கநீக்கத்தின் அடிப்படை அங்கமாகும். அறிவு வளங்கள் அர்த்தங்களை சொற்களுடன் தொடர்புபடுத்துவதற்கு அவசியமான தரவை வழங்குகின்றன. அவை புலக்குறிப்பு செய்யப்படாத அல்லது சொல் அர்த்தங்களால் அடையாளப்படுத்தபட்ட தரவுத்தொகுதிகள் முதல் இயந்திரத்தால் படிக்கக்கூடிய அகராதிகள், சொற்களஞ்சியம், பொருள்விளக்கச் சொற்கோவைகள், முலப்பொருண்மையியல்கள் போன்றவை வரை

மாறுபடலாம். அவை பின்வருமாறு வகைப்படுத்தப்படலாம் (Litkowski 2005: 753–761, Agirre & Stevenson 2006, pp. 217–251):

கட்டமைக்கப்பட்ட (Structured):

இயந்திரம் படிக்கக்கூடிய அகராதிகள் (Machine-readable dictionaries (MRDs)

மூலப்பொருண்மியியல்கள் (Ontologies)

சொற்களஞ்சியங்கள் Thesauri

கட்டமைக்கப்படாதது (Unstructured):

கூட்டு வளங்கள் (Collocation resources)

பிற வளங்கள் (Other resources) (சொல் அதிர்வெண் பட்டியல்கள், நிறுத்தப்பட்டியல்கள், பொருட்புலப் புலக்குறிப்புகள், போன்றவை) (such as word frequency lists, stoplists, domain labels,etc.)

தரவுத்தொகுதிகள் (Corpora): கச்சா தரவுத்தொகுதிகள் (raw corpora) மற்றும் அர்த்தம் அடையாளப்படுத்தப்பட தரவுத்தொகுதிகள் (sense-annotated corpora)

6.4.14. சொல் வலைத் திட்டமிடல் (WordNet design)

சொல்வலை என்பது 200க்கும் மேற்பட்ட மொழிகளில் உள்ள சொற்களுக்கு இடையிலான பொருண்மையியல்சார் உறவுகளின் ஒரு சொல்சார் தரவுத்தளமாகும். சொல்வலை சொற்களை ஒருபொருள்பன்மொழிகள், உள்ளடங்குமொழிகள், சினைமொழிகள் என்பன உள்ளடங்கிய பொருண்மையியசார் உறவுகளுடன் இணைக்கிறது. ஒருபொருள்பன்மொழிகள் குறுகிய வரையறைகள் மற்றும் பயன்பாட்டு எடுத்துக்காட்டுகளுடன் ஒருபொருள்பனமொழியக் குழுமங்களாகத் தொகுக்கப்பட்டுள்ளன. சொல்வலையை ஒரு அகராதி மற்றும் சொற்களஞ்சியத்தின் கலவையாகவும் நீட்டிப்பாகவும் காணலாம். இது ஒரு வலை உலாவி வழியாக மனித பயனர்களுக்கு அணுகக்கூடியதாக இருந்தாலும், இதன் முதன்மை பயன்பாடு தானியங்கி உரை பகுப்பாய்வு மற்றும் செயற்கை நுண்ணறிவு பயன்பாடுகளில் உள்ளது. சொல்வலை முதன்முதலில் ஆங்கில மொழியில் உருவாக்கப்பட்டது (Miller et al 1990) மற்றும் ஆங்கிலச் சொல்வலைத் தரவுத்தளம் மற்றும் மென்பொருள் கருவிகள் பி.எஸ்.டி பாணி உரிமத்தின் கீழ் வெளியிடப்பட்டுள்ளன; மேலும் அவை அந்த சொல்வலைத்தளத்திலிருந்து பதிவிறக்கம் செய்ய இலவசமாகக் கிடைக்கின்றன.

வரலாறு மற்றும் குழு உறுப்பினர்கள்

சொல்வலை முதன்முதலில் ஆங்கிலத்தில் பிரின்ஸ்டன் பல்கலைக்கழகத்தின் (Princeton University) புலனறிவு அறிவியல் ஆய்வகத்தில் (Cognitive Science Laboratory) மட்டுமே உருவாக்கப்பட்டது; இது உளவியல் பேராசிரியர் ஜார்ஜ் ஆர்மிட்டேஜ் மில்லரின் (George Armitage Miller) வழிகாட்டுதலின் கீழ் 1985ஆம் ஆண்டு தொடங்கப்பட்டு சமீபத்திய ஆண்டுகளில் கிறிஸ்டியன் ஃபெல்பாம்-இன் (Christiane Fellbaum) இயக்கத்தின் கீழ் இயங்குகிறது. இந்தத் திட்டத்திற்கு ஆரம்பத்தில் யு.எஸ். கடற்படை ஆராய்ச்சி அலுவலகத்தாலும் பின்னர் தர்பா (DARPA), தேசிய அறிவியல் அறக்கட்டளை (National Science Foundation), சீர்குலைக்கும் தொழில்நுட்ப அலுவலகம் (Disruptive Technology Office) [(முன்னர் மேம்பட்ட ஆராய்ச்சி (Advanced Research) மற்றும் மேம்பாட்டு செயல்பாடு (Development Activity)] மற்றும் ரெஃப்லெக்ஸ் (REFLEX) உள்ளிட்ட பிற யு.எஸ் அரசு முகமையகங்களாலும் நிதிநல்கப்பட்டது. சொல்வலை உருவாக்கிய செயல்பாட்டிற்காக ஜார்ஜ் மில்லர் மற்றும் கிறிஸ்டியன் ஃபெல்பாம் ஆகியோருக்கு 2006 அன்டோனியோ ஜாம்பொல்லி பரிசு (Antonio Zampolli Prize) வழங்கப்பட்டது.

குளோபல் வேர்ட்நெட் அசோசியேஷன் (Global WordNet Association) என்பது வணிகமற்ற ஒரு அமைப்பாகும், இது உலகின் அனைத்து மொழிகளுக்கும் வேர்ட்நெட்களைப் பற்றி விவாதிக்க, பகிர மற்றும் இணைக்க ஒரு தளத்தை வழங்குகிறது; மேலும் கிறிஸ்டியன் ஃபெல்பாம் மற்றும் பீக் தி.ஜே.எம். வோசன் (Piek Th.J.M. Vossen) மற்றும் இணைத் தலைவர்களும் இதில் இருக்கின்றனர்.

தரவுத்தள உள்ளடக்கங்கள்

தரவுத்தளத்தில் மொத்தம் 207 016 சொல்-அர்த்த இணைகளுக்கு 175 979 ஒத்திசைவுகளில் ஏற்பாடு செய்யப்பட்ட 155 327 சொற்கள் உள்ளன; சுருக்கப்பட்ட வடிவத்தில், இது சுமார் 12 மெகாபைட் அளவு.

வேர்ட்நெட்டில் பெயர்ச்சொற்கள், வினைச்சொற்கள், பெயரடைகள், வினையடைகள் என்பன உட்படுத்தப்பட்டுள்ளன; ஆனால் முன்னொட்டுகள், அடைகொளி அடைகள் மற்றும் பிற செயல்பாட்டு சொற்கள் புறக்கணிக்கப்பட்டுள்ளன.

தோராயமாக ஒருபொருள்பன்மொழிகளாக இருக்கும் ஒரே சொல்வகைப்பாட்டைச் சேர்ந்த சொற்கள் ஒருபொருள்பன்மொழிக் குழுமங்களாகத் (சின்செட்களாக) தொகுக்கப்பட்டுள்ளன. ஒருபொருள்பன்மொழியக் குழுமங்கள் சிம்ப்ளக்ஸ் சொற்களையும் "eat out" மற்றும் "car pool"

போன்ற இணைவமைதிகளையும் (collocations) அடக்கும். பல்பொருள் ஒருமொழியச் சொல் (polysemous word) வடிவத்தின் வெவ்வேறு அர்த்தங்கள் வெவ்வேறு ஒருபொருள்பன்மொழியக் குழுமங்களுக்கு ஒதுக்கப்பட்டுள்ளன. ஒரு சின்செட்டின் பொருண்மை குறுகிய வரையறுக்கும் பொருள் மற்றும் ஒன்று அல்லது அதற்கு மேற்பட்ட பயன்பாட்டு எடுத்துக்காட்டுகளுடன் மேலும் தெளிவுபடுத்தப்படுகிறது. ஒரு ஒருபொருள்பன்மொழியக் குழுமத்தின் எடுத்துக்காட்டு பின்வருமாறு அமையும்:

> good, right, ripe – (most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes")

அனைத்து சின்செட்களும் பொருண்மைசார் உறவுகளின் மூலம் பிற சின்செட்களுடன் இணைக்கப்பட்டுள்ளன. எல்லா லெக்சிக்கல் வகைகளாலும் பகிரப்படாத இந்த உறவுகள் பின்வருமாறு:

பெயர்ச்சொற்கள் (Nouns)

• உள்ளடக்கு மொழிகள் hypernyms: ஒவ்வொரு X ஒரு (வகையான) Y ஆக இருந்தால் Y என்பது X இன் உள்ளடக்கு மொழி ஆகும் (canine என்பது dog-இன் உள்ளடக்கு மொழி)

• உள்ளடங்குமொழிகள் hyponyms: ஒவ்வொரு Y ஒரு (வகையான) X என்றால் Y என்பது X இன் ஒரு பெயராகும் (dog என்பது canine என்பதன் உள்ளங்குமொழி)

•ஒருங்கிணைப்பு மொழிகள் (coordinate terms): X மற்றும் Y உள்ளடக்கு மொழியைப் பகிர்ந்து கொண்டால் Y என்பது X இன் ஒருங்கிணைப்புச் சொல் (wolf என்பது dogஇன் ஒருங்கிணைந்த சொல், மற்றும் dog என்பது wolf ஒரு ஒருங்கிணைந்த சொல்)

• சினைமொழி (meronym): Y என்பது X இன் ஒரு பகுதியாக இருந்தால் Y என்பது X இன் சினைமொழியாகும் (window என்பது building-இன் ஒரு சினைமொழியாகும்)

• முழுமொழி (holonym): X என்பது Y இன் ஒரு பகுதியாக இருந்தால் Y என்பது X இன் முழுமொழி ஆகும் (building என்பது window ஒரு முழுமொழி ஆகும்).

வினைச்சொற்கள் (Verbs)

உள்ளடக்குமொழி (hypernym): வினை X ஒரு (வகையான) Y என்றால் வினை Y, வினை Xஇன் உள்ளடக்கு மொழி ஆகும் (*to perceive* என்பது *to listen* என்பதன் உள்ளடக்குமொழியம்)

விதமொழி (troponym): Yஇன் செயல்பாடு ஏதேனும் ஒரு விதத்தில் X செய்கிறதென்றால் வினை Y, வினை Xஇன் ஒரு விதமொழி ஆகும் (*to lisp* 'ஸகரத்தை தகரமாக உச்சித்துப் பேசு' என்பது *to talk* 'பேசு' என்பதன் விதமொழி ஆகும்)

உட்படுமொழி (entailment): நீங்கள் X-ஐச் செய்வதன் மூலம் Yஐச் செய்ய வேண்டும் என்றால் Y என்ற வினை X க்கு உட்பட்டது (to sleep 'உறங்கு' என்பது to snore 'குறட்டைவிடு' என்பதால் உட்படுத்தப்படும்)

ஒருங்கிணைப்பு மொழிகள்: பொதுவான வினைச்சொல்லைப் பகிரும் வினைச்சொற்கள் (to lisp 'ஸகரத்தை தகரமாக உச்சித்துப் பேசு' மற்றும் to yell 'கத்து')

இந்தப் பொருண்மைசார் உறவுகள் இணைக்கப்பட்ட ஒருபொருள்பன்மொழியக் குழுமங்களின் அனைத்து உறுப்பினர்களிடமும் உள்ளன. தனிப்பட்ட சின்செட் உறுப்பினர்கள் (சொற்கள்) சொல்சார் உறவுகளுடன் இணைக்கப்படலாம். எடுத்துக்காட்டாக, "director" என்ற பெயர்ச்சொல் அதில் இருந்து "உருபப்பொருண்மையியல்" இணைப்பு வழியாக பெறப்படுகிற "direct" என்ற வினைச்சொல்லுடன் (ஒரு அர்த்தம்) இணைக்கப்பட்டுள்ளது.

தரவுத்தளத்துடன் விநியோகிக்கப்பட்ட மென்பொருளின் உருபனியல் செயல்பாடுகள் பயனரின் உள்ளீட்டிலிருந்து ஒரு வார்த்தையின் லெம்மா அல்லது பகுதி வடிவத்தைக் குறைக்க முயற்சிக்கின்றன. ஒழுங்கற்ற படிவங்கள் ஒரு பட்டியலில் சேமிக்கப்படுகின்றன, மேலும் "ate" என்பதற்கு "eat" என்று திரும்பும்

அறிவு அமைப்பு

பெயர்ச்சொற்கள் மற்றும் வினைச்சொற்கள் இரண்டுமே படிநிலைகளாக ஒழுங்கமைக்கப்பட்டுள்ளன, அவை உள்ளடக்குமொழியம் அல்லது 'IS A' உறவுகளால் வரையறுக்கப்படுகின்றன. உதாரணமாக, உள்ளடக்குமொழிய படிநிலையத் தொடர்ந்து என்ற dog iசொல்லின் ஒரு அர்த்தம் காணப்படுகிறது; அதே மட்டத்தில் உள்ள சொற்கள் சின்செட் உறுப்பினர்களைக் குறிக்கும். ஒவ்வொரு ஒருபொருள் பனமொழியக் குழுமங்களும் ஒரு தனித்துவமான அட்டவணையைக் (unique index) கொண்டுள்ளன.

dog, domestic dog, Canis familiaris

- canine, canid
    - carnivore
        - placental, placental mammal, eutherian, eutherian mammal

- mammal
  - vertebrate, craniate
    - chordate
      - animal, animate being, beast, brute, creature, fauna
        - ...

மேல் மட்டத்தில், இந்த படிநிலைகள் பெயர்ச்சொற்களுக்கு 25 தொடக்க "கிளைகள்" மற்றும் வினைச்சொற்களுக்கு 15 தொடக்க "கிளைகள்" (பராமரிப்பு மட்டத்தில் அகராதியியல்சார் கோப்புகள் என அழைக்கப்படுகின்றன) என ஒழுங்கமைக்கப்பட்டுள்ளன. அனைத்தும் ஒரு தனித்துவமான தொடக்க ஒருபொருள்பன்மொழியக் குழுமம், "entity" உடன் இணைக்கப்பட்டுள்ளன. வினைச்சொல் படிநிலைகளை விட பெயர்ச்சொல் படிநிலைகள் மிகவும் ஆழமானவை.

பெயரடைகள் படிநிலை கிளைகளாக ஒழுங்கமைக்கப்படவில்லை. அதற்குப் பதிலாக, "hot" மற்றும் "cold" போன்ற இரண்டு "மைய" எதிர்ச்சொற்கள் பைனரி துருவங்களை உருவாக்குகின்றன, அதே நேரத்தில் "செயற்கைகோள்" ஒருபொருள்பன்பொழிகளான "steaming" மற்றும் "chilly" ஆகியவை அந்தந்த துருவங்களுடன் "ஒற்றுமை" உறவுகள் வழியாக இணைகின்றன. பெயரடைகளை இந்த வழியில் ”கிளைகள்” ("trees") என்று இல்லாமல் உடுக்கைகள் ("dumbbells") என்று காட்சிப்படுத்தலாம்.

உளவியல் அம்சங்கள்

வேர்ட்நெட் திட்டத்தின் ஆரம்ப குறிக்கோள் 1960களின் பிற்பகுதியில் உருவாக்கப்பட்ட மனித பொருண்மையியல்சார் நினைவகத்தின் கோட்பாடுகளுடன் பொருந்தக்கூடிய ஒரு சொல்சார் தரவுத்தளத்தை உருவாக்குவதாகும். உளவியல் சோதனைகள் பேசுபவர்கள் தங்கள் கருத்துக்களைப் பற்றிய அறிவை பொருளாதார, படிநிலை பாணியில் ஒழுங்கமைத்துள்ளனர் என்பதைக் குறிக்கின்றன. கருத்துசார் அறிவை அணுகுவதற்கு மீட்டெடுக்கும் நேரம், அறிவை அணுக "பயணிக்க" பேசுபவருக்குத் தேவைப்படும் படிநிலைகளின் எண்ணிக்கையுடன் நேரடியாக தொடர்புடையதாகத் தோன்றும். எனவே, canaries can sing என்பதை பேசுபவர்கள் விரைவாக சரிபார்க்க முடியும், ஏனெனில் கேனரி ஒரு பாடல் பறவை, ஆனால் canaries can fly என்பதை சரிபார்க்க சற்று அதிக நேரம் தேவைப்படுகிறது (அங்கு அவர்கள் "பறவை" என்ற கருத்தை சூப்பர்

ஆர்டினேட் மட்டத்தில் அணுக வேண்டியிருந்தது) மேலும் அதிக நேரம் canaries have skin என்பதைச் சரிபார்க்கவும் ("animal" வரை உள்ளங்குமொழிகளின் பல நிலைகளைக் காண வேண்டும்). இத்தகைய உளவியல் மொழி சோதனைகள் மற்றும் உள்ளார்ந்த கோட்பாடுகள் விமர்சனங்களுக்கு உட்பட்டிருந்தாலும், சொல்வலையின் சில அமைப்பு சோதனைச் சான்றுகளுடன் ஒத்துப்போகிறது. எடுத்துக்காட்டாக, சொல்வலைப் படிநிலை, ஒரு குறிப்பிட்ட பொருண்மைசார் வகைப்பாட்டிலிருந்து சொற்களை உருவாக்கும் பேசுபவர்களின் திறனை அனோமிக் அஃபாசியா (anomic aphasia) தேர்ந்தெடுத்து பாதிக்கும். எதிர்பதப் பெயரடைகள் (Antonymous adjectives) (டம்பல்/உடுக்கை கட்டமைப்பில் சொல்வலையின் மையப் பெயரடைகள்) வாய்ப்பை விட அடிக்கடி நிகழ்கின்றன; இது பல மொழிகளில் இருப்பது கண்டறியப்பட்டுள்ளது.

ஒரு சொல்சார் மெய்மையியல்

சொல்வலை சில நேரங்களில் மெய்மையியல்/இருப்பியல் என்று அழைக்கப்படுகிறது, அதன் படைப்பாளிகள் செய்யாத ஒரு தொடர்ச்சியான கூற்று. பெயர்ச்சொல் ஒருபொருள்பன்மொழியக் குழுமங்களுக்கு இடையிலான உள்ளடக்குமொழி / உள்ளடங்குமொழிய உறவுகள் கருத்துசார் வகைப்பாடுகளுக்களுகு இடையிலான சிறப்பு உறவுகள் என்று பொருள் கொள்ளலாம். வேறு வார்த்தைகளில் கூறுவதானால், சொல்வலையைக் கணினி அறிவியல் அர்த்தத்தில் ஒரு சொல்சார் மெய்மையியல் (lexical ontology) என விளக்கிப் பயன்படுத்தலாம். இருப்பினும், அத்தகைய மெய்மையியல் பயன்படுத்தப்படுவதற்கு முன்பு சரிசெய்யப்பட வேண்டும், ஏனெனில் இது நூற்றுக்கணக்கான அடிப்படை பொருண்மைசார் முரண்பாடுகளைக் கொண்டுள்ளது; எடுத்துக்காட்டாக, (i) பிரத்தியேக வகைப்பாடுகளுக்கான பொதுவான சிறப்பாக்கங்கள் மற்றும் (ii) சிறப்பாக்கப் படிநிலையில் மிகைகள். மேலும், சொல்வலையை அறிவு உருப்பத்தத்திற்குப் பயன்படக்கூடிய ஒரு சொல்சார் மெய்மையியல் ஆக மாற்றுவது பொதுவாக (i) சிறப்பு உறவுகளை துணைவகை (subtypeOf) மற்றும் எடுத்துக்காட்டு உறவுகள் (instanceOf) என வேறுபடுத்துவது மற்றும் (ii) ஒவ்வொரு வகைப்பாடிலும் உள்ளுணர்வு தனித்துவமான அடையாளங்காட்டிகளை இணைப்பது ஆகியவை அடங்கும். WebKB-2 இன் ஒத்துழைப்புடன் புதுப்பிக்கத்தக்க அறிவுத் தளமாக வேர்ட்நெட் 1.7ஐ ஒருங்கிணைப்பதன் ஒரு பகுதியாக இத்தகைய திருத்தங்களும் மாற்றங்களும் செய்யப்பட்டு ஆவணப்படுத்தப்பட்டிருந்தாலும், அறிவு சார்ந்த பயன்பாடுகளுக்கு சொல்வலையை மீண்டும்

பயன்படுத்துவதாகக் கூறும் பெரும்பாலான திட்டங்கள் (பொதுவாக, அறிவு சார்ந்த தகவல் மீட்டெடுப்பு) அதை நேரடியாக மீண்டும் பயன்படுத்துகிறது.

சொல்வலையில் இருந்து இணைவு உறவுகளை (association relations) தானாகவே பிரித்தெடுப்பதற்கான ஒரு கலப்பின கீழ்-கீழ் மேல்-கீழ் நெறிமுறையின் (hybrid bottom-up top-down methodology) மூலம் சொல்வலை ஒரு முறையான விவரக்குறிப்பாக (formal specification) மாற்றப்பட்டுள்ளது, மேலும் இந்த இணைவுகளை (associations) முறையாக டோல்ஸ் (DOLCE) அடித்தள மெய்மையியல் வரையறுக்கப்பட்ட கருத்துசார் உறவுகளின் (conceptual relations) குழுமத்தின்/கணத்தின் அடிப்படையில் விளக்குகிறது (Gangemi et al 2003).

சொல்வலை மெய்மையியலில் ஒருங்கிணைத்ததாகக் கூறும் பெரும்பாலான படைப்புகளில், சொல்வலையின் உள்ளடக்கம் அவசியமாகத் தோன்றும்போது அதை சரிசெய்யவில்லை; அதற்குப் பதிலாக, சொல்வலை பெரிதும் மறு-விளக்கம் செய்யப்பட்டு பொருத்தமான போதெல்லாம் புதுப்பிக்கப்படுகிறது. உதாரணமாக, ஒன்டோக்ளீன் (OntoClean) அடிப்படையிலான அணுகுமுறையின்படி அல்லது சென்சஸ் மெய்மையியலின் (SENSUS ontology) கீழ் வகுப்புகளை உருவாக்குவதற்கு சொல்வலை முதன்மை ஆதாரமாக பயன்படுத்தப்பட்டபோது, சொல்வலையின் உயர்மட்ட மெய்மையியல் (Oltramari et al 2002) மீண்டும் கட்டமைக்கப்பட்டபோது இதுதான்.

பயன்பாடுகள்

சொல்வலைத் தகவல் ஒழுங்குமுறைகளில் பல நோக்கங்களுக்காகப் பயன்படுத்தப்படுகிறது,; இதில் சொல்-அர்த்த பொருண்மைமயக்க நீக்கம், தகவல் மீட்டெடுப்பு, தானியங்கி உரை வகைப்பாடு, தானியங்கி உரைச் சுருக்கம், இயந்திர மொழிபெயர்ப்பு மற்றும் தானியங்கி குறுக்கெழுத்து புதிர் உருவாக்கம் ஆகியவை அடங்கும்.

சொல்வலையின் பொதுவான பயன்பாடு சொற்களுக்கு இடையிலான ஒற்றுமையை தீர்மானிப்பதாகும். சொல்வலையின் வரைபடக் கட்டமைப்பில் சொற்கள் மற்றும் ஒருபொருள்பன்மொழியக் குழுமங்களில் விளிம்புகளின் எண்ணிக்கையை கணக்கிடுவது போன்ற ஒருபொருள்பன்மொழியக் குழுமங்களுக்கு இடையிலான தூரத்தை அளவிடுவது உட்பட பல்வேறு வழிமுறைகள் முன்மொழியப்பட்டுள்ளன. உள்ளுணர்வு என்னவென்றால், இரண்டு சொற்கள் அல்லது ஒருபொருள்பன்மொழியக் குழுமங்கள் நெருக்கமாக இருப்பதால், அவற்றின்

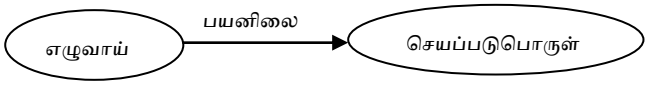
பொருண்மை நெருக்கமாக இருக்கும். சொல்வலை அடிப்படையிலான பல ஒற்றுமை வழிமுறைகள் (similarity algorithms) சொல்வலை::ஒற்றுமை எனப்படும் பெர்ல் தொகுப்பிலும், இயற்கை மொழிக் கருவித்தொகுதி (Natural Language Toolkit (NLTK/என்.எல்.டி.கே)) எனப்படும் பைதான் தொகுப்பிலும் செயல்படுத்தப்படுகின்றன. பிற அதிநவீன சொல்வலை அடிப்படையிலான ஒற்றுமை நுட்பங்கள் ADWஐ (Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity) (Pilehvar et al 2013) உட்படுத்தும்; இதன் செயல்படுத்தல் ஜாவாவில் கிடைக்கிறது. சொல்வலையைப் பிற சொற்களஞ்சியங்களை இணைக்கவும் பயன்படுத்தலாம்.

6.4.15. பொருண்மைசார் வலை/வலையைப்பு(Semantic Web or Semantic Net)

ஒரு பொருண்மை வலையமைப்பு (semantic network) அல்லது அல்லது சட்டக வலையமைப்பு (frame network) என்பது ஒரு வலையமைப்பில் உள்ள கருத்துகளுக்கு இடையிலான பொருண்மைசார் உறவுகளை குறிக்கும் அறிவுத் தளமாகும். இது பெரும்பாலும் அறிவு உருப்படுத்தத்தின் ஒரு வடிவமாகப் பயன்படுத்தப்படுகிறது. பொருண்மைசார் களங்க பொருத்தும் அல்லது இணைக்கும் இது கருத்துருக்கள், இடையிலான பொருண்மைசார் உறவுகளைக் உருப்படுத்தம் செய்யும் கருத்துருக்களையும் விளிம்புகளையும் உருப்படுத்தம் செய்யும் செங்குத்துக்களைக்கொண்ட ஒரு திசைநோக்கு அல்லது திசைநோக்கா வரைபடமாகும் (Sowa 1987).

எடுத்துக்காட்டான தரப்படுத்தப்பட்ட பொருண்மை வலையமைப்புகள் பொருண்மைசார் மும்மடங்குகளாக (semantic triples) வெளிப்படுத்தப்படுகின்றன. [மும்மடங்கு/டிரிபிள் என்பது எழுவாய்-பயனிலை-செயப்படுபொருள் வெளிப்பாடுகள் (எ.கா. "ராமன் ஒரு ஆசிரியர்", அல்லது "ராமனுக்குச் சீதையைத் தெரியும்")] வடிவத்தில் பொருண்மைசார் தரவைப் பற்றிய கூற்றை குறியமாக்கம் செய்யும் மூன்று இருப்புப்பொருள்களின் குழுமம் ஆகும்.

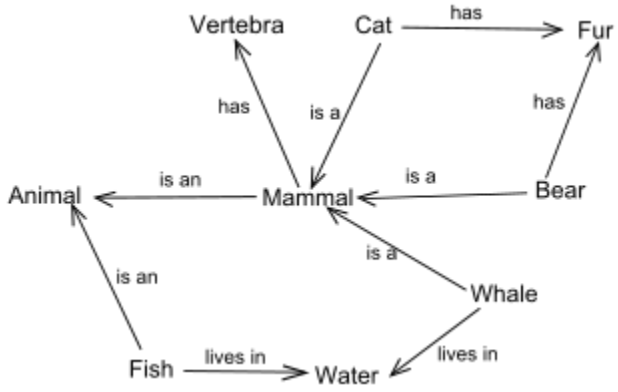
அடிப்படை பொருண்மைசார் மூன்றுமடங்கு மாதிரி.



பொருண்மைசார் பகுப்பாய்வு (semantic parsing) (Poon and Domingos 2009) மற்றும் சொற்பொருண்மை மயக்கநீக்கம் (word-sense disambiguation (WSD)) போன்ற இயற்கையான

மொழி ஆய்வுப் பயன்பாடுகளில் பொருண்மைசார் வலையமைப்புகள் பயன்படுத்தப்படுகின்றன (Sussna 1993).

பொருண்மைசார் வலையமைப்பின் எடுத்துக்காட்டு (விக்கிபீடியாவிலிருந்து எடுத்தாளப்பட்டுள்ளது)



வரலாறு

தர்க்கத்தில் பொருண்மைசார் வலையமைப்புகளின் பயன்பாட்டின் எடுத்துக்காட்டுகள், பல நூற்றாண்டுகளுக்கு முன்பு திசைநோக்கிய நினைக்குறிப்புக் கருவியாகப் பயன்படுத்திய அசைக்ளிக் வரைபடங்கள் ஆகும். கி.பி மூன்றாம் நூற்றாண்டில் அரிஸ்டாட்டில் வகைபாடுகளைப் பற்றிய கிரேக்க தத்துவஞானி போர்பிரியின் வர்ணனையே முந்தைய ஆவணப்படுத்தப்பட்ட பயன்பாடாகும்.

கணினிபயன்படுத்தல் வரலாற்றில், 1956ஆம் ஆண்டில் கேம்பிரிட்ஜ் மொழி ஆராய்ச்சி பிரிவின் ரிச்சர்ட் எச். ரிச்சன்ஸ் என்பவரால் முன்மொழியப்பட்ட கால்குலஸிற்கான "பொருண்மைசார் வலை அமைப்புகள்" இயற்கை மொழிகளின் இயந்திர மொழிபெயர்ப்பிற்கான "இடைமொழி" ஆகச் செயல்படுத்தப்பட்டன (Lehmannச்& Rodin 1992). இந்தச் செயல்பாட்டின் மற்றும் கேம்பிரிட்ஜ் மொழி ஆராய்ச்சி பிரிவின் முக்கியத்துவம் தாமதமாக உணரப்பட்டது.

விக்டர் யங்வேவின் செயல்விளக்கத்தால் ஈர்க்கப்பட்ட பின்னர் முதல் வரிசைப் பயனிலைக் கால்குலஸை (first order predicate calculus) அடிப்படையாகப் பயன்படுத்தி ராபர்ட் எஃப். சிம்மன்ஸ் (Simmons 1963) மற்றும் ஷெல்டன் க்ளெய்ன் ஆகியோரால் பொருண்மைசார் வலையமைப்புகள் தனித்தனியாகச் செயல்படுத்தப்பட்டன. கணினியியல் மொழியியல் சங்கத்தின் முதல் தலைவரான விக்டர் யங்வே என்பவரால் "ஆராய்ச்சிக் கோடு உருவானது, அவர் 1960 ஆம் ஆண்டில் ஒரு சொற்றொடர் கட்டமைப்பு இலக்கணத்தைப் பயன்படுத்துவதற்கான

வழிமுறைகளின் விளக்கங்களை செயற்கையாக நன்கு உருவாக்கிய முட்டாள்தனமான வாக்கியங்களை உருவாக்கினார். நான் 1962-1964 பற்றி நுட்பத்தால் ஈர்க்கப்பட்டேன், மேலும் சொற்களின் சொற்பொருள் சார்புகளை உரையில் நிகழ்ந்ததைப் பொறுத்து அவை உருவாக்கப்பட்டவற்றின் உணர்வைக் கட்டுப்படுத்துவதற்கான ஒரு முறைக்கு பொதுமைப்படுத்தின. " (Simmons 1982) மற்ற ஆராய்ச்சியாளர்கள், குறிப்பாக எம். ரோஸ் குயிலியன் (Quillian, 1963) மற்றும் சிஸ்டம் டெவலப்மென்ட் கார்ப்பரேஷனில் உள்ள மற்றவர்கள் 1960களின் முற்பகுதியில் SYNTHEX திட்டத்தின் ஒரு பகுதியாக தங்கள் பணிகளுக்கு பங்களிக்க உதவினார்கள். எஸ்.டி.சி.யில் இந்த வெளியீடுகளிலிருந்தே "சொற்பொருள் நெட்வொர்க்" என்ற வார்த்தையின் பெரும்பாலான நவீன வழித்தோன்றல்கள் அவற்றின் பின்னணியாகக் குறிப்பிடுகின்றன. பிற்காலத்தில் முக்கியமான படைப்புகள் ஆலன் எம். காலின்ஸ் மற்றும் குயிலியன் (எ.கா., காலின்ஸ் மற்றும் குயிலியன்; (Collins &. Quillian 1969, 1970) காலின்ஸ் மற்றும் லோஃப்டஸ் [Collins & Loftus 1975) குயிலியன் (Quillian 1966, 1967, 1968, 1969). 2006ஆம் ஆண்டில், ஹெர்மன் ஹெல்பிக் மல்டிநெட்டை முழுமையாக விவரித்தார் (Helbig, 2006).

1980களின் பிற்பகுதியில் க்ரோனிங்கன் மற்றும் ட்வென்டே (Groningen and Twente) இரண்டு நெதர்லாந்து பல்கலைக்கழகங்கள் கூட்டாக அறிவு வரைபடங்கள் (Knowledge Graphs) என்று அழைக்கப்படும் ஒரு திட்டத்தைத் தொடங்கினர்; ஆனால் அவை வரைபடத்தில் (Van de Riet. 1992) இயற்கணிதங்களை எளிதாக்குவதற்கு விளிம்புகள் ஒரு குறிப்பிட்ட சாத்தியமான உறவுகளின் குழுமத்திலிருந்து எடுக்கப்பட்டவை என்ற வரம்புக்குட்பட்டவை என்ற கூடுதல் தடைகளுடன் கூடிய பொருண்மைசார் நெட்வொர்க்குகள் ஆகும். அடுத்தடுத்த தசாப்தங்களில், பொருண்மைசார் நெட்வொர்க்குகள் மற்றும் அறிவு வரைபடங்களுக்கிடையிலான வேறுபாடு மங்கலாகிவிட்டது (Hulpus & Prangnawarat, 2015; McCuske & Chastain 2016). 2012 ஆம் ஆண்டில் கூகிள் அதன் அறிவு வரைபடத்திற்கு அறிவு வரைபடம் என்ற பெயரைக் கொடுத்தது.

பொருண்மைசார் இணைப்பு வலையமைப்பு (Semantic Link Network) ஒரு சமூக பொருண்மைசார் வலையமைப்பாக்க முறையாக ஆய்வு செய்யப்பட்டது. அதன் அடிப்படை மாதிரியானது பொருண்மைசார் கணுக்கள், கணுக்களுக்கிடையேயான பொருண்மைசார் இணைப்புகள் மற்றும் கணுக்கள் மற்றும் இணைப்புகளின் பொருண்மையிலை வரையறுக்கும் ஒரு பொருண்மைசார் இடத்தையும் பொருண்மைசார் இணைப்புகளில் பகுத்தறிவு விதிகளையும்

கொண்டுள்ளது. முறையான கோட்பாடு மற்றும் மாதிரி 2004 இல் வெளியிடப்பட்டது (Zhuge 2004). இந்த ஆராய்ச்சி திசையானது 1998இல் திறமையான மாதிரி மீட்டெடுப்பிற்கான பரம்பரை விதிகளின் வரையறையையும் (Zhuge 1998) மற்றும் செயலில் உள்ள ஆவண கட்டமைப்பின் ADF ஐயும் அறியலாம் (Zhuge, 2003). 2003 முதல், சமூக சொற்பொருள் வலையமைப்பை நோக்கி ஆராய்ச்சி வளர்ந்துள்ளது (Zhuge and Zheng 2003). பொருண்மைசார் வலையின் (நெட்வொர்க்) பயன்பாடு அல்லது எளிய நீட்டிப்பைக் காட்டிலும் உலகளாவிய வலை மற்றும் உலகளாவிய சமூக வலைப்பின்னல் காலத்தில் இந்தச் செயல்பாடு ஒரு திட்டமிட்ட கண்டுபிடிப்பு. அதன் நோக்கம் மற்றும் வாய்ப்பு பொருண்மைசார் வலையிலிருந்து (அல்லது நெட்வொர்க்) வேறுபட்டது (Zhuge 2012). பகுத்தறிவு மற்றும் பரிணாம வளர்ச்சிக்கான விதிகள் மற்றும் மறைமுக இணைப்புகளை தானியக்கமாகக் கண்டுபிடிப்பது ஆகியவை பொருண்மைசார் இணைப்பு வலையமைப்பில் முக்கிய பங்கு வகிக்கின்றன (Zhuge and Zheng 2004; Zhuge 2009). சமீபத்தில் இது சைபர்-இயற்பியல்-சமூக நுண்ணறிவை (Cyber-Physical-Social Intelligence) (Zhuge 2011) ஆதரிக்கும் வகையில் உருவாக்கப்பட்டது. இது ஒரு பொதுவான சுருக்கம் (Zhuge 2016) முறையை உருவாக்கப் பயன்படுத்தப்பட்டது. சுய-ஒழுங்கமைக்கப்பட்ட பொருண்மைசார் இணைப்பு நெட்வொர்க் பல்பரிமாண வகைப்பாட்டு இடைவெளியுடன் ஒருங்கிணைக்கப்பட்டு பல்பரிமாண சுருக்கங்கள் மற்றும் சுய-ஒழுங்கமைக்கப்பட்ட பொருண்மைசார் இணைப்புகள் கொண்ட மேம்பட்ட பயன்பாடுகளை ஆதரிக்க ஒரு பொருண்மைசார் இடத்தை உருவாக்குகிறது (Zhuge, 2008). (Zhuge and.Xing, 2012) சொற்பொருள் இணைப்பு நெட்வொர்க் என்று சரிபார்க்கப்பட்டது உரை சுருக்கம் பயன்பாடுகள் மூலம் புரிந்துகொள்ளுதல் மற்றும் பிரதிநிதித்துவப்படுத்துவதில் முக்கிய பங்கு வகிக்கிறது. சிறப்பு சமூக சொற்பொருள்களை விசாரிக்க, போட்டி உறவு மற்றும் கூட்டுவாழ்வு உறவு மற்றும் சமூகத்தை வளர்ப்பதில் அவற்றின் பங்கு ஆகியவை வளர்ந்து வரும் தலைப்பில் ஆய்வு செய்யப்பட்டன: சைபர்-இயற்பியல்-சமூக நுண்ணறிவு (Zhuge 2020)

குறிப்பிட்ட பயன்பாட்டிற்காக பொருண்மைசார் வலையமைப்புகளின் சிறப்பு வடிவங்கள் உருவாக்கப்பட்டுள்ளன. எடுத்துக்காட்டாக, 2008ஆம் ஆண்டில், ஃபாவ்ஸி பெண்டெக்கின் பிஎச்.டி ஆய்வறிக்கை பொருண்மையியல்சார் ஒற்றுமை வலையமைப்பை (Semantic Similarity Network (SSN/எஸ்எஸ்என்)) முறைப்படுத்தியது, இது பொருண்மையியல்சார் ஒற்றுமை

பிரதிநிதித்துவம் (semantic similarity representation) மற்றும் கணக்கீடுகளை எளிதாக்குவதற்கு சிறப்பு உறவுகள் மற்றும் பரப்புதல் வழிமுறைகளைக் கொண்டுள்ளது. (Bendeck, 2008).

சொற்பொருண்மைசார் வலையமைப்புகளின் (நெட்வொர்க்குகளின்) அடிப்படைகள்

ஒருவருக்கொருவர் தொடர்புடைய கருத்துருக்களின் குழுமமாக/கணமாக சிறந்த முறையில் புரிந்துகொள்ளக்கூடிய அறிவு இருக்கும்போது பொருண்மைசார் வலையமைப்பு (நெட்வொர்க்) பயன்படுத்தப்படுகிறது. பெரும்பாலான பொருண்மைசார் வலையமைப்புகள் அறிவாற்றல் அடிப்படையிலானவை. அவை வளைவுகள் மற்றும் கணுக்களையும் கொண்டிருக்கின்றன, அவை ஒரு வகைபிரித்தல் வரிசைக்கு ஒழுங்கமைக்கப்படலாம். பொருண்மைசார் வலையமைப்புகள் (நெட்வொர்க்குகள்) செயலாக்கம், மரபுப்பேறு மற்றும் கணுக்களை தொடக்கநிலைப் பொருள்களாக (proto-objects) பரப்புவதற்கான யோசனைகளை வழங்கின.

எடுத்துக்காட்டுகள்

லிஸ்பில் பொருண்மைசார் வலை

இணைவு (association) பட்டியலைப் பயன்படுத்துதல்

```
(setq *database*
'((canary  (is-a bird)
           (color yellow)
           (size small))
  (penguin (is-a bird)
           (movement swim))
  (bird    (is-a vertebrate)
           (has-part wings)
           (reproduction egg-laying))))
```

"canary" வகையைப் பற்றிய அனைத்து தகவல்களையும் பிரித்தெடுக்க "canary"-இன் கீ/விசையுடன் " assoc" செயல்பாட்டைப் பயன்படுத்துவீர்கள்.

சொல்வலை

ஒரு பொருண்மைசார் வலையமைப்பின் எடுத்துக்காட்டு சொல்வலை ஆகும், ஆங்கிலத்தின் ஒரு சொற்பொருள் தரவுத்தளம். இது ஆங்கில சொற்களை ஒருபொருள்பன்மொழிகளின் குழுமங்கள் (sets of synonyms) எனப்படும்

ஒருபொருள்பன்மொழியக் குழுமங்கத் தொகுக்கிறது, குறுகிய, பொதுவான வரையறைகளை வழங்குகிறது; மேலும் இந்த ஒருபொருள்பன்மொழியக் குழுமங்களுக்கு இடையிலான பல்வேறு பொருண்மைசார் உறவுகளை பதிவு செய்கிறது. (பார்க்க: முன்னர் விளக்கப்பட்ட சொல்வலைத் தலைப்பு).

பிற எடுத்துக்காட்டுகள்

சார்லஸ் சாண்டர்ஸ் பியர்ஸின் (Charles Sanders Peirce) இருத்தலியல் வரைபடங்கள் (existential graphs) அல்லது தொடர்புடைய ஜான் எஃப். சோவாவின் (John F. Sowa) கருத்துருசார் வரைபடங்கள் (conceptual graphs) போன்ற பொருண்மைசார் வலையமைப்புகளைப் பயன்படுத்தி தர்க்கரீதியான விளக்கங்களை உருப்படுத்தம் செய்ய இயலும்.

பொருண்மைசார் வலையமைப்புகளின் பிற எடுத்துக்காட்டுகள் கெல்லிஷ் மாதிரிகள் ஆகும். கெல்லிஷ் ஆங்கிலம் அதன் கெல்லிஷ் ஆங்கில அகராதியுடன், இது ஒரு வடிவம்சார் மொழியாகும் (formal language), இது கருத்துருக்களுக்கும் கருத்துருக்களின் பெயர்களுக்கும் இடையிலான உறவுகளின் வலையமைப்பாக வரையறுக்கப்படுகிறது.

ஸ்கிகிரஞ் (SciCrunch) என்பது அறிவியல் வளங்களுக்கான ஒத்துழைப்புடன் தொகுத்தமைக்கப்பட்ட அறிவுத் தளமாகும். இது மென்பொருள், ஆய்வக கருவிகள் போன்றவற்றிற்கான தெளிவான அடையாளங்காட்டிகளை (ஆராய்ச்சி வள அடையாளங்காட்டிகள் அல்லது RRIDகள்) வழங்குகிறது, மேலும் இது RRID களுக்கும் சமூகங்களுக்கும் இடையிலான இணைப்புகளை உருவாக்குவதற்கான விருப்பங்களையும் வழங்குகிறது.

வகைப்பாடு கோட்பாட்டின் அடிப்படையில் பொருண்மைசார் நெட்வொர்க்குகளின் மற்றொரு எடுத்துக்காட்டு ஒலாக்ஸ் (ologs) ஆகும். இங்கே ஒவ்வொரு வகையும் பொருள்களின்/சாதனங்களின் குழுமத்தை (set of things) உருப்படுத்தம் செய்யும் ஒரு பொருளாகும் (object), மேலும் ஒவ்வொரு அம்பும் ஒரு செயல்பாட்டைக் குறிக்கும் ஒப்புமை அமைப்பு (morphism) ஆகும். பொருண்மையிலைக் கட்டுப்படுத்த கருத்துப்பரிமாற்ற வரைபடங்களும் (Commutative diagrams) பரிந்துரைக்கப்படுகின்றன.

சமூக அறிவியலில் மக்கள் சில சமயங்களில் பொருண்மைசார் வலையமைப்பு என்ற சொல்லை இணை நிகழ்வு வலையமைப்புகளைக் (co-occurrence networks) குறிக்க

பயன்படுத்துகின்றனர் (Atteveldt 2008). உரையின் ஒரு அலகுடன் இணைந்த சொற்கள், எ.கா. ஒரு வாக்கியம், சொற்பொருளோடு ஒன்றுடன் ஒன்று தொடர்புடையது. இணை நிகழ்வுகளை அடிப்படையாகக் கொண்ட உறவுகள் பின்னர் பொருண்மைசார் வலையமைப்பு உருவாக்கப் பயன்படும்.

6.4.16. சொல் வகைப்பாடு அடையாளப்படுத்தும் ஒழுங்குமுறை (Parts-of-Speech Tagging system)

சொல்வகைப்பாடு (Parts-of-Speech)

மரபு இலக்கணத்தில் ஒத்த இலக்கணப் பண்புகளைக் கொண்ட சொற்களின் வகைப்பாடு (அல்லது, பொதுவாக, சொல்சார் கூறுகள்) சொல்வகைப்பாடு எனப்படும். ஒரே சொல்வகைப்பாட்டிற்கு ஒதுக்கப்பட்ட சொற்கள் பொதுவாக ஒத்த வாக்கிய நடத்தைகளைக் காண்பிக்கின்றன; அவை வாக்கியங்களின் இலக்கண கட்டமைப்பிற்குள் ஒத்த பங்களிப்பை வகிக்கின்றன; சில சமயங்களில் ஒத்த உருபன்களுக்கான ஒத்த பண்புகளுக்கு அவை திர்புறுகின்றன. (எடுத்துக்காட்டாக, பெயர்ச்சொல் ஒருமை, பன்மை, வேற்றுமை இவற்றிற்காகத் திரிபுறும்: ஆனை, ஆனைகள், ஆனையை, ஆனையால், ஆனையோடு பெயர்சொல் காலத்திற்காகதிரிபுறும்: வந்தான், வருகிறான், வருவான்) .

ஆங்கிலத்தில் பொதுவாகப் பட்டியலிடப்படும் சொல்வகைப்பாடுகள் பின்வருவனாகும்: பெயர்ச்சொல், வினைச்சொல், பெயரடைகள், வினையடைகள், பிரதிபெயர்/மாற்றுப்பெயர், முன்னொட்டு, இணைப்புக்கிளவி, உணர்வு இடைச்சொல், எண், அடைகொளி அடை. பிற இந்தோ-ஐரோப்பிய மொழிகளிலும் இந்த சொல் வகுப்புகள் அனைத்தும் உள்ளன; இந்த பொதுமைப்படுத்துதலுக்கு ஒரு விதிவிலக்கு என்னவென்றால், பெரும்பாலான ஸ்லாவிக் மொழிகள் மற்றும் லத்தீன் மற்றும் சமஸ்கிருதங்களில் 'articles' இல்லை. இந்தோ-ஐரோப்பிய குடும்பத்திற்கு அப்பால், ஹங்கேரிய மற்றும் பின்னிஷ் போன்ற பிற ஐரோப்பிய மொழிகள் இரண்டும் யூராலிக் குடும்பத்தைச் சேர்ந்தவை; இவை முற்றிலும் முன்னொட்டுக்களைக் கொண்டிருக்கவில்லை அல்லது அவற்றில் மிகக் குறைவானவை மட்டுமே உள்ளன; மாறாக, அவைகளுக்கு பின்னொட்டுக்கள் உள்ளன. திராவிட மொழிகளும் முன்னொட்டுக்களைக் கொண்டிருக்கவில்லை; அவை மாறாக பின்னொட்டுக்களைக் கொண்டிருக்கின்றன.

சொல்வகைப்பாட்டைத் தவிர பிற சொற்கள், குறிப்பாக நவீன மொழியியல் வகைப்பாடுகளில் மரபுத் திட்டத்தை விட துல்லியமான வேறுபாடுகளைச் செய்கின்றன;

அவற்றில் சொல் வகுப்பு word class, சொல்சார் வகுப்பு lexical class மற்றும் சொல்சார் வகை (lexical category) ஆகியவை அடங்கும். சில ஆசிரியர்கள் ஒரு குறிப்பிட்ட வகை தொடரியல் வகையை மட்டுமே குறிக்க 'சொல்சார் வகை' என்ற வார்த்தையைக் கட்டுப்படுத்துகின்றனர்; அவர்களுக்கு இந்த மாற்றுப்பெயர் போன்ற செயல்பாடுசார்ந்தவை என்று கருதப்படும் சொல்வகைப்பாட்டை விலக்குகிறது. வடிவ வகுப்பு (form class) என்ற சொல்லும் பயன்படுத்தப்படுகிறது, இருப்பினும் இது பல்வேறு முரண்பாடான வரையறைகளைக் கொண்டுள்ளது. சொல் வகுப்புகள் திறந்த (open) அல்லது மூடியவை (closed) என வகைப்படுத்தப்படலாம்: திறந்த வகுப்புகள் (open classes) (பெயர்ச்சொற்கள், வினைச்சொற்கள் மற்றும் பெயரடைகள் போன்றவை) தொடர்ந்து புதிய உறுப்பினர்களைப் பெறுகின்றன; அதே நேரத்தில் மூடிய வகுப்புகள் (மாற்றுப்பெயர்கள் மற்றும் இணைப்புக்கிளவிகள் போன்றவை) புதிய உறுப்பினர்களை எப்போதாவது பெறுகின்றன.

ஏறக்குறைய எல்லா மொழிகளிலும் வகுப்புகள் பெயர்ச்சொல் மற்றும் வினைச்சொல் உள்ளன; ஆனால் இந்த இரண்டையும் தாண்டி வெவ்வேறு மொழிகளில் குறிப்பிடத்தக்க வேறுபாடுகள் உள்ளன. எடுத்துக்காட்டாக,

ஜப்பானிய மொழியில் மூன்று வகை பெயரடைகள் உள்ளன; ஆனால் ஆங்கிலத்தில் ஒன்று மட்டுமே உள்ளது (ஏழு வகையான ஆங்கில பெயரடைகளுடன் குழப்பமடையக்கூடாது; அல்லது ஆங்கில பெயரடைகள் பெயர்ச்சொற்களையும் மாற்றுப் பெயர்களையும் அடைசெய்யலாம்);

சீன, கொரிய, ஜப்பானிய மற்றும் வியட்நாமிய மொழிகளில் பெயர்சார் வகைப்படுத்திகள் (nominal classifiers) உள்ளன; மற்றும்

பல மொழிகள் பெயரடைகளுக்கும் வினையடைகளுக்கும் அல்லது பெயரடைகளுக்கும் வினைச்சொற்களுக்கும் இடையில் வேறுபாடு காட்டுவதில்லை (நிலையான வினைச்சொல்லைப் பார்க்கவும்).

வகைகளின் எண்ணிக்கையிலும் அவற்றின் அடையாளம் காணும் பண்புகளிலும் இத்தகைய மாறுபாடு இருப்பதால், ஒவ்வொரு தனி மொழிக்கும் சொல்வகைப்பாட்டு பகுப்பாய்வு செய்யப்பட வேண்டும். ஆயினும்கூட, ஒவ்வொரு வகைக்கும் புலாக்குறிப்புகள் உலகளாவிய அளவுகோல்களின் அடிப்படையில் ஒதுக்கப்படுகின்றன.

சொல்வகைப்பாடு அடையாளப்படுத்தல்

தரவுத்தொகுதி மொழியியலில், சொல்வகைப்பாடு அடையாளப்படுத்தல் (part-of-speech tagging or POS tagging or PoS tagging or POST) இலக்கண அடையாளப்படுத்தல் என்றும் அழைக்கப்படுகிறது, இது ஒரு உரையில் (தரவுத்தொகுதியில்) ஒரு சொல்லை அதன் வரையறை மற்றும் சூழல் இரண்டையும் அடிப்படையாகக் கொண்டு ஒரு குறிப்பிட்ட சொல்வகைப்பாட்டிற்கு ஒத்ததாகக் அடையாளப்படுத்தும் செயல்முறையாகும். இதன் எளிமையான வடிவம் பொதுவாக பள்ளி வயது குழந்தைகளுக்கு கற்பிக்கப்படும் சொற்களை பெயர்ச்சொற்கள், வினைச்சொற்கள், பெயரடைகள், வினையடைகள் எனஅடையாளம் காண்பதாகும்.

ஒரு காலத்தில் கையால் செய்யப்பட்ட சொல்வகைப்பாடு அடையாளப்படுத்தல் இப்போது தனித்துவமான சொற்களையும் மறைக்கப்பட்ட சொல்வகைப்பாட்டையும் இணைக்கும் வழிமுறைகளைப் பயன்படுத்தி விளக்க அடையாளங்களின் குழுமத்தால் கணினி மொழியியலின் சூழலில் செய்யப்படுகிறது. சொல்வகைப்பாடு அடையாளப்படுத்தும் வழிமுறைகள் இரண்டு தனித்துவமான குழுக்களாகின்றன: விதி அடிப்படை (rule-based) மற்றும் புள்ளியியல் அடிப்படை (stochastic). முதலாவதான மற்றும் மிகவும் பரவலாகப் பயன்படுத்தப்படும் ஆங்கில சொல்வகைப்பாட்டு அடையாளப்படுத்திகளில் ஒன்றான ஈ. பிரில்லின் டேக்கர் (E. Brill's tagger) விதி அடிப்படையிலான வழிமுறைகளைப் பயன்படுத்துகிறது.

கொள்கை

சொல்வகைப்பாடு அடையாளப்படுத்தல் என்பது சொற்களின் பட்டியலையும் அவற்றின் சொல்வகைப்பாடுகளையும் வைத்திருப்பதை விட கடினமானது; ஏனென்றால் சில சொற்கள் வெவ்வேறு நேரங்களில் ஒன்றுக்கு மேற்பட்ட சொல்வகைப்பாடுகளைக் குறிக்கக்கூடும்; மேலும் சில சொல்வகைப்பாடுகள் சிக்கலானவை அல்லது தெளிவற்றவை. இது அரிதானது அல்ல; இயற்கை மொழிகளில் (பல செயற்கை மொழிகளுக்கு மாறாக), சொல் வடிவங்களில் பெரும் சதவீதம் பொருண்மை மயக்கம் உள்ளவை. எடுத்துக்காட்டாக, பொதுவாக ஒரு பன்மை பெயர்ச்சொல் என்று கருதப்படும் " dogs" கூட ஒரு வினைச்சொல்லாக இருக்கலாம்:

The sailor dogs the hatch.

சரியான இலக்கண குறிச்சொல் "dogs" இங்கே ஒரு வினைச்சொல்லாக பயன்படுத்தப்படுகிறது என்பதை பிரதிபலிக்கும், இது மிகவும் பொதுவான பன்மை பெயர்ச்சொல்லாக அல்ல. இதை தீர்மானிக்க இலக்கண சூழல் ஒரு வழி; " sailor" மற்றும் " hatch" ஆகியவை "dogs" 1) என கடல் சூழலில் குறிக்கின்றன மற்றும் 2) " hatch" என்ற செயப்படுபொருளுக்கு பயன்படுத்தப்படும் ஒரு

செயல் (இந்த சூழலில், " dogs" ஒரு கடல்துரைசார்ந்த அர்த்தம் (nautical term meaning) " fastens (a watertight door) securely"; "தமிழில் நீர்புக இயலாத கதவு பாதுகாப்பாக மூடப்பட்டது" எனப் பொருள்படும்).

அடையாளக் குழுமங்கள் Tag sets

ஆங்கிலத்தில் 9 சொல்வகைப்பாடுகள் (parts of speech) இருப்பதாக பள்ளிகள் பொதுவாகக் கற்பிக்கின்றன: பெயர்ச்சொல், வினைச்சொல், ஆர்டிகிள், பெயரடை, முன்னடை, மாற்றுப்பெயர், பெயரடை, இணைப்புக்கிளவி மற்றும் உணர்ச்சி இடைச்சொல். இருப்பினும், இன்னும் பல பிரிவுகள் மற்றும் துணை பிரிவுகள் தெளிவாக உள்ளன. பெயர்ச்சொற்களைப் பொறுத்தவரை, பன்மை, உடைமை மற்றும் ஒருமை வடிவங்களை வேறுபடுத்தி அறியலாம். பல மொழிகளில் சொற்கள் அவற்றின் "வேற்றுமை" (எழுவாய், செயப்படுபொருள் போன்றவற்றின் பங்கு), இலக்கண பாலினம் மற்றும் பலவற்றிற்கும் குறிக்கப்பட்டுள்ளன; வினைச்சொற்கள் காலம் tense), வினையாற்றுவகை (aspect) மற்றும் பிற விஷயங்களுக்கு குறிக்கப்பட்டுள்ளன. சில குறியீட்டு/அடையாளப்படுத்தல் ஒழுங்குமுறைகளில், ஒரே வேர் சொல்லின் வெவ்வேறு திரிபுகள் வெவ்வேறு சொல்வகைப்பாடுகளைப் பெறும்; இதன் விளைவாக அதிக எண்ணிக்கையிலான குறிச்சொற்கள்/அடையாளங்கள் கிடைக்கும். எடுத்துக்காட்டாக, ஒருமைப் பொதுவான பெயர்ச்சொற்களுக்கு NN, பன்மை பொதுவான பெயர்ச்சொற்களுக்கு NNS, ஒருமை இயற்பெயர்களுக்கு NP (பிரவுன் தரவுத்தொகுதில் பயன்படுத்தப்படும் சொல்வகைப்பாட்டு அடையாளங்களை/குறிச்சொற்களைப் பார்க்கவும்). பிற அடையாளப்படுத்தல் ஒழுங்குமுறைகள் குறைந்த எண்ணிக்கையிலான அடையாளங்களைப்/குறிச்சொற்களைப் பயன்படுத்துகின்றன மற்றும் சிறந்த வேறுபாடுகளைப் புறக்கணிக்கின்றன அல்லது சொல்வகைப்பாட்டிலிருந்து ஓரளவு சுதந்திரமான பண்புக்கூறுகளாக அவற்றை வடிவமைக்கின்றன.

கணினி மூலம் சொல்வகைப்பாட்டு அடையாளப்படுத்தலில் ஆங்கிலத்திற்காக 50 முதல் 150 தனிச் சொல்வகைப்பாடுகளை வேறுபடுத்துவது பொதுவானது. கொய்ன் கிரேக்கத்தை (DeRose 1990) அடையாளப்படுத்துவதற்கான/குறியிடுவதற்கான புள்ளியியல்சார் முறைகள் 1,000 சொல்வகைப்பாடுகளுக்கு மேல் பயன்படுத்தியுள்ளன; மேலும் ஆங்கிலத்தில் உள்ளதைப் போல அந்த மொழியில் பல சொற்கள் பொருண்மைமயக்கம் உள்ளதாக இருப்பதைக் கண்டறிந்துள்ளது. உருபனியல் அடிப்படையில் வளமான மொழிகளின் விஷயத்தில் ஒரு உருபனியல்-தொடரியல்சார் விளக்கி (morphosyntactic descriptor) வகைப்பாட்டுக்கு Ncmsan =

பெயர்ச்சொல், வகை = பொதுவானது, பாலினம் = ஆண்பால், எண் = ஒருமை, வேற்றுமை = செயப்பொடுபொருள், விலங்கினம் = இல்லை. போன்ற மிகக் குறுகிய நினைவூட்டல்களைப் பயன்படுத்தி வெளிப்படுத்தப்படுகிறது.

அமெரிக்க ஆங்கிலத்திற்கான சொல்வகைப்பாட்டு அடையாளப்படுத்தகான மிகவும் பிரபலமான "அடையாளக் குழுமம் " என்பது பென் கிளைவங்கித் திட்டத்தில் (Penn Treebank project) உருவாக்கப்பட்ட பென் அடையாளக் குழுமம் (Penn tag set) ஆகும். இது பெரும்பாலும் முந்தைய பிரவுன் தரவுத்தொகுதி (Brown Corpus) மற்றும் லாப் தரவுத்தொகுதி (LOB Corpus) அடையாளக் குழுமங்களுடன் ஒத்திருக்கிறது, இது மிகவும் சிறியதாக இருந்தாலும். ஐரோப்பாவில், ஈகிள்ஸ் வழிகாட்டுதல்களிலிருந்து (Eagles Guidelines) அடையாளக் குழுமங்கள் பரந்த பயன்பாட்டைக் காண்கின்றன மற்றும் பன் மொழிகளுக்கான பதிப்புகளை உள்ளடக்குகின்றன.

சொல்வகைப்பாடு அடையாளப்படுத்தல் செயல்பாடு பல்வேறு மொழிகளில் செய்யப்பட்டுள்ளது, மேலும் பயன்படுத்தப்படும் சொல்வகைப்பாடு அடையாளங்களின் குழுமங்கள் மொழிக்குமொழி பெரிதும் மாறுபடும். அடையாளங்கள்/குறிச்சொற்கள் பொதுவாக வெளிப்படையான உருவ வேறுபாடுகளை உள்ளடக்கும் வகையில் வடிவமைக்கப்பட்டுள்ளன, இருப்பினும் இது ஆங்கிலத்தில் பிரதிபெயர்களுக்கான (ஆனால் பெயர்ச்சொற்களுக்கு அல்ல) வேற்றுமையைக் குறிப்பிடுதல் போன்ற முரண்பாடுகளுக்கும் மிகப் பெரிய மொழி கடந்த வேறுபாடுகளுக்கும் (cross-language differences) வழிவகுக்கிறது. கிரேக்க மற்றும் லத்தீன் போன்ற பெரிதும் திரிபுறும் மொழிகளுக்கான அடையாளக் குழுமங்கள் மிகப் பெரியதாக இருக்கும்; இன்யூட் மொழிகள் (Inuit languages) போன்ற ஒட்டுநிலை மொழிகளில் சொற்களைக் குறிப்பது/அடையாளப்படுத்துவது கிட்டத்தட்ட சாத்தியமற்றது. மற்றொரு தீவிரத்தில், பெட்ரோவ் மற்றும் பலர் (Petrov et al 2011). 12 வகைகளைக் கொண்ட "உலகளாவிய" அடையாளக் குழுமத்தை ("universal" tag set) முன்மொழிந்துள்ளது (எடுத்துக்காட்டாக, பெயர்ச்சொற்கள், வினைச்சொற்கள், நிறுத்தற்குறி போன்றவை; ஆங்கிலத்தில் "to" வினை எச்சக் குறியீடாக அல்லது முன்னொட்டாக எந்த வேறுபாடும் இல்லை போன்றவை). மிகச் சிறிய குழுமம் விரும்பத்தக்கதா மிகப் பெரிய துல்லியமான குழுமம் விரும்பத்தக்கதா என்பது கையில் இருக்கும் நோக்கத்தைப் பொறுத்தது. சிறிய அடையாளக் குழுமங்களில் தானியங்கி அடையாளப்படுத்தல் எளிதானது.

வரலாறு

பிரவுன் கார்பஸ் (The Brown Corpus)

சொல்வகைப்பாடு அடையாளப்படுத்தல் குறித்த ஆராய்ச்சி தரவுத்தொகுதி மொழியியலுடன் நெருக்கமாகப் பிணைக்கப்பட்டுள்ளது. கணினிப் பகுப்பாய்விற்கான ஆங்கிலத்தின் முதல் பெரிய தரவுத்தொகுதி 1960 களின் நடுப்பகுதியில் ஹென்றி குசெரா மற்றும் டபிள்யூ. நெல்சன் பிரான்சிஸ் (Henry Kučera and W. Nelson Francis) ஆகியோரால் பிரவுன் பல்கலைக்கழகத்தில் உருவாக்கப்பட்ட பிரவுன் தரவுத்தொகுதி ஆகும். இது தோராயமாக தேர்ந்தெடுக்கப்பட்ட வெளியீடுகளிலிருந்து 500 மாதிரிகளால் ஆன தொடர்ச்சியான ஆங்கில உரைநடை உரையின் சுமார் 1,000,000 (ஒரு மில்லியன்) சொற்களைக் கொண்டுள்ளது. ஒவ்வொரு மாதிரியும் 2,000 அல்லது அதற்கு மேற்பட்ட சொற்கள் (2,000 சொற்களுக்குப் பிறகு முதல் வாக்கியத்தின் முடிவில் முடிவடைகிறது, இதனால் தரவுத்தொகுதியில் முழுமையான வாக்கியங்கள் மட்டுமே உள்ளன).

பிரவுன் தரவுத்தொகுதி பல ஆண்டுகளாக சொல்வகைப்பாட்டு அடையாளங்களால் மிகக் கடினமான முயற்சியோடு அடையாளப்படுத்தப்பட்டது. கிரீன் மற்றும் ரூபின் ஆகியோரால் ஒரு திட்டத்துடன் முதல் தோராயமாக்கல் செய்யப்பட்டது, இதில் என்னென்ன வகைகள் ஒன்றிணைந்து நிகழக்கூடும் என்பதற்கான ஒரு பெரிய கையால் செய்யப்பட்ட பட்டியலைக் கொண்டிருந்தது. எடுத்துக்காட்டாக, ஆர்டிக்கிளின் பின்னர் பெயர்ச்சொற்கள் ஏற்படலாம், ஆனால் ஆர்டிகிள் வினைச்சொல் (விவாதிக்கக்கூடியது) ஏற்படாது. நிரல் சுமார் 70% சரியானது. அதன் முடிவுகள் மீண்டும் மீண்டும் மதிப்பாய்வு செய்யப்பட்டு கையால் திருத்தப்பட்டன; பின்னர் பயனர்கள் பிழைகள் அனுப்பினர்; இதனால் 70களின் பிற்பகுதியில் அடையாளப்படுத்தல் கிட்டத்தட்ட சரியாக இருந்தது (மனித பேச்சாளர்கள் கூட ஒப்புக் கொள்ளாத சில சந்தர்ப்பங்களை அனுமதிக்கிறது).

மறைக்கப்பட்ட மார்கோவ் மாதிரிகளின் பயன்பாடு (Use of hidden Markov models)

1980 களின் நடுப்பகுதியில், ஐரோப்பாவில் ஆராய்ச்சியாளர்கள் பிரிட்டிஷ் ஆங்கிலத்தின் லான்காஸ்டர்-ஒஸ்லோ-பெர்கன் தரவுத்தொகுதியை அடையாளப்படுத்தச் செயல்படும்போது போது, சொல்வகைப்பாட்டின் மயக்கத்தை நீக்க மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகளைப் (hidden Markov models (HMMs/எச்.எம்.எம்) பயன்படுத்தத் தொடங்கினர். எச்.எம்.எம்.கள் வேற்றுமைகளை எண்ணுவது (பிரவுன் கார்பஸிலிருந்து போன்றவை) மற்றும் சில

தெடர்வரிசைகளில் நிகழ்தகவுகளின் அட்டவணையை உருவாக்குவது ஆகியவை அடங்கும். எடுத்துக்காட்டாக, ' the' போன்ற ஒரு ஆர்டிகிளை நீங்கள் பார்த்தவுடன் அடுத்த சொல் 40% தடவை பெயர், 40% பெயரடை, மற்றும் 20% ஒரு எண் எனக் கணிக்க இயலும். இதை அறிந்தால், "the can" என்பதில் வரும் "can" என்பது ஒரு வினைச்சொல் அல்லது ஒரு வினைநோக்கைக் காட்டிலும் ஒரு பெயர்ச்சொல்லாக இருக்க வாய்ப்புள்ளது என்று ஒரு நிரல் தீர்மானிக்க முடியும். பின்வரும் முறையைப் பற்றிய அறிவிலிருந்து பயனடைய அதே முறையைப் பயன்படுத்தலாம்.

எச்.எம்.எம்.கள் புள்ளியியல்சார் அடையாளப்படுத்திகளின் (stochastic taggers) செயல்பாட்டை உள்ளுறை செய்கின்றன; மேலும் அவை பல்வேறு வழிமுறைகளில் பயன்படுத்தப்படுகின்றன; மிக அதிக அளவில் பயன்படுத்தப்படுவது இரு-திசை அனுமான வழிமுறை (bi-directional inference algorithm) ஆகும்.

டைனமிக் நிரலாக்க முறைகள் (Dynamic programming methods)

1987 ஆம் ஆண்டில், ஸ்டீவன் டிரோஸ் (Steven DeRose) மற்றும் கென் சர்ச் (Ken Church) ஆகியோர் ஒரே சிக்கலை மிகக் குறைந்த நேரத்தில் தீர்க்க தனித்தனியாக டைனமிக் நிரலாக்க வழிமுறைகளை உருவாக்கினர். அவற்றின் முறைகள் மற்ற துறைகளில் சில காலம் அறியப்பட்ட விட்டர்பி வழிமுறையைப் (Viterbi algorithm) போலவே இருந்தன. டிரோஸ் இணைகளின் அட்டவணையைப் பயன்படுத்தினார், அதே நேரத்தில் சர்ச் மும்மடங்குகளின் அட்டவணையையும் பிரவுன் கார்பஸில் அரிதான அல்லது இல்லாத மும்மடங்களுக்கான மதிப்புகளை மதிப்பிடும் முறையையும் பயன்படுத்தினார் (மூன்று நிகழ்தகவுகளின் உண்மையான அளவீட்டுக்கு மிகப் பெரிய கார்பஸ் தேவைப்படும்). இரண்டு முறைகளும் 95%க்கும் அதிகமான துல்லியத்தை அடைந்தன. பிரவுன் பல்கலைக்கழகத்தில் டெரோஸின் 1990ஆம் ஆண்டு ஆய்வு ஏட்டில் குறிப்பிட்ட பிழை வகைகள், நிகழ்தகவுகள் மற்றும் பிற தொடர்புடைய தரவுகளின் பகுப்பாய்வுகளும் அடங்கியிருந்தன; மேலும் கிரேக்க மொழியில் அவரது செயல்பாட்டை திரும்பச்செய்தார்; அது இதேபோல் பயனுள்ளதாக இருந்தது.

மேற்பார்வை செய்யப்படாத அடையாளங்கள் (Unsupervised taggers)

ஏற்கனவே விவாதிக்கப்பட்ட முறைகள் அடையாளம் நிகழ்தகவுகளைக் கற்றுக்கொள்வதற்கு முன்பே இருக்கும் தரவுத்தொகுதியிலிருந்து (pre-existing corpus) செயல்படுவதை உள்ளடக்கியது. இருப்பினும், "மேற்பார்வை செய்யப்படாத" அடையாளப்படுத்தலைப் பயன்படுத்தி பூட்ஸ்ட்ராப் செய்ய முடியும். மேற்பார்வை செய்யப்படாத

அடையாளப்படுத்தல் நுட்பங்கள் அவற்றின் பயிற்சித் தரவுகளுக்கு ஒரு அடையாளப்படுத்தப்படாத தரவுத்தொகுதியைப் பயன்படுத்துகின்றன மற்றும் தூண்டல் மூலம் அடையாளக் குழுமங்களை உருவாக்குகின்றன. அதாவது, அவர்கள் சொல் பயன்பாட்டில் வடிவங்களைக் கவனித்து, சொல்வகைப்பாட்டை ஆக்குகின்றார்கள். எடுத்துக்காட்டாக, புள்ளிவிவரங்கள் "the", "a" மற்றும் "an" ஆகியவை ஒத்த சூழல்களில் நிகழ்கின்றன; அதே நேரத்தில் " eat" என்பது மிகவும் வித்தியாசமானவற்றில் நிகழ்கிறது. போதுமான மறுசெய்கையுடன், மனித மொழியியலாளர்கள் எதிர்பார்ப்பதைப் போலவே சொற்களின் ஒற்றுமை வகுப்புகள் (similarity classes of words) வெளிப்படுகின்றன; மற்றும் வேறுபாடுகள் சில நேரங்களில் மதிப்புமிக்க புதிய நுண்ணறிவுகளை பரிந்துரைக்கின்றன.

இந்த இரண்டு வகைகளையும் விதி அடிப்படையிலான, புள்ளியியல் அடிபடையிலான மற்றும் நரம்பியல் அடிப்படையிலான அணுகுமுறைகளாக மேலும் பிரிக்கலாம்.

பிற அடையாளப்படுத்திகள் மற்றும் முறைகள்

சொல்வகைப்பாடு அடையாளப்படுத்தலின் சில தற்போதைய முக்கிய வழிமுறைகளில் விட்டர்பி வழிமுறை (Viterbi algorithm), பிரில் டேக்கர் (Brill tagger), கட்டுப்பாட்டு இலக்கணம் (Constraint Grammar) மற்றும் பாம்-வெல்ச் வழிமுறை (Baum-Welch algorithm) [முன்னோக்கி-பின்நோக்கி வழிமுறை (forward-backward algorithm) என்றும் அழைக்கப்படுகிறது] ஆகியவை அடங்கும். மறைக்கப்பட்ட மார்க்கோவ் மாதிரி (Hidden Markov model) மற்றும் புலப்படும் மார்கோவ் மாதிரி (visible Markov model) அடையாளப்படுத்திகள் இரண்டையும் விட்டர்பி வழிமுறையைப் பயன்படுத்தி செயல்படுத்தலாம். விதி அடிப்படையிலான பிரில் டேக்கர் அசாதாரணமானது, இது ஒரு விதிமுறை அமைப்பொழுங்குகளைக் கற்றுக்கொள்கிறது, பின்னர் புள்ளிவிவர அளவை மேம்படுத்துவதை விட அந்த அமைப்பொழுங்குகளைப் பயன்படுத்துகிறது. விதிகள் தொடர்ச்சியாக வரிசைப்படுத்தப் செய்யப்படும் பிரில் டேக்கரைப் (RDRPOSTagger) போலன்றி, சொல்வகைப்பாடு மற்றும் உருபனியல் அடையாளப்படுத்தல் கருவித்தொகுதி ஆர்.டி.ஆர்.பொஸ்டாகர் (RDRPOSTagger) ரிபிள்-டவுண் ரூல்ஸ் டிரீ (ripple-down rules tree) வடிவத்தில் விதிகளை சேகரிக்கின்றன.

சொல்வகைப்பாட்டு அடையாளப்படுத்தலின் சிக்கலுக்கு பல இயந்திர கற்றல் முறைகளும் பயன்படுத்தப்பட்டுள்ளன. எஸ்.வி.எம். (Support Vector Mahcine (SVM)), அதிகபட்ச என்ட்ரோபி வகைப்படுத்தி (maximum entropy classifier), பெர்செப்டிரான் (perceptron) மற்றும் அருகிலுள்ள

அண்டை (nearest-neighbor) போன்ற முறைகள் அனைத்தும் முயற்சிக்கப்பட்டுள்ளன; மேலும் பெரும்பாலானவை 95% க்கு மேல் துல்லியத்தை அடைய இயலும்.

பல முறைகளின் நேரடி ஒப்பீடு ACL விக்கியில் (குறிப்புகளுடன்) தெரிவிக்கப்படுகிறது. இந்த ஒப்பீடு சில பென் ட்ரீபேங்க் தரவுகளில் அமைக்கப்பட்ட பென் அடையாளக் குழுமங்களைப் பயன்படுத்துகிறது, எனவே முடிவுகள் நேரடியாக ஒப்பிடத்தக்கவை. இருப்பினும், பல குறிப்பிடத்தக்க அடையாளப்படுத்திகள் சேர்க்கப்படவில்லை (ஒருவேளை இந்த குறிப்பிட்ட தரவுகுழுமத்திற்காக அவற்றை மறுசீரமைப்பதில் ஈடுபட்டுள்ள உழைப்பு காரணமாக இருக்கலாம்). எனவே, இங்கே தெரிவிக்கப்பட்ட முடிவுகள் கொடுக்கப்பட்ட அணுகுமுறையால் அடையக்கூடிய சிறந்தவை என்று கருதக்கூடாது; கொடுக்கப்பட்ட அணுகுமுறையால் அடையப்பட்ட சிறந்தவை கூட இல்லை.

2014 ஆம் ஆண்டில், சொல்வகைப்பாடு அடையாளப்படுத்தல் கட்டமைப்பு ஒழுங்குமுறை முறையைப் (structure regularization method) பயன்படுத்தி ஒரு காகித அறிக்கை, நிலையான பெஞ்ச்மார்க் தரவுக்குழுமத்தில் 97.36%ஐ அடைகிறது (Xu Sun (2014).

6.3.17. ஓரிடம்சார் சொற்களை குழுமும் ஒழுங்குமுறை (Local Word Grouping system).

இந்திய அரசின் நிதிநல்கையின் கீழ் நடைபெற்ற 'இந்திய மொழிகளிலிருந்து இந்திய மொழிகளின் இயந்திரமொழிபெயர்ப்பு' என்ற ஆய்வுத்திட்டத்தின் இலக்கண வடிவவாதமாக பாணினியின் இலக்கணம் பயன்படுத்தப்பட்டது. இது ஆங்கிலம் போன்ற மொழிகளுக்காக உருவாக்கப்பட்ட தொடரமைப்பு அல்லது தொடரியல் கட்டமைப்பு அடிப்படையில் அமையாமல் வேற்றுமை இலக்கணம் அடிப்படையில் அமைந்தது. இது வாக்கியங்களைத் தொடர்களாகப் பிரிக்காமல் சில சொல் குழுமங்களாகப் (ஓரிடம்சார் சொற் குழுமுமங்கள் (Local Word Groups (LWG)) பகுத்து ஆய்ந்தன. இந்த இலக்கண வடிவவாதம் வாகியங்களை பெயர்த்தொடர், வினைத்தொடர் எனப் பிரிக்காமல் பெயர்த்தொடர் குழுமம், வினைதொடர் குழுமம் என ஓரடம்சார் சொற்களைக் குழுமும் ஒழுங்குமுறைக்கு முக்கியத்துவம் அளித்தது. ஓரிடம்சார் சொற்களைக் குழுமும் ஒழுங்குமுறை பற்றி பிரதிப்தா ரஞ்சன் ரே மற்றும் பிறர் தங்கள் கட்டுரையில் விரிவாகக் கூறுகின்றனர். இப்பகுதியில் வரும் செய்திகள் அதன் அடிப்படையில் அமையும்.

கணினிகள் இயற்கை மொழியைப் புரிந்துகொள்ள வேண்டிய அவசியம், குறிப்பாக வட்டாரப் பேச்சு மொழிகள், மனித-கணினி இடைமுகங்கள் மிகவும் உள்ளுணர்வாக மாறுவதால்,

வலையில் பன்மொழி உள்ளடக்கம் அதிகரிப்பதால் இயந்திர மொழிபெயர்ப்பு முக்கியத்துவம் பெறுகிறது. இயற்கை மொழிகள் நிலையான சொல் வரிசை மொழிகளாக (எ.கா. ஆங்கிலம்) மற்றும் சுதந்திர சொல் வரிசை மொழிகளாக (எ.கா. சமஸ்கிருதம்) பரவலாக வகைப்படுத்தப்படலாம். நிலையான சொல் வரிசை மொழிகளில், ஒரு வாக்கியத்தை உருவாக்கும் சொற்களை சில நிலையான வழிகளில் இலக்கண விதிகளின்படி ஒரு வாக்கியத்தில் நிலைநிறுத்தலாம் (எ.கா. ஆங்கிலத்தில் எழுவார்-வினை-செயப்படுபொருள் அமைப்பு). சுதந்திரச்சொல் வரிசை மொழிகளில், சொற்களின் வரிசையில் நிலையான வரிசைப்படுத்தல் எதுவும் விதிக்கப்படவில்லை.

ஒரு இயல்பான மொழி வாக்கியத்தை பாகுபடுத்துவது ஒரு வாக்கியத்தின் பொருளை அடையாளம் காண அனுமதிக்கும் கட்டமைப்பு அலகுகளை துண்டிக்க அல்லது அங்கீகரிப்பதாக கருதப்படலாம் (Manning and Schutze, 1999). இத்தகைய துண்டிக்கப்படுதல் வெவ்வேறு முறைகளைப் பயன்படுத்தி அடையப்படலாம். நிலையான சொல் வரிசை மொழிகளுக்கு, சூழல் சுதந்திர இலக்கணங்கள் (context free grammars) மற்றும் கிளை சேர்க்கும் இலக்கணங்கள் (tree adjoining grammars) போன்ற ஆக்கமுறை இலக்கணங்கள் (generative grammars) கட்டமைப்புகளை மாடலிங் செய்ய பயன்படுத்தப்படுகின்றன (Charniak, 1997; Joshi and Schabes, 1997). சுதந்திர சொல் ஒழுங்கு நிகழ்வுகளைப் பிடிக்கத் தேவையான விரித்தெழுதும் விதிகளின் எண்ணிக்கை மிகப் பெரியதாக இருப்பதால், இத்தகைய நுட்பங்கள் சுதந்திர சொல் வரிசை மொழிகளுடன் (free word order languages) சிறப்பாக செயல்படாது. அத்தகைய மொழிகளுக்கான பொதுவான செயலாக்க நுட்பங்கள் கட்டுப்பாடு அடிப்படையிலான பாகுபடுத்தல் (constraint based parsing) ஆகும்; இது ஒரு வாக்கியத்தின் கூறுகளுக்கு இடையிலான உறவுகளைக் கட்டுப்பாடுகள் முன்னிலையில் அடையாளம் காண்பதை உள்ளடக்குகிறது. வேற்றுமை இலக்கணம் (Case grammar) மற்றும் கணினிசார் பாணினியன் மாதிரி (computational Paninian model) இத்தகைய கட்டுப்பாட்டு அடிப்படையிலான பாகுபடுத்தலுக்கான எடுத்துக்காட்டுகள் ஆகும் (Bharati et al, 1995; Bharati and Sangal, 1993; Tapanainen and Jarvinen, 1997). நிலையான ஒழுங்கு மொழிகளை (fixed order languages) பகுப்பாய்வுசெய்வதற்கும் இத்தகைய வடிவவாதங்கள் பயன்படுத்தப்பட்டுள்ளன (Bharati et al, 1997).

இந்தோ-ஆரிய மொழிகளில், சமஸ்கிருதம் முற்றிலும் சுதந்திர ஒழுங்கு மொழி, ஆனால் இந்தி மற்றும் பெங்காலி போன்ற சில இந்தோ-ஆரிய மொழிகள் பரிணாம வளர்ச்சியின் போது சுதந்திர சொல் வரிசையை ஓரளவு இழந்துவிட்டன. இந்தி மற்றும் பெங்காலி மொழிகளில், "சொல் குழுக்கள்" (word groups) கட்டற்ற/சுதந்திர ஒழுங்கு உள்ளவை; ஆனால் சொல் குழுக்களின் உள் அமைப்பு நிலையான-வரிசை உள்ளவை; அதாவது எந்த வாக்கியத்தையும் சொல் குழுக்களின் "பை" என்று பார்க்கலாம். இத்தகைய மொழிகளில் இந்த சொல் குழுக்களை சரியாகவும் அதிகபட்சமாகவும் அடையாளம் காண்பது கணிசார் பாணினியன் மாதிரியின் முக்கிய பாகுபடுத்தியின் (core parser) அதிகப்படியான வேலையைக் குறைக்கிறது. பாகுபடுத்தி பின்னர் வினை குழு மற்றும் பெயர்ச்சொல் குழுக்களுக்கு இடையிலான உறவை சிக்கலை ஒரு முழு நிரலாக்க அல்லது இருதரப்பு வரைபட பொருத்தச் சிக்கலாக மாதிரியாகக் கொண்டு அடையாளம் காட்டுகிறது.

சொற்களின் வரிசைகளை ஒன்றிணைத்து ஒரு சொல் குழுவை உருவாக்கலாம். உருபொலியனியல் ஆய்வியைப் பயன்படுத்தி ஒவ்வொரு தனிப்பட்ட சொற்களிலிருந்தும் பெறப்பட்ட தகவல்களைப் பயன்படுத்தி ஒரு சொல் குழுவில் உள்ள தகவல்களை உருவாக்க முடியும். சொல் குழுவில் உள்ள தகவல்கள் தனிப்பட்ட சொற்களிலிருந்து பெறப்பட்ட தகவல்களையும், சொற்களுக்கு இடையிலான உறவின் தன்மையையும் பொறுத்தது. சொல் குழுக்கள் ஒன்றுக்கொன்று ஒன்றிணைந்து பெரிய சொல் குழுக்களை உருவாக்கலாம். ஒரு சுருக்க வடிவத்தில், இந்தக் குழுக்களின் கட்டமைப்பை ஒரு அடை- அடைசெய்யப்பட்ட கட்டமைப்பாக (modifier-modified structure) நாம் பார்க்கலாம். இயற்றப்பட்ட கட்டமைப்பின் பொருள் அடைசெய்யப்பட்ட அல்லது தலையின் பொருளைப் போன்றது, ஆனால் வெவ்வேறு அடைகளால் வெவ்வேறு வழிகளில் அடைசெயப்படுகிறது. இந்தியில் 10 வகையான மாற்றியமைப்பாளர் - மாற்றியமைக்கப்பட்ட கட்டமைப்புகள் இருக்கலாம். (Bharati et al, 1995)

இந்தியில் உள்ள உள்ளூர் சொல் குழுக்கள் வாக்கியத்தின் நிலையான ஒழுங்கு கூறுகள் ஆகும்; அவை சொற்களின் குழுக்கள் ஆகும். அவை உறுப்புச் சொற்கள் தொடர்ச்சியாக வைக்கப்படுகின்றன, மற்றும் சொற்களின் குழு நிலையான சொல் வரிசைக் கொண்டிருக்கின்றது என்பதைக் காட்டுகின்றன என்ற கட்டுப்பாட்டிற்குக் கீழ்படிகின்றது. இந்தியில் சொல் குழுக்கள் மூன்று முக்கிய வகைகளாக இருக்கலாம் - பெயர்ச்சொல் குழுக்கள், வினை குழுக்கள் மற்றும் வினையடைச் சொல் குழுக்கள். இணைப்புகள் மற்றும் பிற அவ்யாக்கள் தனித்த சொல்

குழுக்களை உருவாக்குகின்றன. குழுத் தலைகள் என்ற வார்த்தையாக குழுவானது வலமிருந்து இடமாக செய்யப்படுகிறது - பெயர்ச்சொல் குழுக்களில் பெயர்ச்சொற்கள், வினை குழுக்களில் வரையறுக்கப்பட்ட வினைச்சொற்கள் மற்றும் வினையடைச் குழுக்களில் உள்ள வினையடைச் சொற்கள் அனைத்தும் அந்தந்த சொல் குழுக்களின் தீவிர வலது முனையில் நிகழ்கின்றன. உள்ளூர் சொல் குழுக்கள் நான்கு கட்டங்களாக இயங்குகின்றன: கிளப்பிங், விதைத்தல், குழுமம் மற்றும் வடிகட்டுதல். இது சொல்வகை அடையாளத்தால் உருவாக்கப்படும் ஒவ்வொரு சொல்வகை அடையாளப்படுத்தப்பட்ட வரிசையிலும் இயங்குகிறது.

ஒன்றிணைப்பி (Clubber)

முன்கூட்டியே செயலாக்க கட்டம் அல்லது ஒன்றிணைப்பி அடிப்படையில் சில சொல் கோர்வைகளைக் குழுவாகக் கொண்டிருக்க வேண்டும், ஏனெனில் அவை குழுமும் போது ஒற்றை அலகு என்று கருதப்பட வேண்டும். இது இரண்டு வகையான குழுமலைக் கொண்டுள்ளது:

(1) இரட்டித்த சொற்களை ஒன்றிணைதல் (Clubbing duplicated words): *ஓட்டிஓடி* போன்ற இரட்டித்த வினைச்சொற்கள் ஒன்றிணைந்து ஒரு வினையடையை உருவாக்குகின்றன. எந்தவொரு தொடர்ச்சியான முடிவற்ற வினைச்சொற்களும் ஒரு வினையடையாக இணைக்கப்படலாம்.

(2) அவ்யா வரிசைகளை ஒன்றிணைத்தல் (Clubbing *avyaya* sequences): சிறப்பு அவ்யா கோர்வைகள் ஒன்றிணைக்கப்பட்டு ஒற்றை முன்னொட்டுகளாகச் செயல்படுகின்றன. இத்தகைய அவ்யா வரிசைகள் ஒரு அட்டவணையில் இருந்து பார்க்கப்படுகின்றன மற்றும் இந்தியில் *ke liye, ke dvArA, ke kAraNa, ke mAdyama se* போன்ற வரிசைகளை உள்ளடக்கியது. இந்தியில் *ke madhya*, *ke bicha, ke niche* போன்றவை பின்னொட்டுக்களாகச் செயல்படும் அவ்யா கோர்வைகளை உருவாக்கலாம். இத்தகைய கோர்வைகள் வழக்கமாக பின்னொட்டுசார் அவயாக்களுடன் தொடர்புடைய *meM, para, se, taka* போன்ற பின்னொட்டுகளுடன் முடிவடையும்.

ஒன்றிணைப்பதற்கான வழிமுறை எளிதானது, அது பின்வருமாறு அமையும்:

அல்காரிதம் க்ளப் (CLUB)

அனைத்து இரட்டித்த சொற்களையும் அடையாளம் காணவும்.

ஒரு அட்டவணையைத் உருவாக்குவதன் மூலம் ஒன்றிணைக்கபட வேண்டிய அனைத்து அவ்யா வரிசைகளையும் அடையாளம் காணவும்.

அடையாளம் காணப்பட்ட சொற்களை ஒன்றிணைப்பிகள் அதனுடன் தொடர்புடைய சொல் வகைப்பாட்டுடன் ஒற்றைச் சொல் அலகாகக் கருதும்.

விதைப்பி (Seeder)

விதை விதை சொல் குழுக்களின் அனைத்து இறுதி புள்ளிகளையும் அடையாளம் காட்டுகிறது. இது வெறுமனே ஒவ்வொரு சொல்லின் சொல்சார் அடையாளதைப் பார்த்து, அதன்படி சொல்லை ஒரு விதை என்று குறிக்கலாமா வேண்டாமா என்பது குறித்து ஒரு முடிவை எடுக்கிறது. விதைப்பி அனைத்து பெயர்ச்சொற்கள், வினைச்சொற்கள், வினையடைகள், இணைப்புக்கிளவிகள், பின்னொட்டுகள் மற்றும் பிற அவயாக்கள் ஆகியவற்றைக் குறிக்கிறது - ஒரு குழுவின் ஒவ்வொரு வலப்பக்க உறுப்பும் விதை என்று அழைக்கப்படுகிறது.

குழுமி (Grouper)

ஒவ்வொரு வகையான உள்ளூர் சொல் குழுவிற்கும் பொருத்தமாக நாம் ஒரு சீரான வெளிப்பாட்டை (regular expression (RE)) பெற்றுள்ளோம். அந்த வகையின் எந்தவொரு சொல் குழுவும் அந்த சீரான வெளிப்பாட்டை பூர்த்தி செய்ய வேண்டும். இருப்பினும், சீரான வெளிப்பாட்டை திருப்திப்படுத்துவது சொற்களின் வரிசைக்குத் தொடர்புடைய சொல் குழுவாக இருக்க போதுமான நிபந்தனை அல்ல. எனவே, குழுமல் மயக்கம் ஏற்படும் போது, அனைத்து குழுக்களும் செல்லுபடியாகாது.

குழுவானது குழுவாக வலதுபுறம் குழுவாக இல்லாத விதைகளைத் தேர்வுசெய்கிறது மற்றும் அதன் விதைகளின் சொல் வகையைப் பொறுத்து தொடர்புடைய முற்று நிலை தானியங்கியை (Finite State Automata (FSA)) அழைக்கிறது. விதைகளில் இருந்து பின்னோக்கித் தொடங்கி சொற்களின் சொல் வகைகளின் வரிசையில் குழுமி முற்று நிலை தானியங்கியை இயக்குகிறது மற்றும் அது தவறான மாற்றத்தை (invalid transition) எதிர்கொள்ளும் வரை தொடர்கிறது, அல்லது வாக்கியம் நிறைவடைகிறது. முற்று நிலை தானியங்கி இறுதி நிலைக்கு நுழைந்த கடைசி குழுவான சொல் எந்த கட்டத்திலும் நினைவில் வைக்கப்படுகிறது – செண்டினல் (sentinel). தவறான மாற்றம் அல்லது வாக்கியத்தின் தொடக்கத்தை எதிர்கொள்ளும்போது, செண்டினல் வரை அதிகபட்சமாக குழுவாக்கம் செய்யப்படுகிறது.

பின்வருபவை ஏற்பட்டால் குழுமத்தின் போது மயக்கம் ஏற்படலாம் - முற்று நிலை தானியங்கி ஒரு இறுதி நிலைக்கு நுழைகிறது, மேலும் முற்று நிலை தானியங்கிக்கு அடுத்த உள்ளீடு ஒரு விதை. பின்னர் குழுமியிடம் இரண்டு விருப்பங்கள் இருக்கும்: தற்போதைய

குழுவுடன் தொடர வேண்டும் அல்லது தற்போதைய குழுவை முடித்துவிட்டு ஒரு புதிய குழுவைத் தொடங்க வேண்டும். இதுபோன்ற சந்தர்ப்பங்களில், குழுமி இரண்டு குழுக்களையும் வெளியிடுகிறது; எது சரியானது என அடுத்த கட்டத்தில் தீர்மானிக்கப்படுகிறது (முடிந்தால்).

வடிகட்டி

குழுமி வெளியீடும் குழுமல்கள் அனைத்தும் சரியாக இருக்கவேண்டும் என்பதில்லை. எனவே, வடிகட்டி பின்வரும் கட்டுப்பாடுகளின் அடிப்படையில் சில குழுமல்களை நிராகரிக்கிறது: குழுமப்படாத சொற்கள் உள்ள எந்த குழுமல்களும் நிராகரிக்கப்படுகின்றன.

- பின்வரும் கட்டுப்பாடுகளுக்குக் கீழ்ப்படியாத பெயர்ச்சொல் குழுக்களின் எந்தக் குழுமல்களும் நிராகரிக்கப்படுகின்றன: ஒரு பெயர்ச்சொல் குழுவில் ஒன்றுக்கு மேற்பட்ட பெயர்ச்சொற்கள் இருந்தால் (இணைப்புக்கிளவியால் சேர்க்கப்பட்ட), அனைத்து பெயர்ச்சொற்களின் திரிபுகளும்/பின்னொட்டுகளும் ஒரே மாதிரியாக இருக்க வேண்டும், அல்லது கடைசி பெயர்ச்சொல் மட்டுமே திரிபுறலாம்/பின்னொட்டுக்களைக் கொண்டிருக்கலாம்.
- பின்வரும் கட்டுப்பாடுகளுக்குக் கீழ்ப்படியாத வினை குழுக்களின் எந்தவொரு குழுமல்களும் நிராகரிக்கப்படுகின்றன: எந்தவொரு நெருங்கிய ஜோடி இணைப்புக்கிளவி குழுக்களுக்கிடையில் அல்லது ஒரு இணைப்புக்கிளவி குழுவிற்கும் ஒரு வாக்கியத்தின் தொடக்கதிற்கும் / முடிவிற்கும் இடையில், குறைந்தது ஒரு வினை குழுவாக இருக்க வேண்டும் மற்றும் அதிகபட்சம் ஒரு முற்று வினை குழு இருக்க வேண்டும்.

முடிவுரை

சொல்வகைப்பாடு அடையாளப்படுத்தி (பிஓஎஸ் டேக்கர்) மற்றும் இருப்பிடம்சார் சொல் குழுமி (எல்.டபிள்யூ.ஜி) ஆகியவற்றின் செயல்திறனைப் புள்ளிவிவர அடிப்படையில் பரிசோதனைசெய்யச் சொல்சார் வளங்கள் தேவை; ஆனால் இது பரந்த அளவிலான மொழி நிகழ்வுகளை உள்ளடக்கியது மற்றும் இந்தியில் உள்ள நான்கு முக்கிய உள்ளூர் சார்புகளை துல்லியமாகப் பிடிக்கிறது. இருப்பிடம்சார் சொல் குழுமி இந்திய மொழிகளில் இயற்கையான பேச்சை உருவாக்க அனுமதிக்கும். மேலும், ஒரு திறன் மிக்க இருப்பிடம்சார் சொல் குழுமி அதைப் பின்தொடரும் எந்தப் பாகுபடுத்தியின் சுமையையும் பெரிதும் குறைக்கிறது. உள்ளூர் நிலையான ஒழுங்கு கட்டமைப்புகள் பற்றிய ஒரு யோசனை ஆங்கிலத்தைப் போலவே மொழி உருவாக்கத்திற்கும் உதவுகிறது (Knight and Hatzivassiloglou, 1995). எதிர்கால வேலை வினை

குழும கட்டுப்பாடுகளை உருவாக்குவதிலும், சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கான விதி அடிப்படையின் அளவை அதிகரிப்பதிலும் உள்ளது. அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியை உருவாக்குவதே முக்கிய முன்முயற்சியாகும், அதன்படி ஒரு பெரிய அடையாளக் குழுமத்தைப் பயன்படுத்தி சொல்வகைப்பாடு அடையாளப்படுத்தலின் புள்ளிவிவர மாதிரிகளை உருவாக்குதல்.

6.4.18. பெயரிடப்பட்ட-சொல் அறிதல் ஒழுங்குமுறைகள் (Named-entity recognition NER Systems)

பெயரிடப்பட்ட-சொல் அறிதல் (Named-entity recognition NER) (சொல் அடையாளம்காணல்/entity identification, சொல் பிரித்தல்/entity chunking மற்றும் சொல் பிரித்தெடுத்தல்/entity extraction என்றும் அழைக்கப்படுகிறது) என்பது தகவல் பிரித்தெடுத்தலின் ஒரு துணைப் பணியாகும்; இது கட்டமைக்கப்படாத உரையில் குறிப்பிடப்பட்டுள்ள பெயரிடப்பட்ட சொற்களை நபர் பெயர்கள், நிறுவனங்கள், இருப்பிடங்கள், மருத்துவ குறியீடுகள், நேர வெளிப்பாடுகள், அளவுகள், பண மதிப்புகள், சதவீதங்கள் போன்றவைகளில் இடத்தைத் தேடி அறிந்து வகைப்படுத்துவதாகும்.

பெயரிடப்பட்ட-சொல் அறிதல் கணினிகளைப் பற்றிய பெரும்பாலான ஆராய்ச்சிகள் இது போன்ற அறிவிக்கப்படாத உரையை எடுத்துக்கொள்வதாக கட்டமைக்கப்பட்டுள்ளன:

ஜிம் 2006 இல் ஆக்மி கார்ப் நிறுவனத்தின் 300 பங்குகளை வாங்கினார்.

நிறுவனங்களின் பெயர்களை முன்னிலைப்படுத்தும் உரையின் சிறுகுறிப்பு தொகுதியை உருவாக்குதல்:

[ஜிம்]$_{\text{நபர்}}$ [2006]$_{\text{காலம்}}$ -இல் [ஆக்மி கார்ப் நிறுவனம்]$_{\text{அமைப்பு}}$ –இன் 300 பங்குகளை வாங்கினார்.

இந்த எடுத்துக்காட்டில், ஒரு டோக்கன், இரண்டு டோக்கன் நிறுவனத்தின் பெயர் மற்றும் ஒரு தற்காலிக வெளிப்பாடு ஆகியவற்றைக் கொண்ட ஒரு நபரின் பெயர் கண்டறியப்பட்டு வகைப்படுத்தப்பட்டுள்ளது.

ஆங்கிலத்திற்கான அதிநவீன என்.இ.ஆர் அமைப்புகள் மனிதனின் செயல்திறனை உருவாக்குகின்றன. எடுத்துக்காட்டாக, MUC-7 க்குள் நுழையும் சிறந்த அமைப்பு 93.39% எஃப்-அளவையும், மனித அடையாளப்படுத்துநர் 97.60% மற்றும் 96.95% மதிப்பெண்களையும் பெற்றனர்.

பெயரிடப்பட்ட சொல் அறிதல் தளங்கள்

குறிப்பிடத்தக்க NER தளங்களில் பின்வருவன அடங்கும்:

- கேட்/GATE பல மொழிகள் மற்றும் களங்களில் பெட்டியின் வெளியே NER ஐ ஆதரிக்கிறது, இது ஒரு வரைகலை இடைமுகம் மற்றும் ஜாவா API வழியாக பயன்படுத்தக்கூடியது.
- ஓபன்என்எல்பி/OpenNLP விதி அடிப்படையிலான மற்றும் புள்ளிவிவர பெயரிடப்பட்ட-சொல் அறிதலை உள்ளடக்கியது.
- ஸ்பேசி/SpaCy வேகமான புள்ளிவிவர என்.இ.ஆர்./NER மற்றும் திறந்த-மூல பெயரிடப்பட்ட-சொல் கண்டறிவான்களைக் (named-entity visualizer) கொண்டுள்ளது.

'பெயரிடப்பட்ட சொல்' என்ற வெளிப்பாட்டில் பெயரிடப்பட்ட சொல், சொற்கள் அல்லது சொற்றொடர்கள் போன்ற ஒன்று அல்லது பல சரங்களைச் சுட்டிக்காட்டித் தொடர்ந்து (நியாயமான முறையில்) நின்று அந்த சொற்களின் எல்லையைக் கட்டுப்படுத்துகிறது.

அணுகுமுறைகள்

மொழியியல் இலக்கண அடிப்படையிலான நுட்பங்களையும் இயந்திர கற்றல் போன்ற புள்ளியியல்சார் மாதிரிகளையும் பயன்படுத்தும் NER ஒழுங்குமுறைகள் உருவாக்கப்பட்டுள்ளன. ஆனால், கையால் அனுபவம் வாய்ந்த கணினி மொழியியலாளர்களால் குறைந்த நினைவுகூரல் மற்றும் கால அளவு வேலைகளின் செலவில் வடிவமைக்கப்பட்ட இலக்கண அடிப்படையிலான அமைப்புகள் பொதுவாக சிறந்த துல்லியத்தைப் பெறுகின்றன. புள்ளியியல்சார் NER அமைப்புகளுக்கு பொதுவாக கைமுறையாக சிறுகுறிப்பு செய்யப்பட்ட பயிற்சி தரவு தேவைப்படுகிறது. சிறுகுறிப்பு முயற்சியின் ஒரு பகுதியைத் தவிர்க்க அரை மேற்பார்வை அணுகுமுறைகள் பரிந்துரைக்கப்பட்டுள்ளன. நிபந்தனை சீரற்ற புலங்கள் ஒரு பொதுவான தேர்வாக (conditional random fields) எந்திரக் கற்ற NERஐச் செய்ய பல வேறுபட்ட வகைப்படுத்தி வகைகள் பயன்படுத்தப்பட்டுள்ளன,

6.5 மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளுக்கு மூலவளமாகப் பெருந்தரவு

மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளுக்கு மூலவளமாகப் பெருந்தரவு என்ற தலைப்பின் கீழ் பின்வருவனவற்றை உள்ளடக்க இயலும்:

1. மொழி மூலவளம் பெறும் ஒழுங்குமுறை
2. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை

3. பன்மொழித் தகவல் பெறும் ஒழுங்குமுறை

4. மொழிகடந்த தகவல் மீட்பு ஒழுங்குமுறை

6.5.1. மொழி மூலவளம் பெறும் ஒழுங்குமுறைகள் (Language resource access systems)

மொழி வளம் என்ற சொல், பேச்சு அல்லது மொழித் தரவு மற்றும் விளக்கங்களை இயந்திரம் படிக்கக்கூடிய வடிவத்தில் குறிக்கிறது, இது இயற்கை மொழி மற்றும் பேச்சு வழிமுறைகள் அல்லது ஒழுங்குமுறைகளை உருவாக்குவதற்கும் மேம்படுத்துவதற்கும் மதிப்பீடு செய்வதற்கும் அல்லது மென்பொருள் உள்ளூர்மயமாக்கல் மற்றும் மொழி சேவைத் தொழில்கள், மொழி ஆய்வுகள், மின்னணு வெளியீடு, சர்வதேச பரிவர்த்தனைகள் ஆகியவற்றிற்கும் பொருள் பகுதி நிபுணர்கள் மற்றும் இறுதி பயனர்கள் ஆகியோருக்கும் முக்கிய வளங்களாகப் பயன்படுத்தப்படுகிறது

மொழி வளங்களின் எடுத்துக்காட்டுகள் எழுதப்பட்ட மற்றும் பேச்சு தரவுத்தொகுதி, கணினிசார் அகராதி, கலைச்சொல் தரவுத்தளங்கள், பேச்சு சேகரிப்பு போன்றவை. இந்த மொழி வளங்கள் மற்றும் பிற வளங்களைக் கையகப்படுத்துதல், தயாரித்தல், சேகரித்தல், மேலாண்மை, தனிப்பயனாக்குதல் மற்றும் பயன்படுத்துவதற்கு அடிப்படை மென்பொருள் கருவிகள் முக்கியம்.

பல மொழிகளைக் கையாள்வது சிக்கலான போதிலும், கலாச்சார மற்றும் மொழி வேறுபாடுகள் பாதுகாக்கப்பட வேண்டிய ஒரு தனித்துவமான சொத்து என்பது உணரப்படவேண்டும். இயற்கையான மொழியைக் கையாளும் தொழில்துறை தொழில்நுட்ப பயன்பாடுகளில் பெரும்பாலானவை, அதாவது இயந்திர மொழிபெயர்ப்பு, குறுக்கு தகவல் மீட்டெடுப்பு, பன்மொழி தகவல் பிரித்தெடுத்தல், தானியங்கி ஆவண அட்டவணைப்படுத்தல், கேள்வி பதில், இயற்கை மொழி இடைமுகங்கள் போன்றவை, மொழி வளங்களை முக்கியமான கூறுகளாக உள்ளடக்குகின்றன. மொழி தொழில்நுட்பங்கள் மொழி சுதந்திர இயந்திரங்களைக் கொண்டிருக்கலாம் என்றாலும், அவை நிஜ வாழ்க்கையை செயல்படுத்த மொழி வளங்களின் வடிவத்தின் கீழ் மொழி சார்ந்த அறிவைப் பெறுவதைப் பொறுத்தது. அதே நேரத்தில், மொழி வளங்களின் ஒரு முக்கியமான வெகுஜன ஆராய்ச்சி மற்றும் தொழில்நுட்ப வளர்ச்சியில் முன்னேற்றத்தை சாத்தியமாகவும் விரைவாகவும் செய்ய முடியும் என்பது நிரூபிக்கப்பட்டுள்ளது. பயிற்சி ஒழுங்குமுறைகளுக்கு அதிக அளவு தரவுகளை சேகரிக்க வேண்டிய தேவை 1992 ஆம் ஆண்டில் மொழியியல் தரவு கூட்டமைப்பை (Linguistic Data Consortium) உருவாக்கியது.

பன்மொழிய இந்தியாவிற்கான இயற்கைமொழிய் ஆய்வுப் பயன்பாடுகளின் வளர்ச்சிக்கு மொழி வளங்கள் முக்கியம். பொருத்தமான பன்மொழி மொழி வள உள்கட்டமைப்பை உணர அனைத்து பங்குதாரர்களிடமிருந்தும் ஒரு கூட்டு மற்றும் ஒருங்கிணைந்த முயற்சி தேவை. கடந்த தசாப்தத்தில் தொழில்நுட்ப வளர்ச்சிகளில் கணிசமான முன்னேற்றம் காணப்பட்டாலும், அனைத்து மொழிகளுக்கும் மொழி தொழில்நுட்ப சமூகத்தின் தற்போதைய துண்டாக்கம் மற்றும் ஏற்றத்தாழ்வைக் கடக்கும் குறிப்பிடத்தக்க சவால் இன்னும் ஒரு பிரச்சினையாகவே உள்ளது. இன்றைய முன்னுரிமைகள் பற்றிய பகிரப்பட்ட பார்வையின் குடையின் கீழ், திட்டங்கள், செயல்கள் மற்றும் செயல்பாடுகளை ஒருங்கிணைப்பதன் மூலம் மொழி வளங்கள் மற்றும் தொழில்நுட்பங்களின் முழு வரிசைப்படுத்தல் ஒருங்கிணைக்கப்படும் எதிர்காலத்தை வடிவமைக்க முடியும்.

மூலவள ஆவணம்

"மொழி வளங்கள் துல்லியமாகவும் நம்பகத்தன்மையுடனும் ஆவணப்படுத்தப்பட்டுள்ளன என்பதை உறுதிப்படுத்தவும்" மொழி வளங்களின் துல்லியமான மற்றும் நம்பகமான ஆவணங்கள் மறுக்க முடியாத தேவை. அதற்குப் பதிலாக, இன்றைய நிலவரப்படி, மொழி வளங்கள் இன்னும் மோசமாக ஆவணப்படுத்தப்பட்டுள்ளன அல்லது ஆவணப்படுத்தப்படவில்லை, மேலும் கிடைக்கும்போது கூட ஆவணங்கள் பெரும்பாலும் கண்டுபிடிக்க எளிதானது அல்ல.

ஆவணப்படுத்தல் மொழி வளங்களை வடிவமைத்து உருவாக்கியவர்களிடமிருந்து வேறுபட்ட நபர்களால் பயன்படுத்த அனுமதிக்கிறது. "ஆராய்ச்சியாளர்கள் மற்றும் பயிற்சியாளர்கள் தங்களுக்குத் தேவையான தரவைக் கண்டுபிடித்து, அணுகலாம் மற்றும் செயலாக்க முடியும். தரவைப் பயன்படுத்துவதற்கும் புரிந்து கொள்வதற்கும் அவர்களின் திறனில் அவர்கள் நம்பிக்கையுடன் இருப்பார்கள், மேலும் தரவை எந்த அளவிற்கு நம்பலாம் என்பதை அவர்கள் மதிப்பீடு செய்யலாம்.

ஆவணப்படுத்தல் முடிந்தவரை முழுமையானதாக இருக்க வேண்டும், மேலும் தரவு வடிவம் மற்றும் தரவு உள்ளடக்கம், உற்பத்தி சூழல் மற்றும் ஏற்கனவே உள்ள பயன்பாடுகள் பற்றிய தகவல்களை உள்ளடக்கியிருக்க வேண்டும். மனித பயனர்களுக்கு உதவும் தகவல் தேவை:

அ) ஒரு ஆதாரத்தைக் கண்டுபிடித்து, கொடுக்கப்பட்ட பயன்பாட்டிற்கான அதன் பயனை மதிப்பிடுதல்

ஆ) உற்பத்தி செயல்முறை, சிறந்த நடைமுறைகளின் பயன்பாடு மற்றும் நோக்கம் கொண்ட பயன்படுத்தல் ஆகியவற்றைப் புரிந்து கொள்ளுதல்

இ) ஒரு வளத்தின் தரத்தை மதிப்பிடுதல்

ஈ) செயல்முறைகளை நகலெடு மற்றும் முடிவுகள்

இ) தனித்துவமான அல்லது ஆவணப்படுத்தப்பட்ட பிழைகளைக் கையாளுதல். இயந்திரங்களுக்கு இவை தேவை (இயந்திரம் புரிந்துகொள்ளக்கூடியவை) தகவல்:

அ) வளங்களைக் கண்டுபிடித்து ஒப்பிடுதல்

ஆ) வடிவங்கள் மற்றும் சிறுகுறிப்புகளை சரிபார்த்தல்

இ) சிறுகுறிப்புகளை சரியான முறையில் ஆய்தல்

ஈ) கொடுக்கப்பட்ட பயன்பாட்டிற்காக ஒரு வளத்தின் தொடர்புடைய பகுதிகளை மீட்டெத்தல்

இ) இன்னும் ஆராயப்படாத புதிய பயன்பாடுகளை இயக்குதல்.

6.5.2. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள்

இயந்திர மொழிபெயர்ப்பு, சில நேரங்களில் (Machine translation) MT/எம்டி என்ற சுருக்கத்தால் குறிப்பிடப்படுகிறது (கணினி உதவிபெறும் மொழிபெயர்ப்பு (computer-aided translation), இயந்திர உதவி பெறும் மனித மொழிபெயர்ப்பு (machine-aided human translation (MAHT)) அல்லது ஊடாடும் மொழிபெயர்ப்பு (interactive translation) ஆகியவற்றுடன் குழப்பமடையக்கூடாது), இது உரை அல்லது பேச்சை ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு மொழிபெயர்ப்பதற்கு மென்பொருளின் பயன்பாட்டை ஆராயும் கணினி மொழியியலின் துணைத் துறையாகும்.

ஒரு அடிப்படை மட்டத்தில், இயந்திர மொழிபெயர்ப்பு ஒரு மொழியில் சொற்களுக்கு இயந்திர மாற்றீட்டை மற்றொரு மொழியில் செய்கிறது, ஆனால் அது மட்டும் ஒரு நல்ல மொழிபெயர்ப்பை அரிதாகவே உருவாக்குகிறது, ஏனெனில் இலக்கு மொழியில் முழு சொற்றொடர்களையும் அவற்றின் நெருங்கிய சரிநிகர்களையும் புரிந்துகொள்வது தேவை. ஒரு மொழியில் உள்ள எல்லா சொற்களுக்கும் மற்றொரு மொழியில் சமமான சொற்கள் இல்லை, பல சொற்களுக்கு ஒன்றுக்கு மேற்பட்ட அர்த்தங்கள் உள்ளன. கூடுதலாக, கொடுக்கப்பட்ட இரண்டு மொழிகளில் முற்றிலும் மாறுபட்ட கட்டமைப்புகள் இருக்கலாம்.

தரவுத்தொகுதி புள்ளிவிவர மற்றும் நரம்பியல் நுட்பங்களுடன் இந்த சிக்கலைத் தீர்ப்பது வேகமாக வளர்ந்து வரும் ஒரு துறையாகும்; இது மொழியியல் வகைப்பாட்டியல் (linguistic typology) வேறுபாடுகள், மரபுச்சொற்களின் (மரபுத்தொடர்களின்) மொழிபெயர்ப்பு மற்றும் முரண்பாடுகளைத் தனிமைப்படுத்தல் ஆகியவற்றைக் கையாண்டு சிறந்த மொழிபெயர்ப்புகளுக்கு வழிவகுக்கிறது (Thomas Fritz 2012).

தற்போதைய இயந்திர மொழிபெயர்ப்பு மென்பொருள் பெரும்பாலும் களம் (டொமைன்) அல்லது தொழில் (வானிலை அறிக்கைகள் போன்றவை) மூலம் தனிப்பயனாக்கத்தை அனுமதிக்கிறது; அனுமதிக்கக்கூடிய மாற்றீடுகளின்/பதிலீடுகளின் வரம்பைக் கட்டுப்படுத்துவதன் மூலம் வெளியீட்டை மேம்படுத்துகிறது. முறையான (formal) அல்லது வாய்பாட்டு/சூத்திர மொழி (formulaic language) பயன்படுத்தப்படும் களங்களில் இந்த நுட்பம் குறிப்பாக பயனுள்ளதாக இருக்கும். அரசாங்க மற்றும் சட்ட ஆவணங்களின் இயந்திர மொழிபெயர்ப்பு, உரையாடல் அல்லது குறைந்த தரப்படுத்தப்பட்ட உரையை விட எளிதில் பயன்படுத்தக்கூடிய வெளியீட்டை உருவாக்குகிறது.

மேம்பட்ட வெளியீட்டுத் தரத்தையும் மனித தலையீட்டால் அடைய முடியும்: எடுத்துக்காட்டாக, உரையில் எந்தச் சொற்கள் இயற்பெயர்கள் (proper names) என்பதை பயனர் பொருண்மைமயக்கமின்றி அடையாளம் கண்டால் சில ஒழுங்குமுறைகள் இன்னும் துல்லியமாக மொழிபெயர்க்க இயலும். இந்த நுட்பங்களின் உதவியுடன், மனித மொழிபெயர்ப்பாளர்களுக்கு உதவுவதற்கான ஒரு கருவியாக இயந்திர மொழிபெயர்ப்பு பயனுள்ளதாக நிரூபிக்கப்பட்டுள்ளது, மிகக் குறைந்த எண்ணிக்கையிலான சந்தர்ப்பங்களில், பயன்படுத்தக்கூடிய வெளியீட்டைக் கூட உருவாக்க முடியும் (எ.கா., வானிலை அறிக்கைகள்).

இயந்திர மொழிபெயர்ப்பின் முன்னேற்றமும் ஆற்றலும் அதன் வரலாற்றின் வழி அதிகம் விவாதிக்கப்பட்டுள்ளன. 1950களில் இருந்து, பல அறிஞர்கள், முதல் மற்றும் குறிப்பாக யெஹோசுவா பார்-ஹில்லெல் (Yehoshua Bar-Hillel 1964) உயர்தர முழுமையான தானியங்கி இயந்திர மொழிபெயர்ப்பை அடைவதற்கான சாத்தியத்தை கேள்விக்குள்ளாக்கியுள்ளனர். மொழிபெயர்ப்பு செயல்முறையை தானியக்கமாக்குவதற்கு கொள்கை ரீதியான தடைகள் இருப்பதாக சில விமர்சகர்கள் கூறுகின்றனர் (Madsen, 2009).

மொழிபெயர்ப்பு செயல்முறை (Translation process)

மனித மொழிபெயர்ப்பு செயல்முறை இவ்வாறு விவரிக்கப்படலாம்:

1. மூல உரையின் பொருளை குறித்திறவு (டிகோடிங்/decoding) செய்தல்; மற்றும்
2. இலக்கு மொழியில் இந்த பொருளை மீண்டும் குறியமாக்கம் (re-encoding) செய்தல்.

இந்த வெளிப்படையான எளிய நடைமுறைக்கு பின்னால் ஒரு சிக்கலான அறிவாற்றல் செயல்பாடு உள்ளது. மூல உரையின் பொருளை முழுவதுமாக குறித்திறவு/டிகோட் செய்ய, மொழிபெயர்ப்பாளர் உரையின் அனைத்து பண்புக்கூறுகளையும் விளக்கி பகுப்பாய்வு செய்ய வேண்டும்; இது மூல மொழியின் இலக்கணம், பொருண்மையியல், தொடரியல், மரபுத்தொடர் போன்றவற்றின் ஆழமான அறிவு தேவைப்படும் ஒரு செயல்முறையாகும்; அத்துடன் பேசுபவர்களின் கலாச்சாரம் பற்றிய அறிவும் வேண்டும். இலக்கு மொழியில் பொருளை மீண்டும் குறிமாக்கம் செய்ய மொழிபெயர்ப்பாளருக்கு அதே ஆழமான அறிவு தேவை.

இயந்திர மொழிபெயர்ப்பில் உள்ள சவால் இதில் உள்ளது: ஒரு நபரைப் போலவே ஒரு உரையை "புரிந்துகொள்ளும்" மற்றும் ஒரு நபரால் எழுதப்பட்டதைப் போல இலக்கு மொழியில் ஒரு புதிய உரையை "உருவாக்கும்" ஒரு கணினியை எவ்வாறு நிரல் செய்வது,

அதன் பொதுவான பயன்பாட்டில், இது தற்போதைய தொழில்நுட்பத்திற்கு அப்பாற்பட்டது. இது மிக வேகமாக இயங்கினாலும், எந்தவொரு தானியங்கி மொழிபெயர்ப்பு நிரலும் அல்லது நடைமுறையும், மனித பங்களிப்பு இல்லாமல், ஒரு மனித மொழிபெயர்ப்பாளர் உருவாக்கக்கூடிய தரத்திற்கு மிக அருகில் கூட வெளியீட்டை உருவாக்க முடியாது. எவ்வாறாயினும், அது என்ன செய்ய முடியும் என்பது ஒரு பொதுவான, அபூரணமாக இருந்தாலும், அசல் உரையின் தோராயமாக, அதன் "சாராம்சத்தை" பெறுகிறது (ஒரு செயல்முறை "ஜிஸ்டிங்" என்று அழைக்கப்படுகிறது). மனித மொழிபெயர்ப்பாளரின் வரையறுக்கப்பட்ட மற்றும் விலையுயர்ந்த நேரத்தை சிறப்பாகப் பயன்படுத்துவது உட்பட பல நோக்கங்களுக்காக இது போதுமானது, மொத்த துல்லியம் இன்றியமையாத அந்த நிகழ்வுகளுக்கு ஒதுக்கப்பட்டுள்ளது. இந்த சிக்கலைப் பல வழிகளில் அணுகலாம்; பரிணாம வளர்ச்சியின் மூலம் துல்லியம் மேம்பட்டுள்ளது.

அணுமுறைகள் (Approaches)

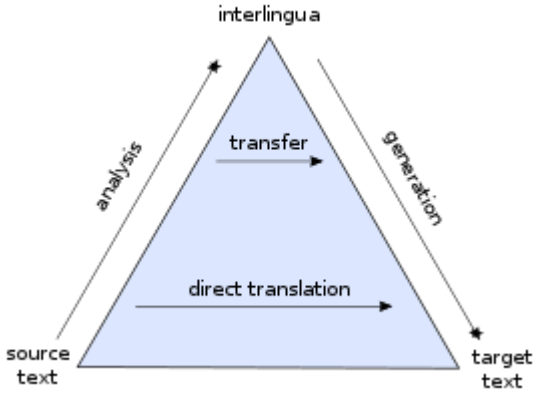
இயந்திர மொழிபெயர்ப்பு மொழியியல் விதிகளின் அடிப்படையில் ஒரு முறையைப் பயன்படுத்தலாம்; அதாவது சொற்கள் மொழியியல் முறையில் மொழிபெயர்க்கப்படும் - அதாவது இலக்கு மொழியின் மிகவும் பொருத்தமான (வாய்வழியாகப் பேசும்) சொற்கள் மூல மொழியில் உள்ளவற்றை இடம்பெயர்க்கும். இயந்திர மொழிபெயர்ப்பின் வெற்றிக்கு இயற்கையான மொழி

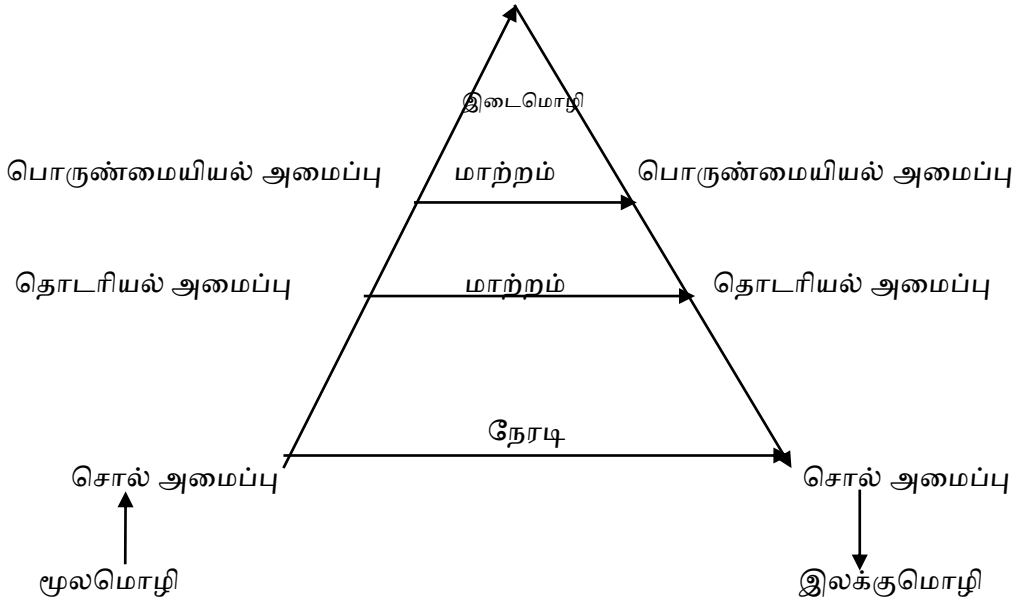
புரிதலின் சிக்கல் முதலில் தீர்க்கப்பட வேண்டும் என்று பெரும்பாலும் வாதிடப்படுகிறது (John Lehrberger 1988).

பொதுவாக, விதி அடிப்படையிலான முறைகள் ஒரு உரையை பகுப்பாய்வு செய்கின்றன; இதன் விளைவாக ஒரு இடைத்நிலை குறியீட்டு உருப்படுத்தத்தை (intermediary, symbolic representation) உருவாக்குகின்றன; இதிலிருந்து இலக்கு மொழியில் உள்ள உரை உருவாக்கப்படுகிறது. இடைநிலை உருப்படுத்தத்தின் தன்மைக்கு ஏற்ப, ஒரு அணுகுமுறை இடைமொழி இயந்திர மொழிபெயர்ப்பு (interlingual machine translation) அல்லது பரிமாற்ற அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (transfer-based machine translation) என விவரிக்கப்படுகிறது. இந்த முறைகளுக்கு உருபனியல், தொடரியல் மற்றும் பொருண்மையியல் தகவல்கள் மற்றும் பெரிய விதிமுறைகள் கொண்ட விரிவான அகராதிகள் தேவைப்படுகின்றன.

போதுமான தரவு கொடுக்கப்பட்டால், இயந்திர மொழிபெயர்ப்பு நிரல்கள் பெரும்பாலும் ஒரு மொழியின் சொந்தமொழி பேசுபவருக்கு பிற சொந்தமொழி பேசுபவரால் எழுதப்பட்டவற்றின் தோராயமான பொருளைப் பெறுவதற்கு போதுமானதாக இருக்கும். குறிப்பிட்ட முறையை ஆதரிக்க சரியான வகையான போதுமான தரவைப் பெறுவது சிரமம். எடுத்துக்காட்டாக, புள்ளிவிவர முறைகள் செயல்படத் தேவையான தரவுகளின் பெரிய பன்மொழியத் தரவுத்தொகுதி இலக்கண அடிப்படையிலான முறைகளுக்கு அவசியமில்லை. ஆனால், இலக்கண முறைகளுக்கு அவர்கள் பயன்படுத்தும் இலக்கணத்தை கவனமாக வடிவமைக்க திறமையான மொழியியலாளர் தேவை. நெருங்கிய தொடர்புடைய மொழிகளுக்கு இடையில் மொழிபெயர்க்க, விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு என குறிப்பிடப்படும் நுட்பம் பயன்படுத்தப்படலாம்.

கீழே தரப்பட்டுள்ள படம் பெர்னார்ட் வாகோயிஸின் பிரமிடு (Bernard Vauquois' pyramid) ஆகும். இது இடைநிலை உருப்படுத்தத்தின் ஒப்பீட்டு ஆழத்தைக் காட்டுகிறது: உச்சத்தில் உள்ளது இடைமொழி இயந்திர மொழிபெயர்ப்பு (intelingua machine translation), அதைத் தொடர்ந்து பரிமாற்ற அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (transfer machine translation, பின்னர் நேரடி இயந்திர மொழிபெயர்ப்பு (direct machine translation).





விதி அடிப்படையிலானது

விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (rule-based machine translation (RBMT) முன்னுதாரணத்தில் பரிமாற்ற அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு, இடைமொழி இயந்திர மொழிபெயர்ப்பு மற்றும் அகராதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஆகியவை அடங்கும். இந்த வகை மொழிபெயர்ப்பு பெரும்பாலும் அகராதிகள் மற்றும் இலக்கண நிரல்களை உருவாக்குவதில் பயன்படுத்தப்படுகிறது. மற்ற முறைகளைப் போலன்றி, மூல மற்றும் இலக்கு மொழிகளின் மொழியியல் பற்றிய கூடுதல் தகவல்களை

உள்ளடக்கியது, இரு மொழிகளின் உருபனியல் மற்றும் தொடரியல் விதிகள் மற்றும் பொருண்மையியல் பகுப்பாய்வுகளைப் பயன்படுத்துகிறது. அடிப்படை அணுகுமுறை உள்ளீட்டு வாக்கியத்தின் கட்டமைப்பை ஒரு பகுப்பான் parser மற்றும் மூல மொழிக்கான ஒரு பகுப்பாய்வி (analyzer), இலக்கு மொழிக்கான உருவாக்கி (generator) மற்றும் உண்மையான மொழிபெயர்ப்பிற்கான பரிமாற்ற அகராதி (transfer lexicon) ஆகியவற்றை வெளியீட்டு வாக்கியத்தின் கட்டமைப்போடு இணைப்பதை உள்ளடக்குகிறது. விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் மிகப்பெரிய வீழ்ச்சி என்னவென்றால், எல்லாவற்றையும் வெளிப்படையாகச் செய்ய வேண்டும்: எழுத்துக்கூட்டுமுறைசார் மாறுபாடு (orthographical variation) மற்றும் தவறான உள்ளீடு ஆகியவற்றைச் சமாளிக்க மூல மொழி பகுப்பாய்வியின் ஒரு பகுதியாக மாற்றப்பட வேண்டும், மேலும் பொருண்மைமயக்கமுள்ள அனைத்து நிகழ்வுகளுக்கும் சொல்சார் தேர்வு விதிகள் (lexical selection rules) எழுதப்பட வேண்டும். புதிய களங்களுடன் பொருத்திக்கொள்வது அவ்வளவு கடினமானதல்ல, ஏனெனில் முக்கிய இலக்கணம் களங்களில் ஒரே மாதிரியாக இருப்பதால், களம்-குறிப்பிட்ட சரிசெய்தல் சொல்சார் தேர்வு சரிசெய்தலுடன் (exical selection adjustment) மட்டுப்படுத்தப்பட்டுள்ளது.

விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு வகைகள்

விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அமைப்புகளில் மூன்று வெவ்வேறு வகைகள் உள்ளன:

- நேரடி ஒழுங்குமுறைகள் (Direct Systems) (அகராதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு) (Dictionary Based Machine Translation) அடிப்படை விதிகளுடன் உள்ளீட்டை வெளியீட்டுடன் பொருத்தும்.
- பரிமாற்ற விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் ((Transfer Based Machine Translation) உருபனியல் மற்றும் தொடரியல் பகுப்பாய்வைப் பயன்படுத்துகின்றன.
- இடைமொழி (Interlingua) விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு சுருக்க அர்த்தத்தைப் பயன்படுத்துகிறது (Koehn 2010; Sergei 1989).

விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் (எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்புExample Based Machine Translation) எதிர் அமைப்புகளாக வகைப்படுத்தலாம்,

அதேசமயம் கலப்பின இயந்திர மொழிபெயர்ப்பு அமைப்புகள் விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பிலிருந்து பெறப்பட்ட பல கொள்கைகளைப் பயன்படுத்துகின்றன.

பரிமாற்ற அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு

பரிமாற்ற அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு என்பது ஒரு வகை இயந்திர மொழிபெயர்ப்பு (MT) ஆகும். இது தற்போது இயந்திர மொழிபெயர்ப்பின் மிகவும் பரவலாகப் பயன்படுத்தப்படும் முறைகளில் ஒன்றாகும். இயந்திர மொழிபெயர்ப்பு எளிமையான நேரடி மாதிரிக்கு மாறாக, இடமாற்ற இயந்திர மொழிபெயர்ப்பு மொழிபெயர்ப்பை மூன்று கட்டங்களாக பிரிக்கிறது: அதன் இலக்கண கட்டமைப்பைத் தீர்மானிக்க மூல மொழி உரையின் பகுப்பாய்வு, இதன் விளைவாக வரும் கட்டமைப்பை இலக்கு மொழியில் உரையை உருவாக்க ஏற்ற கட்டமைப்பிற்கு மாற்றுவது மற்றும் இறுதியாக உருவாக்கம் இந்த உரையின். பரிமாற்ற அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மூல மற்றும் இலக்கு மொழிகளின் அறிவைப் பயன்படுத்த வல்லவை. (Jurafsky & Martin 2009).

இடைமொழிய இயந்திர மொழிபெயர்ப்பு

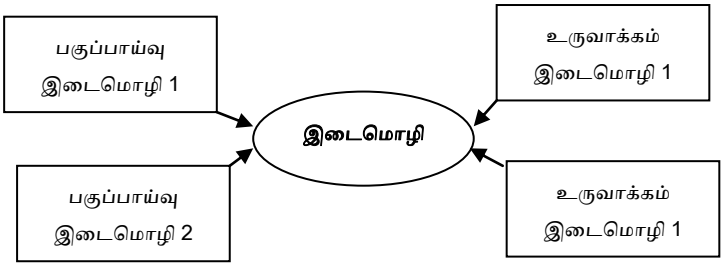
இயந்திர மொழிபெயர்ப்பிற்கான உன்னதமான அணுகுமுறைகளில் ஒன்று இடைமொழிய இயந்திர மொழிபெயர்ப்பு ஆகும். இந்த அணுகுமுறையில், மூல மொழி, அதாவது மொழிபெயர்க்கப்பட வேண்டிய உரை ஒரு இடைமொழியாக (இன்டர்லிங்குவாக/interlingua) மாற்றப்படுகிறது, அதாவது, ஒரு சுருக்க மொழி-சுதந்திரமான உருப்படுத்தம். இலக்கு மொழி பின்னர் இடைமொழியிலிருந்து உருவாக்கப்படுகிறது. விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு முன்னுதாரணத்திற்குள் இடைமொழி அணுகுமுறை, நேரடி அணுகுமுறை மற்றும் பரிமாற்ற அணுகுமுறைக்கு மாற்றாகும்.

நேரடி அணுகுமுறையில், கூடுதல் உருப்படுத்தத்தை கடந்து செல்லாமல் வார்த்தைகள் நேரடியாக மொழிபெயர்க்கப்படுகின்றன. பரிமாற்ற அணுகுமுறையில், மூல மொழி ஒரு அருவத்தன்மையான, குறைந்த மொழிதிடவட்டமான (language-specific) உருப்படுத்தமாக மாற்றப்படுகிறது. மொழி இணகளுக்கு குறிப்பிட்ட மொழியியல் விதிகள் பின்னர் மூல மொழி உருபடுத்தத்தை ஒரு அருவத்தன்மையான இலக்கு மொழி உருப்படுத்தமாக மாற்றுகின்றன, இதிலிருந்து இலக்கு மொழி வாக்கியம் உருவாக்கப்படுகிறது.

இயந்திர மொழிபெயர்ப்பிற்கான இடைமொழி அணுகுமுறையில் நன்மைகளும் தீமைகளும் உள்ளன. நன்மைகள் என்னவென்றால், ஒவ்வொரு மூல மொழியையும் ஒவ்வொரு

இலக்கு மொழியுடனும் தொடர்புபடுத்துவதற்கு குறைவான கூறுகள் தேவைப்படுகின்றன, புதிய மொழியைச் சேர்க்க இது குறைவான கூறுகளை எடுக்கும்; இது மூல மொழியில் உள்ளீட்டின் பெயர்ப்புரைகளை/பொழிப்புரைகளை ஆதரிக்கிறது; இது பகுப்பாய்விகள் மற்றும் உருவாக்கிகள் (ஜெனரேட்டர்கள்) இரண்டையும் ஒருமொழிய ஒழுங்குமுறை உருவாக்குநர்கள் (monolingual system developers) எழுத அனுமதிக்கிறது அனுமதிக்கிறது; மேலும் இது ஒன்றுக்கொன்று மிகவும் வேறுபட்ட மொழிகளைக் கையாளுகிறது (எ.கா. ஆங்கிலம் மற்றும் அரபு) (Abdel Monem et al 2008). வெளிப்படையான குறைபாடு என்னவென்றால், ஒரு இடைமொழியின் (இன்டர்லிங்குவாவின்) வரையறை கடினம் மற்றும் ஒரு பரந்த களத்திற்கு கூட சாத்தியமற்றது. இடைநிலை இயந்திர மொழிபெயர்ப்பிற்கான சிறந்த சூழல் மிகவும் குறிப்பிட்ட களத்தில் பன்மொழி இயந்திர மொழிபெயர்ப்பாகும் (multilingual machine translation).

பின்வரும் படம் இணைக்கும்/இடை மொழியைப் பயன்படுத்தி மொழிபெயர்க்கும் செயல்பாட்டில் பயன்படுத்தப்படும் மொழிகளின் செயல்விளக்கத்தைக் காட்டும்.



இடைநிலை இயந்திர மொழிபெயர்ப்பு என்பது விதி அடிப்படையிலான இயந்திர-மொழிபெயர்ப்பு அணுகுமுறைகளின் ஒரு எடுத்துக்காட்டு. இந்த அணுகுமுறையில், மூல மொழி, அதாவது மொழிபெயர்க்கப்பட வேண்டிய உரை, ஒரு மொழி மொழியாக மாற்றப்படுகிறது, அதாவது எந்தவொரு மொழியிலிருந்தும் சுதந்திரமான "மொழி நடுநிலை" உருப்படுத்தம். இலக்கு மொழி பின்னர் இடைமொழியிலிருந்து உருவாக்கப்படுகிறது. இந்த ஒழுங்கமைப்பின் முக்கிய நன்மைகளில் ஒன்று என்னவென்றால், இலக்கு மொழிகளின் எண்ணிக்கையை அதிகரிப்பதால் இடைமொழி மிகவும் மதிப்புமிக்கதாகிறது. இருப்பினும், வணிக மட்டத்தில் செயல்பட்டு வரும் ஒரே ஒரு மொழி மொழிபெயர்ப்பு முறை KANT ஒழுங்குமுறை (Nyberg and Mitamura, 1992) ஆகும்; இது காட்டர்பில்லர் தொழில்நுட்ப ஆங்கிலத்தை (Caterpillar Technical English (CTE)) பிற மொழிகளில் மொழிபெயர்க்க வடிவமைக்கப்பட்டுள்ளது.

அகராதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு

இயந்திர மொழிபெயர்ப்பு அகராதி உள்ளீடுகளை அடிப்படையாகக் கொண்ட ஒரு முறையைப் பயன்படுத்தலாம்; அதாவது சொற்கள் ஒரு அகராதியைப் போலவே மொழிபெயர்க்கப்படும் - அதாவது வார்த்தை மூலம், பொதுவாக அவற்றுக்கிடையே அதிக தொடர்பு இல்லாமல். அகராதி தேடல்கள் உருபனியல் பகுப்பாய்வு (morphological analysis) அல்லது பகுதியாக்கம் (லெமடிசேஷன்/lemmatisation) இல்லாமல் செய்யப்படலாம். இயந்திர மொழிபெயர்ப்புக்கான இந்த அணுகுமுறை அநேகமாக மிகக் குறைவான நவீனமானது என்றாலும், அகராதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு துணை வாக்கிய (அதாவது, ஒரு முழு வாக்கியம் அல்ல) மட்டத்தில் நீண்ட சொற்றொடர்களின் மொழிபெயர்ப்பிற்கு மிகவும் பொருத்தமானது, எ.கா. சரக்குகள் அல்லது தயாரிப்புகள் மற்றும் சேவைகளின் எளிய பட்டியல்கள் (Uwe Muegge 2006). கையேடு மொழிபெயர்ப்பை விரைவுபடுத்துவதற்கும் இது பயன்படுத்தப்படலாம், அதைச் செயல்படுத்துபவர் இரு மொழிகளிலும் சரளமாக இருப்பதால், தொடரியல் மற்றும் இலக்கணத்தை சரிசெய்யும் திறன் கொண்டவர்.

புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பு

கனடிய ஹான்சார்ட் தரவுத்தொகுதி (Canadian Hansard corpus), கனேடிய பாராளுமன்றத்தின் ஆங்கிலம்-பிரெஞ்சு பதிவு மற்றும் ஐரோப்பிய நாடாளுமன்றத்தின் பதிவான யூரோபார்ல் (EUROPARL) போன்ற இருமொழி உரைத் தரவுத்தொகுதிகளின் அடிப்படையில் புள்ளிவிவர முறைகளைப் பயன்படுத்தி புள்ளிவிவர இயந்திரமொழிபெயர்ப்பு மொழிபெயர்ப்புகளை உருவாக்க முயற்சித்தது. இதுபோன்ற தரவுத்தொகுதிகள் கிடைக்குமிடத்தில், ஒத்த உரைகளை மொழிபெயர்ப்பதில் நல்ல முடிவுகளை அடைய முடியும், ஆனால் இதுபோன்ற தரவுத்தொகுதிகள் பல மொழி இணைகளுக்கு இன்னும் அரிதானவை. முதல் புள்ளிவிவர இயந்திர மொழிபெயர்ப்பு மென்பொருள் ஐபிஎம்மில் (IBM) இருந்து கேண்டிட் (CANDIDE) ஆகும். கூகிள் பல ஆண்டுகளாக சிஸ்ட்ரானைப் (SYSTRAN) பயன்படுத்தியது; ஆனால் அக்டோபர் 2007இல் ஒரு புள்ளியியல்சார் மொழிபெயர்ப்பு முறைக்கு மாறியது (Chitu, Alex 2007). 2005ஆம் ஆண்டில், கூகிள் ஐக்கிய நாடுகள் சபையின் மூலவளத்திலிருந்து சுமார் 200 பில்லியன் சொற்களைப் பயன்படுத்துவதன் மூலம் அதன் உள் மொழிபெயர்ப்பு திறன்களை மேம்படுத்தியது; மொழிபெயர்ப்பு துல்லியம் மேம்படுத்தப்பட்டது ("Google Translator: The Universal Language". Blog.outer-court.com. 2007). கூகிள் மொழிபெயர்ப்பும் (Google Translate)

இதேபோன்ற புள்ளிவிவர மொழிபெயர்ப்பு திட்டங்களும் மனிதர்களால் முன்னர் மொழிபெயர்க்கப்பட்ட நூற்றுக்கணக்கான மில்லியன் ஆவணங்களில் வடிவங்களைக் கண்டறிந்து கண்டுபிடிப்புகளின் அடிப்படையில் புத்திசாலித்தனமான யூகங்களை உருவாக்குவதன் மூலம் செயல்படுகின்றன. பொதுவாக, கொடுக்கப்பட்ட மொழியில் அதிகமான மனித மொழிபெயர்க்கப்பட்ட ஆவணங்கள் கிடைக்கும்போது, மொழிபெயர்ப்பு நல்ல தரத்துடன் இருக்கும் ("Inside Google Translate – Google Translate). METIS II மற்றும் PRESEMT போன்ற புள்ளிவிவர இயந்திர மொழிபெயர்ப்பில் புதிய அணுகுமுறைகள் குறைந்தபட்சத் தரவுத்தொகுதி அளவைப் பயன்படுத்துகின்றன; அதற்கு பதிலாக அமைப்பொழுங்கு புரிதல் மூலம் தொடரியல் கட்டமைப்பைப் பெறுவதில் கவனம் செலுத்துகின்றன. மேலும் வளர்ச்சியுடன், புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பு ஒரு ஒருமொழி உரை தரவுத்தொகுதியிலிருந்து (monolingual text corpus) செயல்பட அனுமதிக்கும் (http://www.mt-archive.info/10/HyTra-2013-Tambouratzis.pdf.). புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பின் (Statisticsl Machine Translation (SMT) மிகப் பெரிய வீழ்ச்சி, இது பெரிய அளவிலான இணை உரைகளைச் சார்ந்தது, உருபனியல் வளமுள்ள மொழிகளுடனான பிரச்சினைகள் (குறிப்பாக இதுபோன்ற மொழிகளில் மொழிபெயர்ப்பது) மற்றும் ஒற்றைப் பிழைகளை (singleton errors) சரிசெய்ய இயலாமை ஆகியவை அடங்கும்.

எடுத்துக்காட்டு அடிப்படையிலானது (Example-based)

எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (ஈபிஎம்டி) அணுகுமுறை 1984 இல் மாகோடோ நாகோவால் முன்மொழியப்பட்டது (Nagao, 1981). எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒப்புமை (analogy) என்ற கருத்தை அடிப்படையாகக் கொண்டது. இந்த அணுகுமுறையில், பயன்படுத்தப்பட்ட தரவுத்தொகுதி ஏற்கனவே மொழிபெயர்க்கப்பட்ட உரைகளைக் கொண்டுள்ளது. மொழிபெயர்க்கப்பட வேண்டிய ஒரு வாக்கியத்தின் அடிப்படையில், இந்த தரவுத்தொகுதியிலிருந்து ஒத்த துணை-வாக்கியசார் கூறுகளைக் (sub-sentential components) கொண்டுள்ள வாக்கியங்கள் தேர்ந்தெடுக்கப்பட்டன ("Kitt.cl.uzh.ch [CL Wiki]" (PDF)). ஒத்த வாக்கியங்கள் அசல் வாக்கியத்தின் துணை-வாக்கியம்சார் கூறுகளை இலக்கு மொழியில் மொழிபெயர்க்கப் பயன்படுத்தப்பட்டன; மேலும் இந்த சொற்றொடர்கள் முழுமையான மொழிபெயர்ப்பை உருவாக்க ஒன்றாக இணைக்கப்பட்டன.

ஒப்புமை மூலம் மொழிபெயர்ப்பு

எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் அடித்தளத்தில் ஒப்புமை மூலம் மொழிபெயர்ப்பின் கருத்து உள்ளது. மனித மொழிபெயர்ப்பின் செயல்முறைக்கு பயன்படுத்தப்படும் போது, மொழிபெயர்ப்பு ஒப்புமை மூலம் நடைபெறுகிறது என்ற எண்ணம் ஆழ்ந்த மொழியியல் பகுப்பாய்வு செய்வதன் மூலம் மக்கள் வாக்கியங்களை மொழிபெயர்க்கும் கருத்தை நிராகரிப்பதாகும். அதற்குப் பதிலாக, ஒரு வாக்கியத்தை முதலில் சில சொற்றொடர்களாக சிதைப்பதன் மூலமும், பின்னர் இந்த சொற்றொடர்களை மொழிபெயர்ப்பதன் மூலமும், இறுதியாக இந்தத் துண்டுகளை ஒரு நீண்ட வாக்கியமாக ஒழுங்காக அமைப்பதன் மூலமும் மக்கள் மொழிபெயர்க்கிறார்கள் என்ற நம்பிக்கையின் அடிப்படையில் இது நிறுவப்பட்டுள்ளது. முந்தைய மொழிபெயர்ப்புகளுக்கு ஒப்புமை மூலம் தொடர்சார் மொழிபெயர்ப்புகள் மொழிபெயர்க்கப்பட்டுள்ளன. ஒப்புமை மூலம் மொழிபெயர்ப்பின் கொள்கை அத்தகைய அமைப்பைப் பயிற்றுவிக்கப் பயன்படும் எடுத்துக்காட்டு மொழிபெயர்ப்புகளின் மூலம் எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்புடன் குறியிடப்பட்டுள்ளது.

இயந்திர இயந்திர மொழிபெயர்ப்பு தொடர்பான பிற அணுகுமுறைகள், புள்ளியியல் இயந்திர மொழிபெயர்ப்பு உட்பட, மொழிபெயர்ப்பின் செயல்முறையை அறிய இருமொழி தரவுத்தொகுதியைப் பயன்படுத்துகின்றன.

வரலாறு

எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு முதன்முதலில் 1984இல் மாகோடோ நாகோவால் பரிந்துரைக்கப்பட்டது. இது குறிப்பாக ஆங்கிலம் மற்றும் ஜப்பானிய போன்ற முற்றிலும் மாறுபட்ட இரண்டு மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பிற்கு ஏற்றது என்று அவர் சுட்டிக்காட்டினார். இந்த நேர்வில், ஒரு வாக்கியத்தை மற்றொரு மொழியில் பல நன்கு கட்டமைக்கப்பட்ட வாக்கியங்களாக மொழிபெயர்க்கலாம், எனவே, விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் ஆழமான மொழியியல் பகுப்பாய்வு பண்புகளைச் செய்வதால் எந்தப் பயனும் இல்லை.

கலப்பின இயந்திர மொழிபெயர்ப்பு

கலப்பின இயந்திர மொழிபெயர்ப்பு (Hybrid machine translation (HMT)) புள்ளிவிவர மற்றும் விதி அடிப்படையிலான மொழிபெயர்ப்பு முறைகளின் பலத்தை மேம்படுத்துகிறது (Adam Boretz 2009). பல இயந்திர மொழிபெயர்ப்பு நிறுவனங்கள் விதிகள் மற்றும் புள்ளிவிவரங்கள்

இரண்டையும் பயன்படுத்தும் கலப்பின அணுகுமுறையைக் கோருகின்றன. அணுகுமுறைகள் பல வழிகளில் வேறுபடுகின்றன:

- புள்ளிவிவரங்களால் பிந்தைய செயலாக்கப்பட்ட விதிகள் (Rules post-processed by statistics): விதிகள் அடிப்படையிலான இயந்திரத்தைப் பயன்படுத்தி மொழிபெயர்ப்புகள் செய்யப்படுகின்றன. விதிகள் இயந்திரத்திலிருந்து வெளியீட்டை சரிசெய்ய / சரிசெய்யும் முயற்சியில் புள்ளிவிவரங்கள் பயன்படுத்தப்படுகின்றன.
- விதிகளால் வழிநடத்தப்பட்ட புள்ளிவிவரங்கள் (Statistics guided by rules): புள்ளிவிவர இயந்திரத்தை சிறப்பாக வழிநடத்தும் முயற்சியில் தரவை முன்கூட்டியே செயலாக்க விதிகள் பயன்படுத்தப்படுகின்றன. இயல்பாக்கம் போன்ற செயல்பாடுகளைச் செய்ய புள்ளிவிவர வெளியீட்டை பிந்தைய செயலாக்க விதிகள் பயன்படுத்தப்படுகின்றன. மொழிபெயர்க்கும்போது இந்த அணுகுமுறை அதிக சக்தி, நெகிழ்வுத்தன்மை மற்றும் கட்டுப்பாட்டைக் கொண்டுள்ளது. மொழிபெயர்ப்புக்கு முந்தைய (எ.கா. உள்ளடக்கத்தின் மார்க்அப் மற்றும் மொழிபெயர்க்க முடியாத சொற்கள்) மற்றும் மொழிபெயர்ப்புக்கு பிந்தைய (எ.கா. பிந்தைய மொழிபெயர்ப்பு திருத்தங்கள் மற்றும் சரிசெய்தல்) இரண்டிலும் உள்ளடக்கம் செயலாக்கப்படும் முறை குறித்த விரிவான கட்டுப்பாட்டை இது வழங்குகிறது.

மிக சமீபத்தில், நரம்புசார் இயந்திர மொழிபெயர்ப்பின் (Neural Machine Translation (NMT)) வருகையுடன், கலப்பின இயந்திர மொழிபெயர்ப்பின் புதிய பதிப்பு உருவாகி வருகிறது, இது விதிகள், புள்ளிவிவர மற்றும் நரம்பியல் இயந்திர மொழிபெயர்ப்பின் நன்மைகளை ஒருங்கிணைக்கிறது. அணுகுமுறை ஒரு விதி வழிகாட்டப்பட்ட பணிப்பாய்வுகளில் முன் மற்றும் பிந்தைய செயலாக்கத்திலிருந்து பயனடைவதோடு நரம்புசார் இயந்திர மொழிபெயர்ப்பு (என்எம்டி) மற்றும் புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பிலிருந்து (எஸ்எம்டி) பயனடைய அனுமதிக்கிறது. எதிர்மறையானது உள்ளார்ந்த சிக்கலானது, இது குறிப்பிட்ட பயன்பாட்டு நிகழ்வுகளுக்கு மட்டுமே அணுகுமுறையை பொருத்தமானதாக ஆக்குகிறது. சிக்கலான பயன்பாட்டு நிகழ்வுகளுக்கான இந்த அணுகுமுறையின் ஆதரவுகளில் ஒன்று ஓம்னிசியன் டெக்னாலஜி (Omniscien Technologies).

நரம்புசார் இயந்திர மொழிபெயர்ப்பு (Neural machine translation (NMT))

நரம்புசார் இயந்திர மொழிபெயர்ப்பு என்பது இயந்திர மொழிபெயர்ப்பிற்கான ஒரு அணுகுமுறையாகும், இது ஒரு செயற்கை நரம்பியல் வலையமைப்பை (artificial neural network) பயன்படுத்துகிறது, இது சொற்களின் வரிசையின் சாத்தியக்கூறுகளை கணிக்க, பொதுவாக முழு வாக்கியங்களையும் ஒரே ஒருங்கிணைந்த மாதிரியில் மாதிரியாகக் கொண்டுள்ளது.

மரபு புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பு (statistical machine translation (SMT)) மாதிரிகளுக்குத் தேவைப்படும் நினைவகத்தின் ஒரு பகுதியை மட்டுமே அவை தேவைப்படுகின்றன. மேலும், வழக்கமான மொழிபெயர்ப்பு முறைகளைப் போலன்றி, மொழிபெயர்ப்புச் செயல்திறனை அதிகரிக்க நரம்பியல் மொழிபெயர்ப்பு மாதிரியின் அனைத்து பகுதிகளும் கூட்டாக (இறுதி முதல் இறுதி வரை) பயிற்சி அளிக்கப்படுகின்றன (Kalchbrenner & Philip 2013).

வரலாறு

ஆழ்ந்த கற்றல் பயன்பாடுகள் (Deep learning applications) 1990களில் பேச்சு புரிதலில் முதலில் தோன்றின. இயந்திர மொழிபெயர்ப்பில் நரம்பியல் வலையமைப்புகளைப் பயன்படுத்துவது பற்றிய முதல் விஞ்ஞான கட்டுரை 2014இல் வெளிவந்தது; அதைத் தொடர்ந்து அடுத்த சில ஆண்டுகளில் நிறைய முன்னேற்றங்கள் ஏற்பட்டன. 2015இல் பொது இயந்திர மொழிபெயர்ப்பு போட்டியில் (ஓபன்எம்டி15/OpenMT'15) நரம்புசார் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை முதல் முறையாகப் பங்கேற்றது. உலக இயந்திரமொழிபெயர்ப்புப் போட்டியும் (WMT'15) முதல் முறையாக நரம்புசார் இயந்திர மொழிபெயர்ப்பு போட்டியாளரைக் கொண்டிருந்த்தது; அடுத்த ஆண்டு அதன் வெற்றியாளர்களிடையே ஏற்கனவே 90% நரம்புசார் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் இருந்தன.

செயல்பாடுகள்

தனித்தனியாக வடிவமைக்கப்பட்ட துணைக் கூறுகளைப் பயன்படுத்தும் சொற்றொடர் அடிப்படையிலான புள்ளிவிவர அணுகுமுறைகளிலிருந்து (phrase-based statistical approaches) நரம்பியல் இயந்திர மொழிபெயர்ப்பு வேறுபடுகிறது (Krzysztof & Krzysztof 2015). நரம்பியல் இயந்திர மொழிபெயர்ப்பு (என்எம்டி) என்பது புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பில் (எஸ்எம்டி) பாரம்பரியமாக செய்யப்பட்டுள்ளதைத் தாண்டிய கடுமையான நடவடிக்கை அல்ல. சொற்கள் மற்றும் உள் நிலைகளுக்கு திசையன் உருப்படுத்தங்களை (vector representations) ["உட்பொதிப்புகள்" ("embeddings"), "தொடர்ச்சியான இடைவெளி உருப்படுத்தங்கள்"

("continuous space representations")] பயன்படுத்துவதே இதன் முக்கிய வேறுபாடு ஆகும். மாதிரிகளின் அமைப்பு சொற்றொடர் அடிப்படையிலான மாதிரிகளை விட எளிமையானது. தனி மொழி மாதிரி, மொழிபெயர்ப்பு மாதிரி மற்றும் மறுவரிசைப்படுத்தல் மாதிரி எதுவும் இல்லை, ஆனால் ஒரே நேரத்தில் ஒரு சொல்லை முன்கணிக்கும் ஒரே வரிசை மாதிரி இருக்கிறது. இருப்பினும், இந்த வரிசை முன்கணிப்பு முழு மூல வாக்கியத்திலும், ஏற்கனவே தயாரிக்கப்பட்ட இலக்கு வரிசையிலும் கட்டுப்படுத்தப்பட்டுள்ளது. நரம்பியல் இயந்திர மொழிபெயர்ப்பு மாதிரிகள் ஆழமான கற்றல் மற்றும் பிரதிநிதித்துவக் கற்றலைப் (representation learning) பயன்படுத்துகின்றன.

வரிசை மாடலிங் என்ற சொல் முதலில் மீண்டும் மீண்டும் நிகழும் நரம்பியல் வலையமைப்பை (recurrent neural network (RNN/ஆர்.என்.என்) பயன்படுத்தி செய்யப்பட்டது. இலக்கு மொழியில் சொற்களைக் கணிக்கப் பயன்படும் டிகோடர் எனப்படும் இரண்டாவது RNNக்கு மூல வாக்கியத்தை குறியாக்க நரம்பியல் நெட்வொர்க்கால் ஒரு குறியாக்கி என அழைக்கப்படும் இருதரப்பு தொடர்ச்சியான நரம்பியல் பிணையம் பயன்படுத்தப்படுகிறது (Dzmitry Bahdanau et al 2014). தொடர்ச்சியான நரம்பியல் நெட்வொர்க்குகள் நீண்ட உள்ளீடுகளை ஒற்றை திசையன் குறியீடாக்குவதில் சிக்கல்களை எதிர்கொள்கின்றன. கவனத்தை ஈர்க்கும் பொறிமுறையால் இதை ஈடுசெய்ய முடியும் (Dzmitry Bahdanau et al 2014); இது வெளியீட்டின் ஒவ்வொரு வார்த்தையையும் உருவாக்கும் போது டிகோடரை உள்ளீட்டின் வெவ்வேறு பகுதிகளில் கவனம் செலுத்த அனுமதிக்கிறது. மேலதிக மொழிபெயர்ப்பு மற்றும் மொழிபெயர்ப்புக்கு வழிவகுக்கும் கடந்த சீரமைப்பு தகவல்களை புறக்கணிப்பது போன்ற கவனத்தை ஈர்க்கும் வழிமுறைகளில் சிக்கல்களை எதிர்கொள்ளும் மேலும் பாதுகாப்பு மாதிரிகள் உள்ளன (Tu et al 2014).

கன்வல்யூஷனல் நரம்பியல் வலையமைப்பு (கன்வல்யூஷனல் நியூரல் நெட்வொர்க்குகள்/ Convolutional Neural Networks (கன்வெனெட்ஸ்/Convnets)) நீண்ட தொடர்ச்சியான வரிசைகளுக்கு கொள்கை அடிப்படையில் ஓரளவு சிறந்தவை, ஆனால் ஆரம்பத்தில் பல பலவீனங்கள் காரணமாக அவை பயன்படுத்தப்படவில்லை. இவை 2017ஆம் ஆண்டில் "கவனத்தை ஈர்க்கும் வழிமுறைகளை" ("attention mechanisms") பயன்படுத்தி வெற்றிகரமாக ஈடுசெய்யப்பட்டன (Devin (2017). கவனத்தை அடிப்படையாகக் கொண்ட மாதிரி, மின்மாற்றி

கட்டமைப்பு (transformer architecture) (Vaswani et al 2017), பல மொழி இணைகளுக்கு ஆதிக்கம் செலுத்துகிறது.

6.5.3. பன்மொழித் தகவல் பெறும் ஒழுங்குமுறை (Multilingual information access systems)

உலகெங்கிலும் தகவல்களைத் தொடர்புகொள்வதற்கும் பரப்புவதற்கும் உலகளாவிய வலை மற்றும் இணையத்தின் பயன்பாடு மற்றும் பிரபலத்தின் விரைவான விரிவாக்கம் என்பது மின்னணு முறையில் அணுகக்கூடிய தகவல்கள் இப்போது அதிகரித்து வரும் மொழிகளில் இப்போது கிடைக்கின்றன என்பதாகும். முதல் வலைத்தளங்கள் ஆங்கிலத்தில் தகவல்களை வழங்குவதற்காக முற்றிலும் அர்ப்பணிக்கப்பட்டன, மேலும் ஆங்கிலம் பேசும் சமூகத்தின் (எ.கா. லைகோஸ்/Lycos, ஆல்டாவிஸ்தா/AltaVista, யாகூ!/Yahoo!) தேவைகளைப் பூர்த்தி செய்வதற்காக முதல் தேடல் சேவைகள் (சுமார் 1995 இல்) செயல்படுத்தப்பட்டன. இந்த சேவைகளைப் பயன்படுத்துபவர்கள் முக்கியமாக கல்விப் பின்னணியைக் கொண்டிருந்தனர் மற்றும் ஆங்கிலத்தில் அர்த்தமுள்ள வினவல்களை உருவாக்குவதற்கும் மீட்டெடுக்கப்பட்ட ஆவணங்களைப் புரிந்து கொள்வதற்கும் போதுமான ஆங்கில மொழித் திறன்களைக் கொண்டிருந்தனர். இருப்பினும், கடந்த சில ஆண்டுகளில், நிலைமை மிகவும் மாறிவிட்டது. வலை என்பது கல்வி நோக்கங்களுக்காக மட்டுமல்ல, வர்த்தகம், பொழுதுபோக்கு, செய்தி, சுற்றுலா, வங்கி, நிதி போன்றவற்றுக்காகவும் பயன்படுத்தப்படுகிறது. தகவல் பெருகிய முறையில் நபர் அல்லது அமைப்பின் சொந்த மொழியில் தகவல் வெளியிடப்படுகிறது மற்றும் பயனரின் சொந்த மொழியில் தகவல் தேடப்படுகிறது. பல நாடுகள் இப்போது தங்கள் சொந்த வலை தேடல் சேவைகளைக் கொண்டுள்ளன. சராசரி வலை பயனரின் சுயவிவரத்தை, அவர்களின் தேவைகள் மற்றும் மொழித் திறன்களை வழங்குவது மிகவும் கடினம். ஆங்கிலம் அல்லாத வலைப்பக்கங்களின் எண்ணிக்கையைப் போலவே ஆங்கிலத்தில் புலமை இல்லாத இணைய பயனர்களின் எண்ணிக்கையும் வேகமாக வளர்ந்து வருவதாக சமீபத்திய ஆய்வுகள் காட்டுகின்றன. கணிப்பு என்னவென்றால், 2005 ஆம் ஆண்டளவில், 78% இணைய பயனர்கள் ஆங்கிலம் அல்லாதவர்கள், அதே சமயம் “49%” வலை உள்ளடக்கம் ஆங்கிலத்தில் இருக்கும் (Oard : http://www.clis.umd.edu/dlrg/filter/papers/).

இணையம் பல மொழிகளில் ஆவண சேகரிப்பிற்கான மிகப்பெரிய களஞ்சியமாக இருந்தாலும், அது மட்டும்தான் என்பதில்லை இல்லை. பல பெரிய சர்வதேச பொது மற்றும் தனியார் அமைப்புகளின் அகங்களும் பெருகிய முறையில் பன்மொழி தகவல்களைக்

கொண்டிருக்கின்றன, ஏனெனில் ஆர்வங்களும் செயல்பாடுகளும் தேசிய எல்லைகளை மீறுகின்றன, மேலும் ஒரு பொதுவான மொழியைப் பயன்படுத்துவது எப்போதும் ஏற்றுக்கொள்ளப்படாது. பல மொழிகளில் தகவல்களை அணுகுவதற்கான பிற பகுதிகள் பெருகிய முறையில் முக்கியத்துவம் பெறுகின்றன: உதாரணமாக: டிஜிட்டல் நூலகங்கள் மற்றும் பொருள் நுழைவாயில்கள் (subject gateways), உலகம் முழுவதும் பள்ளிகள் மற்றும் பள்ளி நெட்வொர்க்குகள், சட்டங்கள் மற்றும் ஒழுங்குமுறைகளின் பன்மொழி சேகரிப்புகள், கலாச்சார பாரம்பரியத்தின் பன்மொழி சேகரிப்புகள்.

இவை அனைத்தும் ஒரு பன்மொழி ஆவண சேகரிப்பை வினவும் பணியை ஒரு பயனர் எதிர்கொள்ளும் சூழ்நிலைகள் பெருகிய முறையில் பொதுவானதாகி வருகின்றன என்பதாகும். பல பயனர்களுக்கு சில வெளிநாட்டு மொழி அறிவு உள்ளது, ஆனால் அவர்களின் தகவல் தேவைகளைச் சரியான முறையில் வெளிப்படுத்த வினவல்களை வகுக்க அவர்களின் திறமை போதுமானதாக இருக்காது. அத்தகைய பயனர்கள் தங்கள் கேள்விகளை தங்கள் சொந்த மொழியில் உள்ளிட முடிந்தால் அவர்கள் பெரிதும் பயனடைவார்கள், ஏனென்றால் அவை மொழிபெயர்க்கப்படாவிட்டாலும் தொடர்புடைய ஆவணங்களை ஆராய முடிகிறது. ஒருமொழி பயனர்கள், மறுபுறம், மொழிபெயர்ப்புக் கருவியைப் பயன்படுத்தி, அவர்களின் தேடல் முடிவுகளை இரண்டாவது மொழியில் புரிந்துகொள்ள உதவலாம். சுருக்கமாக, தகவல்களை எந்த மொழியில் சேமித்து வைத்தாலும், மொழி எல்லைகளில் தொடர்புடைய தகவல்களைக் கண்டுபிடித்து மீட்டெடுப்பதற்கான திறமையான வழிகளைக் கண்டுபிடிப்பதற்கான தேவை அதிகரித்து வருகிறது. பன்மொழி தகவல் அணுகல் என்பது இதுதான்.

பன்மொழி தகவல் அணுகல் (Multilingual Information Access (MLIA)) என்ற வார்த்தையை அதன் பரந்த பொருளில் பயன்படுத்த இயலும். எந்தவொரு மொழியிலும் சேகரிப்பிலிருந்து தகவல்களை அணுகல், வினவல் மற்றும் மீட்டெடுப்பதில் உள்ள சிக்கலை பன்மொழி தகவல் அணுகல் எந்தவொரு குறிப்பிட்ட அளவிலும் நேர்கொள்கிறது மற்றும் பன்மொழி தகவல்களின் ஒட்டுமொத்த நிர்வாகத்தை உள்ளடக்கிய அனைத்து சிக்கல்களையும் உள்ளடக்கியது, அதாவது எழுத்து குறியீட்டு முறை, மொழி அடையாளம் காணல், பல மொழிகளில் சேகரிப்புகளின் அட்டவணைப்படுத்தல் போன்றவை. குறுக்கு மொழி தகவல் மீட்டெடுப்பு (Cross-Language or Cross-Lingual Information Retrieval (CLIR) தொழில்நுட்பங்கள் பிற மொழிகளில் தொடர்புடைய ஆவணங்களை மீட்டெடுப்பதற்காக ஒரு மொழியில் ஒரு

பன்மொழி சேகரிப்பை வினவுவதில் அக்கறை கொண்டுள்ளன, மேலும் அத்தகைய ஆவணங்களை வடிகட்டுதல், தேர்ந்தெடுப்பது மற்றும் தரவரிசைப்படுத்தும் பணியை நிவர்த்தி செய்கின்றன. வினவலின் முடிவுகளின் விளக்கக்காட்சி (presentation), சுருக்கம் (summarization) மற்றும் மொழிபெயர்ப்பு தொடர்பான பிரச்சினைகள் புறமானவை ஆனால் வலுவாக தொடர்புடையவை. இந்த கலைச்சொற்களைப் பயன்படுத்துவதில் பன்மொழி தகவல் அணுகல் என்பது குறுக்கு மொழி தகவல் மீட்டெடுப்பு என்பதை உட்படுத்தும். இரு கலைச்சொற்களும் பன்மொழி உரை மீட்டெடுப்பு முன்னுதாரணத்தைக் குறிக்க இன்னும் அடிக்கடி பயன்படுத்தப்படுகின்றன என்றாலும், அவை படம், வீடியோ, பேச்சு போன்ற பிற ஊடகங்களையும் உள்ளடக்குகின்றன.

6.5.4. மொழிகடந்த தகவல் மீட்பு ஒழுங்குமுறை (Cross language information retrieval systems)

குறுக்கு மொழி தகவல் மீட்டெடுப்பு (Cross language information retrieval (CLIR/ சி.எல்.ஐ.ஆர்) என்பது பயனரின் வினவலின் மொழியிலிருந்து வேறுபட்ட மொழியில் எழுதப்பட்ட தகவல்களை மீட்டெடுப்பதைக் கையாளும் தகவல் மீட்டெடுப்பின் துணைத் துறையாகும் (Wang, & Oard 2012) "குறுக்கு மொழி தகவல் மீட்டெடுப்பு" ("cross-language information retrieval") என்ற கலைச்சொல் பல ஒருபொருள்பன்மொழிகளைக் கொண்டுள்ளது, அவற்றில் பின்வருபவை பெரும்பாலும் நிகழ்கின்றன: குறுக்கு மொழி தகவல் மீட்டெடுப்பு (cross-lingual information retrieval), மொழிபெயர்ப்பு தகவல் மீட்டெடுப்பு (translingual information retrieval), பன்மொழி தகவல் மீட்டெடுப்பு (multilingual information retrieval). "பன்மொழி தகவல் மீட்டெடுப்பு" என்ற சொல் பொதுவாக பன்மொழி சேகரிப்புகளை மீட்டெடுப்பதற்கான தொழில்நுட்பம் மற்றும் ஒரு மொழியில் உள்ள பொருளை மற்றொரு மொழியில் கையாள உருவாக்கப்பட்ட தொழில்நுட்பம் ஆகியவற்றைக் குறிக்கிறது. பன்மொழி தகவல் மீட்டெடுப்பு (Multilingual Information Retrieval (MLIR/ எம்.எல்.ஐ.ஆர்.) என்ற சொல் பல்வேறு மொழிகளில் உள்ள தகவல்களுக்கான வினவல்களை ஏற்றுக் கொள்ளும் அமைப்புகளின் ஆய்வு மற்றும் பல்வேறு மொழிகளின் பொருள்களை (உரை மற்றும் பிற ஊடகங்கள்) பயனரின் மொழியில் மொழிபெயர்க்கப்பட்டுள்ளது. குறுக்கு மொழி தகவல் மீட்டெடுப்பு பயனர்கள் தங்கள் மொழியின் தேவையை ஒரு மொழியில் வகுக்கும் பயன்பாட்டு முறைக்கு மேலும் குறிப்பாக கணினி தொடர்புடைய ஆவணங்களை மற்றொரு மொழியில் மீட்டெடுக்கிறது. அவ்வாறு செய்ய,

பெரும்பாலான சி.எல்.ஐ.ஆர் அமைப்புகள் பல்வேறு மொழிபெயர்ப்பு நுட்பங்களைப் பயன்படுத்துகின்றன. குறுக்கு மொழி தகவல் மீட்டெடுப்பு நுட்பங்களை வெவ்வேறு மொழிபெயர்ப்பு வளங்களின் அடிப்படையில் வெவ்வேறு வகைகளாக வகைப்படுத்தலாம் (Thai).

- அகராதி அடிப்படையிலான குறுக்கு மொழி தகவல் மீட்டெடுப்பு நுட்பங்கள்
- இணைத் தரவுத்தொகுதிகள் அடிப்படையிலான குறுக்கு மொழி தகவல் மீட்டெடுப்பு நுட்பங்கள்
- ஒப்பிடக்கூடிய தரவுத்தொகுதிகள் அடிப்படையிலான குறுக்கு மொழி தகவல் மீட்டெடுப்பு நுட்பங்கள்
- இயந்திர மொழிபெயர்ப்பாளர் அடிப்படையிலான குறுக்கு மொழி தகவல் மீட்டெடுப்பு நுட்பங்கள்

குறுக்கு மொழி தகவல் மீட்டெடுப்பு ஒழுங்குமுறைகள் மிகவும் மேம்பட்டுள்ளன, இன்று மிகவும் துல்லியமான பல மொழி மற்றும் குறுக்கு மொழி தற்காலிக தகவல் மீட்டெடுப்பு அமைப்புகள் ஒருமொழி அமைப்புகளைப் போலவே கிட்டத்தட்ட பயனுள்ளதாக இருக்கும் (Oard, 2011). மீடியா கண்காணிப்பு (media monitoring), தகவல் வடிகட்டுதல் (information filtering) மற்றும் ரூட்டிங் (routing), சென்டிமென்ட் பகுப்பாய்வு (sentiment analysis) மற்றும் தகவல் பிரித்தெடுத்தல் (information extraction) போன்ற, பிற தொடர்புடைய தகவல் அணுகல் செயல்பாடுகளுக்கு அதிக அதிநவீன மாதிரிகள் தேவைப்படுகின்றன, மேலும் பொதுவாக ஆர்வமுள்ள தகவல் பொருட்களின் செயலாக்கம் மற்றும் பகுப்பாய்வு தேவைப்படுகிறது. அந்த செயலாக்கத்தின் பெரும்பகுதி அது பயன்படுத்தப்பட்ட இலக்கு மொழிகளின் தனித்தன்மைகளை அறிந்திருக்க வேண்டும்.

பெரும்பாலும், மனித மொழியின் மாறுபாட்டின் பல்வேறு வழிமுறைகள் தகவல் மீட்டெடுப்பு முறைகளுக்கு பரப்பெல்லை (கவரேஜ்) சவால்களைத் தருகின்றன: ஒரு தொகுப்பில் உள்ள உரைகள் ஆர்வமுள்ள ஒரு தலைப்பைக் கருத்தில் கொள்ளலாம், ஆனால் பயனரால் வழங்கப்பட்ட தகவல் தேவையின் வெளிப்பாட்டுடன் பொருந்தாத சொற்கள் அல்லது வெளிப்பாடுகளைப் பயன்படுத்தலாம். ஒரு ஒரு-மொழி நேர்வில் கூட இது உண்மையாக இருக்கலாம், ஆனால் இது குறுக்கு மொழி தகவல் மீட்டெடுப்பில் இது குறிப்பாக உண்மை ஆகும்; பயனர்கள் இலக்கு மொழியை ஓரளவிற்கு மட்டுமே அறிந்திருக்கலாம். இலக்கு மொழியில் குறைவான முதல் மிதமான திறன் கொண்ட பயனர்களுக்கு குறுக்கு மொழி தகவல் மீட்டெடுப்பு

தொழில்நுட்பத்தின் நன்மைகள் இலக்குமொழியில் சரளமாக இருப்பவர்களைக் காட்டிலும் அதிகமாக இருப்பதாகக் கண்டறியப்பட்டுள்ளது (Airio 2008). குறுக்கு மொழி தகவல் மீட்டெடுப்பு சேவைகளுக்கான குறிப்பிட்ட தொழில்நுட்பங்களில் கலவை சொற்களைக் கையாள்வதற்கான உருபனியல் பகுப்பாய்வு, கூட்டுச்சொற்களைப் பிரித்தல் மற்றும் ஒரு வினவலை ஒரு மொழியிலிருந்து இன்னொரு மொழியில் மொழிபெயர்க்கும் மொழிபெயர்ப்பு வழிமுறைகள் ஆகியவை அடங்கும்.

குறுக்கு மொழி தகவல் மீட்டெடுப்பு பற்றிய முதல் பட்டறை சூரிச்சில் எஸ்.ஐ.ஜி.ஆர் -96 மாநாட்டின் போது நடைபெற்றது. குறுக்கு மொழி மதிப்பீட்டு மன்றத்தின் (Cross Language Evaluation Forum (CLEF)) கூட்டங்களில் 2000 முதல் ஆண்டுதோறும் பட்டறைகள் நடத்தப்படுகின்றன. ஆராய்ச்சியாளர்கள் வருடாந்திர உரை மீட்டெடுப்பு மாநாட்டில் (Text Retrieval Conference (TREC)) வெவ்வேறு அமைப்புகள் மற்றும் தகவல்களை மீட்டெடுக்கும் முறைகள் குறித்து தங்கள் கண்டுபிடிப்புகளைப் பற்றி விவாதிக்கிறார்கள், மேலும் இந்த மாநாடு குறுக்கு மொழி தகவல் மீட்டெடுப்பு துணைத் துறைக்கான குறிப்பு புள்ளியாக செயல்பட்டது (Olvera-Lobo). கூகிள் தேடலில் குறுக்கு மொழி தேடல் அம்சம் இருந்தது, அது 2013 இல் அகற்றப்பட்டது.

6.6. மனித-இயந்திர இடைமுக ஒழுங்குமுறைகளுக்கு மூலவளமாக பெருந்தரவு

இது பின்வருவனவற்றை உள்ளடக்கும்:

1. ஒலியால் வருடி எழுத்துக்களைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்
2. குரலைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்
3. உரையிலிருந்து பேச்சு ஒழுங்குமுறைகள்
4. வலை (இணையத்தளம்) அடிப்படையிலான கற்றல் ஒழுங்குமுறைகள்
5. கேள்வி-பதில் ஒழுங்குமுறைகள்
6. கணினி உதவியுடன் கட்டளைகள்
7. கணினியின் உதயுடன் மொழிக் கல்வி
8. உரை உருவாக்கம்

6.6.1. ஒலியால் வருடி எழுத்துக்களைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்

. ஒலியால் வருடி எழுத்துக்களைப் புரிந்துக்கொள் அல்லது ஆப்டிகல் கேரக்டர் ரீடர் (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன்/ Optical character recognition or optical character reader

ஒளிமூலம் எழுத்துப் புரிவான் (OCR/ஓ.சி.ஆர்) என்பது தட்டச்சு செய்யப்பட்ட, கையால் எழுதப்பட்ட அல்லது அச்சிடப்பட்ட உரையின் படங்களை இயந்திர குறியீட்டு உரையாக மாற்றும் மின்னணு அல்லது இயந்திர மாற்றமாகும், இது ஸ்கேன் செய்யப்பட்ட ஆவணத்திலிருந்து, ஆவணத்தின் புகைப்படம், காட்சி-புகைப்படம் (எடுத்துக்காட்டாக குறியீடுகளிலான உரை, ஒரு நிலப்பரப்பு புகைப்படத்தில் உள்ள விளம்பர பலகை) அல்லது ஒரு படத்தின் எழுத்தப்பட்ட துணைதலைப்பு உரை (எடுத்துக்காட்டாக: ஒரு தொலைக்காட்சி ஒளிபரப்பிலிருந்து). ("OCR Introduction". Dataid.com).

பாஸ்போர்ட் ஆவணங்கள், விலைப்பட்டியல், வங்கி அறிக்கைகள், கணினிமயமாக்கப்பட்ட ரசீதுகள், வணிக அட்டைகள், அஞ்சல், நிலையான-தரவின் அச்சுநகல்கள் அல்லது பொருத்தமான ஆவணங்கள் எதுவாக இருந்தாலும் - அச்சிடப்பட்ட காகிதத் தரவுப் பதிவுகளிலிருந்து தரவு நுழைவு வடிவமாக பரவலாகப் பயன்படுத்தப்படுகிறது; இது அச்சிடப்பட்ட உரைகளை டிஜிட்டல் மயமாக்குவதற்கான பொதுவான முறையாகும்; டிஜிட்டல் மயமாக்கப்பட்ட உரைகளை மின்னணு முறையில் திருத்த இயலும், தேட இயலும், மேலும் சுருக்கமாக சேமிக்க இயலும், ஆன்லைனில் இயலும்; மேலும் புலனறிவு கணினியாக்கம், இயந்திர மொழிபெயர்ப்பு, (பிரித்தெடுக்கப்பட்ட) உரையிலிருந்து பேச்சு, முக்கிய தரவு மற்றும் உரையை ஆழ்ந்தெடுத்தல் போன்ற இயந்திர செயல்முறைகளிலும் ஓசிஆர் பயன்படுத்தப்படுகின்றது. ஒளிமூலம் எழுத்துப் புரிவான் என்பது மாதிரி புரிதல், செயற்கை நுண்ணறிவு மற்றும் கணினிப் பார்வை ஆகியவற்றில் ஆராய்ச்சித் துறையாகும்.

ஆரம்பகால பதிப்புகள் ஒவ்வொரு எழுத்துக்களின் படங்களுடன் பயிற்சியளிக்கப்பட வேண்டும், மேலும் ஒரு நேரத்தில் ஒரு எழுத்துருவில் வேலை செய்ய வேண்டும். பெரும்பாலான எழுத்துருக்களுக்கு அதிக அளவிலான புரிதல் துல்லியத்தை உருவாக்கும் திறன் கொண்ட மேம்பட்ட அமைப்புகள் இப்போது பொதுவானவை, மேலும் பலவிதமான டிஜிட்டல் படக் கோப்பு வடிவ உள்ளீடுகளுக்கான ஆதரவுடன். சில அமைப்புகள் வடிவமைக்கப்பட்ட வெளியீட்டை மீள் உருவாக்கம் செய்யும் திறன் கொண்டவை, அவை படங்கள், நெடுவரிசைகள் மற்றும் பிற உரை அல்லாத கூறுகள் உள்ளிட்ட அசல் பக்கத்தை நெருக்கமாக மதிப்பிடுகின்றன.

வரலாறு

ஆரம்பகால ஒளிமூலம் எழுத்துப் புரிவான் தந்தி சம்பந்தப்பட்ட தொழில்நுட்பங்கள் மற்றும் பார்வையற்றோருக்கான வாசிப்புச் சாதனங்களை உருவாக்குதல் ஆகியவற்றைக் கண்டறியலாம்

(Schantz, 1982). 1914ஆம் ஆண்டில், இமானுவேல் கோல்ட்பர்க் (Emanuel Goldberg) ஒரு இயந்திரத்தை உருவாக்கி, எழுத்துக்களைப் படித்து அவற்றை நிலையான தந்தி குறியீடாக மாற்றினார் (Dhavale, 2017). ஒரே நேரத்தில், எட்மண்ட் ஃபோர்னியர் டி ஆல்பே (Edmund Fournier d'Albe) ஒரு கையடக்க ஸ்கேனரான ஆப்டோஃபோனை (Optophone) உருவாக்கினார், இது அச்சிடப்பட்ட பக்கத்தின் குறுக்கே நகரும்போது, குறிப்பிட்ட எழுத்துக்கள் அல்லது எழுத்துக்களுக்கு ஒத்த டோன்களை உருவாக்கியது (d'Albe, 1914).

1920களின் பிற்பகுதியிலும் 1930களில் இமானுவேல் கோல்ட்பர்க் (Emanuel Goldberg) ஒரு ஆப்டிகல் குறியீடு புரிந்துகொள்ளும் முறையைப் பயன்படுத்தி மைக்ரோஃபில்ம் காப்பகங்களைத் தேடுவதற்காக "புள்ளிவிவர இயந்திரம்" என்று அழைக்கப்பட்ட ஒன்றை உருவாக்கினார். 1931ஆம் ஆண்டில் அவரது கண்டுபிடிப்புக்காக அமெரிக்காவின் காப்புரிமை எண் 1,838,389 வழங்கப்பட்டது. காப்புரிமையை ஐ.பி.எம்.-ஆல் பெறப்பட்டது.

பார்வையற்ற மற்றும் பார்வை குறையுள்ள பயனர்கள்

1974ஆம் ஆண்டில், ரே குர்ஸ்வீல் (Ray Kurzweil) குர்ஸ்வீல் கம்ப்யூட்டர் தயாரிப்புகள், இன்க் (Kurzweil Computer Products, Inc.) என்ற நிறுவனத்தைத் தொடங்கினார் மற்றும் ஓம்னி-எழுத்துரு ஒளிமூலம் எழுத்துப் புரிவானின் (omni-font OCR) தொடர்ச்சியான உருவாக்கத்தைத் தொடர்ந்தார், இது எந்தவொரு எழுத்துருவிலும் அச்சிடப்பட்ட உரையைப் புரிய இயலும் (குர்ஸ்வீல் பெரும்பாலும் ஓம்னி-எழுத்துரு ஒளிமூலம் எழுத்துப் புரிவானைக் கண்டுபிடித்த பெருமைக்கு உள்ளாகிறார்; அது 1960களின் பிற்பகுதியிலும் 1970களின் பிற்பகுதியிலும் கம்ப்யூஸ்கான் (CompuScan) உள்ளிட்ட நிறுவனங்களின் பயன்பாட்டில் இருந்தது). பார்வையற்றோருக்கு ஒரு வாசிப்பு இயந்திரத்தை உருவாக்குவதே இந்தத் தொழில்நுட்பத்தின் சிறந்த பயன்பாடாகும் என்று குர்ஸ்வீல் முடிவு செய்தார், இது பார்வையற்றவர்களுக்கு கணினி வாசிக்கும் உரையைச் சத்தமாக வைத்திருக்க அனுமதிக்கும் ("The History of OCR"). இந்த சாதனத்திற்கு இரண்டு செயல்படுத்தும் தொழில்நுட்பங்களின் கண்டுபிடிப்பு தேவை - சிசிடி பிளாட்பெட் ஸ்கேனர் மற்றும் உரையிலிருந்து பேச்சு உருவாக்கம். ஜனவரி 13, 1976 அன்று, குர்ஸ்வீல் மற்றும் பார்வையற்றோரின் தேசிய கூட்டமைப்பின் தலைவர்கள் தலைமையில் பரவலாக அறிவிக்கப்பட்ட செய்தி மாநாட்டின் போது வெற்றிகரமான முடிக்கப்பட்ட தயாரிப்பு வெளியிடப்பட்டது. 1978ஆம் ஆண்டில், குர்ஸ்வீல் கணினி தயாரிப்புகள் ஒளிவழி எழுத்து புரிதல் கணினி நிரல் (optical character recognition computer program) வணிகப் பதிப்பை விற்பனை

செய்யத் தொடங்கின. லெக்சிஸ்நெக்ஸிஸ் (LexisNexis) முதல் வாடிக்கையாளர்களில் ஒன்றாக இருந்தது; மேலும் சட்ட ஆவணங்கள் மற்றும் செய்தி ஆவணங்களை அதன் புதிய ஆன்லைன் தரவுத்தளங்களில் பதிவேற்றுவதற்கான திட்டத்தை வாங்கினார். இரண்டு ஆண்டுகளுக்குப் பிறகு, குர்ஸ்வீல் தனது நிறுவனத்தை ஜெராக்ஸுக்கு விற்றார்; ஜெராக்ஸ் காகிதத்திலிருந்து கணினி உரை மாற்றத்தை மேலும் வணிகமயமாக்குவதில் ஆர்வம் கொண்டிருந்தது. ஜெராக்ஸ் இறுதியில் அதை ஸ்கேன்சாஃப்ட் (Scansoft) என்று அழைத்தது; இது நுவான்ஸ் கம்யூனிகேஷனுடன் (Nuance Communications) இணைந்தது.

2000களில், ஒளிமூலம் எழுத்துப் புரிவான் ஆன்லைனில் ஒரு சேவையாக (WebOCR), கிளவுட் கம்ப்யூட்டிங் சூழலில், மற்றும் ஸ்மார்ட்போனில் (smartphone) வெளிநாட்டு மொழிக் குறியீகளின் (foreign-language signs) நிகழ்நேர மொழிபெயர்ப்பு போன்ற மொபைல் பயன்பாடுகளில் கிடைக்கச் செய்யப்படது. ஸ்மார்ட் போன்கள் மற்றும் ஸ்மார்ட் கிளாஸின் (Smartglasses) வருகையுடன், கருவியின் கேமராவைப் பயன்படுத்தி பெறப்பட்ட உரையை பிரித்தெடுக்கும் இணையம் இணைக்கப்பட்ட மொபைல் கருவிப் பயன்பாடுகளில் ஓசிஆர்-ஐப் பயன்படுத்தலாம். இயக்க முறைமையில் (operating system) கட்டமைக்கப்பட்ட ஓசிஆர் செயல்பாடு இல்லாத இந்த கருவிகள் பொதுவாக ஒளிமூலம் எழுத்துப் புரிவான் எபிஐ-ஐப் (OCR API) பயன்படுத்தி கருவியால் பெறப்பட்ட மற்றும் வழங்கப்பட்ட படக் கோப்பிலிருந்து உரையைப் பிரித்தெடுக்கும். அசல் படத்தில் கண்டறியப்பட்ட உரையின் இருப்பிடம் பற்றிய தகவலுடன், மேலும் செயலாக்கத்திற்காக (உரையிலிருந்து பேச்சு போன்றவை) அல்லது காட்சிக்கு கருவிப் பயன்பாட்டிற்கு OCR API திரும்பப் பெறுகிறது.

லத்தீன், சிரிலிக், அரபு, ஹீப்ரு, இந்திக், பெங்காலி (பங்களா), தேவநாகரி, தமிழ், சீன, ஜப்பானிய மற்றும் கொரிய எழுத்துக்கள் உள்ளிட்ட பொதுவான வணிக அமைப்புகளுக்கு பல்வேறு வணிக மற்றும் திறந்த மூல ஒளிமூலம் எழுத்துப் புரிதல் ஒழுங்குமுறைகள் கிடைக்கின்றன.

வகைகள்

கீழ் வருவன ஒளிவழி எழுத்துப் புரிதலின் (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன் (OCR)) வகைகளாகப் பட்டியலிடப்பட்டுள்ளன.

- ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன் (Optical character recognition (OCR)) - தட்டச்சு செய்யப்பட்ட உரை, ஒரு நேரத்தில் ஒரு கிளிஃப் அல்லது எழுத்தை குறிவைக்கிறது.

- ஒளிவழி சொல் புரிதல் Optical word recognition - தட்டச்சு செய்யப்பட்ட உரையை குறிவைக்கிறது, ஒரு நேரத்தில் ஒரு சொல் (இடத்தை சொல் பிரிப்பியாகப் பயன்படுத்தும் மொழிகளுக்கு). (பொதுவாக "OCR" என்று அழைக்கப்படுகிறது.)
- நுண்ணறிவு எழுத்துப் புரிதல் (Intelligent word recognition (IWR)) - கையால் எழுதப்பட்ட அச்சுஎழுத்து (handwritten printscript) அல்லது கர்சீவ் உரையை (cursive text) ஒரு நேரத்தில் ஒரு கிளிஃப் அல்லது எழுத்தை குறிவைக்கிறது; பொதுவாக இயந்திர கற்றல் இதில் அடங்கும்.
- நுண்ணறிவு சொல் புரிதல் (IWR) - கையால் எழுதப்பட்ட அச்சுஎழுத்து அல்லது கர்சீவ் உரையையும் குறிவைக்கிறது, ஒரு நேரத்தில் ஒரு சொல். கர்சீவ் எழுத்துவடிவில் கிளிஃப்கள் கூறிடப்படாத மொழிகளுக்கு இது மிகவும் பயனுள்ளதாக இருக்கும்.

ஓசிஆர்பொதுவாக ஒரு "ஆஃப்லைன்" செயல்முறையாகும், இது ஒரு நிலையான ஆவணத்தை பகுப்பாய்வு செய்கிறது. ஆன்லைன் ஒளிமூலம் எழுத்துப் புரிவான் எபிஐ (OCR API) சேவையை வழங்கும் கிளவுட் அடிப்படையிலான சேவைகள் உள்ளன. கையெழுத்து இயக்கப் பகுப்பாய்வு (Handwriting movement analysis), கையெழுத்துப் புரிதலுக்கான (handwriting recognition) உள்ளீடாகப் பயன்படுத்தப்படலாம் (Tappert et al 1990). கிளிஃப்கள் மற்றும் சொற்களின் வடிவங்களைப் பயன்படுத்துவதற்குப் பதிலாக, இந்த நுட்பம், கூறுகள் வரையப்பட்ட வரிசை, திசை, மற்றும் பேனாவை கீழே வைத்து தூக்கும் அமைபொழுங்கு போன்ற இயக்கங்களை ஈட்ட இயலும். இந்தக் கூடுதல் தகவல் இறுதிக்கு-இறுதி செயல்முறையை மிகவும் துல்லியமாக்குகிறது. இந்தத் தொழில்நுட்பம் "ஆன்-லைன் எழுத்துக்குறி புரிதல்" ("on-line character recognition"), "டைனமிக் கேரக்டர் புரிதல்" ("dynamic character recognition"), "நிகழ்நேர எழுத்துத் புரிதல்" ("real-time character recognition") மற்றும் "புத்திசாலித்தனமான எழுத்துப் புரிதல்" (intelligent character recognition") என்றும் அழைக்கப்படுகிறது.

நுட்பங்கள் (Techniques)

முன் செயலாக்கம் Pre-processing

ஒளிமூலம் எழுத்துப் புரிவான் (ஓசிஆர்/OCR) மென்பொருள் பெரும்பாலும் வெற்றிகரமான புரிதலுக்கான வாய்ப்புகளை மேம்படுத்த படங்களை "முன் செயலாக்குகிறது" ("pre-processes"). நுட்பங்கள் பின்வருமாறு (Optical Character Recognition (OCR) – How it works". Nicomsoft.com):

- டி-ஸ்கேவ் (De-skew)- ஸ்கேன் செய்யும் போது ஆவணம் சரியாக ஒழுங்குபடுத்தப்படவில்லை எனில், உரையின் வரிகளை கிடைமட்டமாக அல்லது செங்குத்தாக மாற்ற சில டிகிரி கடிகார திசையில் அல்லது கடிகார திசையில் சாய்ந்து கொள்ள வேண்டியிருக்கும்.
- டெஸ்பெகிள் (Despeckle) - நேர்மறை மற்றும் எதிர்மறை புள்ளிகள், மென்மையான விளிம்புகளை அகற்றவும்
- பைனரைசேஷன் (Binarisation) - ஒரு படத்தை வண்ணம் அல்லது கிரேஸ்கேலில் இருந்து கருப்பு மற்றும் வெள்ளை நிறமாக மாற்றவும் (இரண்டு வண்ணங்கள் இருப்பதால் "பைனரி படம்" என்று அழைக்கப்படுகிறது). உரையை (அல்லது வேறு ஏதேனும் விரும்பிய படக் கூறுகளை) பின்னணியில் இருந்து பிரிப்பதற்கான எளிய வழியாக பைனரைசேஷன் பணி செய்யப்படுகிறது (Sezgin & Sankur 2004). பெரும்பாலான வணிக அங்கீகார வழிமுறைகள் பைனரி படங்களில் மட்டுமே செயல்படுவதால் பைனரைசேஷனின் பணி அவசியம். ஏனெனில் அவ்வாறு செய்வது எளிது என்பதை நிரூபிக்கப்படுள்ளது Gupta, (Maya et al 2007). கூடுதலாக, பைனரைசேஷன் படியின்/நடபடிக்கையின் செயல்திறன் ஒரு குறிப்பிடத்தக்க அளவிற்கு எழுத்து புரிதல் கட்டத்தின் தரத்தை அதிகாரம்செய்கிறது மற்றும் ஒரு குறிப்பிட்ட உள்ளீட்டு பட வகைக்கு பயன்படுத்தப்படும் பைனரைசேஷனைத் தேர்ந்தெடுப்பதில் கவனமாக முடிவுகள் எடுக்கப்படுகின்றன; பைனரி முடிவைப் பெறுவதற்குப் பயன்படுத்தப்படும் பைனரைசேஷன் முறையின் தரம் உள்ளீட்டுப் படத்தின் வகையைப் பொறுத்தது (ஸ்கேன் செய்யப்பட்ட ஆவணம், காட்சி உரை படம், வரலாற்று சீரழிந்த ஆவணம் போன்றவை) (Trier & Jain 1995; Milyaev et al 2013)
- வரி நீக்கம் (Line removal) - கிளிஃப் அல்லாத பெட்டிகளையும் கோடுகளையும் சுத்தம் செய்கிறது
- தளவமைப்பு பகுப்பாய்வு அல்லது "ஸோனிங்" (Layout analysis or "zoning") - நெடுவரிசைகள், பத்திகள், தலைப்புகள் போன்றவற்றைத் தனித்துவமான தொகுதிகளாக அடையாளம் காட்டுகிறது. பல நெடுவரிசை தளவமைப்புகள் மற்றும் அட்டவணைகளில் குறிப்பாக முக்கியமானது.

- வரி மற்றும் சொல் கண்டறிதல் (Line and word detection) - சொல் மற்றும் எழுத்து வடிவங்களுக்கான அடிப்படைகளை நிறுவுகிறது, தேவைப்பட்டால் சொற்களைப் பிரிக்கிறது.
- எழுத்துவடிவத்தைப் புரிதல் (Script recognition) - பன்மொழி ஆவணங்களில், எழுத்துவடிவம் சொற்களின் மட்டத்தில் மாறக்கூடும், எனவே, குறிப்பிட்ட எழுத்துவடிவத்தை கையாளச் சரியான ஓசிஆர்-ஐப் (OCR) பயன்படுத்துவதற்கு முன்பு, எழுத்துவடிவத்தை அடையாளம் காண்பது அவசியம்.
- எழுத்து தனிமைப்படுத்தல் அல்லது "கூறுபடுத்தல்" (Character isolation or "segmentation") - ஒவ்வொரு எழுத்துக்குறி ஓசிஆருக்கு (per-character OCR), படக் கலைப்பொருட்கள் காரணமாக இணைக்கப்பட்டுள்ள பல எழுத்துக்கள் பிரிக்கப்பட வேண்டும்; செயற்கைப்பொருள் (artifacts) காரணமாக பல துண்டுகளாக உடைக்கப்பட்ட ஒற்றை எழுத்துக்கள் இணைக்கப்பட வேண்டும்.
- தோற்ற விகிதம் மற்றும் அளவை இயல்பாக்குதல்

நிலையான-இசைமை எழுத்துருக்களின் (fixed-pitch fonts) கூறாக்கம் ஒப்பீட்டளவில் வெறுமனே படத்தை ஒரு சீரான கட்டத்துடன் (uniform grid) செங்குத்து கட்டக் கோடுகள் எப்போதாவது கருப்பு பகுதிகளை இடைவெட்டுகின்றன என்பதன் அடிப்படையில் இணைப்பதன் மூலம் நிறைவேற்றப்படுகிறது. விகிதாசார எழுத்துருக்களுக்கு, அதிநவீன நுட்பங்கள் தேவைப்படுகின்றன, ஏனென்றால் எழுத்துக்களுக்கு இடையில் உள்ள இடைவெளி சில நேரங்களில் சொற்களுக்கு இடையில் இருப்பதை விட அதிகமாக இருக்கலாம், மேலும் செங்குத்து கோடுகள் ஒன்றுக்கு மேற்பட்ட எழுத்துக்களை இடைவெட்டக்கூடும் (Ray Smith 2007).

உரை புரிதல் (Text recognition)

மைய ஒளிமூலம் எழுத்துப் புரிவான் (கோர் ஓ.சி.ஆர்.) வழிமுறையின் இரண்டு அடிப்படை வகைகள் உள்ளன, அவை தேர்வுக்குரிய எழுத்துக்களின் தரவரிசைப் பட்டியலை உருவாக்கக்கூடும் ("OCR Introduction". Dataid.com.)

மேட்ரிக்ஸ் பொருத்தம் (Matrix matching) என்பது ஒரு படத்தை பிக்சல்-பை-பிக்சல் அடிப்படையில் சேமிக்கப்பட்ட கிளிஃபுடன் ஒப்பிடுவதை உள்ளடக்குகிறது; இது "அமைப்பொழுங்குப் பொருத்தம்" ("pattern matching"), "அமைப்பொழுங்குப் புரிதல்" ("pattern recognition") அல்லது "படத் தொடர்பு", ("image correlation") என்றும் அழைக்கப்படுகிறது. இது

உள்ளீட்டுக் கிளிஃபைப் படத்தின் பிற பகுதிகளிலிருந்து சரியாக தனிமைப்படுத்தப்படுவதையும், சேமிக்கப்பட்ட கிளிஃப் ஒத்த எழுத்துருவிலும் அதே அளவிலும் இருப்பதையும் நம்பியுள்ளது. இந்த நுட்பம் தட்டச்சு செய்யப்பட்ட உரையுடன் சிறப்பாகச் செயல்படுகிறது, மேலும் புதிய எழுத்துருக்களை எதிர்கொள்ளும்போது நன்றாக வேலை செய்யாது. ஆரம்பகால இயற்பியல் ஒளிச்சேர்க்கை அடிப்படையிலான ஓசிஆர் (physical photocell-based OCR) நேரடியாக செயல்படுத்தப்பட்ட நுட்பமாகும்.

பண்புக்கூறு பிரித்தெடுத்தல் கோடுகள் (lines), மூடிய கண்ணிகள்/வளையங்கள் (closed loops), வரி திசை (line direction) மற்றும் வரி குறுக்குவெட்டுகள் (line intersections) போன்ற "பண்புக்கூறுகளாக" கிளிஃப்களை சிதைக்கிறது. பிரித்தெடுக்கும் பண்புக்கூறுகள் உருப்படுத்தத்தின் பரிமாணத்தை குறைக்கிறது மற்றும் புரிதல் செயல்முறையைக் கணினி/கணக்கீட்டு அடிப்படையில் திறம்படச் செய்கிறது. இந்தப் பண்புக்கூறுகள் ஒரு எழுத்தின் சுருக்கத் திசையன் போன்ற (abstract vector-like) உருப்படுத்தத்துடன் ஒப்பிடப்படுகின்றன, இது ஒன்று அல்லது அதற்கு மேற்பட்ட கிளிஃப் மூலமுன்மாதிரிகளாகக் (glyph prototypes) குறையக்கூடும். கணினி பார்வையில் பண்புக்கூறுறைக் கண்டறிவதற்கான பொதுவான நுட்பங்கள் இந்த வகை ஓசிஆர்-க்கு பொருந்தும், இது பொதுவாக "அறிவார்ந்த" கையெழுத்துப் புரிதல் ("intelligent" handwriting recognitionமற்றும்) உண்மையில் நவீன ஓசிஆர்/OCR மென்பொருளில் காணப்படுகிறது. கே-அருகிலுள்ள அண்டைகள் வழிமுறை (k-nearest neighbors algorithm) போன்ற அருகிலுள்ள அண்டை வகைப்படுத்திகள் (Nearest neighbour classifiers) பட பண்புக்கூறுகளைச் சேமிக்கப்பட்ட கிளிஃப் பண்புக்கூறுகளுடன் ஒப்பிட்டு அருகிலுள்ள பொருத்ததைத் (nearest match) தேர்வுசெய்யப் பயன்படுத்தப்படுகின்றன.

கியூனிஃபார்ம் (Cuneiform) மற்றும் டெசராக்ட் (Tesseract) போன்ற மென்பொருள்கள் எழுத்துப் புரிதலுக்கு (character recognition) இரண்டு-பாஸ் அணுகுமுறையைப் பயன்படுத்துகின்றன. இரண்டாவது பாஸ் "தகவமைப்பு புரிதல்" ("adaptive recognition") என்று அழைக்கப்படுகிறது மற்றும் இரண்டாவது பாஸில் மீதமுள்ள எழுத்துக்களைச் சிறப்பாக அடையாளம் காண முதல் பாஸில் அதிக நம்பிக்கையுடன் புரியப்பட்ட எழுத்து வடிவங்களைப் பயன்படுத்துகிறது. எழுத்துரு சிதைந்த அசாதாரண எழுத்துருக்கள் அல்லது குறைந்த தரமான ஸ்கேன்களுக்கு (எ.கா. மங்கலான அல்லது மங்கிப்போன) இது சாதகமானது (Ray Smith 2007).

நவீன ஒளிமூலம் எழுத்துப் புரிதல் (OCR) மென்பொருளானது ஓசிஆர்ஓபஸ் OCRopus அல்லது தெஸ்ஸெராக்ட் (Tesseract) போன்ற நரம்பியல் வலையமைப்புகளைப் (நெட்வொர்க்குகளைப்) பயன்படுத்துகிறது, அவை ஒற்றை எழுத்துக்களில் கவனம் செலுத்துவதற்குப் பதிலாக உரையின் முழு வரிகளையும் அடையாளம் காணப் பயிற்சி பெற்றன.

ஓசிஆர்/OCR முடிவை தரப்படுத்தப்பட்ட ஆல்டோ (ALTO) வடிவத்தில் சேமிக்க முடியும், இது யுனைடெட் ஸ்டேட்ஸ் லைப்ரரி ஆஃப் காங்கிரஸால் (United States Library of Congress) பராமரிக்கப்படும் ஒரு பிரத்யேக எக்ஸ்எம்எல் (XML) திட்டமாகும். பிற பொதுவான வடிவங்களில் hOCR மற்றும் PAGE XML ஆகியவை அடங்கும்.

ஒளிவழி எழுத்துப் புரிதல் (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன்) மென்பொருளின் பட்டியலுக்கு ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன் மென்பொருளின் ஒப்பீடு பார்க்கவும்.

பின் செயலாக்கம் (Post-processing)

வெளியீட்டை ஒரு அகராதி (ஒரு ஆவணத்தில் நிகழ அனுமதிக்கப்பட்ட சொற்களின் பட்டியல்) மூலம் கட்டுப்படுத்தினால் ஒளிமூலம் எழுத்துப் புரிவான் துல்லியத்தை அதிகரிக்க இயலும் ("Optical Character Recognition (OCR) – How it works". Nicomsoft.com.). எடுத்துக்காட்டாக, இது ஆங்கில மொழியில் உள்ள அனைத்து சொற்களும் அல்லது ஒரு குறிப்பிட்ட புலத்திற்கான தொழில்நுட்ப அகராதியாக இருக்கலாம். இயற்பெயர்களைப் போல, அகராதியில் இல்லாத சொற்கள் ஆவணத்தில் இருந்தால் இந்த நுட்பம் சிக்கலாக இருக்கும். மேம்பட்ட துல்லியத்திற்காக, எழுத்து கூறாக்க நடத்தையை ஊகுவிக்க டெசராக்ட் அதன் அகராதியைப் பயன்படுத்துகிறது (Ray Smith 2007).

வெளியீட்டு ஒழுக்கு ஒரு எளிய உரை ஒழுக்கு அல்லது எழுத்துகளின் கோப்பாக இருக்கலாம், ஆனால் மிகவும் அதிநவீன ஓசிஆர் அமைப்புகள் பக்கத்தின் அசல் தளவமைப்பைப் பாதுகாத்து உருவாக்கலாம்; எடுத்துக்காட்டாக, பக்கத்தின் அசல் படம் மற்றும் தேடக்கூடிய உரை உருப்படுத்தம் ஆகிய இரண்டையும் உள்ளடக்கிய ஒரு அடையாளப்படுத்தப்பட்ட PDF .

"அருகிலுள்ள அண்டை பகுப்பாய்வு" ("Near-neighbor analysis") சில சொற்கள் பெரும்பாலும் ஒன்றாகக் காணப்படுவதைக் குறிப்பிடுவதன் மூலம் பிழைகளைச் சரிசெய்ய இணைநிகழ்வு அதிர்வெண்களைப் (co-occurrence frequencies) பயன்படுத்தலாம் ("How does OCR document scanning work?"). எடுத்துக்காட்டாக, "Washington, D.C." பொதுவாக "Washington DOC" - ஐ விட ஆங்கிலத்தில் மிகவும் பொதுவானது.

ஸ்கேன் செய்யப்படும் மொழியின் இலக்கணத்தைப் பற்றிய அறிவு ஒரு சொல் வினைச்சொல்லாகவோ அல்லது பெயர்ச்சொல்லாகவோ இருக்க முடியுமா என்பதை தீர்மானிக்க உதவும், எடுத்துக்காட்டாக, அதிக துல்லியத்தை அனுமதிக்கிறது.

OCR API-இன் முடிவுகளை மேலும் மேம்படுத்த ஒளிமூலம் எழுத்துப் புரிவான் பிந்தைய செயலாக்கத்திலும் லெவன்ஸ்டீன் தொலைநிலை வழிமுறை பயன்படுத்தப்பட்டுள்ளது ("How to optimize results from …. Community").

பயன்பாடு சார்ந்த மேம்படுத்தல்கள் (Application-specific optimizations)

சமீபத்திய ஆண்டுகளில், முக்கிய ஒளிமூலம் எழுத்துப் புரிவான் தொழில்நுட்ப வழங்குநர்கள் குறிப்பிட்ட வகை உள்ளீடுகளை மிகவும் திறமையாகக் கையாள ஒளிமூலம் எழுத்துப் புரிதல் ஒழுங்குமுறைகளை மாற்றத் தொடங்கினர். பயன்பாட்டு-குறிப்பிட்ட அகராதிக்கு (application-specific lexicon) அப்பால், வணிக விதிகள், நிலையான வெளிப்பாடு, அல்லது வண்ணப் படங்களில் உள்ள வளமான தகவல்களை கணக்கில் எடுத்துக்கொள்வதன் மூலம் சிறந்த செயல்திறனைப் பெறலாம். இந்த மூலோபாயம் "பயன்பாடு சார்ந்த ஒளிமூலம் எழுத்துப் புரிவான்" ("Application-Oriented OCR") அல்லது "தனிப்பயனாக்கப்பட்ட ஒளிமூலம் எழுத்துப் புரிவான்" ("Customized OCR") என்று அழைக்கப்படுகிறது; மேலும் இது உரிமத் தகடுகள், விலைப்பட்டியல், ஸ்கிரீன் ஷாட்கள், அடையாள அட்டைகள், ஓட்டுநர் உரிமங்கள் மற்றும் ஆட்டோமொபைல் உற்பத்தி ஆகியவற்றில் ஒளிமூலம் எழுத்துப் புரிவான் பயன்படுத்தப்படுகிறது.

நியூயார்க் டைம்ஸ் ஓசிஆர் தொழில்நுட்பத்தை அவர்கள் வைத்திருக்கும் தனியுரிம கருவியாக ஆவண உதவியாளராக (Document Helper) மாற்றியமைத்துள்ளது, இது அவர்களின் ஊடாடும் செய்திக் குழுவை மதிப்பாய்வு செய்ய வேண்டிய ஆவணங்களின் செயலாக்கத்தை துரிதப்படுத்த உதவுகிறது. நிருபர்கள் உள்ளடக்கங்களை மறுஆய்வு செய்வதற்கான தயாரிப்பில் ஒரு மணி நேரத்திற்கு 5,400 பக்கங்கள் வீதம் செயலாக்க இது உதவுகிறது என்பதை அவர்கள் குறிப்பிடுகிறார்கள் (Fehr 2019)

6.6.2. குரலைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள் (Voice recognition systems)

பேச்சுப் புரிந்துகொள்ளுதல் (Speech recognition) என்பது கணினி அறிவியல் மற்றும் கணக்கீட்டு மொழியியலின் ஒரு இடைநிலை துணைத் துறையாகும், இது கணினிகள் மூலம் பேசும் மொழியை உரையாகப் புரிந்துகொள்ளவும் மொழிபெயர்க்கவும் உதவும் முறைகள் மற்றும் தொழில்நுட்பங்களை உருவாக்குகிறது. இது தானியங்கிப் பேச்சு புரிந்துகொள்ளல் (automatic

speech recognition (ASR),) அல்லது கணினி பேச்சு அங்கீகாரம் அல்லது பேச்சு முதல் உரை (computer speech recognition or speech to text (STT)) என்றும் அழைக்கப்படுகிறது. இது கணினி அறிவியல், மொழியியல் மற்றும் கணினி பொறியியல் துறைகளில் அறிவு மற்றும் ஆராய்ச்சியை ஒருங்கிணைக்கிறது.

சில பேச்சுப் புரிதல் ஒழுங்குமுறைகளுக்கு ஒரு தனிப்பட்ட பேச்சாளர் உரை அல்லது தனிமைப்படுத்தப்பட்ட சொற்றொகையை ஒழுங்குமுறையில் படிக்கும் "பயிற்சி" "training" ("சேர்க்கை" ("enrollment") என்றும் அழைக்கப்படுகிறது) தேவைப்படுகிறது,. கணினி நபரின் குறிப்பிட்ட குரலை பகுப்பாய்வு செய்து, அந்த நபரின் பேச்சின் அங்கீகாரத்தை நன்றாகப் பயன்படுத்துவதற்குப் பயன்படுத்துகிறது, இதன் விளைவாக துல்லியம் அதிகரிக்கும். பயிற்சியைப் பயன்படுத்தாத அமைப்புகள் "பேச்சாளர் சுதந்திரமான" ("speaker independent") ("Speaker Independent …. Corporation") அமைப்புகள் என்று அழைக்கப்படுகின்றன. பயிற்சியைப் பயன்படுத்தும் அமைப்புகள் "பேச்சாளர் சார்பு" ("speaker dependent") என்று அழைக்கப்படுகின்றன.

பேச்சு புரிதல் பயன்பாடுகளில் குரல் பயனர் இடைமுகங்களான (voice user interfaces) குரல் டயலிங் (voice dialing) (எ.கா. "வீட்டிற்கு அழைப்பு"), அழைப்பு ரூட்டிங் (routing) (எ.கா. "நான் ஒரு சேகரிப்பு அழைப்பை செய்ய விரும்புகிறேன்"), டோமோடிக் பயன்பாட்டுக் கட்டுப்பாடு (domotic appliance control), தேடல் முக்கிய சொற்கள் (எ.கா. குறிப்பிட்ட சொற்களைக் கொண்ட போட்காஸ்டைக் கண்டுபிடி பேசப்பட்டது), எளிய தரவு உள்ளீடு (எ.கா., கிரெடிட் கார்டு எண்ணை உள்ளிடுதல்), கட்டமைக்கப்பட்ட ஆவணங்களைத் தயாரித்தல் (எ.கா. ஒரு கதிரியக்க அறிக்கை), பேச்சாளர் பண்புகளை தீர்மானித்தல் (Nguyen 2010)), பேசிலிருந்து உரை செயலாக்கம் (எ.கா., சொல் செயலிகள் அல்லது மின்னஞ்சல்கள்), மற்றும் விமானம் (பொதுவாக நேரடி குரல் உள்ளீடு என்று அழைக்கப்படுகிறது).

குரல் புரிதல் (voice recognition) ("British English definition of voice recognition", "voice recognition, definition of") அல்லது பேச்சாளர் அடையாளம் காணல் (Speaker recognition) (Reynolds & Rose 1995; "Speaker Identification (WhisperID))" என்பது அவர்கள் சொல்வதைக் காட்டிலும் பேச்சாளரை அடையாளம் காண்பதைக் குறிக்கிறது . பேச்சாளரை அங்கீகரிப்பது ஒரு குறிப்பிட்ட நபரின் குரலில் பயிற்சியளிக்கப்பட்ட அமைப்புகளில் பேச்சை மொழிபெயர்க்கும்

பணியை எளிதாக்குகிறது அல்லது பாதுகாப்புச் செயல்பாட்டின் ஒரு பகுதியாக பேச்சாளரின் அடையாளத்தை அங்கீகரிக்க அல்லது சரிபார்க்க இது பயன்படுத்தப்படலாம்.

தொழில்நுட்ப கண்ணோட்டத்தில், பேச்சு அங்கீகாரம் பல பெரிய கண்டுபிடிப்புகளுடன் நீண்ட வரலாற்றைக் கொண்டுள்ளது. மிக சமீபத்தில், ஆழ்ந்த கற்றல் மற்றும் பெரிய தரவுகளின் முன்னேற்றங்களிலிருந்து இந்தத் துறை பயனடைந்துள்ளது. இந்த துறையில் வெளியிடப்பட்ட கல்வித் தாள்களின் எழுச்சியால் மட்டுமல்லாமல், மிக முக்கியமாக உலகளாவிய தொழில்துறை பேச்சு அங்கீகார முறைகளை வடிவமைப்பதிலும் பயன்படுத்துவதிலும் பல்வேறு ஆழமான கற்றல் முறைகளைப் பின்பற்றுவதன் மூலம் முன்னேற்றங்கள் சாட்சியமளிக்கின்றன.

6.6.3. உரையிலிருந்து பேச்சு ஒழுங்குமுறைகள் (Text to speech systems)

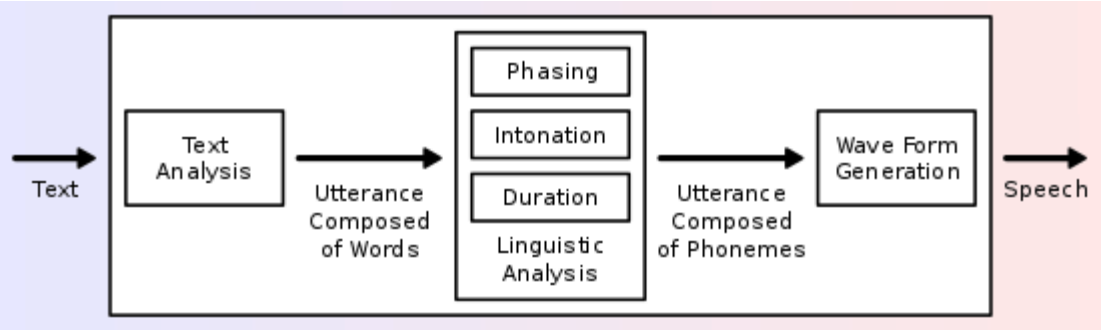
பேச்சு உருவாக்கம் (Speech synthesis) என்பது மனித பேச்சின் செயற்கை உற்பத்தி ஆகும். இந்த நோக்கத்திற்காக பயன்படுத்தப்படும் ஒரு கணினி அமைப்பு பேச்சு கணினி அல்லது பேச்சு சின்தசைசர் என அழைக்கப்படுகிறது, மேலும் இது மென்பொருள் அல்லது வன்பொருள் தயாரிப்புகளில் செயல்படுத்தப்படலாம். ஒரு உரையிலிருந்து பேச்சு (text-to-speech (TTS/டி.டி.எஸ்) ஒழுங்குமுறை சாதாரண மொழி உரையைப் பேச்சாக மாற்றுகிறது; பிற அமைப்புகள் ஒலிப்பு டிரான்ஸ்கிரிப்ஷன்கள் (phonetic transcriptions) போன்ற குறியீட்டு மொழியியல் உருப்படுத்தங்களை (symbolic linguistic representations) பேச்சாக வழங்குகின்றன Allen et al 1987).

ஒரு தரவுத்தளத்தில் சேமிக்கப்பட்டுள்ள பதிவுசெய்யப்பட்ட பேச்சின் துண்டுகளை ஒன்றிணைப்பதன் மூலம் ஒருங்கிணைந்த பேச்சை உருவாக்க முடியும். சேமிக்கப்பட்ட பேச்சு அலகுகளின் அளவுகளில் அமைப்புகள் வேறுபடுகின்றன; ஒலிகளை (phones) அல்லது ஈரொலிகளைச் (diphones) சேமிக்கும் அமைப்பு மிகப்பெரிய வெளியீட்டு வரம்பை வழங்குகிறது, ஆனால் தெளிவு இல்லாமல் இருக்கலாம். குறிப்பிட்ட பயன்பாட்டு களங்களுக்கு, முழு சொற்கள் அல்லது வாக்கியங்களின் சேமிப்பு உயர்தர வெளியீட்டை அனுமதிக்கிறது. மாற்றாக, ஒரு சின்தசைசர் முற்றிலும் "செயற்கை" குரல் வெளியீட்டை உருவாக்க குரல் பாதை மற்றும் பிற மனித குரல் பண்புகளின் மாதிரியை இணைக்க முடியும் (Rubin et al 1981).

பேச்சு உருவாக்கியின் தரம் மனித குரலுடன் அதன் ஒற்றுமை மற்றும் தெளிவாக புரிந்துகொள்ளும் திறன் ஆகியவற்றால் தீர்மானிக்கப்படுகிறது. புத்திசாலித்தனமான உரையிலிந்து பேச்சுத் திட்டம் பார்வைக் குறைபாடுகள் (visual impairments) அல்லது வாசிப்பு

குறைபாடுகள் (reading disability) உள்ளவர்கள் வீட்டு கணினியில் எழுதப்பட்ட சொற்களைக் கேட்க அனுமதிக்கிறது. பல கணினி இயக்க முறைமைகள் 1990களின் முற்பகுதியில் இருந்து பேச்சு உருவாக்கிகளை உள்ளடக்கியுள்ளன.

பின்வரும் படம் விக்கிபீடியாவிலிருந்து எடுத்தாளப்பட்டுள்ளது.



ஒரு உரையிலிருந்து பேச்சு ஒழுங்குமுறை (அல்லது "இயந்திரம்") இரண்டு பகுதிகளைக் கொண்டது (van Santen et al 1997): ஒரு முன்-முனை (front-end) மற்றும் பின்-முனை (back-end). முன் இறுதியில் இரண்டு முக்கிய செயல்பாடுகள் உள்ளன. முதலில், இது எண்கள் மற்றும் சுருக்கங்கள் (abbreviations) போன்ற குறியீடுகளைக் கொண்ட மூல உரையை எழுதப்பட்ட சொற்களுக்குச் சமமாக மாற்றுகிறது. இந்தச் செயல்முறை பெரும்பாலும் உரை இயல்பாக்கம் normalization, முன்செயலாக்கம் pre-processing அல்லது டோக்கனைசேஷன் (tokenization) என்று அழைக்கப்படுகிறது. முன் இறுதியில் ஒவ்வொரு வார்த்தைக்கும் ஒலிப்பு வரிவடிவாக்கங்களை (phonetic transcriptions) ஒதுக்குகிறது, மேலும் உரையை சொற்றொடர்கள், எச்சத்தொடர்கள் மற்றும் வாக்கியங்கள் போன்ற மீக்கூறு அலகுகளாகப் (prosodic units) பிரித்து குறிக்கிறது. சொற்களுக்கு ஒலிப்பு வரிவடிவாக்கங்களை ஒதுக்கும் செயல்முறையை உரையிலிருந்து ஒலியன் (text-to-phoneme) அல்லது எழுத்திலிருந்து ஒலியன் மாற்றம் (grapheme-to-phoneme conversion) எனஅழைக்கப்படுகிறது. ஒலிப்பு வரிவடிவாக்கங்களும் மீக்கூறு தகவல்களும் சேர்ந்து குறியீட்டு மொழியியல்சார் உருப்படுத்தத்தை (symbolic linguistic representation) உருவாக்குகின்றன; இது முன் இறுதியின் வெளியீடு ஆகும். பெரும்பாலும் உருவாக்கி எனக் குறிப்பிடப்படும் பின்-இறுதி (synthesizer) பின்னர் குறியீட்டு மொழியியல் உருப்படுத்தத்தை ஒலியாக மாற்றுகிறது. சில ஒழுங்குமுறைகளில், இந்தப் பகுதி வெளியீட்டு உரையில் திணிக்கப்படும் இலக்கு மீக்கூறின் கணக்கீடு/கணினியாக்கத்தை (இசைமை, காண்டூர், ஒலியன் கால அளவுகள்) (Van Santen, 1994) உட்படுத்தும்.

தொகுப்பு தொழில்நுட்பங்கள்

பேச்சுத் தொகுப்பு அமைப்பின் மிக முக்கியமான குணங்கள் இயல்பான தன்மை மற்றும் புத்திசாலித்தனம் (Taylor 2009). இயல்பானது வெளியீடு மனித பேச்சைப் போல எவ்வளவு நெருக்கமாக ஒலிக்கிறது என்பதை விவரிக்கிறது, அதே நேரத்தில் புத்திசாலித்தனம் என்பது வெளியீட்டைப் புரிந்துகொள்வதற்கான எளிமை. சிறந்த பேச்சு தொகுப்பு இயற்கையானது மற்றும் புரியக்கூடியது. பேச்சுத் தொகுப்பு அமைப்புகள் பொதுவாக இரு குணாதிசயங்களையும் அதிகரிக்க முயற்சிக்கின்றன.

செயற்கைப் பேச்சு அலைவடிவங்களை உருவாக்கும் இரண்டு முதன்மை தொழில்நுட்பங்கள் ஒன்றிணைத்த தொகுப்பு (concatenative synthesis) மற்றும் வடிவத் தொகுப்பு (formant synthesis) ஆகும். ஒவ்வொரு தொழில்நுட்பத்திற்கும் பலங்களும் பலவீனங்களும் உள்ளன, மேலும் ஒரு தொகுப்பு அமைப்பின் நோக்கம் பொதுவாக எந்த அணுகுமுறை பயன்படுத்தப்படுகிறது என்பதைத் தீர்மானிக்கும்.

ஒன்றிணைத்த ஒருங்கிணைப்பு

ஒன்றிணைத்த ஒருங்கிணைப்பு என்பது பதிவுசெய்யப்பட்ட பேச்சின் கூறுகளின் ஒன்றிணைப்பை (அல்லது ஒன்றாகக் கோர்த்தல்) அடிப்படையாகக் கொண்டது. பொதுவாக, ஒன்றிணைத்த தொகுப்பு மிகவும் இயல்பாக ஒலிக்கும் ஒன்றிணைத்த உரையை உருவாக்குகிறது. இருப்பினும், பேச்சில் இயற்கையான மாறுபாடுகள் மற்றும் அலைவடிவங்களை கூறுபடுத்துவபதற்கான தானியங்கு நுட்பங்களின் தன்மை ஆகியவற்றுக்கு இடையிலான வேறுபாடுகள் சில நேரங்களில் வெளியீட்டில் கேட்கக்கூடிய குறைபாடுகளை ஏற்படுத்துகின்றன. ஒன்றிணைத்த தொகுப்பின் மூன்று முக்கிய துணை வகைகள் உள்ளன: அலகுத்தேர்வுத் தொகுப்பு, இரும ஒலி தொகுப்பு, களம் குறிப்பிட்ட தொகுப்பு.

அலகுத் தேர்வு ஒருங்கிணைப்பு (Unit selection synthesis)

அலகு தேர்வு ஒருங்கிணைப்பு பதிவு செய்யப்பட்ட பேச்சின் பெரிய தரவுத்தளங்களைப் பயன்படுத்துகிறது. தரவுத்தள உருவாக்கத்தின் போது பதிவுசெய்யப்பட்ட ஒவ்வொரு சொற்களும் பின்வருவனவற்றில் சில அல்லது அனைத்திலும் கூறிடப்படுகின்றன: தனிப்பட்ட ஒலிகள், இருமை ஒலிகள், அரை ஒலிகள், அசைகள், உருபன்கள், தொடர்கள், மற்றும் வாக்கியங்கள். பொதுவாக, கூறுகளாகப் பிரித்தல் சிறப்பாக மாற்றியமைக்கப்பட்ட பேச்சு புரிதலைப் பயன்படுத்தி "கட்டாய ஒழுங்மைப்பு" பயன்முறையில் அமைக்கப்பட்ட பின்னர் சில மனித

முயற்சித் திருத்தங்களுடன், அலைவடிவம் (waveform) மற்றும் ஸ்பெக்ட்ரோகிராம் and (spectrogram) போன்ற காட்சி உருப்படுத்தங்களைப் பயன்படுத்திச் செய்யப்படுகிறது (Black, 2002). பேச்சு தரவுத்தளத்தில் உள்ள அலகுகளின் உள்ளடக்க அட்டவணை (index) அடிப்படை அதிர்வெண் (இசைமை), கால அளவு, அசையில் நிலை மற்றும் அண்டை ஒலிகள் போன்ற கூறாக்கம் மற்றும் ஒலி அளவுருக்களின் அடிப்படையில் உருவாக்கப்படுகிறது. இயக்க நேரத்தில், தரவுத்தளத்திலிருந்து (அலகு தேர்வு) தேர்வுக்குரிய அலகுகளின் சிறந்த சங்கிலியை தீர்மானிப்பதன் மூலம் விரும்பிய இலக்கு உச்சரிப்பு உருவாக்கப்படுகிறது. இந்த செயல்முறை பொதுவாக விசேஷமாக நிறைசெய்த தீர்மனக் கிளையை (decision tree) பயன்படுத்தி அடையப்படுகிறது.

அலகுத் தேர்வு மிகப் பெரிய இயல்பை வழங்குகிறது, ஏனெனில் இது பதிவுசெய்யப்பட்ட பேச்சுக்கு ஒரு சிறிய அளவு டிஜிட்டல் சிக்னல் செயலாக்கத்தை (digital signal processing (DSP/ டிஎஸ்பி)) மட்டுமே பயன்படுத்தும். டிஜிட்டல் சிக்னல் செயலாக்கம் பெரும்பாலும் பதிவுசெய்யப்பட்ட பேச்சு ஒலியைக் குறைவாக இயல்பாக்குகிறது, இருப்பினும் சில அமைப்புகள் அலைவடிவத்தை மென்மையாக்குவதற்கு ஒரு சிறிய அளவிலான சமிக்ஞைச்/குறிகைச் செயலாக்கத்தை ஒன்றிணைக்கும் கட்டத்தில் பயன்படுத்துகின்றன. சிறந்த அலகு-தேர்வு அமைப்புகளின் வெளியீடு பெரும்பாலும் உண்மையான மனித குரல்களிலிருந்து பிரித்தறிய முடியாதது, குறிப்பாக உரையிலிருந்து பேச்சு (டி.டி.எஸ்) ஒழுங்குமுறை சரிசெய்யப்பட்ட சூழல்களில். இருப்பினும், அதிகபட்ச இயல்புக்குப் பொதுவாக அலகு-தேர்வு பேச்சு தரவுத்தளங்கள் மிகப் பெரியதாக இருக்க வேண்டும், சில அமைப்புகளில் டஜன் கணக்கான மணிநேர பேச்சை உருப்படுத்தம்செய்யும் பதிவுசெய்யப்பட்ட தரவுகளின் ஜிகாபைட் வரை இருக்க வேண்டும் (John Kominek and Black 2003). மேலும், தரவுத்தளத்தில் ஒரு சிறந்த தேர்வு இருக்கும்போது கூட, சிறந்த தொகுப்பைக் காட்டிலும் குறைவான (எ.கா. சிறிய சொற்கள் தெளிவாகத் தெரியவில்லை) ஒரு இடத்திலிருந்து கூறுகளைத் தேர்ந்தெடுப்பது அலகு தேர்வு வழிமுறைகள் அறியப்படுகின்றன (Julia Zhang. 2002). சமீபத்தில், அலகு-தேர்வு பேச்சுத் தொகுப்பு அமைப்புகளில் இயற்கைக்கு மாறான கூறுகளைக் கண்டறிய ஆராய்ச்சியாளர்கள் பல்வேறு தானியங்கி முறைகளை முன்மொழிந்தனர் (William Yang Wang and Kallirroi Georgila. 2011).

ஈரொலி ஒருங்கிணைப்பு (Diphone synthesis)

ஈரொலித் தொகுப்பு ஒரு மொழியில் நிகழும் அனைத்து ஈரொலிகளையும் (ஒலி-க்கு-ஒலி மாற்றங்கள்) கொண்ட குறைந்தபட்சப் பேச்சுத் தரவுத்தளத்தைப் பயன்படுத்துகிறது. ஈரொலிகளின் எண்ணிக்கை மொழியின் ஒலியியல் சார்ந்த தன்மையைப் பொறுத்தது: எடுத்துக்காட்டாக, ஸ்பானிஷ் சுமார் 800 ஈரொலிகளையும், ஜெர்மன் 2500ஐயும் கொண்டுள்ளது. ஈரொலித் தொகுப்பில், ஒவ்வொரு ஈரொலிக்கும் ஒரு எடுத்துக்காட்டு மட்டுமே பேச்சுத் தரவுத்தளத்தில் உள்ளது. இயக்க நேரத்தில், நேரியல் முன்கணிப்பு குறியீட்டு முறை (linear predictive coding), PSOLA (பி.எஸ்.ஓ.எல்.ஏ) அல்லது MBROLA (எம்.பி.ஆர்.ஓ.எல்.ஏ.) போன்ற டிஜிட்டல் சிக்னல் செயலாக்க நுட்பங்கள் மூலம் அல்லது தனித்துவமான கொசைன் உருமாற்றத்தைப் (discrete cosine transform) பயன்படுத்தி மூலக் களத்தில் இசைமை மாற்றம் போன்ற சமீபத்திய நுட்பங்கள் மூலம் ஒரு வாக்கியத்தின் இலக்கு மீக்கூறு (புரோசோடி) இந்தக் குறைந்தபட்ச அலகுகள் மீது மேற்படியப்படுத்தப்பட்டுள்ளது (Dutoit et al 1996). (Muralishankar et al 2004). ஈரொலித் தொகுப்பு ஒன்றிணைத்த தொகுப்பின் சோனிக் குறைபாடுகள் (sonic glitches) மற்றும் வடிவத் தொகுப்பின் ரோபோ-ஒலிக்கும் தன்மை (robotic-sounding) ஆகியவற்றால் பாதிக்கப்படுகிறது; மேலும் சிறிய அளவைத் தவிர இரண்டில் ஏதேனுமொரு அணுகுமுறையின் சில நன்மைகள் உள்ளன. எனவே, வணிக பயன்பாடுகளில் அதன் பயன்பாடு குறைந்து வருகிறது; இருப்பினும் ஏராளமான இலவச மென்பொருள் செயலாக்கங்கள் உள்ளதால் இது ஆராய்ச்சியில் தொடர்ந்து பயன்படுத்தப்படுகிறது. ஈரொலித் தொகுப்பின் ஆரம்ப எடுத்துக்காட்டு மைக்கேல் ஜே. ஃப்ரீமேன் (Michael J. Freeman) கண்டுபிடித்த லீச்சிம் (leachim) என்ற கற்பித்தல் ரோபோ ஆகும் (Rudy the Robot - Michael Freeman). லீச்சிம் வகுப்பு பாடத்திட்டங்கள் மற்றும் கற்பிக்க திட்டமிடப்பட்ட 40 மாணவர்களைப் பற்றிய சில வாழ்க்கை வரலாற்று தகவல்கள் பற்றிய தகவல்களைக் கொண்டிருந்தது. இது நியூயார்க்கின் பிராங்க்ஸில் (the Bronx, New York) நான்காம் வகுப்பு வகுப்பறையில் சோதிக்கப்பட்டது.

களம்-குறிப்பிட்ட ஒருங்கிணைப்பு

களம்-குறிப்பிட்ட ஒருங்கிணைப்பு முழுமையான சொற்களை உருவாக்க முன்பே பதிவுசெய்யப்பட்ட சொற்களையும் சொற்றொடர்களையும் இணைக்கிறது. போக்குவரத்து அட்டவணை அறிவிப்புகள் அல்லது வானிலை அறிக்கைகள் போன்ற ஒரு குறிப்பிட்ட களத்திற்கு கணினி வெளியிடும் பல்வேறு உரைகள் வரையறுக்கப்பட்ட பயன்பாடுகளில் இது பயன்படுத்தப்படுகிறது (Lamel et al 1993). தொழில்நுட்பம் செயல்படுத்த மிகவும் எளிதானது,

மற்றும் பேசும் கடிகாரங்கள் மற்றும் கால்குலேட்டர்கள் போன்ற சாதனங்களில் நீண்ட காலமாக வணிகப் பயன்பாட்டில் உள்ளது. இந்த ஒழுங்குமுறைகளின் இயல்பானதன்மையின் நிலை மிக அதிகமாக இருக்கலாம்; ஏனெனில் பல்வேறு வகையான வாக்கிய வகைகள் குறைவாகவே உள்ளன, மேலும் அவை அசல் பதிவுகளின் மீக்கூறுடனும் இசையோட்டத்துடனும் நெருக்கமாக பொருந்துகின்றன.

இந்த ஒழுங்குமுறைகள் அவற்றின் தரவுத்தளங்களில் உள்ள சொற்கள் மற்றும் சொற்றொடர்களால் கட்டுப்படுத்தப்பட்டுள்ளதால், அவை பொதுவான நோக்கம் கொண்டதல்ல, மேலும் அவை முன்கூட்டியே திட்டமிடப்பட்ட சொற்கள் மற்றும் சொற்றொடர்களின் சேர்க்கைகளை மட்டுமே ஒருங்கிணைக்க முடியும். இயற்கையாகவே பேசப்படும் மொழியில் சொற்களைக் கலக்கும்போது பல வேறுபாடுகளைக் கணக்கில் எடுத்துக் கொள்ளாவிட்டால் இன்னும் சிக்கல்களை ஏற்படுத்தும். எடுத்துக்காட்டாக, ஆங்கிலத்தின் ரோட்டிக் அல்லாத கிளைமொழிகளில் (non-rhotic dialects) (குறிப்பிட இடங்களில் ”r” உச்சரிக்கப்படாத) "clear" /ˈklɪə/ போன்ற சொற்களில் "r" என்பது அதன் பின்வரும் வார்த்தையில் முதல் எழுத்து ஒரு உயிரெழுத்து ஆக இருக்கும்போது மட்டுமே பொதுவாக உச்சரிக்கப்படுகிறது (எ.கா. " clear out " /ˌklɪəɹˈʌʊt/). அதேபோல் பிரெஞ்சு மொழியில் பல இறுதி மெய்யெழுத்துக்கள் கொண்ட சொற்களைத் தொடர்ந்து உயிரெழுத்துடன் தொடங்கும் சொல் தொடர்ந்து வந்தால் இறுதி மெய்யெழுத்துக்கள் மௌனமாக வராமல் உச்சரிக்கப்படும்; இந்த விளைவு லெய்சன் (liaison) என்று அழைக்கப்படுகிறது. இந்த மாற்றீட்டை ஒரு எளிய சொல்-இணைத்தல் முறையால் மீண்டும் உருவாக்க இயலாது; இது சூழல்-உணர்திறன் கொண்டதாக இருக்க கூடுதல் கலவைத்தன்மை தேவைப்படும்.

வடிவ ஒருங்கிணைப்பு (Formant synthesis)

வடிவத் ஒருங்கிணைப்பு மனித பேச்சுப் பதக்கூறுகளை (speech samples) இயக்க நேரத்தில் பயன்படுத்தாது. அதற்குப் பதிலாக, சேர்க்கை தொகுப்பு (additive synthesis) மற்றும் ஒலியியக்க மாதிரி (acoustic model) (இயற்பியல் மாதிரித் தொகுப்பு (physical modelling synthesis)) ஆகியவற்றைப் பயன்படுத்தி உருவாக்கப்படுகிறது. செயற்கை பேச்சின் அலைவடிவத்தை உருவாக்க அடிப்படை அதிர்வெண், குரல் மற்றும் சத்தம் அளவுகள் போன்ற அளவுருக்கள் காலப்போக்கில் மாறுபடும். இந்த முறை சில நேரங்களில் விதிகள் சார்ந்த தொகுப்பு என்று அழைக்கப்படுகிறது; இருப்பினும், பல ஒன்றிணைக்கப்பட ஒழுங்குமுறைகளும் விதிகளை

அடிப்படையாகக் கொண்ட கூறுகளைக் கொண்டுள்ளன. வடிவத் தொகுப்பு தொழில்நுட்பத்தை அடிப்படையாகக் கொண்ட பல ஒழுங்குமுறைகள் செயற்கையான, ரோபோ-ஒலிக்கும் பேச்சை உருவாக்குகின்றன (robotic-sounding speech); அவை மனித பேச்சாக ஒருபோதும் தவறாகக் கருதப்படாது. இருப்பினும், அதிகபட்ச இயல்பான தன்மை எப்போதும் பேச்சு தொகுப்பு அமைப்பின் குறிக்கோள் அல்ல; மேலும் வடிவத் தொகுப்பு ஒழுங்குமுறைகள் ஒன்றிணைக்கப்பட்ட ஒழுங்குமுறைகளை விட நன்மைகளைக் கொண்டுள்ளன. வடிவ தொகுக்கப்பட்ட பேச்சு நம்பத்தகுந்த வகையில் புரியக்கூடியது; மிக அதிக வேகத்தில் கூட, பொதுவாக ஒன்றிணைக்கப்பட்ட ஒழுங்குமுறைகளைப் பாதிக்கும் ஒலிக் குறைபாடுகளைத் தவிர்க்கிறது. ஸ்கிரீன் ரீடரைப் பயன்படுத்தி கணினிகளை விரைவாக வழிநடத்த பார்வை குறைபாடுள்ளவர்களால் அதிவேக ஒன்றிணைக்கப்பட்ட பேச்சு பயன்படுத்தப்படுகிறது. பேச்சுத் பதக்கூறுகளின் தரவுத்தளம் இல்லாததால், வடிவமைத்தல் ஒழுங்குமுறைகள் பொதுவாக ஒன்றிணைக்கப்பட்ட ஒழுங்குமுறைகளை விட சிறிய நிரல்களாகும். எனவே அவை நினைவகம் மற்றும் நுண்செயலி சக்தி (microprocessor power) குறிப்பாகக் குறைவாகவே இருக்கும் உட்பொதிக்கப்பட்ட ஒழுங்குமுறைகளில் (embedded systems) பயன்படுத்தப்படலாம்,. வடிவமைப்பு அடிப்படையிலான ஒழுங்குமுறைகள் வெளியீட்டு உரையின் அனைத்து பண்புக்கூறுகளையும் முழுமையான கட்டுப்பாட்டில் வைத்திருப்பதால், பலவிதமான மீக்க்கூகளும் இசையோட்டங்களும் வெளியீடாக இருக்கலாம், இது கேள்விகள் மற்றும் அறிக்கைகள் மட்டுமல்ல, பலவிதமான உணர்ச்சிகளையும் குரல்களின் சுரங்களையும் வெளிப்படுத்துகிறது.

நிகழ்நேரம் அல்லாத ஆனால் மிகவும் துல்லியமான இசையோட்டக் கட்டுப்பாட்டுக்கான எடுத்துக்காட்டுகள், 1970களின் பிற்பகுதியில் டெக்சாஸ் இன்ஸ்ட்ரூமென்ட்ஸ் பொம்மை ஸ்பீக் & ஸ்பெல் (Texas Instruments toy Speak & Spell) மற்றும் 1980களின் முற்பகுதியில் சேகா ஆர்கேட் மெஷின்கள் (Sega arcade machines) மற்றும் TMS5220 LPC சில்லுகளைப் பயன்படுத்துகின்ற பல அடாரி, இன்க் (Atari, Inc) ஆர்கேட் கேம்கள் ஆகியவற்றில் செய்யப்பட்ட வேலைகளை உட்படுத்தும். இந்தத் திட்டங்களுக்குச் சரியான இசையோட்டத்தை உருவாக்குவது கடினமானது, மேலும் விளைவுகள் நிகழ்நேர உரையிலிருந்து பேச்சு இடைமுகங்களால் இன்னும் பொருந்தவில்லை.

ஒலிப்புசார் ஒருங்கிணைப்பு (Articulatory synthesis)

ஒலிப்புசார் ஒருங்கிணைப்பு என்பது மனித குரல்வழியின் மாதிரிகள் மற்றும் அங்கு நிகழும் வெளிப்பாடு செயல்முறைகளின் அடிப்படையில் பேச்சை ஒருங்கிணைப்பதற்கான கணக்கீட்டு நுட்பங்களைக் குறிக்கிறது. ஆய்வகச் சோதனைகளுக்கு தவறாமல் பயன்படுத்தப்படும் முதல் உச்சரிப்புசார் ஒருங்கிணைப்பி (articulatory synthesizer) 1970களின் நடுப்பகுதியில் ஹாஸ்கின்ஸ் ஆய்வகங்களில் பிலிப் ரூபின், டாம் பேர் மற்றும் பால் மெர்மெல்ஸ்டீன் (Philip Rubin, Tom Baer, and Paul Mermelstein) ஆகியோரால் உருவாக்கப்பட்டது. ASY என அழைக்கப்படும் இந்த ஒருங்கிணைப்பி, 1960கள் மற்றும் 1970களில் பெல் ஆய்வகங்களில் பால் மெர்மெல்ஸ்டீன், சிசில் கோக்கர் மற்றும் சக ஊழியர்களால் (Paul Mermelstein, Cecil Coker, and colleagues) உருவாக்கப்பட்ட குரல் பாதை மாதிரிகளை அடிப்படையாகக் கொண்டது.

சமீப காலம் வரை ஒலிப்புசார் ஒருங்கிணைப்பு மாதிரிகள், வணிகரீதியான பேச்சு ஒருங்கிணைப்பு ஒழுங்குமுறைகளில் இணைக்கப்படவில்லை. ஒரு குறிப்பிடத்தக்க விதிவிலக்கு, அசல் ஆராய்ச்சி அதிகம் நடத்தப்பட்ட கல்கேரி பல்கலைக்கழகத்தின் (University of Calgary) ஸ்பின்-ஆஃப் நிறுவனமான ட்ரில்லியம் சவுண்ட் ரிசர்ச் (Trillium Sound Research) என்பதால் முதலில் உருவாக்கப்பட்டு விற்பனை செய்யப்பட்ட நெக்ஸ்டி அடிப்படையிலான ஒழுங்குமுறை (NeXT-based system) ஆகும். நெக்ஸ்டியின் பல்வேறு அவதாரங்களின் மறைவைத் தொடர்ந்து (1980களின் பிற்பகுதியில் ஸ்டீவ் ஜாப்ஸால் தொடங்கப்பட்டு 1997இல் ஆப்பிள் கம்ப்யூட்டருடன் இணைக்கப்பட்டது), ட்ரில்லியம் மென்பொருள் குனு/ஜி.என்.யு. பொதுவான பொதுமக்கள் உரிமத்தின் கீழ் வெளியிடப்பட்டது, இது தொடர்ந்து குனுஸ்பீச்சாக (Gnuspeech) இருந்தது. 1994ஆம் ஆண்டில் முதன்முதலில் சந்தைப்படுத்தப்பட்ட இந்த ஒழுங்குமுறை, காரின் "தனித்துவமான பிராந்திய மாதிரியால்" (Carré's "distinctive region model") கட்டுப்படுத்தப்படும் மனித வாய்வழி மற்றும் நாசிப் பாதைகளின் அலை வழிகாட்டி அல்லது பரிமாற்ற-வரி அனலாக் பயன்படுத்தி முழு சொற்பொருள் அடிப்படையிலான உரையிலிருந்து பேச்சு மாற்றத்தை வழங்குகிறது.

HMM- அடிப்படையிலான ஒருங்கிணைப்பு (HMM based synthesis)

HMM-அடிப்படையிலான ஒருங்கிணைப்பு என்பது மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகளை அடிப்படையாகக் கொண்ட ஒரு ஒருங்கிணைப்பு முறையாகும், இது புள்ளியியல்சார் அளவுரு ஒருங்கிணைப்பு (Statistical Parametric Synthesis) என்றும் அழைக்கப்படுகிறது. இந்த அமைப்பில், அதிர்வெண் ஸ்பெக்ட்ரம் (frequency spectrum) (குரல் பாதை/vocal tract),

அடிப்படை அதிர்வெண் (குரல் மூலம்) மற்றும் பேச்சின் காலம் (மீக்கூறு) ஆகியவை ஒரே நேரத்தில் எச்.எம்.எம். பேச்சு அலைவடிவங்கள் அதிகபட்ச வாய்ப்பு அளவுகோலின் அடிப்படையில் HMM களில் இருந்து உருவாக்கப்படுகின்றன (Statistical Parametric Synthesis). fundamental frequency (voice source).

6.6.4. வலை (இணையத்தளம்) அடிப்படையிலான கற்றல் ஒழுங்குமுறைகள் (Web-based learning systems)

இணைய அடிப்படையிலான கற்றல் பெரும்பாலும் ஆன்லைன் கற்றல் அல்லது எலார்னிங் (e-learning) என்று அழைக்கப்படுகிறது, ஏனெனில் இது ஆன்லைன் பாட உள்ளடக்கத்தை உள்ளடக்கியது. மின்னஞ்சல், வீடியோ கான்ஃபரன்சிங் மற்றும் நேரடி விரிவுரைகள் (வீடியோ ஸ்ட்ரீமிங்) வழியாக கலந்துரையாடல் மன்றங்கள் அனைத்தும் இணையம் வழியாக சாத்தியமாகும். இணைய அடிப்படையிலான படிப்புகள் அச்சிடப்பட்ட பாடப் பொருட்கள் போன்ற நிலையான பக்கங்களையும் வழங்கக்கூடும். பாடநெறிப் பொருட்களை அணுக இணையத்தைப் பயன்படுத்துவதன் மதிப்புகளில் ஒன்று, வலைப்பக்கங்கள் வலையின் பிற பகுதிகளுக்கான ஹைப்பர்லிங்க்களைக் கொண்டிருக்கக்கூடும், இதனால் ஏராளமான இணைய அடிப்படையிலான தகவல்களை அணுக முடியும். ஒரு "மெய்நிகர்" கற்றல் சூழல் (virtual” learning environment (VLE)) அல்லது நிர்வகிக்கப்பட்ட கற்றல் சூழல் (managed learning environment MLE) என்பது ஒரு கற்பித்தல் மற்றும் கற்றல் மென்பொருள் தொகுப்பில் உள்ளது. ஒரு "மெய்நிகர்" கற்றல் சூழல் பொதுவாக விவாத பலகைகள், அரட்டை அறைகள், ஆன்லைன் மதிப்பீடு, மாணவர்களின் வலையைப் பயன்படுத்துவதைக் கண்காணித்தல் மற்றும் பாடநெறி நிர்வாகம் போன்ற செயல்பாடுகளை ஒருங்கிணைக்கிறது. மெய்நிகர்" கற்றல் சூழல்கள் வேறு எந்த கற்றல் சூழலிலும் செயல்படுகின்றன, அவை கற்பவர்களுக்கு தகவல்களை விநியோகிக்கின்றன. மெய்நிகர்" கற்றல் சூழல்கள், எடுத்துக்காட்டாக, திட்டங்களில் ஒத்துழைக்கவும் தகவல்களைப் பகிரவும் கற்பவர்களுக்கு உதவும். இருப்பினும், இணைய அடிப்படையிலான படிப்புகளின் கவனம் எப்போதும் கற்பவர் மீது இருக்க வேண்டும் - தொழில்நுட்பம் பிரச்சினை அல்ல, அல்லது அவசியமான பதில் அல்ல (Javed Wasim et al, 2014).

இணைய அடிப்படையிலான கற்றலின் மாதிரிகள் வலை அடிப்படையிலான கற்றலை உருவாக்க மற்றும் வழங்க பல அணுகுமுறைகளைப் பயன்படுத்தலாம். இவற்றை தொடர்ச்சியாகக் காணலாம். ஒரு முனையில் “தூய்மையான” தொலைதூரக் கற்றல் (இதில்

பாடநெறி, மதிப்பீடு மற்றும் ஆதரவு அனைத்தும் ஆன்லைனில் வழங்கப்படுகின்றன, மாணவர்கள் மற்றும் ஆசிரியர்களிடையே நேருக்கு நேர் தொடர்பு இல்லாமல்). மறுமுனையில் ஒரு நிறுவன இன்ட்ராநெட் உள்ளது, இது ஒரு அச்சிடப்பட்ட பாடநெறி பொருட்களை ஆன்லைனில் நகலெடுக்கிறது. இருப்பினும், கற்றல், தகவல் தொடர்பு மற்றும் மதிப்பீட்டு நடவடிக்கைகளுக்கான இணைப்புகள் இல்லாமல், அறிவின் களஞ்சியங்களாக இருக்கும் வலைத்தளங்கள் கற்போரை மையமாகக் கொண்டவை அல்ல, அவை உண்மையான இணைய அடிப்படையிலான கற்றல் படிப்புகளாக கருதப்படாது.

ஒரு பொதுவான வலை அடிப்படையிலான பாடத்தின் பண்புக்கூறுகள்

• பாடநெறி தகவல், அறிவிப்பு பலகை, கால அட்டவணை

• பாடத்திட்ட வரைபடம், ஸ்லைடுகள், கையேடுகள், கட்டுரைகள் போன்ற கற்பித்தல் பொருட்கள், மின்னஞ்சல் மற்றும் விவாதப் பலகைகள் வழியாகத் தொடர்பு

• வடிவமைத்தல் மற்றும் சுருக்க மதிப்பீடுகள்

• மாணவர் மேலாண்மைக் கருவிகள் (பதிவுகள், புள்ளிவிவரங்கள், மாணவர் கண்காணிப்பு) பயனுள்ள உள் மற்றும் வெளிப்புற வலைத்தளங்களுக்கான இணைப்புகள் எடுத்துக்காட்டாக, நூலகம், ஆன்லைன் தரவுத்தளங்கள் மற்றும் பத்திரிகைகள்

வலை அடிப்படையிலான பாடத்திட்டத்தை வடிவமைப்பதற்கான முதல் படி கற்றவர்களின் தேவைகளை அடையாளம் காண்பது மற்றும் கற்பவர்கள் ஒரு குழுவின் ஒரு பகுதியாக அல்லது தனிப்பட்ட கற்பவர்களாகக் கருதப்பட வேண்டுமா?. தனிமைப்படுத்தப்பட்ட கற்பவர்களை "மெய்நிகர்" குழுக்களில் ஒன்றிணைக்க வலை ஒரு பயனுள்ள கருவியாக இருக்கலாம். எடுத்துக்காட்டாக, ஒரு விவாத மன்றத்தின் மூலம். இணைய அடிப்படையிலான கற்றல் திட்டங்களை எவ்வாறு வடிவமைப்பது என்பதில் பல ஆன்லைன் ஆதாரங்கள் உள்ளன.

ஒரு வலை அடிப்படையிலான கற்றல் திட்டத்தை தொடங்குவதற்கு முன் கேட்க வேண்டிய கேள்விகள்

- இணைய அடிப்படையிலான கற்றல் திட்டத்தின் கல்வி நோக்கம் என்ன?
- ஆன்லைன் கற்றல் பாடநெறிக்கு அல்லது மாணவர்களுக்கு என்ன கூடுதல் மதிப்பைக் கொடுக்கும்?
- இணைய அடிப்படையிலான கற்றல் குறித்த எந்த வளங்களும் நிபுணத்துவமும் நிறுவனத்தில் உள்ளன?

- திட்டமிடப்பட்ட பாடத்திட்டத்தை சக ஊழியர்களும் நிறுவனமும் அறிந்திருக்கிறார்களா? (நீங்கள் முயற்சியை நகலெடுப்பதைத் தவிர்க்க வேண்டும் மற்றும் நிறுவனத்தின் கணினி அமைப்பு பாடத்திட்டத்தை ஆதரிக்க முடியும் என்பதை உறுதிப்படுத்திக் கொள்ளுங்கள்).
- ஆரம்பகால வளர்ச்சியின் பின்னர் இருக்கும் கற்பித்தல் வளங்கள் மற்றும் தற்போதைய பராமரிப்பு செலவுகளை இந்த திட்டம் கணக்கில் எடுத்துக்கொண்டதா?
- பொருட்களை (Materials) உருவாக்க அல்லது மறுவடிவமைக்க போதுமான நேரத்தை நீங்கள் அனுமதித்தீர்களா?
- இணைய அடிப்படையிலான கற்றல் படிப்புகளின் குறிப்பிட்ட வடிவமைப்பு மற்றும் மாணவர் ஆதரவு தேவைகள் கணக்கில் எடுத்துக் கொள்ளப்பட்டுள்ளதா?

ஒருங்கிணைந்த திட்டங்களுக்கு வலை அடிப்படையிலான கற்றல் ஒரு நிறுவனத்தில் வலை அடிப்படையிலான கற்றல் பெரும்பாலும் வழக்கமான, நேருக்கு நேர் கற்பித்தலுடன் ஒருங்கிணைக்கப்படுகிறது. இது பொதுவாக ஒரு இன்ட்ராநெட் வழியாகச் செய்யப்படுகிறது; இது வழக்கமாக “கடவுச்சொல் வழி பாதுகாக்கப்படுகிறது” மற்றும் பதிவுசெய்யப்பட்ட பயனர்களுக்கு மட்டுமே அணுகக்கூடியது. இதனால் ஆன்லைன் பொருட்களின் அறிவுசார் சொத்துக்களைப் பாதுகாக்கவும், மாணவர்களிடையே ரகசிய தகவல்தொடர்பு பரிமாற்றத்தை ஆதரிக்கவும் முடியும். அடிப்படை அறிவியல் மற்றும் மருத்துவ கற்பித்தல் இரண்டிலும் மருத்துவம் ஆன்லைன் கற்றலுக்கு பல எடுத்துக்காட்டுகளைக் கொண்டுள்ளது. மாணவர்கள் பொதுவாக அடிப்படை அறிவியல் கற்பிப்பதற்காக பெரிய குழுக்களாக இருப்பதால், வழக்கமான திட்டங்களை நிறைவு செய்வதற்கும் சுய மதிப்பீட்டை இயக்குவதற்கும் கற்றல் பொருட்களை வழங்க வலை அடிப்படையிலான கற்றல் பயன்படுத்தப்படலாம். எடுத்துக்காட்டாக, நோயியல் படிப்புகள் கற்பிப்பதற்கான உடற்கூறியல் தளங்கள் மற்றும் பட வங்கிகளுக்கான அணுகல். கற்பவர்கள் புவியியல் ரீதியாக பிரிக்கப்படும் போது மருத்துவ கற்பித்தலை ஆதரிக்க வலை அடிப்படையிலான கற்றல் பயனுள்ளதாக இருக்கும். எடுத்துக்காட்டாக, வீடியோ செயல்விளக்கங்கள் மூலம் மருத்துவ திறன்களைக் கற்றுக்கொள்ள. கீழ்வரும் அட்டவணை பாரம்பரிய கற்றல் மற்றும் வலை அடிப்படையிலான கற்றல் ஆகியவற்றுக்கு இடையிலான வேறுபாட்டை நிரூபிக்கிறது.

பாரம்பரிய கற்றல் மற்றும் வலை அடிப்படையிலான கற்றல் ஆகியவற்றுக்கு இடையிலான வேறுபாடு

பாரம்பரிய கற்றல் வலை	மின் கற்றல் (ஐடியைப் பயன்படுத்துதல்)
ஆசிரியர்களை மையமாகக் கொண்ட அறிவுறுத்தல்	மாணவர்களை மையமாகக் கொண்ட அறிவுறுத்தல்
ஒற்றை உணர்வு தூண்டுதல்	மல்டிசென்சரி தூண்டுதல்
ஒற்றை பாதை முன்னேற்றம்	மல்டிபாத் முன்னேற்றம்
ஒற்றை ஊடகம்	மல்டிமீடியா
தனிமைப்படுத்தப்பட்ட வேலை	கூட்டு வேலை
தகவல் விநியோகம்	தகவல் பரிமாற்றம்
செயலற்ற கற்றல்	/செயல்/ஆய்வு/விசாரணை அடிப்படையிலான கற்றல்
உண்மை, அறிவு சார்ந்த கற்றல்	விமர்சன சிந்தனை மற்றும் தகவலறிந்த முடிவெடுதல்
தனிமைப்படுத்தப்பட்ட, செயற்கை சூழல்	உண்மையான, நிஜ உலக சூழல்

இணைய அடிப்படையிலான நிரல்களை வடிவமைக்கும்போது (எந்தவொரு கற்றல் திட்டத்தையும் போல), கற்பவர்களின் தேவைகளும் அனுபவமும் கணக்கில் எடுத்துக்கொள்ளப்பட வேண்டும். இணைய அடிப்படையிலான அல்லது ஆன்லைன் கற்றலில் இருந்து சிறந்ததைப் பெற பொருத்தமான தொழில்நுட்பமும் நியாயமான கணினி திறன்களும் தேவை. வெவ்வேறு தொழில்நுட்ப விவரக்குறிப்புகள் மற்றும் மென்பொருளின் பதிப்புகளுக்கு இடமளிக்கும் வகையில் நிகழ்ச்சிகள் மற்றும் வலைப்பக்கங்களை வடிவமைக்க முடியும். இருப்பினும், கற்றவர்களுக்கு அவர்கள் மெதுவான அணுகலுடன் இணையத்தில் வேலை செய்ய முயற்சிக்கிறார்களோ அல்லது அவர்களுக்குத் தேவையான படங்களையும் வீடியோக்களையும் பதிவிறக்கம் செய்ய முடியாவிட்டால் அது வெறுப்பாக இருக்கிறது. மறுபுறம், இணைய அடிப்படையிலான நிரல்கள், மேலும் சுதந்திரமான மற்றும் சுறுசுறுப்பான கற்றலை ஊக்குவிக்கக்கூடும், மேலும் அவை பெரும்பாலும் பாடப்பொருட்களை வழங்குவதற்கான திறமையான வழிமுறையாகும். தகவல் அடிப்படையிலான விநியோகம், தகவல் தொடர்பு, ஊடாடும் திறன், புவியியல் சுதந்திரம், தற்காலிக சுதந்திரம் ஆகியவற்றை ஒருங்கிணைப்பதன் மூலம் வலை அடிப்படையிலான திட்டங்கள் கற்பித்தல் மற்றும் கற்றலை மேம்படுத்த இயலும் (McCormack & D Jones, 1998).

முடிவுரை

இணையத்தின் விரைவான விரிவாக்கம் மற்றும் மென்பொருள் திறன்களை அதிகரிப்பது பல்வேறு நிலைகளில் கற்பித்தல் மற்றும் கற்றலின் இயக்கவியலை பாதிக்கிறது. வலை

அடிப்படையிலான கற்றல் கருவிகள் டெவலப்பர்களால் அவற்றின் செயல்திறனை மேம்படுத்துவதற்காக தொடர்ந்து வடிவமைக்கப்படுகின்றன. வெப்சிடி (WebCT) and மற்றும் பிளாக்போர்டு (Blackboard) இரண்டுமே படிப்பிற்குப் பயன்படுத்தப்பட்டதை விட அவற்றின் பாடநெறி கருவிகளின் புதிய பதிப்புகளைக் கொண்டுள்ளன. சில ஆய்வின் முடிவுகளில் விளக்கப்பட்டுள்ளபடி, கருவியின் வடிவமைப்பு மற்றும் கருவியின் வடிவமைப்பு உள்ளிட்ட பல்வேறு காரணிகளைப் பொறுத்து, கருவியின் பயனும் செயல்திறனும் சூழல் சார்ந்ததாகும். மேலும் கருவி மேம்பாட்டிற்கான உள்ளீட்டை வழங்க மாணவர்கள் போன்ற 'உண்மையான' பயனர்களிடமிருந்து வரும் கருத்து முக்கியமானது. துரதிர்ஷ்டவசமாக, கல்வி நிறுவனங்களில் இந்த கருவிகளின் பயனர்கள் இந்த செயல்பாட்டில் அரிதாகவே சேர்க்கப்படுகிறார்கள் (Storey et al).

வலை அடிப்படையிலான கற்றல் கற்றல் மற்றும் பரந்த அளவிலான அறிவு மற்றும் தகவல்களை அணுகுவதற்கான பெரிய வாய்ப்புகளை வழங்குகிறது. வழங்கப்பட்ட கற்றல் சூழல் கற்பவர்களின் தேவைகளை கணக்கில் எடுத்துக்கொள்வதையும், அவை திறம்பட தயாரிக்கப்பட்டு ஆதரிக்கப்படுவதையும் உறுதி செய்வதே ஆசிரியர்களின் பங்கு. ஆன்லைன் கற்றலுக்கு நன்மைகள் உள்ளன, ஆனால் இணைய அடிப்படையிலான கற்றல் எப்போதும் தேர்வு செய்யும் முறையாக பார்க்கப்படக்கூடாது, ஏனெனில் தடைகள் (போதிய உபகரணங்கள் போன்றவை) மாணவர்களை கற்றலில் இருந்து எளிதில் விலகிவிடும். எனவே தொழில்நுட்பம் சரியான முறையில் பயன்படுத்தப்பட வேண்டும்; அது கிடைக்கிறது மற்றும் புதியது அல்லது மாணவர்கள் மற்றும் ஆசிரியர்கள் நிச்சயமாக வழங்குவதற்கான இந்த வழிமுறைகளில் குறிப்பிட்ட எதிர்பார்ப்புகளைக் கொண்டிருக்கிறது என்பதால் மட்டும் பயன்படுத்தப்படக்கூடாது (Storey et al). கணினிகள் மற்றும் வீடியோ கான்பரன்சிங் போன்ற புதிய தொழில்நுட்பங்கள் பழைய தொழில்நுட்பங்களை விட கற்பித்தல் அல்லது கற்றலுக்கு சிறந்தவை (அல்லது மோசமானவை) அல்ல. அவை வேறுபட்டவை. தொழில்நுட்பத்தின் தேர்வு கற்றவர்களின் தேவைகள் மற்றும் நாம் பணிபுரியும் சூழலால் இயக்கப்பட வேண்டும், அதன் புதுமையால் அல்ல (Bates 1995).

6.6.5. கேள்வி-பதில் ஒழுங்குமுறைகள் (Question-answering systems)

கேள்வி பதில் (Question answering (QA)) என்பது தகவல் மீட்டெடுப்பு மற்றும் இயற்கை மொழிச் செயலாக்கம் (natural language processing (NLP/என்.எல்.பி) ஆகிய துறைகளில் உள்ள ஒரு கணினி அறிவியல் துறையாகும்; இது இயற்கை மொழியில் மனிதர்கள் எழுப்பும்

கேள்விகளுக்குத் தானாகவே பதிலளிக்கும் கட்டமைப்புகளில் அக்கறை கொண்டுள்ளது (Philipp Cimiano et al 2014).

கண்ணோட்டம்

கேள்விப்பதில் செயல்படுத்தல், பொதுவாக ஒரு கணினி நிரல், அறிவு அல்லது தகவலின் கட்டமைக்கப்பட்ட தரவுத்தளத்தை (பொதுவாக அறிவுத்தளதை) வினவுவதன் மூலம் அதன் பதில்களை உருவாக்கலாம். மிகவும் பொதுவாக, கேள்வி பதிலளிக்கும் அமைப்புகள் இயற்கையான மொழியின் கட்டமைக்கப்படாத தொகுப்பிலிருந்து பதில்களை எடுக்கலாம் (இது பதிப்புரிமை)

கேள்வி பதில் அமைப்புகளுக்குப் பயன்படுத்தப்படும் இயற்கை மொழி ஆவணச் சேகரிப்புகளின் சில எடுத்துக்காட்டுகள் பின்வருமாறு:

- நோக்கீடு உரைகளின் வட்டாரத் தொகுப்பு
- உள் அமைப்பு ஆவணங்கள் மற்றும் வலைப்பக்கங்கள்
- தொகுக்கப்பட்ட நியூஸ்வைர் (newswire) அறிக்கைகள்
- விக்கிபீடியா (Wikipedia) பக்கங்களின் தொகுப்பு
- உலகளாவிய வலைப்பக்கங்களின் துணைத் தொகுப்பு

கேள்வி பதில் ஆராய்ச்சி பல வகையான கேள்விகளைக் கையாள முயற்சிக்கிறது: உண்மை, பட்டியல், வரையறை, எப்படி, ஏன், கற்பனையான, பொருண்மையியல் அடிப்படையில் கட்டுப்படுத்தப்பட்ட மற்றும் குறுக்கு மொழி/மொழிகடந்த கேள்விகள்.

மூடிய-களக் (Closed-domain) கேள்வி பதில் ஒரு குறிப்பிட்ட களத்தின் கீழ் உள்ள கேள்விகளைக் கையாளுகிறது (எடுத்துக்காட்டாக, மருத்துவம் அல்லது வாகன பராமரிப்பு), மேலும் களம்-குறிப்பிட்ட அறிவை அடிக்கடி ஆன்டாலஜிஸில் முறைப்படுத்தலாம். மாற்றாக, மூடிய-களம் ஒரு குறிப்பிட்ட வகைக் கேள்விகளை மட்டுமே ஏற்றுக் கொள்ளும் சூழ்நிலையைக் குறிக்கலாம், அதாவது நடைமுறை தகவல்களைக் காட்டிலும் விளக்கமாகக் கேட்கும் கேள்விகள். இயந்திர வாசிப்பு பயன்பாடுகளின் சூழலில் கேள்வி பதிலளிக்கும் அமைப்புகள் மருத்துவ களத்திலும் கட்டப்பட்டுள்ளன, எடுத்துக்காட்டாக அல்சைமர் நோய் (Alzheimers disease) தொடர்பானவை (Roser Morante et al 2014)

திறந்த-களக் (Open-domain) கேள்வி பதில் கிட்டத்தட்ட எதையும் பற்றிய கேள்விகளைக் கையாளுகிறது, மேலும் பொதுவான இருப்பியல்கள் (ontologies) மற்றும் உலக அறிவை மட்டுமே

சார்ந்திருக்க இயலும். மறுபுறம், இந்த அமைப்புகள் வழக்கமாகப் பதிலைப் பிரித்தெடுப்பதற்கு அதிகமான தரவுகளைக் கொண்டுள்ளன.

பன்மாதிரி/மல்டிமாடல் (Multimodal) கேள்வி பதில் உரை மற்றும் படங்கள் போன்ற கேள்விகளுக்கு பதிலளிக்க பயனர் உள்ளீட்டின் பல முறைகளைப் பயன்படுத்துகிறது (Mittal et al. 2011).

வரலாறு

இரண்டு ஆரம்ப கேள்வி பதில் அமைப்புகள் பேஸ்பால் BASEBALL (GREEN JR, Bert F; et al.) (1961)மற்றும் லுனார் LUNAR (Woods & Kaplan (1977) என்பன ஆகும். ஒரு ஆண்டுக் காலப்பகுதியில் அமெரிக்க பேஸ்பால் லீக் குறித்த கேள்விகளுக்கு BASEBALL/பேஸ்பால் பதிலளித்தது. அப்பல்லோ நிலவு பயணங்க மூலம் (Apollo moon missions) கிடைக்கப்பெற்ற பாறைகளின் புவியியல் பகுப்பாய்வு குறித்த கேள்விகளுக்கு LUNAR/லூனார் பதிலளித்தது. இரண்டு கேள்விப் பதிலல் ஒழுங்கமைப்புகளும் அவை தேர்ந்தெடுத்த களங்களில் மிகவும் பயனுள்ளதாக இருந்தன. உண்மையில், 1971ஆம் ஆண்டில் சந்திர அறிவியல் மாநாட்டில் (lunar science convention) LUNAR/லுனார் செயல்விளக்கம் செய்யப்பட்டது, மேலும் அதன் களத்தில் 90% கேள்விகளுக்கு பதிலளிக்க முடிந்தது. மேலும் கட்டுப்படுத்தப்பட்ட-களக் கேள்வி பதில் அமைப்புகள் அடுத்த ஆண்டுகளில் உருவாக்கப்பட்டன. இந்த எல்லா அமைப்புகளின் பொதுவான அம்சம் என்னவென்றால், அவை ஒரு முக்கிய தரவுத்தளம் அல்லது அறிவு அமைப்பைக் கொண்டிருந்தன, அவை தேர்ந்தெடுக்கப்பட்ட களத்தின் நிபுணர்களால் கையால் எழுதப்பட்டன. BASEBALL மற்றும் LUNARஇன் மொழித் திறன்கள் முதல் சாட்டர்போட் திட்டங்களான எலிசா/ELIZA மற்றும் டாக்டர்/DOCTOR)போன்ற நுட்பங்களைப் பயன்படுத்தின.

SHRDLU என்பது 1960களின் பிற்பகுதியிலும் 1970களின் முற்பகுதியிலும் டெர்ரி வினோகிராட் (Terry Winograd) உருவாக்கிய மிக வெற்றிகரமான கேள்வி பதில் திட்டமாகும். இது ஒரு பொம்மை உலகில் ("தொகுதிகள் உலகம்"/"blocks world") ஒரு ரோபோவின் செயல்பாட்டை உருவகப்படுத்தியது, மேலும் இது உலகின் நிலை குறித்து ரோபோ கேள்விகளைக் கேட்கும் வாய்ப்பை வழங்கியது. மீண்டும், இந்த அமைப்பின் வலிமை ஒரு குறிப்பிட்ட டொமைன் மற்றும் கணினி நிரலில் குறியாக்க எளிதான இயற்பியல் விதிகளைக் கொண்ட மிக எளிய உலகத்தைத் தேர்ந்தெடுப்பதாகும்.

1970களில், அறிவின் குறுகிய களங்களை குறிவைக்கும் அறிவுத் தளங்கள் உருவாக்கப்பட்டன. இந்த நிபுணர் ஒழுங்குமுறைகளுடன் (expert systems) இடைமுகமாக உருவாக்கப்பட்ட கேள்வி பதில் ஒழுங்குமுறைகள் அறிவின் ஒரு பகுதிக்குள் உள்ள கேள்விகளுக்கு மீண்டும் மீண்டும் செய்யக்கூடிய மற்றும் சரியான பதில்களை உருவாக்கியது. இந்த நிபுணர் அமைப்புகள் அவற்றின் உள் கட்டமைப்பைத் தவிர நவீன கேள்வி பதில் அமைப்புகளை நெருக்கமாக ஒத்திருந்தன. நிபுணர் ஒழுங்குமுறைகள் நிபுணர்களால் கட்டமைக்கப்பட்ட மற்றும் ஒழுங்கமைக்கப்பட்ட அறிவுத் தளங்களைப் பெரிதும் நம்பியுள்ளன; அதேசமயம் பல நவீன கேள்வி பதில் ஒழுங்குமுறைகள் ஒரு பெரிய, கட்டமைக்கப்படாத, இயற்கை மொழி உரை தரவுத்தொகுதியின் புள்ளிவிவர செயலாக்கத்தை நம்பியுள்ளன.

1970கள் மற்றும் 1980களில் கணக்கீட்டு மொழியியலில் விரிவான கோட்பாடுகளின் வளர்ச்சியைக் கண்டது, இது உரை புரிந்துகொள்ளுதல் மற்றும் கேள்வி பதிலளிப்பதில் லட்சிய திட்டங்களை உருவாக்க வழிவகுத்தது. அத்தகைய ஒழுங்குமுறையின் ஒரு எடுத்துக்காட்டு 1980களின் பிற்பகுதியில் யுனிக்ஸ் ஆலோசகர் (Unix Consultant (UC/யு.சி), ராபர்ட் விலென்ஸ்கி (Robert Wilensky) யு.சி. பெர்கிலியில் (U.C. Berkeley) உருவாக்கியது.. கணினி யூனிக்ஸ் இயக்க முறைமை தொடர்பான கேள்விகளுக்கு பதிலளித்தது. இது அதன் களத்தின் விரிவான கையால் வடிவமைக்கப்பட்ட அறிவுத் தளத்தைக் கொண்டிருந்தது, மேலும் இது பல்வேறு வகையான பயனர்களுக்கு இடமளிக்கும் வகையில் பதிலை வடிவமைப்பதை நோக்கமாகக் கொண்டது. மற்றொரு திட்டம் ஒரு ஜெர்மன் நகரத்தில் சுற்றுலா தகவல்களின் களத்தில் இயங்கும் உரை புரிந்துகொள்ளும் ஒழுங்குமுறை LILOG/ லிலோக் ஆகும். யு.சி மற்றும் லிலோக் திட்டங்களில் உருவாக்கப்பட்ட ஒழுங்குமுறைகள் ஒருபோதும் எளிய செயல்விளக்கங்களின் கட்டத்தை கடந்ததில்லை, ஆனால் அவை கணினி மொழியியல் மற்றும் பகுத்தறிவு பற்றிய கோட்பாடுகளின் வளர்ச்சிக்கு உதவின.

உடல்நலம் மற்றும் வாழ்க்கை விஞ்ஞானிகளுக்கான EAGLi போன்ற சிறப்பு இயற்கை மொழி கேள்வி பதில் அமைப்புகள் உருவாக்கப்பட்டுள்ளன, மேலும் வோல்ஃப்ராம்|ஆல்ஃபா (Wolfram|Alpha) எனப்படும் ஒரு ஆன்லைன் கணக்கீட்டு அறிவு இயந்திரம், வெளிப்புற ஆதாரங்களில் இருந்து தரவுகளைக் கணக்கிடுவதன் மூலம் உண்மை கேள்விகளுக்கு நேரடியாக பதிலளிக்கும்.

கட்டமைப்பு

2001 ஆம் ஆண்டு நிலவரப்படி, கேள்வி பதில் அமைப்புகளில் பொதுவாக கேள்வி வகை மற்றும் பதில் வகையை தீர்மானிக்கும் கேள்வி வகைப்படுத்தல் தொகுதி அடங்கும் (Hirschman & Gaizauskas 2001). ஒரு பன்முக கேள்வி-பதில் கட்டமைப்பு முன்மொழியப்பட்டது; இதில் ஒவ்வொரு களமும் ஒரு குறிப்பிட்ட முகவரியால் குறிப்பிடப்படுகின்றன, இது அதன் குறிப்பிட்ட அறிவைக் கணக்கில் எடுத்துக்கொண்டு கேள்விகளுக்குப் பதிலளிக்க முயற்சிக்கிறது; ஒரு மெட்டா-முகவர் (meta–agent) கேள்வி பதிலளிக்கும் முகவர்களுக்கு இடையிலான ஒத்துழைப்பைக் கட்டுப்படுத்துகிறது மற்றும் மிகவும் பொருத்தமான பதிலைத் (பதில்களை) தேர்வு செய்கிறது (Galitsky & Pampapathi 2005)

கேள்வி பதிலளிக்கும் முறைகள் (Question answering methods)

கேள்விக்கு பதிலளிப்பது ஒரு நல்ல தேடல் தரவுத்தொகுதியைச் (search corpus) சார்ந்துள்ளது - ஏனென்றால் பதிலைக் கொண்ட ஆவணங்கள் இல்லாமல், எந்தவொரு கேள்விக்கும் பதிலளிக்கும் ஒழுங்குமுறையையும் உருவாக்க இயலாது. கேள்விக் களம் சேகரிப்புக்கு ஆர்த்தோகனலாக இருந்தால் மட்டுமே பெரிய சேகரிப்பு அளவுகள் பொதுவாக சிறந்த கேள்விக்கு பதிலளிக்கும் செயல்திறனைக் கொடுக்கும் என்பதை இது அர்த்தப்படுத்துகிறது. வலை போன்ற பெரும் அளவிலான சேகரிப்புகளில் தரவு மிகைமை என்ற கருத்து, மாறுபட்ட சூழல்களிலும் ஆவணங்களிலும் சீராக்கப்படாத தகவல்களைப் பல வழிகளில் வடிவமைக்க வாய்ப்புள்ளது; இது இரண்டு நன்மைகளுக்கு வழிவகுக்கிறது (Lin 2002):

1. சரியான தகவல்கள் பல வடிவங்களில் தோன்றுவதன் மூலம், உரையைப் புரிந்துகொள்ள சிக்கலான என்.எல்.பி நுட்பங்களைச் செய்வதற்கான கேள்வி பதில் அமைப்பின் சுமை குறைகிறது.
2. தவறான பதில்களை விட ஆவணங்களில் அதிக முறை தோன்றுவதற்குச் சரியான பதிலை நம்புவதன் மூலம் சரியான பதில்களை தவறான நேர்மறைகளிலிருந்து வடிகட்டலாம்.

சில கேள்வி பதிலளிக்கும் ஒழுங்குமுறைகள் தானியங்கி பகுத்தறிவைப் (automated reasoning) பெரிதும் நம்பியுள்ளன (Moldovan et al. 2003; Furbach et al. 2010). பல கேள்வி பதில் ஒழுங்குமுறைகள் செயற்கை நுண்ணறிவுடன் தொடர்புடைய தர்க்க நிரலாக்க மொழியான (logic programming language) புரோலாகில் (Galitsky, Boris 2003) வடிவமைக்கப்பட்டவை ஆகும்.

திறந்த களக் கேள்வி பதிலளிப்பு

தகவல் மீட்டெடுப்பில், திறந்த டொமைன் கேள்வி பதிலளிக்கும் அமைப்பு பயனரின் கேள்விக்கு பதிலளிக்கும் வகையில் பதிலைத் திருப்புவதை நோக்கமாகக் கொண்டுள்ளது. திரும்பிய பதில் தொடர்புடைய ஆவணங்களின் பட்டியலைக் காட்டிலும் குறுகிய நூல்களின் வடிவத்தில் உள்ளது (Sun et al.). கணினி மொழியியல், தகவல் மீட்டெடுப்பு மற்றும் பதில்களைக் கண்டுபிடிப்பதற்கான அறிவு உருப்படுத்தம் ஆகியவற்றின் நுட்பங்களின் கலவையைப் பயன்படுத்துகிறது.

கணினி ஒரு முக்கியச் சொற் குழுமத்தைக் (set of keywords) காட்டிலும் ஒரு முக்கிய மொழி கேள்வியை உள்ளீடாக எடுத்துக்கொள்கிறது, எடுத்துக்காட்டாக, "When is the national day of China?"/"சீனாவின் தேசிய நாள் எப்போது?" வாக்கியம் அதன் தர்க்க வடிவத்தின் மூலம் வினவலாக மாற்றப்படுகிறது. இயற்கை மொழிக் கேள்வியின் வடிவத்தில் உள்ளீட்டைக் கொண்டிருப்பது கணினியை அதிகப் பயனர் நட்புடன் ஆக்குகிறது; ஆனால் செயல்படுத்த கடினமாக உள்ளது, ஏனெனில் பல்வேறு கேள்வி வகைகள் உள்ளன, மேலும் விவேகமான பதிலைக் கொடுக்க கணினி சரியான ஒன்றை அடையாளம் காண வேண்டும். கேள்விக்கு ஒரு கேள்வி வகையை ஒதுக்குவது ஒரு முக்கியமான பணியாகும்; முழு பதில் பிரித்தெடுக்கும் செயல்முறையும் சரியான பதில் வகை கிடைக்கும் சரியான கேள்வி வகையைக் கண்டுபிடிப்பதை நம்பியுள்ளது.

உள்ளீட்டு கேள்வி வகையை அடையாளம் காண்பதற்கான முதல் படி முக்கிய சொல் பிரித்தெடுத்தல் (Keyword extraction) (Harabagiu & Hickl 2006) ஆகும். சில சந்தர்ப்பங்களில், கேள்வி வகையை நேரடியாகக் குறிக்கும் தெளிவான சொற்கள் உள்ளன. அதாவது "யார்", "எங்கே" அல்லது "எத்தனை", இந்தச் சொற்கள் முறையே பதில்கள் முறையே "நபர்", "இருப்பிடம்", "எண்" வகையாக இருக்க வேண்டும் என்று கூறுகின்றன. மேலே உள்ள எடுத்துக்காட்டில், "எப்போது" என்ற சொல் பதில் "தேதி" வகையாக இருக்க வேண்டும் என்பதைக் குறிக்கிறது. பதில் வகையைத் தீர்மானிக்கச் சொல்வகை அடையாளப்படுத்தல் (part-of-speech/POS tagging) மற்றும் தொடரியல் பாகுபடுத்தல் (syntactic parsing) நுட்பங்களையும் பயன்படுத்தலாம். இந்த வழக்கில், பொருள் "சீன தேசிய தினம்", ஆங்கில வாக்கியத்தில் பயனிலை"is" ஆகும்; "எப்போது" வினையடை ஆகும்; எனவே பதில் வகை "தேதி" வகை ஆகும். துரதிர்ஷ்டவசமாக, "எது", "என்ன" அல்லது "எப்படி" போன்ற சில கேள்விக்குரிய சொற்கள் தெளிவான பதில் வகைகளைத் தருவதில்லை. இந்த சொற்கள் ஒவ்வொன்றும் ஒன்றுக்கு மேற்பட்ட வகைகளைக் குறிக்கும். இது

போன்ற சூழ்நிலைகளில், கேள்வியின் பிற சொற்களைக் கருத்தில் கொள்ள வேண்டும். முதலில் செய்ய வேண்டியது கேள்வியின் பொருளைக் குறிக்கக்கூடிய சொற்களைக் கண்டுபிடிப்பதுதான். வேர்ட்நெட்/சொல்வலை போன்ற ஒரு சொல்சார் அகராதி பின்னர் சூழலைப் புரிந்துகொள்ள பயன்படுத்தப்படலாம்.

கேள்வி வகை அடையாளம் காணப்பட்டதும், சரியான முக்கியச் சொற்களைக் கொண்ட ஆவணங்களின் தொகுப்பைக் கண்டுபிடிக்க ஒரு தகவல் மீட்டெடுப்பு ஒழுங்குமுறை (information retrieval system) பயன்படுத்தப்படுகிறது. கண்டுபிடிக்கப்பட்ட ஆவணங்களில் சரியான நிறுவனங்கள் மற்றும் உறவுகள் குறிப்பிடப்பட்டுள்ளதா என்பதை சரிபார்க்க ஒரு அடையாளப்படுத்தி/டேகர் மற்றும் பெயர்த்தொடர் (NP) வினைக்குழு (Verb Group) கூறாக்கியைப் (chunker) பயன்படுத்தலாம். "யார்" அல்லது "எங்கே" போன்ற கேள்விகளுக்கு, மீட்டெடுக்கப்பட்ட ஆவணங்களிலிருந்து பொருத்தமான "நபர்" மற்றும் "இருப்பிடம்" பெயர்களைக் கண்டுபிடிக்க பெயரிடப்பட்ட-உருப்பொருள் உணரி (named-entity recognizer) பயன்படுத்தப்படுகிறது. தரவரிசைக்கு தொடர்புடைய பத்திகள் மட்டுமே தேர்ந்தெடுக்கப்படுகின்றன.

ஒரு திசையன் இடைவெளி மாதிரியை (vector space model) தேர்வுக்குரியப் பதில்களை வகைப்படுத்துவதற்கான ஒரு மூலோபாயமாகப் பயன்படுத்தலாம். கேள்வி வகை பகுப்பாய்வு கட்டத்தில் தீர்மானிக்கப்பட்டபடி பதில் சரியான வகையா என்று பரிசோதிக்கவும். தேர்வுக்குரியப் பதில்களைச் சரிபார்க்க ஒரு அனுமான நுட்பத்தையும் பயன்படுத்தலாம். இந்த தேர்வுக்குரியவைகள் ஒவ்வொன்றிற்கும் அதில் உள்ள கேள்விச் சொற்களின் எண்ணிக்கையைப் பொறுத்தும் இந்த சொற்கள் தேர்வுக்குரியவைகளுடன் எவ்வளவு நெருக்கமாக இருக்கின்றன என்பதைப் பொறுத்தும் ஒரு மதிப்பெண் வழங்கப்படுகிறது; மேலும் மேலும் நெருக்கமாக இருப்பது சிறப்பாக அமையும். பாகுபடுத்துவதன் மூலம் பதில் ஒரு சிறிய மற்றும் அர்த்தமுள்ள உருப்படுத்தமாக மொழிபெயர்க்கப்படுகிறது. முந்தைய எடுத்துக்காட்டில், எதிர்பார்க்கப்படும் வெளியீட்டுப் பதில் "1 அக்டோபர்" என்பதாகும்.

கணித கேள்வி பதில்

அஸ்கி பிளாட்டிபஸ் மற்றும் விக்கிடேட்டாவை அடிப்படையாகக் கொண்ட ஒரு திறந்த மூல கணித-விழிப்புணர்வு கேள்வி பதில் அமைப்பு 2018இல் வெளியிடப்பட்டது. [14] கணினி ஒரு ஆங்கில அல்லது இந்தி இயற்கை மொழி கேள்வியை உள்ளீடாக எடுத்து விக்கிடேட்டாவிலிருந்து

பெறப்பட்ட கணித சூத்திரத்தை சுருக்கமான பதிலாக வழங்குகிறது. இதன் விளைவாக வரும் சூத்திரம் கணக்கிடக்கூடிய வடிவத்தில் மொழிபெயர்க்கப்பட்டுள்ளது, இது பயனர்களுக்கு மாறிகள் மதிப்புகளைச் செருக அனுமதிக்கிறது. மாறிகள் மற்றும் பொதுவான மாறிலிகளின் பெயர்கள் மற்றும் மதிப்புகள் விக்கிடேட்டாவிலிருந்து கிடைத்தால் மீட்டெடுக்கப்படுகின்றன. இந்த அமைப்பு ஒரு சோதனை தொகுப்பில் வணிக கணக்கீட்டு கணித அறிவு இயந்திரத்தை விஞ்சிவிடும் என்று கூறப்படுகிறது.

முன்னேற்றம்

அறிவின் கூடுதல் களங்களை உள்ளடக்குவதற்காகக் கேள்வி பதில் ஒழுங்குமுறைகள் சமீபத்திய ஆண்டுகளில் நீட்டிக்கப்பட்டுள்ளன (Paşca, Marius 2005). எடுத்துக்காட்டாக, தற்காலிக மற்றும் புவியியல் கேள்விகள், வரையறை மற்றும் கலைச் சொற்களின் கேள்விகள், சுயசரிதை கேள்விகள், பன்மொழிக் கேள்விகள் மற்றும் உள்ளடக்கம் பற்றிய கேள்விகளுக்கு தானாகவே பதிலளிக்க ஒழுங்குமுறைகள் உருவாக்கப்பட்டுள்ளன. ஆடியோ, படங்கள் (Anderson et al. 2018), மற்றும் வீடியோ (Zhu et al. 2017) ஆராய்ச்சித் தலைப்புக்களுக்கான கேள்விப் பதில்கள் பின்வருவற்றை உட்படுத்தும்:

- ஊடாடுதல் - கேள்விகள் அல்லது பதில்களை தெளிவுபடுத்துதல் (Quarteroni and Suresh Manandhar 2009)
- பதில் மறுபயன்பாடு அல்லது தற்காலிக சேமிப்பு
- பொருண்மையியல்சார் பாகுபடுத்தல் (Yih et al. 2014)
- பதில் விளக்கக்காட்சி (Perera et al. 2017)
- அறிவு பிரதிநிதித்துவம் மற்றும் பகுத்தறிவு
- கேள்வி பதில் அமைப்புகளுடன் சமூக ஊடக பகுப்பாய்வு
- உணர்வுப் பகுப்பாய்வு ("BitCrawl by Hobson Lane")
- கருப்பொருள் பாத்திரங்களின் பயன்பாடு (Perera and Perera 2012)
- பொருண்மையியல்சார் தீர்மானம்: செயற்கையாக வேறுபட்ட கேள்விகளுக்கும் பதில் கொண்டிருக்கும் உரைகளுக்கும் இடையிலான இடைவெளியைக் குறைக்க (Bahadorreza Ofoghi et al. 2008)
- வேர்ட்நெட் (WordNet), ஃபிரேம்நெட் (FrameNet) போன்ற மொழியியல் வளங்களைப் பயன்படுத்துதல் (Bahadorreza Ofoghi et al. 2009).

- காட்சி கேள்விக்கு பதிலளிப்பதற்கான பட தலைப்பு (Anderson, Peter, et al. 2018)

ஐபிஎம்-இன் கேள்விப் பதில் ஒழுங்குமுறையான வாட்சன் (Watson) இரண்டு பெரிய இடையூறான சாம்பியன்களான பிராட் ரட்டர் மற்றும் கென் ஜென்னிங்ஸ் குறிப்பிடத்தக்க வித்தியாசத்தில் தோற்கடித்தது (Markoff (2011: 02-16). பேஸ்புக் ஆராய்ச்சி அதன் DrQA ஒழுங்குமுறையை ("DrQA") திறந்த மூல உரிமத்தின் கீழ் கிடைக்கச் செய்துள்ளது. விக்கிபீடியாவை அறிவு மூலமாகப் பயன்படுத்தி திறந்த டொமைன் கேள்விக்குப் பதிலளிக்க இந்த ஒழுங்குமுறை பயன்படுத்தப்பட்டுள்ளது (Chen et al. 2017).

6.6.6. கணினி உதவியுடன் கற்பித்தல் (Computer aided instruction)

அறிமுகம்

கணினி அடிப்படையிலான கல்வி (Computer-based education (CBE/சிபிஇ) மற்றும் கணினி அடிப்படையிலான கற்பித்தல் (computer-based instruction (CBI/சிபிஐ) ஆகியவை பரந்த சொற்கள் மற்றும் கல்வி அமைப்புகளில் கிட்டத்தட்ட எந்த வகையான கணினி பயன்பாட்டையும் குறிக்கலாம். கணினி உதவி கற்பித்தல் (CAI) என்பது ஒரு குறுகிய சொல் மற்றும் பெரும்பாலும் டிரில் மற்றும் பயிற்சி, பயிற்சி அல்லது உருவகப்படுத்துதல் செயல்பாடுகளைக் குறிக்கிறது. கணினி-நிர்வகிக்கப்பட்ட கற்பித்தல் (Computer-managed instruction/சி.எம்.ஐ) கணினி நிர்வகிக்கப்படும் கற்பித்தல் என்பது கற்றல் நோக்கங்கள், கற்றல் வளங்கள், பதிவுகளை வைத்திருத்தல், முன்னேற்ற கண்காணிப்பு மற்றும் கற்பவரின் செயல்திறனை மதிப்பீடு செய்தல் ஆகியவற்றை வழங்க கணினி பயன்படுத்தும் ஒரு கற்பித்தல் உத்தி ஆகும். கணினி அடிப்படையிலான கருவிகள் மற்றும் பயன்பாடுகள் ஆசிரியர் அல்லது பள்ளி நிர்வாகிக்கு கற்பவர் மற்றும் அறிவுறுத்தல் செயல்பாட்டை நிர்வகிக்க உதவுகின்றன.

கணினி உதவி கற்பித்தல் (CAI), கற்பித்தல் மற்றும் கற்றல் செயல்முறைக்கு உதவும் கணினி தொழில்நுட்பங்களின் மாறுபட்ட மற்றும் வேகமாக விரிவடையும் ஸ்பெக்ட்ரம் ஆகும். Computer aided instruction (CAI) என்பது computer-assisted instruction என்றும் அழைக்கப்படுகிறது. கணினி உதவி கற்பித்தல் (CAI) பயன்பாடுகளின் எடுத்துக்காட்டுகளில் வழிகாட்டப்பட்ட தொடர்பயிற்சிகள் மற்றும் பழகு பயிற்சிகள் (practice exercises), சிக்கலான பொருட்களின் கணினி காட்சிப்படுத்தல் மற்றும் மாணவர்கள் மற்றும் ஆசிரியர்களிடையே கணினி வசதியுள்ள தொடர்பு ஆகியவை அடங்கும்.

கணினி உதவியால் கற்றல் (CAI)

கணினி உதவியால் கற்றல் (கம்ப்யூட்டர்-அசிஸ்டட் இன்ஸ்ட்ரக்ஷன் (CAI) பொதுவாக ஆஃப்லைனில் அல்லது ஆன்லைனில் திட்டமிடப்பட்ட கற்பித்தல் பொருட்களுடன் மாணவரின் ஊடாட்டத்தை உட்படுத்தும் ஒரு சுய கற்றல் நுட்பம் ஆகும். கணினி உதவியால் கற்றல் (CAI) என்பது ஒரு ஊடாடும் கற்பித்தல் நுட்பமாகும்; இதன் மூலம் ஒரு கணினி, கற்பித்தல் பொருளை முன்வைக்கவும் நடைபெறும் கற்றலை கண்காணிக்கவும் பயன்படுத்தப்படுகிறது. கற்றல் செயல்முறையை மேம்படுத்துவதில் உரை, கிராபிக்ஸ், ஒலி மற்றும் வீடியோ ஆகியவற்றின் கலவையை கணினி உதவியால் கற்றல் பயன்படுத்துகிறது. கணினி வகுப்பறையில் பல நோக்கங்களைக் கொண்டுள்ளது, மேலும் பாடத்திட்டத்தின் அனைத்து பகுதிகளிலும் ஒரு மாணவருக்கு உதவ இதைப் பயன்படுத்தலாம். கணினி உதவி அறிவுறுத்தல் கணினியைப் பயன்படுத்துவதை ஒரு கருவியாகக் குறிக்கிறது. கணினி உதவி அறிவுறுத்தல் திட்டங்கள் பயிற்சிகள், டிரில் மற்றும் பயிற்சி, உருவகப்படுத்துதல் மற்றும் தற்போதைய தலைப்புகளுக்கு சிக்கல் தீர்க்கும் அணுகுமுறைகளைப் பயன்படுத்துகின்றன, மேலும் அவை மாணவர்களின் புரிதலை சோதிக்கின்றன.

வழக்கமான கணினி உதவி அறிவுறுத்தலின் வழங்குகிறது

1. உரை அல்லது மல்டிமீடியா உள்ளடக்கம்
2. பல்தேர்வு கேள்விகள்
3. பிரச்சினைகள்
4. உடனடி கருத்து
5. தவறான பதில்கள் பற்றிய குறிப்புகள்
6. மாணவர்களின் செயல்திறனைச் சுருக்கமாகக் கூறுகிறது
7. பயிற்சிக்கான பயிற்சிகள்
8. பணித்தாள்கள் மற்றும் சோதனைகள்.

கணினி உதவி அறிவுறுத்தலின் வகைகள்

1. டிரில் மற்றும் பயிற்சி: டிரில் மற்றும் பயிற்சி முன்னர் வழங்கப்பட்ட திறன்களை மீண்டும் மீண்டும் பயிற்சி செய்வதற்கான வாய்ப்புகள் அல்லது மாணவர்களுக்கு வழங்குகிறது, மேலும் தேர்ச்சிக்கு மேலும் பயிற்சி அவசியம்.

2. டுடோரியல்: டுடோரியல் செயல்பாட்டில் தகவல்களை வழங்குதல் மற்றும் அதன் நீட்டிப்பு ஆகிய இரண்டையும் உள்ளடக்கியது, இதில் டிரில் மற்றும் பயிற்சி, விளையாட்டுகள் மற்றும் உருவகப்படுத்துதல் ஆகியவை அடங்கும்.

3. விளையாட்டு: விளையாட்டு மென்பொருள் பெரும்பாலும் அதிக மதிப்பெண் பெற ஒரு போட்டியை உருவாக்குகிறது மற்றும் மற்றவர்களை வெல்லலாம் அல்லது கணினியை வெல்லும்.

4. உருவகப்படுத்துதல்: நிஜ வாழ்க்கையின் செலவு அல்லது அதன் அபாயங்கள் தேவையில்லாத யதார்த்தத்தின் தோராயத்தை உருவகப்படுத்துதலை மென்பொருள் வழங்க முடியும்.

5. கண்டுபிடிப்பு: கண்டுபிடிப்பு அணுகுமுறை ஒரு பாடநெறி அல்லது உள்ளடக்க பகுதிக்கு குறிப்பிட்ட தகவல்களின் பெரிய தரவுத்தளத்தை வழங்குகிறது மற்றும் தரவின் ஆய்வுகளின் அடிப்படையில் பகுப்பாய்வு செய்ய, ஒப்பிட்டு, ஊகிக்க மற்றும் மதிப்பீடு செய்ய கற்றவருக்கு சவால் விடுகிறது.

6. சிக்கல் தீர்க்கும் முறை: இந்த அணுகுமுறை குழந்தைகளுக்கு குறிப்பிட்ட சிக்கல் தீர்க்கும் திறன்கள் மற்றும் உத்திகளை வளர்க்க உதவுகிறது

கணினி உதவி அறிவுறுத்தலின் வகைகள்

தொடர்புகளை கற்பிக்க அல்லது ஊக்குவிக்கும் தகவல்களைக் கணினிகளில் உரை வடிவில் அல்லது பல்லூடக/மல்டிமீடியா வடிவங்களில் வழங்கலாம்; இதில் புகைப்படங்கள், வீடியோக்கள், அனிமேஷன், பேச்சு மற்றும் இசை ஆகியவை அடங்கும். வழிகாட்டப்பட்ட தொடர்பயிற்சி/டிரில் (guided drill) என்பது கணினி நிரலாகும்; இது மாணவர்களுக்கு கேள்விகளை எழுப்புகிறது, கருத்துக்களை வழங்குகிறது, மேலும் மாணவர்களின் பதில்களின் அடிப்படையில் கூடுதல் கேள்விகளைத் தேர்ந்தெடுக்கிறது. சமீபத்திய வழிகாட்டுதல் தொடர்பயிற்சி/டிரில் ஒழுங்குமுறைகள் கணினி நிரலில் பொருள் அறிவுக்குக் (subject matter knowledge) கூடுதலாகக் கல்வியின் கொள்கைகளையும் இணைத்துள்ளன.

கணினிகள் மாணவர்களுக்குக் கடினமான அல்லது பார்க்க முடியாத பொருட்களைக் காட்சிப்படுத்த உதவும். எடுத்துக்காட்டாக, மனித உடற்கூறியல், மூலக்கூறு கட்டமைப்புகள் (molecular structures) அல்லது சிக்கலான வடிவியல் பொருள்களைக் (complex geometrical objects) காட்ட கணினிகள் பயன்படுத்தப்படலாம். உருவகப்படுத்தப்பட்ட சூழல்களின் ஆய்வு மற்றும் கையாளுதல் கணினி உதவி கற்பித்தல்-மெய்நிகர் ஆய்வகச் சோதனைகள் முதல் பள்ளி சூழல் முதல், விமான விமான சிமுலேட்டர்களில் பயன்படுத்தப்படுவது போன்ற சிக்கலான

மெய்நிகர் உலகங்கள் வரை நிகழ்த்துவது மிகவும் கடினமானதாக, விலையுயர்ந்ததாக அல்லது ஆபத்தானதாக இருக்கலாம்.

சொல் செயலிகள் (word processors), விரிதாள்கள் (spreadsheets) மற்றும் தரவுத்தளங்கள் (databases போன்ற) கணினி உதவி கற்பித்தல் கருவிகள், தகவல்களை சேகரித்தல், ஒழுங்கமைத்தல், பகுப்பாய்வு செய்தல் மற்றும் பரிமாற்றம் செய்தல் இவற்றை செய்யும். அவை மாணவர்களிடையேயும், மாணவர்கள் மற்றும் பயிற்றுநர்களிடையேயும், வகுப்பறைக்கு அப்பால் தொலைதூர மாணவர்கள், பயிற்றுநர்கள் மற்றும் நிபுணர்களிடமும் தொடர்பு கொள்ள உதவுகின்றன.

பாடத்தின் முன்னேற்றத்தை யார் கட்டுப்படுத்துகிறார்கள் என்பதன் அடிப்படையில் CAI அமைப்புகளை வகைப்படுத்தலாம். ஆரம்பகால அமைப்புகள் தகவல்களின் நேரியல் விளக்கக்காட்சிகள் மற்றும் வழிகாட்டப்பட்ட டிரில்/தொடர்பயிற்சி, மற்றும் கட்டுப்பாடு என்பன மென்பொருளின் ஆசிரியரால் இயக்கப்பட்டது. நவீன அமைப்புகளில், குறிப்பாகக் காட்சிப்படுத்தல் அமைப்புகள் மற்றும் உருவகப்படுத்தப்பட்ட சூழல்களில், கட்டுப்பாடு பெரும்பாலும் மாணவரிடமோ அல்லது பயிற்றுவிப்பாளரிடமோ இருக்கும். இது தகவல்களை மறுபரிசீலனை செய்யவோ அல்லது ஆய்வு செய்யவோ அனுமதிக்கிறது. தொடர்புடைய பொருள் கூட ஆராயப்படலாம். சில குழு அறிவுறுத்தல் நடவடிக்கைகளில், குழுவின் இயக்கவியல் படி பாடம் முன்னேறலாம்.

நன்மைகள் மற்றும் தீமைகள்

கணினி உதவி கற்றல் ஒரு மாணவரின் தகவலுக்கான அணுகலை வியத்தகு முறையில் அதிகரிக்க முடியும். இந்தத் திட்டம் தனிப்பட்ட மாணவரின் திறன்களுக்கும் விருப்பங்களுக்கும் ஏற்ப மாற்றலாம் மற்றும் ஒரு மாணவர் பெறும் தனிப்பயனாக்கப்பட்ட அறிவுறுத்தலின் அளவை அதிகரிக்கலாம். பல மாணவர்கள் கணினி தொடர்புகளின் உடனடி பதிலளிப்பிலிருந்து பயனடைகிறார்கள் மற்றும் சுய-வேக மற்றும் தனியார் கற்றல் சூழலைப் பாராட்டுகிறார்கள். மேலும், கணினி கற்றல் அனுபவங்கள் பெரும்பாலும் மாணவர்களின் ஆர்வத்தில் ஈடுபடுகின்றன, மேலும் அவர்களைக் கற்றுக்கொள்ள தூண்டுகின்றன, மேலும் சுதந்திரம் மற்றும் கல்விக்கான தனிப்பட்ட பொறுப்பை அதிகரிக்கின்றன.

எந்தவொரு கல்வி முறையின் செயல்திறனையும் மதிப்பிடுவது கடினம் என்றாலும், தேர்வு மதிப்பெண்களை உயர்த்துவதிலும், மாணவர்களின் மனப்பான்மையை மேம்படுத்துவதிலும், சில

விஷயங்களை மாஸ்டர் செய்வதற்குத் தேவையான நேரத்தைக் குறைப்பதிலும் கணினி உதவி கற்றல் வெற்றிகரமாக இருப்பதாகப் பல ஆய்வுகள் தெரிவிக்கின்றன. ஆய்வு முடிவுகள் பெரிதும் மாறுபடும் அதே வேளையில், கணினி உதவி கற்றல் அனைத்து கல்வி மட்டங்களிலும் கற்றலை மேம்படுத்த முடியும் என்பதற்கு கணிசமான சான்றுகள் உள்ளன.

சில பயன்பாடுகளில், குறிப்பாகச் சுருக்கப் பகுத்தறிவு மற்றும் சிக்கல் தீர்க்கும் செயல்முறைகள் சம்பந்தப்பட்டவையில் கணினி உதவி கற்றல் மிகவும் பயனுள்ளதாக இல்லை. மோசமாக வடிவமைக்கப்பட்ட கணினி உதவி கற்றல் அமைப்புகள் கல்வி அனுபவத்தை மனிதநேயமற்றதாகவோ அல்லது கட்டுத்திட்டமாக ஆக்கவோ முடியும் என்றும் அதன் மூலம் மாணவர்களின் ஆர்வத்தையும் ஊக்கத்தையும் குறைக்க முடியும் என்றும் விமர்சகர்கள் கூறுகின்றனர். கணினி உதவி கற்றலின் பிற குறைபாடுகள் தேவையான கணினி அமைப்புகளை செயல்படுத்துவதற்கும் பராமரிப்பதற்கும் உள்ள சிரமம் மற்றும் செலவில் இருந்து உருவாகின்றன. சில மாணவர் தோல்விகளை கணினி உதவி கற்றல் அமைப்புகளில் போதாத ஆசிரியர் பயிற்சியில் காணலாம். கணினி தொழில்நுட்பத்தில் மாணவர் பயிற்சியும் தேவைப்படலாம், மேலும் இந்தச் செயல்முறை முக்கியக் கல்விச் செயல்முறையிலிருந்து திசை திருப்பலாம். கற்பித்தல் மற்றும் கற்றல் பற்றிய நிபுணத்துவ அறிவைப் பயன்படுத்த எளிதானது மற்றும் இணைத்துக்கொள்ள எளிதான CAI அமைப்புகளை வளர்ப்பதில் அதிக முயற்சி மேற்கொள்ளப்பட்டாலும், அத்தகைய அமைப்புகள் அவற்றின் முழு திறனை அடைவதில் இருந்து இன்னும் வெகு தொலைவில் உள்ளன.

வரலாறு

1950களின் நடுப்பகுதியிலும் 1960களின் முற்பகுதியிலும் கலிபோர்னியாவின் ஸ்டான்போர்ட் பல்கலைக்கழக கல்வியாளர்களுக்கும் சர்வதேச வர்த்தக இயந்திரக் கழகத்திற்கும் (International Business Machines Corporation IBM/ஐபிஎம்) உள்ள ஒத்துழைப்பு கணினி உதவி கற்றலை தேர்ந்தெடுக்கப்பட்ட தொடக்கப் பள்ளிகளில் அறிமுகப்படுத்தியது. ஆரம்பத்தில், கணினி உதவி கற்றல் திட்டங்கள் தொடர்பயிற்சி/டிரில் மற்றும் பயிற்சி அமர்வுகளுடன் தகவல்களின் நேரியல் விளக்கக்காட்சியாக இருந்தன. இந்த ஆரம்ப கணினி உதவி கற்றல் அமைப்புகள் அந்த நேரத்தில் கிடைத்த கணினிகளைப் பெறுவதற்கும் பராமரிப்பதற்கும் பயன்படுத்துவதற்கும் சிரமம் மற்றும் செலவால் எல்லைபடுத்தப்பட்டன. 1960களின் முற்பகுதியில் இல்லினாய்ஸ் பல்கலைக்கழகத்தில் (University of Illinois) தொடங்கப்பட்ட மற்றும்

கட்டுப்பாட்டு தரவுக் கழகத்தால் உருவாக்கப்பட்ட மற்றொரு ஆரம்ப கணினி உதவி கற்றல் அமைப்பான தானியங்கி கற்பித்தல் நடவடிக்கைகளுக்கான அமைப்பு திட்டமிடப்பட்ட தர்க்கம் (Programmed Logic for Automatic Teaching Operations (PLATO/பிளாட்டோ)) உயர் கற்றலுக்குப் பயன்படுத்தப்பட்டது. இது ஒரு மெயின்பிரேம் கணினியைக் கொண்டிருந்தது, இது தனிப்பட்ட மாணவர்களின் பயன்பாட்டிற்காக 1000 டெர்மினல்களை ஆதரித்தது. 1985 வாக்கில் அமெரிக்காவில் 100க்கும் மேற்பட்ட பிளாட்டோ (PLATO)அமைப்புகள் இயங்கி வந்தன. 1978 முதல் 1985 வரை பயனர்கள் 40 மில்லியன் மணிநேரங்களை பிளாட்டோ (PLATO) கணினிகளில் பதிவு செய்தனர். நவீன மின்னணு அஞ்சல்களின் முன்னோடியாக இருந்த மாணவர்களிடையே ஒரு தகவல்தொடர்பு முறையையும் பிளாட்டோ அறிமுகப்படுத்தியது (கணினியிலிருந்து கணினிக்கு மின்னணு முறையில் அனுப்பப்பட்ட செய்திகள்). டைம்-ஷேர்டு இன்டராக்டிவ் கம்ப்யூட்டர்-கன்ட்ரோல்ட் இன்ஃபர்மேஷன் டெலிவிஷன் (Time-shared Interactive Computer-Controlled Information Television TICCIT/டிஐசிஐடி) அமைப்பு என்பது உட்டாவில் உள்ள மிட்டர் கார்ப்பரேஷன் மற்றும் ப்ரிகாம் யங் பல்கலைக்கழகம் உருவாக்கிய கணினி உதவி கற்றல் திட்டமாகும். தனிப்பட்ட கணினி மற்றும் தொலைக்காட்சி தொழில்நுட்பத்தின் அடிப்படையில், 1970களின் முற்பகுதியில் டி.ஐ.சி.ஐ.டி புதிய மாணவர் அளவிலான கணிதம் மற்றும் ஆங்கில படிப்புகளைக் கற்பிக்க பயன்படுத்தப்பட்டது. 1980களில் மலிவான மற்றும் சக்திவாய்ந்த தனிநபர் கணினிகளின் வருகையுடன், கணினி உதவி கற்றலின் பயன்பாடு வியத்தகு அளவில் அதிகரித்தது. 1980ஆம் ஆண்டில் தொடக்கப் பள்ளிகளில் 5 சதவீதமும், அமெரிக்காவில் 20 சதவீத மேல்நிலைப் பள்ளிகளும் மட்டுமே பயிற்றுவிப்பிற்கு உதவ கணினிகள் இருந்தன. மூன்று ஆண்டுகளுக்குப் பிறகு, இரு எண்களும் ஏறக்குறைய நான்கு மடங்காக உயர்ந்தன, தசாப்தத்தின் முடிவில் அமெரிக்காவிலும், பெரும்பாலான தொழில்மயமான நாடுகளிலும் கிட்டத்தட்ட அனைத்து பள்ளிகளும் கற்பித்தல் கணினிகளைக் கொண்டிருந்தன. கணினி உதவி கற்றலுக்கு நீண்டகால தாக்கங்களைக் கொண்ட சமீபத்திய வளர்ச்சியானது இணையத்தின் பரந்த விரிவாக்கம் ஆகும், இது ஒன்றோடொன்று இணைக்கப்பட்ட கணினிகளின் கூட்டமைப்பு ஆகும். உலகெங்கிலும் மில்லியன் கணக்கான கணினிகளை இணைப்பதன் மூலம், இந்த நெட்வொர்க்குகள் மாணவர்களின் பெரிய தகவல்களை அணுக உதவுகின்றன, இது அவர்களின் ஆராய்ச்சி திறன்களை பெரிதும் மேம்படுத்துகிறது.

6.6.7. கணினியின் உதவியுடன் மொழிக் கற்றல்

கணினி உதவியுடன் மொழி கற்றல் (Computer-assisted language learning (CALL) என்று பிரிட்டிஷ் நடையிலும்), கணினி உதவியுடன் கற்பித்தல் (Computer-Aided Instruction (CAI) அல்லது கணினி உதவியுடன் மொழி கற்பித்தல் (Computer-Aided Language Instruction (CALI) என்று அமெரிக்கன் நடையிலும்) (Higgins 1983) குறிப்பிடப்படுகின்ற செயல்பாடு சுருக்கமாக லெவியால் (Levy 1997:1) "மொழி கற்பித்தல் மற்றும் கற்றலில் கணினியின் பயன்பாடுகளுக்கான தேடல் மற்றும் ஆய்வு" என்று வரையறுக்கப்படுகின்றது. 1960கள் மற்றும் 1970களில் கணினியின் உதயுடன் மொழிக் கல்வியை வகைப்படுத்திய "பாரம்பரிய" பயிற்சி மற்றும் நடைமுறை திட்டங்களிலிருந்து (drill-and-practice programs) கணினியின் உதவியுடன் மொழிக் கற்றலின் சமீபத்திய வெளிப்பாடுகள் வரை வெளிநாட்டு மொழிகள் கற்பித்தல் மற்றும் கற்றுக்கொள்வதற்கான பரந்த அளவிலான தகவல் மற்றும் தகவல் தொடர்பு தொழில்நுட்ப பயன்பாடுகள் மற்றும் அணுகுமுறைகளை கணினியின் உதவியுடன் மொழிக் கல்வி ஏற்றுக்கொள்கிறது, எ.கா. ஒரு மெய்நிகர் கற்றல் சூழலில் மற்றும் வலை அடிப்படையிலான தொலைதூரக் கற்றலில் பயன்படுத்தப்படுகிறது. இது தரவுத்தொகுதி மற்றும் ஒத்திசைவாளர்கள் (corpora and concordancers), ஊடாடும் ஒயிட் போர்டுகள் (interactive whiteboards) (Schmid Euline Cutrim 2009), கணினி-இடையீடான தகவல் தொடர்பு (Computer-mediated communication (CMC/சிஎம்சி) (Lamy & Hampel 2007), மெய்நிகர் உலகங்களில் மொழி கற்றல், மற்றும் மொபைல் உதவி மொழி கற்றல் (mobile-assisted language learning (MALL)) ஆகியவற்றின் பயன்பாட்டிற்கும் நீண்டுள்ளது (Shield & Kukulska-Hulme 2008).

CALI (கணினி உதவியுடன் மொழி கற்பித்தல்) என்ற சொல் கணினியின் உதயுடன் மொழிக் கல்விக்கு முன்னர் பயன்பாட்டில் இருந்தது, இது CAI (கணினி உதவி கற்பித்தல்) என்ற பொதுச் சொல்லின் துணைக்குழுவாக அதன் தோற்றத்தைப் பிரதிபலிக்கிறது. இருப்பினும், CALI மொழி ஆசிரியர்களிடையே ஆதரவை இழந்தது, ஏனெனில் இது ஆசிரியர்களை மையமாகக் கொண்ட அணுகுமுறையைக் (கற்பித்தல்) குறிக்கிறது, அதே சமயம் மொழி ஆசிரியர்கள் மாணவர்களை மையமாகக் கொண்ட அணுகுமுறையை விரும்புவதில் அதிக விருப்பம் கொண்டுள்ளனர், கற்பிப்பதை விட கற்றலில் கவனம் செலுத்துகிறார்கள். 1980 களின் முற்பகுதியில் CALIஐ மாற்றுவதற்கு CALL தொடங்கியது (டேவிஸ் & ஹிக்கின்ஸ் 1982: பக். 3) (Davies & Higgins 1982) மேலும் இது இப்போது உலகளவில் வளர்ந்து வரும் தொழில்முறை சங்கங்களின் பெயர்களில் இணைக்கப்பட்டுள்ளது.

1990களின் முற்பகுதியில் தொழில்நுட்ப-மேம்பட்ட மொழி கற்றல் (TELL), (Bush & Terry 1997) ஒரு மாற்று சொல் தோன்றியது: எ.கா. டெல் கூட்டமைப்பு திட்டம், ஹல் பல்கலைக்கழகம்.

கணினியின் உதவியுடன் மொழிக் கற்றலின் தற்போதைய தத்துவம், மாணவர்களை மையமாகக் கொண்ட பொருட்களுக்கு (student-centred materials) வலுவான முக்கியத்துவத்தை அளிக்கிறது, இது கற்பவர்களுக்குச் சொந்தமாக செயல்பாடு செய்ய அனுமதிக்கிறது. இத்தகைய பொருட்கள் கட்டமைக்கப்பட்டதாகவோ அல்லது கட்டமைக்கப்படாமலோ (structured or unstructured) இருக்கலாம், ஆனால் அவை பொதுவாக இரண்டு முக்கியமான அம்சங்களைக் கொண்டுள்ளன: ஊடாடும் கற்றல் (interactive learning) மற்றும் தனிப்பயனாக்கப்பட்ட கற்றல் and (individualised learning). கணினியின் உதவியுடன் மொழிக் கற்றல் என்பது அடிப்படையில் மொழி கற்றல் செயல்முறையை எளிதாக்க ஆசிரியர்களுக்கு உதவும் ஒரு கருவியாகும். வகுப்பறையில் ஏற்கனவே கற்றுக்கொண்டவற்றை வலுப்படுத்த அல்லது கூடுதல் ஆதரவு தேவைப்படும் கற்பவர்களுக்கு உதவ ஒரு தீர்வுக் கருவியாக இதைப் பயன்படுத்தலாம்.

கணினியின் உதவியுடன் மொழிக் கல்வி பொருட்களின் வடிவமைப்பு பொதுவாக மொழி கற்பித்தல் மற்றும் முறையின் கொள்கைகளைக் கருத்தில் கொள்கிறது, அவை வெவ்வேறு கற்றல் கோட்பாடுகளிலிருந்து (எ.கா. நடத்தையியல், புலனறிவுவியல், அமைப்பியலார்/behaviourist, cognitive, constructivist) மற்றும் ஸ்டீபன் க்ராஷனின் மானிட்டர் கருதுகோள் (Stephen Krashen's monitor hypothesis) போன்ற இரண்டாம் மொழி கற்றல் கோட்பாடுகளிலிருந்து பெறப்படலாம்.

நேருக்கு நேர் கற்பித்தல் (face-to-face teaching) மற்றும் கணினியின் உதயுடன் மொழிக் கல்வி ஆகியவற்றின் கலவை பொதுவாகக் கலப்பு கற்றல் (blended learning) எனக் குறிப்பிடப்படுகிறது. கலப்பு கற்றல் கற்றல் திறனை அதிகரிக்க வடிவமைக்கப்பட்டுள்ளது மற்றும் இது தூய கணினியின் உதவியுடன் மொழிக் கல்வி விடப் பொதுவாகக் காணப்படுகிறது (Pegrum 2009: 27).

வரலாறு

கணினியின் உதயுடன் மொழிக் கற்றல் பல்கலைக்கழக மெயின்பிரேம் கணினிகளில் 1960களில் முதன்முதலில் அறிமுகப்படுத்தப்பட்டதில் இருந்து பயன்பாட்டில் இருக்கிறது. 1960இல் இல்லினாய்ஸ் பல்கலைக்கழகத்தில் தொடங்கப்பட்ட PLATO திட்டம், கணினியின் உதயுடன் மொழிக் கல்வியின் ஆரம்ப வளர்ச்சியில் ஒரு முக்கியமான அடையாளமாகும் (Marty 1981). 1970களின் பிற்பகுதியில் மைக்ரோகம்ப்யூட்டரின் வருகை ஒரு பரந்த பார்வையாளர்களின்

வரம்பிற்குள் கம்ப்யூட்டிங்கைக் கொண்டுவந்தது, இதன் விளைவாக கணினியின் உதயுடன் மொழிக் கல்வி திட்டங்களின் வளர்ச்சியில் ஏற்றம் ஏற்பட்டது மற்றும் 1980களின் முற்பகுதியில் கணினியின் உதவியுடன் மொழிக் கற்றல் பற்றிய புத்தகங்களின் வெளியீடுகள் பெருகின.

டஜன் கணக்கான கணினியின் உதவியுடன் மொழிக் கற்றல் திட்டங்கள் இலவசத்திலிருந்து விலையுயர்ந்த விலையில் தற்போது இணையத்தில் கிடைக்கின்றன; மற்றும் பிற திட்டங்கள் பல்கலைக்கழக மொழி படிப்புகள் மூலம் மட்டுமே கிடைக்கின்றன.

கணினி உதவியுடன் மொழிக் கற்றல் வரலாற்றை ஆவணப்படுத்த பல முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. சாண்டர்ஸ் (Sanders 1995) 1960களின் நடுப்பகுதியிலிருந்து 1990களின் நடுப்பகுதி வரையிலான காலத்தை உள்ளடக்கியுள்ளார், இது வட அமெரிக்காவில் கணினி உதவியுடன் மொழிக் கற்றலை மையமாகக் கொண்டது. டெல்க்லோக் (Delcloque 2000) 1960களில் கணினி உதவியுடன் மொழிக் கற்றலின் தொடக்கத்திலிருந்து புதிய மில்லினியத்தின் விடியல் வரை உலகளவில் கணினி உதவியுடன் மொழிக் கற்றலின் வரலாற்றை ஆவணப்படுத்துகிறார். டேவிஸ் (2005) கணினியின் உதயுடன் மொழிக் கற்றலின் கடந்த காலத்தை திரும்பிப் பார்த்து, அது எங்கு செல்கிறது என்பதைக் கணிக்க முயற்சிக்கிறார். ஹப்பார்ட் (Hubbard 2009) 74 முக்கிய கட்டுரைகள் மற்றும் புத்தகப் பகுதிகளின் தொகுப்பை வழங்குகிறார்; இது முதலில் 1988-2007 ஆண்டுகளில் வெளியிடப்பட்டது; கணினியின் உதயுடன் மொழிக் கல்வியின் வளர்ச்சியில் செல்வாக்கு செலுத்திய பரந்த அளவிலான முன்னணி யோசனைகள் மற்றும் ஆராய்ச்சி முடிவுகளின் விரிவான கண்ணோட்டத்தை இது அளிக்கிறது. மேலும் எதிர்காலத்தில் அவ்வாறு செய்வதன் நம்பிக்கையைக் காட்டுகிறது. ஹப்பார்ட்டின் சேகரிப்பு பற்றிய வெளியிடப்பட்ட மதிப்பாய்வை மொழி கற்றல் மற்றும் தொழில் நுட்பத்தில் (Language Learning & Technology 14, 3 2010) காணலாம்.

பட்லர்-பாஸ்கோ (Butler-Pasco 2011) கணினியின் உதயுடன் மொழிக் கற்றலின் வரலாற்றை வேறுபட்ட கண்ணோட்டத்தில் பார்க்கிறார்; அதாவது கல்வித் தொழில்நுட்பத்தின் இரட்டை துறைகளில் கணினியின் உதயுடன் மொழிக் கல்வியின் பரிணாமம் மற்றும் இரண்டாம்/வெளிநாட்டு மொழி ஈட்டம் மற்றும் வழியில் அனுபவிக்கும் முன்னுதாரண மாற்றங்கள் (paradigm shifts)

வகைப்பாட்டியல் மற்றும் கட்டங்கள்

1980கள் மற்றும் 1990களில், ஒரு கணினியின் உதவியுடன் மொழிக் கற்றல் வகைப்பாட்டியல் நிறுவ பல முயற்சிகள் மேற்கொள்ளப்பட்டன. டேவிஸ் & ஹிக்கின்ஸ் (Davies & Higgins 1985), ஜோன்ஸ் & ஃபோர்டெஸ்க்யூ (Jones & Fortescue 1987), ஹார்டிஸ்டி & விண்டீட் (Hardisty & Windeatt 1989) [21] மற்றும் லெவி (Levy 1997: 118ff.) ஆகியோரால் பல்வேறு வகையான கணினியின் உதயுடன் மொழிக் கல்வி திட்டங்கள் அடையாளம் காணப்பட்டன. இடைவெளி நிரப்புதல் மற்றும் மூடு நிரல்கள், பல தேர்வுத் திட்டங்கள், இலவச-வடிவம் (உரை-நுழைவு) நிரல்கள், சாகசங்கள் மற்றும் உருவகப்படுத்துதல்கள், செயல் பிரமைகள், வாக்கிய-மறுசீரமைப்புத் திட்டங்கள், ஆய்வுத் திட்டங்கள் - மற்றும் கற்றவர் முழு உரையையும் புனரமைக்க வேண்டிய "மொத்த மூடு" (total Cloze(ஆகியவை இதில் அடங்கும். இந்த ஆரம்ப திட்டங்களில் பெரும்பாலானவை நவீனமயமாக்கப்பட்ட பதிப்புகளில் உள்ளன.

1990களில் இருந்து, வலைப்பதிவுகள், விக்கிகள், சமூக வலைப்பின்னல், போட்காஸ்டிங், வலை 2.0 பயன்பாடுகள், மெய்நிகர் உலகங்களில் மொழி கற்றல் மற்றும் ஊடாடும் ஒயிட் போர்டுகள் (டேவிஸ் மற்றும் பலர். Davies et al. 2010: பிரிவு 3.7).

வார்சவுர் (Warschauer 1996) மற்றும் வார்சவுர் & ஹீலி (Warschauer & Healey 1998) வேறுபட்ட அணுகுமுறையை எடுத்தனர். கணினியின் உதவியுடன் மொழிக் கற்றலின் வகைப்பாட்டியல் மீது கவனம் செலுத்துவதற்குப் பதிலாக, அவர்கள் கணினியின் உதயுடன் மொழிக் கல்வியின் மூன்று வரலாற்று கட்டங்களை அடையாளம் கண்டனர், அவற்றின் அடிப்படை கல்வி மற்றும் வழிமுறை அணுகுமுறைகளின்படி வகைப்படுத்தப்பட்டனர்:

- நடத்தை கணினியின் உதவியுடன் மொழிக் கற்றல் (Behavioristic CALL): 1950 களில் கருத்தரிக்கப்பட்டு 1960 கள் மற்றும் 1970 களில் செயல்படுத்தப்பட்டது.
- தகவல்தொடர்பு கணினியின் உதவியுடன் மொழிக் கற்றல் (Communicative CALL): 1970கள் முதல் 1980கள் வரை.
- ஒருங்கிணைந்த கணினியின் உதவியுடன் மொழிக் கல்வி (Integrative CALL): பல்லூடகம் (Multimedia) and மற்றும் இணையத்தைத் (Internet) தழுவுதல்: 1990கள்.

வார்சவுர் & ஹீலியின் முதல் கட்டமான நடத்தை கணினியின் உதவியுடன் மொழிக் கற்றல் (1960கள் முதல் 1970கள் வரை) பெரும்பாலான கணினியின் உதவியுடன் மொழிக் கற்றல் திட்டங்கள், தொடர்ப்பயிற்சி மற்றும் பயிற்சிப் பொருள்களைக் கொண்டிருந்தன, இதில் கணினி ஒரு தூண்டுதலை வழங்கியது மற்றும் கற்பவர் ஒரு பதிலை வழங்கினார். முதலில், இரண்டையும்

உரை மூலம் மட்டுமே செய்ய முடியும். கணினி மாணவர்களின் உள்ளீட்டை ஆராய்ந்து கருத்துத் தெரிவிக்கும், மேலும் அதிநவீன நிரல்கள் மாணவர்களின் தவறுகளுக்குத் திரைகள் மற்றும் தீர்வு நடவடிக்கைகளுக்கு உதவுவதன் மூலம் செயல்படும். இத்தகைய திட்டங்கள் மற்றும் அவற்றின் அடிப்படைக் கல்வி கற்பித்தல் இன்றும் இருந்தபோதிலும், மொழி கற்றலுக்கான நடத்தை அணுகுமுறைகள் பெரும்பாலான மொழி ஆசிரியர்களால் நிராகரிக்கப்பட்டுள்ளன, மேலும் கணினி தொழில்நுட்பத்தின் அதிகரித்துவரும் நுட்பம் பிற சாத்தியக்கூறுகளுக்கு கணினியின் உதவியுடன் மொழிக் கற்றல் விடுத்துள்ளது.

1970களின் பிற்பகுதியிலும் 1980களின் பிற்பகுதியிலும் (அண்டர்வுட்/Underwood 1984) முக்கியத்துவம் பெற்ற தகவல்தொடர்பு அணுகுமுறையை அடிப்படையாகக் கொண்ட தகவல்தொடர்பு அழைப்பின் அடிப்படையில் வார்சவுர் & ஹீலி விவரித்த இரண்டாம் கட்டம். தகவல்தொடர்பு அணுகுமுறையில், மொழியின் பகுப்பாய்வைக் காட்டிலும் மொழியைப் பயன்படுத்துவதில் கவனம் செலுத்தப்படுகிறது, மேலும் இலக்கணம் வெளிப்படையாகக் கற்பிக்கப்படுவதைக் காட்டிலும் மறைமுகமாக கற்பிக்கப்படுகிறது. மொழியின் மாணவர் வெளியீட்டில் அசல் மற்றும் நெகிழ்வுத்தன்மையையும் இது அனுமதிக்கிறது. தகவல்தொடர்பு அணுகுமுறை தனிநபர் கணினியின் (Personal Computer (PC)) வருகையுடன் ஒத்துப்போனது, இது கம்ப்யூட்டிங் மிகவும் பரவலாகக் கிடைக்கச் செய்தது மற்றும் மொழி கற்றலுக்கான மென்பொருளின் வளர்ச்சியில் ஏற்றம் கண்டது. இந்த கட்டத்தில் முதல் கணினியின் உதவியுடன் மொழிக் கற்றல் மென்பொருள் திறன் பயிற்சியைத் (skill practice) தொடர்ந்து வழங்கியது, ஆனால் ஒரு தொடர்பயிற்சி வடிவத்தில் (drill format) இல்லை - எடுத்துக்காட்டாக: வேகமான வாசிப்பு, உரை புனரமைப்பு மற்றும் மொழி விளையாட்டுகள் - ஆனால் கணினி ஆசிரியராகவே இருந்தது. இந்தக் கட்டத்தில், கணினிகள் மாணவர்களுக்கு மொழியைப் பயன்படுத்துவதற்கான சூழலை வழங்கின, அதாவது ஒரு இடத்திற்கான திசைகளைக் கேட்பது, மற்றும் சிம் சிட்டி, ஸ்லூத் Sleuth மற்றும் வேர் இன் தி வேர்ல்டு கார்மென் சாண்டிகோ? (Where in the World is Carmen Sandiego?) போன்ற மொழிக் கற்றலுக்காக வடிவமைக்கப்படாத திட்டங்கள் மொழிக் கற்றலுக்குப் பயன்படுத்தப்பட்டன. இந்த அணுகுமுறையின் விமர்சனங்கள், மொழி கற்பித்தலின் மைய நோக்கங்களைக் காட்டிலும் அதிக விளிம்பு நோக்கங்களுக்காகக் கணினியை ஒரு குறிப்பிட்ட நோக்கத்திற்குரிய மற்றும் தொடர்பற்ற முறையில் பயன்படுத்துவதும் அடங்கும்.

1990களில் தொடங்கி வார்சவுர் & ஹீலி விவரித்த ஒருங்கிணைந்த கணினியின் உதவியுடன் மொழிக் கற்றலின் மூன்றாம் கட்டம், செயல்முறையையும் ஒத்திசைவையும் தருவதற்காக மொழித் திறன்களைக் கற்பிப்பதைச் செயல்பாடுகள் அல்லது திட்டங்களாக ஒருங்கிணைப்பதன் மூலம் தகவல்தொடர்பு அணுகுமுறையின் விமர்சனங்களைத் தீர்க்க முயன்றது. இது பல்லூடகத் தொழில்நுட்பம் (உரை, கிராபிக்ஸ், ஒலி மற்றும் அனிமேஷனை வழங்குதல்) மற்றும் கணினி-இடைப்பட்ட தொடர்பு (Computer-mediated communication (CMC/சிஎம்சி) ஆகியவற்றின் வளர்ச்சியுடனும் ஒத்துப்போனது. இந்தக் காலகட்டத்தில் கணினியின் உதவியுடன் மொழிக் கற்றல், தொடர்பயிற்சி மற்றும் பயிற்சி நோக்கங்களுக்காகக் கணினியைப் பயன்படுத்துவதிலிருந்து (கணினியை ஒரு குறிப்பிட்ட செயல்பாட்டிற்காக வரையறுக்கப்பட்ட, அதிகாரபூர்வமான தளமாக) வகுப்பறைக்கு அப்பால் கல்வியை விரிவுபடுத்துவதற்கான ஒரு ஊடகமாக ஒரு உறுதியான மாற்றத்தைக் கண்டது. பல்லூடக கணினியின் உதயுடன் மொழிக் கல்வி மான்டெவிடிஸ்கோ (ஷ்னீடர் & பென்னியன் Schneider & Bennion 1984) மற்றும் எ லா ரென்காண்ட்ரே டி பிலிப் (A la rencontre de Philippe) (ஃபியூஸ்டன்பெர்க் Fuerstenberg 1993) போன்ற ஊடாடும் லேசர் வீடியோடிஸ்களுடன் (interactive laser videodiscs) தொடங்கியது; இவை இரண்டும் கற்பவர் முக்கிய பங்கு வகித்த சூழ்நிலைகளின் உருவகப்படுத்துதல்கள் ஆகும். இந்தத் திட்டங்கள் பின்னர் குறுவட்டுகளுக்கு (CD-ROM) மாற்றப்பட்டன; மேலும் Who is Oscar Lake?/ஆஸ்கார் ஏரி யார்? போன்ற புதிய ரோல்-பிளேமிங் கேம்கள் (role-playing games (RPGs/ஆர்பிஜிக்கள்). வெவ்வேறு மொழிகளின் பரப்பெல்லையில் அவற்றின் தோற்றத்தை உருவாக்கின.

பின்னர் வெளியான ஒரு பதிப்பில், வார்சவுர் முதல் கட்டத்தின் பெயரை நடத்தை அழைப்பிலிருந்து கட்டமைப்பு கணினியின் உதயுடன் மொழிக் கல்வி என மாற்றினார், மேலும் மூன்று கட்டங்களின் தேதிகளையும் திருத்தியுள்ளார் (வார்ஷவுர்/ Warschauer 2000):

- கட்டமைப்பு கணினியின் உதவியுடன் மொழிக் கற்றல் (Structural CALL): 1970 கள் முதல் 1980 கள் வரை.
- தகவல்தொடர்பு கணினியின் உதவியுடன் மொழிக் கற்றல் (Communicative CALL): 1980 கள் முதல் 1990 கள் வரை.
- ஒருங்கிணைந்த கணினியின் உதவியுடன் மொழிக் கற்றல் (Integrative CALL): 2000 முதல்.

பாக்ஸ் (Bax 2003) வார்சவுர் & ஹேலி (Warschaue r& Healey 1998) மற்றும் வார்சவுர் (Warschauer 2000) ஆகியோருடன் சிக்கலை எடுத்து இந்த மூன்று கட்டங்களை முன்மொழிந்தார்:

- கட்டுண்ட கணினியின் உதயுடன் மொழிக் கல்வி (Restricted CALL) - முக்கியமாக நடத்தை: 1960கள் முதல் 1980கள் வரை.
- திறந்த கணினியின் உதவியுடன் மொழிக் கற்றல் (Open CALL) - உருவகப்படுத்துதல்கள் மற்றும் விளையாட்டுகள் உட்பட மாணவர்கள், மென்பொருள் வகைகள் மற்றும் ஆசிரியரின் பங்கு ஆகியவற்றின் பின்னணியில் திறந்திருக்கும்: 1980கள் முதல் 2003 வரை (அதாவது பாக்ஸின் கட்டுரையின் தேதி).
- ஒருங்கிணைந்த கணினியின் உதவியுடன் மொழிக் கற்றல் (Integrated CALL) - இன்னும் அடையப்பட வேண்டும். மொழி ஆசிரியர்கள் எழுதும் நேரத்தில் இன்னும் திறந்த அழைப்பு கட்டத்தில் இருப்பதாக பாக்ஸ் வாதிட்டார், ஏனெனில் கணினியின் உதவியுடன் மொழிக் கற்றல் "இயல்பாக்கம்" ("normalisation") என்ற நிலையை எட்டியபோது மட்டுமே உண்மையான ஒருங்கிணைப்பு அடையப்பட்டதாகக் கூற முடியும் - எ.கா. கணினியின் உதயுடன் மொழிக் கல்வியைப் பயன்படுத்தும் போது பேனாவைப் பயன்படுத்துவது சாதாரணமானது.

ஃப்ளாஷ் கார்டுகள் (Flashcards)

கணினியின் உதயுடன் மொழிக் கற்றலின் அடிப்படை பயன்பாடு ஃபிளாஷ் கார்டுகளைப் பயன்படுத்தி சொற்றொகையை ஈட்டுவதில் உள்ளது; இதற்கு மிகவும் எளிய நிரல்கள் தேவை. இத்தகைய திட்டங்கள் பெரும்பாலும் இடைவெளி மீள்மையைப் (spaced repetition) பயன்படுத்துகின்ற ஒரு நுட்பமாகும், இதன் மூலம் கற்பவருக்குச் சொற்றொகை அலகுகள் வழங்கப்படுகின்றன, அவை நீண்டகால தக்கவைப்பு அடையும் வரை அதிக இடைவெளியில் நினைவகத்திற்கு உறுதியளிக்க வேண்டும். இது குறிப்பாக வெளிநாட்டு மொழிகளைக் கற்பவர்களுக்குத் திட்டமிடப்பட்ட பொதுவான அன்கி அல்லது சூப்பர்மெமோ தொகுப்பு (Anki or SuperMemo package) மற்றும் BYKI மற்றும் கட்டம்-6 (phase-6) போன்ற திட்டங்கள் உட்பட இடைவெளி மீள்மை அமைப்புகள் (spaced repetition systems (SRS)) என அழைக்கப்படும் பல பயன்பாடுகளின் வளர்ச்சிக்கு வழிவகுத்தது.

மென்பொருள் வடிவமைப்பு மற்றும் கற்பித்தல் (Software design and pedagogy)

எல்லாவற்றிற்கும் மேலாக, கணினியின் உதவியுடன் மொழிக் கற்றல் மென்பொருளை வடிவமைப்பதில் கல்வியியலுக்குக் கவனமான கருதல் தரப்பட வேண்டும்; ஆனால் கணினி உதவியுடன் மொழிக் கற்றல் மென்பொருளின் வெளியீட்டாளர்கள் அதன் விருப்பத்தை பொருட்படுத்தாமல் சமீபத்திய போக்கைப் பின்பற்ற முனைகிறார்கள். மேலும், இலக்கண-மொழிபெயர்ப்பு அணுகுமுறையிலிருந்து (grammar-translation approach), நேரடி முறை (direct method), கேட்பொலி மொழிய அணுகுமுறை (audio-lingualism/audio-lingual method) மற்றும் பலவிதமான அணுகுமுறைகள் வழி மிக சமீபத்திய தகவல்தொடர்பு அணுகுமுறை மற்றும் ஆக்கபூர்வவாதம் (constructivism) வரை வெளிநாட்டு மொழிகளைக் கற்பிப்பதற்கான அணுகுமுறைகள் தொடர்ந்து மாறிக்கொண்டே இருக்கின்றன, (டெகோ/Decoo 2001 2001).

கணினி உதவியுடன் மொழிக் கற்றல் மென்பொருளை வடிவமைப்பதும் உருவாக்குவதும் மிகவும் தேவையான செயல்பாடாகும்; இது பலவிதமான திறன்களை வேண்டுகிறது. முக்கிய கணினியின் உதயுடன் மொழிக் கல்வி மேம்பாட்டுத் திட்டங்கள் பொதுவாக ஒரு குழுவினரால் நிர்வகிக்கப்படுகின்றன:

- ஒரு பொருள்/பாட நிபுணர் (subject specialist) (உள்ளடக்க வழங்குநர் என்றும் அழைக்கப்படுகிறார்) - பொதுவாக ஒரு மொழி ஆசிரியர் - உள்ளடக்கம் மற்றும் கற்பித்தல் உள்ளீட்டை வழங்குவதற்கு பொறுப்பானவர். பெரிய கணினியின் உதயுடன் மொழிக் கல்வி திட்டங்களுக்கு ஒன்றுக்கு மேற்பட்ட பொருள்/பாட வல்லுநர்கள் தேவை.
- தேர்ந்தெடுக்கப்பட்ட நிரலாக்க மொழி அல்லது எழுதும் கருவியை நன்கு அறிந்த ஒரு புரோகிராமர்.
- ஒரு கிராஃபிக் டிசைனர்: படங்கள் மற்றும் ஐகான்களை தயாரிக்கவும் மற்றும் எழுத்துருக்கள், வண்ணம், திரை தளவமைப்பு போன்றவற்றுக்கு ஆலோசனை வழங்கவும்.
- ஒரு தொழில்முறை புகைப்படக்காரர் அல்லது, குறைந்தபட்சம், ஒரு நல்ல அமெச்சூர் புகைப்படக்காரர். கிராஃபிக் வடிவமைப்பாளர்கள் பெரும்பாலும் புகைப்படத்திலும் பின்னணி கொண்டவர்கள்.
- ஒலி பொறியாளர் மற்றும் வீடியோ தொழில்நுட்ப வல்லுநர்: தொகுப்பில் கணிசமான அளவு ஒலி மற்றும் வீடியோ இருக்க வேண்டுமானால் தேவைப்படுவார்கள்.
- ஒரு கற்பித்தல் வடிவமைப்பாளர்: ஒரு கணினியின் உதயுடன் மொழிக் கற்றல் தொகுப்பை உருவாக்குவது என்பது ஒரு உரை புத்தகத்தைக் கணினியில் வைப்பதை விட அதிகம்

கடினமானது. ஒரு கற்பித்தல் வடிவமைப்பாளர் புலனறிவு உளவியல் மற்றும் ஊடக தொழில்நுட்பத்தில் ஒரு பின்னணியைக் கொண்டிருப்பார், மேலும் தேர்ந்தெடுக்கப்பட்ட தொழில்நுட்பத்தின் (கிமெனோ & டேவிஸ்/Gimeno-Sanz & Davies 2010) பொருத்தமான பயன்பாடு குறித்து அணியில் உள்ள பொருள் வல்லுநர்களுக்கு ஆலோசனை வழங்க முடியும்.

எக்பர்ட் மற்றும் பிறர் (Egbert et al 2007) எட்டு நிபந்தனைகளின் இறுதி எனக் குறிப்பிடும் "உகந்த மொழி கற்றல் சூழல்களுக்கான நிபந்தனைகள்" ("Conditions for Optimal Language Learning Environments") என்ற கற்றல் சுயாட்சியை கணினி உதவியுடன் மொழிக் கற்றல் இயல்பாகவே ஆதரிக்கிறது. (2007). கற்பவர் தன்னுரிமையை (learner autonomy) உறுதியாகக் கட்டுப்படுத்துகிறது, இதனால் அவர் அல்லது அவள் "கற்றல் குறிக்கோள்களை தீர்மானிக்கிறார்கள்" (எக்பர்ட் மற்றும் பலர்/Egbert et al., 2007:8).

ஒரு எளிய படைபாக்கக் கருவியை (authoring tool) பயன்படுத்தி, வசதியான பாதையில் செல்லவும், பல தேர்வு மற்றும் இடைவெளியை நிரப்பும் பயிற்சிகளை உருவாக்கவும் கணினியின் உதயுடன் மொழிக் கல்வி மென்பொருளை வடிவமைக்கும்போது இது மிகவும் எளிதானது (பேங்க்ஸ்/Bangs 2011); ஆனால் கணினியின் உதவியுடன் மொழிக் கற்றல் இதை விட அதிகம் வலுவானது; எடுத்துக்காட்டாக, ஸ்டெப்-க்ரீனி (Stepp-Greany 2002) ஒரு ஆக்கபூர்வமான மற்றும் முழு மொழி தத்துவத்தையும் உள்ளடக்கிய ஒரு சூழலை உருவாக்குவதையும் நிர்வகிப்பதையும் விவரிக்கிறரார். ஆக்கபூர்வமான கோட்பாட்டின் (constructivist theory) படி, கற்பவர்கள் தங்கள் முந்தைய அனுபவத்திலிருந்து பெறப்பட்ட புதிய அறிவை "கட்டமைக்கும்" செயல்பாடுகளில் செயல்திறமுடைய பங்கேற்பாளர்கள் ஆவர். கற்றவர்களும் தங்கள் கற்றலுக்கான பொறுப்பை ஏற்றுக்கொள்கிறார்கள், மேலும் ஆசிரியர் அறிவைத் தூண்டுவதைக் (purveyor) காட்டிலும் வசதி செய்பவர் (facilitator) ஆவார். முழு மொழிக் கோட்பாடு (Whole language theory) ஆக்கபூர்வ வாதத்தைத் (constructivism) தழுவி, புரிந்துகொள்ளுதல், பேசுதல் மற்றும் எழுதுதல் ஆகியவற்றின் உயர் திறன்களை நோக்கி வழிநடத்த துணை திறன்களை வளர்ப்பதை விட, மொழி கற்றல் முழுமையிலிருந்து பகுதிக்கு (whole to the part) நகர்கிறது என்று கூறுகிறது. புரிந்துகொள்வது, பேசுவது, வாசிப்பது மற்றும் எழுதும் திறன் ஆகியவை ஒன்றோடொன்று தொடர்புடையவை மற்றும் சிக்கலான வழிகளில் ஒன்றையொன்று வலுப்படுத்துகின்றன என்பதையும் இது வலியுறுத்துகிறது. எனவே, மொழி ஈட்டல் என்பது செயலில் உள்ள ஒரு செயல்திறமுடைய

செயல்முறையாகும்; இதில் கற்றவர் குறிப்புகள் மற்றும் பொருண்மையில் (cues and meaning) கவனம் செலுத்துகிறார் மற்றும் அறிவார்ந்த யூகங்களை உருவாக்குகிறார். ஆக்கபூர்வமான மற்றும் முழு மொழிக் கோட்பாடுகளை உள்ளடக்கிய தொழில்நுட்ப சூழலில் பணிபுரியும் ஆசிரியர்கள் மீது கூடுதல் கோரிக்கைகள் வைக்கப்படுகின்றன. ஆசிரியர்களின் தொழில்முறை திறன்களின் வளர்ச்சியில் புதிய கல்வி மற்றும் தொழில்நுட்ப மற்றும் மேலாண்மை திறன் இருக்க வேண்டும். அத்தகைய சூழலில் ஆசிரியர் வசதி குறித்த பிரச்சினை குறித்து, ஆசிரியருக்கு முக்கிய பங்கு உண்டு, ஆனால் கற்பவரின் சுதந்திரத்திற்கான ஒரு சூழ்நிலையை உருவாக்கும் நோக்கத்திற்கும் ஆசிரியரின் இயல்பான பொறுப்பு உணர்வுகளுக்கும் இடையே ஒரு மோதல் இருக்கக்கூடும். கற்பவர்களின் எதிர்மறையான கருத்துக்களைத் தவிர்ப்பதற்காக, ஆசிரியர் தொடர்ந்து தங்கள் தேவைகளை, குறிப்பாகக் குறைந்த திறன் கொண்ட கற்றவர்களின் தேவைகளை நிவர்த்தி செய்வது மிகவும் முக்கியமானது என்று ஸ்டெப்-கிரேனி சுட்டிக்காட்டுகிறார்.

பல்லுடகம் (மல்டிமீடியா)

மொழி ஆசிரியர்கள் மிக நீண்ட காலமாக தொழில்நுட்பத்தைப் பயன்படுத்துபவர்களாக உள்ளனர். சொந்த மொழிபேசுபவர்களின் குரல்களின் பதிவுகளை மாணவர்களுக்கு வழங்குவதற்காக மொழி ஆசிரியர்களால் பயன்படுத்தப்பட்ட முதல் தொழில்நுட்ப உதவிகளில் கிராமபோன் பதிவுகள் இருந்தன; மேலும் வெளிநாட்டு வானொலி நிலையங்களிலிருந்து ஒளிபரப்புகள் ரீல்-டு-ரீல் டேப் ரெக்கார்டர்களில் பதிவு செய்யப் பயன்படுத்தப்பட்டன. ஸ்லைடு ப்ரொஜெக்டர்கள், ஃபிலிம்-ஸ்ட்ரிப் ப்ரொஜெக்டர்கள், ஃபிலிம் ப்ரொஜெக்டர்கள், வீடியோ கேசட் ரெக்கார்டர்கள் மற்றும் டிவிடி பிளேயர்கள் ஆகியவை வெளிநாட்டு மொழி வகுப்பறையில் பயன்படுத்தப்பட்டுள்ள தொழில்நுட்ப எய்ட்ஸின் பிற எடுத்துக்காட்டுகள். 1960களின் முற்பகுதியில், ஒருங்கிணைந்த படிப்புகள் (integrated courses) (அவை பெரும்பாலும் மல்டிமீடியா படிப்புகள் என்று விவரிக்கப்பட்டன) தோன்றத் தொடங்கின. அத்தகைய படிப்புகளின் எடுத்துக்காட்டுகள் ஈகூட்டர் எட் பார்லர் (Ecouter et Parler) (ஒரு பாடநூல் மற்றும் டேப் பதிவுகளை உள்ளடக்கியது) மற்றும் டாய்ச் டர்ச் டை ஆடியோவிசுவேல் முறை (Deutsch durch die audiovisuelle Methode) (ஒரு விளக்கப்பட பாடநூல், டேப் பதிவுகள் மற்றும் ஒரு திரைப்பட-துண்டு ஆகியவற்றைக் கொண்டுள்ளது - இது ஸ்ட்ரக்சுரோ-குளோபல் ஆடியோ-விஷுவலை அடிப்படையாகக் கொண்டது முறை).

1970கள் மற்றும் 1980களில் நிலையான மைக்ரோ கம்ப்யூட்டர்கள் ஒலியை உருவாக்க இயலாது, மேலும் அவை மோசமான கிராபிக்ஸ் திறனைக் கொண்டிருந்தன. இது மொழி ஆசிரியர்களுக்கு ஒரு படி பின்னோக்கி பிரதிபலித்தது, இந்த நேரத்தில் வெளிநாட்டு மொழி வகுப்பறையில் பல்வேறு ஊடகங்களைப் பயன்படுத்துவதற்குப் பழக்கமாகிவிட்டது. 1990களின் முற்பகுதியில் பல்லூடகக் கணினியின் வருகை ஒரு பெரிய திருப்புமுனையாக இருந்தது, ஏனெனில் இது உரை, படங்கள், ஒலி மற்றும் வீடியோ ஆகியவற்றை ஒரு சாதனத்தில் இணைக்க உதவியது மற்றும் கேட்பது, பேசுவது, வாசித்தல் மற்றும் எழுதுதல் ஆகிய நான்கு அடிப்படை திறன்களை ஒருங்கிணைத்தது (டேவிஸ்/Davies 2011:பிரிவு 1).

1990களின் நடுப்பகுதியில் இருந்து சிடி-ரோம் மற்றும் டிவிடியில் வெளியிடப்பட்ட பல்லூடக கணினிகளுக்கான கணினியின் உதயுடன் மொழிக் கல்வி திட்டங்களின் எடுத்துக்காட்டுகள் டேவிஸ் (Davies 2010: பிரிவு 3) விவரித்தன. கணினியின் உதயுடன் மொழிக் கல்வி நிரல்கள் இன்னும் குறுவட்டு மற்றும் டிவிடியில் வெளியிடப்படுகின்றன; ஆனால் வலை அடிப்படையிலான பல்லூடக கணினியின் உதயுடன் மொழிக் கல்வி இப்போது இந்த ஊடகங்களைக் கிட்டத்தட்ட மாற்றியமைத்துள்ளது.

பல்லூடக கணினி உதவியுடன் மொழிக் கற்றலின் வருகையைத் தொடர்ந்து, கல்வி நிறுவனங்களில் பல்லூடக மொழி மையங்கள் தோன்றத் தொடங்கின. பல்லூடக வசதிகள் உரை, படங்கள், ஒலி மற்றும் வீடியோ ஆகியவற்றின் ஒருங்கிணைப்புடன் மொழி கற்றலுக்கான பல வாய்ப்புகளை வழங்கினாலும், இந்த வாய்ப்புகள் பெரும்பாலும் முழுமையாகப் பயன்படுத்தப்படவில்லை. கணினி உதவியுடன் மொழிக் கற்றலின் முக்கிய வாக்குறுதிகளில் ஒன்று கற்றலைத் தனிப்பயனாக்கும் திறன், ஆனால், 1960கள் மற்றும் 1970களில் கல்வி நிறுவனங்களில் அறிமுகப்படுத்தப்பட்ட மொழி ஆய்வகங்களைப் போலவே, பல்லூடக மையங்களின் வசதிகளின் பயன்பாடு பெரும்பாலும் ஒரே தொடர்பயிற்சிகளைச் செய்யும் மாணவர்களின் வரிசைகளாக மாற்றப்பட்டுள்ளது (டேவிஸ்/Davies 2010: பிரிவு 3.1). எனவே பல்லூடக மையங்கள் மொழி ஆய்வகங்களைப் போலவே செல்லக்கூடும் என்ற ஆபத்து உள்ளது. 1970களில் ஒரு ஏற்ற காலத்தைத் தொடர்ந்து, மொழி ஆய்வகங்கள் விரைவாக வீழ்ச்சியடைந்தன. டேவிஸ் (Davies 1997: பக். 28) முக்கியமாக மொழி ஆய்வகங்களைச் செயல்பாட்டின் அடிப்படையிலும் புதிய வழிமுறைகளை வளர்ப்பதன் அடிப்படையிலும் பயன்படுத்த ஆசிரியர்களுக்குப் பயிற்சியளிக்கத் தவறியதன் மீது குற்றம் சாட்டினார்; ஆனால் மோசமான

நம்பகத்தன்மை, பொருட்களின் பற்றாக்குறை மற்றும் நல்ல யோசனைகள் இல்லாதது போன்ற பிற காரணிகள் இருந்தன (Davies 1997). திறனை

ஒரு பல்லூடக மொழி மையத்திற்கு வெளிநாட்டு மொழிகள் மற்றும் மொழி கற்பித்தல் முறை பற்றிய அறிவைப் பெற்ற ஊழியர்கள் மட்டுமல்லாமல், தொழில்நுட்ப அறிவு மற்றும் வரவுசெலவுத்திட்ட மேலாண்மைத் திறன் மற்றும் இவை அனைத்தையும், தொழில்நுட்பம் வழங்க இயலுவதை சாதமாகப் பயன்படுத்துவதன் ஆக்கபூர்வமான வழிகளாக ஒன்றிணைக்கும் திறன் கொண்ட ஊழியர்களும் தேவைப்படுகிறார்கள். ஒரு மைய மேலாளருக்கு பொதுவாகத் தொழில்நுட்ப ஆதரவு, வளங்களை நிர்வகித்தல் மற்றும் மாணவர்களைப் பயிற்றுவித்தல் போன்ற உதவியாளர்கள் தேவை. பல்லூடக மையங்கள் சுய ஆய்வு மற்றும் சுய இயக்க கற்றலுக்குத் தங்களை கடன் கொடுக்கின்றன, ஆனால் இது பெரும்பாலும் தவறாக புரிந்து கொள்ளப்படுகிறது. பல்லூடக மையத்தின் எளிமையான இருப்பு தானாகவே மாணவர்கள் சுயாதீனமாக கற்க வழிவகுக்காது. பொருட்களின் வளர்ச்சிக்கும், சுய ஆய்வுக்கு உகந்த சூழ்நிலையை உருவாக்குவதற்கும் நேரத்தின் குறிப்பிடத்தக்க முதலீடு அவசியம். துரதிர்ஷ்டவசமாக, நிர்வாகிகள் பெரும்பாலும் வன்பொருள் வாங்குவது மையத்தின் தேவைகளைப் பூர்த்திசெய்யும், அதன் பட்ஜெட்டில் 90% வன்பொருளுக்கு ஒதுக்குகிறது மற்றும் மென்பொருள் மற்றும் ஊழியர்களின் பயிற்சி தேவைகளை கிட்டத்தட்ட புறக்கணிக்கிறது (டேவிஸ் மற்றும் பலர்/Davies et al.. 2011: முன்னுரை). சுய அணுகல் மொழி கற்றல் மையங்கள் (Self-access language learning centres) அல்லது சுதந்திர கற்றல் மையங்கள் ஓரளவுக்குச் சுதந்திரமாகவும் ஓரளவுக்கு இந்தச் சிக்கல்களுக்குப் பதிலளிக்கும் வகையிலும் வெளிப்பட்டுள்ளன. சுய அணுகல் கற்றலில், வகுப்பறை கற்றலுக்கு மாறாக (அல்லது ஒரு நிரப்பியாக) சுய-இயக்க கற்றல் (self-directed learning) மூலம் கற்றல் சுயாட்சியை வளர்ப்பதில் கவனம் செலுத்தப்படுகிறது. பல மையங்களில் கற்பவர்கள் பொருட்களை அணுகி, தங்கள் கற்றலை சுதந்திரமாக நிர்வகிக்கிறார்கள், ஆனால் அவர்களுக்கு உதவிக்கு ஊழியர்களுக்கும் அணுகல் உள்ளது. பல சுய அணுகல் மையங்கள் தொழில்நுட்பத்தின் அதிக பயனர்கள் மற்றும் அவற்றில் அதிக எண்ணிக்கையிலானவர்கள் இப்போது ஆன்லைன் சுய அணுகல் கற்றல் வாய்ப்புகளை வழங்குகிறார்கள். சில மையங்கள் மொழி வகுப்பறையின் சூழலுக்கு வெளியே மொழி கற்றலை ஆதரிக்கும் புதிய வழிகளை உருவாக்கியுள்ளன ('மொழி ஆதரவு'/'language support' என்றும் அழைக்கப்படுகிறது) மாணவர்களின் சுய இயக்கிய கற்றலை கண்காணிக்க மென்பொருளை உருவாக்குவதன் மூலமும்

ஆசிரியர்களிடமிருந்து ஆன்லைன் ஆதரவை வழங்குவதன் மூலமும். சுய-இயக்கிய கற்றலில் (self-directed learning) மாணவர்களின் முயற்சிகளை ஆதரிப்பதற்காக மைய மேலாளர்கள் மற்றும் ஆதரவு ஊழியர்கள் புதிய பங்களிப்புகளை வரையறுக்க வேண்டியிருக்கலாம்: வி. மொஸன்-மெக்பெர்சன் & விஸ்மன்ஸ் (Mozzon-McPherson M. & Vismans R. 2001), ஒரு புதிய வேலை விளக்கத்தைக் குறிப்பிடும், அதாவது "மொழி" ஆலோசகர்" ("language adviser").

இணையதளம்

1990களின் முற்பகுதியில் உலகளாவிய வலை (World Wide Web) (இப்போது "வலை" என்று அழைக்கப்படுகிறது) தோன்றியது அனைத்து கணினி பயனர்களுக்கும் தகவல் தொடர்பு தொழில்நுட்பத்தைப் (communications technology) பயன்படுத்துவதில் குறிப்பிடத்தக்க மாற்றத்தைக் குறித்தது. மின்னஞ்சல் மற்றும் பிற வகையான மின்னணு தகவல்தொடர்புகள் (electronic communication) பல ஆண்டுகளாக இருந்தன; ஆனால் முதல் வரைகலை வலை உலாவியான (graphical Web browser) மொசைக் (Mosaic) 1993இல் தொடங்கப்பட்டது; நாம் மின்னணு முறையில் தொடர்பு கொள்ளும் வழிகளில் ஒரு தீவிர மாற்றத்தைக் கொண்டு வந்தது. பொது அரங்கில் வலை தொடங்கப்பட்டது உடனடியாக மொழி ஆசிரியர்களின் கவனத்தை ஈர்க்கத் தொடங்கியது. பல மொழி ஆசிரியர்கள் ஏற்கனவே தனிக் கணினிகளில் ஹைப்பர் டெக்ஸ்ட் என்ற கருத்தை நன்கு அறிந்திருந்தனர்; இது மொழி கற்பவர்களுக்குத் தொடர்ச்சியின்றி கட்டமைக்கப்பட்ட வாசிப்பு நடவடிக்கைகளை அமைப்பதைச் சாத்தியமாக்கியது; அதில் அவைகள் கணினி திரையில் காட்டப்படும் ஒரு பக்கத்தில் உரை அல்லது படங்களின் ஐடங்களை/உருப்படிகளைச் சுட்டிக்காட்ட முடியும் மற்றும் எந்த பக்கங்களுக்கும் கிளை விட இயலும், எ.கா. ஆப்பிள் மேக் கணினிகளில் ஹைப்பர் கார்டு திட்டத்தில் (HyperCard program) செயல்படுத்தப்பட்ட "ஸ்டேக்" ("stack") என்று அழைக்கப்படுகிறது. உலகளாவிய ஹைபர்டெக்ஸ்ட் (hypertext) அமைப்பை உருவாக்குவதன் மூலம் வலை இந்த ஒரு கட்டத்தை மேலும் எடுத்தது, இது ஒரு உரை அல்லது படத்தைச் சுட்டிக்காட்டி கிளிக் செய்வதன் மூலம் உலகில் எங்கிருந்தும் கணினிகளில் வெவ்வேறு பக்கங்களுக்குக் கிளைக்க பயனருக்கு உதவுகிறது. இது ஆசிரியர்கள் மற்றும் மாணவர்களுக்கு பல்வேறு வழிகளில் பயன்படுத்த இயலும் ஆயிரக்கணக்கான உண்மையான வெளிநாட்டு மொழி வலைத்தளங்களுக்கான அணுகலைத் திறந்தது. எவ்வாறாயினும் ஒரு சிக்கல் என்னவென்றால், வலை உலாவல் கட்டமைக்கப்படாத வழியில் பயன்படுத்தப்பட்டால் இது நல்ல நேரத்தை வீணடிக்க வழிவகுக்கும் (டேவிஸ்/ Davies

1997: பக். 42-43); மொழி ஆசிரியர்கள் கூடுதல் கட்டமைக்கப்பட்ட செயல்பாடுகள் மற்றும் ஆன்லைன் பயிற்சிகளை உருவாகுவதன் மூலம் இதற்குப் பதிலளித்ததனர் (லெலூப் & பொன்டெரியோ LeLoup & Ponterio 2003). டேவிஸ் (Davies 2010) 500க்கும் மேற்பட்ட வலைத்தளங்களை பட்டியலிடுகிறது, அங்கு ஆன்லைன் அகராதிகள் மற்றும் கலைக்களஞ்சியங்கள், ஒத்திசைவாளர்கள், மொழிபெயர்ப்பு கருவிகள்/எய்ட்ஸ் மற்றும் மொழி ஆசிரியர் மற்றும் கற்பவர்களுக்கு ஆர்வமுள்ள பிற இதர வளங்களுக்கான இணைப்புகளுடன் ஆன்லைன் பயிற்சிகளுக்கான இணைப்புகளைக் காணலாம்.

முதலில் யூரோகால் 1998 (EUROCALL 1998) மாநாட்டில் பகிரங்கமாக செயல் விளக்கமளி (இலவச) சூடான உருளைக்கிழங்கு (Hot Potatoes) (ஹோம்ஸ் & ஆர்னெய்ல் Holmes & Arneil) என்ற படைப்பாக்கக் கருவியின் (authoring tool வெளியீடு), மொழி ஆசிரியர்கள் தங்களது சொந்த ஆன்லைன் ஊடாடும் பயிற்சிகளை உருவாக்குவதைச் சாத்தியப்படுத்தியது. பிற பயனுள்ள கருவிகள் அதே ஆசிரியர்களால் தயாரிக்கப்பட்டன.

வலை அதன் ஆரம்ப நாட்களில் குறுவட்டு-ரோம் மற்றும் டிவிடியில் பல்லுடக கணினி உதவியுடன் மொழிக் கற்றல் உடன் தீவிரமாகப் போட்டியிட முடியவில்லை. ஒலி மற்றும் வீடியோ தரம் பெரும்பாலும் மோசமாக இருந்தது; மேலும் தொடர்பு மெதுவாக இருந்தது. ஆனால் இப்போது வலை முன்னேறியுள்ளது. ஒலி மற்றும் வீடியோ உயர் தரமானவை மற்றும் தொடர்பு மிகவும் மேம்பட்டது; இருப்பினும் இது போதுமான அலைவரிசை கிடைப்பதைப் பொறுத்தது; இது எப்போதும் இல்லை, குறிப்பாக தொலைதூர கிராமப்புறங்கள் மற்றும் வளரும் நாடுகளில். குறுந்தகடுகள் மற்றும் டிவிடிகள் இன்னும் உயர்ந்ததாக இருக்கும் ஒரு பகுதி கேட்பது/ பதிலளிப்பது/பின்னணி செயல்பாடுகளை வழங்குவதில் உள்ளது; இருப்பினும் வலையில் இதுபோன்ற நடவடிக்கைகள் தொடர்ந்து மேம்படுகின்றன.

2000களின் முற்பகுதியில் இருந்து வலை 2.0 பயன்பாடுகள் (Web 2.0 applications) என்று அழைக்கப்படுபவற்றின் வளர்ச்சியில் ஏற்றம் காணப்படுகிறது. பிரபலமான கருத்துக்கு மாறாக, வலை 2.0 என்பது வலையின் புதிய பதிப்பு அல்ல; மாறாக இது அடிப்படையில் ஒரு வழிச் செயல்முறையான (வலையிலிருந்து இறுதி பயனருக்கு) வலை உலாவலில் இருந்து பயன்படுத்துவதற்கு டெஸ்க்டாப் கணினியில் பயன்பாடுகளைப் உபயோகப்படுத்துவதைப் போலவே வலை பயன்பாடுகளின் உபயோகத்தின் முக்கியத்தும் மாற்றம் பெறுவதைக் குறிக்கிறது. இது அதிக தொடர்பு மற்றும் பகிர்வையும் குறிக்கிறது. வாக்கர், டேவிஸ் & ஹெவர்

(Davies & Hewer 2011: பிரிவு 2.1) மொழி ஆசிரியர்கள் பயன்படுத்தும் வலை 2.0 பயன்பாடுகளின் பின்வரும் எடுத்துக்காட்டுகளை பட்டியலிடுகின்றனர்:

- பட சேமிப்பு மற்றும் பகிர்வு
- சமூக புக்மார்க்கிங் (Social bookmarking)
- கலந்துரையாடல் பட்டியல்கள், வலைப்பதிவுகள், விக்கிகள், சமூக வலைப்பின்னல்
- அரட்டை அறைகள் (Chat rooms), MUD கள், MOO கள் மற்றும் MUVE கள் (மெய்நிகர் உலகங்கள்)
- பாட்காஸ்டிங் (Podcasting)
- ஆடியோ கருவிகள் (Audio tools)
- வீடியோ பகிர்வு பயன்பாடுகள் மற்றும் திரை பிடிப்பு கருவிகள் (வீடியோ திரை பிடிப்பு கருவிகள் மற்றும் ஸ்கிரீன்ஷாட் கருவிகள் இரண்டையும் குறிக்கும்)
- அனிமேஷன் கருவிகள் - காமிக் கீற்றுகள், திரைப்படங்கள் போன்றவை.
- மாஷப்ஸ் (Mashups)
- வலைப்பதிவு உதவி மொழி கற்றல் (Blog assisted language learning (BALL))

மொழி ஆசிரியர்களுக்கு வலை ஒரு முக்கிய மையமாக நிரூபிக்கப்பட்டுள்ளது என்பதில் சந்தேகமில்லை, அவர்கள் அதன் பரந்த அளவிலான வசதிகளை கற்பனையாகப் பயன்படுத்துகின்றனர்: டுடெனே (Dudeney 2007) மற்றும் தாமஸ்ஐப் (Thomas 2008) பார்க்கவும். எல்லாவற்றிற்கும் மேலாக, வலை 2.0 கருவிகளின் பயன்பாடு வகுப்பறையில் ஆசிரியரின் பங்கை கவனமாக மறுபரிசீலனை செய்ய வேண்டும் (ரிச்சர்ட்சன்/Richardson 2006).

6.6.8. உரை உருவாக்கம் (Text generation)

இரண்டு வகையான நடைமுறை உரை உருவாக்கும் நுட்பங்கள் ஏற்கனவே பொதுவான பயன்பாட்டில் உள்ளன மற்றும் நன்கு புரிந்து கொள்ளப்பட்டுள்ளன. முதலாவது முன்னர் தயாரிக்கப்பட்ட உரையைக் (அல்லது பதிவு செய்யப்பட்ட உரை) காட்டுவதாகும், இரண்டாவதாக அறிவு கட்டமைப்புகளின் நேரடி மொழிபெயர்ப்பின் மூலம் உரையை உருவாக்குகிறது. கணினி கணினி உரையை உருவாக்குவதற்கான எளிய மற்றும் பொதுவாக பயன்படுத்தப்படும் வழி, கணினியை செயல்படுத்துபவர்களுக்கு எந்த வகையான மொழியின் (ஆங்கிலம், தமிழ் போன்ற மொழிகளின்) வெளியீடு தேவைப்படும் என்பதை முன்கூட்டியே கண்டுபிடித்து அதை உரை சரங்களாக சேமிக்கவும். கணினி வெறுமனே சேமிக்கப்பட்ட உரையை

காண்பிக்கும். (எடுத்துக்காட்டாக, கிட்டத்தட்ட எல்லா பிழைச் செய்திகளும் இந்த வழியில் தயாரிக்கப்படுகின்றன.) ஒரு நிரல் இந்த வழியில் ஒரு மொழியின் உரையக் உருவாக்குவது ஒப்பீட்டளவில் எளிதானது; மேலும் விரும்பினால் சிக்கலான மற்றும் நேர்த்தியாக உரை எழுதப்படலாம். துரதிர்ஷ்டவசமாக, நிரல் பயன்படுத்தக்கூடிய எந்தவொரு அறிவு கட்டமைப்புகளிலிருந்தும் உரைச் சரங்களை சுயாதீனமாக மாற்ற முடியும் என்பதால், நிரல் என்ன செய்கிறது மற்றும் அது என்ன சொல்கிறது என்பதற்கு இடையில் நிலைத்தன்மைக்கு எந்த உத்தரவாதமும் இல்லை. பதிவு செய்யப்பட்ட உரையின் மற்றொரு சிக்கல் என்னவென்றால், எல்லா கேள்விகளும் பதில்களும் முன்கூட்டியே எதிர்பார்க்கப்பட வேண்டும்; பெரிய அமைப்புகளுக்கு, அது சாத்தியமற்றது என்பதை நிரூபிக்கலாம். இறுதியாக, ஒரு உரை சரம் கணினியைப் பொருத்தவரை வேறு எதையும் போல தோற்றமளிப்பதால், கணினி நிரல் அது என்ன சொல்கிறது என்பதற்கான கருத்தியல் மாதிரியை எளிதில் கொண்டிருக்க முடியாது. இதன் பொருள் ஒருவர் அதிக மூடுதலைக் காண்பார் என்று எதிர்பார்க்கக்கூடாது: உரைக்கான 100 தேவைகளை பூர்த்தி செய்வது இரண்டாவது 100ஐ மிகவும் எளிதாக்காது.

ஒரு மொழியின் வெளியீட்டை வழங்குவதற்கான மற்றொரு அணுகுமுறை நிரலின் அறிவு கட்டமைப்புகளை நேரடியாக அம்மொழிக்கு மொழிபெயர்ப்பதன் மூலம் உரையை உருவாக்குகிறது. இந்த முறை பதிவு செய்யப்பட்ட உரையில் பல சிக்கல்களை எதிர்கொள்கிறது, அதே நேரத்தில் அதன் சொந்த சிலவற்றை அறிமுகப்படுத்துகிறது. மாற்றப்படும் கட்டமைப்புகள் (அல்லது மொழிபெயர்க்கப்பட்டவை) நிரலின் பகுத்தறிவு செயல்பாட்டில் பயன்படுத்தப்படுவதால், நிலைத்தன்மையை உறுதிப்படுத்த முடியும். மூடுதலை உணர முடியும், ஏனென்றால் பெரிய அளவிலான அறிவு கட்டமைப்புகளைக் கையாள மாற்றங்கள் செய்யப்படுகின்றன. இருப்பினும், உருவாகும் மாற்றங்கள் பொதுவாக ஒப்பீட்டளவில் எளிமையானவை என்பதால், உரையின் தரம் அறிவு எவ்வாறு கட்டமைக்கப்பட்டுள்ளது என்பதைப் பொறுத்தது. உரை புரிந்துகொள்ளக்கூடியதாக இருக்க வேண்டுமானால், நிரல் பயன்படுத்தும் அறிவு கட்டமைக்கப்பட்டிருக்க வேண்டும், இதனால் அது உடனடியாக புரிந்துகொள்ளப்படும். இறுதியாக, இந்த நுட்பத்தைப் பயன்படுத்தும் அமைப்புகள் பொதுவாக மிகக் குறைந்த மொழியியல் அறிவைக் கொண்டிருக்கின்றன, எனவே அவை வாசிக்கக்கூடியவை, சொற்கள், தேவையற்றவை, வாசிக்கக்கூடியவை.

உரை உருவாக்கத்தின் நடைமுறை, சில சிறப்பியல்புகளைப் பகிர்ந்து கொள்ளும்:

1. அவற்றுக்கு குறுகிய உரைகள் தேவை: ஒன்று முதல் மூன்று வாக்கியங்கள்.

2. அவை விரும்பிய உரைகளுக்கு மிகவும் எளிமையான வழிகளில் நன்கு விரிவான நிரல் தரவு கட்டமைப்புகளைக் கொண்டுள்ளன.

3. முக்கியமான அறிவைத் தற்போதைய நுட்பங்களுடன் நன்கு குறிப்பிடலாம்.

4. வெளியீட்டின் வரையறுக்கப்பட்ட சரளமானது ஏற்றுக்கொள்ளத்தக்கது.

ஆங்கிலத்தில் அவற்றின் பகுத்தறிவை விளக்கும் சில "நிபுணர் அமைப்புகள்" ("expert systems") என்று அழைக்கப்படுபவை இந்த பண்புகளைக் கொண்டுள்ளன.

எதிர்காலத்தில் உரை உருவாக்கும் அமைப்புகள் அறிவுக் கட்டமைப்புகளை மொழிபெயர்ப்பதன் மூலம் உரையை உருவாக்கும் செயல்முறைகளைத் தொடர்ந்து சேர்க்கும் என்று நம்ப இயலும். இருப்பினும், அவை விரிவான மொழியியல் அறிவு, ஒரு கருத்தாடல் மாதிரி, வாசகரின் மாதிரி மற்றும் மேம்பட்ட அறிவு பிரதிநிதித்துவங்களைப் பயன்படுத்தும் பிற செயல்முறைகளுடன் ஒருங்கிணைக்கப்படும்.

தற்போதைய நுட்பங்களின் மட்டுப்படுத்தப்பட்ட திறன்களின் காரணமாக, உரை உருவாக்கும் பகுதியில் ஒரு பெஞ்ச்மார்க் பயன்பாட்டுத் திட்டத்தை உருவாக்கும் நோக்கில் ஒரு புதிய திட்டம் தற்போது எதிர்விளைவிக்கும் (counterproductive), ஏனெனில் இது சிறிய அல்லது மாற்றத்தக்க தொழில்நுட்பத்தை உருவாக்கும் மற்றும் பொதுவான சிக்கலான சமூகத்தின் முன்னேற்றத்திலிருந்து முன்னேறும் திறனிலிருந்து விலகிவிடும்.

உரை உருவாக்கும் வசதிக்கான அடிப்படை கூறுகள்

இப்போது கிடைக்கக்கூடிய மிகக் குறைந்த திறன்களை புதிய பணிகளுக்கு எளிதில் பயன்படுத்தக்கூடிய சரளமான, சக்திவாய்ந்த உரை உருவாக்கும் முறைகளாக எவ்வாறு உருவாக்க முடியும்? தற்போது கருதப்படும் அடிப்படை மாதிரி, தற்போதைய ஆராய்ச்சி நோக்கி நகர்கிறது, பின்வரும் பண்புகள் உள்ளன:

1. உரை உருவாக்கத்திற்கான பொறுப்பு பயன்பாட்டு புள்ளிகளில் சிதறடிக்கப்படுவதை விட உரை உருவாக்கம் தொகுதியில் உள்ளது.

2. உரை உருவாக்கம் தொகுதியின் ஒரு முக்கிய பகுதி சிறியது மற்றும் பல அமைப்புகள் மூலம் ஒட்டுமொத்தமாக உருவாக்கப்பட்டுள்ளது. போர்ட்டபிள் கூறுகளில் இலக்கு மொழியின் பொதுவான அறிவைக் குறிக்கும் ஒரு இலக்கணம் மற்றும் மொழியியல், பணி-சுயாதீனமான தகவல்களைக் கையாளும் செயல்முறைகள் உள்ளன.

ஒரு திறமையான உரை உருவாக்கும் வசதி பின்வரும் நான்கு அடையாளம் காணக்கூடிய கூறுகளைக் கொண்டிருக்க வேண்டும் என்பதையும், இவற்றின் வரம்புகள் எதிர்வரும் எதிர்காலத்திற்கான கலையின் ஒட்டுமொத்த நிலைக்கு வரம்புகளாக இருக்கும் என்பதையும் நாம் உறுதியாக உணராம்:

1. ஒரு விரிவான, மொழியியல் ரீதியாக நியாயப்படுத்தப்பட்ட இலக்கணம்.
2. பல்வேறு வகையான தகவல்களைக் குறியாக்கக்கூடிய அறிவு-பிரதிநிதித்துவ வடிவவாதம்.
3. உரையின் நோக்கம் கொண்ட வாசகரின் மாதிரி.
4. கருத்தாடல் அமைப்பு மற்றும் கட்டுப்பாட்டின் மாதிரி.

இவை ஒவ்வொன்றும் தற்போதுள்ள கணக்கீட்டு அல்லாத முன்னோடிகளை ஈர்க்கின்றன, மேலும் ஒவ்வொன்றிற்கும் உரை உருவாக்கும் பணிக்கு சில சிறப்பு தழுவல் தேவைப்படுகிறது.

6.7. பேச்சுத் தொழில்நுட்பத்தில் தரவுத்தொகுதி

பேச்சுப் புரிதல், பேச்சு தொகுப்பு மற்றும் பேச்சு உரையாடல் ஒழுங்குமுறைகள் போன்ற பேச்சு தொழில்நுட்பப் பயன்பாடுகளுக்குப் பெரும்பாலும் தனிப்பயனாக்கப்பட்ட விவரக்குறிப்புகளின் அடிப்படையில் தரவுத்தொகுதி தேவைப்படுகிறது. சமூகத்திற்குக் கிடைக்கக்கூடிய தற்போதைய தரவுத்தொகுதிகளான டிமிட் (TIMIT) மற்றும் எல்.டி.சி (LDC) மற்றும் எல்.டி.ஏ (ELDA) ஆகியவற்றால் விநியோகிக்கப்படும் பிற தரவுத்தொகுதிகள், அத்தகைய பயன்பாடுகளின் தேவைகளை எப்போதும் பூர்த்தி செய்யாது. இதுபோன்ற சந்தர்ப்பங்களில், டெவலப்பர்கள் தங்கள் சொந்த தரவுத்தொகுதியை உருவாக்க வேண்டும். எவ்வாறாயினும், மிகவும் தனிப்பயனாக்கப்பட்ட பேச்சு தரவுத்தொகுதியை உருவாக்குவது மிகவும் சிறிய மற்றும் நேரத்தைச் செலவழிக்கும் பணியாக இருக்கலாம், குறிப்பாக சிறிய நிறுவனங்களுக்கு. தரவுத்தொகுதி வடிவமைப்பு, மனித பொருள் ஆட்சேர்ப்பு, பதிவு செய்தல், தர உத்தரவாதம் மற்றும் சில சந்தர்ப்பங்களில், பிரிவு, படியெடுத்தல் மற்றும் சிறுகுறிப்பு போன்ற துணை பணிகளை உள்ளடக்கியுள்ளதால் இதற்கு மொழியியல், மேலாண்மை மற்றும் பொறியியல் ஆகியவற்றில் பலதரப்பட்ட நிபுணத்துவம் தேவைப்படுகிறது. ஒரு புதிய தரவு உருவாக்கம் மற்றும் உரிம மாதிரியின் கீழ் ஒரு வணிக நிறுவனத்திற்கான குறைந்த விலை மற்றும் அதிக-தனிப்பயனாக்கப்பட்ட பேச்சு தரவுத்தொகுதியை உருவாக்குவதில் எல்.டி.சி. ஈடுபட்டுள்ளது. இது

குறிப்பிட்ட தரவு கோருபவருக்கும் பொதுவான மொழியியல் தரவு பயனர் சமூகத்திற்கும் பயனளிக்கிறது.

பேச்சுத் தரவுத்தொகுதி

பேச்சுத் தரவுத்தொகுதி (speech corpus or spoken corpus) என்பது பேச்சு ஆடியோ கோப்புகள் மற்றும் உரை டிரான்ஸ்கிரிப்ஷன்களின் தரவுத்தளமாகும். பேச்சு தொழில்நுட்பத்தில், ஒலி மாதிரிகளை உருவாக்கப் பேச்சு தரவுத்தொகுதிகளைப் பயன்படுத்தப்படுகிறது (பின்னர் அவை பேச்சு அங்கீகாரம் அல்லது பேச்சாளர் அடையாள இயந்திரத்துடன் பயன்படுத்தப்படலாம்). மொழியியலில், ஒலியியல், உரையாடல் பகுப்பாய்வு, கிளைமொழியியல் மற்றும் பிற துறைகளில் ஆராய்ச்சி செய்ய பேசும் தரவுத்தொகுதிகள் பயன்படுத்தப்படுகிறது.

ஒரு தரவுத்தொகுதி அத்தகைய தரவுத்தளமாகும். கார்போரா (corpora) என்பது கார்பஸின் (corpus) பன்மை (அதாவது இது போன்ற பல தரவுத்தளங்கள்) வடிவமாகும்.

பேச்சு தரவுத்தொகுதிகளில் இரண்டு வகைகள் உள்ளன: வாசிப்புப் பேச்சு, தன்னிச்சையான பேச்சு.

வாசிப்புப் பேச்சு - இதில் பின்வருவன அடங்கும்:

புத்தக பகுதிகள்

செய்தி ஒளிபரப்பு

சொற்களின் பட்டியல்கள்

எண்களின் வரிசைமுறைகள்

தன்னிச்சையான பேச்சு - இதில் பின்வருவனஅடங்கும்:

உரையாடல்கள் (Dialogs) - இரண்டு அல்லது அதற்கு மேற்பட்ட நபர்களுக்கு இடையில் (கூட்டங்கள் அடங்கும்);

விவரிப்புகள் (Narratives) - ஒரு கதையைச் சொல்லும் ஒருவர் (அத்தகைய ஒரு கார்பஸ் பக்கி கார்பஸ்);

வரைபட பணிகள் (Map-tasks) - ஒரு நபர் ஒரு வரைபடத்தில் ஒரு வழியை இன்னொருவருக்கு விளக்குகிறார்;

நியமனம்-பணிகள் (Appointment-tasks) - தனிப்பட்ட அட்டவணைகளின் அடிப்படையில் ஒரு பொதுவான சந்திப்பு நேரத்தைக் கண்டுபிடிக்க இரண்டு பேர் முயற்சி செய்கிறார்கள்.

ஒரு சிறப்பு வகையான பேச்சுத் தரவுத்தொகுதிகள் வெளிநாட்டு உச்சரிப்புடன் பேச்சைக் கொண்டிருக்கும் சொந்தமற்ற பேச்சு தரவுத்தளங்கள் ஆகும்.

பேச்சுத் தரவுத்தொகுதி பின்வருவனவற்றை உருவாக்கப் பயன்படுத்தப்படும்.

1. பேச்சுத் தொழில்நுட்பத்திற்குப் பொதுவான சட்டகத்தின் உருவாக்கம் (Development of general framework for speech technology)

2. கிளைமொழிகளில் ஒலியியல், சொல்லியியல் மற்றும் உச்சரிப்பியல் வேறுபாடுகள்(Phonetic, lexical, and pronunciation variability in dialectal versions)

3. தானியங்கு பேச்சு புரிந்துக்கொள்ளுதல் செயல்பாடு (Automatic speech recognition tasks)

4. தானியங்கு பேச்சு உருவாக்கும் ஒழுங்குமுறைகள் (Automatic speech synthesis systems)

5. தானியங்கு பகுப்பாய்வு ஒழுங்குமுறைகள் Automatic speech processing systems

6. தானியங்கு பேசுபவரை அடையாளம் காணும் ஒழுங்குமுறைகள் (Automatic speaker identification system)

7. பேச்சுக் குறைபாடுகளைச் சரிசெய்தல் (Repairing speech disorders)

6.7.1. பேச்சுத் தொழில்நுட்பத்திற்குப் பொதுவான சட்டகத்தின் உருவாக்கம் (Development of general framework for speech technology)

பேச்சுத் தொழில்நுட்பத்திற்குப் பொதுவான சட்டகத்தின் உருவாக்கத்திற்குத் தரவுத்தொகுதி பயன்படுத்தப்படுகின்றது. குஹ்னின் (Kuhn, 2012) அறிவியல் புரட்சிகளின் கோட்பாட்டின் படி, அறிவியல் நடைமுறையில் உள்ள அறிவியல் முன்மாதிரிகளின் புரட்சிகர மாற்றங்கள் மூலம் முன்னேறுகிறது, அங்கு ஒரு முன்னுதாரணம் ஒரு அறிவியல் சமூகத்திற்கு பொதுவான நம்பிக்கைகள் மற்றும் மதிப்புகள் மற்றும் தொழில்நுட்ப மற்றும் வழிமுறை நடைமுறைகளை குறிக்கிறது. அறிவியல் சவால்களைத் தீர்ப்பதற்கான பிரேம்கள்/சட்டகங்கள் மற்றும் மாதிரிகளை முன்னுதாரணங்கள் வரையறுக்கின்றன. புதிய தீர்வுகள் புதிய தலைமுறையினருடன் வந்துள்ளன, அவை புதிய உண்மைகளையும் இடைநிலை அணுகுமுறைகளையும் ஏற்கத் தயாராக உள்ளன. புதிய முன்னுதாரணங்கள் (paradigms) திடீரென தோன்றும் மற்றும் ஒரு அறிவியல் சிக்கலுக்குப் புதிய விளக்கங்க்களை வழங்குகின்றன, இது ஒரு குறிப்பிட்ட மற்றும் சிறப்பு அறிவின் சினெர்ஜி அடிப்படையில் செயல்பாட்டு மற்றும்

ஒத்திசைவான ஒற்றுமைக்கு ஒருங்கிணைக்கப்படுகிறது. பேச்சு தொழில்நுட்ப சமூகம் பேசும் மொழிச் செயலாக்கத்தை ஒரு இடைநிலை ஆராய்ச்சி பகுதி என ஆராய்கிறது. பேச்சு உற்பத்தி மற்றும் செவிவழி உணர்வின் அறிவை அடிப்படையாகக் கொண்ட முக்கிய அறிவியல் முன்மாதிரிகளின் ஒரு குறுகிய பின்னோக்கிப் பிறகு, நரம்பியல் மற்றும் மேம்பட்ட சமிக்ஞை செயலாக்கம் தொடர்பான புதிய இயந்திர கற்றல் முன்னுதாரணத்தின் அடிப்படையில் புதிய சாதனைகள் மற்றும் முன்னோக்குகளை முன்வைக்கிறது.

பேச்சு சமிக்ஞை செயலாக்க ஆராய்ச்சியின் வேர்கள் பேச்சு சமிக்ஞை டிஜிட்டல் மயமாக்கலின் தேவைகளுடன் நெருக்கமாக தொடர்புடையவை. நேச நாடுகளுக்கிடையில் பாதுகாப்பான தகவல்தொடர்பு தேவைப்படுவதால் இரண்டாம் உலகப் போரின்போது முன்னோடி தீர்வுகள் பயன்படுத்தப்பட்டன. இந்த அமைப்புக்கு SIGSALY என்று பெயரிடப்பட்டது, மேலும் இது டிஜிட்டல் கருவிகளைப் பயன்படுத்தி முதல் குரல் பரிமாற்றத்தை செயல்படுத்த துடிப்பு-குறியீடு பண்பேற்றத்தை (pulse-code modulation (PCM) (Paul 2019) பயன்படுத்தியது. அடுத்த தசாப்தங்களில், பரந்த அளவிலான பேச்சு சமிக்ஞை மாறுபாடுகளில் (Jayant and Noll 2012; Chu 2003; Hanzo, Somerville, and Woodard 2007; Kondoz, 2004) புனரமைக்கப்பட்ட பேச்சு சமிக்ஞையின் உயர் தரத்தை வழங்குவதற்காக டிஜிட்டல் தொலைபேசியின் விதிகளை தரப்படுத்துவதில் ஆராய்ச்சியாளர்களின் கவனம் இருந்தது. இந்த அமைப்புகள் தொடர்பான சுருக்க முன்னுதாரணங்கள் பல தசாப்தங்களாக கணிசமாக மாறவில்லை. குறிப்பாக, பெரும் முடிவில் சமிக்ஞை தரத்தை மேம்படுத்துவதற்கு அல்லது தேவையான பிட் வீதத்தைக் குறைப்பதை நோக்கி ஆராய்ச்சியின் கவனம் சற்று நகர்த்தப்பட்டுள்ளது (Peric and Nikolic 2012; Nikolic, Peric, and Jovanovic 2016; Nikolic and Peric 2008; Peric and Nikolic 2012; Ordentlich and Erez 2019; Farias and J. M. Brossier 2016). இருப்பினும், கடந்த தசாப்தத்தில் கணினி தொழில்நுட்பத்தின் குறிப்பிடத்தக்க வளர்ச்சியானது தகவமைப்பு இயந்திர கற்றல் முறைகள் (adaptive machine learning methods) (Farias & Brossier 2016) உள்ளிட்ட மேம்பட்ட பேச்சு சமிக்ஞை செயலாக்கத்திற்கான புதிய அணுகுமுறைகள் குறித்த ஆராய்ச்சியை செயல்படுத்த உதவியது (McLoughlin 2016). சமீபத்திய போக்குகள் அறிவாற்றல் பேச்சு குறியீட்டு முறையை உள்ளடக்குகின்றன, இதனால் புலனுணர்வு (செவிப்புலன்) இலிருந்து அறிவாற்றல் (செவிவழி மற்றும் கார்டிகல்) பேச்சு சமிக்ஞை செயலாக்கம் (Cernak, Asaei, and Hyafil 2018) நோக்கி ஒரு முன்னுதாரண மாற்றம் உள்ளது.

நவீன பேச்சு தொழில்நுட்ப அமைப்புகள் மல்டிமோடல் சிக்னல் செயலாக்கம் (multimodal signal processing) மற்றும் செயற்கை நுண்ணறிவு (artificial intelligence) ஆகிய துறைகளில் இடைநிலை ஆராய்ச்சியை நம்பியுள்ளன, மேலும் பல்வேறு சிக்கல்களைத் தீர்க்கும் நோக்கத்துடன் பல முறைகள் மற்றும் வழிமுறைகள் உருவாக்கப்பட்டுள்ளன: பேச்சு புரிதல் (speech recognition) மற்றும் தொகுப்பு (synthesis) அடிப்படையிலான உரையாடல் அமைப்புகள், உணர்ச்சிபூர்வமான பேச்சு உட்பட, பேச்சாளர் அடையாளம் மற்றும் சரிபார்ப்பு (speaker identification and verification), அத்துடன் பேச்சு சமிக்ஞை குறியீட்டு மற்றும் பரிமாற்றம் (speech signal coding and transmission), சப்தம் குறைத்தல், சத்தத்தின் முன்னிலையில் சிக்னல்களைக் கண்டறிதல், தர மேம்பாடு, மற்றும் மனித குரலின் பகுப்பாய்வின் அடிப்படையில் மருத்துவ கண்டறியும் முன்னிலையில் சிக்னல்களைக் கண்டறிதல். இந்த பேச்சு தொழில்நுட்ப தலைப்புகளில் பெரும்பாலானவற்றின் சமீபத்திய முன்னேற்றம் பின்வரும் பிரிவுகளில் மேலும் விவரங்களில் விவாதிக்கப்படும்.

பேச்சு மொழிச் செயலாக்கம் (Spoken language processing (SLP/எஸ்.எல்.பி) என்பது ஒரு இடைநிலை ஆராய்ச்சி பகுதி; இது கணக்கீட்டு நுண்ணறிவின் பண்புகளைக் கொண்டுள்ளது. பேச்சு மொழிச் செயலாக்கம் மொழியியல், உளவியல், பொறியியல் மற்றும் செயற்கை நுண்ணறிவு (AI) ஆகியவற்றின் ஊடறுத்தலில் (intersections) உள்ளது (Moore, 2005). மேம்பட்ட சமிக்ஞை செயலாக்கம் மற்றும் இயந்திர கற்றல் முறைகள் பேச்சு மொழிச் செயலாக்கத்தின் இடைநிலை தன்மைக்கு ஏற்றுக்கொள்ளப்பட்ட பார்வையில் நிலைநிறுத்தப்படுகின்றன, மேலும் வெவ்வேறு பிரிவுகளின் ஒன்றோடொன்று இணைப்புகள் மற்றும் ஊடறுத்தல்கள் ஒரு புதிய பார்வையில் காட்டப்பட்டு வழங்கப்படுகின்றன. “பேட்டர்ன் பிராசசிங்” என்ற அசல் சொல்லைப் பயன்படுத்துவதற்குப் பதிலாக, பொறியியல் மற்றும் AI துறைகளின் சமூகத்திற்கு இடையிலான ஒன்றுடன் ஒன்று குறிக்கும் “சிக்னல் பிராசசிங் மற்றும் மெஷின் கற்றல் (SP&ML)” என்ற பொதுவான சொல் தேர்ந்தெடுக்கப்பட்டு பயன்படுத்தப்படுகிறது. மொழியியல் அம்சங்கள் சேர்க்கப்பட்டுள்ளதால், அவை இயற்கையான மொழி செயலாக்க (என்.எல்.பி) புலத்தை உருவாக்குகின்றன. மனித-கணினி தொடர்பு (Human-computer interaction (HCI/எச்.சி.ஐ) பொறியியல் மற்றும் உளவியல் துறைகளில் இருந்து அனுபவத்தையும் முறையையும் ஈர்க்கிறது, மேலும் மொழியியலில் உள்ள அறிவுடன், அவை உரையாடல் அமைப்புகளின் ஆய்வு மற்றும் வளர்ச்சிக்கு ஒரு அடிப்படையாக அமைகின்றன.

பேச்சு தொழில்நுட்பங்கள் பேச்சுச் சமிக்ஞை செயலாக்கத்தை அடிப்படையாகக் கொண்டவை; அவை பரந்த அளவிலான தலைப்புகளைக் கொண்டுள்ளன, அதே நேரத்தில் அதிக நிபுணத்துவம் வாய்ந்த மூன்று பகுதிகளில் உள்ளது:

- அடிப்படைத் தலைப்புகள் (பேச்சு பகுப்பாய்வு மற்றும் தொகுப்பு, ஒலி அலைகள் மற்றும் பேச்சு அம்சங்கள், பேச்சு உற்பத்தி, செவிப்புலன் கருத்து மற்றும் மொழியியல் அம்சம் உட்பட புலனறிவு)
- பேச்சு புரிதல் மற்றும் உரையிலிருந்து-பேச்சு தொகுப்பு (உணர்ச்சி பேச்சு அங்கீகாரம் மற்றும் குரல் மற்றும் பாணி மாற்றம் உள்ளிட்ட உரை-க்கு-பேச்சு தொகுப்பு) அடிப்படையிலான உரையாடல் அமைப்புகள்
- பேச்சு குறியீட்டு முறை (Speech coding), சுருக்கம் மற்றும் பரிமாற்றம்

பேச்சு தொழில்நுட்ப புலங்கள் ஒரு ஒருங்கிணைந்த கட்டமைப்பாக வழங்கப்படுகின்றன, அவை மூடப்பட்ட தலைப்புகளை உள்ளடக்கியது, அவற்றின் நிரப்புத்தன்மை, வரம்புகள் மற்றும் எல்லைகள், ஒன்றோடொன்று இணைப்புகள் மற்றும் பேச்சு மொழிச் செயலாக்கத்தின் இடைநிலை பகுதியில் உள்ள ஊடறுத்தல்களைக் காட்டுகிறது.

பேச்சு தொழில்நுட்பங்கள் மற்றும் அவற்றுடன் தொடர்புடைய பிற அறிவியல் பகுதிகளின் சமீபத்திய வளர்ச்சி, பெரும்பாலும் புதிய இயந்திர கற்றல் முன்னுதாரணத்தின் வளர்ச்சியின் காரணமாக, இந்த களத்தில் மிகப்பெரிய தாக்கத்தை ஏற்படுத்தியுள்ளது. இயற்கையான பேச்சு உற்பத்தி மற்றும் பேச்சு உணர்தல் தவிர, பேசு மொழி புரிதல் மற்றும் உருவாக்கம் மொழி தொழில்நுட்பங்கள் உள்ளிட்ட எதிர்கால எச்.சி.ஐ அமைப்புகளுக்கு பேச்சு தகவல்தொடர்பு அறிவாற்றல் அம்சங்களைப் புரிந்துகொள்வது மிகவும் முக்கியமானது. இயந்திர கற்றல் முன்னுதாரணம் தானியங்கி பேச்சு புரிதல் (automatic speech recognition (ASR/ஏ.எஸ்.ஆர்) மற்றும் உரையிலிருந்து பேச்சு உருவாக்கம் (text-to-speech synthesis (TTS/டி.டி.எஸ்) ஆகியவற்றில் அடிப்படை பேச்சு தொழில்நுட்பங்களாக பெரும் தாக்கத்தை ஏற்படுத்தியுள்ளது. ஆழ்ந்த கற்றல் மற்றும் தகவமைப்பு வழிமுறைகளை அடிப்படையாகக் கொண்ட ஏ.எஸ்.ஆர் அமைப்புகள் சிக்கலான ஒலியியல் சூழல்களில் தன்னிச்சையான பேச்சை அங்கீகரிக்க முடியும் என்று எதிர்பார்க்கப்படுகிறது, துல்லியத்துடன் மனிதர்களின் தொடர்புடைய திறனை விஞ்சிவிடும். செயற்கை பேச்சு ஏற்கனவே மனித பேச்சிலிருந்து வேறுபடுத்துவது கடினம் அல்லது சாத்தியமற்றது போன்ற தரத்தை அடைந்துள்ளது. பேச்சாளர் மற்றும் பாணியை மாற்றுவதற்கான நெகிழ்வுத்தன்மையுடன், மனித-கணினி ஊடாட்டம் (Human-computer interaction) மனித-மனித ஊடாட்டங்களைப் போலவே இனிமையாகவும் இயற்கையாகவும் மாறி வருகிறது. மேற்பார்வை

செய்யப்படாத மற்றும் வலுவூட்டல் அடிப்படையிலான இயந்திர கற்றல் வழிமுறைகளும் மேலும் மேம்படும், இது பெரிய தரவுத் தொகுப்புகள் கிடைக்காத குறைந்த மூலவள மொழிகளுக்கான (under-resourced languages) பேச்சு புரிதல் மற்றும் பேச்சு உருவாக்கதிற்கான பகுப்பாய்விலும் முன்னேற்றத்தைக் கொண்டுவரும். துல்லியமான தானியங்கி பேச்சு சமிக்ஞை கண்டறிதல் மற்றும் தொகுப்புக்கு டிஜிட்டல் மயமாக்கப்பட்ட மற்றும் சுருக்கப்பட்ட பேச்சு சமிக்ஞையின் தரம் முக்கியமானது என்பதால் பேச்சு குறியீட்டு நுட்பங்கள் மற்றும் தகவமைப்பு அளவீட்டு அளவீடுகளின் தற்போதைய முன்னேற்றம் பற்றிய ஒரு கண்ணோட்டம் தேவை. இந்த நுட்பங்கள் பரந்த அளவிலான பேச்சு சமிக்ஞை மாறுபாடுகளில் வலுவானதாக வடிவமைக்கப்படலாம் அல்லது சட்ட-தகவமைப்புக்கு ஏற்றதாக இருக்க முடியும் என்றாலும், பிரபலமடைவதற்கான இயந்திர கற்றல் கருவிகள் புதிய தீர்வுகளுக்கு வழிவகுக்கும் என்று ஒருவர் எதிர்பார்க்கலாம், இது பல்வேறு அமைப்புகளின் செயல்திறனை மேம்படுத்தும் முன்கணிப்பு குணகங்களைத் தழுவுதல். முடிவுக்கு, இயந்திர கற்றல் முன்மாதிரிகள் காரணமாக பேச்சு சமிக்ஞை செயலாக்கத் துறையில் பெருகிய முறையில் விரைவான முன்னேற்றத்தை நாங்கள் காண்கிறோம், மேலும் அவை அடுத்து என்ன கொண்டு வரும், எவ்வளவு விரைவில் எதிர்பார்க்கலாம் என்று கணிப்பது மிகவும் கடினமாகத் தெரிகிறது.

6.7.2. கிளைமொழிகளில் ஒலியியல், சொல்லியியல் மற்றும் உச்சரிப்பியல் வேறுபாடுகள் (Phonetic, lexical, and pronunciation variability in dialectal versions)

கிளைமொழிகளில் ஒலியியல், சொல்லியியல் மற்றும் உச்சரிப்பியல் வேறுபாடுகளை துல்லியமாக அறிந்துகொள்ளவும் கணினிமொழியியல் அல்லது மொழித்தொழில் நுட்பம் அடிப்படையிலான ஆய்வுகளை மேற்கொள்ளவும் கிளைமொழிகளின் தரவுத்தொகுதிகள் பெரிதும் பயன்படும். நிலைபேறுபெற்ற மொழிகளுக்கு மட்டுமன்றி கிளைமொழிகளுக்கும் டிஜிட்டல் தரவுத்தொகுதிகளின் சேகரிப்பும் இலக்கணத்தகவல்களுக்கான அடையாளப்படுத்தலும் மிக முக்கியமாகும். கணினியின் வரவுக்கு முன்னால் கிளைமொழிகளின் தரவுகள் மனித முயற்சியால் திரட்டப்பட்டு காகிதங்களில் எழுதப்பட்டு ஆயப்பட்டு வந்தன. கணினியின் வரவுக்குப் பின்னர் இத்தகைய தரவுகள் தரவுத்தொகுதிகளாகத் திரட்டப்பட்டு கணினியில் சேகரிக்கப்பட்டு இலக்கணக் குறிப்புகளுக்காக அடையாளப்படுத்தப்பட்டு ஆயப்படுகின்றன. இத்தகைய டிஜிட்டல்/மின் தரவுத்தொகுதி கிளைமொழிகளில் ஒலியியல், சொல்லியியல் மற்றும் உச்சரிப்பியல் வேறுபாடுகளை அறிந்துகொள்ளவும் அதன் அடிப்படையில் பேசுத் தொழிநுட்ப ஆய்வுகளை மேற்கொள்ளவும் பயனுள்ளதாக அமையும்.

6.7.3. தானியங்குப் பேச்சு புரிந்துக்கொள்ளுதல் செயல்பாடு (Automatic speech recognition tasks)

பேச்சு புரிந்துக்கொள்ளுதல் என்பது கணினி அறிவியல் மற்றும் கணினி மொழியியலின் ஒரு இடைநிலை துணைத் துறையாகும், இது கணினிகள் மூலம் பேசும் மொழியை உரையாக புரிந்துகொள்ளவும் மொழிபெயர்க்கவும் உதவும் முறைகள் மற்றும் தொழில்நுட்பங்களை உருவாக்குகிறது. இது தானியங்கி பேச்சு புரிந்துகொள்ளுதல் (Automatic speech recognition ஏ.எஸ்.ஆர்), கணினி பேச்சு அங்கீகாரம் அல்லது பேச்சிலிருந்து உரை (speech to text (STT/ எஸ்.டி.டி) என்றும் அழைக்கப்படுகிறது. இது கணினி அறிவியல், மொழியியல் மற்றும் கணினி பொறியியல் துறைகளில் அறிவு மற்றும் ஆராய்ச்சியை ஒருங்கிணைக்கிறது.

சில பேச்சுப் புரிந்துக்கொள்ளுதல் அமைப்புகளுக்கு "பயிற்சி" ("பதிவு" என்றும் அழைக்கப்படுகிறது) தேவைப்படுகிறது; இதில் ஒரு தனிப்பட்ட பேசுபவர் உரை அல்லது தனிமைப்படுத்தப்பட்ட சொற்றொகையைக் கணினியில் படிக்கிறார். கணினி நபரின் குறிப்பிட்ட குரலை பகுப்பாய்வு செய்து, அந்த நபரின் பேச்சின் புரிதலை செம்மைசெய்யப் பயன்படுத்துகிறது, இதன் விளைவாக துல்லியம் அதிகரிக்கும். பயிற்சியைப் பயன்படுத்தாத அமைப்புகள் "பேசுபவர் சுதந்திர" அமைப்புகள் என்று அழைக்கப்படுகின்றன. பயிற்சியைப் பயன்படுத்தும் அமைப்புகள் "பேசுபவர் சார்பு" என்று அழைக்கப்படுகின்றன.

பேச்சு புரிந்துக்கொள்ளுதல் பயன்பாடுகளில் குரல் பயனர் இடைமுகங்களான குரல் டயலிங் (எ.கா. "வீட்டிற்கு அழை"), அழைப்பு ரூட்டிங் (எ.கா. "நான் ஒரு சேகரிப்பு அழைப்பை செய்ய விரும்புகிறேன்"), டோமோடிக் பயன்பாட்டுக் கட்டுப்பாடு, தேடல் முக்கிய சொற்கள் (எ.கா. பேசப்பட்ட குறிப்பிட்ட சொற்களைக் கொண்ட போட்காஸ்டைக் கண்டுபிடி), எளிய தரவு உள்ளீடு (எ.கா., கிரெடிட் கார்டு எண்ணை உள்ளிடுதல்), கட்டமைக்கப்பட்ட ஆவணங்களைத் தயாரித்தல் (எ.கா. கதிரியக்க அறிக்கை), பேச்சாளர் பண்புகளை தீர்மானித்தல் (Nguyen 2010), பேச்சிலிருந்து-உரை செயலாக்கம் speech-to-text processing (எ.கா., சொல் செயலிகள் அல்லது மின்னஞ்சல்கள்), மற்றும் விமானம் (பொதுவாக நேரடி குரல் உள்ளீடு என்று அழைக்கப்படுகிறது).

குரல் புரிந்துக்கொள்ளுதல் அல்லது பேசுபவர் அடையாளம் காணல் என்பது அவர்கள் சொல்வதைக் காட்டிலும் பேச்சாளரை அடையாளம் காண்பதைக் குறிக்கிறது. பேச்சாளரை புரிந்துக்கொள்ளுதல் ஒரு குறிப்பிட்ட நபரின் குரலில் பயிற்சியளிக்கப்பட்ட ஒழுங்குமுறைகளில் பேச்சை மொழிபெயர்க்கும் பணியை எளிதாக்குகிறது அல்லது பாதுகாப்புச் செயல்பாட்டின் ஒரு பகுதியாக பேசுபவரின் அடையாளத்தை புரிந்துகொள்ள அல்லது சரிபார்க்க இது பயன்படுத்தப்படலாம்.

தொழில்நுட்ப கண்ணோட்டத்தில், பேச்சு புரிந்துகொள்ளுதல் பல பெரிய கண்டுபிடிப்புகளுடன் நீண்ட வரலாற்றைக் கொண்டுள்ளது. மிக சமீபத்தில், ஆழ்ந்த கற்றல் மற்றும் பெரிய தரவுகளின் முன்னேற்றங்களிலிருந்து இந்தத் துறை பயனடைந்துள்ளது. இந்த துறையில் வெளியிடப்பட்ட கல்வித் தாள்களின் எழுச்சியால் மட்டுமல்லாமல், மிக முக்கியமாக உலகளாவிய தொழில்துறை பேச்சு புரிந்துகொள்ளுதல் முறைகளை வடிவமைப்பதிலும் பயன்படுத்துவதிலும் பல்வேறு ஆழமான கற்றல் முறைகளைப் பின்பற்றுவதன் மூலம் முன்னேற்றங்கள் சாட்சியமளிக்கின்றன.

வரலாறு

வளர்ச்சியின் முக்கிய பகுதிகள்: சொற்றொகை அளவு, பேச்சாளர் சுதந்திரம் மற்றும் செயலாக்க வேகம். (விக்கிபீடியாவில் “Speech recognition” என்ற தலைப்பில் கூறப்பட்ட செய்திகளிலிருந்து எடுத்தாளப்பட்டுள்ளது.)

1970க்கு முன்

- 1952 - மூன்று பெல் லேப்ஸ் ஆராய்ச்சியாளர்கள், ஸ்டீபன் பாலாஷேக், ஆர். பிதுல்ப், மற்றும் கே. எச். டேவிஸ் ஆகியோர் ஒற்றை பேச்சாளர் இலக்கப் புரிந்துகொள்ளலுக்காக (single-speaker digit recognition) "ஆட்ரி" ("Audrey") என்ற அமைப்பை உருவாக்கினர். அவற்றின் அமைப்பு ஒவ்வொரு உரையின் சக்தி நிறமாலையில் ஒத்ததிர்வு தொகுப்புகளை (formants) இடங்கண்டது.
- 1960 - குன்னர் ஃபாண்ட் (Gunnar Fant) பேச்சு உற்பத்தியின் மூல-வடிகட்டி மாதிரியை உருவாக்கி (source-filter model of speech production) வெளியிட்டார்.
- 1962 - ஐபிஎம் அதன் 16-சொல் "ஷூ பாக்ஸ்" (16-word "Shoebox") இயந்திரத்தின் பேச்சு புரிதல் திறனை 1962 உலகக் கண்காட்சியில் செயல்விளக்கமளித்தது (Melanie Pinola 2011).
- 1966 - பேச்சு குறியீட்டு முறையான நேரியல் முன்கணிப்பு குறியீட்டு முறை (Linear predictive coding (LPC/எல்பிசி)) (Gray 2010). பேச்சு புரிதலில் பணிபுரியும் போது முதன்முதலில் நாகோயா பல்கலைக்கழகத்தின் (Nagoya University) ஃபுமிடாடா இடாகுரா (Fumitada Itakura) மற்றும் நிப்பான் டெலிகிராப் மற்றும் டெலிபோனின் (Nippon Telegraph and Telephone (NTT/என்.டி.டி) ஷூசோ சைட்டோ (Shuzo Saito) ஆகியோரால் முன்மொழியப்பட்டது.

- 1969 - பெல் லேப்ஸில் நிதி பல ஆண்டுகளாக வறண்டு போனது, 1969 இல் செல்வாக்கு மிக்க ஜான் பியர்ஸ் (John Pierce) பேச்சு புரிதல் ஆராய்ச்சியை விமர்சித்தது ஒரு திறந்த கடிதத்தை எழுதினார்; இதனால் பேச்சுப் புரிதல் ஆராய்ச்சிக்கு நிதி திரும்பப் தரப்பட்டது (John R. Pierce 1969). பியர்ஸ் ஓய்வுபெற்று ஜேம்ஸ் எல். ஃபிளனகன் (James L. Flanagan) பொறுப்பேற்கும் வரை இந்த நிதி திரும்பத்தரல் நீடித்தது.

1960களின் பிற்பகுதியில் ஸ்டான்போர்ட் பல்கலைக்கழகத்தில் பட்டதாரி மாணவராக தொடர்ச்சியான பேச்சு புரிதலைப் ஆய்ந்த முதல் நபர் ராஜ் ரெட்டி ஆவார். முந்தைய அமைப்புகள் பயனர்கள் ஒவ்வொரு வார்த்தையின் பின் இடைநிறுத்தப்பட வேண்டும். ரெட்டியின் அமைப்பு சதுரங்கம் விளையாடுவதற்கு பேசும் கட்டளைகளை வெளியிட்டது.

இந்த நேரத்தில் சோவியத் ஆராய்ச்சியாளர்கள் டைனமிக் டைம் வார்பிங் (dynamic time warping (DTW/ டி.டி.டபிள்யூ)) வழிமுறையை கண்டுபிடித்தனர் மற்றும் 200-சொற்களின் சொற்றொகையில் இயங்கக்கூடிய ஒரு பேச்சுப்புரிவானை உருவாக்க இதைப் பயன்படுத்தினர் (Benesty, Sondhi & Huang, 2008). டி.டி.டபிள்யூ பதப்படுத்தப்பட்ட உரையை குறுகிய சட்டங்களாகப் பிரிப்பதன் மூலமும் (எ.கா. 10 மிசெ கூறுகள்) ஒவ்வொரு சட்டத்தையும் ஒரு யூனிட்டாக செயலாக்குகிறது. டி.டி.டபிள்யூ பிற்கால வழிமுறைகளால் முறியடிக்கப்பட்டாலும், நுட்பம் தொடர்ந்தது. பேசுபவர் சுதந்திரத்தை அடைவது இந்தக் காலகட்டத்தில் தீர்க்கப்படாமல் இருந்தது.

1970-1990

1971 - தர்பா DARPA (The Defense Advanced Research Projects Agency) பேச்சு புரிந்துணர்வு ஆராய்ச்சிக்கு (Speech Understanding Research) ஐந்து ஆண்டுகள் நிதியளித்தது, குறைந்தபட்ச சொற்றொகை அளவு 1,000 சொற்களைக் கோரும் பேச்சு புரிதல் ஆராய்ச்சி. பேச்சு புரிதலில் முன்னேற்றம் அடைவதற்கு பேச்சு புரிதல் முக்கியமாக இருக்கும் என்று அவர்கள் நினைத்தார்கள்; இது பின்னர் பொய்யானது என்று நிரூபிக்கப்பட்டது (John Makhoul 2018). பிபிஎன், ஐபிஎம், கார்னகி மெலன் மற்றும் ஸ்டான்போர்ட் ஆராய்ச்சி நிறுவனம் அனைவரும் இந்த நிகழ்ச்சியில் பங்கேற்றனர் (Blechman & Blechman 2008; Klatt 1977). இது புத்துயிர் பெற்ற பேச்சு புரிதல் ஆராய்ச்சி ஜான் பியர்ஸின் கடிதம் இடுகை செய்தது.

1972 - மாசசூசெட்ஸின் நியூட்டனில் IEEE ஒலியியக்கம், பேச்சு மற்றும் சிக்னல் செயலாக்கக் குழு (Acoustics, Speech, and Signal Processing group) ஒரு மாநாட்டை நடத்தியது.

1976 முதல் ஐ.சி.ஏ.எஸ்.எஸ்.பி ICASSP பிலடெல்பியாவில் நடைபெற்றது, அதன் பின்னர் பேச்சு புரிதல் குறித்த ஆராய்ச்சியை வெளியிடுவதற்கான முக்கிய இடமாக இருந்தது (Rabiner (1984).

1960களின் பிற்பகுதியில் லியோனார்ட் பாம் (Leonard Baum) பாதுகாப்புப் பகுப்பாய்வு நிறுவனத்தில் (Institute for Defense Analysis) மார்கோவ் சங்கிலிகளின் கணிதத்தை (mathematics of Markov chains) உருவாக்கினார். ஒரு தசாப்தத்திற்குப் பிறகு, சி.எம்.யுவில், ராஜ் ரெட்டியின் மாணவர்களான ஜேம்ஸ் பேக்கர் (James Baker) மற்றும் ஜேனட் எம். பேக்கர் (Janet M. Baker) ஆகியோர் பேச்சு புரிதலுக்காக மறைக்கப்பட்ட மார்க்கோவ் மாதிரியை (Hidden Markov Model (HMM/ எச்.எம்.எம்)) பயன்படுத்தத் தொடங்கினர். ஜேம்ஸ் பேக்கர் தனது இளங்கலை கல்வியின் போது பாதுகாப்பு பகுப்பாய்வு நிறுவனத்தில் (Institute of Defense Analysis) கோடைகால வேலையிலிருந்து HMMகளைப் பற்றி அறிந்து கொண்டார். HMMகளின் பயன்பாடு, ஆராய்ச்சியாளர்களை ஒலியியக்கம், மொழி மற்றும் தொடரியல் போன்ற பல்வேறு அறிவு மூலங்களை ஒன்றிணைந்த நிகழ்தகவு மாதிரியில் இணைக்க அனுமதித்தது.

1980களின் நடுப்பகுதியில், ஐபிஎம்மின் ஃப்ரெட் ஜெலினெக்கின் குழு (Fred Jelinek's team) 20,000 சொற்களைக் கொண்ட சொற்றொகையக் கையாளக்கூடிய டாங்கோரா (Tangora) என்ற குரல் செயல்படுத்தப்பட்ட தட்டச்சுப்பொறியை (voice activated typewriter) உருவாக்கியது ("Pioneering Speech Recognition"). ஹெலினெக்கின் புள்ளிவிவர அணுகுமுறை (Jelinek's statistical approach) மனித மூளை செயலாக்கப்படுவதைப் பின்பற்றுவதில் குறைந்த முக்கியத்துவத்தை அளிக்கிறது மற்றும் எச்.எம்.எம் போன்ற புள்ளிவிவர மாடலிங் நுட்பங்களைப் பயன்படுத்துவதற்கு ஆதரவாக பேச்சைப் புரிந்துகொள்கிறது. இருப்பினும், மாடலிங் பேச்சுக்கு எச்.எம்.எம் மிகவும் பயனுள்ள வழியாக நிரூபிக்கப்பட்டது மற்றும் 1980களில் டைனமிக் டைம் வார்பிங்கை (dynamic time warping) இடம் பெயர்த்து அதிகாரமுள்ள பேச்சுப் புரிதல் வழிமுறையாக மாறியது.

1982 - ஐபிஎம்மின் சில போட்டியாளர்களில் ஒன்றான டிராகன் சிஸ்டம்ஸ், ஜேம்ஸ் மற்றும் ஜேனட் எம். பேக்கர் ஆகியோரால் நிறுவப்பட்டது ("History of Speech Recognition").

நடைமுறை பேச்சுப் புரிதல் Practical speech recognition

1980 களில் என்-கிராம் மொழி மாதிரி அறிமுகப்படுத்தப்பட்டது.

- 1987 - பேக்-ஆஃப் மாதிரி மொழி மாதிரிகள் பன்நீள n-கிராமை பயன்படுத்த அனுமதித்தது, மேலும் CSELT மொழிகளை அங்கீகரிக்க HMM ஐப் பயன்படுத்தியது (மென்பொருள் மற்றும் வன்பொருள் சிறப்பு செயலிகளில், எ.கா. RIPAC).

இந்த துறையில் அதிக முன்னேற்றம் என்பது கணினிகளின் விரைவாக அதிகரித்து வரும் திறன்களுக்கு கடமைப்பட்டிருக்கிறது. 1976ஆம் ஆண்டில் தர்பா திட்டத்தின் (DARPA program) முடிவில், ஆராய்ச்சியாளர்களுக்கு கிடைக்கக்கூடிய சிறந்த கணினி 4 எம்பி ராம் கொண்ட பிடிபி - 10 ஆகும். இது விநாடிகள் பேச்சைக் குறியத்திறவு செய்ய 100 நிமிடங்கள் எடுக்கலாம் (Kevin McKean 1980).

இரண்டு நடைமுறை தயாரிப்புகள்:

- 1987 - குர்ஸ்வீல் அப்ளைடு இன்டலிஜென்ஸில் (Kurzweil Applied Intelligence) இருந்து பேச்சுப் புரிவான்.
- 1990 - 1990 இல் வெளியிடப்பட்ட நுகர்வோர் தயாரிப்பு டிராகன் டிக்டேட் (Melanie Pinola 2011; "Ray Kurzweil biography" 2014) மனித ஆபரேட்டரைப் பயன்படுத்தாமல் தொலைபேசி அழைப்புகளை வழிநடத்த AT&T 1992இல் குரல் புரிதல் அழைப்பு செயலாக்க சேவையை (Voice Recognition Call Processing service) பயன்படுத்தியது (Juang & Rabiner 2017). இந்தத் தொழில்நுட்பத்தை லாரன்ஸ் ராபினெர் மற்றும் பலர் பெல் லேப்ஸில் உருவாக்கினர்.

இந்த கட்டத்தில், வழக்கமான வணிகப் பேச்சுப் புரிதல் அமைப்பின் சொற்றொகை சராசரி மனித சொற்றொகையை விடப் பெரியதாக இருந்தது (Xuedong Huang; James Baker; Raj Reddy 2015). ராஜ் ரெட்டியின் முன்னாள் மாணவர், சியூடோங் ஹுவாங், சி.எம்.யுவில் ஸ்பிங்க்ஸ்- II ஒழுங்குமுறையை உருவாக்கினார். ஸ்பீங்க்ஸ்- II ஒழுங்குமுறை முதன்முதலில் பேசுபவர்-சுதந்திரமான, பெரிய சொற்றொகை, தொடர்ச்சியான பேச்சு புரிதல் ஆகியவற்றைச் செய்தது மற்றும் இது தர்பாவின் (DARPA) 1992 மதிப்பீட்டில் சிறந்த செயல்திறனைக் கொண்டிருந்தது. ஒரு பெரிய சொற்றொகையுடன் தொடர்ச்சியான பேச்சைக் கையாள்வது பேச்சுப் புரிதல் வரலாற்றில் ஒரு முக்கிய மைல்கல்லாகும். 1993ஆம் ஆண்டில் மைக்ரோசாப்ட் நிறுவனத்தில் பேச்சு புரிதல் குழுவை ஹுவாங் கண்டுபிடித்தார். ராஜ் ரெட்டியின் மாணவர் கை-ஃபூ லீ (Kai-Fu Lee) ஆப்பிள் நிறுவனத்தில் சேர்ந்தார், அங்கு 1992 இல், காஸ்பர் (Casper) எனப்படும் ஆப்பிள் கணினிக்கான பேச்சு இடைமுக முன்மாதிரி ஒன்றை உருவாக்க உதவினார்.

பெல்ஜியத்தை தளமாகக் கொண்ட பேச்சுப் புரிதல் நிறுவனமான லெர்னவுட் & ஹவுஸ்பி (Lernout & Hauspie) 1997இல் குர்ஸ்வீல் அப்ளைடு இன்டலிஜென்ஸ் (Kurzweil Applied Intelligence) மற்றும் 2000ஆம் ஆண்டில் டிராகன் சிஸ்டம்ஸ் (Dragon Systems) உள்ளிட்ட பல நிறுவனங்களை வாங்கியது. விண்டோஸ் எக்ஸ்பி இயக்க முறைமையில் (Windows XP operating system) எல் அண்ட் எச் L&H பேச்சு தொழில்நுட்பம் பயன்படுத்தப்பட்டது. 2001ஆம் ஆண்டில் ஒரு கணக்கு ஊழல் முடிவுக்கு வரும் வரை எல் அண்ட் எச் (L&H) ஒரு தொழில்துறைத் முதன்மையாக இருந்தது. எல் அண்ட் எச்-இன் (L&H) பேச்சு தொழில்நுட்பம் ஸ்கேன்சாஃப்ட் நிறுவனத்தால் வாங்கப்பட்டது, இது 2005 இல் நுணுக்கமாக மாறியது. ஆப்பிள் முதலில் அதன் டிஜிட்டல் உதவியாளர் சிரிக்கு (Siri) பேச்சுப் புரிதல் திறனை வழங்குவதற்காக நுவான்ஸிலிருந்து (Nuance) மென்பொருளை உரிமம் பெற்றது.

2000 கள்

2000களில் தர்பா (DARPA) இரண்டு பேச்சு புரிதல் திட்டங்களை வழங்கியது: 2002 இல் பயனுள்ள மலிவு மறுபயன்பாட்டு பேச்சிலிருந்து-உரை (Effective Affordable Reusable Speech-to-Text (EARS)) மற்றும் உலகளாவிய தன்னாட்சி மொழி தேட்டம் (Global Autonomous Language Exploitation GALE). EARS திட்டத்தில் நான்கு அணிகள் பங்கேற்றன: ஐபிஎம் (IBM), லிம்ஸி (LIMSI) மற்றும் பிட்ஸ்பர்க் பல்கலைக்கழகம், கேம்பிரிட்ஜ் பல்கலைக்கழகம் இவற்றுடன் பிபிஎன் BBN தலைமையிலான குழு மற்றும் ஐ.சி.எஸ்.ஐ., எஸ்.ஆர்.ஐ. மற்றும் வாஷிங்டன் பல்கலைக்கழகம் ஆகியவற்றைக் கொண்ட குழு. 500 க்கும் மேற்பட்ட பேசுபவர்களிடமிருந்து 260 மணிநேர பதிவு செய்யப்பட்ட உரையாடல்களைக் கொண்ட ஸ்விட்ச்போர்டு தொலைபேசி பேச்சு தரவுத்தொகுதி சேகரிப்புக்கு EARS நிதியளித்தது. GALE திட்டம் அரபு மற்றும் மாண்டரின் ஒளிபரப்பு செய்தி உரையை மையமாகக் கொண்டது. பேச்சு புரிதலில் கூகிளின் முதல் முயற்சி 2007இல் நுவான்ஸில் இருந்து சில ஆராய்ச்சியாளர்களைப் பணியமர்த்திய பின்னர் வந்தது (Jason Kincaid 2015). முதல் தயாரிப்பு GOOG-411, தொலைபேசி அடிப்படையிலான அடைவு சேவையாகும் (directory service). GOOG-411இன் பதிவுகள், கூகிள் அதன் பேச்சுப் புரிதல் ஒழுங்குமுறைகளை மேம்படுத்த உதவிய மதிப்புமிக்க தரவை உருவாக்கியது. கூகிள் குரல் தேடல் (Google Voice Search) இப்போது 30க்கும் மேற்பட்ட மொழிகளில் ஆதரிக்கப்படுகிறது.

யுனைடெட் ஸ்டேட்ஸில், தேசிய பாதுகாப்பு நிறுவனம் (National Security Agency) குறைந்தபட்சம் 2006-இலிருந்து முக்கிய சொற்களைக் கண்டுபிடிப்பதற்கான ஒரு வகை பேச்சுப்

புரிதலைப் பயன்படுத்தியுள்ளது (Froomkin 2015). இந்த தொழில்நுட்பம் ஆய்வாளர்களை பதிவுசெய்த உரையாடல்களின் பெரிய தொகுதிகளில் தேடவும், முக்கிய வார்த்தைகளின் குறிப்புகளைத் தனிமைப்படுத்தவும் அனுமதிக்கிறது. பதிவுகளை அட்டவணைப்படுத்தலாம் மற்றும் ஆர்வமுள்ள உரையாடல்களைக் கண்டறிய ஆய்வாளர்கள் தரவுத்தளத்தில் கேள்விகளை இயக்கலாம். சில அரசாங்க ஆராய்ச்சி திட்டங்கள் பேச்சுப் புரிதலின் நுண்ணறிவு பயன்பாடுகளில் கவனம் செலுத்தின, எ.கா. தர்பாவின் (DARPA) EARSஇன் திட்டம் மற்றும் IARPAஇன் பாபல் திட்டம் (Babel program).

2000களின் முற்பகுதியில், முன் எதிர்பார்ப்பு செயற்கை நரம்பியல் நெட்வொர்க்குகளுடன் (feedforward artificial neural networks) Herve (Bourlard and Nelson Morgan 1994) இணைந்து மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகள் போன்ற பாரம்பரிய அணுகுமுறைகளால் பேச்சு புரிதல் இன்னும் ஆதிக்கம் செலுத்தப்பட்டது. இருப்பினும், இன்று, பேச்சுப் புரிதலின் பல அம்சங்கள் 1997இல் செப் ஹோக்ரீட்டர் & ஜூர்கன் ஷ்மிதூபரால் (Sepp Hochreiter & Jürgen Schmidhuber) வெளியிடப்பட்ட (Sepp Hochreiter & Schmidhuber 1997) மீள்நிகழும் நரம்பியல் வலையமைப்பான (a recurrent neural network (RNN)) நீண்ட குறுகிய கால நினைவகம் (Long short-term memory (LSTM/எல்.எஸ்.டி.எம்.) என்ற ஆழமான கற்றல் முறையால் (deep learning method) கையகப்படுத்தப்பட்டுள்ளன. எல்.எஸ்.டி.எம் ஆர்.என்.என்-கள் (LSTM RNNs) மறைந்துபோகும் சாய்வுச் சிக்கலைத் தவிர்க்கின்றன, மேலும் பேச்சுக்கு முக்கியமான ஆயிரக்கணக்கான தனித்துவமான நேர படிகளுக்கு முன்பு நடந்த நிகழ்வுகளின் நினைவுகள் தேவைப்படுகிற "மிக ஆழமான கற்றல்" பணிகளைக் ("Very Deep Learning" tasks) (Schmidhuber 2015) கற்றுக் கொள்ள இயலும். 2007ஆம் ஆண்டில், இணைப்பாளர் தற்காலிக வகைப்பாடால் (Connectionist Temporal Classification (CTC/சி.டி.சி.) (Alex Graves, Santiago Fernandez, Faustino Gomez, and Jürgen Schmidhuber 2006) பயிற்சியளிக்கப்பட்ட எல்.எஸ்.டி.எம் சில பயன்பாடுகளில் பாரம்பரிய பேச்சுப் புரிதலை விஞ்சத் தொடங்கியது (Santiago Fernandez, Alex Graves, and Jürgen Schmidhuber (2007). 2015ஆம் ஆண்டில், கூகிளின் பேச்சுப் புரிதல் சி.டி.சி-பயிற்சி பெற்ற எல்.எஸ்.டி.எம் மூலம் 49% வியத்தகு செயல்திறன் அதிகரித்ததாக கூறப்படுகிறது, இது இப்போது கூகிள் குரல் மூலம் அனைத்து ஸ்மார்ட்போன் பயனர்களுக்கும் கிடைக்கிறது (Haşim Sak, Andrew Senior, Kanishka Rao, Françoise Beaufays and Johan Schalkwyk 2015).

பேச்சுப் புரிதலின் நீண்ட வரலாற்றில், 1980கள், 1990கள் மற்றும் சில ஆண்டுகளில் 2000களில் செயற்கை நரம்பியல் நெட்வொர்க்குகளின் ஆழமற்ற வடிவம் மற்றும் ஆழமான வடிவம் (எ.கா. . வலைகள்) பல ஆண்டுகளாக ஆராயப்பட்டன (Morgan, Renals, Franco 1993); (Robinson 1992); (Waibel, Hinton, Lang 1989). ஆனால் இந்த முறைகள் ஒருபோதும் பாகுபாடின்றி பேச்சு பயிற்சியளிக்கப்பட்ட ஆக்கமுறை மாதிரிகளை அடிப்படையாகக் கொண்ட ஒரே மாதிரி அல்லாத உள்-கைவினையாக்க காஸியன் கலவை மாதிரி (Gaussian mixture model)/ மறைக்கப்பட்ட மார்க்கோவ் மாதிரி /Hidden Markov model (GMM-HMM/ஜிஎம்எம்-எச்எம்எம்)) தொழில்நுட்பத்தை வென்றெடுக்கவில்லை (Baker , Li Deng, Glass, Khudanpur, Chin-Hui Lee, Morgan, O'Shaughnessy 2009). 1990களில் பல முக்கிய சிரமங்கள் முறைப்படி பகுப்பாய்வு செய்யப்பட்டன, இதில் சாய்வு குறைதல் (gradient diminishing) (Sepp Hochreiter 1991) மற்றும் நரம்பியல் முன்கணிப்பு மாதிரிகளில் (neural predictive models) பலவீனமான தற்காலிக தொடர்பு அமைப்பு (weak temporal correlation structure) ஆகியவை அடங்கும் (Bengio1991) (Deng, Hassanein, Elmasry 1994). இந்த சிரமங்கள் அனைத்தும் இந்த ஆரம்ப நாட்களில் பெரிய பயிற்சித் தரவு மற்றும் பெரிய கணினி சக்தி இல்லாததால் கூடுதலாக இருந்தன. இத்தகைய தடைகளைப் புரிந்து கொண்ட பெரும்பாலான பேச்சு புரிதல் ஆராய்ச்சியாளர்கள், பின்னர் நரம்பியல் வலைகளிலிருந்து விலகி, மாடலிங் அணுகுமுறைகளைத் தொடர 2009-2010ஆம் ஆண்டு தொடங்கி ஆழ்ந்த கற்றலின் சமீபத்திய எழுச்சி வரை இந்த சிரமங்களை எல்லாம் சமாளித்தனர். ஹிண்டன் மற்றும் பலர். மற்றும் டெங் மற்றும் பலர் (Hinton et al. and Deng et al.) ஒருவருக்கொருவர் மற்றும் பின்னர் நான்கு குழுக்களில் (டொராண்டோ பல்கலைக்கழகம், மைக்ரோசாப்ட், கூகிள் மற்றும் ஐபிஎம்) சக ஊழியர்களுடனான ஒத்துழைப்பு எவ்வாறு ஆழ்ந்த ஃபீட்பார்வேர்ட் நரம்பியல் நெட்வொர்க்குகளின் (deep feedforward neural networks) பயன்பாடுகளை பேச்சுப் புரிதலுக்குத் தூண்டியது என்பது பற்றி இந்த சமீபத்திய வரலாற்றின் ஒரு பகுதியை மதிப்பாய்வு செய்தனர். (Hinton, Geoffrey; Deng, Li; Yu, Dong; Dahl, George; Mohamed, Abdel-Rahman; Jaitly, Navdeep; Senior, Andrew; Vanhoucke, Vincent; Nguyen, Patrick; Sainath, Tara; Kingsbury, Brian 2012); Morgan, Bourlard, Renals, Cohen, Franco 1993; Keynote talk by Geoff Hinton 2013; Keynote talk by Li Deng 2014).

2010கள்

2010ஆம் ஆண்டின் முற்பகுதியில், குரல் புரிதல் என்றும் அழைக்கப்பட்ட பேச்சு புரிதல், பேசுபவரை புரிதலிருந்து தெளிவாக வேறுபடுத்தப்பட்டது, மேலும் பேசுபவர் சுதந்திரம் ஒரு பெரிய திருப்புமுனையாகக் கருதப்பட்டது. அதுவரை, ஒழுங்குமுறைகளுக்கு "பயிற்சி" காலம் தேவைப்பட்டது. 1987ஆம் ஆண்டு ஒரு பொம்மைக்கான விளம்பரம் "இறுதியாக, உங்களைப் புரிந்துகொள்ளும் பொம்மை" என்ற கோஷத்தைக் கொண்டிருந்தது. - "இது அவர்களின் குரலுக்கு பதிலளிக்க எந்த குழந்தைகள் பயிற்சியளிக்க முடியும்" என்று விவரிக்கப்பட்ட போதிலும் (Melanie Pinola 2011).

2017ஆம் ஆண்டில், மைக்ரோசாப்ட் ஆராய்ச்சியாளர்கள் பரவலாக பெஞ்ச்மார்க் செய்யப்பட்ட சுவிட்ச்போர்டு பணியில் உரையாடல் தொலைபேசி உரையை படியெடுக்கும் வரலாற்று மனித சமநிலை மைல்கல்லை எட்டினர். பேச்சு புரிதல் துல்லியத்தை மேம்படுத்த பல ஆழமான கற்றல் மாதிரிகள் பயன்படுத்தப்பட்டன. பேச்சு புரிதல் சொல் பிழை வீதம் 4 தொழில்முறை மனித டிரான்ஸ்கிரிபர்கள் ஒரே அளவுகோலில் ஒன்றாக வேலை செய்வதாகக் கூறப்பட்டது, இது ஐபிஎம் வாட்சன் பேச்சுக் குழுவால் அதே பணியில் நிதியளிக்கப்பட்டது.

மாதிரிகள், முறைகள் மற்றும் வழிமுறைகள்

ஒலியியக்க மாதிரியாக்கம் (acoustic modeling) மற்றும் மொழி மாதிரியாக்கம் (language modeling) இரண்டும் நவீன புள்ளியியல் அடிப்படையிலான பேச்சுப் புரிதல் வழிமுறைகளின் முக்கியமான பகுதிகள் ஆகும். மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகள் (HMMs) பல அமைப்புகளில் பரவலாகப் பயன்படுத்தப்படுகின்றன. ஆவண வகைப்பாடு அல்லது புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பு போன்ற பல இயற்கை மொழி செயலாக்க பயன்பாடுகளிலும் மொழி மாடலிங் பயன்படுத்தப்படுகிறது.

மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகள் (Hidden Markov models)

நவீன பொது நோக்கத்திற்கான பேச்சு அறிதல் ஒழுங்குமுறைகள் மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகளை அடிப்படையாகக் கொண்டவை. இவை குறியீடுகள் அல்லது அளவுகளின் வரிசையை வெளியிடுகின்றன புள்ளியியல்சார் மாதிரிகள் ஆகும். பேச்சு அறிதலில் HMMகள் பயன்படுத்தப்படுகின்றன; ஏனெனில் பேச்சு சமிக்ஞையை ஒரு கூறுவாரியான நிலையான சமிக்ஞையாக அல்லது குறுகிய கால நிலையான சமிக்ஞையாக பார்க்க முடியும். குறுகிய நேர அளவிலான (எ.கா., 10 மில்லி விநாடிகள்), பேச்சு ஒரு நிலையான செயல்முறையாக

தோராயமாக மதிப்பிடப்படலாம். பேச்சு பல சீரற்ற நோக்கங்களுக்காக ஒரு மார்கோவ் மாதிரியாகக் கருதப்படலாம்.

எச்.எம்.எம்-கள் பிரபலமாக இருப்பதற்கான மற்றொரு காரணம், அவை தானாகவே பயிற்சியளிக்கப்படலாம், மேலும் அவை எளிமையானவை மற்றும் கணக்கீட்டு/கணினி ரீதியாக பயன்படுத்தக்கூடியவை. பேச்சு அறிதலில், மறைக்கப்பட்ட மார்க்கோவ் மாதிரியானது n-பரிமாண உண்மையான மதிப்புள்ள திசையன்களின் (n-dimensional real-valued vectors) வரிசையை வெளியிடும் (n என்பது 10 போன்ற சிறிய முழு எண்ணாக இருப்பதால்), இவற்றில் ஒன்றை ஒவ்வொரு 10 மில்லி விநாடிகளிலும்வெளியிடுகிறது. திசையன்கள் செப்ஸ்ட்ரல் குணகங்களைக் கொண்டிருக்கும், அவை குறுகிய கால பேச்சு சாளரத்தின் ஃபோரியர் உருமாற்றத்தை (Fourier transform) எடுத்தும் கோசைன் உருமாற்றத்தைப் (cosine transform) பயன்படுத்தி ஸ்பெக்ட்ரத்தை தொடர்பறுத்தும் பின்னர் முதல் (மிக முக்கியமான) குணகங்களை எடுத்தும் பெறப்படுகின்றன. மறைக்கப்பட்ட மார்க்கோவ் மாதிரியானது ஒவ்வொரு நிலையிலும் ஒரு புள்ளிவிவர விநியோகத்தைக் கொண்டிருக்கும், இது மூலைவிட்ட கோவாரன்ஸ் காஸியன்களின் (diagonal covariance Gaussians) கலவையாகும், இது ஒவ்வொரு கவனிக்கப்பட்ட திசையனுக்கும் ஒரு வாய்ப்பை வழங்கும். ஒவ்வொரு சொல்லுக்கும் அல்லது (கூடுதல் பொதுவான பேச்சு அறிதல் அமைப்புகளுக்கு) ஒவ்வொரு ஒலியனுக்கும் வெவ்வேறு வெளியீட்டு விநியோகம் இருக்கும்; சொற்கள் அல்லது ஒலியன்களின் வரிசைக்கு ஒரு மறைக்கப்பட்ட மார்க்கோவ் மாதிரி தனித்தனி சொற்கள் மற்றும் ஒலியன்களுக்கு தனிப்பட்ட பயிற்சி பெற்ற மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகளை இணைப்பதன் மூலம் செய்யப்படுகிறது.

பேச்சு அறிதலுக்கான மிகவும் பொதுவான, HMM- அடிப்படையிலான அணுகுமுறையின் முக்கிய கூறுகள் மேலே விவரிக்கப்பட்டுள்ளன. நவீன பேச்சு அறிதல் ஒழுங்குமுறைகள் மேலே விவரிக்கப்பட்ட அடிப்படை அணுகுமுறையின் முடிவுகளை மேம்படுத்த பல நிலைபேறுபெற்ற நுட்பங்களின் பல்வேறு சேர்க்கைகளைப் பயன்படுத்துகின்றன. ஒரு பொதுவான பெரிய-சொற்றொகை ஒழுங்குமுறைக்கு ஒலியன்களுக்கான சூழல் சார்பு தேவைப்படும் (எனவே வெவ்வேறு இடது மற்றும் வலது சூழலுடன் கூடிய ஒலியன்கள் எச்.எம்.எம் நிலைகளைப் போல வெவ்வேறு உருப்படுத்தங்களைக் கொண்டுள்ளன); வெவ்வேறு பேசுபவர் மற்றும் பதிவு நிலைமைகளை இயல்பாக்குவதற்கு இது செப்ஸ்ட்ரல் இயல்பாக்கலைப் (cepstral normalization) பயன்படுத்தும்; மேலும் பேசுபவர் இயல்பாக்குதலுக்காக, ஆண்-பெண் இயல்பாக்குதலுக்காக

குரல் பாதை நீள இயல்பாக்கம் (vocal tract length normalization (VTLN/வி.டி.எல்.என்) மற்றும் அதிகப் பொது பேசுபவர் தழுவலுக்கு அதிகபட்ச வாய்ப்பு நேரியல் பின்னடைவு (maximum likelihood linear regression (MLLR/எம்.எல்.எல்.ஆர்) ஆகியவற்றைப் பயன்படுத்தலாம். பண்புக்கூறுகள் பேச்சு இயக்கவியலைப் (speech dynamics) கைக்கொள்ள டெல்டா மற்றும் டெல்டா-டெல்டா குணகங்கள் என்று அழைக்கப்படுகின்றவற்றை கொண்டிருக்கும்; கூடுதலாக ஹீட்டோரோசெஸ்டாஸ்டிக் நேரியல் பாகுபாடு பகுப்பாய்வைப் (heteroscedastic linear discriminant analysis (HLDA/எச்.எல்.டி.ஏ) பயன்படுத்தலாம்; அல்லது டெல்டா மற்றும் டெல்டா-டெல்டா குணகங்களைத் தவிர்த்து, பிணைத்தல் மற்றும் ஹீட்டோரோசெஸ்டாஸ்டிக் நேரியல் பாகுபாட்டுப் பகுப்பாய்வு (heteroscedastic linear discriminant analysis) அல்லது உலகளாவிய அரை-பிணைந்த இணை மாறுபாடு மாற்றம் (global semi-tied co variance transform) [அதிகபட்ச சாத்தியக்கூறு நேரியல் மாற்றம் (maximum likelihood linear transform) அல்லது MLLT/எம்.எல்.எல்.டி என்றும் அழைக்கப்படும்] என்பதால் தொடரப்படும் எல்.டி.ஏ-அடிப்படையிலான முந்திட்டத்தைப் (LDA-based projection) பயன்படுத்தலாம். பல ஒழுங்குமுறைகள், எச்.எம்.எம் அளகை மதிப்பீட்டிற்கான முற்றிலும் புள்ளிவிவர அணுகுமுறையை விட்டுவிடும் பாகுபாடுகாட்டுகிற பயிற்சி நுட்பங்களைப் (discriminative training techniques) பயன்படுத்தும்; மற்றும் அதற்குப் பதிலாகப் பயிற்சித் தரவின் சில வகைப்பாடு தொடர்பான அளவை மேம்படுத்தும். எடுத்துக்காட்டுகள் அதிகபட்ச பரஸ்பர தகவல் (maximum mutual information (MMI/எம்எம்ஐ), குறைந்தபட்ச வகைப்பாட்டுப் பிழை (minimum classification error (MCE/எம்சிஇ) மற்றும் குறைந்தபட்ச ஒலிப் பிழை (minimum phone error (MPE/எம்.பி.இ) என்பனவாகும்.

பேச்சின் குறியத்திறவு/டிகோடிங் (கணினி ஒரு புதிய கூற்றுடன் வழங்கப்படும்போது என்ன நடக்கும் என்பதற்கான சொல் மற்றும் பெரும்பாலும் மூல வாக்கியத்தை கணக்கிட வேண்டும்) சிறந்த வழியைக் கண்டுபிடிக்க விட்டர்பி வழிமுறையைப் (Viterbi algorithm) பயன்படுத்தக்கூடும்; மற்றும் இங்கே ஒலி மற்றும் மொழி மாதிரி தகவல்கள் இரண்டையும் உள்ளடக்கிய இயக்கவாற்றலில் சேர்க்கை மறைக்கப்பட்ட மார்க்கோவ் மாதிரி மற்றும் அதை முன்பே நிலையான முறையில் இணைத்தல் (முற்றுநிலை மாற்றி/finite state transducer அல்லது எஃப்எஸ்டி/ FST, அணுகுமுறை) இவற்றிற்கிடையில் தேர்வு உள்ளது.

குறியத்திறவுக்கான (decoding) சாத்தியமான முன்னேற்றம் என்னவென்றால், சிறந்த தேர்வுக்குரியதை வைப்பதற்கு பதிலாக ஒரு நல்ல தேர்வுக்குரியவைகளின் தொகுப்பை வைத்திருப்பது, மேலும் இந்த நல்ல தேர்வுக்குரியவைகளை மதிப்பிடுவதற்கு சிறந்த மதிப்பெண் செயல்பாட்டை (மறு மதிப்பெண்) பயன்படுத்துவதன் மூலம் இந்த சுத்திகரிக்கப்பட்ட மதிப்பெண்ணின் படி சிறந்த ஒன்றை நாம் தேர்வு செய்யலாம். தேர்வுக்குரியவைகளின் தொகுப்பை ஒரு பட்டியலாக (N-சிறந்த பட்டியல் அணுகுமுறை) அல்லது மாதிரிகளின் துணைக்குழுவாக (ஒரு லட்டு) வைக்கலாம். பேய்ஸ் அபாயத்தைக் (Bayes risk) குறைக்க (அல்லது அதன் தோராயமாக்கல்) முயற்சிப்பதன் மூலம் மறு மதிப்பெண் வழக்கமாக செய்யப்படுகிறது (Goel & Byrne 2000): மூல வாக்கியத்தை அதிகபட்ச நிகழ்தகவுடன் எடுத்துக்கொள்வதற்குப் பதிலாக, நாம் சாத்தியமான அனைத்து வரிவடிவாக்கத்துடனும் (transcriptions) கொடுக்கப்பட்ட இழப்புச் செயல்பாட்டின் எதிர்பார்ப்பைக் குறைக்கும் வாக்கியத்தை எடுக்க முயற்சிக்கிறோம் (அதாவது, அவற்றின் மதிப்பிடப்பட்ட நிகழ்தகவு மூலம் எடையிட்ட/அளவிட்ட பிற சாத்தியமான வாக்கியங்களுக்கான சராசரி தூரத்தை குறைக்கும் வாக்கியத்தை நாம் எடுத்துக்கொள்கிறோம்). இழப்புச் செயல்பாடு பொதுவாக லெவன்ஸ்டீன் தூரமாகும் (Levenshtein distance) இருப்பினும் இது குறிப்பிட்ட பணிகளுக்கு வெவ்வேறு தூரங்களாக இருக்கலாம்; சாத்தியமான எழுத்தாக்கங்களின் (டிரான்ஸ்கிரிப்ஷன்களின்) தொகுப்பு, நிச்சயமாக, கண்டுபிடிப்பதற்காகப் (டிராக்டபிலிட்டிக்காக) பராமரிக்கக் கத்தரிக்கப்படுகிறது. சில அனுமானங்களைச் சரிபார்க்கும் ஒரு முற்றுநிலை மாற்றியாகத் தங்களைத்தாமே உருப்படுத்தம் செய்யப்படும் சீராக்க (எடிட்) தூரங்களைக் கொண்ட எடையுள்ள முற்றுநிலை மாற்றிகளாக உருப்படுத்தம் செய்யப்படும் அணிக்கோவைளை மறு மதிப்பெண்பெற திறமையான வழிமுறைகள் வகுக்கப்பட்டுள்ளன (Mohri 2002).

இயக்கம்சார் காலப் பொதிதல் (Dynamic time warping (DTW/டி.டி.டபிள்யூ) அடிப்படையிலான பேச்சு அறிதல்

இயக்கம்சார் காலப் பொதிதல் என்பது ஒரு அணுகுமுறையாகும், இது வரலாற்று ரீதியாக பேச்சு அறிதலுக்காகப் பயன்படுத்தப்பட்டது, ஆனால் இப்போது மிகவும் வெற்றிகரமான HMM-அடிப்படையிலான அணுகுமுறையால் இடம்பெயர்க்கப்பட்டுள்ளது.

இயக்கம்சார் காலப் பொதிதல் என்பது நேரம் அல்லது வேகத்தில் மாறுபடக்கூடிய இரண்டு வரிசைகளுக்கு/கோவைகளுக்கு இடையிலான ஒற்றுமையை அளவிடுவதற்கான ஒரு

வழிமுறையாகும். எடுத்துக்காட்டாக, ஒரு வீடியோவில் நபர் மெதுவாக நடந்து கொண்டிருந்தாலும், மற்றொரு வீடியோவில் அவர் விரைவாக நடந்து கொண்டிருந்தாலும், அல்லது ஒரு உற்றுநோக்கலின் போது முடுக்கம் மற்றும் வீழ்ச்சி ஏற்பட்டாலும் கூட, நடை முறைகளில் ஒற்றுமைகள் கண்டறியப்படும். வீடியோ, ஆடியோ மற்றும் கிராபிக்ஸ் ஆகியவற்றில் டி.டி.டபிள்யூ பயன்படுத்தப்பட்டுள்ளது - உண்மையில், ஒரு நேரியல் உருப்படுத்தமாக மாற்றக்கூடிய எந்த தரவையும் டி.டி.டபிள்யூ மூலம் பகுப்பாய்வு செய்யலாம்.

வெவ்வேறு பேசும் வேகத்தை சமாளிக்கும் தானியங்கி பேச்சு அறிதல் நன்கு அறியப்பட்ட பயன்பாடு ஆகும். பொதுவாக, சில கட்டுப்பாடுகளுடன் கொடுக்கப்பட்ட இரண்டு கோவைகளுக்கு (எ.கா. நேரத் தொடர்கள்) இடையே உகந்த பொருத்தத்தைக் கண்டறிய கணினியை அனுமதிக்கும் ஒரு முறை ஆகும். அதாவது, கோவைகள் நேரியல் அல்லாமல் ஒன்றுக்கொன்று பொருந்தக்கூடிய வகையில் "திசைதிருப்பப்பட்டவை" ("warped"). மறைக்கப்பட்ட மார்கோவ் மாதிரிகளின் சூழலில் இந்த வரிசை சீரமைப்பு முறை பெரும்பாலும் பயன்படுத்தப்படுகிறது.

நரம்பியல் வலையமைப்புகள்

1980களின் பிற்பகுதியில் தானியக்க பேச்சு அறிதலில் (Automatic Speech reconition (ASR) நரம்பியல் நெட்வொர்க்குகள் ஒரு கவர்ச்சியான ஒலியியக்க மாடலிங் அணுகுமுறையாக (acoustic modeling approach) வெளிப்பட்டன. அப்போதிருந்து, ஒலியன் வகைப்பாடு (Waibel, Hanazawa, Hinton, Shikano, Lang (1989), பன்நோக்கம் கொண்ட பரிணாம வளர்ச்சி வழிமுறைகள் (multi-objective evolutionary algorithms) (Bird, Wanner, Ekárt, Faria 2020) மூலம் ஒலியன் வகைப்பாடு, தனிமைப்படுத்தப்பட்ட சொல் அறிதல் (isolated word recognition) (Wu & Chan 1993), ஆடியோவிஷுவல் பேச்சு அறிதல், ஒலிகட்புல speech recognition, ஒலிகட்புல பேசுபவர் அறிதல் audiovisual speaker recognition (AVSR) மற்றும் பேசுபவர் தழுவல் (speaker adaptation) போன்ற பேச்சு அறிதலின் பல அம்சங்களில் நரம்பியல் நெட்வொர்க்குகள் பயன்படுத்தப்படுகின்றன. தழுவல்.

நரம்பியல் வலைப்பின்னகள் எச்.எம்.எம்-களைக் காட்டிலும் பண்புக்கூறு புள்ளியியல்சார் பண்புகளைப் பற்றி குறைவான வெளிப்படையான அனுமானங்களைச் செய்கின்றன; மற்றும் அவைஅவற்றைப் பேச்சு அறிதலுக்கான கவர்ச்சிகரமான அறிதல் மாதிரிகளாக ஆக்கும் பல குணங்களைக் கொண்டுள்ளன. பேச்சு பண்புக்கூறு பிரிவின் நிகழ்தகவுகளை மதிப்பிடுவதற்குப்

பயன்படுத்தும்போது, நரம்பியல் நெட்வொர்க்குகள் இயற்கையான மற்றும் திறமையான முறையில் பாகுபாடுகாட்டுகிற பயிற்சியை அனுமதிக்கின்றன. இருப்பினும், தனிப்பட்ட ஒலியன்கள் மற்றும் தனிமைப்படுத்தப்பட்ட சொற்கள் போன்ற குறுகிய கால அலகுகளை வகைப்படுத்துவதில் (Zahorian, Zimmer and Meng 2002) அவற்றின் செயல்திறன் இருந்தபோதிலும், தொடக்ககால நரம்பியல் நெட்வொர்க்குகள் காலம்சார் சார்புகளை மாதிரிப் படுத்துவதற்கான அவற்றின் வரபிற்குட்பட்ட திறன் காரணமாகத் தொடர்ச்சியான அறிதல் பணிகளுக்கு அவை அரிதாகவே வெற்றி பெற்றன.

இந்த வரம்புக்கு ஒரு அணுகுமுறை, நரம்பியல் நெட்வொர்க்குகளை ஒரு முன் செயலாக்கம், பண்புக்கூறு மாற்றம் அல்லது பரிமாணக் குறைப்பு (Hu, Zahorian & Stephen 2010), HMM அடிப்படையிலான அறிதலுக்கு முன் பயன்படுத்துவது ஆகும். இருப்பினும், மிக சமீபத்தில், நீண்ட-குறுகிய கால நினைவகம் (Long short-term memory (LSTM/எல்.எஸ்.டி.எம்)) மற்றும் தொடர்புடைய மீள்நிழும் நரம்பியல் நெட்வொர்க்குகள் (recurrent neural networks RNN ஆர்.என்.என்) (Hochreiter & Schmidhuber 1997); (Schmidhuber 2015); (Fernandez, Graves, & Schmidhuber 2007) (Graves, Mohamed, & Hinton 2013) மற்றும் நேர தாமத நரம்பியல் நெட்வொர்க்குகள் (Time Delay Neural Networks TDNN/ டி.டி.என்.என்) (Waibel, Alex 1989) ஆகியவை இந்த பகுதியில் மேம்பட்ட செயல்திறனை நிரூபித்துள்ளன.

ஆழமான எதிர்பார்ப்பு மற்றும் மீள்நிகழ்வு நரம்பியல் வலையமைப்புகள்

ஆழமான நரம்பியல் வலையமைப்புகள் மற்றும் டெனோசிங் ஆட்டோஎன்கோடர்கள் (Maas, Le, O'Neil, Vinyals, Nguyen & Ng, 2012) ஆகியவை விசாரணையில் உள்ளன. ஒரு ஆழமான எதிர்பார்பு நரம்பியல் வலையமைப்பு (deep feedforward neural network (DNN/டி.என்.என்)) என்பது ஒரு செயற்கை நரம்பியல் வலையமைப்பாகும் (artificial neural network); இது உள்ளீடு மற்றும் வெளியீட்டு அடுக்குகளுக்கு இடையில் பல மறைக்கப்பட்ட அடுக்குகளைக் கொண்ட அலகுகளைக் கொண்டுள்ளது (Hinton, Deng, Yu, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath, Kingsbury 2012). மேலோட்டமான நரம்பியல் வலையமைப்புகளைப் போலவே, டி.என்.என் களும் சிக்கலான நேரியல் அல்லாத உறவுகளை மாதிரியாகக் கொள்ளலாம். டி.என்.என் கட்டமைப்புகள் தொகுப்பாக்க மாதிரிகளை உருவாக்குகின்றன, அங்கு கூடுதல் அடுக்குகள் கீழ் அடுக்குகளிலிருந்து பண்புக்கூறுகளின்

தொகுப்பை செயல்படுத்துகின்றன, இது ஒரு பெரிய கற்றல் திறனைக் கொடுக்கும், இதனால் பேச்சுத் தரவின் சிக்கலான வடிவங்களை மாடலிங் செய்யும் திறன் உள்ளது (Deng, & Yu 2014).

பெரிய சொற்றொகை பேச்சு அறிதலில் டி.என்.என்-களின் வெற்றி 2010இல் கல்வி ஆராய்ச்சியாளர்களுடன் இணைந்து தொழில்துறை ஆராய்ச்சியாளர்களால் நிகழ்ந்தது; இதில் தீரமானக் கிளைகளால் (decision trees) கட்டப்பட்ட சூழல் சார்ந்த எச்.எம்.எம். நிலைகளின் அடிப்படையில் அமைந்த டி.என்.என்-இன் பெரிய வெளியீட்டு அடுக்குகள் ஏற்றுகொள்ளப்பட்டுள்ளன (Yu, Deng, Dahl 2010) (Dahl, Yu, Deng & Acero 2012) (Deng, Li, Huang, Yao, Yu, Seide, et al. 2013) மைக்ரோசாப்ட் ரிசர்ச்சின் சமீபத்திய ஸ்பிரிங்கர் புத்தகத்தில் அக்டோபர் 2014 நிலவரப்படி இந்த வளர்ச்சி மற்றும் தற்போதைய நிலை பற்றிய விரிவான மதிப்புரைகளைக் காணலாம் (Yu, Deng, 2014). சமீபத்திய மேலோட்டக் கட்டுரைகளில் தானியங்கி பேச்சு அறிதலின் பின்னணி மற்றும் குறிப்பாக ஆழமான கற்றல் உட்பட பல்வேறு இயந்திர கற்றல் முன்மாதிரிகளின் தாக்கத்தையும் காண்லாம் (Deng & Li 2013) (Schmidhuber 2015).

ஆழ்ந்த கற்றலின் ஒரு அடிப்படைக் கொள்கை, கையால் வடிவமைக்கப்பட்ட பண்புக்கூறு பொறியியலை நீக்குவதும், மூல பண்புக்கூறுகளைப் பயன்படுத்துவதும் ஆகும். இந்த கொள்கை முதன்முதலில் "மூல" ஸ்பெக்ட்ரோகிராம் அல்லது நேரியல் வடிகட்டி-வங்கி பண்புக்கூறுகளில் (linear filter-bank features) ஆழ்ந்த ஆட்டோஎன்கோடரின் கட்டமைப்பில் (architecture of deep autoencoder) வெற்றிகரமாக ஆராயப்பட்டது (Deng, Seltzer, Yu, Acero, Mohamed, and Hinton 2010); இது ஸ்பெக்ட்ரோகிராம்களிலிருந்து நிலையான மாற்றத்தின் சில கட்டங்களைக் கொண்டுள்ள மெல்-செப்ஸ்ட்ரல் பண்புக்கூறுகளை (Mel-Cepstral features) விட அதன் மேன்மையைக் காட்டுகிறது. அலைவடிவங்களான பேச்சின் உண்மையான "மூல" பண்புக்கூறுகள், மிகச் சிறந்த பெரிய அளவிலான பேச்சு அறிதல் முடிவுகளைத் தருவதாக சமீபத்தில் காட்டப்பட்டுள்ளன (Tüske, Golik, Schlüter, Ney, 2014).

இறுதியிலிந்து இறுதி தானியக்க பேச்சு அறிதல் (End-to-end automatic speech recognition)

2014 முதல், இறுதியிலிந்து இறுதி தானியக்க பேச்சு அறிதலில் (ASR/ ஏ.எஸ்.ஆர்.) அதிக ஆராய்ச்சி ஆர்வம் உள்ளது. பாரம்பரிய ஒலிப்பு அடிப்படையிலான (அதாவது, அனைத்து எச்.எம்.எம்-அடிப்படையிலான மாதிரி) அணுகுமுறைகளுக்கு உச்சரிப்பு, ஒலி மற்றும் மொழி மாதிரிக்கு தனித்தனி கூறுகள் மற்றும் பயிற்சி தேவை. பேச்சு அறிதலின் அனைத்து கூறுகளையும்

முடிவிலிருந்து இறுதி மாதிரிகள் கூட்டாகக் கற்றுக்கொள்கின்றன. இது பயிற்சி செயல்முறை மற்றும் வரிசைப்படுத்தல் செயல்முறையை எளிதாக்குவதால் இது மதிப்புமிக்கது. எடுத்துக்காட்டாக, அனைத்து எச்.எம்.எம்-அடிப்படையிலான அமைப்புகளுக்கும் ஒரு என்-கிராம் மொழி மாதிரி தேவைப்படுகிறது, மேலும் ஒரு பொதுவான என்-கிராம் மொழி மாதிரி பெரும்பாலும் பல ஜிகாபைட் நினைவகத்தில் எடுத்துக்கொள்கிறது, எனவே மொபைல் சாதனங்களில் பயன்படுத்துவது சாத்தியமற்றது (Jurafsky 2016). இதன் விளைவாக, கூகிள் மற்றும் ஆப்பிள் (Google and Apple) (2017 நிலவரப்படி) ஆகியவற்றிலிருந்து நவீன வணிக ஏ.எஸ்.ஆர் அமைப்புகள் மேகக்கட்டத்தில் பயன்படுத்தப்படுகின்றன, மேலும் உள்நாட்டில் சாதனத்திற்கு மாறாக பிணைய இணைப்பு தேவைப்படுகிறது.

கூகிள் டீப் மைண்டின் (Google DeepMind) அலெக்ஸ் கிரேவ்ஸ் (Alex Graves) மற்றும் டொராண்டோ பல்கலைக்கழகத்தின் (University of Toronto) நவ்தீப் ஜெட்லி (Navdeep Jaitly)ஆகியோரால் 2014இல் அறிமுகப்படுத்தப்பட்ட இணைப்பாளர் காலம்சார் வகைப்பாடு (Connectionist Temporal Classification (CTC/சி.டி.சி)) அடிப்படையிலான அமைப்புகளுடன் ஏ.எஸ்.ஆரின் முதல் முயற்சி இருந்தது (Graves (2014). இந்த மாதிரி மீள்நிகழ்வு நரம்பியல் வலையமைப்புகள் (recurrent neural networks) மற்றும் சி.டி.சி லேயரைக் கொண்டிருந்தது. கூட்டாக, ஆர்.என்.என்-சி.டி.சி (RNN-CTC) மாதிரி, உச்சரிப்பு மற்றும் ஒலி மாதிரியை ஒன்றாகக் கற்றுக்கொள்கிறது, இருப்பினும் ஒரு எச்.எம்.எம் போன்ற நிபந்தனைக்குட்பட்ட சுதந்திர அனுமானங்களின் (conditional independence assumptions) காரணமாக மொழியைக் கற்க இயலாது. இதன் விளைவாக, சி.டி.சி மாதிரிகள் பேச்சு ஒலியியலை ஆங்கில எழுத்துக்களுடன் பொருத்தக் கற்றுக் கொள்ளலாம், ஆனால் மாதிரிகள் பல பொதுவான எழுத்துப்பிழைகளைச் செய்கின்றன மற்றும் டிரான்ஸ்கிரிப்ட்களை சுத்தம் செய்ய ஒரு தனி மொழி மாதிரியை நம்பியிருக்க வேண்டும். பின்னர், பைடு (Baidu) மிகப் பெரிய தரவுத்தொகுப்புகளுடன் இந்த வேலையை விரிவுபடுத்தியது மற்றும் சீன மாண்டரின் மற்றும் ஆங்கிலத்தில் வணிகரீதியான வெற்றியைக் காட்டியது (Amodei 2016). 2016ஆம் ஆண்டில், ஆக்ஸ்போர்டு பல்கலைக்கழகம் லிப்நெட் ("LipNet: How easy do you think lipreading is? 2017) முதல் இறுதி முதல் இறுதி வாக்கிய அளவிலான லிப் ரீடிங் மாதிரியை வழங்கியது; இடகால திணிப்புகளுடன் (spatiotemporal convolutions) ஒரு ஆர்.என்.என்-சி.டி.சி கட்டமைப்பைப் பயன்படுத்தி, கட்டுப்படுத்தப்பட்ட இலக்கண தரவுத்தொகுப்பில் மனித அளவிலான செயல்திறனை விஞ்சியது

(Shillingford, et al 2018). மனித நிபுணர்களை விட 6 மடங்கு சிறந்த செயல்திறனை அடைந்த ஒரு பெரிய அளவிலான சி.என்.என்-ஆர்.என்.என்-சி.டி.சி கட்டமைப்பு 2018இல் கூகிள் டீப் மைண்ட்-ஆல் (DeepMind) வழங்கப்பட்டது (Shillingford et al 2018).

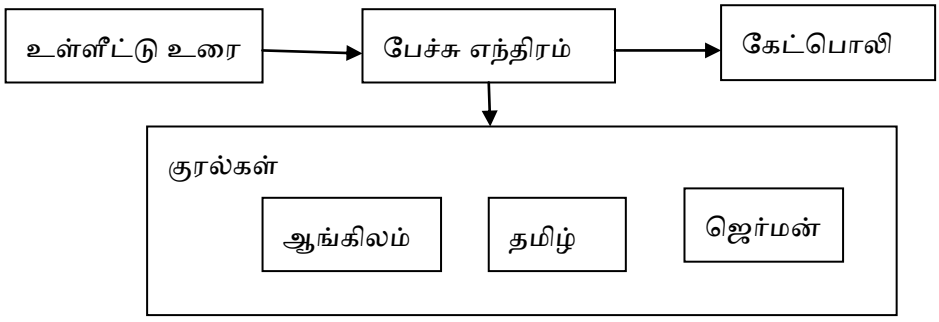
of and and of the

சி.டி.சி அடிப்படையிலான மாதிரிகளுக்கு மாற்று அணுகுமுறை கவனத்தை அடிப்படையாகக் கொண்ட மாதிரிகள் (attention-based models) ஆகும். கவனம் சார்ந்த ஏ.எஸ்.ஆர் மாதிரிகள் ஒரே நேரத்தில் கார்னகி மெலன் பல்கலைக்கழகத்தின் (Carnegie Mellon University) கூகிள் பிரைனின் (Google Brain) சான் மற்றும் பலராலும் (Chan et al. 2016) மற்றும் மாண்ட்ரீல் பல்கலைக்கழகத்தின் (University of Montreal) பஹ்தானாவ் மற்றும் பலராலும் (Bahdanau et al. 2016) 2016இல் அறிமுகப்படுத்தப்பட்டன. ”கேள், கவனி, உச்சரி” ("Listen, Attend and Spell") (LAS)" (எல்ஏஎஸ்) என்று பெயரிடப்பட்ட மாதிரி, ஒலி சமிக்ஞையை உண்மையில் "கேட்கிறது", சிக்னலின் வெவ்வேறு பகுதிகளுக்கு "கவனம் செலுத்துகிறது" மற்றும் ஒரு நேரத்தில் டிரான்ஸ்கிரிப்ட் ஒரு எழுத்தை "உச்சரிக்கிறது". சி.டி.சி அடிப்படையிலான மாதிரிகள் போலல்லாமல், கவனத்தை அடிப்படையாகக் கொண்ட மாதிரிகள் நிபந்தனை-சுதந்திர அனுமானங்களைக் கொண்டிருக்கவில்லை, மேலும் உச்சரிப்பு, ஒலி மற்றும் மொழி மாதிரி உள்ளிட்ட பேச்சு அறிதலின் அனைத்து கூறுகளையும் நேரடியாகக் கற்றுக்கொள்ள முடியும். இதன் பொருள், வரிசைப்படுத்தலின் போது, ஒரு மொழி மாதிரியைச் சுற்றிச் செல்ல வேண்டிய அவசியமில்லை, இது எல்லைக்குட்பட்ட நினைவகம் கொண்ட பயன்பாடுகளுக்கு மிகவும் நடைமுறைக்குரியது. 2016ஆம் ஆண்டின் இறுதியில், கவனத்தை அடிப்படையாகக் கொண்ட மாதிரிகள் சி.டி.சி மாதிரிகளை (வெளிப்புற மொழி மாதிரியுடன் அல்லது இல்லாமல்) விஞ்சுவது உட்பட கணிசமான வெற்றியைக் கண்டன (Chorowski & Jaitly 2016). அசல் ”கேள், கவனி, உச்சரி” (LAS) மாதிரியிலிருந்து பல்வேறு நீட்டிப்புகள் முன்மொழியப்பட்டுள்ளன. ஆங்கில எழுத்துக்களை விட இயற்கையான துணைச் சொல் அலகுகளை நேரடியாக வெளியேற்ற கார்னகி மெலன் பல்கலைக்கழகம், எம்ஐடி மற்றும் கூகிள் மூளை (Google Brain) ஆகியவற்றால் உள்ளுறைந்த வரிசை சிதைவுகள் (Latent Sequence Decompositions (LSD/எல்.எஸ்.டி) முன்மொழியப்பட்டது (Chan et al. 2016); ஆக்ஸ்போர்டு பல்கலைக்கழகமும் கூகிள் டீப் மைண்டும் மனித அளவிலான செயல்திறனை மிஞ்சும் உதடு வாசிப்பைக் கையாள ”கேள், கவனி,

உச்சரி”-ஐ (LAS) "பார், கேள், கவனி, உச்சரி" ("Watch, Listen, Attend and Spell" (WLAS) என்பதற்கு நீட்சி செய்தது (Chung et al. 2016).

6.7.4. **தானியக்க உரையிலிருந்து பேச்சுக் கூட்டிணைப்பாக்கம் (**Automatic speech synthesis systems)

பேச்சுக் கூட்டிணைப்பாக்கம் மனிதப் பேச்சின் செயற்கைத்தன்மையான உற்பத்தி; இது படத்தில் காட்டியுள்ளது போன்று பேச்சு குறிகையை எழுத்துவடிவப் பனுவலாக மாற்றும் செயல்பாடாகும். மேலும் நாம் தமிழ் பனுவல்-பேச்சு ஒழுங்குமுறை உருவாக்குவதாய் இருந்தால் இவ்வொழுங்குமுறை எந்த எழுத்துவடிவ உள்ளீட்டையும் ஏற்கவேண்டும். இதைக் கருத்திற்கொண்டு கணினி பேச்சைப் பற்றி ஒரு சில வேறுபடுத்தல்கள் செய்வது தேவையாகும். ஏதாவது பேச்சை ஒலிப்பதிவு செய்து கணிப்பொறியில் சேமித்துவைத்து ஒலிக்கச்செய்து கேட்கலாம். பல ஒழுங்குமுறைகள் இதைத்தான் செய்கின்றன; விடை எந்திரம் (answer machine) பதிவு செய்யப்பட்ட செய்திகளைத் திரும்ப ஒலிக்கின்றது; வானொலிப்பெட்டி ஏற்கனவே பதிவுசெய்யப்பட்ட நேர்காணலைத் திரும்ப ஒலிக்கின்றது; இவ்வாறு பல ஒழுங்குமுறைகள் உள்ளன. உரை-பேச்சுக்குப் பின்னணியாக வரும் கருத்து என்னவென்றால் "திரும்ப ஒலிக்கப்படும்” செய்திகள் உண்மையிலேயே ஒலிப்பதிவு செய்யப்பட்டவை அல்ல. திரும்ப ஒலிக்கச் செய்யும் நடத்தையிலிருந்து அடுத்தபடியான நடவடிக்கை பொதுவான சொற்களை ஒலிப்பதிவு செய்து திரும்பச் சேர்ப்பதாகும்; இந்த நுட்பம் தொலைபேசி உரையாடல் சேவைகளில் அடிக்கடி பயன்படுத்தப்படுகின்றது. சில வேளைகளில் பயன்விளைவு ஏற்றுக்கொள்ளக்கூடியதாய் இருக்கும்; சில வேளைகளில் ஏற்றுக்கொள்ள இயலாததாய் இருக்கும்; செயற்கையாக ஒன்றுசேர்க்கப்பட்டப் பேச்சு செயற்கையாக இருக்கும். மாறாக உரை-பேச்சு விரும்பப்படுகிற செய்தி உண்மையிலேயே பேசப்பட்டாலும் இல்லாவிட்டாலும் எதையும் பேச்சும் இலக்கைக் கொண்டது.



பேச்சு

பேச்சுக் குறிகை (speech signal) ஒலி மூலத்திற்கும் கேட்பவருக்கும் இடையிலுள்ள ஊடகத்தில் ஏற்படும் அழுத்த மாற்றங்களின் வரிசையாகும். பேச்சுக் குறிகையின் மிகப் பொதுவான உருப்படுத்தம் பெரும்பாலும் அலைவடிவம் (wave form) என்று அழைக்கப்படும் ஆசிலோகிராம் (ocillogram) ஆகும். பேச்சு மனித முன் தொண்டை, தொண்டை, வாய், நாக்கு இவற்றின் வழியாகச் செல்லும் காற்றின் போக்கில் ஏற்படும் தடைகளால் உற்பத்தி செய்யப்படுகின்றது. உயிரொலிகளை உள்ளடக்கிய குரல் ஒலிகள் ஓரளவு சீரான அமைப்பொழுங்கைக் காட்டுகின்றன. குரலொலிகள் இயல்பாக அதிக சக்தியைக் கொண்டிருக்கின்றது; *ஸ்* போன்ற குரலிலா ஒலிகள் சீரற்ற அமைப்பொழுங்கைக் கொண்டிருக்கும். ஒரு பேச்சு குறிகையை ஒலியன்களின் அல்லது நோட்டங்களின் வரிசையாகப் புரிந்துகொள்ளப்படலாம். பொதுவாகக் கூற்று அதிக அளவில் வழக்கமாகப் பயன்படுத்தப்படுகின்ற ஐ.பி.எ. ஒலியியல் நெடுங்கணக்கால் (IPA Phoneic Alphabet) ஒலிபெயர்க்கப்படுகின்றது.

பேச்சு குறிகையை உருப்படுத்தம் செய்யும் மற்றொருவழி இசைமை ஆய்வால் உற்பத்திசெய்யபடும் ஒன்று. பேச்சு இரு பாகங்களைக் கொண்ட இயற்பியல் செயற்பாங்கு ஆகும்: ஒலி மூலத்தின் (குரல்வளை மடல்கள்) மற்றும் வடிகட்டுதலின் (நாக்கு, உதடுகள், பல் போன்றவற்றால்) உற்பத்திப் பொருள். இசைமை ஆய்வு இறுதி பேச்சுக் கூற்றை ஆய்வதால் ஒலி மூலத்தின் அடிப்படை நிகழ்வெண்ணைக் கண்டுபிடிக்க முயலுகின்றது. அடிப்படை நிகழ்வெண் குரல்வளை மடல்களால் உற்பத்தி செய்யப்படும் ஒலியின் முனைப்பான நிகழ்வெண்ணாகும். எவ்வாறு கேட்பவர்கள் பேசுபவர்களின் இசையோட்டதையும் அசை அழுத்தத்தையும்

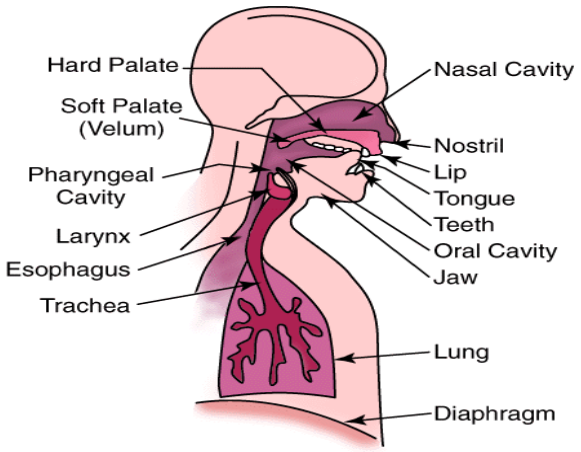
கேட்கின்றனர் என்பதற்கு வலுவான தொடர்புப்பொருத்தம் அடிப்படை நிகழ்வெண்ணாகும். ஆணின் இயல்பு F0 நிகழ்வெல்லை 80-200 Hz ஆகும்; பெண்களுக்கு 150-350 Hz ஆகும்.

பேச்சு உற்பத்தி

பேச்சு நுரையீரலிலிருந்து குரல்வளை, வாயறை, மூக்கறை வழியாகச் செல்லும் காற்றொழுக்கால் உருவாக்கப்படுகின்றது. இது மூன்று செயற்பாங்குகளை உட்படுத்தும்: தொடக்கம் (initiation), இழுமை (phonation), வாய்-மூக்குச் செயற்பாங்கு (oral-nasal process), ஒலிப்பு (articulation). நுரையீரல்களிலிருந்து காற்று வெளியேற்றப்படுவது தொடக்கச் செயற்பாங்கு ஆகும். தொண்டையில் நிகழும் செயற்பாங்கு இழுமை ஆகும்.

தொண்டைக் காற்று வழியில் இரண்டு செங்குத்தான மடல்களைக் கொண்டிருக்கின்றது; அவை குரல்வளை மடல்களாகும் (vocal folds). மடல்களுக்கு இடையிலுள்ள இடைவெளி குரல்வளையாகும் (glottis). குரல்வளையை அதன்வழியே காற்று போகாதபடி மூட இயலும் அல்லது குரல் ஒலிகளை ஒலிக்கும் விதத்தில் குரல்வளை மடல்களை அதிரச்செய்ய இயலும் படி குறுகிய திறப்பைக் கொண்டிருக்கச் செய்யவியலும். இறுதியாக இயல்பான மூச்சுவிடும் போது அதை அதிகமாகத் திறக்க இயலும்; இதனால் குரல்வளை மடல்களின் அதிர்வு குறைக்கப்பட்டு குரலிலா ஒலிகள் உற்பத்தி செய்யப்படும்.

தொண்டை, மேல்தொண்டை வழியாகச் சென்ற பின்னர் காற்று வாயறை அல்லது மூக்கறை வழியாகச் செல்ல இயலும். பின்னண்ணம்/கடையண்ணம் இந்தத் தேர்வுக்குப் பொறுப்பான பகுதியாகும். வாய்-மூக்குச் செயற்பாங்கு மூக்கு மெய்யொலிகள் மற்றும் பிற ஒலிகளைத் தீர்மானிப்பதற்கு உதவும். இறுதிச் செயற்பாங்கு வாயறையில் நிகழும் ஒலிப்புச் செயற்பாங்கு ஆகும்; இந்தச் செயற்பாங்கு பெரும்பாலான பேச்சொலிகளை வேறுபடுத்த மனிதனுக்கு உதவுகின்றது. பேச்சு உற்பத்தியின் உடலியங்கியல்/உடற்கூறியல் கீழே படத்தில் தரப்பட்டுள்ளது. பேச்சு உற்பத்தி பற்றி இயல் 3இல் விரிவாகக் கூறப்பட்டுள்ளது.



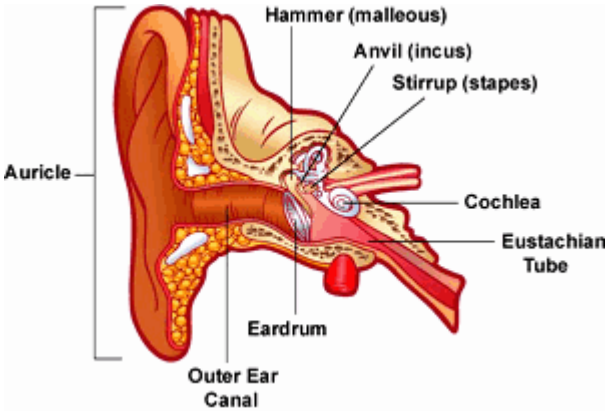
Hard palate ‘வல்லண்ணம்’, trachea ‘முச்சுகுழல், teeth ’பல், Soft palate ‘மெல்லண்ணம்’, nasal cavity ‘மூக்கறை, oral cavity ‘வாயறை’ Pharangeal cavity ‘மேல்தொண்டைஅறை, nostril ‘மூக்குத் துளை, jaw ‘தாடை, Larynx ‘தொண்டை’ lip ‘இதழ்’, lung ‘நுரையீரல்’, Esophagus ‘உணவுகுழல்’, tongue ‘நா’, diapharam ‘உந்தி/உதரவிதானம்’.

பேச்சு கேட்புணர்வு

கேட்புணர்வு ஒழுங்கமைப்பில் இரண்டு முக்கியமான கூறுகள்/பகுதிகள் உள்ளன: வெளிப்படையான கேட்பு உறுப்புகள் (காதுகள்) மற்றும் கேட்பு நரம்பு ஒழுங்குமுறை (மூளை). காது ஒலிப்பு அழுத்த குறிகையை ஆய்கின்றது; முதலில் அதைப் பேசிலியார் தோல்/ஜவ்வு மீது எந்திரத்தன்மையான அதிர்வு அமைப்பொழுங்குகளாக மாற்றுகின்றது; பின்னர் கேட்பு நரம்பால் கடத்தப்பட்ட துடிப்புகளின் வரிசைகளால் உருப்படுத்தம் செய்கின்றது. கேட்புணர்வுத் தகவல் கேட்பு நரம்பு ஒழுங்கமைப்பின் பல்வேறு நிலைகளில் பிரித்தெடுக்கப்படுகின்றது.

மனிதக் காது மூன்று பகுதிகளைக் கொண்டது: வெளிக்காது, நடுக்காது, அகக்காது. வெளிகாது காணக்கூடிய வெளிப்பகுதி, ஒலி பயணிக்கும் குழாய் வடிவ வெளி கேட்புக் குழாய் (external auditory canal) இவற்றைக் கொண்டிருக்கும். குழாய் அதன் ஒரு முனையில் காதுமுரசால் (ear drum) மூடப்பட்டிருக்கும். காற்று அழுத்த வேறுபாடுகள் காதுமுரசை அடைந்தது அதை அதிரச்செய்யும்; இந்த அதிர்வுகளை அதன் எதிர் பக்கம் இருக்கும் எழும்புகளுக்குக் கடத்தும். உள்வரும் ஒலி அழுத்த அலை போன்ற அதே நிகழ்வெண்ணில் (சுருங்குதல், விரிதல் இவை மாறிமாறி நிகழ்தல்) காதுமுரசின் அதிர்வெண் இருக்கும். நடுக்காது காற்று நிரப்பப்பட்ட இடைவெளி அல்லது அறை ஆகும். இந்த அறையுடன் மூக்கையும் தொண்டையையும் இணைக்கும் குழாய்வழி காற்று நடுக்காது அறைக்குப் பயணிக்கும். அகக் காதுக்கு (காக்ளியா)

(cochlea) எலும்பு இடைமுகத்தில் இருக்கும் சிறிய தோல் ஓ வடிவச் சாளரம் (oval window) ஆகும். காக்ளியாவின் சுவர்கள் எலும்பாக இருப்பதால் சக்தியானது ஸ்டேப்களின்/உட்செவி எலும்புகளின் (stapes) எந்திரம்சார் செயல்பாட்டல் ஓ-வடிவ சாளரத்தின் மேல் போர்த்தப்பட்டுள்ள/இழுக்கப்பட்டுள்ள தோல்/ஜவ்வுமீது ஒரு பதிப்பாகக் கடத்தப்படுகின்றது. காதின் அமைப்பு படத்தில் காட்டப்பட்டுள்ளது.



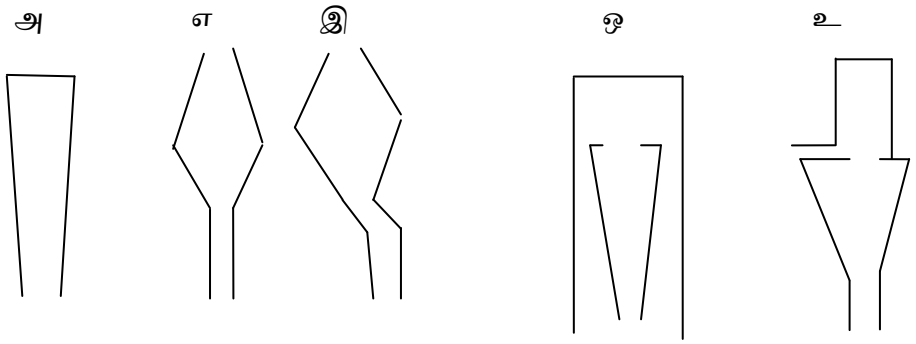
Hammer ‘சுத்தி’, Auricle ‘காதுமடல்’, Outer ear canal ‘வெளிக்காது குழாய், Anvil ‘பட்டறை எலும்பு’, Stirrup ‘அங்கவடி எலும்பு, Cochlea ‘அகக்காது, Ear drum ‘காது முரசு, Eustachian tube ‘நடுச்செவிக் குழல்’

ஒலி கேட்புணர்வுக்கு அகக்காதின்/உட்காதின் தேவையான அமைப்பு/பகுதி காக்ளியா ஆகும்; இது ஒலியின் உருப்படுத்தத்தை மூளைக்கு கடத்தும் கேட்பு நரம்புடன் நேரடியாகத் தொடர்புகொள்கின்றது. சுருள் பகுக்கப்பட்டுள்ளது; முதன்மையாக நீளவாக்கில் ஓடும் பாசிலார் ஜவ்வாலும்/தோலாலும் இரண்டு திரவம் நிரப்பப்பட்ட அறைகளாலும் பகுக்கப்பட்டுள்ளது. காக்ளியாவை ஏறக்குறைய வடிகட்டும் வங்கியாயக் கருத இயலும்; இதன் வெளியீடுகள் அதிர்வெண்-இட மாற்றம் நிறைவேற்றப்பட வேண்டி இடத்தால் ஒழுங்குபடுத்தப்பட்டுள்ளது. காக்ளியாவின் அடிக்கு அண்மையிலிருக்கும் வடிகட்டிகள் உயர்ந்த நிகழ்வெண்களுக்கு பதிலளிப்பு செய்கின்றது; உச்சிக்கு அண்மையில் இருப்பவைகள் குறைந்த நிகழ்வெண்களுக்குப் பதிலளிப்பு செய்கின்றது. நரம்புக் கடத்தல் செயற்பாங்கு இதைத் தொடர்கின்றது மற்றும் ஸ்பெக்ட்ரல் குறிகையைக் (spectral signal) கிட்டத்தட்ட பண்புக்கூறு பிரித்தெடுக்கும் கூறு/பகுதிக்கு பொருத்தமுறும் கேட்பு நரம்பின் (auditory nerve) செயல்பாட்டு குறிகைகளாக மாற்றுகின்றது.

6.7.4.1. **பேச்சுக் கூட்டிணைப்பாக்கத்தின் வரலாறு**

பல நூற்றாண்டுகளாக மனித இனத்தின் கனவாகச் செயற்கைப் பேச்சு இருந்தது. குறிகை ஆய்வைக் கண்டுபிடிப்பதற்கு முன்னர் மனிதப் பேச்சை உருவாக்க இயலும் சில எந்திரங்களை உருவாக்குவதற்கு முயற்சிகள் மேற்கொள்ளப்பட்டன. பேச்சுக் கூட்டிணைப்பாக்கம் (speech synthesis) இயந்திரச்சார்பிலிருந்து மின்சாரம்சார் கூட்டிணைப்பாக மாறியது; இயந்திரம்சார் கூட்டிணைப்பாக்கம் 1779-இல் தொடங்கப்பட்டது. 1779-இல் ரஷ்யப் பேராசிரியரான கிறிஸ்டியன் கிராட்செண்டைன் (Christian Kratzenstein) ஐந்து நெட்டுயிர்களுக்கு (அ, எ, இ, ஒ, உ) இடையேயுள்ள உறுப்பியல்சார் வேறுபாடுகளை விளக்கினார் (History and Development of Speech Synthesis 2005); அவர் அவற்றைச் செயற்கையாக உற்பத்திசெய்யக் கருவியை உருவாக்கினார். கிராட்செண்டைன் படத்தில்காட்டியுள்ளது போன்று மனிதப் பேச்சுக்குழாயுடன் ஒற்றுமையுள்ள ஒலியியக்க ஒலியூக்கிகளை (acoustic resonators) உருவாக்கினார்; இசைக்கருவிகள் போன்று அதிர்வுறும் நாக்கால் ஒலியூக்கிகளைச் செயற்படுத்தினார்.

கிராட்செண்டைனின் ஒலியூக்கிகள் (Kratzenstein's resonators)

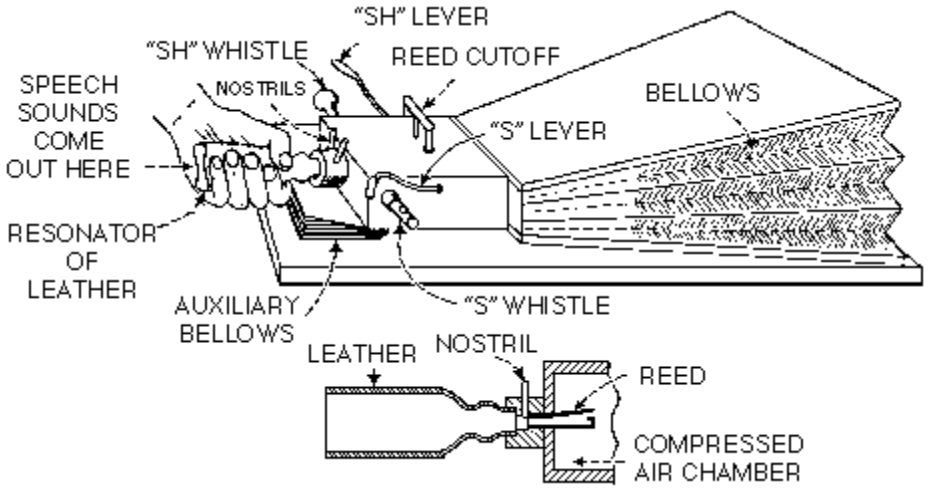


1791-இல் வோல்ஃப்காங்க் வோன் கெம்ப்ளென் (Wolfgang von Kempelen) "ஒலியியக்க எந்திரப் பேச்சு எந்திரம்" (Acoustic-Mechanical Speech Machine) என்பதை அறிமுகப்படுத்தினார். அது சில தனி ஒலிகளையும் ஒலிச் சேர்க்கைகளையும் உற்பத்தி செய்ய முடிந்தது.

வோல்ஃப்காங்க் வோன் கெம்ப்ளென் ஒலியியக்க எந்திர பேச்சு எந்திரம்

அதன்பின்னர் 1837-இல் சார்லஸ் வீட்ஸ்டோன் (Charles Wheatstone) என்பவர் வான் கெம்ப்ளெனின் மாதிரி அடிப்படையில் ஒரு பேசும் எந்திரத்தை உருவாக்கினார். இந்தப் பேச்சு எந்திர வகை உயிரொலிகளையும் பெரும்பாலான மெய்யொலிகளையும் உற்பத்திசெய்ய முடிந்தது; இது ஒலிச்சேர்க்கைகளையும் முழுச் சொற்களையும் உற்பத்தி செய்ய முடிந்தது.

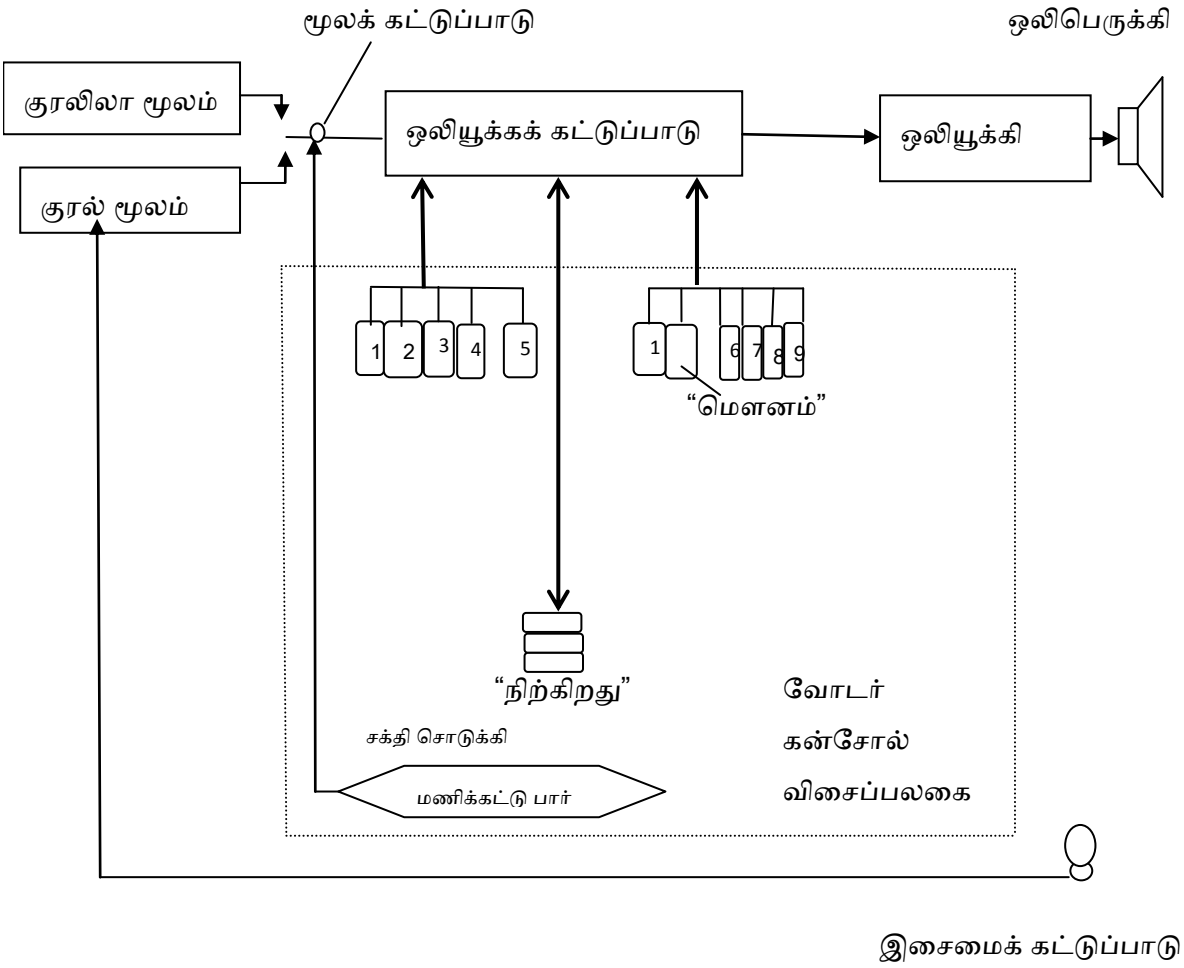
வீட்ஸ்டோனின் பேசும் எந்திரம் (Wheatstone's speaking machine)



மின்னியல்சார் கூட்டிணைப்பாக்கக் கருவி (Electrical synthesis device) 1922-இல் உருவாக்கப்பட்டது. முதல் மின்னியல்சார் கூட்டிணைப்புக் கருவி ஸ்டேவார்ட் (Stewart) என்பவரால் அறிமுகப்படுத்தப்பட்டது. இந்த எந்திரம் தனி நிலையான உயிரொலிகளை உற்பத்தி செய்தது; ஆனால் மெய்யொலிகளையும் கூற்றுக்களையும் உற்பத்தி செய்யவில்லை. 1932-இல் ஒபாட்டா (Obata), டெஷிமா (Teshima) என்ற ஜப்பனிய ஆய்வாளர்கள் உயிரொலிகளில் மூன்றாவது ஒலிச்செறிவைக் கண்டுபிடித்தனர். முதல் மூன்று ஒலிசெறிவுகளும் பொதுவாகப் புரியத்தக்கப் பேச்சு கூட்டிணைப்பாக்கத்திற்குப் போதுமானதாகக் கருதப்பட்டது.

இருப்பினும் பேச்சுக் கூட்டிணைப்பானாகக் (speech synthesizer) கருத்தப்பட்ட முதல் கருவி 1939-இல் ஹோமர் டட்லே (Homer Dudley) என்பவரால் அறிமுகப்படுத்தப்பட்டது. இந்தக் கருவி வோடர் (VODER – Voice Operating Demonstrator) என்று பெயரிடப்பட்டது. இது படத்தில் காட்டப்பட்டுள்ளது. இது விசைப்பலகையால் இயக்கப்படும் மின்னணுசார் பேச்சு ஆய்வியாகும்; கூட்டிணைப்பான் 1930களின் மையத்தில் பெல் லாபால் (Bell lab) உருவாக்கப்பட்டது. வோடரைப் பயன்படுத்துவது கடினம். மட்டுமன்றி இந்த இயந்திரத்தால் உற்பத்திசெய்யப்பட்ட பேச்சின் பண்பும் புரிதிறனும் நல்லதல்ல. இந்த ஒழுங்குமுறை செயற்கைப் பேச்சை உற்பத்தி செய்வதில் அதன் திறன் காரணமாகப் புகழ்பெற்றது.

வோடர் பேச்சுக் கூட்டிணைப்பான் (VODER speech synthesizer)



வோடரின் செய்முறை விளக்கத்திற்குப் பின்னர் அறிவியல் உலகம் பேச்சு கூட்டிணப்பாக்கத்தில் அதிக விருப்பத்தைக் காட்டியது. வோடரின் அடிப்படை அமைப்பும் கருத்தும் பேச்சின் மூலம்-வடிகட்டுவான்–மாதிரி (source-filter-model) தற்போதைய ஒழுங்குமுறைகள் போன்றதே. பிராங்கிளின் கூப்பர் (Franklin Cooper) மற்றும் அவர்தம் உதவியாளர்கள் 1951-இல் ஹாஸ்கின்ஸ் ஆய்வுக்கூடத்தில் (Haskins laboratory) அமைப்பொழுங்கு திருப்பி ஒலிக்கும் கூட்டிணைப்பானை (pattern playback synthesizer) உருவாக்கினர். இது பதிவுசெய்யப்பட்ட நிறமாலை நிழற்பட/ஒலிவர்ணனைப்பட அமைப்பொழுங்கை மூல அல்லது மாற்றப்பட்ட வடிவில்

ஒலிகளாகத் திரும்ப மாற்றியது. நிறமாலை நிழற்பட/ஒலிவர்ணனைப்பட அமைப்பொழுங்கு ஒளிப்புகவல்ல பெல்ட் மீது ஒளிசார்ந்ததாக/ஆப்டிக்கலாகப் பதிவிக்கப்படுகின்றது.

PAT (Pragmatic Artificial Talker) முதல் ஒலிச்செறிவு கூட்டிணைப்பாக்கி ஆகும்; இது வால்டர் லாரன்சால் (Walter Lawrence) 1953-இல் அறிமுகப்படுத்தப்பட்டது. பாட் இணையாக இணைக்கப்பட்ட மின்னணு ஒலிச்செறிவு ஒலியூக்கிகளைக் கொண்டிருக்கின்றது. உள்ளீட்டு குறிகை பஸ்-ஆகவோ இரைச்சலாகவோ (buzz or noise) இருக்கும். நகரக்கூடிய கண்ணாடி ஸ்லைடு (moving glass slide) மூன்று ஒலிச்செறிவு நிகழ்வெண்களை (formant frequencies), குரல், அலைவீச்சு, அடிப்படை அதிர்வெண் (fundamental frequency), இரைச்சல் அலைவீச்சு (noise amplitude) இவற்றைக் கட்டுப்படுத்த ஆறு காலச் செயல்பாடுகளாக வர்ணம்பூசிய அமைப்பொழுங்குகளை மாற்றப் பயன்படுத்தினர். குன்னார் ஃபண்ட் (Kunnar Fant) 1950-இல் முதலில் ஒழுக்கு ஒலிச்செறிவு கூட்டிணைபாக்கி (cascade formant synthesizer) ஓவ் 1 (OVE I) (OVE- Orator Verbis Elecris) அறிமுகப்படுத்தினார்; இது ஒழுக்கில் இணைக்கப்பட்ட ஒலிச்செறிவு ஒலியூக்கிகளைக் கொண்டது. 1962-இல் ஃபாண்டும் மார்ட்டோனியும் (Fant and Martony) மேம்படுத்தப்பட்ட ஓவ் 2 (OVE II) கூட்டிணைப்பாக்கியை அறிமுகப்படுத்தினார்; இது உயிரொலிகள், மூக்கொலிகள், தடைமெய்யொலிகள் இவற்றிற்குப் பேச்சுக்குழலின் கடத்தல் செயல்பாட்டை மாதிரிப்படுத்தத் தனிப் பகுதிகளைக் கொண்டிருக்கின்றது. குரல், உயிர்ப்பு இரைச்சல், உராய்வு இரைச்சல் என்பன சாத்தியமான துண்டல்கள் ஆகும்.

பாட் மற்றும் ஓவ் (PAT and OVE) இவற்றை ஆய்ந்த பின்னர் ஜாண் ஹோல்ம்ஸ் (John Holmes) 1972-இல் இணை ஒலிச்செறிவு கூட்டிணைப்பாக்கியை (paralle formant synthesizer) அறிமுகப்படுத்தினார். 1978-இல் ரிச்சர்ட் காக்னான் (Richard Gagnon) மலிவான வோட்ராக்ஸ் அடிப்படையிலான டைப்-என்–டாக் ஒழுங்குமுறையை (Votrax-based Type-n-Talak system) அறிமுகப்படுத்தினார். 1982-இல் ஸ்ட்ரீட் எலக்ட்ரானிக்ஸ் (Street Electronics) எதிரொலி மலிவு ஒலியிருமை கூட்டிணைப்பாக்கியை (Echo low-cost diaphone synthesizer) அறிமுகப்படுத்தினார். தற்காலப் பேச்சுக் கூட்டிணைப்பாக்கத் தொழில்நுட்பங்கள் மிகச் சிக்கலான நவீனமான நெறிமுறைகளையும் வழிமுறைவரைவுகளையும் உள்ளடக்கியது. ஒன்றிணைப்பாக்கப் பேச்சுக் கூட்டிணைப்பாக்கம், ஹிட்டன் மார்க்கோவ் மாதிரிகள் (HMM models) அடிப்படையிலான கூட்டிணைப்பாக்கம் என்பன அண்மையில் பயன்படுத்தப்பட்டுள்ள தற்கால நெறிமுறைகள் ஆகும்.

6.7.4.2. **பேச்சுக் கூட்டிணைப்பாக அணுகுமுறைகள்**

பேச்சு மனிதவினம் ஒருவருக்கொருவர் கருத்துக்களைப் பரிமாறிக்கொள்ளும் இயற்கையான முறையாகும். பேச்சுக் கூட்டிணைப்பாக்கம் செய்கையாக, குறிப்பாகக் கணினியைப் பயன்படுத்திப் பேச்சைக் கூட்டிணைப்பாக்கம் செய்யும் செயல்முறையாகும். மனிதப் பேச்சை பாவிக்கும்/போலச்செய்யும் ஒலி உருவாக்கம் கீழ் மட்டக் கூட்டிணைப்பாக்கம் என்று கூறப்படுகின்றது. மேல்மட்ட கூட்டிணைப்பாக்கம் எழுதப்பட்ட உரையை/பனுவலை அல்லது குறியீடுகளைக் கீழ்மட்ட உருவாக்க ஒழுங்குமுறையை இயக்கப் பொருத்தமான விருப்பப்பட்ட ஒலியியக்கக் குறியீடுகளின் அருவ உருப்படுத்தமாக மாற்றுவதை விளக்குகின்றது. உற்பத்தியின் நெறிமுறைகளின் அடிப்படையில் பேச்சு கூட்டிணைப்பாக்கத்தை நான்கு முக்கியமான வகைப்பாடுகளாக வகைப்படுத்தலாம்.

1 ஒலிப்பியல் கூட்டிணைப்பாக்கம் (Articulatory synthesis)

2. ஒலிச்செறிவு கூட்டிணைப்பாக்கம் (Forman snynthesis)

3. இணைப்பு கூட்டிணைப்பாக்கம் (Concatenative synthesis)

4. எச்.எம்.எம். அடிப்படையிலான கூட்டிணைப்பாக்கம் (HMM based synthesis)

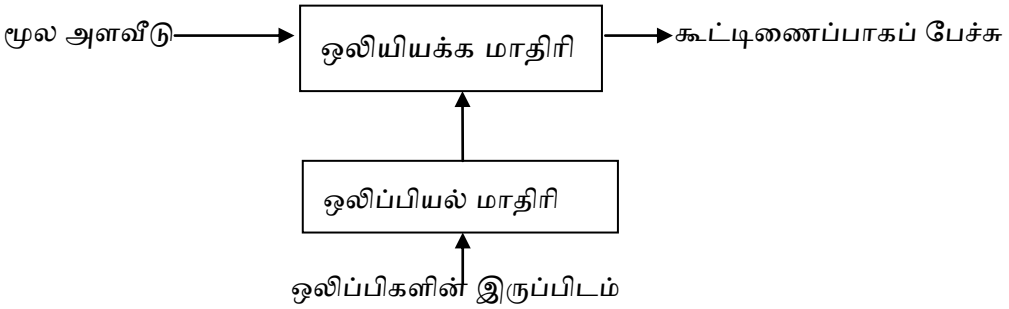
6.7.4.2.1. **ஒலிப்பியல்சார் கூட்டிணைப்பாக்கம்**

பேச்சுக் குழாயின் மாதிரியைப் பயன்படுத்தி பேச்சு ஒலிகளை உற்பத்திசெய்வது ஒலிப்பியல் சார் கூட்டிணைப்பாக்கம் (articulatory synthesis) ஆகும்; இது நேரடியாகவோ மறைமுகமாகவோ பேச்சு ஒலிப்பிகளின் இயக்கத்தை ஊக்குவிக்கின்றது. பேச்சுக்குழாயின் வடிவத்தைப் பல வழிகளில் கட்டுப்படுத்த இயலும்; இது நாக்கு, தாடை, உதடுகள் போன்ற பேச்சு ஒலிப்பிகளின் இருப்பை மாற்றுவதை உட்படுத்தும். பேசுக்குழாயின் உருப்படுத்தம் வழி காற்றின் ஓட்டத்தை டிஜிட்டலாக போலச்செய்வதால்/பாவிப்பதால் பேச்சு உருவாக்கப் படுகின்றது. ஒலிப்பியல்சார் கூட்டிணைப்பாகத்தில் இணை ஒலிப்பு விளைவுகள் இயற்கையாக வருகின்றது. குரல்வளைசார் மூலப் பண்புக்கூறுகள், பேச்சுக்குழாய் மற்றும் குரல்வளை மடல்கள் இவற்றிற்கிடையேயுள்ள ஊடாட்டம், துணை குரல்வளைசார் ஒழுங்குமுறையின் பங்களிப்பு மற்றும் மூக்குக் குழாய் மற்றும் மூக்குத் துளைகள் (sinus cavities) இவற்றுடன் இதைச் சரியாகக் கையாள இயலும்.

ஒலிப்பியல்சார் உருவாக்கம் படத்தில் காட்டியுள்ளது போன்று இரு மதிரிகளை இணைக்கின்றது. ஒலிப்பியல் மாதிரியில் பேச்சுக் குழல் பல சிறிய பகுதிகளாகப்

பகுக்கப்பட்டுள்ளது மற்றும் தொடர்புடைய குறுக்கு வெட்டுப் பகுதிகள் (cross-sectional areas) பேச்சுக் குழாய்ப் சிறப்புப் பண்புகளை உருப்படுத்தம் செய்வதற்கு அளவீடுகளாகப் பயன்படுத்தப்படுகின்றது. ஒலியியக்கவியல் மாதிரியில் ஒவ்வொரு குறுக்கீடான பரப்புகளும் மின்சார அனலாக் கடத்து லைனால் (electrical analog transmission line) தோராயமாக்கப்பட்டுள்ளது.

கீழே தரப்பட்டுள்ள ஒலிப்பியல் சார் மாதிரி பேச்சு உற்பத்தியின் முக்கியமான கூறுகளை உருப்படுத்தம் செய்கின்றது; இது தனது எல்லைகளை முறையே குரல்வளையிலும் வாயிலும் இடமாகக் கொண்ட ஒலியியக்கக் குழாயின் குறுக்குவெட்டுப் பரப்பின் வேறுபாட்டைப் பிரதிபலிக்கும் பரப்புச் செயல்பாட்டின் கணிப்பு அதன்முக்கிய நோக்கமாகும்.



பேச்சுக்குழாயின் ஜியாமண்டரிக் பரப்பு (geomenteric area) ஒலிப்பைப் பற்றிய நமது புரிந்துகொள்ளலுக்கு முக்கியமாகும் மற்றும் அது பேச்சு உருவாக்கத்திற்கு முக்கியக் காரணியாகும். பேச்சு உருவாக்கத்தின் ஒலியியக்கக் கோட்பாடு அடிப்படையில் சீரற்ற மற்றும் கால மாறுபாட்டுடைய குறுக்குவெட்டுப் பரப்புகள் மனிதப் பேச்சுக் குழாயை ஒரு ஒலியக்கக் குழாயாக (acoustic tube) மாதிரியாக்கம் செய்ய இயலும். வேறுபட்ட மொழி ஒலிகளை உற்பத்தி செய்ய அது கிளர்ச்சி மூலத்தை (excitation source) ஒழுங்குபடுத்துகிறது. ஒலிப்பு இடங்கள் குறிப்பிடப்பட்டுவிட்டால் குறுக்குவெட்டுப் பரப்புகள் பேச்சுக்குழாய் எல்லைக்கோட்டின் மீது ஒரு சல்லடை அமைப்பை (grid structure) ஒன்றின் மேல் ஒன்றாக வைப்பதால் கணிக்கப்படுகின்றது. இந்த சல்லடைக் கோடுகள் (grid lines) ஒலிப்பிகளின் இருப்பிடம் அடிப்படையில் வேறுபடுகின்றது (அவைகள் மெர்மெல்ஸ்டென் மாதிரியில் (Mermelstein's model) நிறுவப்படுகின்றது). மொத்தம் 60 பகுதிகள், பேச்சுகுழாய்க்கு 59 பகுதிகள் மற்றும் நமது மாதிரியில் பயன்படுத்தப்படும் குரல்வளையின் வெளியேற்றும் வழிக்கு ஒரு பகுதி (நிலையான

நீளம் மற்றும் பரப்பு) நமது மாதிரியில் பயன்படுத்தப்படுகின்றது. அம்புவடிவ தூரம் (sagittal distance) பின்வய-மேல்வய மற்றும் முன்வய-கீழ்வய வெளிஎல்லைக் கோடுகளுக்கு இடையேயுள்ள சல்லடைக் கோடுப் பகுதியாக வரையறை விளக்கம் செய்யப்படுகின்றது.

பேச்சுக்குழாயின் மையக்கோடு பக்கத்திலுள்ள சல்லடைக் கோடுகளின் மையப் புள்ளிகளை இணைப்பதால் உருவாக்கப்படுகின்றது. மையக் கோட்டின் நீளம் பேச்சுக்குழாயின் நீளத்திற்குச் சமமானதாகக் கருதப்படுகின்றது. அம்புவடிவ தூரம் இறுதியில் அனுபவாத சூத்திரங்களால் (empiric formulas) குறுக்குவெட்டு பரப்புகளாக மாற்றப்படுகின்றது. பேச்சு உற்பத்தியின் ஒலியியக்கக் கோட்பாடு அடிப்படையில் நாம் தரப்பட்ட பேச்சுக்குழல் குறுக்குவெட்டுப் பரப்பிலிருந்து ஒலிச்செறிவு நிகழ்வெண்களைக் கணிக்க இயலும். தரப்பட்ட பேச்சுக்குழலின் ஒலியியக்கவியல் மாற்றச் செயல்பாட்டைக் கணிப்பதிலிருந்து ஒலிச்செறிவு நிகழ்வெண்களை ஆக்கக்கூறுகளாகப் பிரிப்பது எளிது. பேச்சுக்குழல் குறுக்குவெட்டுப் பரப்பு பேச்சுக் குறிகளைப் பற்றிய ஒலியியக்கவியல் தகவல்களைத் (ஒலிச்செறிவு நிகழ்வெண்களை) தருகின்றது. இறுதியாக நாம் அதைக் கட்டுப்பாடுகளுடன் கூடிய பல்பரிமாண வரிசையல்லாத உத்தமமாக்கச் சிக்கலாகக் (multi dimensional nonlinear optimization problem) கருதுகின்றோம் மற்றும் உத்தம ஒலிப்பு வெக்டாராகக் (optimum articulatory vector) கணிக்கின்றோம்.

மனிதப் பேச்சு ஒழுங்கமைப்பின் ஒலியியக்க மாதிரி பல துணைமாதிரிகளைக் கொண்டுள்ளது. பேச்சுக்குழல் மற்றும் மூக்குக்குழல் மாதிரிகள் (vocal tract and nasal tract models) இந்தக் குழல்களில் ஓசை பரவலைப் பாவிக்கின்றது. கிளர்வுறல் மூல மாதிரி பேச்சுக்குழலுக்கு குரல் கிளர்வுறல் அலைவடிவுகளை (voice excitation waveforms) உருவாக்குகின்றது. உரசொலிகள் மற்றும் அடைப்பொலிகளுக்குரிய இடுக்குகளில் கிளர்ந்தெழும் காற்றொழுக்கு ஓசை மூலத்தால் உருவாக்கப்படுகின்றது. பரவல் மாதிரி (radiation model) உதடுகளிலிருந்தும் மூக்குத்துவாரங்களிலிருந்தும் பரவும் ஒலியியக்கச் சக்தியைப் பாவிக்கின்றது. பேச்சுக்குழல், எலும்பு உட்புழைகளுடன் கூடிய மூக்குக் குழல், குரல்வளை தடை, துணைக் குரல்வளை குழல், கிளர்தல் மூலம், மற்றும் கிளர்தல் இரைச்சல் மூலம் இவற்றை உட்படுத்தும் பேச்சு ஒழுங்குமுறையின் கடத்தல்-கோடு சுற்று மாதிரி (transmission-line circuit model) உருவாக்கப்படுகிறது. ஒவ்வொரு பேச்சுக் குழல் துணை ஒழுங்கமைப்பின் ஒலியியக்க மாதிரியும் ஆயப்படுகிறது.

ஒரு நடைமுறை ஒலிப்பு கூட்டிணைப்பாக்கி (practical articulatory synthesizer) போச்சுக்குழல், எலும்பு உட்புழைகளுடன் கூடிய மூக்குக் குழல், குரல்வளைத் தடைகள், துணை குரல்வளை ஒழுங்குமுறை, கிளர்வுறல் மூலம், கிளர்தல் ஓசை மூலம் இவற்றை உட்படுத்தும். பேச்சு ஒழுங்குமுறையின் ஒலியியக்க சமன்பாடுகள் பரிந்துரைக்கப்பட்ட ஒலிப்பு கூட்டிணைப்பாக்கிக்கு வேண்டி உருவாக்கப்படுகின்றது. காலப்புல அணுகுமுறை (time-domain approach) பேச்சு ஒழுங்குமுறையின் இயக்கப் பண்புக்கூறுகளைப் பாவிக்கப் பயன்படுத்தப்படுகின்றது. பேச்சுக் குழல் குறுக்குவெட்டுப் பரப்பு அல்லது ஒலிப்பு அளகைகள் (articulatory parameters), நேர்கோட்டு அல்லது ஆர்க் டான் செயல்பாட்டைப் (arc tan function) பயன்படுத்தி இரு அடுத்தடுத்த சட்டங்களுக்கு இடையில் செருகப்படுகின்றது. முறையாகத் திட்டமிடப்பட்ட ஒலிப்பு உருவாக்கிகள் உரசொலிகள் மற்றும் அடைப்பொலிகளின் எல்லா இயல்பான தேவையான விளைவுகளை மீள்ளுருவாக்கம் செய்யவும் உண்மையான பேச்சு உருவாக்கத்தில் நிகழும் பௌதிகச் செயல்பாடுகளை ஒக்கும் விதத்தில் இணை ஒலிப்பு மாற்றங்களையும் மூலக் குழாய் ஊடாட்டத்தையும் மாதிரிப்படுத்த இயலும்.

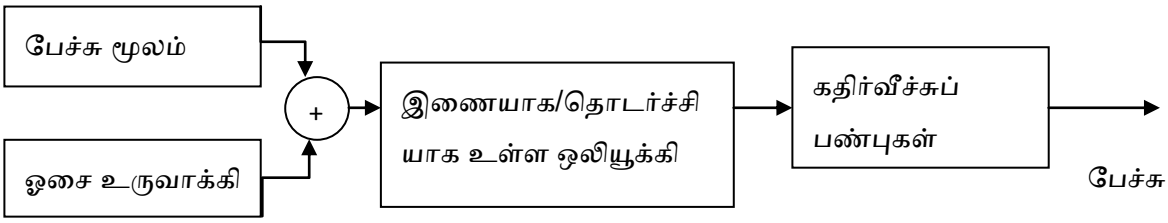
6.7.4.2.2 **ஒலிச்செறிவுக் கூட்டிணைப்பாக்கம்**

ஒலிச்செறிவு கூட்டிணைப்பாக்கம் (formant synthesis) ஓட்டநேரத்தில் மனிதப் பேச்சு மாதிரிகளைப் பயன்படுத்துவதில்லை. மாறாக ஒலியியக்க மாதிரிகளைப் பயன்படுத்தி வெளியீடு கூட்டிணைப்பாக்கம் செய்யப்பட்ட பேச்சு (output synthesized speech) உருவாக்கப்படுகின்றது. அடிப்படை நிகழ்வெண், குரல்படுத்தல் (voicing), இரைச்சல் நிலைகள் (noise levesl) போன்ற அளகைகள் செயற்கைப்பேசின் அலைவடிவை உருவாக்கும் படி கால அடிப்படையில் மாற்றப்படுகின்றது. இந்த அணுகுமுறை சிலவேளைகளில் விதி அடிப்படையிலான உருவாக்கம் (rule-based synthesis) என்று அழைக்கப்படுகின்றது; இருப்பினும் சிலர் பல ஒன்றிணைக்கும் ஒழுங்குமுறைகள் ஒழுங்குமுறையின் சில பகுதிகளுக்கு விதி அடிப்படையிலான உட்கூறுகளை/பகுதிகளை பயன்படுத்துகின்றன என்று வாதிடுகின்றனர்,

ஒலிச்செறிவு உருவாக்கத் தொழில்நுட்பம் அடிப்படையிலான பல ஒழுங்குமுறைகள் கீழ்வரும் படத்தில் காட்டியுள்ளது போல் செயற்கையான, ரபோட்டிக் ஓசைப் பேச்சு மற்றும் வெளியீட்டை உருவாக்குகின்றது; இருப்பினும் அதிக இயற்கைத்தன்மை பேச்சு கூட்டிணைப்பாக்க ஒழுங்குமுறையின் (speech synthesis system) இலக்கு அல்ல; ஒலிச்செறிவு

ஒழுங்குமுறைகள் (formant synthesis systems) ஒன்றிணைக்கும் ஒழுங்குமுறைகளைவிட (concatenative systems) சில அனுகூலங்களைக் கொண்டுள்ளன.

ஒலிச்செறிவு கூட்டிணைப்பாக்கம் செய்யப்பட்ட பேச்சு மிக அதிக வேகத்திலும் ஒன்றிணைப்பு ஒழுங்குமுறைகளைச் சிதைக்கும் ஒலியியக்கத் தடுமாற்றத்தை விலக்குவதால் நம்பகமான அளவில் புரியக்கூடியது. இரண்டாவது ஒலிச்செறிவுக் கூட்டிணைப்பாக்கிகள் ஒன்றிணைப்பு ஒழுங்குமுறைகளைவிட பெரும்பான்மையும் குறைந்த வழியமைப்புவரைவுகள் கொண்டது; ஏனென்றால் அவை பேச்சு மாதிரிகளின் தரவு அடிதளத்தைக் கொண்டிருக்கவில்லை. நினைவக இடம், ஆய்வு ஆற்றல் பெரும்பான்மையும் அரிதாக இருக்கிற உட்படு கணிப்புச் சூழல்களிலும் (embedded computing situations) அவற்றைப் பயன்படுத்த இயலும். இறுதியாக ஒலிச்செறிவு அடிப்படை ஒழுங்குமுறைகள் வெளியீட்டுப் பேச்சின் எல்லா நோக்குகளுடனும் முழுக் கட்டுப்பாடு கொண்டிருப்பதன் காரணமாகக் கேள்விகளையும் கூற்றுகளையும் மட்டுமல்லாமல் பலவகை உணர்ச்சிகளையும் பேச்சின் சுரங்களையும் வெளிப்படுத்தும் மீக்கூறு அல்லது இசையோட்டத்தின் அதிக வகைகள் வெளியீடாக இருக்கவியலும்.



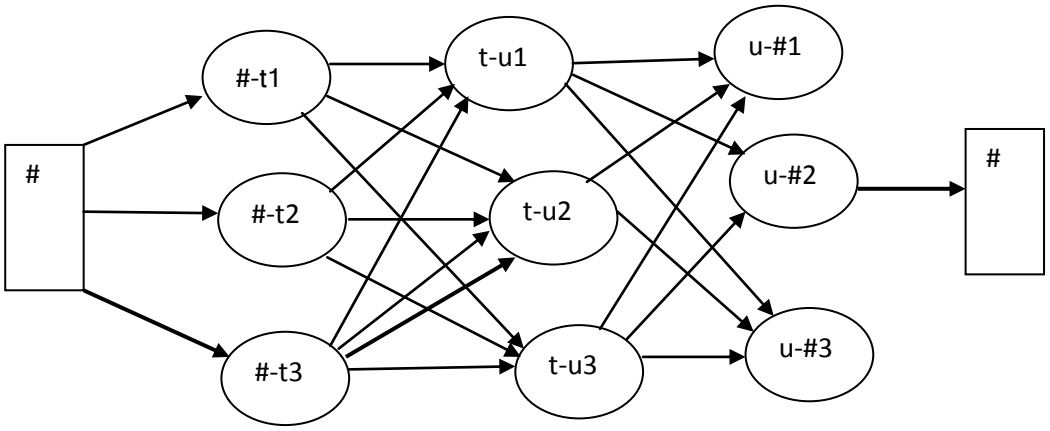
6.7.4.2.3 **ஒன்றிணைப்பு கூட்டிணைப்பாக்கம்**

ஒன்றிணைப்புப் பேச்சு கூட்டிணைப்பாக்கம் (concatenative synthesis) இயல்பான பேச்சிலிருந்து ஆக்கப்பட்ட வேறுபட்ட நீளமுள்ள ஏற்கனவே பதிவுசெய்யப்பட்ட பேச்சு அலகுகளைப் பயன்படுத்தும் செயல்பாடாகும். ஒன்றிணைப்பு உருவாக்கத்தில் முன்பதிவுசெய்யப்பட்ட பேச்சின் தரவுதளத்திலிருந்து உள்ள பேச்சுக் கூறுகள் ஒன்றிணைக்கப்படும். ஒன்றிணைப்பு பேச்சுக் கூட்டிணைப்பாக்க நுட்பங்கள் மூலம் பல வெற்றிகரமான ஒழுங்குமுறைகள் உருவாக்கப்பட்டுள்ளன. இருப்பினும் ஒன்றிணைப்பாக்கத்தில் பல சிக்கல் உள்ளன. அதிக அளபு இயற்கைத்தன்மையான ஒலிகளை உருவாக்கவேண்டி அவைகள் கவனத்தில் கொள்ளப்படவேண்டும். சில சிக்கல்கள் கீழே பட்டியலிடப்பட்டுள்ளன.

1. உருவாக்கப்பட்ட பேச்சில் இணைக்கப்படும் இடங்களில் தொடர்ச்சி இல்லாமை.
2. குறிப்பாக அசைகள், சொற்கள் போன்ற நீண்ட ஒன்றிணைப்பு அலகுகள் பயன்படுத்தப்படும்போது நினைவகத் தேவை பெரும்பான்மையும் அதிகமாகும்.
3. தரவைச் சேகரிப்பதற்கும் அடையாளப்படுத்துவதற்கும் பெரும்பான்மையும் காலம் அதிகமாகத் தேவைப்படும்.
4. வெளியீட்டுப் பேச்சு தேர்ந்தெடுக்கப்பட்ட தரவு அடித்தளத்தை அதிகமாகச் சார்ந்து அமையும்.

கோட்பாட்டு அடிப்படையில் எல்லா சாத்தியமான மாற்றொலிகளும் உட்படுத்தப்படவேண்டும்; இருப்பினும் தரத்திற்கும் மாதிரிகளின் எண்ணிக்கைகளுக்கும் இடையில் சலுகைகள் செய்துகொள்ளவேண்டும். ஒன்றிணைக்கப்படவேண்டிய பேச்சு அலகுகள் பதிவுசெய்யப்பட்ட ஒலியின் வேறுபட்ட இடங்களிலிருந்து பிரித்தெடுக்கப்படுவதால் ஒலியைப் பதிவு செய்யும் போது எப்போதும் இசைமை ஒன்றுபோல் இருக்க இயலாததாலும் வேறுபட்ட சொற்கள் வேறுபட்ட இசைமகளைக் கொண்டிருப்பதாலும் ஒன்றிணைக்கும் இடங்களில் தொடர்ச்சி இன்மை ஏற்படுகின்றது. இயல்பான ஒலியைப் பெறுவதற்காக இணைக்கப்படும் இடங்கள் அகற்றப்படவேண்டும்; இணைக்கப்படும் இரண்டு பேச்சு அலகுகளின் இசைமையைச் சமன்படுத்தி இதைச் செய்யவியலும்.

இந்தச் செயல்பாட்டில் நாம் தேவைப்படும் கூற்றுக்களை உருவாக்குவதற்காக முன்ஒலிப்பதிவு செய்யப்பட்ட தரவு அடித்தளத்திலிருந்து பேச்சின் கூறுகளைத் தெரிந்தெடுத்து ஒன்றோடு ஒன்று இணைக்கின்ற காரணத்தால் பின்வரும் படத்தில் காட்டப்பட்டுள்ள ஒன்றிணைப்பு உருவாக்கம் வெட்டி ஒட்டும் கூட்டிணைப்பாக்கம் (cut and paste synthesis) என்றும் அழைக்கப்படுகின்றது. இந்தச் செயல்பாட்டில் நாம் தேவைப்படும் கூற்றுக்களை உருவாக்குவதற்காக முன்ஒலிப்பதிவு செய்யப்பட்ட தரவு அடித்தளத்திலிருந்து பேச்சின் கூறுகளைத் தெரிந்தெடுத்து ஒன்றோடு ஒன்று இணைக்கின்றோம்.



ஒலிப்பதிவு செய்வதற்குப் பொதுவான தேர்வு ஒலியன்களும் ஒலியிருமைகளும் (diphones) ஆகும்; ஏனென்றால் அவைகள் போதுமான நெகிழ்வு பெறுவதற்கு வேண்டியும் நினைவகத் தேவைகளைப் போதுமான அளவுக்கு வைப்பதற்கு வேண்டியும் குறுகி உள்ளன. பல காரணங்களால் அசைகள் அல்லது சொற்கள் போன்ற நீண்ட அலகுகளைப் பயன்படுத்துவது சாத்தியமில்லாதது மற்றும் நடைமுறைப் படுத்த இயலாதது ஆகும். ஒன்றிணைப்புகளில் ஒலியிருமைகளைப் (diaphones) பயன்படுத்துவது ஒலிப்புகளைக் கையாள நல்ல சாத்தியங்களைத் தருகின்றது; ஏனென்றால் ஒலியிருமைகள் ஒரு ஒலியனிலிருந்து மற்றொரு ஒலியனுக்கு மாறுவதையும் மற்றும் முதல் ஒலியனின் பின் பகுதியையும் இரண்டாம் ஒலியனின் முன் பகுதியையும் கொண்டிருக்கின்றது. இதன் விளைவாக ஒன்றிணைப்பு இடம் ஒவ்வொரு ஒலியனின் மையத்திலும் இடங்காணப்படும் மற்றும் இது ஒலியனின் நிலையான இடமாக இருப்பதன் காரணமாக எல்லைகளில் ஏற்படும் வேறுபாட்டின் அளவு குறைவதை எதிர்பார்க்க இயலும். ஒரு தரவு அடித்தளத்தின் வேறுபட்ட ஒலியன்களின் எண்ணிக்கை கிட்டத்தட்ட 40-இலிருந்து 50-க்குள் இருக்கையில் பொருத்தமுள்ள ஒலியிருமைகளின் எண்ணிக்கை இயல்பாக 1500-இலிருந்து 2000 வரைக்கும் இருக்கும்; இருப்பினும் இந்த அளவு தரவு அடித்தளத்தைக் கொண்ட கூட்டிணைப்பாக்கி பொதுவாகச் செயல்முறைப்படுத்தத் தக்கதாகும்.

பொதுவாக ஒருங்கிணைப்பாக்கம் மிக இயற்கையான உருவாக்கப்பட்ட பேச்சை உருவாக்கும். இருப்பினும் பேச்சில் உள்ள இயற்கையான வேறுபாடுகளுக்கும் அலைவடிவுகளைக் கூறிடுவதற்கான தானியக்கமான நுட்பங்களின் இயல்புக்கும் இடையிலான

வேறுபாடுகள் வெளியீட்டில் கேட்கத்தக்க தடுமாற்றத்தில் விளையும். ஒன்றிணைப்பாகத்தின் மூன்று முக்கியத் துணை வகைகள் உள்ளன.

6.7.4.2.4 **அலகுத் தேர்வு கூட்டிணைப்பாக்கம்**

அலகுத் தேர்வு கூட்டிணைப்பாக்கம் (unit-selection synthesis) பதிவு செய்யப்பட்ட பேச்சின் பெரிய தரவு அடித்தளங்களைப் பயன்படுத்துகின்றது. தரவு அடித்தள உருவாக்கத்தின் போது ஒவ்வொரு பதிவுசெய்யப்பட்ட கூற்றும் பின்வருவனவற்றின் சிலவாக அல்லது எல்லாமாகக் கூறிடப்படும்: தனிப்பட்ட ஒலிகள், ஒலியிருமைகள், அரையொலிகள், அசைகள், உருபன்கள், தொடர்கள், மற்றும் வாக்கியங்கள். இயல்பாகக் கூறுகளாக்கும் பகுப்பு சிறப்பாக மாற்றப்பட்ட ஒரு பேச்சுப் புரிவான் குழுமத்தைப் பயன்படுத்தி அலைவடிவம், நிறமாலை நிழற்படம்/ஒலியியக்க வர்ணனைப்படம் போன்ற காட்சி உருப்படுத்தங்களைப் பயன்படுத்தி பின்னர் சில கையாலான திருத்தங்களுடன் ஒரு "கட்டாயப்படுத்தப்பட்ட வரிசையாக்க" முறைக்குச் செய்யப்படுகின்றது. பேச்சு தரவு அடித்தளத்தில் அலகுகளின் சொல்லடைவு, பகுப்பு மற்றும் அடிப்படை நிகழ்வெண் (இசைமை), காலம், அசையில் இருப்பிடம், அடுத்துவரும் ஒலிகள் போன்ற ஒலியியக்க அளகைகளின் அடிப்படையில் உருவாக்கப்படுகின்றது. வழிமுறைவரைவு இயக்கக் காலகட்டத்தில் (runtime) விரும்பப்பட்ட இலக்குக் கூற்று தரவு அடித்தளத்திலிருந்து தேர்வை நாடும் அலகுகளின் நற் தொடர்ச்சியியை உறுதிசெய்வதால் உருவாக்கப்படுகின்றது. இந்தச் செயல்பாடு சிறப்பாக மதிப்பீடுசெய்யப்பட்ட தீர்மானக் கிளையமைப்பைப் பயன்படுத்தி வழக்கமாகச் சாதித்துக்காட்டப்படுகின்றது.

அலகுத் தேர்வு பதிவுசெய்யப்பட்ட பேச்சுக்கு இலக்கவெண்ணியல்/டிஜிட்டல் குறிகை/சமிக்கைப் பகுப்பாய்வின் (digital signal processing) குறைந்த அளவையே பயன்படுத்துவதன் காரணமாக அதிக இயற்கைத்தன்மையை வழங்குகின்றது. இலக்கவெண்ணியல் குறிகைப் பகுப்பாய்வு பெரும்பான்மையும் பதிவுசெய்யப்பட்ட பேச்சை குறைந்த அளவு இயல்பானதாகச் செய்கின்றது; சில ஒழுங்குமுறைகள் அலைவடிவை நேர்த்திசெய்ய ஒன்றிணைக்கும் இடத்தில் குறிகைப் பகுப்பாய்வின் குறைந்த அளவையே பயன்படுத்துகின்றன. நல் அலகுத்தேர்வு ஒழுங்குமுறைகளின் (best unit selection systems) வெளியீடு முக்கியமாகப் பனுவலிலிருந்து பேச்சு ஒழுங்குமுறை ஒத்திசைவு செய்யப்பட்ட சூழல்களில் பெரும்பான்மையும் உண்மையான மனிதக் குரல்களிலிருந்து வேறுபடுத்த இயலாதனவாய் உள்ளன. இருப்பினும் அதிகப்படியான இயற்கைத்தன்மை இயல்பாக அலகுத்

தேர்வு பேச்சுத் தரவு அடித்தளத்தளங்களை மிகப் பெரிதாக வேண்டுகின்றது; சில ஒழுங்குமுறைகளில் பேசின் நீண்ட மணிநேரங்களை உருப்படுத்தம் செய்யும் கிகாபைட் அளவிலான பதிவு செய்யப்பட்ட தரவை வேண்டும். மேலும் தரவு அடித்தளத்தில் நற் தேர்வு சாத்தியம் இருந்தபோதிலும் குறைந்த பேச்சு கூட்டிணைப்பாக்கத்தை விளைவிக்கும் கூறுகளையே தேர்வுசெய்யும் வழக்கத்திற்கு அலகுத் தேர்வு வழிமுறை வரைவுகள் (unit selection algorithms) அறியப்படுவன ஆகும்.

6.7.4.2.5 **ஒலியிருமை கூட்டிணைப்பாக்கம்** (Diaphone synthesis)

ஒலியிருமைக் கூட்டிணைப்பாக்கம் (Diaphone synthesis) மொழியில் நேரும் எல்லா ஒலியிருமைகளையும் கொண்டிருக்கும் குறைந்த அளவிலான பேச்சுத் தரவு அடித்தளத்தைப் பயன்படுத்துகின்றது. ஒலியிருமைகளின் எண்ணிக்கை மொழியின் ஒலியன் வரன்முறையை (phonotactics) பொறுத்து அமையும். ஒலியிருமைக் கூட்டிணைப்பு உருவாக்கத்தில் ஒவ்வொரு ஒலியிருமைக்கும் ஒரேயொரு எடுத்துக்காட்டுதான் பேச்சுத் தரவு அடித்தளத்தில் இருக்கும். வழியமைப்பு இயக்கக் காலகட்டத்தில் ஒரு வாக்கியத்தின் இலக்கு மீக்கூறு இலக்கவெண்ணியல் குறிகை பகுப்பாய்வு நுட்பங்களின் மூலம் இந்தக் குறைந்த அலகுகள் மீது சுமத்தப்படும். விளைவிக்கப்படும் பேச்சு அலகுத்தேர்வு ஒழுங்குமுறைகளைக் காட்டிலும் பொதுவாக மோசமானதாக இருக்கும்; ஆனால் ஒலிச்செறிவு கூட்டிணைப்பாக்கிகளின் வெளியீட்டைக் காட்டிலும் அதிக இயல்பானதாக இருக்கும். ஒலியிருமை கூட்டிணைப்பாகம் ஒன்றிணைப்பு கூட்டிணைப்பாகத்தின் ஒலிசார் தடுமாற்றங்களாலும் ஒலிச்செறிவு கூட்டிணைப்பாக்கத்தின் செயற்கை ஒலிப்பு இயல்பாலும் குறையுற்றிருக்கும்; சிறிய அளவு என்பதைத் தவிர இவ்விரு அணுகுமுறைகளிலிருந்தும் குறைந்த அளவு அனுகூலங்களையே கொண்டுள்ளது. எனவே இதன் வணிகவயமான பயன்பாடுகள் குறைகின்றன; இருப்பினும் பல மென்பொருள் நடைமுறைப்படுத்தல்கள் இலவசமாகக் கிடைப்பதன் காரணமாக இது தொடர்ந்து ஆய்வில் பயன்படுத்தப்பட்டு வருகின்றது.

6.6.4.2.6 **பொருண்மைக்களச் சிறப்புக் கூட்டிணைப்பாக்கம்** (Domain-specific synthesis)

பொருண்மைக்களச் சிறப்புக் கூட்டிணைப்பாக்கம் (Domain-specific synthesis) முழு கூற்றையும் உருவாக்க முன் ஒலிப்பதிவு செய்யப்பட்ட சொற்களையும் தொடர்களையும் ஒன்றிணைக்கின்றது. இது பயணக் கால அட்டவணை அறிவிப்புகள் அல்லது வானிலை அறிக்கைகள் போன்ற ஒரு குறிப்பிட்ட பொருண்மைக் களத்தை எல்லைப்படுத்தும்

வெளியீடுகளின் பயன்பாடுகளில் பயன்படுத்தப்படுகின்றது. இத்தொழில் நுட்பம் நடைமுறை படுத்துவதற்கு எளியது மற்றும் பேசும் கடிகாரங்கள் கணிப்பான்கள் போன்ற கருவிகளில் நீண்டகால வணிகப் பயன்பாடு கொண்டுள்ளது. வாக்கிய வகைகள் எல்லைக்குட்பட்டு இருப்பதாலும் மற்றும் அவை மூல ஒலிப்பதிவுகளின் மீக்கூறு மற்றும் இசையோட்டத்துடன் நெருக்கமாகப் பொருந்துவதாலும் இவ்வொழுங்கமைப்புகளின் இயற்கைத்தன்மை நிலை மிகவும் உயர்ந்ததாய் இருக்க இயலும்.

இவ்வொழுங்கமைப்புகள் அவற்றின் தரவு அடித்தளங்களில் சொற்கள் மற்றும் தொடர்களால் எல்லைப்படுத்தப்பட்டுள்ளதாலும் அவை பொது நோக்கத்திற்கு உகந்ததல்ல மற்றும் அவை முன் வழியமைப்பு செய்யப்பட்ட சொற்கள் மற்றும் தொடர்களின் ஒன்றிணைப்புகளை மட்டுமே கூட்டிணைப்பு உருவாக்கம் செய்ய இயலும்.

6.7.4.2.**7 எச்.எம்.எம். அடிப்படையிலான கூட்டிணைப்பு உருவாக்கம்** (HMM based synthesis)

எச்.எம்.எம். அடிப்படையிலான கூட்டிணைப்பாக்க அணுகுமுறை (HMM based synthesis) ஹிட்டன் மார்க்கோவ் மாதிரிகள் அடிப்படையிலானது. இந்த ஒழுங்குமுறையில் நிகழ்வெண் நிறமாலை/அலைமாலை (பேச்சுக் குழாய்), அடிப்படை நிகழ்வெண் (பேசு மூலம்), பேச்சின் கால அளவு (மீக்கூறு) என்பன எச்.எம்.எம்.களால் ஒரே நேரத்தில் மாதிரிப்படுத்தப்பட்டுள்ளன. பேச்சு அலைகள் கூடுதல் சாத்திய அளவீடு (maximum likelihood criterion) அடிப்படையில் உருவாக்கப்படுகின்றது.

இவ்வொழுங்கமைப்பு இரண்டு நிலைகளைக் கொண்டுள்ளது: பயிற்சி நிலை, (training stage) கூட்டிணைப்பாக்கம் (சோதனை) (testing). பயிற்சி நிலையில் ஒலியன் எச்.எம்.எம்.கள் பேச்சு தரவு அடித்தளத்தைப் பயன்படுத்திப் பயிற்சி செய்யப்படுகின்றது. நிறமாலை/அலைமாலை (spectrum), அடிப்படை நிகழ்வெண் (F0) என்பன பல்லொழுக்கு எச்.எம்.எம்.களால் மாதிரிப்படுத்தப்பட்டுள்ளன; இதில் நிறமாலை/அலைமாலை மற்றும் F0 பகுதிகளுக்கு வெளியீட்டு வினியோகங்கள் (output distributions) முறையே தொடர்ச்சியான நிகழ்வுத்தகவு வினியோகம் (continuous probability distribution) மற்றும் பல் இடைவெளி நிகழ்வுத்தகவு விநியோகம் (multi-space probability distribution) இவற்றைப் பயன்படுத்தி மாதிரிப்படுத்தப்பட்டுள்ளன

நிறமாலை/அலைமாலை மற்றும் F0 இவற்றின் வேறுபாடுகளை மாதிரிப்படுத்த ஒலியியல், மீக்கூறு, ஒலியன் அடையாளக் காரணிகள், அழுத்தம் தொடர்பான காரணிகள், இடம்சார்

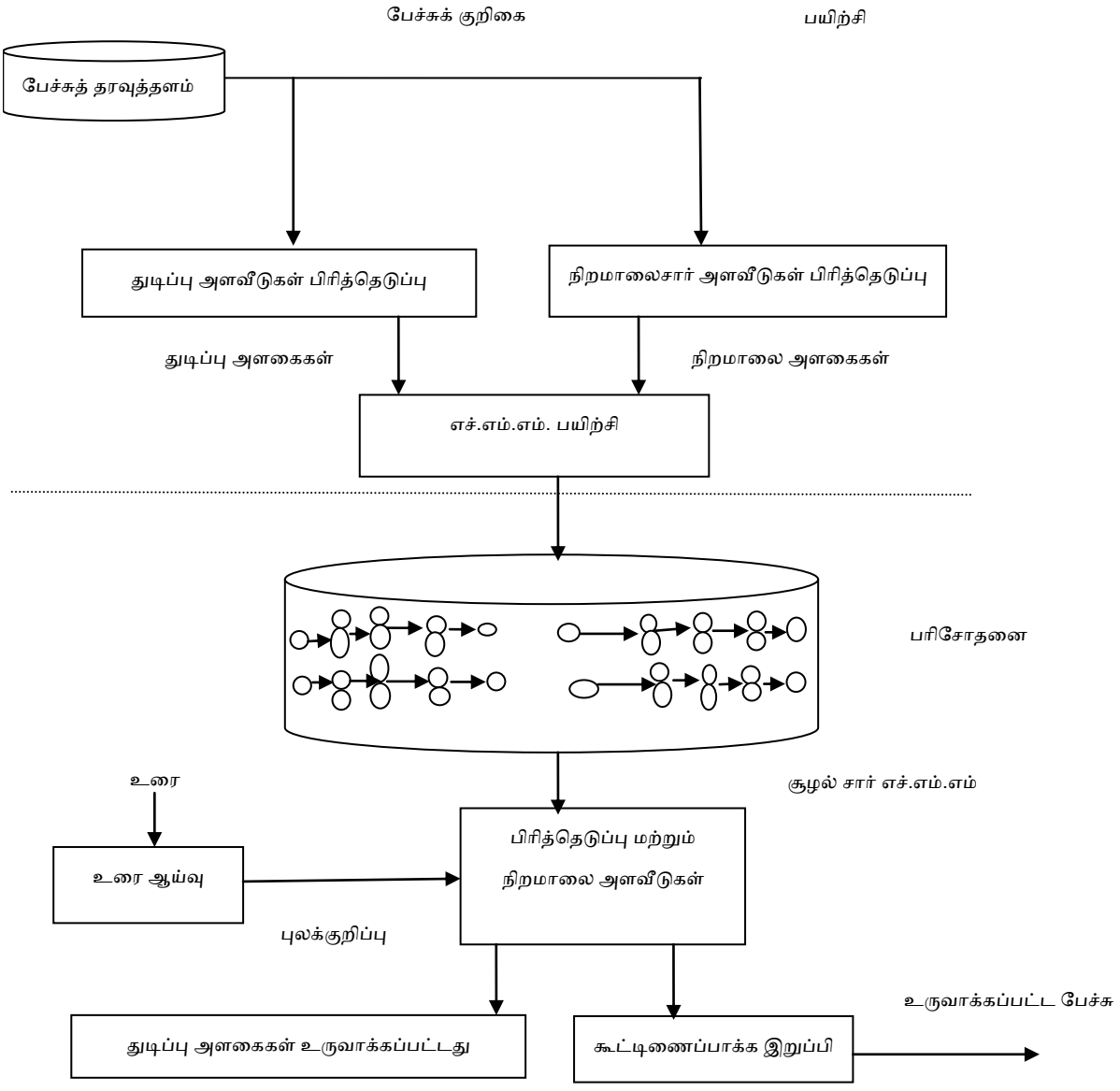
காரணிகள் போன்ற மொழிச் சூழல் காரணிகள் கருத்தில் கொள்ளப்படுகின்றன. பின்னர் தீர்மானக் கிளை அடிப்படையிலான சூழல் கொத்தாக்க நுட்பம் (decision tree based context clustering technique) சூழல் சார் ஒலியன் எச்.எச்.எம்.களின் நிறமாலை/அலைமாலை மற்றும் F0 பகுதிகளுக்குத் தனித்தனியாகப் பயன்படுத்தப்படுகின்றது. இறுதியாக நிலை கால அளவுகள் பல்பரிமாண காசியன் வினியோகங்களால் (multi-dimensional Gaussian distributions) மாதிரிப்படுத்தப்படுகின்றது மற்றும் நிலைக் கொத்தாக்க நுட்பம் (state clustering technique) கால அளவு மாதிரிகளுக்குப் பயன்படுத்தப்படுகின்றது. எச்.எம்.எம்.இன் கட்டு வரைபடம் கீழே தரப்பட்டுள்ளது.

கூட்டிணைப்பாக்க (பரிசோதனை) நிலையில் தரப்பட்ட ஒரு பனுவல் சூழல் சார் ஒலியன் புலக்குறிப்பு தொடர்ச்சியாக மாற்றப்படுகின்றது. புலக்குறிப்புத் தொடர்ச்சி அடிப்படையில் ஒரு வக்கியம் எச்.எம்.எம். சூழல்சார் ஒலியன் எச்.எச்.எம்.களை ஒன்றிணைப்பதால் உருவாக்கப்படுகின்றது. ஒலியன் கால அளவுகள் நிலை கால அளவு வினியோகங்களைப் பயன்படுத்திப் பெறப்படுகின்றது; பின்னர் நிறமாலை/அலைமாலை மற்றும் F0 அளகை தொடர்ச்சிகள் வாக்கிய எச்.எம்.எம்.-இலிருந்து எம்.எல். அளவீடு அடிப்படையில் பெறப்படுகின்றது. இறுதியாக எம்.எல்.எஸ்.எ.-ஐப் பயன்படுத்தி உருவாக்கப்பட்ட மெல்-செப்ட்ரல் (Mel-septral) மற்றும் F0 அளகை தொடர்ச்சிகள் இவற்றிலிருந்து இறுப்புப் பேச்சு (filter speech) கூட்டிணைப்பாக்கப்படுகின்றது.

இந்த அணுகுமுறையில் பேசும் பாணிகள்/நடைகள் இரண்டு வழிகளில் மாதிரிப்படுத்தப்பட்டுள்ளது: "நடை/பாணி சார் மாதிரியாக்கம்" அல்லது "நடை கலவை மாதிரியாக்கம்". நடைசார் மாதிரியாக்கத்தில் ஒவ்வொரு நடையும் தனித்தனியாக மாதிரிப்படுத்தப்படுகின்றன. மாறாக நடைக் கலவை மாதிரியாக்கத்தில் பேசும் பாணிகளும்/நடைகளும் உணர்ச்சி வெளியீடுகளும் சூழல் காரணிகளாகவும் ஒலியியல் மற்றும் மொழியியல் காரணிகளாக எடுத்துக்கொள்ளப்படுகின்றது; மற்றும் எல்லா நடைகளும் ஒரே நேரத்தில் ஒரு தனியான ஒலியியக்க மாதிரியால் மாதிரிப்படுத்தப்பட்டுள்ளன. இவ்விரு மாதிரியாக்க அணுகுமுறைகளுக்கும் ஒரே நிறைவேற்றத்தைக் கொண்டிருக்கும் மற்றும் பதிவு செய்யப்பட்ட பேச்சுடன் ஒற்றுமையுள்ள பேச்சைச் கூட்டிணைப்பாக்கம் செய்வது சாத்தியமாகும்.

நம்மிடம் வேறுபட்ட இரு நபர்களின் இரு பேச்சு நடை மாதிரிகள் இருந்தால் இரு நபர்களின் மாதிரிகளை இடைச் செருகுவதால் இரு பேசுபவர்களுக்கு இடையில் இடைப்பட்ட

குரல் சிறப்புப்பண்புகளைக் கொண்ட பேச்சைக் கூட்டிணைப்பாக்கம் செய்ய இயலும். இதற்கு நாம் கவனக்கணிப்புகளுக்கு (observation) இடைச்சொருகல் எனப்படும் எளிய இடைச்சொருகல் நெறிமுறையைப் (interpolation method) பயன்படுத்துகின்றோம்.



6.7.4.3. பேச்சு கூட்டிணைப்பாக்க நெறிமுறைகளின் நிறைகளும் குறைகளும்

கூட்டிணைப்பாக்க வகை	நிறை	குறை
ஒலிப்பு கூட்டிணைப்பாக்கம்	பேச்சின் பௌதிகப் பாவனை; ரபோட்டிக் பயன்பாடுகளுக்குப் பொருத்தமானது.	இயல்புத்தன்மை/ இயற்கைத் தன்மை குறைவு
ஒலிச்செறிவு கூட்டிணைப்பாக்கம்	வேறுபட்ட பேச்சு நடைகளுகளை உட்படுத்தும் நெகிழ்வுத்தன்மை.	அளகைகளின் கையிலான இசைவிப்பு
ஒன்றிணைப்பு கூட்டிணைப்பாக்கம்	நல்ல இயற்கைத் தன்மை.	குறைவான புரிந்துகொள்கை; அதிக நினைவகம் தேவைப் படுதல்.
எச்.எம்.எம். அடிப்படையிலான கூட்டிணைப்பாக்கம்	நல்ல பண்புள்ள புரியவியலும் பேச்சைப் பெற குறைந்த அளவிலான நினைவகத் தேவைகள்	அலகுத் தேர்வுடன் ஒப்பிடுகையில் இயற்கைத் தன்மையின் தரக்குறைவு.

6.7.4.4. **உரை-பேச்சு ஒழுங்குமுறைகளின் பயன்பாடுகள்**

உரையிலிருந்து பேச்சு ஒழுங்குமுறை இயல்பான மொழியை (உரையை) பேச்சாக மாற்றுகின்றது. தற்காலத்தில் உரையிலிருந்து பேச்சு ஒழுங்குமுறை பல நோக்கங்களுக்காகப் பயன்படுத்தப்படுகின்றது. உரையிலிருந்து பேச்சு ஒழுங்குமுறை சில பயன்பாடுகளுக்குப் பயன்படுத்தப்படுகின்றது. இதன் சில பயன்பாடுகள் கீழே பட்டியலிடப்பட்டுள்ளன.

1.இலக்கவெண்சார் நூல் படிப்பான் (digital reader)

2.கைபேசி செய்தி படிப்பான் (mobile news reader)

3.பேச்சு ஒழுங்குமுறை (speech system)

4. கேளிக்கை/பொழுதுபோக்குத் தொழில் (entertainment industry)

6.7.5. பேசுபவர் அறிதல் ஒழுங்குமுறை

குரல்களின் சிறப்பியல்புகளிலிருந்து ஒரு நபரை அடையாளம் காண்பது பேசுபவர்அறிதல். எனப்படும் (Poddar, Sahidullah, & Goutam 2018) "யார் பேசுகிறார்கள்?" என்ற கேள்விக்கு பதிலளிக்க இது பயன்படுகிறது. குரல் அறிதல் (voice recognition) (Pollack, Pickett, Sumby

1974); Van Lancker and Kreiman (1984) என்ற சொல் பேசுபவர் அறிதல் (speaker recognition) அல்லது பேச்சு அறிதலைக் (speech recognition) குறிக்கலாம். பேச்சாளர் சரிபார்ப்பு (பேசுபவர் உறுதிப்படுத்தல் என்றும் அழைக்கப்படுகிறது) அடையாளத்துடன் முரண்படுகிறது, மேலும் பேசுபவர் அறிதல் ஸ்பீக்கர் டயரிஸேஷனிலிருந்து (speaker diarisation) வேறுபடுகிறது (அதே பேசுபவர் பேசும்போது அறிகிறது).

பேசுபவரை அறிதல் குறிப்பிட்ட குரல்களில் பயிற்சியளிக்கப்பட்ட அமைப்புகளில் பேச்சை மொழிபெயர்க்கும் பணியை எளிதாக்கலாம் அல்லது பாதுகாப்புச் செயல்பாட்டின் ஒரு பகுதியாக பேசுபவரின் அடையாளத்தை அறிய அல்லது சரிபார்க்க இது பயன்படுத்தப்படலாம். பேசுபவர் அறிதல் 2019ஆம் ஆண்டின் நான்கு தசாப்தங்களுக்கு முந்தைய வரலாற்றைக் கொண்டுள்ளது மற்றும் தனிநபர்களிடையே வேறுபடுவதாகக் கண்டறியப்பட்ட பேச்சின் ஒலியியக்கப் பண்புக்கூறுகளைப் பயன்படுத்துகிறது. இந்த ஒலி வடிவங்கள் உடற்கூறியல் மற்றும் கற்றறிந்த நடத்தை முறைகள் இரண்டையும் பிரதிபலிக்கின்றன.

சரிபார்ப்பு மற்றும் அடையாளம் காணல்

பேசுபவர் அறிதல் தொழில்நுட்பங்கள் மற்றும் முறைகளின் இரண்டு முக்கிய பயன்பாடுகள் உள்ளன. பேசுபவர் ஒரு குறிப்பிட்ட அடையாளத்தைக் கொண்டிருப்பதாகக் கூறி, இந்தக் கோரிக்கையைச் சரிபார்க்க குரல் பயன்படுத்தப்பட்டால், இது சரிபார்ப்பு அல்லது அறிதல் என அழைக்கப்படுகிறது. மறுபுறம், அடையாளம் என்பது அறியப்படாத பேசுபவரின் அடையாளத்தை தீர்மானிக்கும் பணியாகும். ஒரு விதத்தில், பேசுபவர் சரிபார்ப்பு என்பது 1: 1 பொருத்தமாகும், அங்கு ஒரு பேசுபவரின் குரல் ஒரு குறிப்பிட்ட டெம்ப்ளேட்டுடன் பொருந்துகிறது, அதே நேரத்தில் ஸ்பீக்கர் அடையாளம் 1: N பொருத்தமாகும், அங்கு குரல் பல வார்ப்புருக்களுடன் ஒப்பிடப்படுகிறது.

பாதுகாப்பு கண்ணோட்டத்தில், அடையாளம் சரிபார்ப்பிலிருந்து வேறுபட்டது. பாதுகாப்பான அமைப்புக்கான அணுகலை வழங்குவதற்காக பேசுபவர் சரிபார்ப்பு வழக்கமாக "வாயிற்காப்போன்"-ஆகப் ("gatekeeper") பயன்படுத்தப்படுகிறது. இந்த அமைப்புகள் பயனர்களின் அறிவுடன் இயங்குகின்றன மற்றும் பொதுவாக அவர்களின் ஒத்துழைப்பு தேவைப்படுகிறது. ஒரு கலந்துரையாடலில் பேசுபவர்களை அடையாளம் காணவும், பேசுபவர் மாற்றங்களின் தானியங்கி அமைப்புகளை எச்சரிக்கவும், ஒரு பயனர் ஏற்கனவே ஒரு கணினியில்

சேர்க்கப்பட்டுள்ளாரா என்பதை சரிபார்க்கவும் பயனரின் அறிவு இல்லாமல் பேசுபவர் அடையாள அமைப்புகளை இரகசியமாக செயல்படுத்த முடியும்.

தடயவியல் பயன்பாடுகளில் (forensic applications) முதலில் "சிறந்த பொருத்தங்களின்" பட்டியலை உருவாக்க பேசுபவர் அடையாளச் செயல்முறையைச் (speaker identification process) செய்வது பொதுவானது, பின்னர் ஒரு உறுதியான பொருத்தத்தைத் தீர்மானிக்க தொடர்ச்சியான சரிபார்ப்பு செயல்முறைகளைச் செய்வது.

பயிற்சி

வணிகமயமாக்க ஆரம்பகால பயிற்சி தொழில்நுட்பங்களில் ஒன்று வேர்ல்ட்ஸ் ஆஃப் வொண்டரின் 1987 ஜூலி பொம்மையில் செயல்படுத்தப்பட்டது. அந்த நேரத்தில், பேச்சாளர் சுதந்திரம் ஒரு நோக்கம் கொண்ட முன்னேற்றமாகும், மேலும் அமைப்புகளுக்கு ஒரு பயிற்சி காலம் தேவைப்பட்டது. 1987 ஆம் ஆண்டு பொம்மைக்கான விளம்பரம் "இறுதியாக, உங்களைப் புரிந்துகொள்ளும் பொம்மை" என்ற கோஷத்தைக் கொண்டிருந்தது. - இது ஒரு தயாரிப்பு என்று விவரிக்கப்பட்ட போதிலும், "குழந்தைகள் தங்கள் குரலுக்கு பதிலளிக்க பயிற்சி அளிக்க முடியும்." (Melanie Pinola 2011). குரல் அறிதல் என்ற சொல், ஒரு தசாப்தத்திற்குப் பிறகும், பேசுபவர் சுதந்திரத்தைக் குறிக்கிறது.

பேசுபவர் அறிதலின் மாறுபாடுகள்

ஒவ்வொரு பேசுபவர் அறிதல் முறைக்கும் இரண்டு கட்டங்கள் உள்ளன: பதிவு மற்றும் சரிபார்ப்பு. சேர்க்கையின் போது, பேச்சாளரின் குரல் பதிவு செய்யப்படுகிறது மற்றும் பொதுவாக குரல் அச்சு, வார்ப்புரு அல்லது மாதிரியை உருவாக்க பல அம்சங்கள் பிரித்தெடுக்கப்படுகின்றன. சரிபார்ப்பு கட்டத்தில், முன்பு உருவாக்கிய குரல் அச்சுக்கு எதிராக ஒரு பேச்சு மாதிரி அல்லது "உச்சரிப்பு" ஒப்பிடப்படுகிறது. அடையாள அமைப்புகளுக்கு, சிறந்த பொருத்தத்தை (கள்) தீர்மானிக்க பல குரல் அச்சிட்டுகளுக்கு எதிராக உச்சரிப்பு ஒப்பிடப்படுகிறது, அதே நேரத்தில் சரிபார்ப்பு அமைப்புகள் ஒரு குரல் அச்சுக்கு எதிராக ஒரு உரையை ஒப்பிடுகின்றன. சம்பந்தப்பட்ட செயல்முறை காரணமாக, சரிபார்ப்பு அடையாளத்தை விட வேகமாக உள்ளது.

பேச்சாளர் அங்கீகார அமைப்புகள் உரை சார்ந்த மற்றும் உரை-சுயாதீனமான இரண்டு வகைகளாகும்.

உரை சார்ந்தவை:

பதிவு மற்றும் சரிபார்ப்புக்கு உரை ஒரே மாதிரியாக இருக்க வேண்டும் என்றால் இது உரை சார்ந்த அங்கீகாரம் என்று அழைக்கப்படுகிறது. [10] உரை சார்ந்த அமைப்பில், அனைத்து பேச்சாளர்களிடமும் (எ.கா. பொதுவான பாஸ் சொற்றொடர்) அல்லது தனித்துவமானது. கூடுதலாக, பல காரணி அங்கீகார காட்சியை உருவாக்க பகிர்வு-ரகசியங்கள் (எ.கா.: கடவுச்சொற்கள் மற்றும் பின்ஸ்) அல்லது அறிவு சார்ந்த தகவல்களைப் பயன்படுத்தலாம்.

உரை-சுதந்திரம்:

உரை-சுதந்திர அமைப்புகள் பெரும்பாலும் பேசுபவரை அடையாளம் காண பயன்படுத்தப்படுகின்றன, ஏனெனில் அவை பேசுபவரின் எந்த ஒத்துழைப்பும் மிகக் குறைவாகவே தேவைப்படுகின்றன. இந்த வழக்கில் பதிவு மற்றும் சோதனையின் போது உரை வேறுபட்டது. உண்மையில், பல தடயவியல் பயன்பாடுகளைப் போலவே, பயனரின் அறிவு இல்லாமல் பதிவு செய்யப்படலாம். உரை-சுயாதீன தொழில்நுட்பங்கள் பதிவு மற்றும் சரிபார்ப்பில் கூறப்பட்டதை ஒப்பிடாததால், சரிபார்ப்பு பயன்பாடுகள் அங்கீகாரத்தின் கட்டத்தில் பயனர் என்ன சொல்கிறார் என்பதைத் தீர்மானிக்க பேச்சு அங்கீகாரத்தையும் பயன்படுத்துகின்றன.

உரை சுதந்திஅ அமைப்புகளில் ஒலியியல் மற்றும் பேச்சு பகுப்பாய்வு நுட்பங்கள் இரண்டும் பயன்படுத்தப்படுகின்றன (Lisa Myers (2004).

தொழில்நுட்பம் (Technology)

பேசுபவர் அறிதல் ஒரு அமைப்பொழுங்கு அறிதல் சிக்கல் ஆகும். குரல் அச்சிட்டுகளை (voice prints) செயலாக்கம் செய்யவும் சேமிக்கவும் பயன்படுத்தப்படும் பல்வேறு தொழில்நுட்பங்கள் அதிர்வெண் மதிப்பீடு (frequency estimation), மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகள் (hidden Markov models), காஸியன் கலவை மாதிரிகள் (Gaussian mixture models), அமைப்பொழுங்கு பொருந்தும் வழிமுறைகள் (pattern matching algorithms), நரம்பியல் வலையமைப்புகள் (neural networks), மேட்ரிக்ஸ் உருப்படுத்தம் (matrix representation), திசையன் அளவு (vector quantization) மற்றும் தீர்மானக் கிளைகள் (decision trees) ஆகியவற்றை உட்படுத்தும். குரல் அச்சிட்டுகளுக்கு எதிரான கூற்றுகளை ஒப்பிடுவதற்கு, கொசைன் ஒற்றுமை (cosine similarity) போன்ற அடிப்படை முறைகள் பாரம்பரியமாக அவற்றின் எளிமை மற்றும் செயல்திறனுக்காகப் பயன்படுத்தப்படுகின்றன. சில அமைப்புகள் கோஹார்ட் மாதிரிகள் மற்றும் உலக மாதிரிகள் போன்ற "எதிர்ப்பு பேசுபவர்" ("anti-speaker") நுட்பங்களையும்

பயன்படுத்துகின்றன. பேசுபவர் சிறப்பியல்புகளை பிரதிநிதித்துவப்படுத்துவதில் ஸ்பெக்ட்ரல் பண்புக்கூறுகள் முக்கியமாக பயன்படுத்தப்படுகின்றன (Sahidullah, Kinnunen 2016). நேரியல் முன்கணிப்பு குறியீட்டு முறை (Linear predictive coding (LPC/எல்பிசி)) என்பது பேசுபவர் அறிதல் மற்றும் பேச்சு சரிபார்ப்பில் பயன்படுத்தப்படும் பேச்சு குறியீட்டு முறையாகும் (speech coding method) (Gupta 2016).

சுற்றுப்புற இரைச்சல் அளவுகள் (Ambient noise levels) ஆரம்ப மற்றும் அடுத்தடுத்த குரல் மாதிரிகளின்/பதக்கூறுகளின் சேகரிப்பைத் தடுக்கலாம். துல்லியத்தை மேம்படுத்த சத்தம் குறைப்பு வழிமுறைகளைப் பயன்படுத்தலாம், ஆனால் தவறான பயன்பாடு எதிர் விளைவை ஏற்படுத்தும். குரலின் நடத்தை பண்புகளில் ஏற்பட்ட மாற்றங்கள் மற்றும் ஒரு தொலைபேசியைப் பயன்படுத்தி பதிவுசெய்தல் மற்றும் மற்றொரு தொலைபேசியில் சரிபார்ப்பு ஆகியவற்றால் செயல்திறன் சீரழிவு ஏற்படலாம். இரண்டு காரணி உறுதிப்படுத்தல் தயாரிப்புகளுடன் ஒருங்கிணைப்பு அதிகரிக்கும் என்று எதிர்பார்க்கப்படுகிறது. வயதானதால் ஏற்படும் குரல் மாற்றங்கள் காலப்போக்கில் கணினி செயல்திறனைப் பாதிக்கலாம். சில வெற்றிகரமான சரிபார்ப்பிற்குப் பிறகு சில அமைப்புகள் குரலில் இத்தகைய நீண்டகால மாற்றங்களைக் கைப்பற்றுகின்றன, இருப்பினும் தானியங்கி தழுவலால் விதிக்கப்படும் ஒட்டுமொத்த பாதுகாப்பு பாதிப்பு குறித்து விவாதம் உள்ளது.

சட்டரீதியான தாக்கங்கள்

ஐரோப்பிய ஒன்றியத்தில் பொது தரவு பாதுகாப்பு ஒழுங்குமுறை (General Data Protection Regulation) மற்றும் அமெரிக்காவில் கலிபோர்னியா நுகர்வோர் தனியுரிமை சட்டம் (California Consumer Privacy Act) போன்ற சட்டங்கள் அறிமுகப்படுத்தப்பட்டதன் காரணமாக, பணியிடத்தில் பேசுபவர் அறித்லைப் பயன்படுத்துவது குறித்து அதிக விவாதம் நடந்துள்ளது. செப்டம்பர் 2019 இல் ஐரிஷ் பேச்சு அங்கீகார டெவலப்பர் சோப் பாக்ஸ் லேப்ஸ் (Soapbox Labs) சம்பந்தப்பட்ட சட்டரீதியான தாக்கங்கள் குறித்து எச்சரித்தன.

பயன்பாடுகள்

முதல் சர்வதேச காப்புரிமை (first international patent) 1983ஆம் ஆண்டில் தாக்கல் செய்யப்பட்டது; தொலைத்தொடர்பு ஆய்வு மையம் மற்றும் ஆய்வகங்களில் (சி.எஸ்.இ.எல்.டி) (Centro Studi e Laboratori Telecomunicazioni (CSELT/சிஎஸ்இஎல்டி) (இத்தாலி) தொலைத்தொடர்பு ஆராய்ச்சியிலிருந்து மைக்கேல் கவாஸ்ஸா (Michele Cavazza) மற்றும் ஆல்பர்டோ சியாரமெல்லா (Alberto Ciaramella) ஆகியோரால் எதிர்கால டெல்கோ

சேவைகளுக்கான இறுதி வாடிக்கையாளர்களுக்கு மற்றும் வலையமைப்புகளைக் கடந்து சத்தம்-குறைப்பு நுட்பங்களை (noise-reduction techniques) மேம்படுத்துவதற்கான அடிப்படையாக வந்தது.

1996 மற்றும் 1998க்கு இடையில், ஸ்கோபி-கொரோனாச் பார்டர் கிராசிங்கில் (Scobey–Coronach Border Crossing) பேசுபவர் அங்கீகார தொழில்நுட்பம் பயன்படுத்தப்பட்டது, பதிவுசெய்யப்பட்ட உள்ளூர்வாசிகளுக்கு கனடா-அமெரிக்காவின் (Canada–United States) எல்லையைத் தாண்டி அறிவிக்க ஒன்றுமில்லாமல், ஆய்வு நிலையங்கள் இரவு முழுவதும் மூடப்பட்டபோது Meyer, Barb (June 12, 1996). இந்த அமைப்பு யு.எஸ். குடிவரவு மற்றும் இயற்கைமயமாக்கல் சேவைக்காக (U.S. Immigration and Naturalization Service) மிச்சிகனில் உள்ள வாரனின் குரல் (Voice Strategies of Warren, Michigan) உத்திகளால் உருவாக்கப்பட்டது.

சாதாரண உரையாடலின் 30 விநாடிகளுக்குள் தொலைபேசி வாடிக்கையாளர்களின் அடையாளத்தை சரிபார்க்க பார்க்லேஸ் செல்வம் (Barclays Wealth) செயலற்ற பேசுபவர் அறிதலைப் பயன்படுத்துவதாக மே 2013இல் அறிவிக்கப்பட்டது. ஆப்பிள் நிறுவனத்தின் சிரி தொழில்நுட்பத்தின் பின்னால் உள்ள நிறுவனமான குரல் அங்கீகார நிறுவனமான நுவான்ஸ் (2011 இல் சி.எஸ்.இ.எல்.டி. கணினிக்கு அழைப்பாளர்களை அடையாளம் காண சரிபார்க்கப்பட்ட குரல்வழி பயன்படுத்தப்பட வேண்டும், மேலும் எதிர்காலத்தில் கணினி முழுவதும் நிறுவனம் உருவாக்கப்படும்.

வாடிக்கையாளர்களை அவர்களின் அழைப்பு மையங்களுக்கு அறிதலுக்கான முதன்மை வழிமுறையாக குரல் பயோமெட்ரிக்ஸை வரிசைப்படுத்திய முதல் நிதி சேவை நிறுவனம் பார்க்லேஸின் (Barclays) தனியார் வங்கி பிரிவு ஆகும். 93% வாடிக்கையாளர் பயனர்கள் இந்த அமைப்பை வேகம், பயன்பாட்டின் எளிமை மற்றும் பாதுகாப்புக்காக "10 இல் 9" என மதிப்பிட்டனர்(Matt Warman May 8, 2013) .

ஜேம்ஸ் ஃபோலி மற்றும் ஸ்டீவன் சோட்லோஃப் ஆகியோரின் 2014 மரணதண்டனை போன்ற குற்றவியல் விசாரணைகளிலும் பேசுபவர் அறிதல் பயன்படுத்தப்படலாம் (Ewen MacAskill).

பிப்ரவரி 2016இல், இங்கிலாந்தின் ஹை-ஸ்ட்ரீட் வங்கி எச்எஸ்பிசி (UK high-street bank HSBC) மற்றும் அதன் இணைய அடிப்படையிலான சில்லறை வங்கி ஃபர்ஸ்ட் டைரக்ட் 15 மில்லியன் வாடிக்கையாளர்களுக்கு கைரேகை அல்லது குரலைப் பயன்படுத்தி ஆன்லைன் மற்றும்

தொலைபேசி கணக்குகளை அணுக அதன் பயோமெட்ரிக் வங்கி மென்பொருளை வழங்குவதாக அறிவித்தது (Julia Kollewe 2016).

6.7.6. பேச்சுக் குறையைச் சரிசெயதல்

எந்த வகையான பேச்சு மற்றும் மொழி கோளாறுகள் குழந்தைகளை பாதிக்கின்றன?

பேச்சு மற்றும் மொழி கோளாறுகள் குழந்தைகள் பேசும், புரிந்துகொள்ளும், பகுப்பாய்வு செய்யும் அல்லது செயலாக்கும் முறையை பாதிக்கும். பேச்சுக் கோளாறுகள் குழந்தையின் பேசும் சொற்களின் தெளிவு, குரல் தரம் மற்றும் சரளமாக அடங்கும். மொழி கோளாறுகள் ஒரு குழந்தையின் அர்த்தமுள்ள உரையாடல்களை நடத்துதல், மற்றவர்களைப் புரிந்துகொள்வது, சிக்கலைத் தீர்ப்பது, படிப்பது மற்றும் புரிந்துகொள்வது மற்றும் பேசும் அல்லது எழுதப்பட்ட சொற்களின் மூலம் எண்ணங்களை வெளிப்படுத்தும் திறன் ஆகியவை அடங்கும்.

பள்ளிகளில் பேச்சு மற்றும் மொழி கோளாறுகளுக்கு எத்தனை குழந்தைகள் சிகிச்சை பெறுகிறார்கள்?

பேச்சு, மொழி மற்றும் செவித்திறன் கோளாறுகள் கற்றலை எவ்வாறு பாதிக்கின்றன?

தகவல்தொடர்பு திறன்கள் வாழ்க்கையின் அனுபவத்தின் மையத்தில் உள்ளன, குறிப்பாக அறிவாற்றல் வளர்ச்சி மற்றும் கற்றலுக்கு முக்கியமான மொழியை வளர்க்கும் குழந்தைகளுக்கு. படித்தல், எழுதுதல், சைகை செய்தல், கேட்பது மற்றும் பேசுவது அனைத்தும் மொழியின் வடிவங்கள் - கருத்துக்களைத் தொடர்புகொள்வதற்கு நாம் பயன்படுத்தக் கற்றுக்கொள்ளும் குறியீடு. தகவல்தொடர்பு செயல்முறை மூலம் கற்றல் நடைபெறுகிறது. ஒரு மாணவர் பள்ளியில் வெற்றிபெற கல்வி அமைப்பில் சகாக்கள் மற்றும் பெரியவர்களுடன் செயலில் மற்றும் ஊடாடும் தகவல்தொடர்புகளில் பங்கேற்கும் திறன் அவசியம்.

பேச்சு மற்றும் மொழித் திறன் கல்வியறிவுக்கு ஏன் மிகவும் முக்கியமானவை?

பேசும் மொழி வாசிப்பு மற்றும் எழுத்தின் வளர்ச்சிக்கான அடித்தளத்தை வழங்குகிறது. பேசும் மற்றும் எழுதப்பட்ட மொழி ஒரு பரஸ்பர உறவைக் கொண்டுள்ளன - ஒவ்வொன்றும் பொதுவான மொழி மற்றும் கல்வியறிவுத் திறனை விளைவிக்கும் வகையில் மற்றொன்றை உருவாக்குகின்றன, ஆரம்பத்தில் தொடங்கி குழந்தை பருவத்திலிருந்தே முதிர்வயது வரை தொடர்கின்றன.

தகவல்தொடர்பு கோளாறு பள்ளி செயல்திறனை பாதிக்கும் அறிகுறிகள் யாவை?

தகவல்தொடர்பு கோளாறுகள் உள்ள குழந்தைகள் அடிக்கடி மோசமான அல்லது போதுமான கல்வி மட்டத்தில் செயல்படுகிறார்கள், வாசிப்பதில் சிரமப்படுகிறார்கள், மொழியைப் புரிந்துகொள்வதிலும் வெளிப்படுத்துவதிலும் சிரமப்படுகிறார்கள், சமூகக் குறிப்புகளை தவறாகப் புரிந்துகொள்கிறார்கள், பள்ளியில் சேருவதைத் தவிர்க்கிறார்கள், மோசமான தீர்ப்பைக் காட்டுகிறார்கள், சோதனைகளில் சிரமப்படுகிறார்கள்.

கேட்க, பேச, படிக்க, அல்லது எழுதக் கற்றுக்கொள்வதில் சிரமம் மொழி வளர்ச்சியில் உள்ள சிக்கல்களால் ஏற்படலாம். ஒலி, எழுத்து, சொல், வாக்கியம் மற்றும் சொற்பொழிவு மட்டங்களில் மொழியின் உற்பத்தி, புரிதல் மற்றும் விழிப்புணர்வு ஆகியவற்றில் சிக்கல்கள் ஏற்படலாம். வாசிப்பு மற்றும் எழுதுதல் சிக்கல்களைக் கொண்ட நபர்கள் தொடர்புகொள்வதற்கும், சிந்திப்பதற்கும், கற்றுக்கொள்வதற்கும் மொழியை மூலோபாய ரீதியாகப் பயன்படுத்துவதில் சிரமங்களை சந்திக்க நேரிடும்.

பேசுக் குறைபாட்டை உள்ளடக்கும் தரவுத்தொகுதியின் முக்கியத்துவம்

மேற்கண்ட பேச்சுக்குறைபாடுகளை நீக்குவதற்கு வேண்டி பேச்சுக்குறைபாடு நீக்க ஒழுங்குமுறைகள் உருவாக்கப்படவேண்டும். இதற்கு பேசுக் குறைபாட்டை உள்ளடக்கும் பேச்சுத் தரவுத்தொகுதிகள் தேவை. இத்தகைய தரவுத்தொகுதிகளைக் கொண்டு பேச்சு ஆய்வு செய்து குறைபாட்டிற்கான அம்சங்களையும் காரணங்களையும் புரிந்துகொண்டு இதற்கான மென்பொருள் உருவாக்க இயலும். இத்தகைய முயற்சிகள் தொடக்க நிலையில் உள்ளன.

6.8. சுருக்கவுரை

இவ்வியலின் தொடக்கத்தில் ஒரு சுருக்கமான அறிமுகம் தரப்பட்டுள்ளது. இதைத் தொடர்ந்து மொழித் தொழில்நுட்பத்தில் தரவுத்தொகுதியின் மதிப்பு, அறிவு வளமாகத் தரவுத்தொகுதி, மொழி தொழில்நுட்ப கருவிகளை வடிவமைப்பதில் தரவுத்தொகுதி, மொழிபெயர்ப்பதற்கான ஆதரவு ஒழுங்குமுறைகளாகத் தரவுத்தொகுதி, மனித-இயந்திர ஊடாடும் ஒழுங்குமுறைகளுக்கான வளமாகத் தரவுத்தொகுதி மற்றும் பேச்சுத் தரவுத்தொகுதி தொழில்நுட்பம் ஆகியன தெளிவாக விளக்கப்பட்டுள்ளன.

இயல் 7

முக்கிய மொழியியலில் தரவுத்தொகுதியின் பயன்பாடு

7.1 அறிமுகம்

தரவுத்தொகுதியானது முக்கிய மொழியியல் ஆய்விலும் பகுப்பாய்விலும் பயன்பாட்டிலும் முதன்மையான மூலவளமாக மாறியுள்ளது. இது மொழியியலில் ஏற்பட்டுள்ள அறிவியல் பார்வையையும் தொழில் நுட்ப பார்வையும் பிரதிபலிக்கிறது. கருத்தியல் மாற்றம் உள்ளுணர்வு அடிப்படையிலான பகுத்தறிவு அனுமானங்களிலிருந்து அனுபவ மொழி ஆய்வை நோக்கி ஒரு முக்கிய திருப்பத்தை ஏற்படுத்துகிறது; தொழில்நுட்ப மாற்றம் அதிக அளவிலான சேமிப்பு வசதிகளையும் கணினிகளால் நிகழ்த்தவியலும் செயலாக்க சக்தியையும் வழங்குகிறது.

ஏறக்குறைய ஓரிரு தசாப்தத்திற்கு முன்னர் சில பெரிய ஆங்கில அகராதிகள் கணினி வழி பெரிய தரவுத்தொகுதியின் விரிவான பகுப்பாய்வு மூலம் உருவாக்கப்பட்டன. இந்த அகராதிகளில் பதிவுச் சொற்களின் பொருள் மற்றும் பயன்பாடு பற்றிய தகவல்களும், சொல்சார் துணைவகைப்பாடு, சூழல் பயன்பாட்டு மாறுபாடு, பல்பொருள் ஒருமொழிய உணர்பொருள் (polysemous connotation) போன்றவற்றின் விரிவான தகவல்களும், தரவுத்தொகுதியிலிருந்து மீட்டெடுக்கப்பட்டன; நிலைபேறான அகராதியில் காணப்படும் அவற்றின் சொற்பிறப்பியல், ஒலியனியல் மற்றும் உருபனியல் தகவல்களும் பிரித்தெடுக்கப்பட்டன.

டிஜிட்டல் தரவுத்தொகுதி, அகராதி தயாரிப்பில் இன்றியமையாததாகிறது; ஏனெனில் இது உள்ளுணர்விலிருந்து பெறமுடியாத நம்பகமான மற்றும் பன்முகப்படுத்தப்பட்ட தகவல்களை வழங்குகிறது. இதன் தாக்கம் மிகவும் சுவாரஸ்யமாக இருக்கிறது. இப்போது தாவுத்தொகுதியைக் குறிப்பிடாமல் அகராதிகளை உருவாக்க யாரும் நினைப்பதிலை.

தரவுத்தொகுதியைக் குறிப்பிடாமல் மொழியை விவரிக்கும் யோசனை சிலருக்கு அறிவியலற்றதாகத் தோன்றுகிறது. கடந்த சில தசாப்தங்களில் கணினி தொழில்நுட்பத்தின் முன்னேற்றம் காரணமாக இந்த அணுகுமுறையில் மாற்றம் ஏற்பட்டுள்ளது. மொழி ஆராய்ச்சியில் கணினியைப் பயன்படுத்துவது முதன்மை மொழியியலில் மொழியைப் பயன்படுத்துவதற்கான பல புதிய சாத்தியங்களைத் திறக்கிறது.

மேலும், எந்தவொரு உள்ளுணர்வு அணுகுமுறையையும் விட மொழி ஆய்வுக்கான அனுபவ அணுகுமுறை மிகவும் நம்பகமானது மற்றும் உண்மையானது என்ற உண்மையை உணர்ந்ததன் காரணமாக பகுத்தறிவுவாதத்திலிருந்து அனுபவவாதத்திற்கு கருத்தியல் மாற்றம்

ஏற்பட்டுள்ளது. 1960களின் முற்பகுதியில், தரவுத்தொகுதி மொழியியலின் முன்னோடிகள் 'பழமையான' தொழில்நுட்பத்துடன் மனநிறைவுகொண்ண்டவராய் இருந்தனர். மேலும், அவர்கள் சாம்ஸ்கியன் விமர்சனத்தின் பெரும் அலைகளை எதிர்க்க வேண்டியிருந்தது; ஏனெனில் அவர்களின் அனுபவ கண்டுபிடிப்புகள் (empirical findings) ஆக்கமுறை இலக்கணக் கோட்பாட்டுன் முரண்பட்டன.

ஒரு சில தசாப்தங்களுக்குள் இந்த காட்சி பெரிதும் மாறியது; ஏனென்றால் தரவுத்தொகுதியின் பகுப்பாய்வு மனித அறிவின் அறிவாற்றல் களத்தில் மனித மூளை எவ்வாறு சிந்திக்கிறது மற்றும் மொழியியல் சமிக்ஞைகளைப் பெறுகிறது என்பதையும், இந்த சமிக்ஞைகள் மூளையில் எவ்வாறு செயலாக்கப்படுகின்றன என்பதையும் புரிந்துகொள்ள நுண்ணறிவுகளை வழங்கின (Winograd 1983: 18).

தரவுத்தொகுதியின் இன்றியமையாமையைப் பற்றிய இந்த விவாதம் முடிவுக்கு வந்தது. இதன் விளைவாக, மொழி தொழில்நுட்ப வல்லுநர்கள் மற்றும் முதன்மை மொழியியலாளர்கள் இருவரும் மொழி ஆராய்ச்சி மற்றும் பயன்பாடு இரண்டிலும் தரவுத்தொகுதியின் பயன்பாட்டை ஒப்புக் கொண்டனர். தரவுத்தொகுதியின் பகுப்பாய்வு மரபான மொழியியல் விளக்கங்களுக்கு மதிப்புமிக்க கண்ணோட்டங்களை வழங்கியதன் காரணமாக குறுகிய காலத்திற்குள் தரவுத்தொகுதி, மொழியியலின் பல களங்களில் இன்றியமையாத ஒன்றாக மாறியது (Biber 1996: 173). இது பயன்பாட்டு மற்றும் விளக்க மொழியியலின் அனைத்து முக்கிய துணை கிளைகள் உட்பட மொழியியலின் பல களங்களில் அனுபவம்சார் தரவுகளைக் (empirical data) கையாளுவதற்கு வழிவகுக்கிறது: அகராதியியல், சொல்லியல், இலக்கணம், பொருண்மையியல், மொழிக் கல்வி, வரலாற்று மொழியியல், கருத்தாடல், பயன்வழியியல், சமூகமொழியியல், உளமொழியியல், மற்றும் பிற (Leech and Fligelstone 1992).

இருப்பினும், அனுபவ மொழியியல் ஆராய்ச்சிக்கான தற்போதைய பாராட்டு மொழியின் அழகைப் பாராட்டுவதற்காக அல்ல, ஆனால் மொழி விளக்கம் மற்றும் பயன்பாடு ஆகிய இரண்டிலும் தரவுத்தொகுதியின் பல புதிய சாத்தியமான பயன்பாடுகளைக் கண்டுபிடிப்பதற்கான வாய்ப்பின் காரணமாக (Granger and Tyson 2003) என்பதைப் புரிந்து கொள்ள வேண்டும்.

முந்தைய இயல்களிலிருந்து குறிப்புகளை எடுத்துக் கொண்டு, முதன்மை மொழியியலில் தரவுத்தொகுதியின் பயன்பாட்டை முன்னிலைப்படுத்துவதை நோக்கமாகக் கொண்டுள்ளோம். மொழியியலின் பல்வேறு துறைகளில் தரவுத்தொகுதி எவ்வாறு பயன்படுத்தப்படுகிறது என்பதைக்

காட்டுகிறோம். அனைத்து மொழியியல் களங்களிலும் தரவுத்தொகுதியின் பயன்பாட்டின் விளக்கம் இங்கு வழங்க முடியாது என்பதால், அனுபவ தரவு இன்றியமையாததாகக் கருதப்படும் சில முக்கியப் பகுதிகளில் நாங்கள் கவனம் செலுத்துகிறோம்.

7.2 அகராதியியலில் தரவுத்தொகுதி (Corpus In Lexicography)

தரப்படுத்தப்பட்ட அகராதிகள் பொதுவாக ஒரு மொழியில் பயன்படுத்தப்படும் சொற்களின் உருபனியல், சொற்பிறப்பியல், பொருண்மையியல் மற்றும் ஒலியியல் தகவல்களை வழங்குகின்றன. அவற்றின் நோக்கம் பெரும்பாலும் குறைவாக இருப்பதால், அவை துணைவகைப்பாடு, தேர்வு கட்டுப்பாடு மற்றும் சொற்களின் களச்சிறப்பு பயன்பாடு தொடர்பான தகவல்களைக் கொண்டிருக்கத் தவறிவிட்டன. பெரும்பாலான சந்தர்ப்பங்களில் தொகுப்பாளர்கள் சொற்களை அடுக்குகளாக ஒழுங்கமைக்கின்றனர்; அவை பொருத்தமானவை மற்றும் கையாளக்கூடியவை என்று நினைக்கின்றனர். சொற்களின் வடிவம், செயல்பாடு மற்றும் அர்த்தம் இவற்றில் மாற்றம், இயற்கை மொழியைப் புரிந்துகொள்வதற்கு மிகவும் முக்கியமானது என்றாலும், இவை இந்த அகராதிகளில் அரிதாகவே வெளிப்படுகிறது.

சமீபத்திய அனுபவம் தரவுத்தொகுதிமட்டுமே சொற்களின் பயன்பாட்டின் மாறுபாடுகளுக்குள் காணப்பட்ட அனைத்து வகையான தகவலையும் வழங்குகிறது என்பதை வெளிப்படுத்துகிறது. மேலும், சொற்களின் சூழல் அடிப்படையிலான மற்றும் களம் சார்ந்த தகவல்கள் - பயன்பாட்டு அடிப்படையிலான அகராதியின் (usage-based dictionary) இரண்டு முக்கிய பண்புகள் - தரவுத்தொகுதியிலிருந்து சிறந்த முறையில் பெறப்படுகின்றன.

ஒரு தரவுத்தொகுதியைச் செயலாக்கிய பிறகு, வரிசைப்படுத்தப்பட்ட சொற்கள் மற்றும் சொற்றொடர்கள் பல்வேறு வகையான அகராதிகளை உருவாக்கப் பயன்படுத்தப்படுகின்றன (எ.கா. தொழில்நுட்ப சொற்களின் அகராதி, எழுத்துப்பிழைகள், ஒப்புருமொழிகள் (homonyms), ஒருபொருள்பன்மொழிகள் (synonyms) போன்றவை). பி.என்.சி மற்றும் பாங்க் ஆப் ஆங்கிலத்தில் (BNC and Bank of English) இருந்து உருவாக்கப்பட்ட கொலின்ஸ் கோபில்ட் அகராதி (Collins COBUILD dictionary) (1995) அகராதியில் மிகவும் குறிப்பிடப்பட்ட ஒன்றாகும். இது அகராதி ஆராய்ச்சி மற்றும் கல்விக்கு இன்றியமையாத குறிப்பு புத்தகமாகவும் பயன்படுத்தப்படுகிறது.,

தரவுத்தொகுதி அடிப்படையிலான அகராதிகளின் வெற்றி, மொழியின் சாதாரண பயன்பாட்டிற்கு அதிக கவனம் செலுத்த அகராதியலார்களை வழிநடத்துகிறது (Landau 2001: 278). மொழியின் சாதாரண பயன்பாடு எவ்வாறு அகராதியலாரின் உள்ளுணர்வுகளை

துணைநிறைவு செய்யக்கூடும் அல்லது மறுக்கக்கூடும் என்பதற்கான எடுத்துக்காட்டுகளுடன் இது நிரூபிக்கப்பட்டுள்ளது (Atkins and Levin 1995). தரவுத்தொகுதியிலிருந்து பெறப்பட்ட சான்றுகள் பல பொதுவான சொற்களுக்கு, அடிக்கடி வரும் பொருள்/அர்த்தம் முதலில் நம் நினைவுக்கு வருவது அல்ல; அது அகராதிகளில் நடைபெறுகிறது (Sinclair 1991: 39).

ஆகவே, சொற்களின் பாரம்பரிய விளக்கம் தரவுத்தொகுதியிலிருந்து திரட்டப்பட்ட புதிய சான்றுகளால் சவால் விடப்படுகிறது. தரவுத்தொகுதி, அகராதி மற்றும் மொழியியல் ஆகிய இரண்டின் முதிர்ச்சிக்கும்பங்களிப்பு செய்துள்ளது. கொள்கை அடிப்படையில், ஒரு வாழும் மொழியின் அகராதி ஒருபோதும் தகவலால் முழுமையடையாது. வாழ்க்கை, அறிவு மற்றும் நாகரிகத்தின் முன்னேற்றத்தால் பழைய சொற்கள் வழக்கற்றுப் போய்விடும் (எ.கா. இறகுப்பேனா, முக்குப்பேனா போன்றவை) அதே நேரத்தில் புதிய சொற்கள் காலத்தின் புதிய தேவைகளைப் பூர்த்தி செய்ய வருகின்றன (எ.கா. ரிஃபில் , டாட் பேனா முதலியன). மறுபுறம், சொற்றொகையின் மாற்றத்தின் நீரோட்டத்திற்கு எதிராக உயிர்வாழக்கூடிய சொற்கள், அவற்றின் வடிவம், பொருள் மற்றும் பயன்பாடு ஆகியவற்றை காலப்போக்கில் புதிய காலநிலைக்கு ஏற்ப மாற்றும்.

ஒரு அகராதிக்கு இரண்டு முக்கிய பங்குகள் உள்ளன: (அ) இது சொல்சார் பரிணாமத்திற்கான இருகாலத் தேடலின் மூலம் சொல்சார் இழப்பின் தடத்தை மீட்டெடுக்க வேண்டும், மற்றும் (ஆ) சொல் பயன்பாட்டின் சமகால காட்சியை சொல்சார் பட்டியல்களின் ஒருகால முந்திட்டத்தின் மூலம் முன்வைக்க வேண்டும்.

தரவுத்தொகுதியை அணுகி அகராதியலார்கள் அதிலிருந்து ஒரு சொல் பயன்பாட்டின் அனைத்து எடுத்துக்காட்டுகளையும் எளிதில் பெறுகிறார்கள்; சொற்களின் பயன்பாட்டு மாறுபாடுகள் குறித்த புதுப்பித்த தகவல்களை மீட்டெடுத்த பிறகு அகராதிகளை எளிதாகவும் விரைவாகவும் தயாரிக்கவும் திருத்தவும் செய்கிறார்கள். தரவுத்தொகுதியிலிருந்து பெறப்படும் தகவல்கள் அடிப்படையில் அவர்களால் அதிக எண்ணிக்கையிலான எடுத்துக்காட்டுகளை ஆராய இயலுகிறது; எனவே அவர்கள் சொற்களின் வரையறையை மேலும் முழுமையான மற்றும் துல்லியமான முறையில் வழங்குகின்றனர்.

இந்த விஷயத்தில், ஒரு திறந்த (அதாவது தொடர்ந்து வளர்ந்து வரும்) தரவுத்தொகுதிக்கு முக்கிய பங்கு உண்டு; புதிய சொற்கள் மொழியில் நுழைவதைக் கண்டுபிடிக்கவும், அதே நேரத்தில் இருக்கும் சொற்கள் அவற்றின் அர்த்தத்தையும் பயன்பாட்டையும் மாற்றுவதை

அறியவும் அகராதியியலார்களுக்குத் தரவுத்தொகுதி உதவுகிறது. மேலும், அகராதி தயாரிப்பாளர்கள் சொற்களின் அதிர்வெண் பட்டியல்களை உருவாக்கலாம்; இனங்களில் அடிப்படையில் அவற்றின் பயன்பாட்டைச் சமப்படுத்தலாம் மற்றும் மொழி வகைகளில் அடிப்படையில் அவற்றைப் பிரிக்கலாம்.

ஒரு தரவுத்தொகுதியில் சமூகமொழியியல் மாறிகள் (sociolinguistic variables) (எ.கா. இடம், ஆசிரியர், நாள், வகை/இனம், தொழில், பாலினம் போன்றவை) தொடர்பான அதிக அளவு உரைத் தகவல்கள் உள்ளன. குறிப்பிட்ட சொற்கள் அல்லது சொற்றொடர்களின் பயன்பாட்டை குறிப்பிட்ட வகை, இனம், காலம் அல்லது தொழிலுக்கு பொதுவானதாக இணைக்க அகராதி தயாரிப்பாளர்களுக்கு இது உதவுகிறது.

அகராதியலார்கள் பல்வேறு வகையான சொற்களின் கலவையை எளிதில் பெறலாம்; அவற்றின் அடிப்படை பரஸ்பர தகவல்களை பகுப்பாய்வு செய்வதன் மூலம் இணை நிகழும் (co-occurring) சொற்களுக்கு இடையிலான உறவை ஏற்படுத்தலாம். குறிப்பிட்ட சொல் அர்த்தத்தை மீட்டெடுப்பதற்கான முக்கியமான தடயங்களை வழங்குவதால் இது மொழி பயனர்களுக்கு சொற்றொடர்கள், மரபுத்தொடர்கள் (idioms) மற்றும் சொல்லடிவகைப்படு (collocations) என்பனவற்றை மிகவும் முறையாக புரிந்துகொள்ள உதவுகிறது.

இந்திய மொழிகளின் கண்ணோட்டத்தில், பல்வேறு வகையான அகராதிகளை உருவாக்குவதற்கு தரவுத்தொகுதியைப் பயன்படுத்த நாம் நினைக்கிறோம் (எ.கா. ஒருமொழி மற்றும் இருமொழி அகராதி, தொழில்நுட்ப மற்றும் அறிவியல் சொற்களின் அகராதி, எழுத்துப்பிழைகள், ஒப்புருமொழிகள், ஒருபொருள்பன்மொழிக்கள் போன்றவை). இத்தகைய அகராதிகள் சந்தைகளில் கிடைக்கும் அகராதிகளைக் காட்டிலும் மேம்பட்டவை, தகவல் தரும் மற்றும் நடைமுறைக்குரியவை. பாங்க் ஆப் ஆங்கிலத்திலிருந்து பெறப்பட்ட தகவல்கள் ஆங்கில அகராதியின் ஒவ்வொரு பண்புக்கூறையும் மேம்படுத்தியுள்ளன (Rundell 1996). அது எவ்வாறு சாத்தியமானது என்பதை தெளிவுபடுத்துவோம்.

கேள்வி 1: இப்போதெல்லாம் ஆங்கிலத்தில் 'gay' என்ற வார்த்தையை அதன் பழைய அர்த்தத்தில் நாம் எப்போதாவது பயன்படுத்துகிறோமா?

கேள்வி 2: 'different from' என்று சொல்வது சரியானதா அல்லது அது 'different to' என இருக்க வேண்டுமா?

கேள்வி 3: 'cheap' என்ற சொல் எதிர்மறையான பொருளைக் கொண்டிருக்கிறதா?

இந்தக் கேள்விகளுக்கு பதிலளிக்க, சொற்கள் எவ்வாறு பயன்படுத்தப்பட்டு வருகின்றன என்பதை தரவுத்தொகுதியை உற்றுநோக்கி அறியவேண்டும். எந்த அர்த்தத்தில் சொற்கள் உண்மையில் பயன்படுத்தப்படுகின்றன என்பதைத் தரவுத்தொகுதி காட்டும் (கேள்வி 1 இன் பதில்); எந்த வடிவத்தில் எந்த உரை வகையில் மிகவும் பொதுவானது (கேள்வி 2 இன் பதில்); வார்த்தைகள் ஏதேனும் கேவலமான அர்த்தத்தைக் கொண்டிருக்கிறதா (கேள்வி 3 இன் பதில்).

பாங்க் ஆப் ஆங்கிலத்திலிருந்து எடுத்துக்காட்டுகளின் பகுப்பாய்வு 'gay' அதன் பழைய அர்த்தத்தைக் கிட்டத்தட்ட இழந்துவிட்டது என்பதைக் காட்டுகிறது (கேள்வி 1 இன் பதில்); எழுத்திலும் மற்றும் பேச்சிலும் 'different from' மற்றும் 'different to' என்ற சொற்றொடர்களின் குறிப்பிடத்தக்க பயன்பாட்டு மாறுபாடு பராமரிக்கப்பட்டுள்ளது (கேள்வி 2 இன் பதில்); 'cheap' என்பது ஒரு திட்டவட்டமான இழிவான பொருளைக் கொண்டிருந்தாலும் அர்த்தத்தில் நடுநிலையானது (கேள்வி 3 இன் பதில்).

கணேஷ் அம்பேத்கர் (2019) தமது கட்டுரையில் தரவுத்தொகுதியைப் பயன்படுத்தி தமிழில் *போல* மற்றும் *மாதிரி* என்ற பின்னொட்டுகளின் பயன்பாட்டு வேறுபாட்டை விளக்குவது தமிழ் மொழியின் அகராதி உருவாக்கத்தில் தரவுத்தொகுதியின் இன்றியமையாமையைப் பறைசாற்றி நிற்கும். அவர் 'அதிகம்' (more) என்பதற்கும் 'மிக அதிகம்' (most) என்பதற்கும் இடையிலுள்ள ஒப்பீட்டை வெளிப்படுத்த *மாதிரி* என்பது பயன்படுத்தப்படுவதாகவும் 'ஒன்று' (one) என்பதற்கும் 'பல' (many) என்பதற்கும் இடையிலான ஒப்பீட்டை வெளிப்படுத்த *போல* பயன்படுத்தப்படுவதாகவும் முடிவுரை செய்கின்றார்.

கேள்வி 4: பொதுவான அகராதியில் எந்த தொழில்நுட்பச் சொற்கள் சேர்க்கப்பட வேண்டும்?

இது அகராதி அகராதியிலார்களுக்கு ஒரு முக்கியமான கேள்வி ஆகும். ஒரு பதிலைக் கண்டுபிடிக்க, பத்திரிகைகள் மற்றும் செய்தித்தாள்களில் தோன்றும் சொற்கள் பொதுவான மக்களின் சொற்றொகையில் அரிதாகவே சந்திக்கப்படுகின்றன என்பதைக் காட்டுவதற்கான சொற்களுக்கான அதிர்வெண் புள்ளிவிவரங்களை ஆங்கிலத்தின் வங்கி/பாங்க் ஆப் ஆங்கிலம் (Bank of English) அளிக்கிறது. இது ஒரு சாத்தியமான தீர்வாக இருக்கலாம், ஆனால் ஆதாரங்கள் இல்லாமல் கேள்விக்கு பதிலளிக்க கிட்டத்தட்ட சாத்தியமில்லை. பல நபர்கள் இந்த செய்தியைப் பற்றி வெவ்வேறு கருத்துக்களைக் கொண்டிருப்பார்கள், ஆனால் ஒரு அகராதியில் மொழித் தகவல்களைச் சேர்க்கும் போது, ஒரு மொழியைப் பற்றி ஒரு தரவுத்தொகுதி வழங்கும் தரவுகளின் செல்வத்தை நாம் கணக்கில் எடுத்துக்கொள்ள வேண்டும்.

ஆகவே, மேம்பட்ட கற்றவர்களுக்கான கொலின்ஸ் கோபில்ட் ஆங்கில அகராதியின் (Collins COBUILD English Dictionary for Advanced Learners 2000) சமீபத்திய பதிப்பில் பல புதுமையான பண்புக்கூறுகளைச் சேர்க்க முடிந்தது; ஏனெனில் இது இன்றுவரையுள்ள உரைகளில் பயன்பாட்டில் உள்ள மில்லியன் கணக்கான சொற்களில் மேற்கொள்ளப்பட்ட ஆய்வுத் தகவல்களைப் பயன்படுத்தியுள்ளது.

எலக்ட்ரானிக் (போகுரேவ் மற்றும் புஸ்டெஜ்வ்ஸ்கி/ Boguraev and Pustejvsky 1996) அல்லது அச்சிடப்பட்ட வடிவத்தில் (ஓஒய்/Ooi 1997) அகராதியை உருவாக்க விரும்பினால், பின்வரும் செய்திகளை நாம் கருத்தில் கொள்ள வேண்டும்; ஏனென்றால் இவை இல்லாமல் இன்றைய மொழியின் அகராதி முழுமையடையாது. இருப்பினும், மின்னணு பதிப்பைப் பொறுத்தவரை, ஒவ்வொரு கட்டமும் வழக்கமான மனித தலையீட்டிற்கான அனுமதியுடன் அதிகபட்சமாக தானியக்கம் செய்யப்பட வேண்டும்.

- ஒருமொழி அகராதியைப் பொறுத்தவரை, அகராதியலார்கள் பெரிய ஒருமொழி தரவுத்தொகுதியைக் கொண்டிருக்க வேண்டும். இது எழுதப்பட்ட மற்றும் பேசும் வகைகளின் மாதிரிகளை உள்ளடக்கும்; இதனால் அகராதியலார்கள் அகராதியில் சேர்க்க விரும்பும் தேவையான தகவல்களை வழங்க முடியும்.
- தரவுத்தொகுதியிலிருந்து பெறப்பட்ட சொல்லனாக்கம் செய்யப்பட்ட சொற்களின் (lemmatised words) முழுமையான அதிர்வெண் பட்டியல், எந்த சொற்கள் அகராதியில் சேர்க்கப்படும் அவ்வளவு முக்கியத்துவம் வாய்ந்தவை என்பதைத் தீர்மானிக்க அகராதியலார்கள்களுக்கு உதவும்.
- அவர்கள் எழுத்துச் சரிபார்ப்பு முறையைப் பயன்படுத்த வேண்டும்; அது எழுத்து மாறுபாட்டைப் பதிவு செய்கின்ற சொற்களைச் சேகரிக்க உதவும். அதிர்வெண் அடிப்படையில், அகராதியில் சேர்ப்பதற்கு எந்த எழுத்துவடிவம் கருதப்படும் என்பதை அவர்கள் தீர்மானிப்பார்கள். தமிழ் போன்ற ஒரு மொழிக்கு இத்தகைய கருத்தாய்வு மிகவும் முக்கியமானது.
- சொற்களின் உச்சரிப்பில், அகராதியலார்கள் சொற்களின் நிலைபேறுபெற்ற மற்றும் பிராந்திய உச்சரிப்பு மாறுபாட்டைக் கருத்தில் கொள்ள வேண்டும். இதற்காக, அவர்கள் பேசு உரைகளின் தரவுத்தொகுதியைப் பயன்படுத்த வேண்டும்.

- சொற்களின் பிராந்திய மாறுபாட்டைக் உருப்படுத்தம் செய்ய, அகராதி தயாரிப்பாளர்கள் எந்தெந்தச் சொற்கள் எந்தப் பகுதிகளில் பயன்படுத்தப்படுகின்றன, எந்த அதிர்வெண்ணில் பயன்படுத்தப்படுகின்றன என்பதை அடையாளம் காணப் புள்ளியியல்சார் தகவல்களைக் கொண்டிருக்க வேண்டும். ஒரு நிலைபேறுபெற்ற மொழி வகையில் ஒரு சொல்லின் அதிர்வெண்ணுடன் தொடர்புபடுத்தி ஒரு சொல்லுக்குப் பிராந்திய/வட்டார மொழி வகையில் அதிக அல்லது குறைந்த அதிர்வெண் உள்ளதா என்பதை அறிய இது உதவும்.
- சொற்களின் வகை விநியோகத்தை அறிய புள்ளியியல்சார் தகவல்களும் தேவை. அகராதியில் சேர்க்கப்பட்டுள்ள சொற்கள் எல்லா உரை வகைகளிலும் சம அதிர்வெண்ணுடன் பயன்படுத்தப்படுகின்றனவா, அல்லது அவை உரையின் வகையைப் பொறுத்து மாறுபடுகின்றனவா (எ.கா. பேச்சு, எழுத்து, செய்தித்தாள், கற்பனை உரை, தகவல் உரை போன்றவை) என்பது குறித்து அகராதியியலார் அறிந்திருப்பது அவசியம்.
- இலக்கணத் தகவலைப் பொறுத்தவரையில், அகராதியில் சேர்க்கப்பட வேண்டிய சொற்களின் சொல்வகுப்பு, அவற்றின் சொல்லடி வகைப்பாட்டின் அமைப்பொழுங்கு, சொல்லடி வகைப்பாட்டின் நிகழ்வெண், விநியோகத்தின் தன்மை போன்றவற்றை அகராதியலார்கள் அடையாளம் காண வேண்டும்.
- சொற்களின் பொருண்மையியல்சார் பண்புகள் குறித்து அவர்களுக்குப் போதுமான மொழியியல் தகவல்கள் இருக்க வேண்டும். இதில் அகராதியில் சேர்க்கப்பட்டுள்ள சொற்களின் உள்ளடங்குமொழிகள் (hyponyms), ஒருபொருள்பன்மொழிகள் (synonyms), பல்பொருளொருமொழிகள் (polysyms), எதிர்மொழிகள் (antonyms), உள்ளடக்குமொழிகள் (hypernyms) போன்றவை இருக்கலாம்.

அகராதியலார்களின் தேவைகளைப் பூர்த்தி செய்ய, ஒரு தரவுத்தொகுதியின் பொருத்தம் மதிப்பீட்டிற்கு அப்பாற்பட்டது. மேலே கூறப்பட்ட ஒவ்வொரு தேவைகளையும் பூர்த்தி செய்ய, அகராதி தயாரிப்பாளர்கள் தரவுத்தொகுதியை நேர்த்தியுடன் மற்றும் செயல்திறனுடன் பயன்படுத்துவார்கள்.

அவர்கள் தங்கள் உள்ளுணர்வுகளைத் துணைசெய்யவோ அல்லது மறுக்கவோ கூடிய ஆதாரங்களை சேகரிக்கத் சொற்றொகுதியைப் பயன்படுத்துவார்கள். அவரது அகராதியியல் நடவடிக்கைகள் பின்வருவனவற்றை உட்படுத்தும்: காலப்போக்கில் மொழியில் நுழையும் புதிய சொற்களையும் கலைச்சொற்களையும் கண்டறிதல்; இருக்கும் சொற்கள் அவற்றின் அர்த்தங்களை

எவ்வாறு மாற்றுகின்றன என்பதை அடையாளம் காணுதல்; இனங்கள் மற்றும் உரை வகைகளுக்கு ஏற்ப சொற்கள் அவற்றின் பயன்பாட்டை எவ்வாறு சமன் செய்கின்றன என்பதை அடையாளம் காணுதல்; அவற்றின் சூழல் மாறுபாடுகளை அடையாளம் காண சொற்களின் எடுத்துக்காட்டுகளை பெறுதல்; புதிய மற்றும் புதுப்பித்த தரவு மற்றும் தகவலுடன் ஏற்கனவே இருக்கும் அகராதிகளைத் திருத்தல்; பல்வேறு மொழியியல் பொருட்களின் (சொற்களின்) முழுமையான மற்றும் துல்லியமான வரையறைகளை முன்வைத்தல்; மொழியில் ஏதேனும் மாற்றம் அல்லது சொற்களை இழப்பது குறித்த புதுப்பித்த தகவல்களைக் கொடுத்தல்; பகுப்பாய்விற்கான எடுத்துக்காட்டுகளை மிகவும் அர்த்தமுள்ள குழுக்களாக ஏற்பாடு செய்தல்; வெவ்வேறு ஆராய்ச்சி அளவுருக்களின் படி தனிப்பட்ட சொற்களை வகைப்படுத்துதல்; அவற்றின் இணை நிகழ்வை உறுதிப்படுத்தும் உள்ளார்ந்த பரஸ்பர உறவை ஆராய சொற் சேர்க்கைகளை தனிமைப்படுத்துதல்; குறிப்பிட்ட சொல்-அர்த்தத்திற்கு முக்கியமான தடயங்களை வழங்குவதற்காக சொற்றொடர்களையும் சொல்லடி வகைபாடுகளையும் முறையாக நடத்துதல்; மற்றும் சூழ்நிலை தகவல்களை ஆராய்வதன் மூலம் குறிப்பிட்ட பிராந்திய/வட்டார வகைகள் மற்றும் இனங்களுக்கு பொதுவானதாக சொற்கள் மற்றும் சொற்றொடர்களைப் பயன்படுத்துதல்.

இந்த அனைத்து செயல்பாடுகளிலும் தரவுத்தொகுதியின் உள்ளடக்கம் மற்றும் அளவு முக்கியமான சிக்கல்கள் ஆகும். நம்பகத்தன்மையைப் பொறுத்த வரையில், அகராதிப் பதிவுகள் ஒரு எடுத்துக்காட்டை மட்டும் அடிப்படையாகக் கொண்டிருக்க இயலாது. ஒரு குறிப்பிட்ட பண்புக்கூறு அல்லது பயன்பாட்டு முறை வெளிப்படுவதற்கு முன்பு அல்லது அகராதியில் சேர்க்க புள்ளியியல் அடிப்படையில் நம்பகமானதாக மாறுவதற்கு முன்பு குறைந்தது பத்து அல்லது அதற்கு மேற்பட்ட எடுத்துக்காட்டுகள் தேவை. இருப்பினும், பல சந்தர்ப்பங்களில், 10+ டோக்கன்களைக் கொண்ட சொற்கள் பெரும்பாலான அகராதிகளில் சேர்ப்பதற்கு கருதப்படுகின்றன, இருப்பினும் அவை சேர்க்கப்படுவது ஒரு அகராதியை உருவாக்கப் பயன்படுத்தப்படும் தலையங்கக் கொள்கையைப் பொறுத்தது.

தரவுத்தொகுதி அடிப்படையிலான அகராதிகளைத் தொகுக்கத் தொடங்குவதற்கு முன் பில்லியன் கணக்கான சொற்களின் வரம்பில் தரவுத்தொகுதி தேவை. அதுவரை, தரவுத்தொகுதி பெரும்பாலான அகராதிசார் மற்றும் மொழியியல் செயல்பாடுகளில் எளிதாக உதவ முடியும், ஆனால் அது போதுமான அளவு என்று சொல்ல இயலாது.

தரவுத்தொகுதிகள் முதன்முதலில் வெளிநாட்டு மொழியாக ஆங்கிலம் (English as a Foreign Language (EFL/ஈ.எஃப்.எல்) என்ற நிலையில் அகராதி உருவாக்கப் பயன்படுத்தப்பட்டன என்பது தற்செயல் நிகழ்வு அல்ல; அதில் தலைச் சொற்களின் எண்ணிக்கையின் தேவை மிதனானதாக இருந்தது. அப்போதும் கூட, இது பொதுவான பொதுவான சொற்களைத் தவிர மற்ற அனைத்துச் சொற்கள், இலக்கண அமைப்பொழுங்கு மற்றும் சொல்லடிவகைப்பாட்டு நடத்தை பற்றிய நம்பகமான ஆதாரங்களை வழங்க பெரிய தரவுத்தொகுதிகளைப் பயன்படுத்தியது.

சிறிய தரவுத்தொகுதியுடன் பணிபுரியும் அகராதியியலார்களை நாம் குறைகூற வாதிடவில்லை. தரவுத்தொகுதி அடிப்படையிலான அகராதியில் கோபில்டின் (COBUILD's) ஆரம்ப முயற்சி ஒன்று மற்றும் ஏழு மில்லியன் சொற்களைக் கொண்ட தரவுத்தொகுதிகளிலிருந்து சேகரிக்கப்பட்ட எடுத்துக்காட்டுகளின் அடிப்படையில் அமைந்தது. 450 மில்லியன் சொற்களைக் கொண்ட தரவுத்தொகுதியுடன் சரிபார்க்கப்பட்டாலும் கூட, அவற்றில் பல எடுத்துக்காட்டுகள் சாராம்சத்தில் இன்னும் செல்லுபடியாகும். களச்சிறப்பு அகராதிக்கு (domain-specific dictionary) சிறிய தரவுத்தொகுதி போதுமானதாக இருக்கலாம் என்பதை இது குறிக்கிறது, ஆனால் முழு அகராதியைப் பற்றிய பொதுமைப்படுத்தல்களுக்குப் பெரிய தரவுத்தொகுதிசிறந்தது.

7.3 சொல்லியியலில் தரவுத்தொகுதி (Corpus in Lexicology)

சொல்லியியல், சொல்சார் ஆய்வுக்கான பலவிதமான ஆர்வங்களையும் அணுகுமுறைகளையும் உள்ளடக்கியது. சொற்களின் பொருண்மை மற்றும் பொருண்மை மாற்றம், சொல்சார் மாறுபாடு மற்றும் கால அளவிலான மாற்றம், பல நூற்றாண்டுகளாக சொற்றொகையின் பரிணாமவளர்ச்சி, மொழிகளில் நிகழும் புதுச்சொல்லாக்கம் மற்றும் சொல் இழப்பு, காலப்போக்கில் சொல்சார் கடன்வாங்கல் மற்றும் ஆக்கம், சொல் அலகுகளின் கட்டமைப்பு மற்றும் சொற்பிறப்பியல் ஆய்வு ஆகியவை இதில் அடங்கும்; இவ்வாறு சொல்லியியலானது பொருண்மையியல், தொடரியல் மற்றும் பயன்வழியியல் இவற்றுடன் நெருங்கிய இடைமுகத்தைக் கொண்டுள்ளது.

இந்தியாவில் இந்த புலம் இன்னும் செழிக்கவில்லை என்றாலும், ஆங்கிலம் மற்றும் பிற மொழிகளில் பல்வேறு வகையான தரவுத்தொகுதியின் பகுப்பாய்வை அடிப்படையாகக் கொண்டவை சில படைப்புகள் உள்ளன (Coleman and Kay 2000, Vera 2002). தற்போது

கிடைக்கக்கூடிய இந்திய மொழிகளின் தரவுத்தொகுத்யுடன் இதேபோன்ற ஆய்வுகள் எவ்வாறு தொடங்கப்படலாம் என்பதைக் காட்டுவதற்காக இவற்றில் சில இங்கே குறிப்பிடப்படுகின்றன.
and Michigan Early Modern English Materials Corpus along with

ஆங்கில உரைகளின் ஹெல்சின்கி தரவுத்தொகுதி (Helsinki Corpus of English Texts) மற்றும் மிச்சிகன் தொடக்ககால நவீன ஆங்கிலப் பொருட்களின் தரவுத்தொகுதி (Michigan Early Modern English Materials Corpus) மற்றும் ஆக்ஸ்போர்டு ஆங்கில அகராதி (Oxford English Dictionary) ஆகியவை புலனறிவுசார் செயல்முறைகளைக் குறிக்கும் அருவப் பெயர்களில் (abstract nouns) சொற்பொருள்/பொருண்மை மாற்றத்தில் ஏற்படும் திசைகளையும் வழிமுறைகளையும், குறிப்பாக பேசுபவர் கண்ணோட்டத்தில், ஆய்வு செய்யப் பயன்படுகின்றன (Alanko 2000). தரவுத்தொகுதியின் காலவரிசை ஆய்வு, சொற்பொருள் மாற்றத்தை ஆய்வு செய்வதற்கு முன்மாதிரிகள் போன்ற கருவிகள் மற்றும் சொற்பொருள் மாற்றத்தில் திசைகளை நம்பத்தகுந்த முறையில் ஆவணப்படுத்த வரலாற்றுத் தரவுத்தொகுதிகள் போன்ற பொருட்கள் தேவை என்பதை தெளிவுபடுத்துகிறது. தரவுத்தொகுதியைப் பற்றிய கூடுதல் ஆய்வு, பொருண்மை/பொருள் புலத்திலிருந்து (semantic field) அடிப்படை அர்த்தங்கள் மறைந்துவிடாது என்பதைக் காட்டுகிறது; இருப்பினும், அவை தனிப்பட்ட சொற்களின் விஷயத்தில் இழக்கப்படலாம். முன்மாதிரி மையத்திற்குள் சொற்களின் புதிய அர்த்தங்களை உருவாக்குவதில் சொற்பொருள்/பொருண்மை மாற்றத்தின் அமைப்பொழுங்கு மீண்டும் மீண்டும் நிகழ்கிறது; இதனால் வர்ணனை ஒரே மாதிரியாக இருக்கும்.

வரலாற்று தரவுத்தொகுதிகளிலிருந்து பெறப்பட்ட சொற்களின் பெரிய பட்டியல் பொதுவாக ஆங்கிலத்தில் 'குழுஉக்குறிகளை' ('cants'/'கேண்ட்கள்) ஆக்குவதற்குப் பயன்படுத்தப்படும் சொல் உருவாக்கும் செயல்முறைகளைப் பிரதிபலிக்கப் பயன்படுகிறது (Gotti 2002). ஆரம்பகால கேன்ட் அகராதிகளின் தொகுப்பாளர்கள் சொல் உருவாக்கும் செயல்முறைகளைப் பற்றி குறிநிலை மொழியியல் (metalinguistic) விழிப்புணர்வைக் கொண்டிருந்ததாக பகுப்பாய்வு தரவு காட்டுகிறது.

அகராதியில் 'கடினமான சொற்களை' நிரூபிப்பதற்கான முறையை ஆய தரவுத்தொகுதியிலிருந்து வரும் சான்றுகள் பயன்படுத்தப்படுகின்றன (McDermott 2002). தரவுதொகுதி தரவின் பகுப்பாய்வு, மக்களின் உண்மையான சொற்றொகையின் ஒரு பகுதியாக மிகவும் கடினமான சொற்கள் ஒருபோதும் பயன்படுத்தப்படுவதில்லை என்ற உண்மையை

நிறுவுகிறது; இருப்பினும் அவை அகராதியில் முக்கியமான இடங்களை வகிக்கின்றன. கடினமான சொற்களை அகராதியிலிருந்து அகற்றுவதற்கு இது வாதிடுகிறது; ஏனெனில் அவை இயல்பான மொழிப் பயிற்சியில் பயன்படுத்தப்படவில்லை.

ஐ.சி.ஏ.எம் சேகரிப்பு ஆங்கில மொழித் தரவுதொகுதிகள் (ICAME Collection of English Language Corpora 1999) ஆங்கிலத்தில் சொல் இழப்பு மற்றும் சொற்பொருள் மாற்றத்தின் நிகழ்வை விளக்கும் காரணங்களை அடையாளம் காணப் பயன்படுத்தப்படுகிறது (Cabanillas and Martínez 2002). புதிதாக அறிமுகப்படுத்தப்பட்ட சொற்கள் பெறுநரின் மொழியை எவ்வாறு பாதிக்கின்றன மற்றும் சொந்த சொற்களை எவ்வாறு பாதிக்கின்றன என்பதை ஆய்வு காட்டுகிறது. சொற்பொருள் மாற்றத்தின் செயல்பாட்டில் உருவகத்தைப் பயன்படுத்துவதால் சொற்களின் அசல் அர்த்தத்தில் குறிப்பிடத்தக்க மாற்றத்தை இது எடுத்துக்காட்டுகிறது.

டொரொன்டோ மத்தியகால ஆங்கிலத் தரவுத்தொகுதி (Toronto Corpus of Middle English) சில சொற்களின் சொல்சார் புலங்களால் வெளிப்படுத்தப்பட்ட நேரடி (literal), உருவகம்சார் (metaphorical) மற்றும் ஆகுபெயர்சார் (metonymical) அர்தங்களை புனரமைக்க பயன்படுத்தப்பட்டது (Gevaert 2002). கருத்துக்களின் சமநிலையை நிவர்த்தி செய்வதற்காக சொற்களின் அடிப்படை கருத்துருசார் துறைகள் பெரும்பாலும் வெளிநாட்டு கருத்துகளின் செல்வாக்கின் கீழ் மாற்றங்களுக்கு உள்ளாகின்றன என்பதை பகுப்பாய்வு முடிவுகள் காட்டுகின்றன. வரலாற்று தரவுத்தொகுதிகளின் அளவு பகுப்பாய்வின் அடிப்படையில் வரலாற்று, அறிவாற்றல் மற்றும் முன்மாதிரி சொற்பொருள்களை ஒருங்கிணைக்கும் அணுகுமுறையால் மட்டுமே பரிணாம வளர்ச்சியின் சொற்கள் மற்றும் சொற்களின் தொடர்பு அளவிட முடியும் என்று அது சுட்டிக்காட்டுகிறது.

பி.என்.சி. (BNC), ஃப்ளோப் தரவுதொகுதி (FLOB Corpus), ஃப்ரீபர்க்-பிரவுன் தரவுதொகுதி (Freiburg-Brown Corpus) மற்றும் ஆங்கில உரையின் ஹெல்சின்கி தரவுதொகுதி (Helsinki Corpus of English Text) ஆகியவை நவீன ஆங்கிலத்தில் பயன்படுத்தப்படும் சில அருவப் பெயர்களின் இருகால கண்ணோட்டம் நோக்கிச் செல்லும் சொற்பொருள் மாற்றத்தில் முன்மாதிரிகளை ஆராய பயன்படுத்தப்படுகின்றன (Alanko 2002). இருகால தரவுதொகுதிகளிலிருந்து பெறப்பட்ட ஒரு மொழியின் கணிப்புகள், சொற்பொருள் மாற்றத்தில் அதிகரிக்கும் பேசுபவர் கண்ணோட்டதின் (subjectification) அமைப்பொழுங்கு புலனறிவு மொழியியலில் (cognitive linguistics) காணப்படுகிறதா என்பதை ஆராய ஆய்வாளர்களை அனுமதிக்கிறது. முன்மாதிரி சொற்களின்

சொற்பொருள் மாற்றத்தின் செயல்முறை மற்றும் சில வரலாற்று கட்டத்தில் அவற்றின் ஒருபொருள்பன்மொழிகள், சொற்களின் பொருண்மைப் புலங்கள் படிப்படியாகக் குறைக்கப்படுவதைக் காட்டுகின்றன; இருப்பினும் அவற்றின் பல்பொருண்மைகள் (multiple senses) முற்றிலும் மறைந்துவிடாது. அவை அவற்றின் முந்தைய அருகிலுள்ள ஒருபொருள்பன்மொழிகளுக்கு மாற்றப்படுகின்றன; அதே நேரத்தில், சில புதிய சொற்கள் சொற்பொருள் மாற்றத்தில் அதிகரிக்கும் பேசுபவர் கண்ணோட்டதின் அமைப்பொழுங்கை தொடர்ந்து செய்யப் புலத்தில் நுழைகின்றன.

and

ஹெல்சின்கி மத்திய கால ஆங்கிலத் தரவுத்தொகுதி (Helsinki Corpus of Middle English) மற்றும் வரலாற்று ஆங்கில பதிவேடுகளின் ஒரு பிரதிநிதித்வத் தரவுதொகுதி (A Representative Corpus of Historical English Registers) ஆகியவை சொலாக்க அமைப்பொழுங்குகளில் மாற்றங்களைக் கண்டறியவும், சொல் ஆக்கத்தின் பொதுவான இயக்கவியலை (Cowie and Puffer 2002) கையாளவும் ஆய்வு செய்யப்படுகின்றன. சொல் ஆக்கத்தில் உற்பத்தித்திறன் (productivity) முறைகளின் பகுப்பாய்வு (ஒரு தரமான-அளவு மற்றும் இருகாலச் செயல்முறையாக) சொல் ஆக்கத்தில் பல்வேறு செயல்முறைகள் உள்ளன என்பதையும் அவை பொதுவாக வெவ்வேறு காலகட்டங்களில் மாற்றங்களுக்கு உட்படுகின்றன என்பதையும் வெளிப்படுத்துகின்றன. இந்த காரணி காரணமாக, சொல்சார் உற்பத்தித்திறன் ஒரு தூய கோட்பாடுசார் கருத்தாகக் கருதப்படவில்லை. மாறாக, இது காலப்போக்கில் மிகவும் அளவிடக்கூடிய பண்புக்கூறாகும்.

இன்ஸ்ப்ரூக் உரைநடை தரவுதொகுதி (Innsbruck Prose Corpus) - இன்ஷ்ப்ரக் கம்ப்யூட்டர் காப்பகத்தின் இயந்திரம் படிக்கக்கூடிய ஆங்கில உரைகளின் (Innsbruck Computer Archive of Machine Readable English Texts) துணைக்குழு, அகராதியலில் வரலாற்று தரவுதொகுதியின் சாத்தியமான பயன்பாடுகளைக் கருத்தில் கொள்ள பகுப்பாய்வு செய்யப்படுகிறது (Markus 2002). (அ) செயல்பாட்டுச் சொற்களின் தொடரியல், (ஆ) நிலையான வெளிப்பாடுகள் மற்றும் மரபுத்தொடர்கள்களின் பொருண்மையியல் மற்றும் (சி) சொற்களின் பயன்வழியியல் மற்றும் நடையியல் பண்புக்கூகள் (அல்லது சொல் சேர்க்கைகள்) ஆகியவற்றுடன் தொடர்புடைய ஒரு சிக்கலான இடைமுகம், முறையான சொல்லியல் ஆய்வுகளுக்கு அவசியமான பொருட்கள் ஆகும் என்பதை பல்வேறு மொழியியல் துணைப்பிரிவுகளுடன் இணைக்கப்பட்டுள்ள பல எடுத்துக்காட்டுகளின் குறிப்புரை காட்டுகிறது. இயல்பாக்கப்பட்ட மற்றும்

அடையாளப்படுத்தப்பட்ட உரைகளின் தரவுத்தொகுதிகள் இந்த வகைச் சொல்சார் ஆராய்ச்சிக்கு மிகவும் பயனுள்ளதாக இருக்கும் என்பதை வெளிப்படுத்துகிறது.

ஹெல்சின்கி ஆங்கில உரைகளின் தரவுதொகுதி: இருகாலம் மற்றும் கிளைமொழிகள் (Helsinki Corpus of English Texts: Diachronic and Dialects) மத்திய ஆங்கிலத்தில் சொல்சார் கடன் வாங்கலின் போக்கையும் தன்மையையும் அறியப் பயன்படுகிறது (Skaffari 2002). மத்திய ஆங்கிலத்தில் பிரெஞ்சு, ஸ்காண்டிநேவிய மற்றும் லத்தீன் கடன் சொற்களின் (Scandinavian, and Latin loanwords) அளவு-தர ஆய்வு (Quantitative-qualitative examination) பல நூற்றாண்டுகளாக பரவியிருக்கும் சொல்சார் கடன் வாங்கும் செயல்முறையின் மீது இருகாலக் கண்ணோட்டத்தை வெளிப்படுத்துவதில் ஒருகாலத் தரவுகளின் திறனை நிறுவுகிறது. அளவிடப்பட்ட தகவல்கள் அகராதி தொடர்பான பிரச்சினைகள் குறித்து வெளிச்சம் போடத் தவறிவிட்டாலும், இது பல ஆண்டுகளாக ஒரு மொழியின் சொற்றொகையின் வளர்ச்சியை விளக்குகிறது. மேலும், தரவுத்தொகுதியிலிருந்து சேகரிக்கப்பட்ட கடன் சொற்களின் சூழல் பயன்பாட்டின் எடுத்துக்காட்டுகள் இந்த நிகழ்வின் ஆழமான ஆய்வுக்குத் தேவையான மதிப்புமிக்க நுண்ணறிவுகளை வழங்குகின்றன.

நவீன ஆங்கிலத்தின் ஹெல்சின்கி தரவுதொகுதி Helsinki Corpus of Modern English, லாம்பேட்டர் கார்பஸ் Lampeter Corpus, ஆரம்பகால நவீன ஆங்கில கடிதத் தொடர்பு மாதிரிகள் (Corpus of Early Modern English Correspondence Samples), மிச்சிகன் ஆரம்பகால நவீன ஆங்கிலப் பொருட்கள்/தரவுகள் (Michigan Early Modern English Materials) மற்றும் மத்திய ஆங்கில உரைநடை மற்றும் செய்யுளின் தரவுதொகுதி ஆகியவை சொல் அலகுகளின் பொருண்மை மாற்றத்தில் சொல்லாட்சிக் குறிக்கோள் (rhetorical purpose) மற்றும் சூழலின் முக்கியத்துவத்தைச் சரிபார்க்க ஆராயப்படுகின்றன (Lewis 2002). சொல்சார் அலகுகளின் அறிவுசார் மற்றும் மதிப்பீட்டு செயல்பாடுகளை வழங்குவதற்காக பிரதிநிதித்துவ செயல்பாட்டைக் கொண்ட பல்வேறு அளவிடல் தகுதிகள் எவ்வாறு பல்பொருளொருமொழிய வெளிப்பாடுகளாக உருவாக்கப்படலாம் என்பதை இது காட்டுகிறது. வழக்கமான சொல்லாட்சிக் கலை அமைப்பொழுங்குகளில் அவை பயன்படுத்தப்படுவதால், அர்த்தங்களில் பேசுபவரின் கண்ணோட்டம் எவ்வாறு எழுகிறது என்பதையும் இது காட்டுகிறது; இது இறுதியில் அகச்சார்பு ஒப்புமைகள் (local analogies) வழியாக ஒரு பொருண்மை மாற்றத்திற்கு வழிவகுக்கிறது, இது புதிய களங்களில் சொல்சார் அலகுகளை அவற்றின் பயன்பாட்டை நீட்டிக்கக்

கட்டாயப்படுத்துகிறது. வெளிப்பாடுகளின் அளவுசார் ஆய்வு குறிப்பிட்ட சொல்லாட்சித் திறன் அமைபொழுங்குடன் அவற்றின் இணை நிகழ்வு (co-occurrence) புதிய பல்பொருளொருமொழியத்தை உருவாக்குகிறது மற்றும் புதிய தகவல் கட்டமைப்பைப் பெறுகிறது என்பதைக் காட்டுகிறது.

ஆங்கில உரைகளின் ஹெல்சின்கி தரவுதொகுதியின் (Helsinki Corpus of English Texts) தனியார் கடிதங்களின் பெரிய தரவுத்தளம் ஆரம்பகால நவீன ஆங்கிலத்தில் வினைநோக்கு மாற்றத்தின் (modal change) அமைப்பொழுங்குகளைப் ஆய்வதற்கும், இன்றைய ஆங்கிலத்தில் பயன்படுத்தப்படும் வினைநோக்குடன் (modality) ஒப்பிடுவதற்கும் பயன்படுத்தப்படுகிறது (Plaza 2002). சுவாரஸ்யமாக, தரவுதொகுதியிலிருந்து சேகரிக்கப்பட்ட வினைநோக்குக் கூறுகள் இலக்கணமயமாக்கலின் ஒரு இருகாலம்சார் செயல்முறையை எடுத்துக்காட்டுகின்றன; அதிலிருந்து சொல்சார் வினைச்சொற்கள் படிப்படியாக வினைநோக்கு வினைச்சொற்களாக இலக்கண மதிப்புகளைப் பெற்றுள்ளதை அறியலாம். பயன்பாட்டின் இத்தகைய மாற்றம் பழைய மற்றும் புதிய கட்டமைப்புகளுக்கு நெருக்கமாகத் தோன்றும் ஒப்புமைசார் ஆகங்களாலும் (analogic formations) மற்றும் முழுமையற்ற செயல்முறைகளின் பாதுகாப்பின்மையிலிருந்து உருவாக்கப்பட்ட வடிவம்சார் இரண்டகத்தாலும் (formal duplicity) பண்பாக்கம் செய்யப்பட்ட கட்டமைப்பு மிகைக்கு (structural redundancy) உள்ளாக்கப்பட்டுள்ளது.

, Brown Corpus, and LOB Corpus

தொடக்ககால நவீன ஆங்கிலத்தின் ஹெல்சின்கி தரவுதொகுதி (Helsinki Corpus of Early Modern English), பிரவுன் தரவுதொகுதி மற்றும் லாப் தரவுதொகுதி ஆகியவை ஐநூறு ஆண்டுகளில் ஆங்கிலத்தில் LOVE என்பதன் அர்த்தத்தில் ஏற்பட்ட மாற்றங்களை விளக்க ஆய்வு செய்யப்படுகின்றன (Tissari 2000). தரவின் பல பரிமாண பகுப்பாய்வு மூலம், இது அன்பின் ஐந்து வெவ்வேறு கருத்தியல் களங்களை அடையாளம் காட்டுகிறது: 'family love', 'friendship', 'sexual love', 'religious love' மற்றும் 'love of things', இருப்பினும் இந்த வார்த்தையின் பயன்பாட்டில் தெளிவின்மை உள்ளது சில சூழல்களில். இருப்பினும், செயலில் ஈடுபடும் பங்கேற்பாளர்களைப் பற்றிய குறிப்பு, எந்தவொரு இரண்டு வகைகளும் ஒரே நேரத்தில் வார்த்தையால் குறிக்கப்படுவதாகக் கூறுகிறது, இருப்பினும் பங்கேற்பாளர்களைப் பொறுத்தவரை சூழ்நிலை தகவல்கள் வேறுபடலாம். தரவுகளின் எண் பகுப்பாய்வு, 'LOVE' வகைகளின் ஒப்பீட்டு அதிர்வெண் பல ஆண்டுகளாக மாறிவிட்டது என்று கூறுகிறது. 'sexual

love' என்பது பல நூற்றாண்டுகளாக மிகவும் ஆதிக்கம் செலுத்தும் காரணியாக இருந்தாலும், 'family love' மற்றும் 'friendship' ஆகியவை பயன்பாட்டில் குறைவாகவே இருக்கின்றன, அதே நேரத்தில் 'love of things' மற்றும் 'religious love' ஆகியவை பல நூற்றாண்டுகளாக மாறாமல் இருக்கின்றன.

வரலாற்று ஆங்கில ரெஜிஸ்டர்களின் பிரதிநிதி தரவுதொகுதி (Representative Corpus of Historical English Registers) மற்றும் ஆரம்பகால நவீன ஆங்கிலத்தின் ஹெல்சின்கி தரவுதொகுதி (Helsinki Corpus of Early Modern English) ஆகியவை ஆங்கிலத்தில் ஆக்க உருபனியல் (derivational morphology) உற்பத்தித்திறனில் புறமொழியியல்சார்ந்த (extralinguistic) மற்றும் சூழல் சார்ந்த காரணிகளின் பங்கை ஆராய்வதற்கு ஆய்வு செய்யப்படுகின்றன (Cowie 2000). தரவுதொகுதிகளிலிருந்து பெறப்பட்ட சான்றுகள், உருபனியல் உற்பத்தித்திறனின் அதிகரிப்பு மற்றும் குறைவு ஆகியவை ஒரு சமூக கலாச்சார செயல்பாடான புதியசொல்லாக்கத்தால் வலுவாகத் தீர்மானிக்கப்படுகின்றன என்ற வாதத்தைச் சரிபார்க்க இயலச் செய்தது. எவ்வாறாயினும், புதிய சொற்களின் தேவை பயன்வழியல் அடிப்படையிலும் அழகியல் அடிப்படையிலும் உருவாக்கப்படுவதாகக் கண்டுபிடிப்புகள் காட்டுகின்றன. கொள்கை அடிப்படையில் “பேசுவதற்கு புதிய விஷயங்கள் இருக்கும்போது, அவற்றிற்கு பெயரிட புதிய சொற்கள் தேவை. அல்லது சில நேரங்களில், பழைய விஷயங்களைப் பற்றி முற்றிலும் புதிய வழியில் பேச விரும்புகிறோம். சமூகத்தில் ஏற்படும் மாற்றங்கள், அது பொருள்சார்பானதாக இருந்தாலும் அறிவுசார்ந்ததாக இருந்தாலும், புதிய சொற்களுக்கு அழைப்பு விடுகின்றன; மேலும் சமூக மாற்றம் எவ்வளவு தீவிரமாக இருக்கிறதோ, அவ்வளவு புதிய விஷயங்களுக்கு நாம் பெயரிட வேண்டும் அல்லது பழையவற்றை மறுபெயரிட வேண்டும். எனவே, கண்டுபிடிப்பு, கண்டுபிடிப்பு, தேடல் ஆய்வு, போர், வர்த்தகம் மற்றும் புரட்சி அனைத்தும் புத்தாக்கத்தை வளர்க்கின்றன "(Algeo 1991: 14).

'LOOK AT' என்பதன் உட்படுமொழிகள் (troponyms) (எ.கா. stare, gaze, gape, gawp, gawk, goggle, glare, glimpse, glance, peek, peep, peer, squint, leer, gloat, and ogle) சில வரலாற்று தரவுத்தொகுதிகள் அடிப்படையில் வகையில் காட்சிப் புலனுணர்வுக் களத்தில் ஆய்வு செய்யப்படுகிறது (Poch and Clavera 2002). இந்த வினைச்சொற்கள் தோன்றிய பல்வேறு பொருண்மைக் களங்களை இது எடுத்துக்காட்டுகிறது. இது ஒரு களத்திலிருந்து மற்றொரு களத்திற்கு அர்த்தங்களை மாற்ற ஊக்குவிக்கும் காரணிகளில் கவனம் செலுத்துகிறது. இந்த

வினைச்சொற்கள் மட்டுமல்லாமல், காட்சிப் புலன் உணர்வோடு இணைக்கப்பட்ட யாவையும் கண்கள், பார்த்தல் அல்லது நோக்குதல் என்ற அடிப்படை செயல்பாடுகளைத் தவிர, உணர்வுகள், உணர்ச்சிகள் மற்றும் மனப்பான்மைகளையும் வெளிப்படுத்தக்கூடும் என்ற உண்மையை பிரதிபலிக்கிறது என்பது மிகவும் குறிப்பிடத்தக்க உற்றுநோக்கு ஆகும் (Poch and Clavera 2002: 571).

தரவுத்தொகுதி அடிப்படையிலான ஆய்வுகள் பற்றிய குறிப்புரை வரலாற்றுப் பொருண்மையியல் மற்றும் அகராதியலில் குறிப்பிடத்தக்க பங்களிப்புகளைச் செய்ய பல்வேறு வரலாற்று மற்றும் இருகாலத் தரவுத்தொகுதிகள் பயன்படுத்தப்படுகின்றன என்பதைக் குறிக்கிறது. வரலாற்று மொழியியலில் அகராதியின் மறுகண்டுபிடிப்பு பரந்த அளவிலான ஆராய்ச்சி கருவிகளின் வளர்ச்சியிலிருந்து பயனடைகிறது: இருகாலத் தரவுத்தொகுதிகள், அகராதிகள் மற்றும் சொற்களஞ்சியங்கள் (thesauri); இது ஒரு குறிப்பிட்ட மொழியின் பரிணாம வளர்ச்சியை நன்றாகப் பகுப்பாய்வு செய்ய அனுமதிக்கிறது. தரவுத்தொகுதி அடிப்படையிலான அகராதி (மற்றும் வரலாற்று மொழியியல்) ஆய்வு ஆங்கிலம் மற்றும் பிற மொழிகளில் (Hofland and Johansson 1982, Hundt 1997) சில படிகள் முன்னேறியிருந்தாலும், இந்திய மொழிகளில் அதன் பெரிதும் எதிர்பார்க்கப்பட்ட பயணத்தை இன்னும் தொடங்கவில்லை.

7.4. கலைச்சொல் ஆக்கத்தில் தரவுத்தொகுதி (Corpus in Technical Word Formation)

பிற மொழிகளிலிருந்து தமிழில் (அல்லது பிற இந்திய மொழிகளுக்கு) கடன் வாங்கிய தொழில்நுட்ப மற்றும் அறிவியல் சொற்களுக்குக் கலைச்சொல் உருவாக்கம் மற்றும் பொருத்தம் குறித்து விவாதம் நடைபெறுகிறது. தரவுத்தொகுதியிலிருந்து சேகரிக்கப்பட்ட எடுத்துக்காட்டுகள் பல அறிவியல் மற்றும் தொழில்நுட்ப சொற்கள் பல ஆண்டுகளாக மொழியில் கடன் வாங்கப்பட்டுள்ளன என்பதைக் காட்டுகிறது. இவை அவற்றின் தோற்றத்தை கண்டுபிடிப்பது கிட்டத்தட்ட கடினம் என்ற அளவிற்கு சொந்தமொழியாக்கம் (nativization) செய்யப்பட்டுள்ளது.

சுவாரஸ்யமாக, மொழி பயன்பாட்டின் அனைத்து பிரிவுகளிலிருந்தும் உரை மாதிரிகளின் சேகரிப்புகளைக் கொண்ட ஒரு தரவுத்தொகுதி, சொற்களின் உருவாக்கங்களின் புலங்கள் மற்றும் காலங்களைப் பற்றிய ஒருகால மற்றும் இருகால விசாரணைகளை மேற்கொள்வதற்கான கலைச்சொல்லாக்க வல்லுநர்களுக்கான ஒரு புதையல் வீடாகும் (Meyer and Mackintosh 1996).

தரவுத்தொகுதி பல அறிவியல் மற்றும் தொழில்நுட்ப சொற்களை வழங்குகிறது; அவை நிலைபேறுபெற்ற அகராதிகளில் கிடைக்காது. எனவே, இது கலைச்சொல்லாக்க வல்லுநர்களுக்கு

அவர்களின் கலைச்சொல் தரவுவங்கி (terminology databank) மற்றும் அகராதிகளை வடிவமைத்து மேம்படுத்துவதற்கான அனுபவம்சார் தரவுத்தளத்தை (empirical database) வழங்குகிறது. தரவுத்தொகுதி பல்வேறு அறிஞர்களால் வெவ்வேறு காலங்களில் உருவாக்கப்பட்ட சொற்களின் தகுதியை மதிப்பிடுவதற்கான வாய்ப்பையும் வழங்குகிறது.

பயன்பாட்டின் களங்கள் மற்றும் பயன்பாட்டின் தன்மையைப் பொறுத்து விதிமுறைகள் மாறுபடும். ஒரு குறிப்பிட்ட துறையில் ஒரு சிறப்பு அர்த்தத்தைக் கொண்ட ஒரு குறிப்பிட்ட சொல் மற்றொரு துறையில் பயன்படுத்தும்போது அர்த்தத்தில் மாறுபடும். இது இயற்கை மொழியின் உள்ளார்ந்த பண்புக்கூறான சொல்சார் ஆக்கமுறை (lexical generativity) என்ற நிகழ்வுக்கு வழிநடத்துகிறது. எனவே, place 'இடம்' என்பது பல்வேறு களங்களில் அதன் வேறுபட்ட அர்த்தம் காரணமாகப் பின்வருமாறு பயன்படுத்தப்படும்: stadium - விளையாட்டரங்கம், shipyard - கப்பல்தளம், harbor - துறைமுகம், hospital – மருத்துவ மனை, court - நீதிமன்றம், jail - சிறைச்சாலை.

புதிய தொழில்நுட்ப சொற்களை உருவாக்கும் நேரத்தில், கலைச்சொல்லாக்க வல்லுநர்கள் பின்வருவனவற்றை நினைவில் கொள்ள வேண்டும்: (அ) புதிதாக உருவாக்கப்பட்ட சொல் இலக்கணப்படி மொழியில் சீரானதாக இருக்கவேண்டும்; (ஆ) அது குறிப்பிடும் அறிவியல் நிகழ்வு அல்லது பொருளின் கருத்தைக் கொண்டிருக்கவேண்டும்; (இ) அது பொதுவான பயனர்களுக்கு பயன்பாட்டில் எளிதானதாக இருக்கவேண்டும்; (ஈ) உச்சரிப்பதற்கு எளிதாகவும் கேட்பதற்கு நல்லதாகவும் இருக்கவேண்டும்.

முரண்பாடாக, திரட்டப்பட்ட தரவுத்தொகுதி ஆதாரங்கள் புதிய சொற்களின் உருவாக்க விஷயத்தில் இந்த கோட்பாடுகள் அரிதாகவே பின்பற்றப்படுகின்றன என்பதைக் காட்டுகிறது. எடுத்துக்காட்டாக, மொழிபெயர்ப்பாளர்கள் அவசரத்திற்காக பல சொற்களை உருவாக்குவர். இருப்பினும், இவை ஒருபோதும் தரவுத்தொகுதியில் இருப்பதில்லை. மாறாக, பொதுவான மற்றும் எளிதான சொற்கள் அனைத்து உரை வகைகளிலும் மிக உயர்ந்த நிகழ்வுகளைக் கொண்டிருக்கும்.

ஆங்கில *slang* என்பதற்கு தமிழில் மிகவும் பொருத்தமான மற்றும் பொருத்தமான சொல் எது என்று பார்த்தால் அகராதிகளும் தரவுத்தொகுதிகளும் பல சொற்களைத் தரும்: *கொச்சைச்சொல், கொச்சைவழக்கு, கொச்சையான பேச்சு, பச்சையான சொல், பேச்சு வழக்கு, கெட்டவார்த்தை*. முன்னர் பட்டியலிடப்பட்ட சொற்களைக் காட்டிலும் ஸ்லாங் (அதாவது 'slang'-இலிருது எழுத்துப் பெயர்க்கப்பட்டது) சில உரைகளில் பயன்படுத்தப்படுவதைத் தரவுத்தொகுதி

காட்டுகிறது. பின்வரும் வினாக்கள் நம் உள்ளத்தில் ஏற்படும்: ஏற்கனவே உருவாக்கிய சொற்களில் ஏதேனும் ஒன்றை நாம் தேர்ந்தெடுக்கலாமா? மொத்தத் தொகுப்பையும் நிராகரிக்கும் ஒரு புதிய சொல்லை நாம் உருவாக்கலாமா? ஆங்கில ஒலிபெயர்ப்பு வடிவங்களை நாம் ஏற்றுக்கொள்ளலாமா? மொழியில் சொற்களின் பயன்பாட்டின் அதிர்வெண் அடிப்படையிலான ஒரு இணக்கமான தீர்வை நாம் ஏற்கலாம்.

மற்றொரு சுவாரஸ்யமான எடுத்துக்காட்டாக *ice cream*. *ice cream* என்ற சொல்லுக்கு அகராதியிலும் உரைகளிலும் பல விதத்தில் நிகரன்கள் தரப்படுகின்றன: ஐஸ்கிரீம், பனிகம், குளிர்களி, பனிக்குமிழ், பனிக்கூழ், பனிக்குழவு, பனிக்குழைமம், பனிகுழைவு, பனிப்பாலேடு, பனிகுழவு, பனிப்பாகு. தரவுத்தொகுதி *ஐஸ்கிரீம்* என்பதை அதிக அளவில் பயன்படுத்தப்படும் சொல்லாகக் காட்டும். இது நிபுணர்களால் முன்மொழியப்பட்ட சொற்களை விட மிகவும் பொருத்தமானது.

மூல மொழிக்குச் (Soruce Language (SL/எஸ்.எல்.)) சொந்தமான ஒரு குறிப்பிட்ட சொல்லுக்கு இலக்குமொழியில் (Target Language (TL/டி.எல்)) உருவாக்கங்கள் பல்வேறு சமூகவியல் காரணிகளைப் பொறுத்து அமையும்; அவை பயன்பாட்டு களங்களின்படி சரியான சிக்கல்களைக் கட்டுப்படுத்துகின்றன. எ.கா. ஆங்கிலம் 'die' என்பதைக் குறிக்க தமிழில் ஒரு சொல் உருவாக்கப்படும்போது, நபரின் சமூக நிலை, தலைப்பு, குறிப்பு பகுதி போன்ற காரணிகள் உருவாக்கத்தில் குறிப்பிடத்தக்க பங்குகளை வகிக்கின்றன. இவற்றில் அடிப்படையில் தமிழில் சொற்கள் உருவாக்கப்படும்: ஒரு தெய்வீக மனிதருக்கு *பரலோகபதவி அடைதல்* என்ற சொல்லும், ஒரு சிறந்த மனிதருக்கு *உயிர்நீத்தல்* என்ற சொல்லும், ஒரு கௌரவமான மனிதருக்கு *உயிர்துறத்தல்* என்ற சொல்லும் ஒரு சாதாரண மனிதருக்கு *இறத்தல்* என்ற சொல்லும் ஒரு விலக்குக்கு *சாதல்* என்ற சொல்லும் தேர்ந்தெடுக்கப்படலாம்.

ஒரு அர்த்தத்தைக் குறிக்க ஒருபொருள்பன்மொழிகள் இருக்கையில் அதிர்வெண் குறித்த பார்வையைப் பயன்படுத்த இயலும். ஆங்கிலச் சொல்லான algorithm என்பதைக் குறிப்பிட *வழிமுறை, நெறிமுறை, படிமுறை, கணிப்பு நெறி, நிரல் நெறிமுறை* என்ற சொற்கள் பயன்படுத்தப்படுகின்றன. இவை அனைத்தும் algorithm என்ற ஆங்கிலச் சொல்லைக் குறிக்க வல்லுநர்களால் உருவாக்கப்பட்டுள்ளன. *வழிமுறை* பெரும்பாலும் தரவுத்தொகுதியில் பயன்படுத்தப்படுகிறது என்பது கண்டறியப்பட்டுள்ளது. பயன்பாட்டின் அதிர்வெண் மற்றும் சொற்களின் மொழியியல் பொருத்தப்பாட்டைக் கருத்தில் கொண்ட பிறகு, இலக்குமொழியின்

சொல் தரவுத்தளத்தில் அவை ஏற்றுக்கொள்ளப்படுவதற்கோ அல்லது நிராகரிப்பதற்கோ வலியுறுத்தப்படுகின்றது.

7.5. இலக்கண உருவாக்கத்தில் தரவுத்தொகுதி

தரவுத்தொகுதி பல்வேறு வகையான இலக்கண (மற்றும் தொடரியல்) ஆய்வுகளுக்கு அடிக்கடி பயன்படுத்தப்படுகிறது (Halliday 1991). இலக்கண வல்லுநர்களுக்கும் தொடரியல் வல்லுநர்களுக்கும் பல்வேறு கோட்பாடுகளிலிருந்து பெறப்பட்ட முந்தைய கருதுகோள்களைச் சோதிப்பதற்கான மொத்த மொழி வகை மற்றும் உண்மையான அனுபவம்சார் தரவுத்தளங்களைப் பற்றிய அளவு தகவல்களை (quantified information) இது வழங்குகிறது. மொழி பயனர்களின் திறனுக்கான உண்மையான சான்றுகளை விட, இலக்கணத்தைப் பற்றிய ஆராய்ச்சியாளரின் உள்ளுணர்வுகளின் அடிப்படையில் இலக்கணம் மற்றும் தொடரியல் பற்றிய ஆராய்ச்சி மேற்கொள்ளப்படுகிறது. ஆகையால், ஆராய்ச்சியாளர்கள் எந்த முடிவுகளை எடுத்தாலும், அவை எவ்வளவு அருமையாக தோன்றினாலும், இவை உண்மையான பயன்பாட்டில் பிரதிபலிக்கும் செயல்திறனின் சான்றுகளுடன் சரிபார்ப்புக்கு அப்பாற்பட்டவை அல்ல. உண்மையில் நிகழும் எடுத்துக்காட்டுகளுடன் சரிபார்க்கப்படாவிட்டால், ஆக்கமுறை இலக்கணத்தார் (generative grammarians) கூட அனுமானங்களுடன் உடன்பட மறுக்கிறார்கள். உண்மையான அனுபவ சான்றுகள் தரவுத்தொகுதியில் கிடைப்பது, அன்றாட வாழ்க்கையின் போது உண்மையில் மொழி மக்களால் எவ்வாறு பயன்படுத்தப்படுகிறது என்பதை அறிய மொழிப் பயனர்களின் செயல்திறனைப் ஆய்வதைச் சாத்தியமாக்குகிறது (Schmied 1993).

ஒரு மொழியின் எளிமையான விளக்கத்திற்கு தரவைப் பயன்படுத்துவதை விட முறையான இலக்கணக் கோட்பாடுகளின் செல்லுபடியைச் சோதிக்க தரவுத்தொகுதியிலிருந்து நிஜ வாழ்க்கை மொழி தரவை (real-life language data) ஆராய்ச்சியாளர்கள் பயன்படுத்துகின்றனர் (Mair 1991). இலக்கணங்களின் தற்போதைய விவரங்களைக் குறிப்பிட்டு முறையான இலக்கணங்கள் வகுக்கப்படுகின்றன. இந்த இலக்கணங்கள் கணினி பாகுபடுத்தலில் ஏற்றப்படுகின்றன மற்றும் உண்மையான தரவுகளை விதிகள் எவ்வளவு தூரம் கணக்கிட முடியும் என்பதை சரிபார்க்க தரவுத்தொகுதி மூலம் இயக்கப்படுகின்றன. அதன்பிறகு, காணாமல் போயுள்ள அல்லது தவறாகக் காணப்படுகின்ற பகுப்பாய்வுகளைக் கணக்கில் எடுத்துக்கொள்ள இலக்கணங்கள் மாற்றியமைக்கப்படுகின்றன, (Mair 1994, 1996). இவ்வாறு முறையான இலக்கணங்கள் உண்மையான பயன்பாட்டு மொழியைப் பிரதிபலிக்கின்றனவா அல்லது ஆராய்ச்சியாளர்களின்

தற்போக்கான கற்பனையிலிருந்து பெறப்பட்ட எடுத்துக்காட்டுகளைக் குறிக்கும் ஒரு மாதிரியை உருவாக்குகின்றனவா என்பதை சரிபார்க்கப் பரிசோதிக்கப்படுகின்றன.

கடந்த சில ஆண்டுகளில், முறையான இலக்கணங்களைப் பற்றிய சிறிய அளவிலான ஆய்வுகள், கோட்பாடுகளை உறுதிப்படுத்த அல்லது முன்மொழிவுகளை நியாயப்படுத்த தரவுத்தொகுதியிலிருந்து பெறப்பட்ட அளவுசார் பகுப்பாய்வின் (quantitative analysis) முடிவுகளை உள்ளடக்கியது (Mindt 1995). எடுத்துக்காட்டாக, ஓஸ்ட்டிஜ் மற்றும் டிஹான் (Oostdijk and deHaan 1994) பி.என்.சி யிலிருந்து பல்வேறு எச்சத்தொடர் வகைகளின் அதிர்வெண் பற்றிய தகவல்களை ஆங்கிலத்தில் எச்சத்தொடர் அமைபொழுங்குகளை ஆயப் பயன்படுத்தினர். தரவுத்தொகுதியிலிருந்து பெறப்பட்ட தகவல்கள் எவ்வாறு ஆக்கமுறை சட்டங்களை (generative frame) மாதிரியாகக் கொண்ட இலக்கணங்களை எழுத உதவுகின்றன என்பதை இது வெளிப்படுத்துகிறது. இந்த படைப்புகள் 'ஒவ்வொரு (முறையான) இலக்கணமும் ஆரம்பத்தில் உள்ளுணர்வு தரவுகளின் அடிப்படையில் எழுதப்படுகின்றன; கட்டுப்பாடற்ற தரவுத்தொகுதி தரவுகளுடன் (unrestricted corpus data) இலக்கணத்தை எதிர்கொள்வதன் மூலம் அதன் சரியான தன்மை மற்றும் அதன் முழுமை குறித்து பரிசோதிக்க இயலும் ' (Aarts 1991: 48).

இதைத் தொடர்ந்து, இந்திய மொழிகளில் நவீன 'பயன்பாட்டு அடிப்படையிலான இலக்கணங்களை' ('usage-based grammars') எழுதத் தரவுத்தொகுதியைப் பயன்படுத்துவது பற்றி நாங்கள் நினைக்கிறோம். இவை கோட்பாடுகளின் முன்மொழிவில் மட்டுமல்லாமல், விதிகள் மற்றும் கொள்கைகளை உருவாக்குவதிலும் பாரம்பரியமானவற்றிலிருந்து வேறுபடும்.

சிலர் தற்போதுள்ள தமிழ் இலக்கணங்கள் ஆங்கிலம் மற்றும் சமஸ்கிருத இலக்கணங்களின் கலவையைத் தவிர வேறொன்றுமில்லை என்றும் அவை தமிழைப் பற்றி பேசவோ அல்லது நாம் பயன்படுத்தும் மொழியைப் பிரதிபலிக்கவோ இல்லை என்று வாதிடுகின்றனர். தற்போதுள்ள தமிழ் இலக்கணங்கள் நாம் உண்மையில் பயன்படுத்தும் மொழியை உண்மையாக பிரதிபலிக்கத் தவறிவிட்டதால், மொழி பற்றிய அறிவைப் பெறுவதற்கு அவை தேவையான தகவல்களை வழங்காததால், தரவுத்தொகுதியை நோக்கீடுசெய்ய நாம் தயங்கக்கூடாது.

தரவுத்தொகுதி உண்மையில் மக்கள் பயன்படுத்தும் மொழியை உண்மையாகப் பிரதிபலிப்பதால் புதிய இலக்கணத்தை உருவாக்குவதற்கு தரவுத்தொகுதி பயனுள்ளதாக இருக்கும். புத்திசாலித்தனமாகப் பயன்படுத்தினால் மொழி செயலாக்கம் மற்றும் உருவாக்கம்,

இயந்திர மொழிபெயர்ப்பு மற்றும் மொழி கற்பித்தல் ஆகியவற்றிற்கு இலக்கணம் பயனுள்ளதாக இருக்கும்.

7.6. பொருண்மையியல் ஆய்வில் தரவுத்தொகுதி

தரவுத்தொகுதி அடிப்படையிலான மொழி ஆய்வு, மொழியியல் பொருட்கள் மற்றும் உரைக் கூறுகளின் பொருண்மைசார் ஆய்வுக்கு ஒரு புறநிலை அணுகுமுறையை (objective approach) நிறுவுவதற்கு பங்களிக்கிறது. இது சொல்சார் அலகுகளின் (அல்லது எந்த உரைக் கூறின்) உறுதியான பொருண்மை அவை உண்மையில் நடக்கும் சூழலில் இருந்து பெறப்பட்டதாகும் (Schütze 1997: 142) என்று கருத்துரைக்கின்றது.

ஒரு சொல்சார் அலகு அல்லது உரைக் கூறின் பொருண்மை அதன் முந்தைய நிகழ்வின் வரலாற்றை உள்ளடக்கியது; அது அதன் பொருண்மை மற்றும் அது உருவாக்கப்பட்ட பகுதிகளின் பொருண்மை பற்றி அங்கு கூறப்பட்ட அனைத்தையும் உள்ளடக்கியது. சிறந்த அர்த்தங்களின் தடங்கள் (எ.கா. உணர்பொருள், நடையியல், பயன்வழியியல், போன்றவை) மற்றும் குறிநிலைமொழியியல் கூற்றுக்களில் (metalinguistic statements) விவாதிக்கப்பட்ட பொருண்மையியல் பண்புக்கூறுகள் சூழல்களிலிருந்து பிரித்தெடுக்கப்படலாம் என்று வாதிடப்படுகிறது (Teubert 2000). பொதுவாக, இந்தப் பொருண்மைகள் ஒரு சொல்சார் அலகு அல்லது உரைக் கூறின் பொருண்மையை விவரிக்கும் உரையாகச் சுருக்கப்பட்டு மற்றும் பொழிப்புரை செய்யப்படுகின்றன.

பல்வேறு மொழியியல் சொற்களுக்கு அர்த்தங்களை ஒதுக்குவதற்கான புறநிலை அளவுகோல்களை வழங்கத் தரவுத்தொகுதியிலிருந்து பெறப்பட்ட தகவல்கள் எவ்வாறு பயன்படுத்தப்படுகின்றன என்பதை நிரூபிக்க முயற்சிகள் உள்ளன (Mindt 1991). பொருண்மையியல் பகுதியில், மொழியியலாளரின் சொந்த உள்ளுணர்வு மற்றும் ஒரு மொழியைப் பற்றிய அறிவு ஆகியவற்றைக் கொண்டு சொற்களின் பொருண்மைகள் விவரிக்கப்படுகின்றன.

தரவுத்தொகுதியின் பகுப்பாய்வு, சொற்களின் பொருண்மைசார் வேறுபாடுகள் உருபனியல், தொடரியல், மீக்கூறு மற்றும் மரபுத்தொடர் சட்டங்களில் வெளிப்படுத்தப்பட்ட பல்வேறு பண்புரீதியாகக் காணக்கூடிய சூழல்களுடன் உரைகளில் தொடர்புபடுத்தப்பட்டுள்ளது என்பதைக் காட்டுகின்றது. இதேபோல், கூட்டுச் சொற்கள், பல்வேறு பல்சொல் அலகுகள், சொல்லடி வகைப்பாடுகள் மற்றும் தொகுப்பு சொற்றொடர்கள் ஆகியன பொருண்மைசார் பகுப்பாய்வு மற்றும் புரிதலுக்கான தரவுத்தொகுதி அடிப்படையிலான சூழல் தகவலுக்குத்

தகுதியானவை. பல்வேறு மொழியியல் கூறுகள் நிகழும் சூழல்களைக் கருத்தில் கொள்வது பொருண்மைசார் வேறுபாட்டைக் கட்டமைக்க புறநிலை அனுபவம்சார் விளக்கத்தை வழங்கக்கூடும் என்பதையே இது குறிக்கிறது. பொருண்மையியலில் தரவுத்தொகுதியின் மற்றொரு முக்கிய பங்கு சொல்சார் பொருண்மையியலில் (Leech, Francis, and Xu 1994) 'தெளிவில்லாத வகை' ('fuzzy category') (நிச்சயமற்ற தன்மை indeterminacy) மற்றும் 'சாய்வு' ('gradience') என்ற கருத்தை இன்னும் உறுதியாக நிறுவுவதற்கான அதன் முயற்சியில் ஒப்புக் கொள்ளப்படுகிறது.

தத்துவார்த்த மொழியியலின் (theoretical linguistics) கட்டமைப்பிற்குள், சொல்சார் ஐட்டங்களின் வகைகள் வழக்கமாக நிலையானவை: ஒரு சொல்சார் ஐட்டம் ஒரு குறிப்பிட்ட வகையைச் சேர்ந்தது, அல்லது அது இல்லை. இருப்பினும், வகைப்படுத்தலில் உளவியல் சோதனை புலனறிவு வகைப்பாடுகள் பொதுவாக 'நிலையானவை' அல்ல, ஆனால் 'தெளிவில்லாத' எல்லைகளைக் ('fuzzy' boundaries) கொண்டிருக்கின்றன என்று கூறுகின்றன. ஆகையால், ஒரு சொல்சார் ஐட்டம் ஒரு வகையைச் சேர்ந்ததா அல்லது மற்றொன்றுக்கு சொந்தமானதா என்பது கேள்வி அல்ல, ஆனால் மற்றொன்றுக்கு மாறாக இது ஒரு சொல்சார் வகைப்பாட்டுக்குள் எவ்வளவு தடவை அடிக்கடி விழுகிறது என்பது கேள்வியாகும்.

தரவுதொகுதியில் உள்ள ஆதாரங்களை அனுபவபூர்வமாகப் பார்ப்பதன் மூலம், 'தெளிவில்லாத' மாதிரி ('fuzzy' model) தரவுகளுக்கு மிகவும் சிறந்தது என்பது தெளிவாகிறது, ஏனெனில் சொற்களின் வகைகளுக்கு இடையே தெளிவான எல்லை இல்லை. மாறாக, உறுப்பினர்களின் 'சாய்வு' உள்ளன, அவை வழக்கமாக மொழியில் உள்ள சூழல்களில் அவற்றின் பயன்பாட்டின் அடிப்படையில் ஒரு குறிப்பிட்ட வகையை நோக்கி 'சேர்ப்பதற்கான அதிர்வெண்' உடன் இணைக்கப்படுகின்றன.

தரவுத்தொகுதியின் பயன்பாடு சொல்சார் பல்பொருள் ஒருமொழியத் (lexical polysemy) தன்மையைப் புரிந்துகொள்ள உதவுகிறது, இதன் மூலம் சொற்களின் சூழ்நிலை பயன்பாட்டின் மாறுபாடு காரணமாக பல அர்தங்களைக் குறிப்பிட இயலும் (Ravin and Leacock 2000, Bouillon and Busa 2001, Cuyckens and Zawada 2001). ஒரு பெரிய தரவுத்தொகுதியில் காட்டப்படும் சொற்களின் அர்த்த வேறுபாடுகளின் எண்ணிக்கை சாதாரண மனிதனால் உருவாக்கப்பட்ட அகராதிகளில் தரப்பட்ட அர்த்த வேறுபாடுகளின் எண்ணிக்கையை விட அதிகமாக உள்ளது (Fillmore and Atkins 2000, Kilgarriff 2001). மேலும், அவற்றின் அனைத்து பயன்பாட்டு

வகைகளும் தரவுத்தொகுதியிலிருந்து பிரித்தெடுக்கப்பட்டு அவற்றின் பயன்பாட்டின் சூழலுடன் நெருக்கமான குறிப்புடன் பகுப்பாய்வு செய்யப்பட்டால் (Dash 2004) சொற்களின் பல்பொருள்ஒருமொழிய அர்த்தங்கள் சிறப்பாக புரிந்து கொள்ளப்படும். இது இயந்திர மொழிபெயர்ப்பிற்குப் பயன்படுத்தப்படும் இருமொழி மொழிபெயர்ப்பு தரவுத்தொகுதிகளில் (bilingual translation corpora) மொழிபெயர்ப்பு-நிகரன்களுக்கான பொருத்தங்களைக் கண்டறிய சொற்களின் பல அர்த்தங்களின் மாறுபாடுகளை ஈட்டவும் சொற்களின் உண்மையான அர்த்தங்களை அடையாளம் காணவும் உதவுகிறது.

இயற்கையான மொழி செயலாக்கத் திட்டங்களில் சொற்களின் பொருண்மை மயக்கம் என்பது மைய அக்கறை (Schütze 1998). சொல்வலையில் (WordNet) (Miller et al. 1990) முன்மொழியப்பட்ட அணுகுமுறை கணிசமான அளவு சிக்கல்களுக்கு வழிவகுக்கிறது என்பது பெருகிய முறையில் தெளிவாகத் தெரிகிறது, ஏனெனில் பல்பொருளுமொழியச் சொற்களின் அர்த்தங்கள் ஒன்றுக்கொன்று எவ்வாறு தொடர்புபடுகின்றன என்பது பற்றி எதுவும் கூறவில்லை.

சொற்களின் புதிய பயன்பாடுகள் அடிக்கடி நிகழும்போது மற்றும் சொற்களின் ஒற்றை சட்டகத்திற்குள் உருவக அர்த்தங்கள் அடையாளப்படுதப்படும்போது சிக்கல் மேலும் தீவிரமடைகிறது.

இயல்பான கருத்தாடலில் (normal discourse) உருவக மொழி (figurative language) பரவலாக இருப்பதால், உருவகமாகப் பயன்படுத்தப்படும் சொல்லின் மூலப் பொருள் பெரும்பாலும் விரும்பப்படும் இலக்கு அர்த்தத்திலிருந்து வெகு தொலைவில் இருக்கும். இந்தச் சிக்கலை சமாளிப்பதற்கான ஒரு சாத்தியம் அனைத்து வெவ்வேறு அர்த்தங்களையும் பட்டியலிடுவது மட்டுமல்ல, குறைவான அர்த்தங்களைக் கொண்டிருப்பது மற்றும் புதிய அர்த்தங்களை உருவாக்குவதற்கும் அவற்றுக்கிடையிலாயான உறவுகளை கையளுவதற்கும் ஒரு ஆக்கப் பொறிமுறையைப் பயன்படுத்துதல் (generative mechanism) ஆகும் (Pustejovsky 1991). 'ஆக்கமுறை அகராதி' ('Generative Lexicon') (Pustejovsky 1995) திட்டம், அகராதிக்கு ஒரு கட்டமைப்பை ஒதுக்குவதை நோக்கமாகக் கொண்டுள்ளது மற்றும் சூழல்களில் சொற்களின் வெவ்வேறு அர்த்தங்கள் எவ்வாறு இணைக்கப்படுகின்றன என்பதைத் தீர்மானிக்கும் மிகச் சிறந்த வளமான பிரதிநிதித்துவத் திட்டத்தை (representation schema) உருவாக்குகிறது.

ஆகுபெயரின் சில எளிய நிகழ்வுகளுடன் சில வெற்றிகள் எட்டப்பட்டாலும், அணுகுமுறை உருவகங்களை எவ்வாறு சமாளிக்கிறது/கையாளுகிறது என்ற கேள்வி இன்னும் விவாதத்திற்குத்

திறந்திருக்கிறது. மேலும், மொழியில் ஆகுபெயர் மற்றும் உருவகம் ஆகியவற்றுக்கு இடையேயான வேறுபாடுகளைச் செய்ய இது தவறிவிட்டது. சொல் அர்த்தங்கள் அவற்றின் உருவக அல்லது ஆகுபெயர் அர்த்தங்களைப் பற்றிய தகவல்களுடன் குறிக்கப்படவில்லை என்பதால், நேரடி அர்த்தத்தை நேரடி அல்லாத அர்த்தத்திலிருந்து எவ்வாறு வேறுபடுத்துவது என்பது பற்றிய கேள்வி ஒரு முக்கியமான பிரச்சினையாக மாறுகிறது; இது மொழி தரவுத்தொகுதியிலிருந்து சேகரிக்கப்பட்ட சொற்களின் விரிவான பகுப்பாய்விற்கு வாதிடுகிறது. சொற்களின் உருவப் பயன்பாடு அனைத்து வகையான கருத்தாடல்களிலும் பரவலாக உள்ளது, இது மொழியியல், உளவியல், செயற்கை நுண்ணறிவு (artificial intelligence) மற்றும் தத்துவம் உள்ளிட்ட பல்வேறு துறைகளில் இருந்து கணிசமான கவனத்தை ஈர்க்கிறது. உண்மையான பயன்பாட்டின் குறிப்பு மற்றும் பகுப்பாய்வு இல்லாமல் பெரும்பாலான படைப்புகள் தூய மொழியியல் உள்ளுணர்வால் வழிநடத்தப்படுகின்றன. எனவே மொழியில் சொற்களின் உருவகப் பயன்பாட்டை, தரவுத்தொகுதியில் அவற்றின் பயன்பாட்டில் கவனம் செலுத்துவதன் மூலம் அவற்றை ஆராய்வது கட்டாயமாகும்; ஏனெனில் ஒரு தரவுத்தொகுதி சொற்களின் உருவகப் பயன்பாட்டின் பின்வரும் வழிமுறைச் சிக்கல்களை ஆராயத் தேவையான தகவல்களை வழங்குகிறது.

- தரவுத்தொகுதி என்பது நேரடித்தன்மை, உருவகம், ஆகுபெயர், பல்பொருளொருமொழியம், சூழல்-கட்டுண்ட பொருண்மை மற்றும் உருவக மொழியுடனான அவற்றின் உறவுகள் ஆகியவற்றின் கருத்துக்களை வெளிச்சமாக்கும்.
- உருவப் பயன்பாடு, உருவகம், ஆகுபெயர் போன்றவை அமைக்கப்பட்டுள்ள இடை-விளக்கக்குறிப்பான் ஒப்பந்தத்தைப் (inter-annotator agreement) புரிந்துகொள்ள தரவுத்தொகுதி தேவையான தகவல்களை வழங்கும்.
- தரவுத்தொகுதி உருவக மொழிக்கான, அவற்றின் அதிர்வெண்கள், நம்பகத்தன்மை மற்றும் மதிப்பீடு ஆய்வுகள் உள்ளிட்ட குறிப்பிட்ட மொழியியல் குறிப்புகளை வழங்கும்,.
- தரவுத்தொகுதி அடிப்படையிலான ஆய்வுகள் ஒரு மொழியில் சொற்களின் உருவக அம்சங்கள் பற்றிய களம், இனம் அல்லது தரவுத்தொகுதி வகையின் விளைவுகளைக் கண்டுபிடிக்கும்.

- தரவுத்தொகுதியிலிருந்து பெறப்பட்ட முடிவுகள் உருவகப் பயன்பாட்டு விளக்கம் மற்றும் / அல்லது ஆக்க விஷயத்தில் பயன்படுத்தப்படும் கணக்கீட்டு/கணினிசார் மாதிரிகளை மதிப்பீடு செய்யத் தேவையான வழிகாட்டுதலை வழங்கும்.
- தரவுத்தொகுதியிலிருந்து கிடைக்கப்பெறும் முடிவுகள் உருவக மொழிச் செயலாக்கத்தின் உளவியல் மாதிரிகளை உருவாவதற்குக் குறிப்பிடத்தக்க பங்களிப்பை வழங்கும்.

7.7. மொழிக் கற்றலில் தரவுத்தொகுதி

தரவுத்தொகுதி பல வழிகளில் மொழி கற்பிப்பதில் பயன்படுத்தப்படுகிறது (Levy 1997, Wichmann 1997, Botley, McEnery, and Wilson 2000, Granger, Hung, and Tyson 2002). பொதுவாகத் தரவுத்தொகுதியிலிருந்து தொகுக்கப்பட்ட எடுத்துக்காட்டுகள் மொழி கற்பவர்களுக்கு நிஜ வாழ்க்கைச் சூழ்நிலைகளில் தொடர்பு கொள்ளும்போது அவர்கள் சந்திக்கும் மொழி வகைகளுக்குப் (அதாவது வாக்கியம், சொற்றொடர்கள் போன்றவை) பயிற்சி அளிக்கப் பயன்படுகின்றன. அனுபவம்சார் கற்பித்தலுக்கான தரவுகளின் ஆதாரமாக இருப்பதைத் தவிர, தற்போதுள்ள மொழி கற்பித்தல் பொருட்களையும் (language teaching materials) விமர்சன ரீதியாகப் பார்க்கத் தரவுத்தொகுதி பயன்படுத்தப்படுகிறது.

பாடப்புத்தகங்கள் கற்பிப்பதற்கும், சொந்த மொழி பேசுபவர்கள் உண்மையில் மொழியை எவ்வாறு பயன்படுத்துகிறார்கள் என்பதற்கும் கணிசமான வேறுபாடுகள் இருப்பதாக ஆய்வுகள் கண்டறிந்துள்ளன (Ghadessy, Henry, and Roseberry 2001). மொழி கற்பித்தலில் பயன்படுத்தப்படும் பாடப்புத்தகங்களில் சில கண்டுபிடிக்கப்பட்ட (செயற்கையாக உருவாக்கப்பட்ட) எடுத்துக்காட்டுகள் உள்ளன; அவை கண்டுபிடிப்பாளர்களின் உள்ளுணர்வு அல்லது இரண்டாம் நிலை விளக்கங்களை அடிப்படையாகக் கொண்டவை. இந்த எடுத்துக்காட்டுகள் பயன்பாட்டின் முக்கியமான பண்புக்கூறுகளைக் கவனிக்கத் தவறிவிட்டன அல்லது மிகவும் பொதுவானவற்றுப் பதிலாக குறைவான நிகழ்வெண்ணைக் கொண்ட நடையியல்சார் தேர்வுகளை முன்னணிப்படுத்தியுள்ளன எனலாம்.

கொள்கை அடிப்படையில், மொழி கற்பித்தல் பொருட்கள் (language teaching materials) தரவுத்தொகுதி அல்லது நிஜ வாழ்க்கை மொழித் தரவுகளின் பிற மூலங்களிலிருந்து எடுத்துக்காட்டுகள் மற்றும் விளக்கங்களுடன் வெளிப்படையாக அனுபவபூர்வமாக இருக்க வேண்டும்; இதன் காரணமாகப் பயன்பாட்டின் பொதுவான தேர்வுகளுக்குக் குறைந்த பொதுவான தேர்வுகளை விட அதிக கவனம் செலுத்தப்படுகின்றன (Kübler 2002).

வகுப்புச் செயல்பாட்டிற்காகத் தரவு நேரடியாக மாணவர்களுக்கு தரப்படலாம் அல்லது கற்பித்தல் பொருட்களைத் தயாரிப்பதில் பயன்படுத்தப்படலாம். "தரவுத்தொகுதிகள், கற்பவர் அறியவேண்டிய ஒரு மொழியின் அமைப்பொழுங்குகளின் பரப்பெல்லையை மட்டுமல்லாமல், பாடப்பொருட்கள் உருவாக்கம் மற்றும் பாடத்திட்ட வடிவமைப்பில் ஒரு முக்கிய காரணியாக அமையும் அவற்றின் அதிர்வெண்ணையும் வெளிப்படுத்துகின்றன." எனப் பார்லோ (Barlow 1996) வாதிடுகிறார்,

தரவுத்தொகுதியை அடிப்படையாகக் கொண்ட மொழிக் கற்பித்தலுக்கான நூல்கள் வழக்கமாக மிகவும் அரிதான எடுத்துக்காட்டுகளுக்குப் பதிலாக, விளக்கத்தில் மிகவும் துல்லியமாகவும் கற்பிப்பதில் பயனுள்ளதாகவும் இருக்கும் பொதுவான மற்றும் அடிக்கடி நிகழ்கிற எடுத்துக்காட்டுகளைக் குறிப்பிடுகின்றன. பாரம்பரிய முறையில் கூட, சொல் பயன்பாட்டின் வடிவங்களைப் புரிந்து கொள்வதில் சொல்சார் சொல்லடிவகைப்பாடின் பயன்பாடு குறித்த தகவல்களை வழங்க தரவுத்தொகுதி பயனுள்ளதாக இருக்கும் (Gavioli 2004).

தரவுத்தொகுதி கற்பவர்களுக்கு மொழி பயன்பாட்டின் பல்வேறு பண்புக்கூறுகளைப் புரிந்துகொள்ள உதவுகிறது: அதாவது வாக்கியங்களில் மரபுத்தொடர்களைப் பயன்படுத்துவதைக் கட்டுப்படுத்தும் கொள்கைகள்; சொல் பயன்பாட்டின் அமைப்பொழுங்கு மற்றும் அவற்றின் பொருண்மை உறவுகள் தொடர்பான விதிகள்; சொற்றொடர், எச்சத்தொடர் மற்றும் வாக்கியங்கள் உள்ளிட்ட பல்வேறு கட்டுமானங்களின் புறக் கட்டமைப்பை உள்ளுறை செய்யும் லெக்ஸிஸ் மற்றும் இலக்கணத்தின் வலை அமைப்பு; தொகுப்புச் சொற்றொடர்கள், சொற்றொடர்கள் மற்றும் மரபுத்தொடர் வெளிப்பாடுகளின் சூழல் அடிப்படையிலான பயன்பாடு; ரெஜிஸ்டர்கள் மற்றும் உரை வகைகளுக்குக் குறுக்காக மொழிப் பயன்பாட்டின் மாறுபாடு; மற்றும் முதன்மை மற்றும் மேம்பட்ட மட்டங்களில் கற்பவர்களின் மொழித் திறன்களின் இயல்பான வளர்ச்சியில் பொதுவாக பங்களிக்கும் பிற மொழியியல் விதிகள் (Hunston 2002: 176).

பொதுவான முடிவு என்னவென்றால், உள்ளுணர்வு அடிப்படையிலான மொழி கற்பித்தல் தவறானது; தரவுத்தொகுதி அடிப்படையிலான கற்பித்தல் நம்பகமானதாகவும் நம்பகத்தன்மையுடனும் இருக்கின்றது; ஏனென்றால் இது குறைவான பொதுவான பயன்பாடுகளை விட பொதுவான பொதுவான பயன்பாட்டின் தேர்வுகளுக்குக் கவனம் செலுத்துகிறது.

கணினி உதவியுடன் 'ஊடாடும் மொழி கற்றல்' என்பதிலும் தரவுத்தொகுதி பங்களிப்பு செய்கிறது; இதில் கற்பவர்கள் அணுகல், பயன்பாடு மற்றும் குறிப்புக்காக கணினியில் சேமித்து வைக்கப்பட்டுள்ள தரவுத்தொகுதிகளுக்கு வெளிப்படுத்தப்படுகிறார்கள் (Kettemann and Marko 2002).

இரண்டு வெவ்வேறு குழுக்களை உருவாக்கும் மாணவர்களிடையே சொல்வகைப்பாடு கற்பித்தல் குறித்து ஒரு சோதனை மேற்கொள்ளப்பட்டது: ஒரு குழு தரவுத்தொகுதியைப் பற்றிய குறிப்புடன் கற்பிக்கப்படது; மற்ற குழு தரவுத்தொகுதியைக் குறிப்பிடாமல் பாரம்பரிய விரிவுரை அடிப்படையிலான முறைகள் மூலம் கற்பிக்கப்பட்டது. பாரம்பரிய விரிவுரை அடிப்படையிலான முறைகள் மூலம் கற்பிக்கப்பட்ட மாணவர்களை விட தரவுத்தொகுதியுடன் கற்பிக்கப்பட்ட மாணவர்கள் மிகச் சிறப்பாக செயல்படுகிறார்கள் என்பது இவ்வாய்வின் விளைவாக வெளிப்படுத்தப்பட்டது (McEnery, Baker, and Wilson 1995). மற்றொரு ஆய்வில், வாக்கியங்களின் இலக்கண பகுப்பாய்வின் அடிப்படைகளைப் பற்றி இளங்கலை மாணவர்களுக்குத் தரவுத்தொகுதி அடிப்படையிலான கற்பித்தல் பாரம்பரிய விரிவுரை அடிப்படையிலான நுட்பங்களை விட சிறந்த முடிவுகளைத் தருகிறது (McEnery and Wilson 1996: 105).

முதன்மை 'மொழி கற்பவர்களை' மேம்பட்ட 'மொழி ஆராய்ச்சியாளர்களாக' மாற்றுவதற்கு தரவுத்தொகுதியின் மதிப்பு மகத்தானது (Kirk 2002). மேம்பட்ட மொழி கற்றலில், ஆசிரியர்களால் வடிவமைக்கப்பட்ட வகுப்பு நடவடிக்கைகளுக்குப் பதிலாக, கற்பவர்களுக்கு அவர்களின் விருப்பத்தின் தலைப்புகளில் தங்கள் சொந்த ஆராய்ச்சியை மேற்கொள்ள தரவுத்தொகுதி பயனுள்ளதாக இருக்கும். மேம்பட்ட மாணவர்கள் ஒரு மொழித் தொடர்பான பரந்த அளவிலான தலைப்புகளில் சொந்த ஆராய்ச்சியை மேற்கொள்ள எளிய தேடல் கருவிகளுடன் பல்வேறு நீளம் மற்றும் வகைகளின் பல தரவுத்தொகுதிகளைப் பயன்படுத்துகின்றனர். இது மாணவர்களை தரவுத்தொகுதியில் தேட, கேள்விக்குரிய தரவைப் பற்றிய கருதுகோள்களை உருவாக்க, பல்வேறு கருதுகோள்களைச் சோதிக்க உதவுகிறது. இது தரவுத்தொகுதி பயன்பாட்டின் சரியான முறையைப் பின்பற்றி மாணவர்களின் மொழியியல் திறன்களையும் புரிதலையும் மேம்படுத்துவதன் மூலம் அவர்களை எளிமையான 'மொழி கற்பவர்கள்' என்ற நிலையிலிருந்து 'மொழி ஆராய்ச்சியாளர்கள்' என்ற நிலைக்கு ஊக்குவிக்கிறது.

'தரவு-உந்துதல் கற்றல்' 'data-driven learning' (Johns 1991) என்பதில் தரவுத்தொகுதி பொருண்மையை உய்த்தறிய சூழலைப் பயன்படுத்துவதற்கான பொதுவான திறன்களை மேம்படுத்த பயன்படுத்தப்படுகிறது. தரவுத்தொகுதித் தரவைத் தொடரடைவு வரிகள் அல்லது வாக்கியங்களின் வடிவத்தில் படிப்பதன் மூலம் கற்பவர்கள் மொழியைப் பற்றிய கேள்விகளுக்கு தாமே பதிலளிக்கும் சூழ்நிலைகளை அமைப்பது இதில் அடங்கும்.

'தரவு-உந்துதல் கற்றல்'-இல் (Bernadini 2000) சமீபத்திய முன்னேற்றங்கள் மாணவர்கள் தங்கள் சொந்த தரவுத்தொகுதி விசாரணைகளை வடிவமைக்க அவர்களை ஊக்குவிப்பதன் நன்மைகளை வலியுறுத்துகின்றன. அவ்வாறான நிலையில், நிகழ்ச்சி நிரல் மிகவும் உறுதியாக நிர்ணயிக்கப்படாதபோது, ஒரு தரவுத்தொகுதி மூலம் தேடுவதை கற்றவர்கள் பயன்படுத்திக் கொள்கிறார்கள்; இவ்வாறு அவர்கள் நேரிடும் எந்தவொரு ஆர்வமிக்க கவனித்தலையும் அவர்கள் பின்பற்றுகிறார்கள்.

மொழி கற்றலுக்கான அடித்தளங்களை அமைப்பதை விட, தங்கள் அறிவின் இடைவெளிகளை நிரப்புகின்ற மேம்பட்ட கற்றவர்களுக்கு இந்த 'கண்டுபிடிப்பு கற்றல்' ('discovery learning') மிகவும் பொருத்தமான அமைப்பாகும் (Hunston 2002: 171).

தரவுத்தொகுதியிலிருந்து தொடரடைவுளைப் (concordances) பயன்படுத்தி 'கல்வி நோக்கங்களுக்கான ஆங்கில' (English for Academic Purposes (EAP)) மாணவர்களுக்கு அதிக அளவு சொற்றொகையைக் கற்பிக்க சில சுவாரஸ்யமான சோதனைகள் மேற்கொள்ளப்பட்டன (Cobb and Horst 2001). கட்டுப்படுத்தப்பட்ட சூழலில் மாணவர்கள் பிற முறைகளைப் பயன்படுத்துவதை விட தரவுத்தொகுதியிலிருந்து வடிவமைக்கப்பட்ட தொடரடைவுகளை அணுகும்போது சொற்றொகை ஐடங்களின் பட்டியல்களை மிகவும் வெற்றிகரமாகக் கற்றுக்கொள்ள முடியும் என்பதைச் சோதனை நிரூபிக்கிறது. கற்பவர்களால் பயன்படுத்தப்படும் மொழி மாதிரிகளால் செய்யப்பட்ட தரவுத்தொகுதி பல்வேறு ஆராய்ச்சி நோக்கங்களுக்கான முக்கியமான மூலவளங்கள் ஆகும் (Barlow 2000).

பொதுவாக, அத்தகைய தரவுத்தொகுதியில் சேமிக்கப்படும் தரவு, கற்றவர்கள் பெற்றுள்ள மொழியியல் செயல்திறன் மற்றும் அவர்கள் கற்றலில் உள்ள குறைபாடுகள் பற்றிய நம்பகமான சான்றுகளை வழங்குகிறது. வல்லுநர்களால் மேற்கொள்ளப்பட்ட இத்தகைய தரவுத்தொகுதியின் பகுப்பாய்வு இயல்புநிலை கற்பவர்களின் மொழியியல் திறனை மேம்படுத்துவதற்கும் அவர்களின்

எழுத்து மற்றும் பேசும் திறனை மேம்படுத்துவதற்கும் தேவையான நடவடிக்கைகளை எடுக்க தேவையான உத்வேகத்தை வழங்குகிறது.

உண்மையில், இந்தக் குறிப்பிட்ட குறிக்கோளுடன், கற்றல் ஆங்கிலத்தின் சர்வதேசத் தரவுத்தொகுதி (International Corpus of Learner English (ICLE/ஐ.சி.எல்.இ.) உருவாக்கப்பட்டுள்ளது; இதில் வெளிநாட்டு மொழியாக ஆங்கிலம் கற்ற பல்வேறு நாடுகளிலிருந்து வரும் கற்பவர்கள் தயாரிக்கும் எழுத்துக்களின் சாறுகள் உள்ளன.

தற்போது கற்பவர்கள் மொழியில் செயல்திறனை எவ்வாறு பெற்றுள்ளனர் அல்லது பேச்சு மற்றும் எழுத்து இரண்டிலும் அவர்களின் மொழியியல் திறனில் குறைபாடு உள்ளதா என்பதை அறிவது பகுப்பாய்வின் கீழ் உள்ளது. இது இந்திய மொழிகளில் உள்ள தரவுத்தொகுதிகளுக்கு மொழி கற்பித்தல், ஈட்டல் மற்றும் கற்பவர்களின் மொழியியல் திறன்களை மேம்படுத்துதல் ஆகியவற்றில் நம்பகமான தரவுத்தளங்களாகக் கருதப்படுவதற்கான சாத்தியக்கூறுகள் இருப்பதைக் குறிக்கிறது (Mukherjee 2002).

7.8 கிளைமொழி ஆய்வில் தரவுத்தொகுதி

வட்டாரங்களில் மொழி பயன்பாட்டின் மாறுபாடுகளை ஆய்வதற்கு, அதாவது பல்வேறு புவியியல் பகுதிகளில் பயன்படுத்தப்படும் பேச்சுவழக்குகளை ஆய்வதற்குத் தரவுத்தொகுதிகள் இன்றியமையாதது. இது பொது தரவுத்தொகுதி அல்லாத கிளைமொழித் தரவுத்தொகுதி முதன்மை முக்கியத்துவம் வாய்ந்தது, ஏனெனில் இது முழு அனுபவ நம்பகத்தன்மையுடன் கிளைமொழி மாறுபாடுகளின் சான்றுகளை வழங்குகிறது. புவியியல் பகுதிகளுக்குள் மொழி எவ்வாறு மாறுபடுகிறது, அதாவது பல்வேறு புவியியல் பகுதிகளில் பேசப்படும் மொழிகள் நிலைபேறான வடிவத்திலிருந்து எவ்வாறு வேறுபடுகின்றன என்பதை அறிய கிளைமொழியியல் வல்லுநர்கள் அக்கறை கொண்டுள்ளனர்,.

பல நூற்றாண்டுகளாக, மொழி வகைக்கான வினவல், கிளைமொழியியலை மொழியியலின் அனுபவம்சார் (பயன்பாட்டு) துறைகளில் ஒன்றாகச் செய்கிறது, இது உண்மையான பயன்பாட்டின் தரவுத்தொகுதி பயன்படுத்துவதை விட சோதனைகள் மற்றும் கட்டுப்படுத்தப்பட்ட மாதிரிகளில் கவனம் செலுத்துவதை நோக்கமாகக் கொண்டுள்ளது. பல ஆண்டுகளாக, இது குறிப்பிட்ட வட்டாரப் பேச்சு சமூகங்களின் சொற்றொகை மற்றும் உச்சரிப்பில் மட்டுமே கவனம் செலுத்தி, உருபனியல், தொடரியல், பொருண்மையியல் போன்ற பிற பண்புக்கூறுகளைப் புறக்கணித்தது. தற்போது கிளைமொழித் தரவுத்தொகுதி

கிடைப்பதால், ஆராய்ச்சியின் தற்போதைய போக்கு முழு நெறிமுறை மற்றும் திசை மாற்றத்திலும் உள்ளது. தற்போது ஒரு சில கிளைமொழித் தரவுத்தொகுதிகள் மட்டுமே இருந்தாலும், அனைத்து முக்கிய மொழிகளிலும் கிளைமொழித் தரவுத்தொகுதிகளை உருவாக்க வேகம் ஏற்கனவே எடுத்துள்ளது.

ஒரு கிளைமொழி, அதன் இட மற்றும் சொற்பிறப்பியல் அருகாமையின் காரணமாக, நிலைபேறுபெற்ற வகையுடன் மட்டுமல்லாமல் பிற வட்டார வகைகளுடனும் இணைக்கப்பட்டுள்ளது. இது கிளைமொழியை சமூகமொழியியலின் ஒரு முக்கியமான பகுதியாக ஆக்குகின்றது. குறிப்பிட்ட பேச்சு சமூகத்தால் பயன்படுத்தப்படும் மொழி வகையைப் பற்றிய ஆய்வுடன் கிளைமொழியியல் தொடர்புடையது என்பதால், அவர்கள் பயன்படுத்தும் மொழியில் வெளிப்படும் மக்களின் வாழ்க்கை மற்றும் சமூகத்தைப் பற்றிய குறிப்பு இல்லாமல் இது முழுமையடைய முடியாது. சாராம்சத்தில், மக்களின் பன்முக வாழ்க்கை கிளைமொழியில் பிரதிபலிக்கிறது, இதன் பகுப்பாய்வு உண்மையான பயன்பாட்டின் மொழியின் மாதிரிகளால்/ பதக்கூறுகளால் செய்யப்பட்ட பெரிய மற்றும் பல்பரிமாண தரவுத்தொகுதிகளின் (multidimensional corpora) பகுப்பாய்வு இல்லாமல் ஒருபோதும் முழுமையடையாது. இது கிளைமொழித் தரவுத்தொகுதியின் தேவையை நிறுவுகிறது (மின்னணு அல்லது அச்சிடப்பட்ட வடிவத்தில்); இது மக்களின் வாழ்க்கையின் பல்வேறு பண்புக்கூறுகளை ஆராய்ந்து புரிந்துகொள்ள தேவையான தகவல்களையும் தரவையும் பிரித்தெடுக்க பயன்படுத்தப்படலாம். தவிர, தொகுப்பியல் அகலம் மற்றும் பல்வேறு காரணங்களால், கிளைமொழித் தரவுத்தொகுதி பொதுவாக வாழ்க்கை, மொழி மற்றும் சமூகத்தின் பல்வேறு பண்புக்கூறுகளைப் ஆய்வதற்கும், இலக்கு மக்கள் தொகை குறித்து நம்பகமான தரம்சார்-அளவுசார் (qualitative-quantitative) முடிவுகளை எடுப்பதற்கும் பல்வேறு வகையான தகவல்களை வழங்க வல்லது.

இந்திய மொழிகளின் கிளைமொழிகளைப் பொறுத்தவரையில், வட்டார வகைகளை நுணுக்கமான விவரங்களுக்காக ஆய்வு செய்ய தரவுத்தொகுதிகளை உருவாக்க எந்தவொரு உண்மையான முயற்சியும் எடுக்கப்படவில்லை. தமிழின் கிளைமொழிகளும் அதே நிலையைக் கொண்டிருந்தன. கடந்த நூறு ஆண்டுகளாக தமிழின் கிளைமொழிகளை பல்வேறு இடங்களில் ஆய்வு செய்ய சில முயற்சிகள் இருந்தபோதிலும், இவற்றில் பெரும்பாலானவை குறிப்பிட்ட வட்டார வகையைச் சேர்ந்த கைமுறையாக சேகரிக்கப்பட்ட தகவல்களின் சில பகுப்பாய்வுகளின் எளிய பகுப்பாய்வின் அடிப்படையில் அமைந்தவை. இத்தகைய ஆய்வுகளின் வெளியீடுகள்

பெரும்பாலும் சொற்றொகைப் பட்டியல்கள் அல்லது குறிப்பிட்ட கிளைமொழிகளில் பயன்படுத்தப்படும் சில சொற்களின் அகராதிகளின் உருவாக்கத்தில் விளைந்தன .

கிளைமொழிகளின் இலக்கணங்களை உருவாக்குவதற்கான முயற்சிகள் அல்லது கிளைமொழிகளின் சொல்-தொடரியல் பண்புக்கூறுகளை நிலைபேறுபெற்ற மொழியிலிருந்து வேறுபடுத்துவதற்கான முயற்சிகள் அரிதாகவே உள்ளன. எந்தவொரு ஆய்வும் இதுவரை பெரிய, சீரான மற்றும் பிரதிநிதித்துவ கிளைமொழித் தரவைப் பயன்படுத்தவில்லை என்று உறுதியாக வாதிட இயலும்; இதன் காரணமாக ஆய்வுகள் பெரும்பாலும் மக்களின் வாழ்க்கை, மொழி மற்றும் சமுதாயத்தின் பரந்த அளவிலான காட்சியை வெளிப்படுத்த தவறிவிட்டன. நிச்சயமாக, கிளைமொழிகளின் தரவுத்தொகுதி மிகச் சிறந்த விருப்பத் தேர்வாகும்; இதன் பகுப்பாய்வு கிளைமொழிகளையும் முழு மொழிச் சமூகத்தையும் விளக்கி நிற்கும்.

கிளைமொழி ஆய்வின் தற்போதைய உலகளாவிய சூழ்நிலை தரவுத்தொகுதியை அடிப்படையாகக் கொண்ட அணுகுமுறையை வலியுறுத்துகிறது, ஏனெனில் இது கிளைமொழியியல் வல்லுநர்கள் எதிர்கொள்ளும் பிரச்சினைகளுக்கு மிகச் சிறந்த முடிவுகளையும் முன்னோக்கையும் தருகிறது. கிளைமொழித் தரவுத்தொகுதி எழுத்து மற்றும் பேசு உரைகளிலிருந்து மாதிரிகளை/பதக்கூறுகளை மிகவும் சீரான முறையில் பாதுகாக்கிறது. எழுத்து தரவுத்தொகுதி எழுத்து மூலங்களிலிருந்து பதக்கூறுகளைச்/மாதிரிகளைச் சேமிக்கும் அதே வேளையில், உரையாடல் மற்றும் உரையாடல் ஊடாட்டங்களின் பல்வேறு துறைகளிலிருந்து பெறப்பட்ட பதக்கூறுகளைப்/மாதிரிகளைப் பேசும் தரவுத்தொகுதி பாதுகாக்கிறது. எனவே, இலக்கு சமூகத்தின் மொழியைப் பிரதிபலிக்க இரண்டு வகையான உரையின் விகிதாசார பிரதிநிதித்துவம் மிகவும் பயனுள்ளதாக இருக்கும்.

திட்டத்தின் படி, ஒரு பிரதிநிதி கிளைமொழித் தரவுத்தொகுதி இரண்டு பகுதிகளைக் கொண்டுள்ளது: எழுத்து மற்றும் பேச்சு. ஒவ்வொரு பகுதிக்கும் இரண்டு துணைப்பகுதிகள் உள்ளன: கற்பனை உரைகள் மற்றும் தகவல் உரைகள். குறிப்பிட்ட பேச்சு வகையைச் சேர்ந்த மக்களின் வாழ்க்கை மற்றும் சமுதாயத்தின் அகலம் மற்றும் வகையை கணக்கில் எடுத்துக்கொள்வதன் மூலம் 'மொழி'-இன் ஒரு 'கலப்பு காட்சியை' இது முன்வைக்கிறது (அட்டவணை 7.1).

அட்டவணை 7.1: ஒரு பேச்சுவழக்கு கார்பஸின் முன்மொழியப்பட்ட கலவை		
கிளைமொழித் தரவுத்தொகுதிகள்		
உரைகள்	எழுத்துப் பகுதி	பேச்சுப்பகுதி
கற்பனை	புராணம், புவியியல், கலாச்சாரம், புனைவுகள், வரலாறு, விசித்திரக் கதைகள், நாட்டுப்புறக் கதைகள், நாட்டுப்புறக் வழக்காறு, கட்டுக்கதைகள், பொதுக் கதைகள், பேய் கதைகள், பாடல்கள், காதல் கதைகள், நிகழ்வுகள், குழந்தைப்பாடல்கள், நாடகங்கள், கவிதைகள், புதிர்கள், பழமொழிகள், மரபுத்தொடர்கள், கதைப்பாடல்கள், இறங்கற்பாக்கள் போன்றவை.	நாட்டுப்புறப்பாடல்கள் (எ.கா. சண்டை நாடகங்கள், வாய்வழி கதைகள், கதைகள், புதிர்கள், குழந்தைப்பாடல்கள், கதைப்பாடல்கள், இறங்கற்பாக்கள், கவிதைகள், பேய் கதைகள், காதல் கதைகள், புதிர்கள் போன்றவை)
தகவல்சார்ந்தது	வணிகப் பேச்சுக்கள், சமூக வாழ்க்கை, வர்த்தகம், வரலாறு, மதம், நம்பிக்கைகள், வழிபாட்டு முறைகள், சடங்குகள், சுற்றுச்சூழல், இயற்கை, அரசியல், கலாச்சாரம், இலக்கியம், நடைமுறை, விதிமுறைகள், விவசாயம், பழக்கவழக்கங்கள், விருந்துகள், திருவிழாக்கள், விளையாட்டு, விளையாட்டு, மரபுகள், தொழில்கள், சுகாதாரம், சாகுபடி, சுகாதாரம் போன்றவை.	வணிகம், விவசாயம், மதம், சுற்றுச்சூழல், வரலாறு, இயற்கை, விதிமுறைகள், நம்பிக்கைகள், புவியியல், அரசியல், சமூக விதிகள், அமைப்புகள், வழிபாட்டு முறைகள், மரபுகள், சடங்குகள், பழக்கவழக்கங்கள், சமூகமயமாக்கல், கலாச்சாரம், திருவிழாக்கள், நாட்டுப்புற அறிவியல், சுகாதாரம், விளையாட்டு, விளையாட்டு, சுகாதாரம், வியாதிகள் போன்றவை.

அட்டவணை (7.1), எழுத்துப் பகுதியில் எழுத்து உரையின் மாதிரிகள் இருக்கும்போது, பேச்சுப் பகுதியானது வாழ்க்கையின் அனைத்து அம்சங்களுடனும் தொடர்புடைய மக்களின் அன்றாட உரையாடல் ஊடாட்டங்களிலிருந்து அவர்களின் பேச்சு வகைகளைக் கொண்டுள்ளது என்பதைக் காட்டுகிறது.

எழுத்து பதக்கூறுகள்/மாதிரிகள் பேச்சு வடிவத்தில் பெறப்படுகின்றன. எழுத்து வடிவத்தில் சேமிக்கப்பட வேண்டிய உரைகளைப் பேசவோ அல்லது படிக்கவோ தகவலாளர்கள் கேட்கப்படுகிறார்கள். எனவே, எழுத்து மாதிரிகள் தோல்வியடையும் இடங்களில் குறிப்பிட்ட கிளைமொழி வகையின் மொழியைப் புரிந்துகொள்வதற்கான சிறந்த ஆதாரங்கள் பேசும் உரையின் மாதிரிகள்/பதக்கூறுகள் ஆகும். கிளைமொழித் தரவுத்தொகுதியில் சேமிக்கப்பட்டுள்ள தரவின் முறையான மற்றும் அறிவியல்சார் பகுப்பாய்வுகள் இதற்கு முன் கண்டிராத புதிய சான்றுகளையும் எடுத்துக்காட்டுகளையும் உருவாக்கும்.

இந்தச் செயல்பாடு கடினம் மற்றும் நேரம் எடுக்கும். ஆனால் இதை அடைய வேறு மாற்று இல்லை. இதற்கு நீண்டகால திட்டமிடல், பெரிய அளவிலான நிதி முதலீடு, நன்கு பயிற்சி பெற்ற மொழியியலாளர்கள், நிறுவன முன்முயற்சி மற்றும் கருவிகளைப் பயனுள்ள முறையில் பயன்படுத்துதல் மற்றும் மொழித் தொழிநுட்பத்தின் நுட்பங்கள் என்பன தேவை. இந்த இலக்கு அடையப்பட்டால், மொழியியல் ஆராய்ச்சியில் நீண்டகாலம் புறக்கணிக்கப்பட்ட பகுதி அதன் மரியாதை, கௌரவம் மற்றும் மகிமையுடன் தன்னை நிலைநிறுத்த இயலும். பொது மொழியியல் ஆராய்ச்சி மற்றும் பயன்பாட்டின் தற்போதைய சூழ்நிலையில் கிளைமொழித் தரவுத்தொகுதியின் முக்கியத்துவத்தைப் பற்றி ஒருவர் கேள்வி எழுப்ப இயலும். இதற்குப் பதிலளிக்க, கிளைமொழி ஆய்வு மற்றும் விசாரணையில் தரவுத்தொகுதியின் பயன்பாடு எந்தவொரு கேள்விக்கும் அப்பாற்பட்டது என்று நாம் வாதிட இயலும்; ஏனெனில் கிளைமொழியியல் சமூகமொழியியலில் மிக முக்கியமான நிலையை வகிக்கின்றது.

சமூக அறிவியலின் பெரும்பான்மையான கிளைகள் பல்வேறு பேச்சு சமூகங்களின் வாழ்க்கை, மொழி மற்றும் சமூகத்துடன் நேரடியாகவோ அல்லது மறைமுகமாகவோ தொடர்புடையவை என்பதால், பேச்சுவழக்கின் முக்கியத்துவத்தை புறக்கணிக்க இயலாது. இந்தக் காரணிகளை கணக்கில் எடுத்துக்கொள்வதன் மூலம், பின்வரும் வழிகளில் கிளைமொழித் தரவுத்தொகுதியின் பல பயன்பாடுகள் முன்வைக்கப்படுகிறது.

(அ) கிளைமொழித் தரவுத்தொகுதியில் பாதுகாக்கப்பட்டுள்ள அகல மற்றும் பல்வேறு பிரதிநிதித்துவ மாதிரிகள் இரண்டும் தரவுத்தொகுதி சேகரிக்கப்பட்ட இடத்தில் மொழியின் பொதுவான பயன்பாட்டு முறைகளை உண்மையாகப் பிரதிபலிக்கும். இன்றுவரை, பல்வேறு கிளைமொழிகளிலிருந்து சேகரிக்கப்பட்ட தரவு அத்தகைய தரத்தைக் கொண்டிருக்கவில்லை, ஏனெனில் மாதிரிகள்/பதக்கூறுகள் பெரும்பாலும் அதன் தன்மையில் ஒரு பரிமாணமாக இருப்பதால், ஒன்று அல்லது இரண்டு குறிப்பிட்ட அம்சங்களில் (எ.கா. ஒலியன், சொற்கள், முதலியன) மட்டுமே கவனம் செலுத்துகின்றன. உண்மையில், இத்தகைய தரவு ஒருபோதும் கிளைமொழியின் ஒட்டுமொத்த நிலைநிற்பைக் குறிக்காது. ஆகையால், மக்களின் வாழ்க்கையையும் சமூகத்தையும் அவர்களின் மொழி பயன்பாட்டின் மூலம் அறிந்துகொள்வதற்கும், கிளைமொழிக்கும் நிலைபேறுபெற்ற வகைக்கும் இடையிலான வேறுபாட்டின் பண்புகளை அடையாளம் காண்பதற்கும், பொருள் சார்ந்த ஆய்வுகளுக்காக (object-oriented studies) கைமுறையாக சேகரிக்கப்பட்ட சிறிய தரவுகளின் தரவைக் காட்டிலும் தரவுத்தொகுதியை நம்புவது எப்போதும் நல்லது.

(ஆ) கிளைமொழியில் பயன்படுத்தப்படும் ஒலிகள் மற்றும் ஒலியன்கள் தொடர்பான விவாதங்களுக்கு தேவையான தகவல்களை வழங்குவதைத் தவிர, கிளைமொழித் தரவுத்தொகுதி உருபனியல், சொற்கள், மரபுத்தொடர்கள், வாக்கியங்கள், சொற்றொடர்கள் மற்றும் பிற மொழியியல் பண்புகளை விசாரிப்பதற்கான தரவை வழங்குகிறது. தரவுத்தொகுதி புலனாய்வாளர்களுக்கு மொழிப் பண்புகள் குறித்து பல்வேறு புள்ளிவிவர மதிப்பீடுகளைச் செய்ய உதவுகிறது, இதன் அடிப்படையில் இலக்கு மக்கள் தொகை பற்றி பொதுவான மற்றும் குறிப்பிட்ட முடிவான கருத்துக்கள் கூறப்படுகின்றன. ஆங்கில கிளைமொழிகளின் ஹெல்சின்கி தரவுத்தொகுதி (Helsinki Corpus of English Dialects), வடக்கு அயர்லாந்து பேச்சின் உரை தரவுத்தொகுதி) (Northern Ireland Text Corpus of Speech, லான்காஸ்டர்-ஒஸ்லோ-பெர்கன் தரவுத்தொகுதி (Lancaster-Oslo-Bergen Corpus) மற்றும் லண்டன்-லண்ட் பேச்சுத் தரவுத்தொக்தி (London-Lund Speech Corpus) ஆகியவற்றிலிருந்து பெறப்பட்ட தகவல்கள், ஆங்கில கிளைமொழிகளுக்குள்ளும் மற்றும் ஒருபுறம் கிளைமொழிகளுக்கும் மறுபுறம் 'நிலைபேறான வகை'-க்கும் இடையிலும் ஒற்றுமைகள் மற்றும் வேறுபாடுகளைக் கண்டறியவும் பயன்படுகின்றது (McEnery and Wilson 1996: 110). சம்பந்தப்பட்ட மொழிகளின் கிளைமொழித் தரவுத்தொகுதியுடன் இந்திய மொழிகளுக்கும் இதே போன்ற பணிகள் பரிந்துரைக்கப்படலாம்.

(இ) நாகரிகத்தின் வளர்ச்சி மற்றும் முன்னேற்றத்துடன் கிளைமொழிகளின் பல மொழியியல் பொக்கிஷங்கள் இழக்கப்பட்டுள்ளன. இது அநேகமாக அனைத்து இயற்கை மொழிகளிலும் குறிப்பிடப்பட்ட ஒரு உலகளாவிய போக்கு ஆகும். இதன் விளைவாக, பல பாடல்கள், விசித்திரக் கதைகள், குழந்தைப்பாடல்கள், புனைவுகள், கதைகள், விடுகதைகள், புதிர்கள், கட்டுக்கதைகள், பழமொழிகள், நாட்டுப்புறக் கதைகள் போன்றவை ஒரு காலத்தில் கிளைமொழிகளில் பாதுகாக்கப்பட்டிருந்தன. இவை குறிப்பிட்ட வட்டார மொழி சமூகத்திற்கு மட்டுமல்ல, முழு மொழி சமூகத்திற்கும் ஈடுசெய்ய முடியாத இழப்புகள் ஆகும். தரவுத்தொகுதியில் மீதமுள்ள புதையல்களைப் பாதுகாப்பது அவசியம், இதனால் இவை ஒரே விதியை சந்திக்காது, மேலும் அவை பயன்பாட்டிற்குக் கிடைக்கின்றன. இவை வரலாற்றின் உள்ளடக்கம் மட்டுமல்ல, தற்போதைய வாழ்க்கையையும் சமூகத்தையும் புதிய கண்ணோட்டங்களுடன் வடிவமைப்பதில் பங்களிக்கும் கடந்த காலத்தின் அத்தியாவசிய பொருட்கள் ஆகும்.

(ஈ) வட்டார பேச்சு சமூகத்தின் பல பரிமாண வாழ்க்கையின் மொத்த வர்ணனைப் படமும் கற்பனையான உரை பதக்கூறுகளில்/மாதிரிகளில் மட்டுமே பயன்படுத்தப்படும் மொழி மூலம் எப்போதும் முன்னிறுத்தப்பட்டுள்ளது. மக்களின் தினசரி உரையாடலில் பயன்படுத்தப்படும் தகவல் உரையின் மாதிரிகள்/பதகூறுகள் நமக்குத் தேவை. சமூக அறிவியலின் பிற கிளைகளும் சமமாகப் பயனடைவதால், இது மொழியியலில் மட்டும் மட்டுப்படுத்தப்படவில்லை. சமூக அறிவியலார்கள் கிளைமொழித் தரவுத்தொகுதியைப் பயன்படுத்தி வாழ்க்கை, வாழ்க்கை முறை, ஒரு மொழி சமூகத்தின் கலாச்சாரம் ஆகிய அவர்களின் விசாரணையின் எல்லைக்கு உட்பட்டவற்றின் மொத்தப் படத்தை வரையலாம்.

(உ) ஒரு கிளைமொழி மற்றும் 'நிலைபேறு' வகையின் அடிப்படையிலான வேறுபாட்டின் கோடுகளை வரையக் கிளைமொழித் தரவுத்தொகுதி தேவைப்படுகிறது. பல பழைய சொற்கள், இனச் சொற்கள், கருவிகள் மற்றும் நுட்பங்களின் பெயர்கள், சிறப்புச் சொற்கள், வயது முதிர்ந்த குறியீடுகள் மற்றும் குழு வழக்குகள், மறந்துபோன மரபுச்சொற்கள் மற்றும் சொற்றொடர்கள், அரிதாகப் பயன்படுத்தப்படும் எச்சத்தொடர்கள் மற்றும் சொற்றொடர்கள், பண்டைய அடைமொழிகள்/பட்டப்பெயர்கள் மற்றும் பழமொழிகள் போன்றவற்றைப் பாதுகாக்கும் திறன் இதற்கு உண்டு. 'நிலைபேறு' வகை. உள்ளடக்கம், கலவை, வகை மற்றும் அகலம் ஆகியவற்றில் உள்ள தனித்துவத்தின் காரணமாக, கிளைமொழித் தரவுத்தொகுதி, விளக்கமான, வரலாற்று, ஒப்பீட்டு மற்றும் சமூகவியல் அறிவுக்கு இன்றியமையாத மதிப்புமிக்க தகவல்களை வழங்குகிறது.

(ஊ) ஒரு சில எடுத்துக்காட்டுகளுடன், தமிழ் கிளைமொழிகளைப் பற்றிய பெரும்பாலான ஆய்வுகள் மனித முயற்சியால் கள ஆய்வின் மூலம் திரட்டப்பட்ட சிறிய தரவுகள் அடிப்படையிலானவை. அவை கிளைமொழிகளில் உள்ள சொற்கள் பெரும்பாலும் 'நிலைபேறு' வகைகளிலிருந்து உச்சரிப்பில் வேறுபடுகின்றன என்பதையும் அவ்வாறே ஒரு வட்டாரக் கிளைமொழி மற்றொரு வட்டாரக் கிளை மொழியிலிருந்தும் ஒரு சமூகக் கிளைமொழி மற்றொரு சமூகக் கிளைமொழியிலிருந்தும் உயர்வழக்கு தாழ்வழக்கிலிருந்தும் வேறுபடுகின்றன என்பதை நிறுவ முயன்றன.

வட்டாரக் கிளைமொழி 1	**வட்டாரக்கிளைமொழி 2**
அரை	அறை
கரி	கறி
ஏரு	ஏறு
பேரு	பேறு
பளம்	பழம்
வாளைப்பளம்	வாழைப்பழம்
தமிள்	தமிழ்
மீனவர் பேச்சு	**அரிசனர் பேச்சு**
தேழு	தேளு
ராசாழி	ராசாளி
மஞ்செ	மன்செ
பூதா(ன்)	பூரான்
உயர்வழக்கு	**தாழ்வழக்கு**
இன்று	இண்ணைக்கி
உயர்வு	ஒசத்தி
எண்பது	எம்பளது
பிறகு	பொறகு

உச்சிரிப்பில் மட்டும் அன்றி சொல் தேர்விலும் கிளைமொழிகள் வேறுபடுவதை அவைச் சுட்டுக்காட்டின (வாரியல் – விளக்குமாறு – துடைப்பம்; அன்னாசிப்பழம் – பிருத்திச்சக்கை, பலாப்பழம் – சக்கை). வியக்கத்தக்க வகையில், புலனாய்வாளர்களால் பொருத்தமான சொற்களின்

தொகுப்பை அடிப்படையாகக் கொண்ட எடுத்துக்காட்டுகளின் எண்ணிக்கை பொதுவாக பத்து முதல் பதினைந்து வரை வேறுபடும். புலனாய்வாளர்களின் பங்களிப்பை ஒப்புக்கொள்வதன் மூலம் பின்வரும் கேள்விகளை நாம் எழுப்ப இயலாது: ஒலிப்பு வேறுபாடுகள் மட்டுமே கிடைக்குமா? உருபனியல்சார், சொல்சார், தொடரியல்சார், பொருண்மையியல்சார் வேறுபாடுகள் கிடைக்காதா? இந்தச் சிலவற்றைத் தவிர வேறு வகைகள் இல்லையா? நிச்சயமாக, குறிப்பிடப்பட்டதை விட பல்வேறு நிலைகளில் அதிக நிகழ்வுகளும் பல வேறுபாடுகளும் உள்ளன. ஆனால் பெரிய, பிரதிநித்தவ மற்றும் சீரான தரவுத்தொகுதி இல்லாததால் இவற்றை கண்டுபிடிக்க இயலவில்லை. கிளைமொழித் தரவுத்தொகுதி பகுப்பாய்வு அடிப்படையில் கிளை மொழிகளுக்கு இடையில் மொழி அமைப்பின் எல்லா நிலையிலும் (சொல்லியல், உருபனியல், தொடரியல், பொருண்மையியல்) உள்ள தனித்துவமான பண்புக்கூறுகளைப் பெறலாம்.

(எ) ஒரு கிளைமொழிக்கும் நிலைபேறு வழக்குக்கும் இடையிலான சில உருபனியல்சார் வேறுபாடுகளை அறிஞர்கள் குறிப்பிடுகின்றனர். அவர்களைப் பொறுத்தவரை, நிலைபேறு வழக்கில் இல்லாத முன்னொட்டு, பின்னொட்டு, பின்னொட்டு, வேற்றுமைக் குறிகள் போன்றவற்றின் தொகுப்பு ஒரு கிளைமொழியில் உள்ளது. சந்தேகத்திற்கு இடமின்றி, இந்த கவனிப்புகள் சரியானவை, குறைந்தபட்சம் அவற்றின் விசாரணைக்கு பயன்படுத்தப்படும் தரவுத்தளங்களைப் பொறுத்தவரை. நமக்கு சில கேள்விகள் எழலாம்: இந்தப் பண்புக்கூறுகள் எவ்வாறு கண்டுபிடிக்கப்படுகின்றன - வாய்வழி மாதிரிகள் அல்லது எழுதப்பட்ட மாதிரிகளிலிருந்து? மாதிரிகள்/பதக்கூறுகள் எவ்வாறு சேகரிக்கப்பட்டு பகுப்பாய்வு செய்யப்படுகின்றன? மாதிரிகள்/பதக்கூறுகள் முழு மொழிவகையையும் சரியாகக் குறிக்கிறதா? மாதிரிகளின் ஏற்றுக்கொள்ளல் எவ்வாறு தீர்மானிக்கப்பட்டது? முழு மொழிவகையிலும் தரவு எவ்வளவு தூரம் உண்மையானது? பண்புக்கூறுகள் புள்ளியியல் ரீதியாக முக்கியத்துவம் வாய்ந்தவையா அல்லது அவை அரிதானவை, ஆனால் மொழிவகையில் முக்கியமற்றவையா? வேறு சில வினவல்கள்: பண்புக்கூறுகளின் தொகுப்பு மொத்த பண்புக்கூறுகளின் தொகுப்பையும் குறிப்பிடப்படுகிறதா? மொத்த மக்கள்தொகையை உள்ளடக்கும் அளவுக்கு அவை எண்ணிக்கையில் பெரியவையா? சொல்சார், தொடர்சார் மற்றும் தொடரியல் பண்புக்கூறுகள் பற்றி என்ன? தரவுத்தொகுதியிலிருந்து பிரித்தெடுக்கப்பட்ட தரவைக் கொண்டு கவனிப்புகள் செய்யப்பட்டால் இந்த கேள்விகள் அனைத்தும் பொருத்தமற்றதாக இருக்கும். சாராம்சத்தில், முழுமையான தரவுத்தளங்களின் பற்றாக்குறை பெரும்பாலும் புலனாய்வாளர்கள் தங்களது

கவனிப்புகள் மற்றும் வாதங்களை நிறுவுவதற்கான பொருத்தமான மற்றும் நம்பகமான சான்றுகள் மற்றும் ஆதாரங்களை வழங்குவதிலிருந்து அவர்களைத் தடுக்கும்படி கட்டாயப்படுத்துகிறது.

(ஏ) ஒரு பேச்சுவழக்கின் இலக்கணம் நிலைபெறு வழக்கை ஒத்ததா? அவை வேறுபடுகின்றனவா? அப்படியானால், எப்படி, எங்கே, ஏன் அவை வேறுபடுகின்றன? பொதுவாக, ஒரு கிளைமொழியில் பயன்படுத்தப்படும் இலக்கணம் நிலைபெறு வழக்கிலிருந்து வேறுபட்டது என்று கருதப்படுகிறது. இது சரியான கவனிப்பா? இது ஆதாரங்களுடன் நிரூபிக்கப்பட்டுள்ளதா? அல்லது இது ஒரு ஊகமா? தரவுத்தொகுதித் தரவுத்தளத்தைப் பயன்படுத்தி இரு வகைகளையும் ஒப்பிட்டு ஆய யாராவது முயற்சித்திருக்கிறார்களா? புள்ளிவிவர ரீதியாக வேறுபாடுகள் எவ்வாறு வேறுபடுகின்றன என்பதை ஆராய ஏதாவது முயற்சி செய்யப்பட்டுள்ளதா? இரண்டு வகைகளின் தரவுத்தொகுதியும் முறையாக பகுப்பாய்வு செய்யப்படும்போது, ஒப்பிடுகையில் மற்றும் புள்ளிவிவர ரீதியாக அளவிடப்படும் போது அனைத்து கேள்விகளுக்கும் பதில்களை வழங்க முடியும். மொத்தத்தில், கிளைமொழித் தரவுத்தொகுதியைக் குறிப்பிடாமல் கூறப்படுவது பகுதியளவு கவனிப்பால் செய்யப்பட்ட உண்மையை முடக்கிய பொதுமயமாக்கலைத் தவிர வேறில்லை.

(ஐ) கிளைமொழியியல் பல்வேறு சிக்கல்களைப் புரிந்துகொள்ள கிளைமொழித் தரவுத்தொகுதி இன்றியமையாதது. இடம் மற்றும் நேரம் அடிப்படையில் மொழி எவ்வாறு மாறுபடுகிறது, நேர மாற்றம் அடிப்படையில் ஒரு கிளைமொழியின் சொற்றொகை எவ்வாறு அதிகரிக்கிறது அல்லது குறைகிறது, நேரம் மற்றும் நிகழ்வின் மாற்றத்தால் சொற்களின் பொருண்மை எவ்வாறு மாறுகிறது, ஒரே புவியியல் பிராந்தியத்தில் மொழி எவ்வாறு மாறுபடுகிறது; ஒத்த மொழி வகைகளுக்குள் உள்ள ஒரு குறிப்பிட்ட மொழி வகை எவ்வாறு படிப்படியாக தரநிலையாக அங்கீகரிக்கப்படுகிறது, தரநிலைப்படுத்தலின் செயல்பாட்டில் குறிப்பிட்ட வகையின் மொழியியல் பண்புகள் எவ்வாறு பங்களிக்கின்றன போன்றவற்றைப் பற்றி ஆய்வு செய்ய கிளைமொழித் தரவுத்தொகுதி பயன்படுத்தப்படுகிறது. இந்திய சூழலில், இக்கேள்விகளுக்கான பதில்களை பல்வேறு வட்டார வகைகளின் எடுத்துக்காட்டுகளிலிருந்து தயாரிக்கப்படுகிற பெரிய அளவிலான கிளைமொழி தரவுத்தொகுதிகளின் பகுப்பாய்விலிருது பெற இயலும்.

(ஒ) இறுதியாக, சமூகம் மற்றும் கலாச்சாரம்/பண்பாடு பற்றிய கிளைமொழி அடிப்படையிலான ஆய்வில் ஈடுபடுபவர்கள் கிளைமொழித் தரவுத்தொகுதியை மிகவும் பயனுள்ளதாகக் கருதுகின்றனர். உண்மையில், குறிப்பிட்ட வகைகளில் பயன்படுத்தப்படும் சொற்களின் மொத்த

பட்டியலை அவர்கள் சேகரிக்கும் இடத்திலிருந்து கிளைமொழித் தரவுத்தொகுதி மட்டுமே அவர்களுக்கு ஆதாரமாகும். அகராதியை உருவாக்குவதற்காக அல்லது மொழிக்கும் மக்களுக்கும் இடையில் உள்ளார்ந்த இடைமுகத்தை முன்னிலைப்படுத்த சிறப்புச் சொற்கள், கலைச்சொற்கள், மரபுச்சொற்கள், சொற்றொடர்கள் மற்றும் பழமொழிகளின் எடுத்துக்காட்டுகளை அவர் சேகரிக்கின்றனர்.

இந்தியா போன்ற ஒரு நாட்டில், நம் கவனத்தை பேச்சுவழக்கு கார்பஸின் தலைமுறைக்கு திருப்பி விடுகிறோம். இது நாட்டில் காணப்படும் மொத்த மொழி (இயங்கியல்) வகைகளைப் பாதுகாக்க உதவும். மேலும், கிளைமொழித் தரவுத்தொகுதியின் முறையான பகுப்பாய்வு பல புதிய மொழியியல் நுண்ணறிவுகளை வழங்கும், இதன் மூலம் முழு தேசமும் மொழியியல் ரீதியாகவும் கலாச்சார ரீதியாகவும் பயனடைகிறது.

7.9. சமுதாய மொழியியலின் தரவுத்தொகுதி

சமூகமொழியியல் என்பது மொழியியல் ஆராய்ச்சியின் அனுபவம்சார் துறையாகும் (empirical field). இந்த பகுதியிலுள்ள பெரும்பாலான ஆய்வுகள் மொழிக்கும் பாலினத்துக்கும் இடையிலான இடைமுகத்தைக் கண்டறிய சொல்சார் ஆய்வுகளில் அக்கறை கொண்டுள்ளன. இந்த ஆய்வுகள் ஆராய்ச்சி-குறிப்பிட்ட தரவுத்தளங்களின் (research-specific databases) எல்லைக்குட்பட்ட சேகரிப்பைப் பெரிதும் சார்ந்து இருப்பதால், தரவு அளவு சரிபார்ப்பு அல்லது முறையான மாதிரிக்கு/பதமாதிரிக்கு வைக்கப்படவில்லை. மேலும் முழு தரவுத்தளமும் அதன் இயற்கையான பின்னணியில் இருந்து முழுமையாக வெளிப்படுத்தப்படுகிறது; இதன் காரணமாக உற்றுநோக்கு ஒரு பக்கம் சார்ந்தது அல்லது சமநிலையற்றது ஆகும். இதை சமாளிக்க, தரவுத்தொகுதி இயற்கையான தரவுகளின் பெரிய அளவிலான பிரதிநிதித்துவ மாதிரிகளை வழங்குகிறது, இது அளவுசார் அளவீட்டுக்கும் முறையான மாதிரிக்கும் வைக்கப்படுகிறது.

பேசுபவர்கள், பயனர்கள் மற்றும் படைப்பாளிகளைப் பற்றிய பல்வேறு சமூகமொழியியல் தகவல்களைக் கொண்ட ஒரு தரவுத்தொகுதி என்பது சமூகமொழியியல் ஆராய்ச்சியில் சாத்தியமான ஆதாரமாகும். சமீபத்திய ஆய்வில், பிரவுன் மற்றும் லாப் தரவுத்தொகுதி (Brown and LOB Corpus) அமெரிக்க மற்றும் பிரிட்டிஷ் ஆங்கிலத்தில் (American and British English) 'ஆண்பால் சார்பு' கண்டுபிடிக்க ஆராயப்பட்டன. இந்த இரு தரவுத்தொகுதிகளிலும் 'ஆண் பொருட்களை' விட 'பெண் பொருட்களை' பயன்படுத்துவதற்கான அதிர்வெண் மிகக் குறைவு என்று கண்டறியப்பட்டுள்ளது. இருப்பினும், சுவாரஸ்யமாக, அமெரிக்க ஆங்கிலத்தை விட

பிரிட்டிஷ் ஆங்கிலத்தில் பெண் பொருட்கள் அதிகம் காணப்படுகின்றன (கெல்மர்/ Kjellmer 1986). மற்றொரு சுவாரஸ்யமான ஆய்வில், லண்டன் பதினம வயதினர்களின் தரவுதொகுதி (கார்பஸ் ஆஃப் லண்டன் டீனேஜர்ஸ்/Corpus of London Teenagers COLT) இளம் பருவப் பெண்களிடையே (Stenström and Hasund 1996) வாய்மொழி விவாதங்களின் தன்மை மற்றும் மாறுபாட்டை ஆராய பயன்படுத்தப்பட்டது. பி.என்.சியின் பெரும்பகுதி, சில சமூகவியல் மாறிகளால் (எ.கா. பேசுபவர்களின் வயது, பாலினம், சமூக வகுப்பு, எழுத்தாளர்களின் வயது, பாலினம், இருப்பிடம் போன்றவை) குறிக்கப்பட்டுள்ளது; இது பல சமூகமொழியியல் ஆய்வுகள் மற்றும் புலனாய்வுகளுக்குப் பயன்படுத்தப்படுகிறது (Rayson and Hodges 1997).

சமூகமொழியியல் ஆராய்ச்சியின் மற்றொரு முக்கியமான பகுதி, பேச்சு மற்றும் எழுத்து இரண்டிலும் எப்படி, ஏன் மக்கள் மற்றவர்களுக்கு விஷயங்களை விளக்க முயற்சிக்கிறார்கள் என்பது பற்றிய ஆய்வு. பொதுவாக, விளக்கம் (அல்லது கற்பித்துக்கூறுதல்) என்பது மனித உளவியலில் ஒரு முக்கியமான அளவுகோலாகும், ஏனெனில் மக்கள் பொதுவாக அவர்கள் வாழும் சூழலுடன் ஊடாடும் வழிகளை இது வெளிப்படுத்துகிறது. இதைச் சரிபார்க்க, சமூக உளவியலாளர்களுக்கு 'இயற்கைத் தரவுத்தளங்களை' ('natural databases') அணுக வேண்டும்; அது செயற்கை ஆய்வகச் சூழ்நிலைகளில் மீட்டுருவாக்கம் செய்ய இயலாது. அவர்கள் தரமான தரவை நம்புவதை விட அவர்களின் கோட்பாடுகளை அளவிட மற்றும் பரிசோதிக்க விரும்புவதால் அவர்களுக்கு இயற்கையான தரவு அதிக அளவில் தேவைப்படுகிறது. பேச்சு மற்றும் எழுத்து இரண்டிலும் விளக்கத்தின் பின்னணியில் உள்ள காரணங்களைப் ஆய்வற்காக, சமூகமொழியியலாளர்கள் பொதுவாக இயற்கையாக நிகழும் உரைகளான செய்தித்தாள்கள், டைரிகள், நிறுவனத்தின் அறிக்கைகள், அன்றாட உரைகள், வகுப்பறை நடவடிக்கைகள், காவல் நிலையங்களின் அறிக்கைகள், கேள்வி பதில் அளிக்கும் உரைகள் மற்றும் பிற ஆதாரங்களிலிருந்து தரவைப் பயன்படுத்துகின்றனர்.

ஒரு ஆய்வில், லண்டன்-லண்ட் பேச்சுத் தரவுத்தொகுதியின் (London-Lund Speech Corpus) இரண்டு மில்லியன் சொற்களின் உரையாடல் உரைகளின் ஒரு பகுதி, மக்கள் பயன்படுத்திய because (ஆங்கிலத்தில் பொதுவான காரண இணைப்புக் கிளவி) என்பதன் நிகழ்வுகளை மீட்டெடுக்க பகுப்பாய்வு செய்யப்பட்டது. எடுத்துக்காட்டுகளின் ஆரம்ப உற்றுநோக்கலுக்குப் பிறகு, ஒரு மாதிரி/பதகூறு வகைப்பாடுத் திட்டம் வடிவமைக்கப்பட்டது, இதன் மூலம் அனைத்து விளக்கங்களும் விளக்கப்பட்டுள்ளவற்றின் படி வகைப்படுத்தப்பட்டன

(எ.கா. பேசுபவரின் செயல்கள், விவகாரத்தின் பொதுவான நிலை, பிறரின் செயல்கள் போன்றவை). தரவுத்தொகுதியில் உள்ள விளக்க வகைகளின் இறுதி பகுப்பாய்வு, பொது விவகாரங்களுக்கான விளக்கங்கள் மிகவும் பொதுவான வகை விளக்கமாகும், அதைத் தொடர்ந்து பேசுபவர்கள் மற்றும் பிறரின் செயல்கள் (Antaki and Naji 1987). முந்தைய கோட்பாடுகளை இந்த ஆய்வு மறுக்கிறது; இது முன்மாதிரி வகை விளக்கம் ஒரு நபரின் ஒற்றை நடவடிக்கை என்று வாதிட்டது. இது போன்ற செயல்பாடுகள் கோட்பாடுகளின் பரிசோதனை மற்றும் மாற்றியமைப்பதில் தரவுத்தொகுதியின் ஆற்றலை நிறுவ முடியும்;, இது இயற்கையான அளவிடக்கூடிய மொழித் தரவை வழங்காமல் சாத்தியமில்லை.

மொழி வகைகளுக்கு இடையிலான ஒப்பீட்டு ஆய்வுகளையும், இந்த வகைகளின் விளக்கத்தையும் தொடங்குவதற்கான மதிப்புமிக்க வளமாகவும் தரவுத்தொகுதி அங்கீகரிக்கப்பட்டுள்ளது. பொதுவாக, உரை வகைகள், களங்கள், காலங்கள், பிராந்தியங்கள், பேசுபவர்கள், எழுத்தாளர்கள் போன்றவற்றின் குறுக்காக மொழி எவ்வாறு மாறுபடுகிறது என்பதை ஆய ஒரு வகை மற்றவற்றுடன் ஒப்பிடப்படுகிறது. மாறுபாடுகள் ஒரே தரவுத்தொகுதியின் வெவ்வேறு பகுதிகளாக இருக்கலாம் (எ.கா. தமிழ் தரவுத்தொகுதியில் உள்ள அறிவியல் புனைக்கதை உரைகள் Vs காதல் புனைக்கதை உரைகள்) அல்லது வெவ்வேறு தரவுத்தொகுதிகளின் ஒத்த பகுதிகளாக இருக்கலாம் (எ.கா. இந்தி தரவுத்தொகுதியில் உள்ள அறிவியல் புனைகதை நூல்கள் மற்றும் தமிழ் தரவுத்தொகுதியில் உள்ள அறிவியல் புனைக்கதை உரைகள்). இந்திய மொழிகளின் தரவுத்தொகுதிகள் ஒப்பீட்டின் அளபை அதிகரிக்க ஒத்த மாதிரி/பதக்கூறு நடைமுறையைப் பின்பற்றுவதால், அவை நாட்டின் மொழிகளுக்குள் உள்ளார்ந்த பல்வேறு சிக்கலான சமூகமொழியியல் பண்புக்கூறுகளை ஆய்வதற்கான நல்ல ஆதாரங்கள் ஆகும்.

கருத்தாடலில் தரவுத்தொகுதி அடிப்படையிலான ஆராய்ச்சியின் அளவு ஒப்பீட்டளவில் சிறியது (Graf 1996). முதன்மையாக ஆராய்ச்சிக்கு ஏற்ற பொருள்களைக் கண்டுபிடிப்பது கடினம், ஏனென்றால் இந்தத் துறையானது கருத்தாடலின் உண்மையான சூழல்களை பெரிதும் நம்பியுள்ளது, இது தரவுத்தொகுதிக்குள் அரிதாகவே கிடைக்கிறது. தரவுத்தொகுதிக்குள் சேமிக்கப்பட்ட எழுத்து அல்லது பேச்சு உரையின் மாதிரிகள் சமூக மற்றும் உரை சூழல்களில் இருந்து அகற்றப்படுகின்றன. சில சமயங்களில், தொடர்புடைய சமூகமொழியியல் தடயங்களையும் (எ.கா. பாலினம், வகுப்பு, பிராந்தியம், தொழில், கல்வி, கலாச்சாரம், இனம்,

இருப்பிடம் போன்றவை) பல்வேறு கருத்தாடல் தகவல்களையும் (எ.கா. நிகழ்வு, நேரம், பங்கேற்பாளர்கள், சூழல், நிலைமை, முதலியன) தரவுத்தொகுதியில் குறியாக்கம் செய்ய விரும்பினாலும் உண்மையான சூழல் தகவல்களை அதிலிருந்து ஊகிப்பது எளிதல்ல. எவ்வாறாயினும், அதிகமான தரவுத்தொகுதிகள் தொகுக்கப்பட்டு தொடர்புடைய தகவல்களுடன் சிறுகுறிப்பு செய்யப்படுவதால், இந்தப் பகுதியில் மேலும் தரவுத்தொகுதி அடிப்படையிலான ஆராய்ச்சி செய்யப்படும் என்று நம்பலாம். இந்த பகுதிக்குள் ஒரு சில தரவுத்தொகுதிகள் பயன்படுத்தப்படுகின்றன (Stenström 1994, Andersen 1997); இதன் முக்கிய நோக்கம், தொடர்புடைய உரையாடல் செயல்பாடுகளைக் கொண்ட சொல்சார் ஐடங்கள், மரபுச்சொற்கள் மற்றும் சொற்றொடர்களைப் பொறுத்து உரையாடல் எவ்வாறு செயல்படுகிறது என்பதைப் புரிந்துகொள்வது ஆகும். இருப்பினும், உரையாடல் தரவுத்தொகுதி கிடைப்பது மற்றும் பல்வேறு சமூக மற்றும் புவியியல் பரப்பெல்லைகளுடன் பதிவுசெய்யப்பட்ட தரவுத்தளங்களின் அதிகரிப்பு ஆகியவை கருத்தாடல் மற்றும் பயன்வழியியல் குறித்த இத்தகைய ஆய்வுகளை விரிவுபடுத்துவதற்கு அதிக நோக்கங்களை வழங்குகிறது.

ஒப்பிடக்கூடிய தரவுத்தொகுதியின் இருப்பு வெவ்வேறு மாநிலங்களிலும் நாடுகளிலும் மொழி பயன்பாட்டை ஒப்பிட்டுப் பார்க்க இயலச் செய்கின்றது, ஏனெனில் அவை அந்தந்த மொழிப் பயனர்களின் கலாச்சார வேறுபாடுகளைத் தீர்மானிக்க பகுப்பாய்வு செய்யப்பட்டுள்ளன (Lovejoy 1995). லோப் தரவுத்தொகுயின் தொகுப்பிற்குப் பிறகு, அதன் சொற்றொகையை பிரவுன் கார்பஸுடன் ஒப்பிடுவது ஆரம்பகால படைப்புகளில் ஒன்றாகும். இது பல சுவாரஸ்யமான வேறுபாடுகளை வெளிப்படுத்தியது; இது எழுத்துப்பிழை, உருபனியல் அல்லது சொற்கள் (Leech and Fallon 1992) போன்ற தூய மொழியியல் அடிப்படையிலானவற்றைத் தாண்டிச் சென்றது. இதற்கு முன்னர் கண்டிராத இரு நாடுகளுக்கும் (இங்கிலாந்து மற்றும் அமெரிக்கா) இடையிலான பல கலாச்சார வேறுபாடுகளை ஆய்வு வெளிப்படுத்தியது. எடுத்துக்காட்டாக, சுற்றுப்பயணம் மற்றும் பயணம் தொடர்பான சொற்களின் எண்ணிக்கை பிரிட்டிஷ் ஆங்கிலத்தை விட அமெரிக்க ஆங்கிலத்தில் அடிக்கடி காணப்படுகிறது, இது அமெரிக்காவின் பெரிய அளவைப் பற்றி அறிவுறுத்துகிறது. இதேபோல், குற்றம் மற்றும் இராணுவம் தொடர்பான சொற்கள் அமெரிக்க தரவுத்தொகுதியில் அதிகம் காணப்படுகின்றன; அவை மீண்டும் அமெரிக்க 'துப்பாக்கி கலாச்சாரம்' பற்றி குறிக்கின்றன. பொதுவாக, கண்டுபிடிப்புகள் அமெரிக்கக் கலாச்சாரம் பிரிட்டிஷ் கலாச்சாரத்தை விட 'வன்மையானது' மற்றும் 'ஆற்றல்வாய்ந்தது' என்று கூறுகின்றன.

மற்றொரு ஆய்வில், மொழிப் பயனர்களின் கலாச்சாரத்தில் உள்ள வேறுபாடு குறித்த சுவாரஸ்யமான உற்றுநோக்குகள் இந்திய ஆங்கிலத்தின் கோலாப்பூர் தரவுத்தொகுதி (Kolhapur Corpus of Indian English (KCIE)), LOB தரவுத்தொகுதி மற்றும் பிரவுன் தரவுத்தொகுதி (Shastri 1988) ஆகியவற்றுக்கு இடையே ஒரு ஒப்பீட்டு ஆய்வை மேற்கொண்ட பிறகு கிடைக்கப்பெறுகின்றன. இத்தகைய ஆய்வுகள் பல நம்பிக்கைக்குரிய பகுதிகளைத் திறக்கின்றன, அவை இரண்டு அல்லது அதற்கு மேற்பட்ட மொழிகளின் இணைத் தரவுத்தொகுதிகளுடன் (parallel corpora) மிகவும் நெருக்கமாக ஆராயப்படுகின்றன.

நாடுகளைக் கடந்து மொழி மாறுபாட்டின் கோட்பாடுகளை சரிபார்க்கத் தரவுத்தொகுதியின் பங்கு சோதனை படுக்கையாக அதன் பயன்பாட்டில் ஒப்புக் கொள்ளப்பட்டுள்ளது. இந்திய ஆங்கிலத்தின் கோலாப்பூர் தரவுத்தொகுதி (Kolhapur Corpus of Indian English (KCIE) என்பது பிரவுன் மற்றும் லாப் தரவுத்தொகுதிகள் (Shastri 1988) ஆகியவற்றின் தொகுப்பிற்கு இணையாக 1978-இல் வடிவமைக்கப்பட்டது. இவை மூன்றுமே ஒரே மாதிரியான முறையில் வடிவமைக்கப்பட்டுள்ளன; அவை பிரிட்டன், அமெரிக்கா மற்றும் இந்தியாவில் ஆங்கிலத்தின் பயன்பாட்டு முறைகளைப் புரிந்து கொள்ள பல்வேறு வகையான ஒப்பீட்டு ஆய்வுகளுக்கான உதவியாகப் பயன்படுத்தப்பட்டன. ஆங்கிலம் சொந்த மற்றும் சொந்தமற்ற மொழி பேசுபவர்களின் மொழிப் பயன்பாட்டில் பயன்படுத்தப்படும் வேறுபாடுகளை காண்பதற்காக மீள்நிகழும் சொல் சேர்க்கைகளின் (recurrent word combinations) அமைப்பொழுங்குகளை ஆய இவை பயன்படுத்தப்பட்டன (Cock 1998). பிரவுன் மற்றும் லாப் தரவுத்தொகுதியுடன் (Leitner 1991) இந்திய ஆங்கிலத்தின் கோலாப்பூர் தரவுத்தொகுதி அகராதி மற்றும் இலக்கணம் குறித்து ஒரு சுவாரஸ்யமான ஒப்பீட்டு விசாரணை தொடங்கப்பட்டது. இந்த ஆய்விலிருந்து பெறப்பட்ட முடிவுகள் க்யூர்க் மற்றும் பலர் (Quirk et al. 1985) பரிந்துரைத்த பொதுவான மையக் கருதுகோளை (COMMON CORE HYPOTHESIS) ஆதரிக்கின்றது.

மொழிகளைக்கடந்த (Cross-linguistic) மற்றும் கலாச்சாரதைக் கடந்த (cross-cultural) ஆய்வுகள் இருமொழி மற்றும் பன்மொழி தரவுத்தொகுதிகளால் ஒரே மாதிரியான சட்டகம், அளவு மற்றும் அமைப்புடன் அதிகரிக்கப்பட்டுள்ளன. இந்தியா போன்ற ஒரு பன்மொழி நாட்டில், இத்தகைய தரவுத்தொகுதிகள் பிராந்தியங்களுக்கு இடையிலான உறவுகளை மேம்படுத்துவதோடு மட்டுமல்லாமல் தேசிய ஒருமைப்பாட்டையும் பலப்படுத்தும். எடுத்துக்காட்டாக, பெங்காளி, ஒரியா மற்றும் அசாமி மொழிகளில் தயாரிக்கப்பட்ட முத்தொகுப்பு தரவுத்தொகுதிகளின்

பகுப்பாய்விலிருந்து, சொற்கள் பெரும்பாலும் ஒரே மூலத்திலிருந்து எவ்வாறு உருவாகின்றன என்பதைக் காண்கிறோம்; மொழிகள் பரஸ்பரம் புரிந்துகொள்ளக்கூடியவையாக உள்ளன; வாக்கியங்கள் கட்டுமானத்தில் ஒத்தவையாக இருகின்றன; சொல்சார் பொருண்மை கருத்துசார் ஒற்றுமையைக் கொண்டுள்ளது; இலக்கணப் பண்புகள் செயல்பாட்டில் ஒத்தவையாக உள்ளன; மொழிப் பயனர்களுக்கு ஒத்த சுவை, பழக்கம் மற்றும் வாழ்க்கை முறைகள் கிடைக்கின்றன. இத்தகைய தரவுத்தொகுதிகள் மொழி பயனர்களைத் தொடர்புடைய மொழிகளுக்கிடையில் தகவல் பரிமாற்றம், மொழித் தொடர்பு மற்றும் மொழிபெயர்ப்பைத் தொடங்க ஊக்குவிக்கிறது.

7.10. உளமொழியியலில் தரவுத்தொகுதி

சமீபத்திய உளவியல் ஆராய்ச்சி மற்றும் பயன்பாட்டில், ஆய்வக சோதனைகளுக்கான பொருட்கள் உருவாக்கப்படும் தரவுகளின் மூலமாக தரவுத்தொகுதி பயன்படுத்தப்படுகிறது. சொல் அறிதல் (word recognition) உள்ளிட்ட பல புலனறிவு செயல்முறைகளுக்குள் பல்வேறு சொல்சார் அலகுகளின் அதிர்வெண் ஒரு முக்கியமான அளவுகோலாக இருப்பதால், முறையாக மாதிரிப்படுத்தப்பட்ட தரவுத்தொகுதி சொல் பயன்பாட்டின் அதிர்வெண் பற்றிய உறுதியான மற்றும் நம்பகமான தகவல்களை வழங்கப் பயனுள்ளதாகக் கருதப்படுகிறது (வெவ்வேறு அர்த்தங்களின் அதிர்வெண் மற்றும் பொருண்மை மயக்கமுள்ள சொற்களின் சொல்சார் வகுப்புகள் உட்பட). தரவுத்தொகுதியில் பிரதிபலிக்கும் மொழியின் பயன்பாட்டைக் கவனிப்பதன் மூலம், மனித மொழியால் மொழி எவ்வாறு செயலாக்கப்படுகிறது என்பது குறித்த கருதுகோள்களை உருவாக்குவதற்கு உளவியலாளர்கள் பங்களிக்கின்றனர்.

இயல்பான உரையாடல்களில் பேச்சு பிழைகள் ஏற்படுவதை ஆராய்வதில் உளவியலில் தரவுத்தொகுதியின் இன்றியமையாத பயன்பாடு அங்கீகரிக்கப்பட்டுள்ளது. எடுத்துக்காட்டாக, லண்டன்-லண்ட் பேசுத் தரவுத்தொகுதி (London-Lund Speech Corpus) என்பது ஆங்கிலத்தில் இயல்பான உரையாடலில் பேச்சு பிழைகளை ஆய்வதற்கான நம்பகமான தரவுத்தளமாகும். இது ஆய்வுக்குத் தேவையான தரவைச் சரியாக வழங்கியது. பேசுபவர்களின் ஒட்டுமொத்த மொழியியல்சார் வெளியீடு தொடர்பாக இவற்றின் பொதுவான அதிர்வெண் குறித்த உண்மையான மதிப்பீட்டை வழங்க வெவ்வேறு பிழை வகைகளின் அதிர்வெண்களை வகைப்படுத்தவும் எண்ணவும் முடிந்தது (Garnham et al. 1981). இந்த ஆய்வுக்கு முன்னர் அன்றாட மொழியில் பேச்சு பிழைகளின் அதிர்வெண் பற்றி சரியான மதிப்பீடு எதுவும் இல்லை, ஏனெனில் இதுபோன்ற பகுப்பாய்விற்கு இயல்பான உரையாடலில் இருந்து போதுமான அளவு

தரவு தேவைப்படுகிறது; அது கிடைக்கவில்லை. பேச்சு பிழைகள் குறித்த முந்தைய படைப்புகள் பெரும்பாலும் வெவ்வேறு மூலங்களிலிருந்து சிறிய அளவிலான தற்காலிக தரவு சேகரிப்பை அடிப்படையாகக் கொண்டவை.

மொழி நோயியலில் (language pathology) தரவுத்தொகுதியின் பயன்பாடு அங்கீகரிக்கப்பட்டுள்ளது - மொழி அடிப்படையில் பலவீனமானவர்கள் பல்வேறு வகையான மொழியியல் குறைபாட்டால் எப்படி, ஏன் பாதிக்கப்படுகிறார்கள் என்பதைப் புரிந்துகொள்வதை நோக்கமாகக் கொண்ட உளவியல் அறிவியலின் ஒரு முக்கிய களம். இத்தகைய ஆய்வுகளுக்கு இலக்கு பேசுபவர்களால் தயாரிக்கப்பட்ட மாதிரிகளின் பெரிய தரவுத்தளம் தேவைப்படுகிறது, அவை அன்றாட நடவடிக்கைகளின் பல்வேறு சூழ்நிலைகளில் மொழியியல் ரீதியாக தொடர்பு கொள்கின்றன (MacWhinney 1991). மொழி அடிப்படையில் பலவீனமான மக்களால் உருவாக்கப்பட்ட மொழி குறித்த பெரும்பாலான ஆய்வுகள் அளவிடப்பட்ட பிரதிநிதி விளக்கங்களின் அம்சத்தைக் கொண்டிருக்கவில்லை. தரவுத்தொகுதி துல்லியமான 'அசாதாரண தரவு'களின் தேவையை பூர்த்தி செய்துள்ளது, இதன் பகுப்பாய்வு மனித மொழிச் செயலாக்க அமைப்பில் என்ன தவறு இருக்கிறது என்பதைப் பரிசோதனை மற்றும் கருதுகோள்களுக்கான வழிமுறைகளை உருவாக்க ஆய்வாளர்களுக்கு உதவுகிறது. மாதிரி தரவுத்தொகுதிகளுடன் இதுவரை ஒரு சில பணிகள் செய்யப்பட்டிருந்தாலும், இத்தகைய நோயியல் பகுப்பாய்வுகளுக்கான அவற்றின் முக்கியத்துவம் புறக்கணிக்கப்படுகிறது. பலவீனமான மற்றும் இயல்பான குழந்தைகளால் உற்பத்தி செய்யப்படும் பெரிய அளவிலான தரவைக் கொண்ட சில்ட்ஸ் (CHILDES) தரவுத்தளம் இப்பகுதியில் உள்ள அடிப்படை சிக்கல்களைப் புரிந்துகொள்ள அனுபவ ரீதியாக பகுப்பாய்வு செய்யப்பட்டுள்ளது (Biber, Conrad, and Reppen 1998: 177).

7.11. நடையியலில் தரவுத்தொகுதி

வெவ்வேறு வகைகள், களங்கள், ஆசிரியர்கள், ஊடகங்கள் போன்றவற்றிலிருந்து பெரிய அளவிலான உரை மாதிரிகள் கொண்ட தரவுத்தொகுதி, நடையியல் குறித்த ஆராய்ச்சியின் பல புதிய சாத்தியங்களைத் திறக்கிறது. குறிப்பிட்ட நடையியல் அளவுகோல்களுடன் ஆசிரியர்களால் இயற்றப்பட்ட தனிப்பட்ட உரை வகைகள் அல்லது உரைகளில் ஆராய்ச்சியாளர்கள் ஆர்வம் காட்டுகின்றனர் (Stubbs 1996). ஒரு நாட்டின் எழுத்தாளர்களாலும் பிற நாட்டின் எழுத்தாளர்களாலும் இயற்றப்பட்ட உரைகளில் பிரதிபலிக்கும் அடிப்படை நடையியல் வேறுபாடுகளைக் கண்டறிய சிலர் ஆர்வமாக உள்ளனர் (Wilson 1992). மற்றவர்கள் ஒரு

தலைமுறை அல்லது எழுத்தாளர்களின் குழுவின் எழுத்துக்கள் மற்ற தலைமுறை அல்லது எழுத்தாளர்களின் குழுவிலிருந்து எவ்வாறு வேறுபடுகின்றன என்பதை அறிய ஆர்வமாக உள்ளனர். ஆராய்ச்சியாளர்கள் தங்கள் ஆய்வுகளில் தொடர்புடைய பல்வேறு நடையியல் அம்சங்களைக் குறிக்கும் ஒருமொழிய மற்றும் இருமொழிய தரவுத்தொகுதிகளைக் கொண்டிருக்கும்போது மட்டுமே இத்தகைய திசை மற்றும் ஒப்பீட்டு நடையியல் ஆய்வுகள் சாத்தியமாகும்.

சில ஆராய்ச்சியாளர்கள் உரைகளின் இனம் (genre) மற்றும் வகை (type) போன்ற பரந்த சிக்கல்களை ஆய ஆர்வமாக உள்ளனர். மொழியில் நடையியல் நுணுக்கங்களைக் கையாளும் புலனாய்வாளர்கள் பொதுவாக மொழியின் பொதுவான வகைகளைக் காட்டிலும் சில உரை வகைகளின் சில குறிப்பிட்ட அம்சங்களில் கவனம் செலுத்துவார்கள் என்பது குறிப்பிடத்தக்கது. செய்தித்தாள்களில் பயன்படுத்தப்படும் மொழி பல்வேறு அறிவியல் எழுத்துக்களில் பயன்படுத்தப்படும் மொழியிலிருந்து எவ்வாறு மாறுபடுகிறது என்பதை அறிய அவர்கள் தயாராக உள்ளனர். இத்தகைய ஆய்வுகள், உண்மையுள்ள பகுப்பாய்வு மற்றும் நம்பகமான முடிவுகளுக்கு பெரிய தரவுத்தளத்தின் தரவுத்தொகுதியை வேண்டும். பொது மற்றும் சிறப்பு தரவுத்தொகுதிகள் இரண்டும் முக்கியமான ஆதாரங்களாக இருக்கின்றன, ஏனென்றால் அவை தங்களுக்குள்ளும் மற்றவைகளுடனும் ஒப்பிட்டுப் பார்க்கக் குறிப்புச் சட்டமாகச் செயல்படுகின்றன.

நடையியலில் உள்ள பல்வேறு பொருள் சார்ந்த ஆய்வுகள் (object-oriented studies) அந்த தீர்ப்புகளை காப்புப் பிரதி எடுக்க புள்ளிவிவர ரீதியாக சரிபார்க்கப்பட்ட தகவல்கள் தேவைப்படுகின்றன; அவை புலனாய்வாளர்களுக்கு புறவயமானவை (objective) என்பதைக் காட்டிலும் அகவயமானவை (subjective) என்று தோன்றுகின்றன. எழுத்தாளர் பண்புக்கூறு, நடையின் மாற்றங்களை ஆய்தல், சொற்களின் தேர்வு மற்றும் பயன்பாட்டின் மாற்றங்களின் தடங்களைக் கண்காணித்தல் போன்றவற்றுக்கு தரவுதொகுதி பொருத்தமானதாகிறது. ஆசிரியரின் குறிப்பிட்ட நடையை வரையறுக்க, அவரது எழுத்துக்களிலிருந்து தயாரிக்கப்பட்ட தரவுத்தொகுதி அவர் எவ்வாறு வெவ்வேறு வழிகளைப் பின்பற்றுகிறார் என்பதை அடையாளம் காண பயன்படுத்துகிறோம். விஷயங்களை வைப்பது (எ.கா. தொழில்நுட்பதிற்கு எதிராக தொழில்நுட்பமற்ற சொற்கள், சொற்றொகைத் தேர்வு, நீண்ட மற்றும் குறுகிய வாக்கியம், முறையானதற்கு எதிராக முறைசாரா கதை, முதலியன). இத்தகைய ஆராய்ச்சிக்கு தரவுத்தொகுதியிலிருந்து பல்வேறு புள்ளிவிவர ஆதரவு தேவைப்படுகிறது. இதற்கு ஆசிரியரின்

சொந்த படைப்புகளுக்குள் மட்டுமல்லாமல் பிற ஆசிரியர்களின் படைப்புகள் அல்லது முழு மொழி வேறுபாட்டின் விதிமுறைகள் (deHaan 1997) ஆகியவற்றுடன் ஒப்பிட்டுப் பார்க்க வேண்டும். இது பரிசீலனையில் உள்ள ஆசிரியர்களின் எழுதும் பாணியையும், முழு உரையும் எழுதப்பட்ட பாணியையும் காட்டுகிறது (Hoffman 1955).

ஒரே எழுத்தாளரால் எழுதப்பட்ட பல்வேறு உரைகளிலிருந்து பெறப்பட்ட தரவுத்தொகுதியின் ஒப்பீட்டு பகுப்பாய்வு, பல்வேறு சிக்கல்களையும் நிகழ்வுகளையும் (Elliott and Valenza 1996) கவனிக்கும் போது ஆசிரியர் தனது நுட்பம், கதை, சொற்றொகை, வாக்கியம், நடை போன்றவற்றை எவ்வாறு வேண்டுமென்றே மாற்றுகிறார் என்பதைக் காட்டும். ஒரு குறிப்பிட்ட எழுத்தாளரின் எழுத்துக்களைக் கவனியுங்கள். அவரது எழுத்துக்களிலிருந்து தயாரிக்கப்பட்ட பல்வேறு வகையான தரவுத்தளங்களுக்கிடையில் (எ.கா. சிறுகதைகள், நாவல்கள், கட்டுரைகள், தனிப்பட்ட கடிதங்கள், பயணக் குறிப்புகள் போன்றவை) ஒரு எளிய குறுக்கு ஒப்பீடு அவரது கதை நடை, வாக்கிய கட்டுமானத்தின் நுட்பம், உரை பிரதிநிதித்துவ முறை, சொற்றொகைத் தேர்வு போன்றவை ஒரு உரை வகையிலிருந்து மற்றொன்றுக்கு மாறுபடுவதைக் காட்டும்.

இந்திய மொழிகளில் கிடைக்கும் தரவுத்தொகுதிகள் இனம் மற்றும் ரெஜிஸ்டர் மாறுபாட்டின் தகவல்களால் நிறைந்துள்ளது. இவை நடையியல் ஆய்வின் பல புதிய சாத்தியங்களைத் திறக்கலாம். மாதிரிகளில் காணப்பட்ட வகைகளைப் பற்றிய எளிய பொது ஒப்பீடுகளுடன் தொடங்க இந்த தரவுத்தொகுதிகளை நாம் அணுகலாம். எளிதான ஒப்பீடு மற்றும் குறிப்பிடும் திறன் காரணமாக, இவை ஒருமொழி மற்றும் இருமொழி சூழல்களுக்குள் மற்றும் முழுவதும் வெவ்வேறு அம்சங்களை ஆராய்வதற்கான மதிப்புமிக்க வளங்கள். இவை உரை வகைகளை மற்றும் சில ஆசிரியர்களின் குறிப்பிட்ட நடைகளைப் பின்பற்றும் ஆசிரியர்கள் பண்புக்கூறு ஆகியவற்றை ஆய்வதற்கான நம்பகமான தரவுத்தளங்கள்.

7.12. சுருக்கவுரை

இவ்வியலில் முக்கிய மொழியியலில் செயல்பாடுகளில் தரவுத்தொகுதி பற்றி ஒரு சுருக்கமான அறிமுகம் தரப்பட்டுள்ளது. முக்கியமான மொழியியல் தொடர்பான நடத்தைகளில் தரவுதொகுதியின் பயன்பாடும் அது தொடர்பான செய்திகளும் தரப்பட்டுள்ளன. அகாராதி இயலில் தரவுதொகுதி, சொல்லனியலில் தரவுதொகுதி, தொழில்நுட்பச் சொற்களை உருவாக்குவதில் தரவுதொகுதி, இலக்கணம் எழுதுவதில் தரவுதொகுதி, பொருண்மையியல்

ஆய்வில் தரவுதொகுதி, மொழிக் கற்றலில் தரவுதொகுதி, கிளைமொழியியல் ஆய்வில் தரவுதொகுதி, சமூகமொழியியலில் தரவுதொகுதி, உளமொழியியலில் தரவுதொகுதி, நடையிலில் தரவுதொகுதி என்ற தலைப்புகளில் மொழியியல் தொடர்பான வேவேறுவிதமான செயல்பாடுகளில் தரவுத்தொகுதியின் பங்கு பற்றியும் அது தொடர்பான ஒழுங்குமுறைகள் அல்லது மென்பொருள்கள் உருவாக்கம் குறித்தும் விரிவாக விளக்கப்பட்டுள்ளது.

இயல் 8

இயந்திர மொழிபெயர்ப்பில் தரவுத்தொகுதி

8.1. அறிமுகம்

இயந்திர மொழிபெயர்ப்பு செயற்கை அறிவுநுட்பத்தின் (Artificial Intelligence) சிக்கலான ஒரு பகுதியாகும். இது மூலமொழியின் (source language) வாக்கியங்களிலிருந்து, இலக்குமொழி (target language) வாக்கியங்களை உருவாக்குவதற்கான ஒழுங்கு முறையை உருவாக்க வேண்டும். இலக்கு மொழியின் வெளியீடு உயர்ந்த தரமுடையதாக அமையத் தேவையில்லை;· ஓரளவுக்குத் தரம் வாய்ந்ததாய் இருந்தால் போதும். இயந்திர மொழிபெயர்ப்பு மொழிகளுக்கிடையே கருத்துப்பரிமாற்றத்திற்கும் தகவல் பரிமாற்றத்திற்கும் மொழிகளைக் கடந்த தகவல் மீட்புக்கும் மொழி கற்றலுக்கும் முக்கியத் தொழில் நுட்பமாகும். பல விதமான பயன்பாட்டின் சாத்தியம் சாரணமாக இயந்திர மொழிபெயர்ப்பு மின் வணிகம் (e-commerce), தகவல் வட்டாரமாக்கம் (information localization), பல்மொழி ஆவணப்படுத்தல் (multilingual documentation) மற்றும் தகவல் தொழில் நுட்பம் (information technology) போன்றவற்றில் மிகப் பயனுள்ள தொழில் நுட்பமாக விளங்குகிறது. நாம் இயந்திர மொழிபெயர்ப்புத் தொழில் நுட்பத்தில் விரிதரவின் பங்களிப்பைப் பற்றிக் கூறுவதற்கு முன் பின்வரும் கேள்விக்குத் திருப்திகரமான விடை தர வேண்டும்: மனிதர்கள் மொழி பெயர்ப்பைச் செய்யும்போது ஏன் இயந்திர மொழி பெயர்ப்பை நாம் ஏன் உருவாக்க வேண்டும்? இக்கேள்விக்கான பதில் எளிமையானது அல்ல. ஹட்சின் (Hutchin's 1986) பின்வரும் கதரணங்களைத் தருகிறார்:

1. பல்வேறுபட்ட அறிவியல் தொழில் நுட்ப ஆவணங்களை உலகிலுள்ள தொழில் நுட்பவியலாருக்கும் அறிவியலாருக்கும் உடனடியாகத் தர வேண்டுமானால் இயந்திர மொழிபெயர்ப்பு மிகவும் அவசியமாகும்.

2. மனித மொழி பெயர்ப்பாளர்கள் இல்லாத சூழல்களில் இயந்திர மொழிபெயர்ப்பு மிகவும் பயனுள்ளதாக அமையும்.

3. இயந்திர மொழிபெயர்ப்பு மொழிகளைக் கடந்த ஆவண மாற்றங்கள் மூலம் அக நிலையானக் கூட்டுறவை மேம்படுத்த இயலும்.

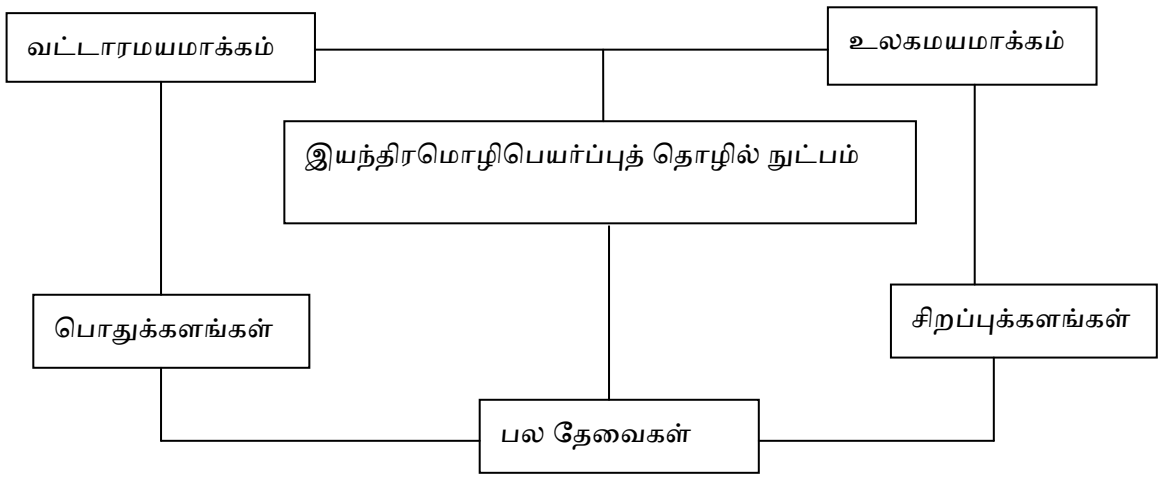
4. இயந்திர மொழிபெயர்ப்பு பல்வேறுபட்ட அறிவியல், தொழில்நுட்பம், விவசாயம் மற்றும் மருத்துவத் தகவல்களை ஏழ்மையான மற்றும் முன்னேறுகின்ற நாடுகளுக்கு விரைவான, எளிதான

மற்றும் மலிவான பரிமாற்றம் அடிப்படையில் மொழித் தடைகளை நீக்கி அக நிலையானக் கூட்டுறவையும் அமைதியையும் மேம்படுத்த இயலும்.

5. மொழிப் பெயர்ப்பு இராணுவ நோக்கங்களுக்கும் தொழில்நுட்ப ஆய்வு மற்றும் வணிக நோக்கங்களுக்கும் மிகவும் பயனுள்ளது.

1980-களில் ஹட்சின் இயந்திர மொழிப்பெயர்ப்புக்குச் சாதகமாகக் கூறிய போது இருந்த சூழல் கடந்த ஆண்டுகளில் மிகவும் மாறியுள்ளது. இந்நூற்றாண்டின் தொடக்கத்தில் மக்களின் வாழ்க்கை நடத்தை காரணமாகத் தாய்மொழி மூலம் தகவலைப் பெறுவது விதியாக மாறியுள்ளது. மேலும் படிப்பறிவின் வளர்ச்சி, செய்தி மற்றும் தகவலின் உலகமயமாக்கம், நாடுகளைக் கடந்த பன்மொழிய தரவுத்தொகுதி, இனங்களுக்கிடையே மொழித் தனித்தன்மையின் அவசியம் மற்றும் பிற காரணிகள் கணினி அறிவியலாரையும் மொழியியலாரையும் இயந்திர மொழிப்பெயர்ப்பு ஒழுங்குமுறையை உருவாக்க இணைந்து பணியாற்றும் சூழலுக்குத் தள்ளியுள்ளது. இக்காரணிகளின் பங்களிப்பைப் பின்வரும் வரைபடத்தின் மூலம் சுருக்கமாகக் கூறலாம்:

படம்: இயந்திர மொழிபெயப்புத் தொழில்நுட்பத்தின் உலகளாவிய எழுச்சிக்கு பின்னால் உள்ள பிணையம்/வலையமைப்பு



இங்கு அனுபவவாதத் தரவுத்தொகுதி அடிப்படையிலான (empirical corpus-based MT (CBMT)). மொழிபெயர்ப்பு முன்மொழியப்படுகின்றது. இது மொழிபெயர்ப்பில் ஈடுபடுத்தப்பட்டுள்ள மொழிகளின் உரைகளிலிருந்து உருவாக்கப்பட்ட மொழி பெயர்க்கப்பட்ட தரவுத்தொகுதிகளின்

பகுப்பாய்விலிருந்து பெறப்பட்ட தகவல்களின் மற்றும் எடுத்துக்காட்டுகளின் அடிப்படையில் அமைந்தது. இங்கு இயந்திர உதவியால் செய்யப்படும் மனித மொழிபெயர்ப்பிற்கு, மொழிபெயர்ப்பு விரிதரவு பயன்படுத்தப்பட விவாதிக்கப்படுகின்றது. எல்லைப்படுத்தப்பட்ட மற்றும் பொது களங்களில் இவ்வொழுங்குமுறையின் பயன்பாடு அழுந்தக்கூறப்படுகின்றது. தகவல் பரிமாற்றத்தின் உலக உயர்வுக்கும் வட்டாரவாக்கத்திற்கும் மொழிபெயர்ப்பு ஒழுங்குமுறையின் தேவை விளக்கப்பட்டு எதிர்கால முன்னேற்றத்திற்கு வேண்டிய வழிமுறைகள் முன்வைக்கப்படுகின்றன. மொழி பெயர்ப்பு வரலாற்றின் மற்றும் தோல்வியில் இன்று கற்றுக்கொண்டவைகளின் அடிப்படையில் இயந்திர மொழிபெயர்ப்பு அமைக்கப்பட்ட ஆலோசனை கூறப்பட்டுள்ளது.

8.2. நோக்கம்

கடந்த 50 ஆண்டுகளில் இயந்திர மொழிபெயர்ப்பு தொழில் நுட்பங்கள் மற்றும் பயன்பாடு இவற்றின் அடிப்படையில் கூறத்தக்க அளவு வளர்ந்துள்ளது. இன்றைய இணையதள கால கட்டத்தில் இணைய தளத்தில் தகவல்கள் பெரிய அளவில் பல மொழிகளில் உள்ளன. அவற்றை எல்லா மொழி பேசுவோரும் பகிர்ந்துகொள்ள வேண்டுமானால் அவை ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு உடனடியாக மொழி மாற்றம் செய்யப்பட வேண்டும். இதை இயந்திர மொழிபெயர்ப்பின் விரைவான முன்னேற்றத்தின் அடிப்படையில் தான் நிறைவேற்ற இயலும். இது நடைமுறையில் சாத்தியமல்லாத செயல்பாடாகத் தற்போது தோன்றினாலும் நாம் இதைச் சாத்தியமாகச் செய்கின்ற இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதில ஆர்வம் காட்டவேண்டும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு வடிவிலிருந்து மற்றொருவடிவுக்கு (அதாவது சொல்லிலிருந்து சொல்லையோ கூட்டுச் சொற்களிலிருந்து சுட்;டுச் சொற்களையோ தொடரிலிருந்து தொடரையோ வாக்கியத்திலிருந்து வாக்கியத்தையோ) இடப்பெயர்ச்சிசெய்யும் அல்லது ஒரு மொழியிலுள்ள உரைகளை மற்றொரு மொழியில் உருப்படுத்தம் செய்யும் எளிய முறையோ அல்ல. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை, உருவாக்கப்பட்ட உரை இலக்குமொழியில் இலக்கண அடிப்படையில் சரியான மற்றும் கருத்துரு அடிப்படையில் ஏற்றுக் கொள்ளத்தக்க வகையில் அமைய உறுதி தருவதைக் குறிப்பிடுகின்றது. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் வெளியீடு அர்த்தம் மற்றும் பொருளடக்க அடிப்படையில் ஒன்றாக இல்லாவிடினும் குறைந்தது மூலமொழிக்கு அண்மையில் இருக்க வேண்டும். மூல மொழியிலுள்ள தகவல் இலக்கு மொழியில் மாற்றப்படும் போது இழப்பு எற்படக் கூடாது. மொழிப்பில்

மூலமொழியை விடக் கூடுதல் தகவல் சேர்க்கப்படலாகாது. இயந்திர மொழிபெயர்ப்பில் துல்லியத்தைப் பெற இவைகள் தடைகளாகும். இயந்திர மொழிபெயர்ப்பு உருபனியல், சொல்லியல், தொடரியல் பொருண்மையியல், பயன்வழியியல், கருத்தாடல், புலனறிவு இவற்றைச் சார்ந்திருக்க வேண்டும். இதன் அர்த்தம் என்னவென்றால் ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு மனித மொழிபெயர்ப்பாளருக்குள்ள திறனைப் பெற்றிருக்க வேண்டும் என்பதாகும். இதைச் சாதிப்பதற்கு நாம் காத்ததிருக்க வேண்டும். இருப்பினும் இருமொழி (bilingual) அல்லது இணை (parallel) மொழிபெயர்ப்பு தரவுத்தொகுதிகள் (translation corpora) இதைச் செய்வதற்குச் சாத்தியமான சூழலைத் தருகிறது. இது நம்மை தரவுத்தொகுதிகள் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அணுகுமுறைக்குக் கொண்டு செல்கிறது. இதன் காரணமாக இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் திறன் கூடுகிறது. வெளியீடானது ஏற்றுக் கொள்ளத் தக்க முறையில் பின் திருத்தம் செய்யப்படுவதன் காரணமாகவும் குறையான தகவலின் மூலமாகத் திருத்தப்படாத வடிவில் பயன்படுத்தப்படுவதன் காலரணமாகவும் இயந்திர மொழிபெயர்ப்பு முற்றிலும் சரியான மொழி பெயர்ப்பை உருவாக்குவதை நோக்கமாகக் கொண்டதல்ல எனக் கூறவியலும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை மொழியியல் அடிப்படையிலும் புலனறிவு அடிப்படையிலும் உத்தேச ஒழுங்குமுறையைத்தான் நோக்கமாகக் கொள்கிறது. முயற்சித்தல் மற்றும் பிழை செய்தல் என்ற நீண்ட செயற்பாங்குகளின் வழி இலக்கை அடையலாம். இதற்கு தற்போதுள்ள அமைப்புகளின் வழக்கமான மதிப்பீடு, தவறுகளை அடையாளம் காணுதல், இருக்கும் நுட்பங்களைச் செம்மைப்படுத்துதல், கடந்த கால அனுபவங்களை மேம்படுத்துதல் மற்றும் மொழியியல் அறிவுத் தளத்தை மேம்படுத்துதல் ஆகியவை தேவை. பாதை முட்கள் நிறைந்ததாக இருக்கிறது, ஆனால் இது இயந்திரமொழிபெயர்ப்பு மென்பொருள் பொறியாளர்கள் மற்றும் கணினி மொழியியலாளர்களின் திறனை சவால் செய்கிறது.

8.3. வரலாற்றிலிருந்து கிடைக்கின்ற பாடம்

1950-களில் இயந்திர மொழி பெயர்ப்பு உலகிலுள்ள பெரும்பாலான எல்லா மொழிகளுக்கும் நேரடியாகவோ மறைமுகமாகவோ மூல மொழியிலிருந்து இலக்கு மொழிக்கு இயந்திர மொழிபெயர்ப்பு செய்ய முயற்சிகள் எடுக்கப்பட்டன. இதுவரை முழு வெற்றி அடைய இயலவில்லை. இதன் அர்த்தம் என்னவென்றால் மூல மொழிக்கும் இலக்கு மொழிக்கும் மொழி பயன்பாட்டாளர்களால் ஏற்றுக் கொள்ளக்கூடிய மொழிபெயர்ப்புகளைத் தானியக்கமாக

உருவாக்கும் ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை இல்லை. இதன் விளைவாக இயந்திர மொழிபெயர்ப்பு ஆய்வாளர்கள் மரபு சார்ந்த விதி அடிப்படையிலான அணுகுமுறையிலிருந்து தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறைக்குத் தங்கள் கவனத்தை திருப்பியுள்ளனர்.

இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சியில் கடந்த ஐந்து தசாப்தங்களின் முடிவுகளை நோக்கிய அவர்களின் பொதுவான ஆர்வமின்மையை இது அம்பலப்படுத்துகிறது. பிற அணுகுமுறைகளிலிருந்து பெறப்பட்ட முடிவுகள் இயந்திர மொழிபெயர்ப்பு வணிகத்தை உலகளாவிய மொழிபெயர்ப்பு சந்தையின் மிகக் குறைந்த பகுதியை விட ஊடுருவ அனுமதிக்கவில்லை.

பல தசாப்தங்களுக்கு முன்னர், மார்ட்டின் கே (Martin Kay 1980) இத்தகைய தோல்வியைப் பற்றி சிந்தித்தார், மேலும் இயந்திர மொழிபெயர்ப்பு சமூகம் தங்கள் இலக்கை மரபு எம்டி அமைப்பிலிருந்து இயந்திர உதவியுடன் மனித மொழிபெயர்ப்பு (machine-aided human translation MAHT) முறைக்கு மாற்றுமாறு வலியுறுத்தினார். கேயால் முன்மொழியப்பட்டவை விவேகமானவை என்பதை ஒரு சில இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சியாளர்கள் மட்டுமே உணர்ந்திருந்தாலும், அவர்களில் பெரும்பாலோர் கே உடன் உடன்பட மறுத்து, தங்களது சொந்த பழக்கமான அமைப்புகளுடன் ஒட்டிக்கொள்ள விரும்பினர். இயந்திர உதவியுடன் மனித மொழிபெயர்ப்புக்கு தீவிர கவனம் செலுத்துவதற்கான தயக்கம் சமகால விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் (rule-based machine translation RBMT) பரவலான வெற்றியின் தாக்கத்தின் விளைவாகும். இருப்பினும், விதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஆராய்ச்சியின் மூன்று தசாப்தங்களுக்குள், இயந்திர உதவியுடன் மனித மொழிபெயர்ப்பின் வெற்றியை சவால் செய்ய திறமையான எந்த ஒரு முடிவையும் கண்டுபிடிப்பது கடினம்.

8.4. தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை

கடந்த நூற்றாண்டுகளாக இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்கு வேண்டி உருவாக்கப்பட்ட விதிகளின் குழுமம் மொழி அமைப்பை முழுவதுமாகப் பிரதிபலிக்கவில்லை. விதிகளின் சில குழுமங்களால் இயற்கை மொழியின் நெகிழ்வான அமைப்பை உருப்படுத்தம் செய்ய இயலாது. இதன் காரணமாக தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழி பெயர்ப்பு உருவானது. தரவுத்தொகுதி அடிப்படையிலான (corpus-Based Approach (CBM) பின்வரும் இயந்திர மொழிபெயர்ப்பு வகைகளை உள்ளடக்கும்.

1. எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Example-based Machine Translation (EBMT) (McLean 1992, Somers 1999)

2.புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (Statistics-based Machine Translation (SBMT)) (Brown, Pietra, and Mercer 1993)

இது கணினி வடிவமைப்பாளர்களுக்கு இன்னும் அடைய இயலாத இலக்கை அடையச் செய்தது.

பிற இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மட்டுப்படுத்தப்பட்ட வெற்றியைக் கொண்டிருக்கும்போது இந்த ஒழுங்குமுறை எவ்வாறு நல்ல அறுவடை செய்யத் துணிந்தது என்பதை ஆராய்வோம். தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் விளைவுகளில் ஒன்று இருமொழி உரைகளின் வாக்கியங்களை வரிசைப்படுத்தும் வழிமுறை வரைவுகளின் உருவாக்கமாகும். இவ்வெளிய விளைவு இயந்திர உதவி வழி மனித மொழி பெயர்ப்பின் அடிப்படை விஷயங்களில் ஒன்றாக அமைந்தது. ஏனென்றால் இது மொழிபெயர்ப்புக்கு உதவும் பல புதிய கருவிகளுக்குப் பொருத்தமான அடிப்படையாக அமைந்தது. தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு மனித மொழிபெயர்ப்பாளர்களால் ஏற்கனவே உருவாக்கப்பட்ட மொழிபெயர்ப்புகளைப் பயன்படுத்தி முழுமையாகவோ பகுதியாகவோ அவற்றின் அக அமைப்புகளைக் கண்டுபிடிக்க முயலுகிறது. இப்பகுப்பாய்வு அடிப்படையிலான அணுகுமுறை மொழி பெயர்ப்பாளர்களுக்கு உதவும் கருவிகளை உருவாக்க உதவியது. இருமொழி அல்லது இணைமொழி பெயர்ப்பு விரிதரவை இயந்திர மொழிபெயர்ப்பில் பயன்படுத்தும் கருத்து புதியது அல்ல. இயந்திர மொழிபெயர்ப்பின் ஆரம்ப காலத்திலேயே இது முயற்சிக்கப்பட்டது (Kay and Röscheisen 1993). பொதுவாக இணை தரவுத்தொகுதிகள் ஒரு மொழி தரவுத்தொகுதிகளை விட மொழிகளைப் பற்றிய விரிவான தகவலைக் கொண்டிருக்கிறது. தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (corpus-based machine translation (CBMT) ஒழுங்குமுறை எடுத்துக்காட்டு இயந்திர மொழிபெயர்ப்பு அணுகுமுறையையும் புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அணுகுமுறையும் இணைத்துப் பயன்படுத்துவதால் இது உயர்ந்ததாக இருக்கிறது. தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (corpus -based machine translation (CBMT) மூல மொழியிலும் இலக்கு மொழியிலும் உள்ள விரிதரவுகளில் முதன்மையான பகுப்பாய்விலிருந்து உருவாக்கப்பட்ட வகைகளின் பரப்பெல்லைகளின் அடித்தளத்தில் அமைந்துள்ளது. பகுப்பாய்வு மொழிபெயர்க்கப்பட்ட தரவுத்தொகுதியில் அடங்கியுள்ள தொடர்கள், மரபுத் தொடர்கள்,

வாக்கியங்கள், பத்திகள் மற்றும் பிற மொழியியல் பண்புக்கூறுகளின் உருபனியல், பொருண்மையியல் மற்றும் புலனறி பொருள்கோள்களை உட்படுத்தும். தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் அடிப்படை நெறிமுறை மொழி பெயர்ப்புக்கு முன் வரையறுக்கப்பட்ட தீர்மானங்கள் இல்லை. ஆனால், மனிதரால் மொழி பெயர்க்கப்பட்ட உரைகளில் மிகச் சாத்தியமான தீர்மானங்கள் காணப்படும் என்ற நம்பிக்கையின் அடிப்படையில் அமைவதாகும். அதாவது மனித மொழிபெயர்ப்பாளர்களின் திறமையின் பெரும்பகுதி மொழி பெயர்க்கப்பட்ட உரைகளில் காணப்படும் மொழி நிகரன்களில் குறியாக்கம் செய்யப்பட்டுள்ளன என்று கூறலாம். எல்லைகுட்பட்ட களங்களில் இவ்வணுகுமுறையால் பெறப்பட்ட சமீபகால வெற்றி (Teubert 2002) பொதுவான களங்களிலும் இம்மாதிரியான வெற்றியைப் பெற மொழி சார்ந்த மற்றும் மொழிக்குப் புறம்பான அறிவுதேவை என்று காட்டுகின்றது. எல்லாக் களங்களிலும் இவ்வணுகுமுறையின் வெற்றி குறித்துத் தீர்மானமாக வருவதுரைக்க இயலாது. இருப்பினும் நன்றாகத் திட்டமிடப்பட்ட மொழிபெயர்ப்பு தரவுத்தொகுதியிலிருந்து பெறப்பட்ட தகவலிருந்து உருவாக்கப்பட்ட நெறிமுறை எல்லைக்குட்பட்ட மற்றும் பொது களங்களில் வெற்றியைத் தரும் என்று கூறவுயலும் (Su and Chang 1992).

8.5. தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறையுடன் தொடர்புடைய சிக்கல்கள்

தரவுத்தொகுதி அடிப்படையிலான மொழிபெயர்ப்பு விரிவான தரவுத்தொகுதி பகுப்பாய்வின் மற்றும் ஆய்வின் வழி கூறத்தக்க முன்னேற்றத்தைப் பெற்றுள்ளது. இந்நாள் வரை பல்வேறு வகையிலான மொழிபெயர்ப்பு விரிதரவுகள் பொருத்தமான ஒழுங்குமுறைகளைத் திட்டமிட உதவும் முக்கியமான உள்ளறிவைத் தரவேண்டி உருவாக்கப்பட்டுப் பகுப்பாயப்பட்டுள்ளன. பொதுவாக இம்மொழிபெயர்ப்பு தரவுத்தொகுதி இயற்கையாகப் பெறப்படும் மொழித் தரவுகளின் பெரிய சேகரிப்பை உருப்படுத்தம் செய்கின்றது. இறுதி பயனர்களின் தேவையைப் பிரதிபலிக்கும் வகையில் உரை மாதிரிகளைச் சேர்ப்பதன் மூலம் அவை குவிக்கப்படுகின்றன. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையைத் தேர்ந்தெடுக்க விரும்பும் இலக்கு இறுதி பயனர்களுக்கு இது மிகவும் முக்கியமானது, இது குறிப்பிட்ட உரை வகைகளை அவற்றின் தேவைகளுக்கு போதுமானதாக மொழிபெயர்க்கும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மொழிபெயர்ப்பு தரவுத்தொகுதியின் பயன்பாடு கட்டாயமானதாகும் என்று தெளிவாக்குகின்றது; ஏனென்றால் அவை மொழி மற்றும் மொழிக்குப்

புறம்மான எடுத்துக்காட்டுகளின் மற்றும் தகவல்களின் பல்வேறு விதமான வகைகளை நமக்குத் தருகிறது.

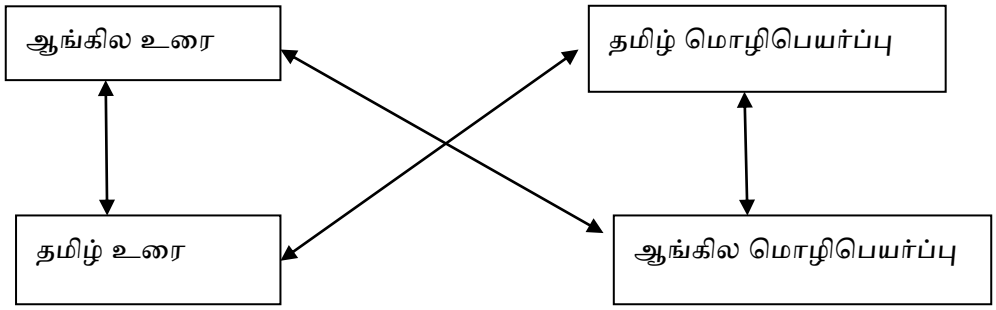
பின்வரும் பிரிவுகளில், தரவுத்தொகுதி அடிப்படையிலான மொழிபெயர்ப்பு அணுகுமுறை தொடர்பான முதன்மை சிக்கல்கள் தமிழ் தரவுத்தொகுதியிலிருந்து பெறப்பட்ட எடுத்துக்காட்டுகள் மற்றும் தேவைப்படும்போது சில சொற்பொருள் வளங்களிலிருந்து விவாதிக்கப்படுகின்றன (Rajendran and Vasuki 2019).

8.5.1 மொழிபெயர்ப்பின் தரவுத்தொகுதி உருவாக்கம்

மொழிபெயர்ப்பு தரவுத்தொகுதி மூல மொழியின் உண்மையான பண்புக் கூறுகளையும் இலக்கு மொழியிலிருந்து பெறப்பட்ட அவற்றின் மொழிபெயர்ப்புகளையும் கொண்டிருக்கும். இவ்விரிதரவு மொழிகளைக் கடந்து சொற்களின் மற்றும் தொடர்களின் பொருண்மையையும் செயல்பாட்டையும் கொண்டிருக்கும். இதன் காரணமாக ஒரே கட்டுபாட்டிற்குள் இரு வேறுபட்ட மொழிகளின் குறிப்பிட்ட பொருண்மைகளின் மெய்யுருவாக்கத்தை ஒப்பிடுவதற்குப் பொருத்தமான அடிப்படையைத் தருகின்றது. மேலும் அவை மொழிகளைக் கடந்த வேறுபட்ட சொற்களைக் கண்டுபிடிப்பதைச் சாத்தியமாக்குகிறது. இவ்வாறு மொழிபெயர்ப்பு தரவுத்தொகுதி மொழிகளைக் கடந்த தரவுப் பகுப்பாய்வுக்கும் மொழிபெயர்ப்புக்குத் தேவையான விதி வடிவாக்கத்திற்கும் பயனுள்ள கூடுதல் மூலவளங்களைத் தருகின்றது (Altenberg and Aijmer 2000: 17). மொழிபெயர்ப்பு விரிதரவின் உருவாக்கம் ஒரு சிக்கலான வேலையாகும். இது தரவுத்தொகுதியின் உருவாக்கம் மற்றும் பகுப்பாய்வி இவற்றின் நீண்ட அனுபவம் உள்ள தரவுத்தொகுதி மொழியிலாரின் நிரந்தரமான மற்றும் கவனமான வழிகாட்டலை வேண்டுகிறது. மொழிபெயர்ப்பு விரிதரவு ஒப்பீட்டு மற்றும் இணை விரிதரவுகளின் (Comparable Parallel Corpora) முன்னேற்றத்திற்கான வழிகளை ஒன்றிணைக்க வேண்டி தகுதியுடையதாக ஆக்க உருவாக்கப்படும். இரு மொழிகளிலிருந்து உரை மாதிரிகள், உரை வகை, பாடப்பொருள், நோக்கம் மற்றும் கலைச்சொல் இவற்றின் அடிப்படையில் இயன்ற வரைக்கும் பொருத்தப்படும்.

இயந்திரமொழிபெயர்ப்புச் செயல்பாட்டின் அடிப்படை நோக்கம் மற்றும் மொழிபெயர்ப்பு தரவுத்தொகுதிக்குள் ஒருங்கிணைக்கப்பட வேண்டிய கூறுகளை மனதில் வைத்துக் கொண்டு, மொழிபெயர்ப்புக் தரவுத்தொகுதிகளின் கட்டமைப்பானது இரண்டு மொழிகளுக்குள் பின்வரும் வழியில் திட்டமிடப்பட்டுள்ளது (படம்). கீழே கொடுக்கப்பட்டுள்ள வரைபடம், இரண்டு மொழிகளுக்கிடையேயான மொழிபெயர்ப்புத் தரவுத்தொகுதிகளை, அவை மேலும்

ஒப்பிடக்கூடிய தரவுத்தொகுதிகளாகவும் (இடதுபக்க கட்டங்களைச் சேர்க்கும் செங்குத்து அம்பு), ஒரே மொழிகளில் உள்ள அசல் மற்றும் மொழிபெயர்க்கப்பட்ட உரைகளை ஒப்பிடுவதற்கு பயன்படுத்தப்படுகின்ற மற்றும் உரைகள் (மூலைவிட்ட அம்புகள்), இரண்டு மொழிகளில் மொழிபெயர்க்கப்பட்ட உரைகளை ஒப்பிடுகின்றன (வலதுபக்க கட்டங்களை இணைகிற செங்குத்து அம்பு) இரு திசை மொழிபெயர்ப்புத் தரவுத்தொகுதிகளாகவும் (கிடைமட்ட அம்புகள்) வடிவமைக்கப்பட்டுள்ளது .



இம்மாதிரியான தரவுத்தொகுதிக்கு உரை மாதிரிகள் தெரிவு பின்வரும் கொள்கைகளால் பொதுவாக வழி நடத்தப்படும்:

1. எழுதப்பட்ட உரைகளில் பயன்படுத்தப்படும் மொழி மாதிரிகள் மட்டும் தான் மொழி பெயர்ப்பு விரிதரவில் உட்படுத்தப்பட்டுள்ளன.

2. பேச்சு கருத்தாடலிலிருந்து பெறப்படும் உரைகளை உட்படுத்தும் சந்தர்ப்பம் இல்லை. ஏனென்றால் தற்போதைய இயந்திர மொழிபெயர்ப்பு எழுதப்பட்ட உரைகளை மட்டுமே இலக்காகக் கொண்டுள்ளது. உட்படுத்தப்பட்டுள்ள உரைகள் தற்கால மொழியைப் பிரதிபலிப்பது எதிர்பார்க்கப்படும். இருப்பினும், பழங்கால நூல்கள் வரலாற்று உரைகளின் மொழி பெயர்ப்புக்குத் தேவையானதாகும்.

3. மொழிபெயர்ப்பு தரவுத்தொகுதி எந்தக் குறிப்பிட்ட வட்டார மொழிக்கோ மொழி வகைக்கோ உரிய உரை வகைக்கு எல்லைப்படுத்தப்படவில்லை. அவை மொழிப் பயன்பாட்டின் எல்லாச் சாத்தியமான களங்கள் மற்றும் துறைகளிலிருந்து பெறப்பட்ட உரை வகைகளின் பரந்த பரப்பெல்லையை உட்படுத்தும்.

4. இரண்டு மொழிகளிலிருந்து எடுக்கப்பட்ட உரைகள் இயன்றவரை ஒப்பிடத் தக்கவையாக இருக்கும். அவை இனம் (எ.கா. செய்தி), வகை (எ.கா.அரசியல்), பொருளடக்கம் (எ.கா. தேர்தல்) மற்றும் வடிவம் (எ.கா. அறிக்கை) இவற்றின் அடிப்படையில் பொருத்தம் பார்க்கப்படும். அவைகள் நோக்கம், பயன்படுத்துபவரின் மொழி இவற்றின் அடிப்படையிலும் பொருத்தமுடையதாக இருக்கும்.

5. மொழிபெயர்ப்பு விரிதரவில் உட்படுத்தப்பட்டுள்ள மாதிரிகள் ஒரு இயல்பான தடை நிலையில் (எ.கா. பகுதி, துணைப்பகுதி, அல்லது பிரிவு, பத்தி போன்றவை) தொடக்கத்திலிருந்து இறுதிவரை எடுக்கப்பட்ட உரைகளில் விரிவான மற்றும் இயற்கையான பாகங்களைக் கொண்டிருக்கும்.

8.5.2 மொழிபெயர்ப்பு தரவுத்தொகுதியைப் பொருத்தமாக வரிசைப்படுத்துதல்

மொழிபெயர்ப்பு தரவுத்தொகுதிகளைப் பொருத்தமாக வரிசைப்படுத்துவது என்பது மூலமொழியின் ஒவ்வொரு மொழிபெயர்ப்பு அலகும், இலக்குமொழியின் ஒவ்வொரு மொழிபெயர்ப்பு அலகுடன் பொருந்த வேண்டும் என்பதாகும். இங்கு மொழிபெயர்ப்பு அலகு என்பது சிறிய தொடர்ச்சிகளான சொற்கள், கூட்டுச்சொற்கள், தொடர்கள் மற்றும் வாக்கியங்கள் (Dagan, Church and Gale 1993) என்பனவற்றுடன் பெரிய தொடர்ச்சிகளான பத்திகள், இயல்கள் (Simard et al. 2000) என்பனவற்றையும் உள்ளடக்கும். மொழிபெயர்ப்பு அலகுகளின் தெரிவு மொழியியல் ஆய்வுக்குத் தெரிந்தெடுக்கப்பட்டுள்ள பார்வையையும் பயன்படுத்தப்பட்டுள்ள தரவுத்தொகுதியின் வகையையும் பொறுத்து அமையும். மொழிபெயர்க்கப்பட்ட தரவுத்தொகுதி மூலத்திலிருந்து (எ.கா. சட்டம் மற்றும் தொழில்நுட்ப விரிதரவுகள்) மாறாத உயர்ந்த நிலையை வேண்டினால் இரு தரவுத்தொகுதிகளின் நெருங்கிய வரிசைப் பொருத்தம் வாக்கியங்கள், சொற்கள் என்ற அளவில் விலகி நிற்கும். தரவுத்தொகுதி தழுவலாக இருந்தால் மூலத்தின் நேரடி மொழிபெயர்ப்பாக அமையாமல் பத்திகள், இயல்கள் என்ற பெரிய அலகுகளை வரிசைப்பொருத்தம் செய்ய முயற்சிக்கப்படும் (Véronis 2000: 12). இவ்வாறு வரிசைப்படுத்தத்தின் செயற்பாடு பயன்படுத்தப்பட்டுள்ள விரிதரவின் வகை அடிப்படையில் நேர்படுத்தப்படும். வரிசையாக வருவதும் மனித மொழிபெயர்ப்பாளர்களின் நம்பகத் தன்மையும் மொழிபெயர்க்கப்பட்ட தரவுத்தொகுதிகளை வரிசைப்பொருத்தம் செய்ய உதவும். இது கூடுதல் தொழில்நுட்ப விரிதரவுக்குப் பகுதி உண்மையாகும். மாறாக, இலக்கிய விரிதரவுகள் (Literary corpora), விரிதரவில் பயன்படுத்தப்படும் நிகரன்கள் முன்னரே முறையாக்கம் செய்யப்பட்டிருந்தால் வாக்கிய மட்டத்திற்கு கீழே உள்ள அலகுகளின் வரிசைப் பொருத்தத்திற்கு

உட்படும் (Chen and Chen 1995). எந்த நிலையிலும் (பத்தி, வாக்கியம், சொல் போன்றவை) மொழிபெயர்க்கப்பட்ட விரிதரவுகள் இணையான அலகுகள் கொண்ட எளிய சொல் தரவு மையங்களாகக் கருதப்படுகிறது. இதன் முக்கிய நோக்கம் இரண்டு மொழிகளுக்கிடையே காணப்படும் அமைப்பு ஒற்றுமைகளைக் காட்டுவதற்கு அல்ல. பயன்வழியியல் அடிப்படையில் மூல உரை அலகுகளுக்கு அண்மைப்பட்டதாகத் தோன்றும் இலக்கு உரை அலகுகளைத் தேடுவதாகும். இதைச் செய்வதற்குத் தொடக்க நிலை, இருமொழிய அகராதிகளின் உதவியால் சொற்களின் ஆரம்ப வரிசைப் பொருத்தத்தைச் செய்வதாகும். இத்தகைய குறைபாடான வரிசைப் பொருத்தங்கள் வாக்கிய மட்டத்தில் திருப்திகரமான முடிவுகளைத் தருகின்றன (Kay and Röscheisen 1993), குறிப்பாக அவை புள்ளியியல்சார் முறைகளுடன் (Brown and Alii 1990) முக்கிய தொடரியல் நிகழ்வுகளின் குறைந்தபட்ச முறைப்படுத்தலுடன் (Brown and Alii 1993) இணைக்கப்படும்போது. இந்த முறையின் முக்கிய நன்மை 'மொழிபெயர்ப்பு நினைவகம்' ('translation memory') அதாவது இருமொழி நூல்களில் காணப்படும் தரவின் ஒருங்கிணைப்பு ஆகும். இரு மொழிகளிலிருந்தும் சிறப்புத் துறைகளின் சில நோக்கீட்டு தரவுத்தொகுதிகளை (reference corpora) (எ.கா. மருத்துவ அறிவியல், சட்ட நடவடிக்கைகள், கணினி அறிவியல் போன்றவை) பயன்படுத்துவதன் மூலம் பணி மேலும் எளிமைப்படுத்தப்படுகிறது.

பயிற்சி கட்டத்தின் போது மனித மொழிபெயர்ப்பாளர்களால் உருவாக்கப்பட்ட தனிப்பயனாக்கப்பட்ட அடிப்படை அகராதி (customised basic dictionary) மற்றும் மொழிபெயர்ப்பு நினைவகத்தைப் பயன்படுத்தி 'இயந்திரம்' அடிப்படையில் மொழிபெயர்க்கப்படுகிறது என்பது செய்தி ஆகும். மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் மற்றொரு முக்கியமான பகுதி, வாக்கிய நிலை ஒழுங்குபடுத்தல் ஆகும். இது பொருத்தங்களை வாக்கிய நிலைக்கு காட்டுவதை நோக்கமாகக் கொண்டுள்ளது; அதற்கு மேல் அல்ல (Brown, Lai and Mercer 1991). இந்த பணியைப் பொறுத்தவரை, ஒரு வலுக்குறைவான மொழிபெயர்ப்பு மாதிரியானது நோக்கத்திற்காக உதவுகிறது, ஏனெனில் இது மொழிபெயர்ப்பு தரவுத்தொகுதிகள் பகுப்பாய்வின் ஆரம்ப கட்டத்தில் தேவைப்படும் முதன்மை கருவிகளில் ஒன்றாகும் (Simard, Foster, and Isabelle 1992).

மொழிபெயர்ப்பு பகுப்பாய்விகளை உருவாக்குவதற்கான ஆராய்ச்சி நடந்து வருகிறது, இது சொற்கள், சொற்றொடர்கள் மற்றும் உருபன்களுக்கு இடையில் சிறந்த மொழிபெயர்ப்பு பொருத்தத் தொடர்புகளைக் கொண்டிருக்கும். சமீபத்தில் உருவாக்கிய சில நுட்பங்கள் (Oakes

and McEnery 2000) ஒரு புதிய புதிய மொழிபெயர்ப்பு ஆதரவுக் கருவிகளுக்கான வழிகளைத் திறக்கின்றன: மொழிபெயர்ப்பு நினைவகப் பயன்பாடு (translation memory application), மொழிபெயர்ப்புச் சரிபார்ப்பு (translation checker) மற்றும் மொழிபெயர்ப்பு எடுத்துக்கூறல் ஒழுங்குமுறை (translation dictation system) போன்றவை. முன்மொழியப்பட்ட முறையின் மற்றொரு சுவாரஸ்யமான அம்சம் தரவுத்தளங்கள் மூலம் பொருட்களைத் தேடுவதற்கு சில புள்ளிவிவர நுட்பங்களைப் பயன்படுத்துவது. புள்ளிவிவர தேடல் வழிமுறைகள் இரண்டு முக்கிய உரை மாதிரிகளிலிருந்து சமமான சொற்றொடர் கூறுகளை மீட்டெடுக்க சில முக்கிய சொற்களைப் பயன்படுத்துகின்றன. இவை கண்டுபிடிக்கப்பட்டதும், இவை இருமொழி மொழிபெயர்ப்பு நினைவகத்தில் சேமிக்கப்படுவதற்கு முன்பு மனித மொழிபெயர்ப்பாளர்களால் மாதிரிகள் என முறைப்படுத்தப்படுகின்றன.

பயிற்சியைத் தானியங்கச் செய்ய மட்டுமே செயல்முறை பரிந்துரைக்கப்படுகிறது; மொழிபெயர்ப்பின் சரிபார்ப்புக்கு அல்ல. இது தானியங்கி இயந்திரமொழிபெயர்ப்பு என அழைக்கப்படும் ஒழுங்குமுறைக்கும் மனித உதவி இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைக்கும் இடையிலான வேறுபாடுகளைக் குறிக்கும் அடிப்படை அளவுகோல்களில் ஒன்றாகும், இது மொழிபெயர்ப்பு தரவுத்தொகுதிகளால் ஆதரிக்கப்படுகிறது.

8.5.3. இன்றைய இந்திய நிலை

விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பில் முக்கியமான தடைகளில் ஒன்று இருமொழிய மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் உருவாக்கமாகும். ஒருமொழிய விரிதரவில் போதுமான அளவு எழுதப்பட்ட உரைகள் இருந்தாலும் பின்வரும் செயல்படுத்துநிலை சிக்கல்களின் காரணமாக இருமொழிய மொழிபெயர்ப்பு தரவுத்தொகுதிகளை மறுதிட்டவரைவு செய்வதற்கு ஆர்வமுள்ள முயற்சிகள் எடுக்கப்படவில்லை:

உருவாக்கம்: இந்திய மொழிகளுக்கிடையே வேறுபட்ட வகையான மொழிபெயர்ப்பு விரிதரவு நீண்டநாள் தேவையாகும். சில ஒருமொழிய தரவுத்தொகுதிகள் ஒரேமாதிரியான திட்டவரைவுக் கொள்கைகளையும் உரை வகைகளையும் கொண்டு உருவாக்கப்பட்டுள்ளன. இவற்றை முறையாக மொழிபெயர்ப்பு தரவுத்தொகுதிகளை உருவாக்கப் பயன்படுத்தலாம்.

ஏற்றுக்கொள்ளல்: மனித முயற்சியால் உருவாக்கப்பட்ட மொழிபெயர்ப்பு தரவுத்தொகுதிகள் சில (எ.கா. National Book Trust மற்றும் Sahitya Academy போன்றவற்றால் வெளியிடப்பட்ட பல

உரைகள்) ஏற்கனவே உருவாக்கப்பட்டுள்ளது என்றாலும் அவை ஏற்றுக்கொள்ளக்கூடிய மின்வடிவில் இல்லை.

மாற்றம்: வேறுபட்ட மூலம், உரைவகை, மாதிரி மற்றும் வடிவமைப்பு கொண்ட பிற விரிதரவுகள் இணை மொழிபெயர்ப்பு தரவுத்தொகுதிகளாக மாற்ற வேண்டி ஆயப்படவேண்டும்.

சுத்தமாக்கல்: மொழிபெயர்ப்பு விரிதரவு எதிர்கால செயற்பாங்கிற்கும் ஏற்றுக்கொள்ளலுக்கும் பயன்பாட்டிற்கும் வேண்டி இயந்திரப் பயன்பாட்டு வடிவமைப்பில் வைத்திருக்க மொழி வல்லுனர்களால் முறையாக தலையிடப்படவும் கட்டுப்படுத்தப்படவும் வேண்டும்.

இணைப்பொருத்தமாக வரிசைப்படுத்தல்: மொழிகளுக்கிடையில் பொருந்தும் பகுதிகளை அடையாளங்காண வேண்டி மொழிபெயர்ப்பு தரவுத்தொகுதி இணைப்பொருத்தமாக வரிசைப்படத்தப்படவேண்டும். இது மொழிபெயர்ப்பு அலகுகளை அடையாளம் காண்பது மற்றும் பல்வேறு அறிவுத் தளங்களை (எ.கா. முறையான, சொல்சார், பொருண்மையியல்சார், புலனறிவுசார் போன்றவை) வரிசைப்படுத்துதல் மற்றும் இறுதியில் தவத்தொகுதிகளில் பயன்படுத்தப்படும் அலகுகளின் இணைத்தல் மற்றும் ஒழுங்குமுறைப்படுத்துதல் ஆகியவற்றை உள்ளடக்கியது.

இந்தச் சிக்கல்களில் பெரும்பாலானவை மொழியியல் இயல்புடையவை, மூல மொழிபெயர்ப்பு தரவுத்தொகுதிகள் கிடைக்கும்போதெல்லாம் அவை கவனத்தில் கொள்ளப்படலாம். தற்போது, முக்கிய சிரமம் மொழிபெயர்ப்பு தரவுத்தொகுதிகளை வாங்குவது, மற்றும் அவற்றின் பொருத்தங்களை அலகுகளுக்குள் வரையறுப்பது. தரவுத்தொகுதிகளை வாங்குதல், மொழியியல் பகுப்பாய்வு மற்றும் அவற்றின் கூறுகளின் ஒழுங்குபடுத்துதல் ஆகியவற்றிற்குப் பிறகு, இவை 'ஐடம் தேடலுக்கு' ('item search') சமர்ப்பிக்கப்படும் - இது ஒரு புள்ளிவிவர நுட்பமாகும், இது மொழிபெயர்ப்பு பகுப்பாய்வு மூலம் வழங்கப்படும் தடைகளை குறைக்கும்.

தரவுத்தொகுதி அடிப்படையிலான இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறைக்கு எதிரான விமர்சனங்கள் செயல்பாட்டுக்காக வடிவமைக்கப்பட்ட இணை நிறுவனங்களின் பற்றாக்குறை என்பதை சுட்டிக்காட்டுகின்றன. தரவுத்தொகுதி அடிப்படையிலான இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையின் ஆராய்ச்சியாளர்களுக்கு மொழிபெயர்ப்புத் தரவுத்தொகுதிகள் மட்டுமல்ல, மனித வல்லுநர்களால் தரவுத்தொகுதிகளின் விரிவான பகுப்பாய்விலிருந்து பெறப்பட்ட

தகவல்களும் தேவைப்படுகின்றன (Elliott 2002). இது அமைப்பை வடிவமைக்க மற்றும் அமைப்பின் நம்பகத்தன்மையை சோதிக்க நிபுணர்களை அனுமதிக்கிறது.

எனவே, தரவுத்தொகுதி அடிப்படையிலான இயந்திரமொழிபெயர்ப்பு படைப்புகளுக்காக இந்திய மொழிகளில் மொழிபெயர்ப்பு தரவுத்தொகுதிகளை தொகுப்பது விவேகமானது, ஆராய்ச்சி தேவைகளை பூர்த்தி செய்வது மட்டுமல்லாமல், ஏற்கனவே உருவாக்கப்பட்டுள்ள பிற ஒழுங்குமுறைகளின் செயல்திறன் மற்றும் பயனை மதிப்பீடு செய்வதும் ஆகும்.

இந்த விஷயத்தில், பிற மொழிகளின் மொழிபெயர்ப்பு தரவுத்தொகுதின் பகுப்பாய்விலிருந்து பெறப்பட்ட தகவல்கள் தொடக்கநிலைக்கு தேவையான வழிகாட்டுதல் திசைகளை வழங்கக்கூடும். எடுத்துக்காட்டாக, ஆங்கிலத்திலிருந்து பிற மொழிகளுக்கான மொழிபெயர்ப்புகளில் தரவுத்தொகுதிகள் எவ்வாறு பயன்படுத்தப்படுகிறது என்பது ஆங்கிலம் மற்றும் இந்திய மொழிகளுடன் பணிபுரியும் மக்களுக்கு பயனளிக்கும் (Rajapurohit 1994).

அந்த மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் தகவல்களுக்கான அணுகல் குறிப்பிட்ட மூல மொழியுடன் (Baker 1993) மொழிகள் எவ்வாறு இணைக்கப்பட்டுள்ளன என்பதையும், மூலமொழி உரைகளில் குறியிடப்பட்ட தகவல்கள் இலக்குமொழி உரைகளுக்கு எவ்வாறு மாற்றப்படுகின்றன என்பதையும் கண்டறிய தேவையான வழிமுறைகளை வழங்கும்.

இத்தகைய ஆய்வுகள் இந்திய மொழிகளுக்காக வடிவமைக்கப்பட்ட தரவுத்தொகுதி அடிப்படையிலான இயந்திரமொழிபெயர்ப்பு ஒழிங்குமுறைகளின் தரத்தை மேம்படுத்த உதவுகின்றன. மேலும், மொழிபெயர்ப்பிற்காகக் கருதப்படும் இரண்டு மொழிகளுக்கு இடையில் இருக்கக்கூடிய சிக்கலான மொழியியல் உறவுகள் குறித்த புதிய நுண்ணறிவுகளை அவை சேகரிக்கக்கூடும்.

8.6. மொழிபெயர்ப்பு தரவுத்தொகுதிகளில் மொழியியல் செயல்பாடுகள்

மொழிபெயர்ப்பு தரவுத்தொகுதி தொகுத்தலுக்கும் வரிசைப்படுத்தலுக்கும் பின்னர் மொழியியல் பகுப்பாய்வின் வேறுபட்ட நிலைகளுக்கு உள்ளாக்கப்படுவதற்கு முன்னர் வரிசைப் பொருத்தத்தின் பல நிலைகளைக் கடக்கும். இது மொழி பெயர்ப்பு நிகரன்களின் முறையாக்கத்தின் (வடிவாக்கத்தின்) நிலைகளுக்கு அடிப்படையாக அமையும். பொதுவாக மொழியியல் ஆய்வு பின்வருவனவற்றை உள்ளடக்கும்:

1. உருபன்களின் வடிவத்தையும் செயல்பாட்டையும் அடையாளம் காண உதவும் சொற்களின் உருபனியல் ஆய்வு (morphological analysis).

2. தொடர்புள்ள தரவுத்தொகுதிகளில் தொடர்களின் வடிவத்தையும் செயல்பாடுகளையும் அடையாளம் காண உதவும் தொடரியல் ஆய்வு (syntactic analysis)

3. விரிதரவுகளில் பயன்படுத்தப்பட்டுள்ள சொற்களின் புறவடிவங்களுக்கு இடையில் உள்ள இடைமுகத்தை ஆய உதவும் உருபனியல்-தொடரியல் ஆய்வு (morphosyntactic analysis) இதன் துல்லியமான மற்றும் திறமையான ஆய்வு பகுப்பாய்வின் பண்பையும் விரைவையும் அதிகரிக்கும்.

4. சொற்கள், தொடர்கள் போன்ற அலகுகளின் பொருண்மையைக் கண்டுபிடிக்கவும் இதில் வரும் பொருள் மயக்கங்களைக் கண்டுபிடிக்கவும் உதவும் பொருண்மையில் ஆய்வு (semantic analysis)

பயனுள்ள மொழியியல் பகுப்பாய்விற்கு, நிகழ்தகவு அளவீட்டுக்கான சில புள்ளியியல் அணுகுமுறைகளுடன், எளிய 'மேலோட்டமான' மற்றும் விளக்கமான உருபனியல்-தொடரியல் (morphosyntactic) அணுகுமுறையை (அதாவது சொல்வகைப்பாடு அடையாளப்படுத்தல் மற்றும் ஆழமில்லாப் பகுப்பாய்வு போன்றவை) பயன்படுத்துவதே சிறந்த முறையாகும்.

முன்னர் செயல்படுத்தப்பட்ட தரவுத்தொகுதிகளிடமிருந்து பெறப்பட்ட அடுக்குப்படுத்தப்பட்ட இலக்கணத்தால் (stratified grammar) பகுப்பாய்வி முன்னுரிமையாக ஆதரவு அளிக்கப்படுகிறது, ஏனெனில் இதுபோன்ற தரவுத்தொகுதிகள் அதன் பயன்பாட்டை அடுத்த கட்டங்களில் மீறுகின்றன.

செயலில் மொழிபெயர்ப்பு நினைவுகளை உருவாக்குவதே முக்கிய நோக்கம். பொருத்தமான இருமொழி சொற்களஞ்சிய பட்டியல்களை உருவாக்குதல், கணினி உதவியுடன் மொழி கற்பிப்பதற்கான பொருத்தமான எடுத்துக்காட்டுகளின் பட்டியல்களைப் பிரித்தெடுப்பது போன்ற பிற பயன்பாடுகளும் பரிந்துரைக்கப்படுகின்றன. சிறுகுறிப்பு செய்யப்பட்டவுடன், இந்த நிறுவனங்கள் மின்னணு அகராதிகளை மேம்படுத்துவதற்கும் இலக்கணங்களை உருவாக்குவதற்கும் அடிப்படையாகப் பயன்படுத்தப்படுகின்றன.

ஒரு சிறந்த சூழ்நிலையில், 'நிகழ்தகவு' நடைமுறையைப் பின்பற்றி இரு நிறுவனங்களிலும் சேர்க்கப்பட்டுள்ள உரை மாதிரிகளை ஒப்பிடுவதன் மூலம் பேச்சு-பகுதி குறியீட்டு செயல்முறை தானாகவே செய்யப்படுகிறது. இந்த வழியில், சில நேரங்களில் பெயரடைகள் பெயர்ச்சொற்கள் அல்லது நேர்மாறாக மொழிபெயர்க்கப்பட்டாலும், சாதாரண குறிப்பு அகராதிகளில் கொடுக்கப்பட்டுள்ள பிரிவுகள் இத்தகைய இலக்கண தெளிவற்ற தன்மைகளைத் தீர்க்க உதவுகின்றன.

பாரம்பரிய இலக்கண வகைகள் பேச்சு-பகுதி குறிச்சொல்லின் தரத்தில் வலுவான தாக்கத்தை ஏற்படுத்துகின்றன, ஏனெனில் குறைவான இலக்கண வகைகளைக் கொண்ட மொழிபெயர்ப்பு முறை, முழுமையான வகைகளின் பட்டியலைக் கொண்ட ஒரு அமைப்பை விட சிறந்த வெற்றி விகிதத்தைக் கொண்டுள்ளது (Chanod and Tapanainen 1995).

8.6.1 மொழிபெயர்ப்பு ஆய்வு

இயந்திர மொழிபெயர்ப்பு ஆய்வாளருக்கு இடையிலான முக்கியமான வாக்குவாதம் மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் ஆய்வில் உட்படும் கலவைத் தன்மையின் நிலையைப் பற்றியாகும். பொதுவான நம்பிக்கை இயற்கையான உரைகளில் வரும் பல எண்ணிக்கையிலான மொழியியல் நடத்தைகள் ஆயப்பட்டு வெளிப்படையாக உருப்படுத்தம் செய்யப்படும் வரை உயர்ந்த தன்மையுள்ள இயந்திர மொழிபெயர்ப்பு எய்த இயலாததாகும் என்பதாகும். சொல் மயக்கம் மற்றும் உறுப்பு பொருத்தம் இவற்றின் சிக்கல்களை மொழிபெயர்ப்பு விரிதரவுகளிலிருந்து பெறப்பட்டு உட்படுத்தப்பட்ட ஒரு மொழியின் அகராதியிலும் இலக்கணத்திலும் சேகரித்து வைக்கப்பட்டுள்ள ஏரளமான அறிவின் உதவியால் தீர்க்க இயலும். இது மொழிபெயர்ப்பு ஆய்வின் கடுமையான செயற்பாங்கின் பயன்படுத்தலை வேண்டும். சுருக்கமாக இயந்திர மொழிபெயர்ப்பில் மொழி ஆய்வு நுட்பத்திற்குப் பல பயன்பாடுகள் இருக்கின்றன. அவை பின்வருமாறு அமையும்:

1. மொழிபெயர்ப்புப் பகுப்பாய்வு நுட்பம் முன்னரே இருக்கக் கூடிய மொழிபெயர்ப்புகளை அமைப்பாக்கம் செய்ய உதவும். இதன் மூலம் அவை புதிய மொழிபெயர்ப்புகளின் உருவாக்கத்தில் மறு உபயோகம் செய்ய இயலும்.

எடுத்துக்காட்டாக, CWARCஇன் டிரான்ஸ் தேடல் அமைப்பு (Isabelle et al. 1993) என்பது இருமொழி தொடரடைவுக் கருவியாகும் (bilingual concordance tool), இது குறிப்பிட்ட மொழிபெயர்ப்புச் சிக்கல்களுக்கு ஆயத்தத் தீர்வுகளுக்காக (ready-made solutions) சிறப்பு நோக்க தரவுத்தளங்களில் (special-purpose databases) மொழிபெயர்ப்பாளர்களைத் தேட அனுமதிக்கிறது. வழக்கமாக, முன்பே இருக்கும் சில மொழிபெயர்ப்பு தரவுத்தொகுதிகளுக்கு வாக்கிய ஒழுங்குபடுத்தல் நுட்பங்களைப் பயன்படுத்துவதற்குப் பிறகுத் தரவுத்தளங்கள் ஆக்கப்படுகின்றன.

2. மொழிபெயர்ப்புப் பகுப்பாய்வு தொழில் நுட்பம் மொழிபெயர்ப்பு தவறுகளின் சில வகைகளைக் கண்டுபிடிக்க வேண்டி தொடக்க நிலை மொழிபெயர்ப்புகளில் பயன்படுத்தப்படுகிறது. மூல

மற்றும் வரைவு இலக்கு உரைக்கு இடையில் மொழிபெயர்ப்பு பொருத்தங்கள் புனரமைக்கப்பட்டவுடன், இந்த பொருத்தங்களுக்கு ஏதேனும் தடைகள் இருந்தால் சரிபார்க்க முடியும். எடுத்துக்காட்டாக, மூல உரையின் பெரிய கூறுகள் (எ.கா. பக்கங்கள், பத்திகள், பிரிவுகள், வாக்கியங்கள் போன்றவை) இலக்கு உரையில் சரியாக மொழிபெயர்க்கப்பட்டிருப்பதால், மொழிபெயர்ப்பு முழுமையானது என்பதை சரிபார்க்க முடியும்.

3. மொழிபெயர்ப்புப் பகுப்பாய்வு நுட்பம் போலியான ஓரினமொழிச் சொற்களால் ('deceptive cognates') ஏற்படும் குறுக்கீட்டு தவறுகளிலிருந்து சுதந்திரமாக மொழிபெயர்ப்பு இருந்தால் அதை பரிசோதிக்கப் பயன்படுத்தப்படும்.

8.6.2 இருமொழிய அகராதியின் உருவாக்கம் (Building Bilingual Dictionary)

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் மற்றொரு பயனுள்ள மற்றும் முக்கிய பகுதி இருமொழி அகராதிகளின் (Geyken 1997) உருவாக்கமாகும். இதன் குறைபாடு இந்திய மொழிகளின் இயந்திர மொழிபெயர்ப்பின் தடைகளில் ஒன்றாகும். மரபு அகராதிகள் இக்குறைபாட்டை நிவர்த்தி செய்ய இயலாது. ஏனென்றால் அவை சொல் துணை வகைப்பாடு, சொல்தெரிவு, சொல்தேர்வுக் கட்டுப்பாடு மற்றும் சொல் அலகுகளின் பயன்பாட்டு களங்கள் இவற்றைப் பற்றிய போதுமான தகவல்களைக் கொண்டிருக்கவில்லை. பல்வேறு புள்ளிவிவர முறைகளைப் பயன்படுத்துவதன் மூலம், சொல்சார் துணை வகைப்படுத்தல் (sub-categorisation) பற்றிய தகவல்கள் இருமொழி அகராதியில் சேர்க்கப்பட, சொல்வகைப்பாட்டிற்கு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகளிலிருந்து தானாகவே பிரித்தெடுக்கப்படுகின்றன (Brown 1999).

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகள் கிடைக்காதபோது, அடையாளப்படுத்தப்படாத தரவுத்தொகுதிகளிலிருந்து தயாரிக்கப்படும் அகராதிகள் தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயப்புக்குச் சமமாக பயனுள்ளதாக இருக்கும். ஓரினம்சார்ந்த மொழிகளுக்கிடையில் (cognate languages) இருமொழியச் சொல் அகராதியை உருவாக்குவது மிகச் சிறந்தது, அவை ஒன்றுக்கொன்று பொதுவாகவும் / அல்லது பரம்பரை ரீதியாகவும் தொடர்புடையவையாகவும் இருக்கும் (எ.கா. தமிழ் மற்றும் மலையாளம், இந்தி மற்றும் உருது போன்றவை); ஏனெனில் இதுபோன்ற ஓரின மொழிகள் பொதுவாக பல பொதுவான பண்புகளைப் பகிர்ந்து கொள்கின்றன (இரண்டும் மொழியியல்சார் மற்றும் மொழியியல்சாரா); பிற மொழிகளில் அரிதாகவே காணப்படுகிறது.

மேலும், ஒலிப்பு மற்றும் எழுத்து பிரதிநிதித்துவங்களில் மட்டுமல்லாமல், அவற்றின் அர்த்தம், உள்ளடக்கம் (பொருண்மை) மற்றும் உட்கோள் (implication) ஆகியவற்றிலும் ஒன்றுக்கொன்று ஒத்த வழக்கமான சொற்றொகையின் பெரிய பகுதி உள்ளது. எடுத்துக்காட்டாக, தமிழ் மற்றும் மலையாளம் ஆகிய இரண்டு ஓரின இந்திய மொழிகளுக்கு ஒத்த வழக்கமான சொற்றொகையின் மாதிரி பட்டியலை மேற்கோள் காட்டும் பின்வரும் அட்டவணையை (அட்டவணை 8.1) கவனியுங்கள்.

அட்டவணை 8.1: தமிழுக்கும் மலையாளத்திற்கும் இடையிலான சொல்லகராதி பட்டியல்களில் ஒற்றுமை வேற்றுமை

சொற்கள்	**தமிழ்**	**மலையாளம்**
உறவுப்பெயர்கள்	அம்மா, அப்பா, அத்தை, அண்ணன், தம்பி, மகன், மகள்	அச்சன், அம்மே, சேட்டன், அனியன், மோன், மோள்
மாற்றுப்பெயர்	அவன், அவள், அவர், அது, நான், நீ,	அவன், அவள், அவர், அது, ஞான், நீ
பெயர்	உலகம், வீடு, கை, மொழி	லோகம், வீடு, கையு, பாஷ
பெயரடை	நல்ல, கெட்ட, இனிய, சிவந்த	நல்ல, சீத்த, மதுரமுள்ள, சொவன்ன
வினை	வா, போ, இரு, கொடு, சொல், அடி, குடி	வரு, போகு, கொடுக்கு, பறயு, அடிக்கு, குடிக்கு
பின்னொட்டு	முன்னால், பின்னால், கீழே, மேலே	முன்பு, புறகே, தாழே, மேலே
திரிபுறாதவை	ஆனால்,	பக்ஷே

இருமொழி அகராதியின் உருவாக்கத்திற்கு, குறிக்கப்பட்ட/அடையாளப்படுத்தப்பட்ட மற்றும் பாகுபடுத்தப்பட்ட தரவுத்தொகுதிகளில் பல்வேறு வழிகளில் பயன்படுத்தப்படும் பல்வேறு புள்ளியியல் முறைகளைப் பயன்படுத்த வேண்டும். புள்ளியியல் நடைமுறைகளைப் பயன்படுத்துவதில் வகைகள் இருந்தாலும், பெரும்பாலான சந்தர்ப்பங்களில், குறிக்கோள் பின்வருமாறு அமையும்:

•இருமொழி மொழிபெயர்ப்பு தரவுத்தொகுதியிலிருந்து ஒப்பிடக்கூடிய பெரிய தொடரியல் தொகுதிகள் (எ.கா. எச்சத்தொடர்கள், சொற்றொடர்கள் போன்றவை) மீட்டெடுப்பு.

•இரண்டு தரவுத்தொகுதிகளின் குறிக்கப்பட்ட/அடையாளப்படுத்தப்பட்ட தரவுத்தளங்களிலிருந்து பல்வேறு துணை வகைப்படுத்தப்பட்ட கூறுகளை (எ.கா. எழுவார், செயப்படுபொருள், பொருள், பயனிலை போன்றவை) பிரித்தெடுத்தல்.

•அடிக்கடி நிகழும் வினையடை எச்சத்தொடர்கள், பெயரெச்சத் தொடர்கள், மரபுச்சொல் வெளிப்பாடுகள் மற்றும் தொகுப்புத் தொடர்கள் போன்றவற்றை பிரித்தெடுப்பது.

•தரவுத்தொகுதிகளிலிருந்து வடிவத்தில் உள்ள ஒற்றுமைகள் மற்றும் பொருண்மை காரணங்களால் மொழிபெயர்ப்புக்கு நிகரன்களாகக் கருதப்படுகிற சொல்சார் ஐட்டங்களைத் தேர்ந்தெடுப்பது.

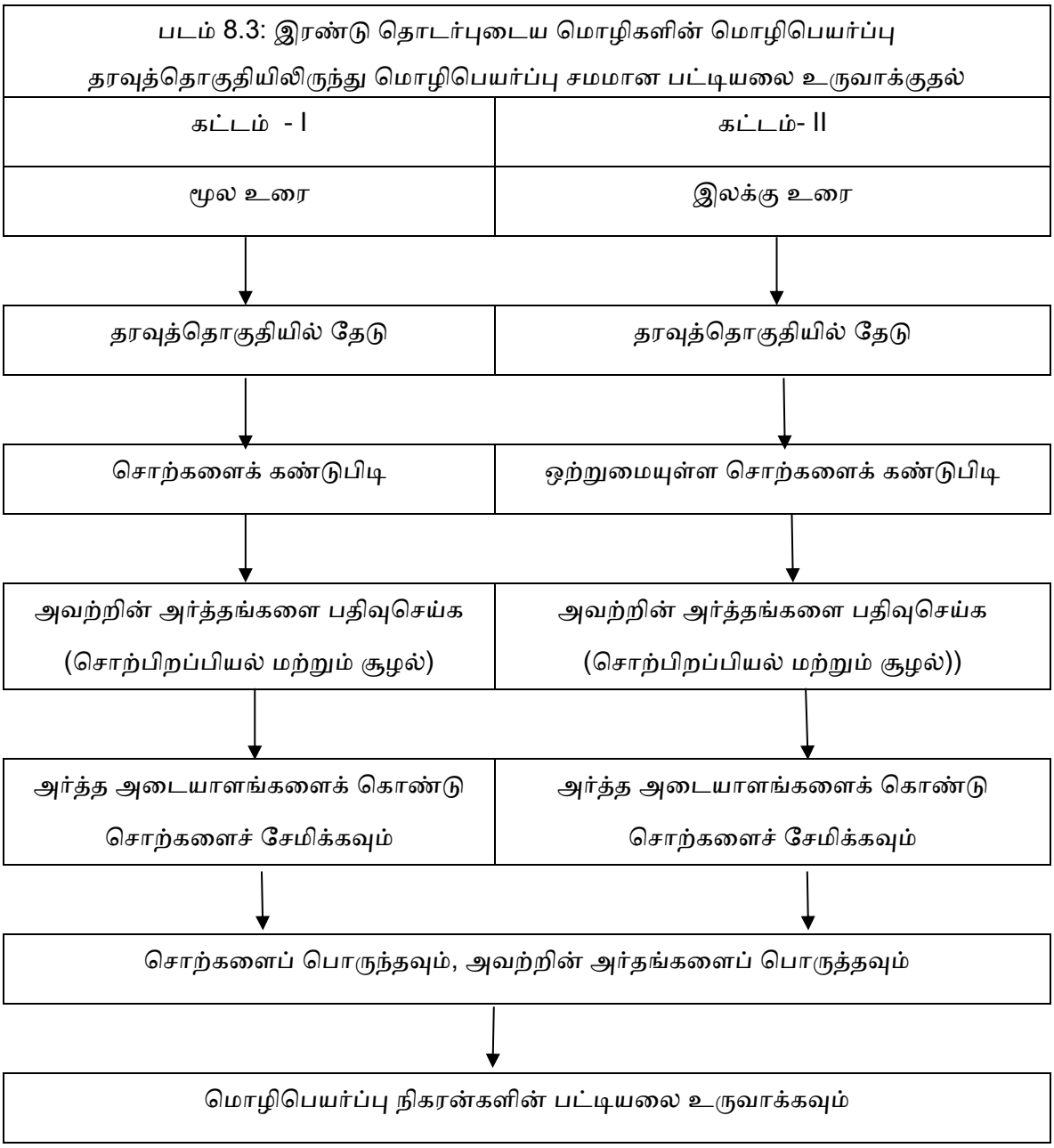
இவை இரண்டு மொழிகளுக்குள் உருபனியல், சொல்சார், தொடரியல்சார் பொருண்மையியல்சார் மற்றும் கருத்துரு மட்டங்களில் நூறு சதவீத ஒற்றுமையை எதிர்பார்க்கவில்லை, இருப்பினும் இவை நெருக்கமாக ஒன்றோடொன்று தொடர்புடையவை. ஆகையால், மொழியியல் கோட்பாடுகள் மற்றும் தரவுத்தொகுதிகளின் அனைத்து அறிவுத் தளங்களுடனும், ஒரு மைய/முக்கிய இலக்கணம் சிறந்த தீர்வாகும், இது இன மற்றும் / அல்லது வகைபாட்டியல் தொடர்பான இந்திய மொழிகளில் இன்னும் உருவாக்கப்படவில்லை. இந்திய மொழிகளிடையே தரவுத்தொகுதியை முறையாக பகுப்பாய்வு செய்வதன் மூலம் தகவல் கிடைப்பதன் மூலம், இருமொழி அகராதி அகராதியின் உருவாக்கத்தில் நாம் ஓரளவு வெற்றியை அடைய முடியும்.

8.6.3 மொழிபெயர்ப்பு நிகரன்களின் பிரித்தெடுப்பு (Extraction of Translation Equivalents)

மொழிபெயர்ப்பு தரவுத்தொகுதிகளில் மொழிபெயர்ப்பு நிகரன்கள் தேடல் மூல மொழியில் ஒரு குறிப்பிட்ட பொருண்மையை அல்லது கருத்துருவை வெளிப்படுத்தும் ஒரு குறிப்பிட்ட வடிவிலிருந்து தொடங்கப்பட வேண்டும். மூல மொழியில் இச்சொல் கண்டுபிடிக்கப்பட்டு தனியான தரவு மையத்தில் பதிவு செய்யப்பட்ட பின் தேடல் செயற்பாங்கின் 2-வது பகுதி தொடங்கும். இது இலக்குமொழியில் ஒரே அர்த்தத்தைக் கொண்டிருக்கிற சொல்லின் அடையாளம் காணலை உட்படுத்தும். இயல்பாகப் பெரும்பாலான மொழிபெயர்ப்பு விரிதரவுகள் மொழிபெயர்ப்பு நிகரன்களின் மிகுந்த பரப்பெல்லையை வெளிப்படுத்த இயலும். அவை மாற்று வடிவின் பரிசோதனைக்கு முக்கியமான மூலமாக அமையும். இருப்பினும் பொருத்தமான நிகரன்களின் தேர்வை உறுதி செய்யும் காரணிகள், சொற்களின் பயன்பாட்டின் மறு நிகழ்வு அமைப்பொழுங்கின் அடிப்படையில் உறுதி செய்யப்படும். தேவைப்பட்டால், ஒவ்வொரு உரையிலும் காணப்படும் இந்த சமமான வடிவங்கள் நூல்கள் குறிப்பிடப்படும் இரண்டு மொழிகளின் அசல் ஒருமொழி நிறுவனத்துடன் மேலும் சரிபார்க்கப்படுகின்றன.

பின்வரும் வரைபடம் (படம்) மொழிபெயர்ப்பு தரவுத்தொகுதிலிருந்து மொழிபெயர்ப்பு நிகரன்களின் பட்டியல் எவ்வாறு உருவாக்கப்படுகிறது என்பது பற்றிய எளிய திட்டத்தைக்

காட்டுகிறது. இருப்பினும், நெருங்கிய தொடர்புடைய இரண்டு மொழிகளில் கூட, மொழிபெயர்ப்பு நிகரன்கள் எல்லா சூழல்களிலும் ஒரே பொருளைக் குறிப்பதில்லை; ஏனெனில் இவை ஒரே மாதிரியான தொடரியல் கட்டுமானத்தில் அரிதாகவே பயன்படுத்தப்படுகின்றன. மேலும், மொழிச் சார்ந்த சூழல்களைப் பொறுத்து அவற்றின் பொருண்மைசார் உணர்பொருள் (semantic connotations) மற்றும் சம்பிரதாயத்தின் அளபு வேறுபடலாம். இலக்கு மொழியில் ஒரு சொல் அலகு மூல மொழியில் சொல்லனுக்குச் (லெம்மா/lemma) சமமான சரியான மொழிபெயர்ப்பு நிகரனாகப் பயன்படுத்தப்பட்டாலும் கூட, அந்தச் சொல் அலகின் மொழிபெயர்ப்பு எப்போதும் மூலமொழியில் சொல்லனாக இருக்காது. ஒரு எளிய, இருவழி மொழிபெயர்ப்பு விலங்குகள் அல்லது கருவிகளின் பெயர்களிலும் சில அறிவியல் சொற்களிலும் சாத்தியமாகும், ஆனால் சாதாரண மொழிகளில் அரிதாகும் (Landau 2001: 319). சாதாரண மொழிகளுக்கான தரவுத்தொகுதிசார் மொழிபெயர்ப்பு ஒழுங்குமுறை பல சிக்கல்களை எதிர்கொள்ளும் என்பதை இது குறிக்கிறது, இது சிறந்த வெளியீடுகளை வழங்குவதற்கு அதிக அளவு மொழியியல் நுட்பம் தேவைப்படும். குறைவான சிக்கல்கள் உள்ள இடங்களில், அறிவியல் மற்றும் தொழில்நுட்ப மொழியைப் பொறுத்தவரை, தரவுத்தொகுதிசார் மொழிபெயர்ப்பு ஒழுங்குமுறை சிறந்த முடிவுகளைக் கொண்டிருக்கும்.



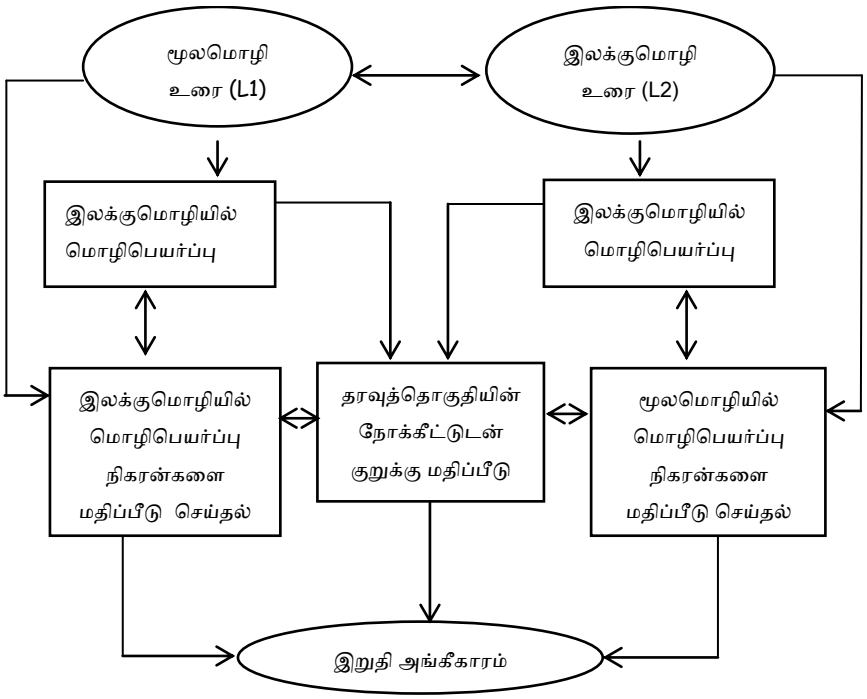
மொழிபெயர்ப்பு, சமமான அலகுகளை மொழிபெயர்ப்புத் தரவுத்தொகுதியிலிருந்து பிரித்தெடுப்பது தரவுத்தொகுதிசார் மொழிபெயர்ப்பில் உள்ள நிபுணர்களுக்கு மட்டுமல்ல, மற்றவர்களுக்கும் பன்மொழிப் பகுதிகளில் வளங்களை உருவாக்க உதவும்; ஏனென்றால் மொழிபெயர்ப்புத் தரவுத்தொகுதிகள் பொருத்தமான மொழிபெயர்ப்பு நிகரன்களைக் கண்டறிய

அவர்களுக்கு அதிகாரம் அளிக்கும். பொதுவாக, மொழிபெயர்ப்புத் தரவுத்தொகுதிகள் மொழி புலனாய்வாளர்களை பின்வருமாறு உதவும்:

- சொற்கள், மரபுச்சொற்கள், கூட்டுச்சொற்கள், சொல்லடிவகைப்பாடு மற்றும் சொற்றொடர்கள் உள்ளிட்ட நல்ல மொழிபெயர்ப்பு நிகரன் அலகுகளை மீட்டெடுத்தல்;
- இலக்குமொழியின் 'இயல்பான தன்மையை' காட்டும் மொழிபெயர்க்கப்பட்ட உரைகளை உருவாக்கத் தரவுத்தொகுதித் தரவு எவ்வாறு உதவுகிறது என்பதை அறிதல்;
- பயனர்களுக்கு குறைந்த அளவிலான கட்டளை கொண்ட வெளிநாட்டு மொழியில் சரியாக மொழிபெயர்க்க உதவும் புதிய மொழிபெயர்ப்புத் தரவுத்தளங்களை உருவாக்குதல்;
- புதிய உரைகளில் உள்ள கலைச்சொல்சார் பொருட்களின் பெரும்பகுதி தரப்படுத்தப்படவில்லை அல்லது 'கால வங்கியில்' பதிவு செய்யப்படவில்லை என்பதால், பெரிய கலைச்சொல்சார் தரவுத்தளத்தை உருவாக்குதல்.

மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து நிகரான வடிவங்களையும் அலகுகளையும் பிரித்தெடுக்கும் செயல்முறை மற்றும் ஒருமொழி தரவுத்தொகுதிகளுடன் அங்கீகாரத்திற்கான அவற்றின் அடுத்தடுத்த சரிபார்ப்பு ஆகியவை பின்வரும் வழியில் விவரிக்கப்பட்டுள்ளன (படம்). மொழிபெயர்ப்புத் தரவுத்தொகுதிகளிலிருந்து நிகரான அலகுகளைக் கண்டுபிடிக்க, ஒப்பிடக்கூடிய அலகுகளின் பொண்மைகளைக் கண்டறிய பல்வேறு தேடல் முறைகள் பயன்படுத்தப்படுகின்றன; அவை பெரும்பாலும் எளிய சொற்களைக் காட்டிலும் பெரியவை மற்றும் சிக்கலானவை. மொழிபெயர்ப்புத் தளங்களில் செயல்படுத்தப்பட்ட இந்தத் தரவுத்தளங்கள் வழக்கமான மொழிபெயர்ப்பு நினைவுகளை (translation memories) விட மொழிபெயர்ப்புகளை எளிதாக்குகின்றன. தரவுத்தொகுதி கண்டுபிடிப்புகளை இருமொழி அகராதிகள், பன்மொழி கலைச்சொல் வங்கிகள் மற்றும் மொழிபெயர்ப்பு நிகரன்களின் தரவுத்தளங்களுடன் ஒருங்கிணைக்க முடியும்.

படம் : அங்கீகாரத்திற்கான மொழிபெயர்ப்பு நிகரன்களின் சரிபார்ப்பு



8.6.4 கலைச்சொல் தகவல் வங்கியின் உருவாக்கம் (Generation Of Terminology Databank)

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்புக்குத் தொழில் நுட்ப மற்றும் அறிவியல் கலைச்சொற்களில் பொருத்தமான ஆக்கத்திற்குக் கலைச்சொல் ஆக்கவியலாரின் எண்ணத்தின் அடிப்படையிலான ஆய்வுத் தேவை. கலைச்சொல் ஆக்கவியலாரின் அடிப்படைச் செயல்பாடு இலக்கு மொழிக்கு அயல் கருத்துக்களையும் கருத்துருக்களையும் உருப்படுத்தம் செய்ய மிகப் பொருத்தமான அல்லது ஓரளவுக்குப் பொருத்தமான சொற்களைத் தேர்வு செய்வதாகும். இதைச் செய்வதற்குக் கலைச்சொல் ஆக்கவியலார் இலக்குமொழி பயன்படுத்துபவர்களுக்கிடையில் கலைச்சொற்களின் பொருத்தமான தன்மை, இலக்கணத் தன்மை, ஏற்றுக்கொள்கை மற்றும் பயன்பாடு போன்ற பல்வேறு காரணிகளை மனதில் கொள்ள வேண்டும். முக்கியமான காரணி கலைச்சொற்களின் தரம் ஆகும். இதன்படி மொழியியல் கொள்கைகளின் அடிப்படையில் பொருத்தமான வழிகளைப் பயன்படுத்தி புதிய சொற்கள்

உருவாக்கப்படும். மொழிபெயர்ப்பு தரவுத்தொகுதி ஒரு கருத்து, நிகழ்ச்சி, விஷயம் அல்லது கருத்துரு என்பதை உருப்படுத்தம் செய்யப் பலரால் உருவாக்கப்பட்ட பல கலைச்சொற்களின் பெரிய பட்டியலிலிருந்து பொருத்தமான ஒன்றைத் தேர்வு செய்யும் முக்கிய பங்களிப்பைச் செய்கின்றது. புதிய சொற்கள் உருவாக்கத்தின் தொடர்ச்சியான முயற்சி விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அமைப்பைத் திட்டமிடுபவர்களுக்கு ஒன்றுக்குப் பதிலாக ஒன்றைத் தேர்ந்தெடுப்பதில் சிக்கலைத் தரும். புதிய சொல்லை உருவாக்குவதா அல்லது இலக்குமொழியில் பயன்பாட்டால் இயல்பாக்கம் செய்யப்பட்ட மூலமொழிச் சொல்லைப் பயன்படுத்துவதா என்ற விவாதம் எழும். சில தொழில நுட்பச் சொற்கள் அவற்றின் உண்மையான மூலத்தைக் கண்டுபிடிக்க இயலாதவாறு இயல்பாக்கம் செய்யப்பட்டிருக்கும். இதுபோன்ற சந்தர்ப்பங்களில், விதிமுறைகள் ஏற்கனவே இலக்குமொழியில் 'உலகளாவிய ஏற்றுக்கொள்ளல்' தரத்தை அடைந்துவிட்டதால் நமக்கு சிக்கல் இல்லை. எடுத்துக்காட்டாகத் தமிழ் பயனர்களுக்கு *computer, mobile, calculator, telephone, tram, bus, cycle, taxi, rickshaw, train, machine, pen, pencil, pant, road, station, platform,* (கம்பியூட்டர், மொபைல், கால்குலேட்டர், டெலிபோன், டிராம், பஸ், சைக்கிள், டாக்ஸி, ரிக்ஷா, டிரையின், மெஷின், பேனா, பென்சில், பேன்ட், ரோட், ஸ்டேஷன், பிளாட்பாரம்) போன்ற சொற்களைப் புரிந்து கொள்வதில் அதைக சிரமம் இல்லை. ஏற்கனவே அந்தப் பொருட்களுடன் தமிழில் அவற்றின் மூலமொழிப் பெயர்களும் ஏற்றுக்கொள்ளப்பட்டுள்ளன. அன்றாட வாழ்க்கையில் அவற்றைப் பயன்படுத்தும் உயர்ந்த நிகழ்வெண் அவற்றைச் சொற்றொகையின் பகுதியாக மாற்றும். எனவே மொழிபெயர்பின் போது அவற்றை இடம் பெயர்க்கத் தேவையில்லை. பொதுவாக இலக்கு மொழியில் மொழி பெயர்க்கப்பட்ட விரிதரவு மூல மொழியிலிருந்து கடன் வாங்கப்பட்ட புதிய கருத்துக்களையும் கருத்துருக்களையும் வெளிப்படுத்த பொருத்தமான சொற்களைத் தேர்ந்தெடுக்க நல்ல மூலவளமாகும்.

இந்தத் தரவுத்தொகுதிகள் விவேகமான முடிவுகளை எடுப்பதற்குத் தேவையான எடுத்துக்காட்டுகள் மற்றும் நோக்கீடுகளுடன் கலைச்சொல்லாக்கவியலார்களை (terminologists) ஆதரிப்பதற்கான பல்வேறு வகையான சொற்கள், மரபுச்சொற்கள் மற்றும் தொகுப்புத் தொடர்களைக் கொண்டுள்ளன . அவை இரண்டு முக்கியமான வழிகளில் பங்களிக்கின்றன: (i) மொழியில் நுழைந்த அனைத்து தொழில்நுட்ப சொற்களையும் சொற்றொடர்களையும் அவற்றின் தேதி, புலம் மற்றும் நுழைவுக் களத்துடன் இணைக்க அவைகள் கலைச்சொல்லாக்கவியலார்களை

ஆதரிக்கின்றன; (ii) அவை கலைச்சொற்கள் மற்றும் சொற்றொடர்களின் அனைத்து சொந்த ஆக்கங்களையும் அவற்றின் தொடர்புள்ள களம் மற்றும் பயன்பாட்டின் அதிர்வெண்ணுடன் வழங்குகின்றன. இந்த இரண்டு காரணிகளும் இறுதியில் கலைச்சொல்லாக்கவியலார்களை ஒப்பீட்டளவில் ஏற்றுக்கொள்வது அல்லது நிராகரிப்பதை தீர்மானிக்க உதவுகின்றன. தரவுத்தொகுதியிலிருந்து எடுக்கப்பட்ட சில நேர்வுகளை ஆராய்வது, பொருத்தமான தொழில்நுட்பச் சொற்களை, அதாவது மொழிபெயர்ப்பிற்கான அத்தியாவசியக் கூறுகளைத் தேர்ந்தெடுப்பதில் தரவுத்தொதொகுதி எவ்வாறு பயன்படுகிறது என்பதைக் காட்டுகிறது.

8.6.5 சொல் தேர்வுக் கட்டுப்பாடு (Lexical selection restriction)

மூலமொழியிலிருந்து தெரிந்தெடுக்கப்பட்ட மிகப் பொருத்தமான நிகரன்களாகக் கருதப்பட வேண்டிய இலக்கு மொழிச் சொற்களின் தெரிவு மூலமொழியிலும் இலக்குமொழியிலும் பட்டறிவுள்ள மொழியியலாரின் திறமையான மற்றும் கவனமான பொருள்கோளை வேண்டும். மற்றொரு கலவைத்தன்மையான செயல்பாடாகும். இதன் பொருள் என்னவென்றால் மொழியியலார் இலக்கு மொழியல் இருக்கின்ற ஒரே பொருண்மையை தருகின்ற வடிவங்களின் பெரிய தொகுப்பிலிருந்து பொருத்தமான சொல்லைத் தேர்ந்தெடுக்க வேண்டும் என்பதாகும். பின்வரும் எடுத்துக்காட்டுகளைக் கவனத்தில் கொள்ளவும்:

ஆங்கிலம் : He ate the food

தமிழ் : அவர் போஜனம் செய்தார். (முனிவர் போன்ற மிக உயர்நிலையில் உள்ள எழுவாய்)

அவர் உணவு உண்டார் (உயர்நிலையில் உள்ள எழுவாய்)

அவன் உணவு சாப்பிட்டான். (சாதரண நிலையில் உள்ள எழுவாய்)

அது நாய் சாப்பாட்டை தின்றது (நாய் போன்ற விலங்கு எழுவாய்)

மேற்சொன்ன எடுத்துக்காட்டுகளில் நாம் தமிழில் எழுவாயின் பண்பின்/அந்தஸ்தின் அடிப்படையில் சரியான நிகரன் சொற்களைத் தெரிந்தெடுக்க வேண்டியுள்ளது. மொழியியலாரின் முக்கியமான செயல்பாடு இரு மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பில் வேறுபட்ட சமூக மொழியியல் காரணிகளைப் பொருத்தமான சொல் அலகுகளைக் காண்பதாகும். இம்மாதிரியான எடுத்துக்காட்டுகள் வெற்றிகரமான மற்றும் அறிவு பூர்வமான மொழிபெயர்ப்பில் சொல் தேர்வு கவனமாகக் கையாளப்பட வேண்டும் என்பதைக் காட்டப்படுகிறது. இச்சிக்கல் மனித மொழிபெயர்ப்பில் கவனமாகக் கையாளப்பட்டாலும் பெரும்பாலான இயந்திர மொழிபெயர்ப்பில் புறக்கணிக்கப்படுகின்றன. இச்சிக்கலைத் தீர்ப்பதற்கான நல்ல வழி இயந்திரத்தால் படிக்க

இயலும் அகராதியின் (machine readable dictionary (MRD)) தனியான இடத்தில் பொருண்மையில் ஒற்றுமையைக் காட்டும் எல்லா வடிவங்களைப் பட்டியலிடுவதாகும். இம்மாதிரியான தரவுமையங்களை மொழிபெயர்ப்பு விரிதரவுகளிலிருந்து பிரிந்தெடுக்க இயலும். பொதுவாக இயந்திரம் படிக்க இயலும் அகராதிகள் பல பாடப்பொருள் களங்களாகப் பிரிக்கப்பட்டிருக்கும். இதனால் ஒரு குறிப்பிட்ட பாடப்பொருளைச் சார்ந்த சொற்களைப் பற்றிய பொருத்தமான தகவலைப் பெற இயலும்.

ஆகையால், ஒரு உரையை ஆயும் போதெல்லாம், வடிவமைப்பாளர்கள் தாலைப்பு பகுதியைத் (subject area) தேர்ந்தெடுப்பர்; இது இலக்குமொழியில் மொழிபெயர்க்கப்பட வேண்டிய உரை வகைக்கு மிகவும் பொருத்தமானது என்று அவர்கள் கருதுகிறார்கள். எடுத்துக்காட்டாக, அரசியல் தொடர்பான ஒரு ஆங்கில உரை தமிழில் மொழிபெயர்க்கப்பட வேண்டுமானால், மீதமுள்ள லெக்சிக்கல் தரவுத்தளத்தைத் தேடுவதற்கு முன்பு, ”அரசியல்’ என்ற பொருளில்/தலைப்பில் கலைச்சொல் தரவுவங்கியில் சேமிக்கப்பட்டுள்ள இலக்குமொழியின் பொருத்தமான சொற்களைத் தேட மொழிபெயர்ப்பு முறைக்கு அறிவுறுத்துவது அர்த்தமுள்ளதாக இருக்கும். எடுத்துக்காட்டாக, morphology என்ற என்ற சொல் அதன் பயன்பாட்டின் பொருள் பகுதியைப் (subject area) பொறுத்து பல அர்த்தங்களைக் கொண்டுள்ளது, அது கீழே காட்டப்பட்டுள்ளது. இங்கு கொடுக்கப்பட்டுள்ள எடுத்துக்காட்டுகள், மொழிபெயர்ப்பாளர்கள் பொருள் பகுதியைக் கருத்தில் கொண்டு மிகவும் பொருத்தமான சொல் அலகைத் தேர்ந்தெடுக்க வேண்டும் என்பதை வலியுறுத்துகின்றன. இந்தச் சிக்கல்களைக் கருத்தில் கொள்ளும் வரை, மூலமொழியிலிருந்து இலக்குமொழியில் பொருத்தமான வெளியீட்டை அடைய இயலாது.

morphology : புறவடிவயியல், புறவமைப்பியல், உருவியல், உடல் உருவ அமைப்பு , உறுப்பு உரு அமைப்பு, உருபனியல்

மொழியியல் பொருள்/தலைப்புச் சூழலில் morphology என்பதற்கு ‘உருபனியல்’ என்ற நிகரன் தேர்ந்தெடுக்கப் படவேண்டும். மனித உடல் சூழலில் ’உருவியல்’ என்ற நிகரர்ன் தேர்ந்தெடுக்கப்படவேண்டும்.

களங்கள் அல்லது கருத்தாடல் துறைகளில் மூலமொழிச் சொற்களைப் பயன்படுத்துவதற்கான சூழல் கருத்தில் கொள்ளப்படாவிடில் பொருத்தமான தொழில்நுட்ப சொற்களை இலக்குமொழியில் தேர்ந்தெடுப்பது மேலும் சிக்கலானதாக முடியும்.

எடுத்துக்காட்டாக, ஆங்கிலத்தில் (மூலமொழி) deliver என்ற சொல் தமிழில் (இலக்குமொழி) மொழிபெயர்க்கப்பட்ட பின்வரும் எடுத்துக்காட்டுகளைக் கவனிக்கவும்.

(1) ஆங்கிலம்: Mrs. Jonathan has delivered the girl child in a hospital.

Tamil: திருமதி ஜோனாதன் ஒரு பெண் குழந்தையை மருத்துவ இல்லத்தில் பிரசவித்திருக்கிறாள்.

(2) ஆங்கிலம்: Prof. Brown delivered a fine lecture on child education.

தமிழ்: பேராசிரியர் பிரவுண் குழந்தைகள் கல்வி பற்றி ஒரு விரிவுரை ஆற்றினார்

(3) ஆங்கில: The courier boy delivered a parcel in the evening.

தமிழ்: கொரியர் பையன் ஒரு பார்சலை மாலையில் வினியோகித்தார்.

மேலே குறிப்பிடப்பட்ட எடுத்துக்காட்டுகள் மூலமொழியில் மூன்று தனித்துவமான அர்த்தங்களைக் கொண்டுள்ளன, அவை இந்த வார்த்தையின் பயன்பாட்டின் கருத்தாட்லைக் கருத்தில் கொண்டு மிகவும் பொருத்தமான முறையைப் பின்பற்றுவதன் மூலம் கைமுறையாக இலக்குமொழியில் மொழிபெயர்க்கப்பட்டுள்லன. பிரசவப் புலத்தில்/துறையில் deliver என்பதற்கு இலக்கு மொழியில் மிகவும் பொருத்தமான சொல் 'பிரசவி' அல்லது 'பெறு' என்பனவாகும்; ஆனால் கூட்டம் அல்லது மக்கள் பேரணி சொற்பொழிவுப் புலத்தில்/தலைப்பில் deliver என்பது இலக்குமொழியில் 'ஆற்று', 'செய்' (முன்னர் வரும் 'விவுரை' என்பதன் ஆதரவுதடன்) என மொழிபெயர்க்கப்படும். கடிதம், பார்சல் விநியோகப் புலத்தில் deliver என்பது 'வினியோகி', 'தரு' என மொழிபெயர்க்கப்படும். மேற்கண்ட எடுத்துக்காட்டுகளில் மிகவும் குறிப்பிடத்தக்க அம்சம் என்னவென்றால் மூலமொழியில் deliver girl child, deliver lecture, deliver parcel என்பன சொல்தேர்வுக் கட்டுப்பாடு அடிப்படையில் மூன்று செயல்களை உணர்த்தும். இந்த சொல்தேர்வுக் கட்டுப்பாடு இலக்குமொழியில் வெவேறு வினைகளின் தேர்வு அடிப்படையில் வெளிப்படையாக உணர்த்தப்படும். அதாவது, மூலமொழியில் சொற்களைப் பயன்படுத்துவது பற்றிய கருத்தாடல் பயன்பாட்டைக் கருத்தில் கொண்டு, மொழிபெயர்ப்பாளர்கள் மூலமொழியிலிருந்து பொருத்தமான சொல்லைத் தேர்ந்தெடுக்க வேண்டும். எல்லா சந்தர்ப்பங்களிலும், தரவுத்தொகுதியிலிருந்து எடுக்கப்பட்ட சான்றுகள் பொருத்தம் மற்றும் ஏற்பு அடிப்படையில் மொழிபெயர்ப்பு வெளியீடுகளை ஏற்றுக்கொள்வதை நியாயப்படுத்தும். பின்வரும் எடுத்துக்காட்டும் மேற்சொன்ன கருத்துக்களை உறுதி செய்யும்.

(1)ஆங்கிலம்: She has conceived the central idea of the poem.

தமிழ்: அவள் பாடலின் மையக்கருத்தைப் புரிந்துகொண்டாள்.

(2) ஆங்கிலம்: She has conceived within a year of her marriage.

தமிழ்: அவள் திருமணமாகி ஒரு ஆண்டுக்குள் கருத்தரித்தாள்.

முதல் எடுத்துக்காட்டில் வகுப்பில் மாணவர்களைப் புரிந்துகொள்வதைக் குறிக்க conceive பயன்படுத்தப்பட்டுள்ளதால் பயன்பாட்டின் கருத்தாடலைப் பொறுத்து, conceive என்பது 'புரிந்துகொள்' எனத் தமிழில் மொழிபெயர்க்கப்பட்டுள்ளது என்பதை எடுத்துக்காட்டுகள் காட்டுகின்றன. புதிதாக திருமணமான மணமகளின் கர்ப்பமாவதைக் குறிக்க அதே வார்த்தையை மீண்டும் 'கர்பந்தரி' என்று மொழிபெயர்க்கப்பட்டுள்ளது. சூழல்கள் conceive என்பதை இரண்டு கருத்துருக்களாகப் பிரிப்பதை உணரலாம்.

பொருத்தமான சொற்றொடர்கள், தொகுப்பு சொற்றொடர்கள், மரபுச்சொற்கள் மற்றும் பழமொழிகளைத் தேர்ந்தெடுப்பது மற்றொரு முக்கியமான பகுதியாகும், இது பொருத்தமான மொழிபெயர்ப்புகளைத் தயாரிப்பதற்கு மொழித் தரவுத்தொகுதிகளின் மூலம் கவனமாகத் தேட வேண்டும் (Geyken 1997). சிறந்த தீர்வு, இதுபோன்ற ஒவ்வொரு வகை வெளிப்படுத்தங்களுக்கும் பொருத்தமான சொல் அலகுகளை இருமொழி அகராதிகளின் உருவாக்குவதும் மற்றும் இயந்திரம் படிக்கவியலும் அகராதியில் பொருத்தமான சூழல் குறிப்புகளுடன் சேகரித்து வைப்பதும் ஆகும்.

அட்டவணை: ஆங்கிலம் மற்றும் தமிழிலிருந்து தொகுப்பு சொற்றொடர்கள் மற்றும் மரபுச்சொற்களின் தேர்வு	
ஆங்கிலம்	**தமிழ்**
White lies	நம்ப தகுந்த பொய்கள்
Crocodile's tear	நீலிக் கண்ணீர்
By hook or by crook	தந்திரமாக
Once in a blue moon	அரிதாக, எப்பொழுதாவது ஒருமுறை
Black sheep	குழு கேடர்
Too many cooks spoil the broth	பல மரம் கண்ட தச்சன் ஒரு மரமும் வெட்டான்
Kicked the bucket	இற
Time and tide wait for none.	ஐயர் வரும்பரை அம்மாவாசை காத்திருக்குமா.

Strike the iron while it is hot.	காற்றுள்ளபோதே தூற்றிக்கொள்.
A white elephant	பலனுக்குமேல் செலவைத்தரும் ஒரு உடைமை

மொழிபெயர்ப்பு தரவுத்தொகுதிகளியிலிருந்து இத்தகைய சொற்றொடர் பட்டியல்களை உருவாக்குவது தரவுத்தொதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் தரம் மற்றும் வலுவான தன்மையை மேம்படுத்துகிறது, ஏனெனில் இந்த தரவுத்தளங்கள் மூலமொழி மற்றும் இலக்கும்மொ ஆகிய இரண்டிலும் வெளிப்படுத்தப்பட்ட உருவக உணர்வை உள்ளடக்கியது.

8.6.6 சொல் மயக்கத்தை நீக்குதல்

இயல்பான சூழலில் மனித கருத்துப் பரிமாற்ற ஒழுங்கமைப்பு மொழியைக் கருவியாக கொண்டு பேசுபவரிடமிருந்து கேட்பவருக்குச் செய்தியைப் பரிமாற்றம் செய்வதை உள்ளடக்கும். சில நேரங்களில் செய்தியின் பரிமாற்றம் பொருண்மை மயக்கத்திலிருந்து விடுபடாமல் இருக்கும். இப்பொருண்மை மயக்கம் மொழியில் மிகப் பரவலான நடப்பாகும். மயக்கமான சொல்லுடன் தொடர்புடைய அகப்பொருண்மையின் போதாமையால் அல்லது கருத்துப் பரிமாற்றத்தின் ஒரு குறிப்பிட்ட நிகழ்வில் கூற்றின் அமைப்பால் ஏற்படும். இவ்வாறு மயக்கங்கள் சொல் மயக்கம் (Lexical ambiguity) என்றும் குறிப்புரை மயக்கம் (Referential ambiguity) என்றும் தொடரியல் மயக்கம் (Syntactic ambiguity) என்றும் மூன்றாகப் பகுக்கப்படும்:

(1) They went to bank – சொல் மயக்கம்

(2) He loves his wife – குறிப்புரை மயக்கம்

(3) He saw a girl in the park with a binocular – தொடரியல் மயக்கம்

முதல் எடுத்துக்காட்டில் bank என்பது பணம் தொடர்புடைய 'வங்கி' என்பதையும் நீர்நிலைகளுடன் தொடர்புடைய 'கரை' என்பதையும் குறிப்பதால் இத்தகைய பொருண்மை மயக்கம் சொல் மயக்கம் எனப்படுகின்றது. இரண்டாவது எடுத்துக்காட்டில் his என்பதை எழுவாயாக வரும் He என்பதுடனும் வாக்கியதிற்கு வெளியிலுள்ள He என்பதுதன் தொடர்புபடுத்த இயலுவதால் இது குறிப்புரை மயக்கம் எனப்படும். He saw a girl in the park with a binocular என்பதை 'அவன் பூங்காவில் உள்ள ஒரு பெண்னை பைனாகுலரால் பார்த்தான்' எனவும் 'அவன் பூங்காவில் பைனாகுலர் வைத்திருந்த ஒரு பெண்ணைப் பார்த்தான்' எனவு பொருள்கொள்ள இயலும் ஆகையால் இது தொடரியல் சார்பான மயக்கம் எனப்படும்.

இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்பு பேசுபவரின் மன உருப்படுத்தத்தின் புலனுணர்வு அடிப்படையில் உருவாக்கப்படுவதால் அது பேசுபவரால் பயன்படுத்தப்படும் சொற்களுக்கும் வாக்கியங்களுக்கும் எல்லைப்படுத்தப்பட்டிருக்கும். இதை நேர்செய்ய மொழிபெயர்ப்பாளர்கள் மூல அகராதியை ஒரு குறிப்பிட்ட சூழலில் பொருத்தமாகத் தோன்றுகின்ற நிகரனை இலக்கு அகராதியால் பொருத்துகிறார்கள். சில நேர்வுகளில் இலக்கு மொழியானது மூலமொழியின் சொல்லை உருப்படுத்தம் செய்ய இயலும் நிகரான சொல் அலகைக் கொண்டிருக்காது. இந்நேர்வுகளில் மொழிபெயர்ப்பாளர்கள் சொற்றொடரைப் பயன்படுத்த முயல்வர் அல்லது நிகரான விளக்கத்தைத் தர முயல்வர். தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பை ஆதரிப்பவர்கள் சொல் மயக்கச் சிக்கலை நீக்க இயற்கைமொழி உரைகளில் கூடுதலாக நிகழும் மயக்க வடிவுகளில் பெரும் எண்ணிக்கையை ஆய்ந்தும் வெளிப்படையாக உருப்படுத்தம் செய்தும் உயர்ந்த தரமான மொழிபெயர்ப்பைத் தரலாம் என்று வாதியிடுகின்றனர். சாத்தியம் என்றால் மயக்கமான வடிவுகளும் மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பெறப்பட்ட விரிந்த அறிவின் அடிப்படையில் ஆயப்பட்டு இயந்திரத்தால் புரிந்து கொள்ளப்படும் வடிவுகளும் அகராதியில் சேமித்து வைக்கப்பட வேண்டும். ஆனால் பகுத்தறிவாதிகள் இவ்விதம் மொழியியல் தகவல்களைப் பெறும் செயல்பாடு சரியானதும் அல்ல சாத்தியமானதும் அல்ல என்று வாதியிடுகின்றனர் (Grishman and Kosaka 1992) களக்கட்டுப்பாடு மொழிபெயர்ப்பு (Domain–Specific Translation) மொழிபெயர்ப்புக்கு எப்போதும் ஆழமான பொருண்மையியல் ஆய்வு தேவையில்லை என்று வெளிப்படுத்துகின்றன. குறிப்பிட்ட மயக்கம் உள்ள சொல் அலகுகளின் ஆழமான பொருண்மையியல் ஆய்வுக்குப் பதிலாக நேரடியாக மூல மொழியிலிருந்து இலக்கு மொழிக்கு பதிலீடு செய்யும் அணுகுமுறையான எளிய ஆய்வைத் தெரிந்தெடுப்பது நல்லது. எடுத்துக்காட்டாக, ஆங்கிலத் head என்பது தமிழில் 'தலை' என மொழிபெயர்க்கப்பட்டுள்ளது; மூலமொழியில் head என்பது பல அர்த்தங்களில் பயன்படுத்தப்பட்டிருந்தாலும் தமிழில்'தலை' என மொழிபெயர்க்கப்பட்டுள்ளது சிக்கலை உருவாக்காது. பொருண்மை மயக்கத்துடன் தொடர்புடைய ஆழமான பொருண்மையியல் பகுப்பாய்விற்கு பதிலாக எளிமையான பகுப்பாய்வு மற்றும் அதிக நேரடி 'மூல மொழியை இலக்காகக் கொள்ள' மாற்று அணுகுமுறையைத் தேர்ந்தெடுப்பது நல்லது.

சில சூழல்களில் மயக்கங்கள் மொழி பெயர்ப்பின் போது அப்படியே இலக்குமொழிக்கு எடுத்துச் செல்லப்படவேண்டும் என்ற எதிர்பார்ப்பு அடிப்படையல் சில சொல் மயக்கங்களைப்

புறக்கணித்தல் சாத்தியமானதும் தேவையானதுமாகும். இருப்பினும் சொல் மயக்கங்களின் ஆய்வு, இலக்குமொழியில் மயக்கமற்ற உருப்படுத்தங்களை உருவாக்கும் என்ற நிலையில் பொதுவான மொழிபெயர்ப்பு நேர்வுகளில் மயக்கத்தைப் புறக்கணிக்கக் கூடாது (Isabelle and Bourbeau 1985: 21).

8.6.7 இலக்கணப் பொருத்தம்

மொழிபெயர்ப்பில் இலக்கண பொருத்தம் (Grammatical Mapping) என்பது மூல மொழியில் ஒரு சொல் இலக்கு மொழியில் உள்ள சொல்லுடன் அர்த்தம் உள்ள மொழிபெயர்ப்பைப் பெற வேண்டி பொருத்தப்படும். மொழிபெயர்ப்பில் பொருத்தத்திற்கு வேறுபட்ட திருத்தங்கள் இருக்கின்றன. அவை சொல் நிலைப் பொருத்தம், உருபன் நிலைப் பொருத்தம், இலக்கண நிலைப் பொருத்தம், தொடர்நிலைப் பொருத்தம், வாக்கியநிலைப் பொருத்தம் என்பனவாகும். மிகப் பொதுவானது மொழிபெயர்ப்புக்கு எடுத்துக்கொள்ளப்பட்ட இரு மொழிகளுக்கிடையில் வினை பங்களிப்பைப் பொருத்துவதாகும். இலக்கண பொருத்தம், வகைப்பாட்டியியல் அடிப்படையில் வேறுபட்ட இரண்டு மொழிகளுக்கிடையில் செய்யப்படும் இயந்திர மொழி பெயர்ப்பில் முக்கியத்துவம் உள்ளதாகும். தற்போதைய சூழலில் நாம் ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்ப்பைப் பற்றி பேசும் பொழுது இது முக்கியத்துவம் வாய்ந்தது. என்னவென்றால் ஆங்கிலத்திற்கு SVO அமைப்பு உண்டு (He is Rama). தமிழுக்கு ளுடீஏ அமைப்பு இருக்கிறது (அவன் சோறு சாப்பிட்டான்). எனவே இலக்கண பொருத்தமும் சொல் வரிசைமுறை மாற்றமும் இலக்கு மொழியில் வெளியீடுகளை வெளியிடுவதற்கு முக்கியத்துவம் வாய்ந்ததாகும்.

எனவே, இலக்குமொழியில் வெளியீடுகளை உருவாக்க ஒரு வகையான இலக்கண மேப்பிங் மற்றும் சொல் வரிசையை மறுசீரமைத்தல் அவசியம். எடுத்துக்காட்டாக, பின்வரும் வாக்கியத்தையும் வரைபட வரைபடத்தையும் கவனியுங்கள் (படம்).

ஆங்கிலம்: All his efforts ended in smoke.

தமிழ்: அவனுடைய எல்லா முயற்சிகளும் தோல்வியில் முடிந்தன.

படம்: மூலமொழி மற்றும் இலக்குமொழியின் வாக்கியங்களுக்கு இடையில் இலக்கண பொருத்தம்						
உள்ளீடு	All	his	Efforts	ended	in	failure
	(a)	(b)	(c)	(d)	(e)	(f)
நேரடி வெளியீடு	எல்லா (1)	அவன் (2)	முயற்சிகளும் (3)	முடிந்தது (4)	-இல் (5)	தோல்வி (6)
உண்மையான வெளியீடு	அவன் (2)	எல்லா (1)	முயற்சிகளும் (3)	தோல்வி (6)	இல் (5)	முடிந்தது (4)
	(2)	(1)	(3)	(7)		

இலக்குமொழியில் ஏற்றுக்கொள்ளக்கூடிய சொல் வரிசையுடன் இறுதி வெளியீட்டை அடைய, மூலமொழியின் வாக்கியத்தில் பயன்படுத்தப்படும் சொற்கள் இலக்குமொழியில் உள்ள சொற்களுடன் பின்வரும் முறையில் குறிக்கப்படுகின்றன:

சொல்சார் பொருத்தல்:

ஆங்கிலம் [a] = தமிழ் [1] (சொலுக்குச் சொல் பொருத்துதல்)

ஆங்கிலம் [b] = தமிழ் [2] (சொல்லுக்குச் சொல் பொருத்துதல்)

ஆங்கிலம் [c] = தமிழ் [3] (சொல்லுக்குச் சொல் சொல் பொருத்துதல்)

ஆங்கிலம் [d] = தமிழ் [4] (சொல்லுக்குச் சொல் சொல் பொருத்துதல்)

ஆங்கிலம் [e] = தமிழ் [5] (முன்னொட்டுக்கு வேற்றுமை உருபின் பயன்பாடு)

ஆங்கிலம் [f] = தமிழ் [6] (சொலுக்குச் சொல் பொருத்துதல்)

இருப்பினும், சரியான மொழிபெயர்ப்பைப் பெறுவதற்கு வெறும் சொல்சார் பொருத்தம் மட்டும் போதாது என்பது குறிப்பிடத்தக்கது. எடுத்துக்காட்டாக இறுதியில் வினைத்தொடர் ended in failure என்பதற்கு ended in smoke என்றிருந்தால் பயன்வழியியல் தகவல் தேவைப்படும். ended in smoke என்பதற்குத் தமிழில் “பயனற்றுப்போது” என்ற மொழிபெயர்ப்பு நிகரன் தேவைப்படும்.

பயன்வழியியல்தகவல்

ஆங்கிலம்: [d-e-f] (ஒரு மரபுத்தொடர் வெளியீடு)

தமிழ்: [பயனற்றுப்போனது (<4-5-6) (ஒத்த மொழிபெயர்ப்பு நிகரன்)

சிறந்த மொழிபெயர்ப்பு வெளியீடுகளைப் பெறுவதற்கு மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் பகுப்பாய்வு சமமான கட்டுமானங்களுக்கு இடையில் பொருத்தத்தை

மேம்படுத்த உதவுகிறது என்பதை ஒப்புக் கொள்ளப்படுகிறது. செயல்முறை என்பது சமமான கட்டுமானங்களின் (எ.கா. பல்சொல் அலகுகள், மரபுச்சொற்கள், சொற்றொடர்கள், எச்சத்தொடர்கள், பெரிய தொடரியல் கட்டமைப்புகள் போன்றவை) கூட்டு பகுப்பாய்வு நேரத்தில் சில வழக்கமான முறையான கட்டமைப்புகளுடன் தொடர்புடையது.

அத்தகைய நடைமுறையின் அடிப்படை நோக்கம், இணைத்தல் பொறிமுறையை முறையான முறையில் மூன்று பகுதிகளாக பிரிக்க அனுமதிப்பது:

- இரண்டு தரவுத்தொகுதிகளில் இணைக்கக்கூடிய அலகுகளை அடையாளம் காணுதல்.
- உருபனியல்-தொடரியல் அடையாளங்களைப் பயன்படுத்துவதன் மூலம் இணைக்கக்கூடிய அலகுகளின் கட்டமைப்புகளை முறைப்படுத்துதல்.
- மொழிபெயர்ப்புத் தரவுத்தொகுதிகளின் பயனுள்ள தரவுகளுடன் ஒப்பிடுவதன் மூலம் முன்மொழியப்பட்ட கட்டமைப்புகளின் நிகழ்தகவைத் தீர்மானித்தல்.

செயல்முறையை மூன்று கட்டங்களாகப் பிரிப்பதன் மூலம், தரவுத்தொகுதியில் காணப்பட்ட உண்மையான மொழிபெயர்ப்புகளின் தத்துவார்த்த பகுப்பாய்வோடு தொடர்புபடுத்தக்கூடிய அலகுகளைத் தீர்மானிக்க ஒப்பீட்டளவில் எளிய மொழிபெயர்ப்பு தொகுதிகள் தயாரிக்கப்படுகின்றன.

எளிதாக்குவதற்கான சாத்தியமான வழிகளில் ஒன்று, 'பயிற்சி தரவுத்தொகுதியில்' சேமிக்கப்பட்டுள்ள தரவுகளின் அடிப்படையில் பகுப்பாய்வு முறைகளை உருவாக்குவது. இத்தகைய முறைகள், மாதிரி பயிற்சியின் அடிப்படையில், பெரும்பாலும் ஒரு முன்னறிவுக்கு கிடைக்கக்கூடிய மொழியியல் தகவல்களின் அளவைப் பொறுத்தது; அதாவது மனித வல்லுநர்களால் முன்னர் உருவாக்கப்பட்ட தொடரியல் விதிகளைப் பொறுத்தது.

மொழிகளின் அனைத்து விதிகளையும் கண்டுபிடிக்க மொழிபெயர்ப்பு தரவுத்தொகுதியில் பயன்படுத்தப்படும் அனைத்து வாக்கியங்களையும் பகுப்பாய்வு செய்ய வேண்டிய அவசியமில்லை என்பதை வலியுறுத்த வேண்டும். முழு டோக்கன்களைக் காட்டிலும், வகைகளின் தொகுப்புகளின் பகுப்பாய்வு, ஆரம்ப நோக்கத்திற்கு உதவும் மற்றும் போதுமானதாக இருக்கும், ஏனெனில்:

- ஒவ்வொரு மொழியிலும், மொழியியல் அலகுகள் உள்ளன, அவை வடிவத்தில் ஒரே மாதிரியாகவும் மற்றவர்களுக்கு கூறுகளாகவும் உள்ளன. அதாவது, ஒரு பெயர்த்தொடர் ஒரு தரவுத்தொகுதிக்குள் வேறு சில பெயர்த்தொடர்களுடன் கட்டமைப்பு ரீதியாக ஒத்திருக்கலாம். மூலமொழி மற்றும் இலக்குமொழி இரண்டிற்கும் இது உண்மை.

- நெருங்கிய தொடர்புடைய இரண்டு சகோதரி மொழிகளிலிருந்து மொழிபெயர்ப்பு தரவுத்தொகுதிகள் உருவாக்கப்பட்டால், மூலமொழி உரையில் உள்ள அலகுகளுக்கும் இலக்குமொழி உரையில் உள்ள அலகுகளுக்கும் இடையிலான வரிசைமுறை மற்றும் தொடர்புகள் ஒன்றுபோல் இருக்கும்.
- சில நிலையான நோக்கீட்டுப் புள்ளிகள் உள்ளன; அவை இரண்டு உரைகளைக் குறிக்கின்றன மற்றும் மொழிபெயர்ப்பு அலகுகளை விரைவாக அடையாளம் காண அனுமதிக்கின்றன. இவை இரண்டு உரை மாதிரிகளின் எண்கள், தேதிகள், சரியான பெயர்ச்சொற்கள், தலைப்புகள் மற்றும் உரை தளவமைப்பு (எ.கா. பத்திகள், பிரிவுகள் போன்றவை) இருக்கும்.

மொழிபெயர்ப்பு தரவுத்தொகுதியிலிருந்து பெறப்பட்ட சமநிலைகளின் பகுப்பாய்வின் அடிப்படையில், பின்வரும் மூன்று வகையான இலக்கண பொருத்தங்கள் காணப்படுகின்றன.

- சொற்களின் எண்ணிக்கை, அவற்றின் வரிசை மற்றும் அவற்றின் பொருண்மை ஒரே மாதிரியாக இருக்கும் வலுவான பொருத்தத்தின் எடுத்துக்காட்டுகள்.
- சொற்களின் எண்ணிக்கையும் அவற்றின் அர்த்தங்களும் ஒரே மாதிரியாக இருக்கும் தோராயமான பொருத்தத்தின் எடுத்துக்காட்டுகள், ஆனால் அவை தோன்றும் வரிசையில் அல்ல.
- ஒழுங்கு மற்றும் சொற்களின் எண்ணிக்கை வேறுபட்டிருக்கும் பலவீனமான பொருத்தத்தின் எடுத்துக்காட்டுகள், ஆனால் அவற்றின் அகராதி அர்த்தங்கள் ஒன்றே

ஆங்கிலத்திலிருந்து தமிழ் இயந்திரமொழிபெயர்ப்பு விஷயத்தில், இரண்டு இலக்கண வரைபடங்கள் பலவீனமான பொருத்தமாக இருக்கும், ஏனெனில் அவை இரண்டு வெவ்வேறு வகைப்பாட்டியல் வகுப்புகளைச் (typological classes) சேர்ந்தவை (ஆங்கிலம் எஸ்.வி.ஓ. வகை/SVO type, தமிழ் எஸ்.ஓ.வி. வகை/ SOV type).

இத்தகைய சூழ்நிலையில், இந்த மொழிகளின் தரவுதொகுதியின் வரிசைபடுத்தல் அந்தந்த உரைகளின் தொடரியல் கட்டமைப்பை மட்டுமே நம்பியிருக்கக்கூடாது; ஆனால் பொருண்மைசார் நங்கூரப் புள்ளிகளுக்கான ஏற்பாட்டைக் கொண்டிருக்க வேண்டும்.

மூலமொழியின் உரையின் ஒரு வாக்கியத்தில் ஐம்பது சதவிகித சொற்கள் மூல மொழி உரையின் ஒரு வாக்கியத்தில் குறைந்தது ஐம்பது சதவிகித சொற்களுடன் பொருண்மை

அடிப்படையில் பொருந்தினால், இரண்டு வாக்கியங்களுக்கு சமநிலை அல்லது மொழிபெயர்ப்பு உறவு இருப்பதாக மட்டுமே நாம் கூறலாம்.

உரைகள் மீது இத்தகைய அமைப்பின் நம்பகத்தன்மை பத்தி அல்லது வாக்கிய மட்டத்தில் இடைநிலை வரிசைப்படுத்தல் மூலம் உத்தரவாதம் அளிக்கப்படுகிறது. ஆகவே, மூலமொழியின் அ-என்ற வாக்கியத்தில் தோன்றும் ஐந்து சொற்கள் இலக்குமொழியின் அ-என்ற வாக்கியத்தில் சமமான அர்த்தங்களின் ஐந்து சொற்களைக் கொண்டிருந்தால், இரண்டு வாக்கிய அலகுகள் மொழிபெயர்ப்பு உறவைக் கொண்டிருப்பதாகக் கருதப்படும்.

நம்பகத்தன்மையை உறுதிப்படுத்துவதற்காக, மிகப் பெரிய மொழிபெயர்ப்பு அலகுகளிலிருந்து (எ.கா. இயல்கள், பத்திகள், முதலியன) சிறியவைகளுக்கு (எ.கா. வாக்கியங்கள், சொற்றொடர்கள், சொற்கள் போன்றவை) மொழிபெயர்ப்புத் தரவுத்தொகுதில் பயன்படுத்தப்படுகிறது.

தரவுத்தொகுதிகளில் 'பிற்போக்கு வரிசைப்படுத்தல் நுட்பத்தை' பயன்படுத்துவது (சிறியது முதல் மிகச் சிறிய அலகுகள் வரை கவனம் செலுத்துதல்) பின்வரும் கருதுகோள்களை சரிபார்க்க புலனாய்வாளர்களுக்கு உதவும்.

- இரண்டு இயல்களின் குறைந்தது ஐந்து பத்திகள் ஒன்றுக்கொன்று ஒத்திருந்தால் அவை மொழிபெயர்ப்பு உறவைக் கொண்டுள்ளன எனலாம்.
- இரண்டு பத்திகளின் குறைந்தது ஐந்து வாக்கியங்கள் ஒன்றுக்கொன்று ஒத்திருந்தால் அவை மொழிபெயர்ப்பு உறவைக் கொண்டுள்ளன எனலாம்.
- இரண்டு வாக்கியங்களின் குறைந்தது ஐந்து சொற்கள் ஒன்றுக்கொன்று ஒத்திருந்தால் அவை மொழிபெயர்ப்பு உறவைக் கொண்டுள்ளன எனலாம்.
- இரண்டு சொற்களின் அர்த்தங்களில் குறைந்தது ஒன்று (இருமொழி அகராதியில் பாதுகாக்கப்படுகிறது) ஒன்றுக்கொன்று ஒத்திருந்தால் அவைகளுக்குள் மொழிபெயர்ப்பு உறவு இருக்கும் எனலாம்.

முறைப்படி, தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அணுகுமுறையுடன் புள்ளியியல்சார் இயந்திர மொழிபெயர்ப்பு அணுகுமுறையை இணைப்பது எதிர்காலச் செயலாக்கம் (future processing) மற்றும் தகவல்களை மீட்டெடுப்பதை (information retrieval) மேம்படுத்துவதற்காக மொழிபெயர்ப்புத் தரவுத்தொகுதிகளின் வரிசைப்படுத்தல் செயல்முறையை சிறப்பாகச் செய்ய உதவுகிறது.

இருப்பினும், இதற்கு 'மொழிபெயர்ப்பு அலகுகள்' அடையாளம் காணப்படுவதும் முறைப்படுத்தப்படுவதும் மற்றும் இருமொழி அகராதிகள் மற்றும் இயந்திரம் படிக்க இயலும் அகராதிகளைப் பயன்படுத்துவதும் தேவை. ஆகையால், ஒவ்வொரு தரவுத்தொகுதி உரையின் முழுமையான உருபனியல்-தொடரியல் அடையாளப்படுத்தல் தேவையில்லை; ஏனென்றால் மொழிபெயர்ப்பு உறவுகளை வெளிப்படுத்திய தரவுத்தொகுதிகளை ஒப்பிடுவதன் மூலம் சமநிலைகளைக் கண்டறிய இயந்திரம் புள்ளியியல் ஆதரவுடன் அதைச் செய்ய இயலும். இருப்பினும், அமைப்பின் தரமான செயல்திறனை உறுதிப்படுத்த, பின்வரும் அம்சங்களை கவனித்துக்கொள்ள வேண்டும்.

- மொழிபெயர்ப்புத் தரவுத்தொகுதிகளின் தரம் அதிகமாக இருக்க வேண்டும். தரவுத்தொகுதிகளின் தரம் மோசமாக இருந்தால் அல்லது மொழியியல் நிபுணர்களால் அவை கடுமையான கட்டுப்பாட்டுக்கு உட்படுத்தப்படாவிட்டால் தரவு (சீரமைக்கப்பட்ட இருமொழி நூல்கள்) சிக்கலை ஏற்படுத்தக்கூடும்.
- இருமொழி அகராதியின் தரம் மற்றும் அளவு மேம்படுத்தப்பட வேண்டும். இலக்கண தகவல்களை வழங்குவதில் அகராதி அடிப்படை ஆதாரமாகும். இருமொழி தரவுத்தொகுதிகளில் காணப்படாத அறியப்படாத சொற்களை ஒருங்கிணைக்க இது ஏற்பாடு கொண்டிருக்க வேண்டும்.
- அமைப்பின் வலுவான தன்மை மற்றும் மொழிபெயர்ப்பின் தரம் ஆகியவை கிடைக்கக்கூடிய பயிற்சி தரவின் அளவைப் பொறுத்தது.
- வெளியீடுகளில் துல்லியத்தின் நிலை மொழிபெயர்ப்பு தரவுத்தொகுதிகளுக்கு இடையிலான ஒத்திசைவு நிலைகளை பெரிதும் நம்பியிருக்கும்.

மேலே உள்ள ஆதாரங்களுடன், ஒரு தரவுத்தொகுதிசார் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறை வலுவானதாக இருக்கும் என்று எதிர்பார்க்கப்படுகிறது; இருப்பினும் பெரிய அளவிலான வெவ்வேறு உரைத் தரவுகள் குறித்த நீண்ட பயிற்சி கட்டம் இன்றியமையாத கட்டாயமாகும். இந்த நிலை முடிந்ததும், மொழிபெயர்ப்பு நினைவகத்தில் சேமிக்கப்பட்ட தகவல்கள் மீண்டும் செயல்படுத்தப்பட்டு, முன்னர் மனித நிபுணர்களின் பிரத்தியேக களமாக இருந்த அனைத்து வகையான மொழிபெயர்ப்பு தீர்வுகளையும் அளிக்கும். முழு அளவிலான வெற்றியை அடைய, செயற்கை நுண்ணறிவின் அளவை கணினியில் ஒருங்கிணைக்க வேண்டியிருக்கும்.

மேலே வழங்கப்பட்ட அடித்தளம் ஒரு பயனுள்ள தரவுத்தொகுதிசார் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதற்கான ஒரு முன்மொழிவாகும்; இது அனைத்து உரை வகைகளையும் ஒரு மொழியிலிருந்து மற்றொரு மொழியில் மொழிபெயர்க்க அனுமதிக்கிறது.

இந்திய மொழிகளின் தரவுத்தொகுதிகள் சில கவனத்தை ஈர்க்கும் சில நடைமுறை சிக்கல்களை முன்வைத்தாலும், பயிற்சி தரவுத்தொகுதிகளின் பயன்பாடு தரவுத்தொகுதிசார் இயந்திரமொழிபெயர்ப்பு ஆராய்ச்சியை மேம்படுத்தக்கூடும் என்று நாம் நம்பலாம். இயந்திரமொழிபெயர்ப்பு ஒரு பயன்பாட்டுத் துறையாகும், மேலும் முன்னேற்றத்திற்கான உந்துதல் வெளிமாநில மூலங்களிலிருந்து பெருமளவில் வருவது மிகவும் பொருத்தமானது.

இது ஒரு சிறப்பு களமாகும்; இது தொடர்புடைய பல்வேறு கோட்பாடுகள் மற்றும் பயன்பாடுகளுக்கான தகுதியான சோதனை படுக்கையாகும். மொழியியலில், தொடரியல், பொருண்மையியல், பயன்வழியியல் மற்றும் கருத்தாடலின் பல்வேறு கோட்பாடுகள் அதற்கு இணக்கமாக இருக்கிறதா என்பதை இது சரிபார்க்கிறது. கணக்கீட்டு முறைகளைப் பொறுத்தவரை, செயலாக்கம், பாகுபடுத்தல், பொருண்மையியல்சார் பகுப்பாய்வு, பயன்வழியியல் விளக்கம் மற்றும் உரை உருவாக்கம் ஆகியவற்றுக்கான பல்வேறு வழிமுறைகள் இதற்குப் பொருந்தினால் அது சவால் விடுகிறது. விளக்க மொழியியல் விஷயத்தில், ஒரு குறிப்பிட்ட மொழிக்கான அகராதி மற்றும் இலக்கணங்கள் இரண்டும் செயல்பாட்டில்/பணியில் பயனுள்ளதாக பயன்படுத்தப்படுமா என்பதை இது சரிபார்க்கிறது. இறுதியாக, மனித பகுத்தறிவு வளங்களை மாதிரியாக்குவதில், அறிவு பிரதிநிதித்துவம் மற்றும் கையாளுதல் ஆகியவை மொழிபெயர்ப்பு வழிமுறைகளில் எந்தவிதமான பொருத்தத்தையும் கொண்டிருக்க முடியுமா என்பதை சோதிப்பதை நோக்கமாகக் கொண்டுள்ளது.

இயந்திர மொழிபெயர்ப்பு, மொழியின் பல்வேறு கணக்கீட்டு கோட்பாடுகளின் விரிவான மதிப்பீட்டிற்கும் இயற்கையான மொழியில் ஏராளமாக நிலவும் குறிப்பிட்ட மொழியியல் நிகழ்வுகளின் பரவலான வளர்ச்சி மற்றும் சோதனைக்கும் ஒரு சிறந்த மற்றும் முக்கியமான சூழலாக மாறுகிறது,

இந்திய மொழிகளின் கண்ணோட்டத்தில் மொழிபெயர்ப்புத் தரவுத்தொகுதிகள் பயனுள்ள தரவுத்தொகுதிசார் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்க பயனுள்ளதாகப் பயன்படுத்தப்படும் முதன்மை முன்நிபந்தனைகள் ஆகும்.

மொழிபெயர்ப்பு தரவுத்தொகுதிகள் இரு வழிப் பங்கைச் செய்ய பயன்படுத்தப்படும்: (அ) இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்குவதற்கான உள்ளீடுகளாகவும், (ஆ) ஏற்கனவே உருவாக்கிய அமைப்புகளை மதிப்பீடு செய்வதற்கான சோதனை படுக்கைகளாகவும். எனவே, மொழிபெயர்ப்பு தரவுத்தொகுதிகள் கிடைப்பது அமைப்புகளின் உண்மையான திறனுக்கு குறிப்பிடத்தக்க பங்களிப்பை வழங்கும். தரவுத்தொகுதி அடிப்படையிலான முறையின் வெற்றி தற்போதைய ஆய்வாளர்களை மரபு விதி அடிப்படையிலான அணுகுமுறைகளுக்குத் துணைபுரிய இந்த அணுகுமுறையைப் பின்பற்றுமாறு வழிநடத்துகிறது; ஏனெனில் மொழிபெயர்ப்புத் தரவுத்தொகுதிகளின் பகுப்பாய்விலிருந்து பெறப்பட்ட தகவல்கள் மூலமொழி மற்றும் இலக்குமொழிக்கு இடையிலான தூரத்தைக் குறைக்க முனைகின்றன.

கட்டுப்படுத்தப்பட்ட மொழி தரவுத்தளங்களைக் கொண்ட அனைத்து டொமைன்-குறிப்பிட்ட இயந்திர மொழிபெயர்ப்பு அமைப்புகளுக்குள்ளும் இது ஒரு நேர்மறையான விளைவாகும்; இதில் அனைத்து வகையான தொடரியல், சொல்சார் மற்றும் தொடர்சார் பொருண்மை மயக்கம் முன்பே அடக்கப்படுகின்றன (Teubert 2000). இத்தகைய அமைப்பு வழக்கமாக மொழிபெயர்ப்பு தரவுத்தொகுதிகளாகப் பயன்படுத்தப்படும் இரண்டு மொழிகளுக்கு இடையில் மொழிபெயர்ப்பை மேம்படுத்த பரஸ்பர நுண்ணறிவின் இடைவெளியைக் குறைக்கிறது.

தற்போதைய விளக்கத்தின் சாராம்சம் என்னவென்றால், இந்திய மொழிகளுக்கான திறமையான இயந்திர மொழிபெயர்ப்பு முறையை உருவாக்க நாம் உண்மையிலேயே உறுதியாக இருந்தால், மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் மதிப்பை நாம் புறக்கணிக்க இயலாது. தமிழுக்கும் இணைத் தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. எடுத்துக்காட்டாக இராசேந்திரன் மற்றும் வாசுகியின் (2019) English to Tamil Machine Translation System Using Parallel Corpus/இணைத் தரவுத்தொகுதியைப் பயன்படுத்தி ஆங்கிலம்-தமிழ் இயந்திர மொழிபெயர்ப்பு என்ற ஆய்வைக் குறிப்பிடலாம்.

8.7. சுருக்கவுரை

இயந்திர மொழிபெயர்ப்பில் தரவுத்தொகுதியின் பயன்பாடு பற்றி விளக்குவது இவ்வியலின் நோக்கமாகும். இவ்வியலின் தொடக்கத்தில் இயந்திர மொழிபெயர்ப்பில் தரவுத்தொகுதியின் பயன்பாடு பற்றி ஒரு சுருக்கமான அறிமுகம் தரப்பட்டுள்ளது. அதைத் தொடர்ந்து குறிக்கோள், வரலாற்றிலிருந்து கிடைத்த பாடங்கள், தரவுத்தொகுதி

அடிப்படையிலான அணுகுமுறை, தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை தொடர்பான சிக்கல்கள், மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் உருவாக்கம் மற்றும் மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் மீது மொழியியல் செயல்பாடுகள் என்பன பற்றி விளக்கங்கள் தரப்பட்டுள்ளன. இவ்வாறு தரவுத்தொகுதி இயந்திர மொழிபெயர்ப்புக்குப் பெரிதும் பயனுள்ளதாக அமைகிறது. இணையான மொழிபெயர்க்கப்பட்ட தரவுத்தொகுதி சொல் பொருத்தம், உருபன் பொருத்தம், இலக்கணப் பொருத்தம், தொடர் பொருத்தம், வாக்கியப் பொருத்தம் என்ற செயல்பாடுகளிலிருந்து மொழியின் சொல் மட்டம், உருபன் மட்டம், தொடரியல் மட்டம் என்ற நிலைகளில் நிகரன்களைக் கண்டறிந்து கூடுதலாக தரவுத்தொகுதிக்குப் பயிற்சி தந்து இயந்திர மொழிபெயர்ப்பை மேம்படுத்தலாம்.

இயல் 9

தரவுத்தொகுதியின் சாத்தியமான பயன்பாடுகள்

9.1 அறிமுகம்

தரவுத்தொகுதி பின்வரும் வல்லுநர்களால் அவர்களின் செயல்பாடுகளுக்குத் துணையாக அதிகமாகப் பயன்படுத்தபடுகின்றது.

1: மொழி வல்லுநர்கள் (நூல்களின் மொழியில் ஆர்வம்)

2: உள்ளடக்க வல்லுநர்கள் (நூல்களின் உள்ளடக்கத்தில் ஆர்வம்), மற்றும்

3: ஊடக வல்லுநர்கள் (மின்னணு சாதனங்களுக்கான சோதனை படுக்கையாக தரவுத்தொகுதியில் ஆர்வம்).

9.2.மொழி வல்லுநர்களின் மத்தியில்

அகராதியலார்கள் (Lexicographers) சொற்கள், சொல்லன்கள், சொற்றொடர்கள், மரபுச்சொற்கள் போன்றவற்றின் உண்மையான மற்றும் குறிப்பிட்ட பயன்பாடு குறித்த தகவல்களுக்குத் தரவுத்தொகுதியைக் கலந்தாலோசிக்கிறார்கள். அவர்கள் சொல்சார் தரவுத்தளங்கள், அகராதிகள், சொற்களஞ்சியங்கள் மற்றும் நோக்கீட்டுப் பொருள்களை உருவாக்கத் தரவுத்தொகுதியை ஆய்கின்றனர்.

கலைச்சொல் வல்லுநர்கள் (Terminologists) மற்றும் தொழில்நுட்ப எழுத்தாளர்கள் (technical writers) தரவுத்தொகுதித் தரவுத்தளத்தை தொழில்நுட்பக் கலைச்சொற்களைத் தரப்படுத்தவும் கலைச்சொல் தரவுத்தளங்களை அதிகரிக்கவும் பயன்படுத்துகின்றனர்.

கோட்பாட்டாளர்கள் (Theoreticians) தரவுத்தொகுதி மொழியின் உண்மைகளின் வெகுஜன பிரதிநிதித்துவத்தின் ஒரு பெரிய அமைப்பாகப் பயன்படுத்துகின்றனர். அவர்களுக்குத் தரவுத்தொகுதி அனைத்து வகையான நிகழ்வுகளின் ஒப்பீட்டு அதிர்வெண் பற்றிய தரவை அளிக்கிறது; மேலும் அவற்றின் சொந்த அல்லது அவற்றின் தகவலறிந்தவர்களின் சான்றுகளைச் சரிபார்க்க வாய்ப்பளிக்கிறது.

பயன்பாட்டு மொழியியலாளர்கள் (Applied linguists) மொழி கற்பிப்பதில் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர்; ஏனெனில் இது சான்றளிக்கப்பட்ட பயன்பாட்டின் அதிகாரத்துடன் மொழியைப் பிரித்தெடுப்பதற்கும் படிப்பதற்கும் ஆதாரங்களை வழங்குகிறது.

9.2. பொருளடக்க வல்லுநர்களின் மத்தியில்

வரலாற்றாசிரியர்கள் (Historians) சொற்கள், சொற்றொடர்கள் மற்றும் அவற்றைக் குறிக்கும் வாக்கியங்களைப் படிப்பதன் மூலம் கருத்துகள் மற்றும் கருத்துக்களின் வளர்ச்சியைக் கண்காணிக்க முடியும். மறைமுகமான நேரம் மற்றும் இடம் குறித்த அடையாளங்களைக் கண்டறிய அவர்கள் தேதியிட்ட மற்றும் பகுப்பாய்வு செய்யப்பட்ட தரவுத்தொகுதியை தோற்றம்/மூலம் தெளிவற்றதாக இருக்கிற ஆவணங்களை அடையாளம் காணப் பயன்படுத்துகின்றனர் பயன்படுத்துகின்றனர.

சொல்-பயன்பாட்டின் புள்ளிவிவர பகுப்பாய்வு, அறியப்பட்ட எழுத்தாளர்களுக்குச் சந்தேகத்திற்குரிய படைப்புகளை நிர்ணயிப்பதில் முக்கிய பங்கு வகிப்பதால் விமர்சகர்கள் (Literary critics) ஸ்டைலோமெட்ரிக்ஸில் (stylometrics) ஆராய்ச்சிக்குத் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். மொழியியலாளர்கள் தனிப்பட்ட சொற்களைக் காட்டிலும் உயர்ந்த மட்டத்தில் குறிப்பிடத்தக்க அம்சங்களைக் கண்டறியும்போது இத்தகைய நுட்பங்கள் மிகவும் பயனுள்ளதாக இருக்கும். நுட்பங்களுக்கான வெகுஜன பயிற்சி களமாகச் செயல்படுவதைத் தவிர, வயது, பாலினம், காலம், பிறந்த நாடு போன்றவற்றால் அடையாளம் காணப்பட்ட வெவ்வேறு குழுக்களை வகைப்படுத்தும் பாணியின்/நடையின் வேறுபாடுகள் குறித்த புள்ளிவிவர தகவல்களுக்கான ஆதாரமாக இது பயன்படுத்தப்படுகிறது.

சமூகவியலாளர்கள் வெவ்வேறு வகை, இனம், மதம், இனம் போன்றவற்றைச் சேர்ந்த வெவ்வேறு குழுக்களை வகைப்படுத்தத் தரவுத்தொகுதியை ஒத்த பாணியில் பயன்படுத்துகின்றனர்.

9.3. தகவல்தொடர்பு வல்லுநர்களின் மத்தியில்

தகவல் மீட்டெடுப்பவர்கள் (Information Retrievers) அறிவுத் தளத்தை உருவாக்குவதற்கு உரைகளிலிருந்து பொருத்தமான தகவல்களைப் பிரித்தெடுக்க இயங்குமுறையை உருவாக்கவும், சொல்லடைவுக்கான பொருட்களின் தகவல்களைக் கண்டுபிடிப்பதற்கும், உரைகளின் முக்கியமான உள்ளடக்கத்தைச் சுருக்கமாகக் கூறுவதற்கும் உரை அமைப்புகளிலிருந்து பொருத்தமான தகவல்களைப் பெறுவதற்கான வழிமுறைகளை வகுக்கவும் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர்.

கணினி மொழியியலாளர்கள் (Computational linguists) தரவுத்தொகுதியைத் தங்கள் படைப்புகளைத் தரவுத்தொகுதியில் காணப்படும் புள்ளியியல் ஒழுங்குமுறைகளுடன் ஒருங்கிணைக்கப் பயன்படுத்துகின்றனர்; இது மொழியைப் பகுப்பாய்வு செய்வதற்கும்

செயலாக்குவதற்கும் ஒரு முக்கிய திறவுகோலாகச் செயல்படுகிறது. மேலும் தரவுத்தொகுதி, தரவு மற்றும் அறிவுத் தளத்தின் ஆதாரமாக மொழியில் சீரான தன்மையின் இருப்பு அல்லது இல்லாமையைப் பரிசோதிக்கப் பயன்படுகிறது; ஏனெனில் சில இலக்கண ரீதியாகப் பகுப்பாய்வு செய்யப்பட்ட தரவுத்தொகுதிகளின் வெளியீடுகளில் பணிபுரியும் போது புள்ளிவிவர நுட்பங்கள் மிகவும் பயனுள்ளதாக இருக்கும்.

இயந்திர மொழிபெயர்ப்பாளர்கள் (Machine Translators) தரவுத்தொகுதியை அணுகுவது தேவையான மற்றும் பொருத்தமான மொழியியல் தகவல்களைப் பெறுவதற்கும் அவற்றின் அமைப்புகளின் செயல்திறனைச் சரிபார்ப்பதற்கும் ஆகும்; ஏனெனில் தரவுத்தொகுதி, ஒழுங்குமுறைகளின் உண்மையான திறனை மேம்படுத்த குறிப்பிடத்தக்கப் பங்களிப்பை செய்கிறது. மேலும், டொமைன் குறிப்பிட்ட தரவுத்தொகுதி மரபு அறிவு அடிப்படையிலான அணுகுமுறைகளுக்கு துணைபுரிய சுய-ஒழுங்குமுறை அணுகுமுறையைப் (self-organising approach) பின்பற்ற அமைப்புகளுக்கு உதவுகிறது.

மொழிச் செயலாக்கம் மக்கள் (Language processing People) பல்வேறு வகையான தரவுத்தொகுதியின் உருவாக்கத்திலிருந்து மேலும் மேலும் பயனடைகிறார்கள்; ஏனெனில் மூல மற்றும் அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகள் மொழி செயலிகளை (language processors) உருவாக்குவதற்குப் பெரிய அளவில் பயன்படுத்தப்படுகின்றன. தரவுத்தொகுதி என்பது ஆராய்ச்சியாளர்கள், தொழில்நுட்ப வல்லுநர்கள், எழுத்தாளர்கள், அகராதியலார்கள், கல்வியாளர்கள், ஆசிரியர்கள், மாணவர்கள், மொழி கற்பவர்கள், அறிஞர்கள், வெளியீட்டாளர்கள் மற்றும் பலர் உட்பட அனைவருக்கும் நன்மை பயக்கும் வளமாகும் என்று சொல்வது போதுமானது.

9.4. சுருக்கவுரை

இந்த இயல் தரவுத்தொகுதியின் சாத்தியமான பயன்பாடு பற்றிய விவரங்களை அளிக்கிறது. தரவுதொகுதியின் பயன்பாடுகளை விரல்விட்டு எண்ணி விட இயலாது. மொழித் தொழில்நுட்ப வளர்ச்சிக்குத் தரவுத்தொகுதிகள் பெரும் பங்காற்றுகின்றன. மொழி வல்லுநர்கள், நூல் உள்ளடக்க வல்லுநர்கள், ஊடக வல்லுநர் முதலானோருக்குத் தரவுத்தொகுதிகள் பெரும் வரப்பிரசாதமாக அமைகின்றன. அகராதிகள், பேச்சு அறிவான்கள், பேச்சு உருவாக்கிகள், இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் போன்ற மொழிக் கருவிகள் உருவாக்கத்திலும் மொழி கற்றலிலிலும் மொழி கற்பித்தலிலும் தரவுத்தொகுதிகள் பெரும் பங்கு வகிக்கின்றன.

இயந்திரம் கற்றலுக்குக் தரவுதொகுதியின் பங்கு போற்றுதற்குரியது. தரவுத்தொகுதிகள் அவற்றில் பொதிந்துள்ள மொழி அறிவைக் கணினிக்கு எளிதில் கற்றுத்தந்து கணினியை ஒரு மனிதனுக்கு இணையான மொழி அறிவுடன் செயலாற்றத் துணை புரிகின்றன.

இயல் 10

முடிவுரை

முந்தைய பகுதிகளில் தரவுத்தொகுதியின் விளக்கம், தரவுத்தொகுதியின் தொடக்க நிலை, விரிதரவின் இன்றைய நிலை, இதுவரை உருவாக்கப்பட்ட தரவுத்தொகுதிகள், தரவுத்தொகுதியின் வகைப்பாடு, எழுத்துவடிவ விரிதரவின் உருவாக்கம், உரை தரவுத்தொகுதியின் ஆய்வு, மொழித் தொழில் நுட்;பத்தில் விரிதரவு, முக்கிய மொழியியலில் தரவுத்தொகுதி, இயந்திர மொழிபெயர்ப்பில் தரவுத்தொகுதி, தரவுத்தொகுதியின் பயன்பாடுகள் என்ற தலைப்புகளில் தரவுத்தொகுதி தொடர்பான செய்திகள் எல்லாம் விளக்கமாக எடுத்துரைக்கப்பட்டுள்ளன. களமொழியலிருந்து தரவுத்தொகுதி மொழியலுக்கான பயணம் பல மயில்கற்களைத் தாண்டி வந்துள்ளது. அன்றைய கால கட்டத்தில் மொழியியலார் கள ஆய்வை மேற்கொண்டு மொழித் தரவுகளைத் திரட்டி ஒரு மொழியின் அமைப்பை அறிய ஒலியியல், ஒலியனியல், உருபனியல், தொடரியல் மற்றும் பொருண்மையியல் மட்டங்களில் ஆய்வு செய்வார்கள். இத்தகைய ஆய்வு தொடக்க காலத்தில் அமைப்புமுறை மொழியியல் ஆய்வாக அமைந்தது. இது நடத்தையிலாரால் ஊக்கப்படுத்தப்பட்ட ஊன்றிக்கவனித்தல் அடிப்படையிலான அனுபவவாத அடிப்படையில் நிகழ்த்தப்பட்டது. சாம்ஸ்கி இத்தகைய ஆய்வுகளைச் செயல்திறன் சார்ந்தது என்றும் குறையுள்ளது என்றும் கூறிப் பகுத்தறிவு வாதத்திற்கு வித்திட்டார். இதன்பயனாகக் கள ஆய்வுகள் ஊக்கமிழந்து உள்ளுணர்வு ஆய்வுகள் ஊக்கமடைந்தன. சாம்ஸ்கி கள ஆய்வாளர்களால் திரட்டப்பட்ட தரவு குறையுடையது என்று வாதிட்டார். ஆனால் தரவுத்தொகுதியின் வரவால் சாம்ஸ்கி கூறிய குறைபாடுகள் நிவர்த்தி செய்யப்பட்டு மொழியில் ஆய்வுகள் கணிமொழியியல் ஆய்வாக மாறியது. உள்ளுணர்வு அடிப்படையிலான பல கருதுகோள்கள் தவறானவை என்பதை தரவுத்தொகுதி ஆய்வு நமக்கு உணர்த்தியது. தரவுத்தொகுதிகள் மொழி வல்லுனர்களால் திறமையாகப் பயன்படுத்தப்படுகின்றது. அகராதிவியலார்கள், கலைச்சொல்லாக்க வல்லுனர்கள், கோட்பாட்டியலார்கள், பயன்பாட்டு மொழியலார்கள் தத்தம் செயல்பாடுகளுக்கு இன்றைய காலகட்டத்தில் தரவுத்தொகுதியைப் பெரிதும் பயன்படுத்துகின்றர். தரவுத்தொகுதியைக் கொண்டு அகராதிகள் உருவாக்கப்படுகின்றன. கலைச்சொல்லாக்கத்திற்காக தரவுத்தொகுதி பெரிதும் உதவுகின்றது. கோட்பாட்டாளர்கள் தங்கள் கோட்பாடுகளை நிரூபிப்பதற்கும் தங்கள் கருதுகோள்களின் உண்மை நிலையை அறியவும் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். பயன்பாட்டு மொழியிலார் பல மொழியில் பயன்பாடுகளுக்கு தரவுத்தொகுதியைப்

பயன்படுத்துகின்றனர். கணினி மொழியிலார்கள், கற்றல் கற்பித்தல் வல்லுனர்கள், சமூக மொழியலார்கள், உளமொழியியலார்கள், மானிட மொழியலார்கள், பண்பாட்டு மொழியியலார்கள் போன்றோர் தரவுத்தொகுதியை தங்கள் ஆய்விற்கும் பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பயன்படுத்துகின்றனர். பொருளடக்க வல்லுனர்களாக வரலாற்றியலார்கள், இலக்கிய திறனாய்வாளர்கள், சமூகவியலார்கள் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். வரலாற்றியலார்களுக்கு தரவுத்தொகுதி பல வழிகளில் கைகொடுக்கின்றது. தரவுத்தொகுதியைப் பயன்படுத்தி ஒரு நிகழ்வின் வரலாற்று காலம், இடம் போன்ற கூறுகளைச் சரியாகக் கணிக்கின்றனர். சமூகவியலார் வேறுபட்ட வகுப்பு, இனம், குடி சார்ந்த உண்மைகளையும் தகவல்களையும் தரவுத்தொகுதியின் உதவியால் சரியாகக் கணிக்கின்றனர். தகவல் வல்லுனர்களாக தகவல் மீட்பாளர்கள், கணினி மொழியியலார்கள், இயந்திர மொழிபெயர்ப்பாளர்கள், மொழியாய்வு ஆய்வாளர்கள் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். தகவல் மீட்பாளர்கள் தரவுத்தொகுதியிலிருந்து தேவையான தகவல்களைப் பெற வேண்டி பல நுட்பங்களை உருவாக்கிச் செயல்படுகின்றனர். அவர்கள் அறிவு அடிப்படை உருவாக்கத்திற்கும் செயற்கை அறிவுக்கும் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். கணினி மொழியியலார்கள் மொழியை ஆயவும் பகுப்பாயவும் முக்கியமான உபாயமாகச் செயல்படும் விரிதரவில் காணப்படும் புள்ளியியல் முறைமையுடன் தங்கள் செயல்பாட்டை ஒருங்கிணைக்க விரிதரவைப் பயன்படுத்துகின்றனர். மேலும் தரவுத்தொகுதித் தரவு மற்றும் அறிவடிப்படையின் மூலவளமாக, புள்ளியியல் உத்திகள் இலக்கண அடிப்படையில் ஆயப்பட்ட தரவுத்தொகுதிகளின் விடுவரல்களின் மீது செயல்படுகையில் கூடுதல் திறனுள்ளதாய் மாறுவதன் காரணமாக மொழியில் உள்ள முறைமைகளின் இருப்பை அல்லது இல்லாமையைப் பரிசோதிக்கப் பயன்படுத்தப்படுகின்றது. இயந்திர மொழிபெயர்ப்பாளர்கள் தரவுத்தொகுதி ஒழுங்குமுறைகளின் உண்மையான திறனை அதிகரிக்க வேண்டி குறிப்பிடத்தக்கப் பங்களிப்பு செய்வதன் காரணமாகத் தேவையான மற்றும் பொருத்தமான மொழியியல் தகவல்களைப் பெறவும் அவர்களின் ஒழுங்குமுறைகளின் திறமையைப் பரிசோதிக்கவும் தரவுத்தொகுதியை நாடுகின்றனர். மொழி ஆய்விகளை உருவாக்கவேண்டி குறிப்புரை செய்யப்படாத மற்றும் குறிப்புரை செய்யப்பட்ட தரவுத்தொகுதிகளைக் கூடுதலாகப் பயன்படுத்தப்படுவதால் மொழி ஆய்வாளர்கள் பல்வேறு வகைப்பட்ட தரவுத்தொகுதிகளின் உருவாக்கத்தால் நன்மை அடைகின்றார்கள். தரவுத்தொகுதி ஆய்வாளர்கள்,

தொழிநுட்பவியலார்கள், எழுத்தாளர்கள், அகராதியலார்கள், கல்வியாளர்கள், ஆசிரியர்கள், மாணவர்கள், மொழிகற்பவர்கள், அறிவாளிகள், வெளியீட்டாளர்கள் மற்றும் பிறர் உள்ளடங்கிய எல்லோருக்கும் பயனுள்ளதாக இருக்கின்றது.

==========

துணைநின்ற நூல்களும் கட்டுரைகளும்

இராசேந்திரன் ச. 2001. தற்காலத் தமிழ்ச் சொற்களஞ்சியம் [Thesaurus for Modern Tamil], தமிழ்ப் பல்கலைக்கழகம், தஞ்சாவூர், 2001.

இராசேந்திரன் ச. மற்றும் ச. பாஸ்கரன் 2006. தமிழ் மின்சொற்களஞ்சியம். [Tamil electronic thesaurus], தமிழ்ப் பல்கலைக்கழகம், தஞ்சாவூர், 2006.

இராசேந்திரன் ச. மற்றும் க. பாக்கியராஜ். 2019. தமிழ் வினைகளின் பொருண்மை மாற்றமும் பொருண்மை நீட்சியும் [Semantic Change and Semantic Extension of Tamil Verbs]. Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:6 June 2019.

இராசேந்திரன் ச. மற்றும் செ. தமிழ்ச்செல்வம். 2019. வரலாற்று மொழியியல் அடிப்படையில் தமிழ்ப் பெயர்சொற்கள் ஆய்வு [A Historical Linguistic Study of Tamil Nouns]. Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:7 July 2019.

இராசேந்திரன் ச. மற்றும் க. அனிதா. 2019. தமிழ்ச் சொற்றொகையின் மூலப்பொருண்மையியல் ஆய்வு [Ontology of Tamil Vocabulary]. Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:8 August 2019.

இராசேந்திரன் ச. மற்றும் அ. தனவள்ளி. 2019. தமிழில் பொருண்மை மயக்க நீக்கம் [Word Sense Disambiguation in Tamil]. Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:9 September 2019.

இராசேந்திரன் ச. 2019. ஒலியியக்கவியலும் உரையிலிருந்து பேச்சாக்கமும் பேச்சிலிருந்து உரையாக்கமும் [Acoustic phonetics and Text to Speech Processing and Speech to Text Processing.] Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:10 October 2019.

இராசேந்திரன் ச. 2019. கணினி மொழியியலும் தமிழ்மொழியின் தொழில் நுட்ப வளர்ச்சியும் [Computational Linguistics and Technological Development of Tamil]. Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:11 November 2019.

இராசேந்திரன் ச. 2019. இயந்திர மொழிபெயர்ப்பு – நேற்று, இன்று, நாளை [Machine Translation – Yesterday, Today and Tomorrow.] Language in India www.languageinindia.com ISSN 1930-2940 Vol. 20:1 January 2020.

இராசேந்திரன் ச. மற்றும் இரா. அமுதா. 2020. தமிழுக்கான எழுத்துப்பிழைத் திருத்தியும் இலக்கணப்பிழைத் திருத்தியும் உருவாகுவதற்கான நுணுக்கங்கள் [Nuances of Making Spell and Grammar Checker for Tamil.] Language in India www.languageinindia.com Vol. 20:2, Feb 2020.

இராசேந்திரன் ச. பொதுமொழியியல் [General Linguistics]. Uploaded in Academia.edu.

இராசேந்திரன் ச. பொருண்மையியல் [Semantics]. Uploaded in Academia.edu.

இராசேந்திரன் ச. தொடரியல் [Syntax]. Uploaded in Academia.edu.

இராசேந்திரன் ச. தமிழில் ஒலியியக்கவியலும் உரையிலிருந்து பேச்சும். [Acoustic phonetics and Text to speech in Tamil]. Uploaded in Academia.edu.

இராசேந்திரன் ச. உளமொழியியல் [Psycholinguistics]. Uploaded in Academia.edu.

இராசேந்திரன் ச. கணினி மொழியியல் [Computational linguistics] Uploaded in Academia.edu.

இராசேந்திரன் ச. இயற்கைமொழி ஆய்வு [Natural language processing] Uploaded in Academia.edu.

இராசேந்திரன் ச. சொல்வலை [Word net]. Uploaded in Academia.edu.

இராசேந்திரன் ச. மூலப்பொருண்மையியல் ஆய்வு [Ontology]. Uploaded in Academia.edu.

இராசேந்திரன் ச. விரிதரவு மொழியியல் [Corpus linguistics]. Uploaded in Academia.edu.

இராசேந்திரன் ச. தமிழ் பெயர்ச்சொற்களின் ஆக்கமுறை அகராதி [Generative semantics of Tamil nouns]. Uploaded in Academia.edu.

இராசேந்திரன் ச. தமிழ் வினைச்சொற்களின் ஆக்கமுறை அகராதி [Generative semantics of Tamil verbs]. Uploaded in Academia.edu.

இராசேந்திரன் ச. தமிழ் வாக்கியங்களின் பொருண்மையியல் அமைப்பு [Semantic structure of Tamil sentences]. Uploaded in Academia.edu.

இராசேந்திரன் ச. தமிழில் ஒலியியக்கவியலும் உரையிலிருந்து பேச்சும் [Acoustic phonetics and Text to speech in Tamil].

இராசேந்திரன் ச. தமிழ் வினைகளின் உருபனியல் பகுப்பாய்வி [Morphological Analyzer for Tamil Verbs]. Uploaded in Academia.edu.

இராசேந்திரன் ச. தமிழ் வாக்கியங்களின் பொருண்மை அமைப்பு [Semantic Structure of Tamil Sentences. Uploaded in Academia.edu.

இராசேந்திரன் ச. தமிழில் இயந்திர மொழிபெயர்ப்பும் அதற்கான கருவிகள் உருவாக்கமும் [Machine translation and Preparation of tools for machine translation]. Uploaded in Academia.edu.

Aarts, J. And Meijs, W. (Eds.) 1984. Corpus Linguistics: Recent Development in the Use of Computer Corpora in English Language Research. Amsterdam-Atlanta, GA.: Rodopi.

Aarts, J. and Meijs, W. (Eds.) 1986. Corpus Linguistics II: New Studies in the Analysis and Explanation of Computer Corpora. Amsterdam-Atlanta, GA.: Rodopi.

Abdel Monem, A., Shaalan, K., Rafea, A., Baraka, H., 2008. Generating Arabic Text in Multilingual Speech-to-Speech Machine Translation Framework, Machine Translation, Springer, Netherlands, 20(4): 205–258, December 2008.

Agirre, E.; Lopez de Lacalle, A.; Soroa, A. (2009). "Knowledge-based WSD on Specific Domains: Performing better than Generic Supervised WSD" (PDF). Proc. of IJCAI.

Agirre, E. Lopez & O. 2003. Clustering WordNet Word Senses. In Proc. of the Conference on Recent Advances on Natural Language (RANLP'03), Borovetz, Bulgaria, pp. 121–130.

Agirre, E.; M. Stevenson. 2006. Knowledge sources for WSD. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY.

Aijmer, K. and Allwood, J. (Eds.) 2004. Dialogic Analysis: New Trends in Dialogue Analysis. Tubingen: Niemeyer.

Aijmer, K. and Altenberg, B. (Eds.) 1991. English Corpus Linguistics: Studies in Honour of Jan Svartvik. London: Longman.

Airio, Eija .2008. "Who benefits from CLIR in web retrieval?". Journal of Documentation. 64 (5): 760–778. doi:10.1108/00220410810899754. Retrieved on May 22 2020.

Albat, Thomas Fritz. 2012."Systems and Methods for Automatically Estimating a Translation Time." US Patent 0185235, 19 July 2012.

Allahyari , Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. Retrieved on March 12 2020.

Allen, J. and A. Frisch.1982. "What's in a Semantic Network". In: Proceedings of the 20th. annual meeting of ACL, Toronto, pp. 19–27.

Allen, Jonathan; Hunnicutt, M. Sharon; Klatt, Dennis .1987. From Text to Speech: The MITalk system. Cambridge University Press. ISBN 978-0-521-30641-6.

Altenberg B. and Granger, S. (Eds.) 2002. Lexis in Contrast: Corpus-based Approaches. Amsterdam/ Philadelphia: John Benjamins.

Anand Kumar M., Dhanalakshmi V., Soman K.P., and Rajendran S. 2009. A Novel Approach for Tamil Morphological Analyzer. In: Proceedings of Tamil Internet Conference 2009. Cologne, Germany, pp. 23-35,October, 2009.

Anand Kumar M., Dhanalakshmi V., Soman K.P., and Rajendran S. 2010. A Sequential Labelling Approach to Morphological analyzer for Tamil Language", (IJCE) International Journal of Computer Science and Engineering. Vol 02, No 06, 2010, 2201-2208.

Anandkumar M, Rajendran S and Soman K.P. 2014. "Tamil Word Sense Disambiguation using Support Vector Machines and Rich Features". International Journal of Applied Engineering Research (IJAER) ISSN 0973-4562 Volume 20 (2014) pp. 7609-7620 @ Research India Publications http://www.ripublication.com

Anand Kumar M, Rajendran S, and Soman K.P. 2015. "Cross-lingual Preposition Disambiguation for Machine Translation", Eleventh International Multi Conference on Information Processing - 2015, August 21st-23rd, 2015, Bangalore [SCOPUS Indexed].

Anand Kumar, M., S. Rajendran, and K. P. Soman. 2014. "Supervised Cross-lingual Preposition Disambiguation for Machine Translation" iDravidian' 2014. Symposium on Natural Language Processing for Dravidian Languages, ICON 2014.

Anand Kumar, M., S. Rajendran, and K. P. Soman. 2014. AMRITA@ FIRE-2014: "Morpheme Extraction for Tamil using Machine Learning". Working notes of MET shared Task, "Forum for Information Retrieval Evaluation (FIRE) 2014.

Anand Kumar M, Rajendran S, and Soman K.P. 2015. "Cross-lingual Preposition Disambiguation for Machine Translation", Procedia Computer Science, volume 54C, 2015, pages 291-300.

Anderson, D.D. 1995. Machine translation as a tool in second language learning. CALICO Journal. 13(1). 68–96.

Andersen, Francis I.; Forbes, A. DEAN, 2003, "Hebrew Grammar Visualized: I. Syntax", Ancient Near Eastern Studies, 40, pp. 43–61.

Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Amodei, Dario.2016. "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin". arXiv:1512.02595 [cs.CL].

"Appendix III of 'The present status of automatic translation of languages', Advances in Computers, vol.1 (1960), p.158-163. Reprinted in Y.Bar-Hillel: Language and information (Reading, Mass.: Addison-Wesley, 1964), p.174-179" (PDF). Retrieved on 12 June 2020.

Arnold, Douglas N. 2000. "Computer-Aided Instruction," Microsoft® Encarta® Online Encyclopedia 2000. http://encarta.msn.com © 1997-2000 Microsoft Corporation.

Aaron van den Oord .2018. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis". arXiv:1711.10433 [cs.CL].

Arık, Sercan Ö.; Chen, Jitong; Peng, Kainan; Ping, Wei; Zhou, Yanqi .2018. "Neural Voice Cloning with a Few Samples", Advances in Neural Information Processing Systems, 31, arXiv:1802.06006

Assael, Yannis; Shillingford, Brendan; Whiteson, Shimon; de Freitas, Nando (5 November 2016). "LipNet: End-to-End Sentence-level Lipreading". arXiv:1611.01599 [cs.CV].

Aston, G. (Ed.) 2004. Learning with Corpora. Cambridge: Cambridge University Press.

Atkins, B. T. S. 1996. Bilingual dictionaries: Past, present and future. In M. Gellerstam, J. Jarborg, S.-G. Malmgren, L. Rogström, & C. R. Papmehl (Eds.), Euralex'96 Proceedings (pp. 515–546). Göteborg: Göteborg University.

Automatic summarization. From Wikipedia, the free encyclopedia as on 26.06.2020.

Atkins, B.T.S. 1994. "A Corpus-Based Dictionary", in: The Oxford-Hachette French Dictionary. OxforcUParis: Oxford University Press/Hachette Livre, pp. xix-xxvi.

Atkins, B. T. S., & Rundell, M. 2008. The Oxford Guide to Practical Lexicography. Oxford University Press.

Atteveldt, Wouter Van 2008. Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content. BookSurge Publishing.

Atwell, E. (Eds.) 1993. Corpus Based Computational Linguistics. Amsterdam: Rodopi.

Babych, Bogdan; Anthony Hartley, and Serge Sharoff .2007. "Translating from under-resourced languages: comparing direct transfer against pivot translation". Proceedings of MT Summit XI, 10–14 September 2007, Copenhagen, Denmark. pp.29—35.

Bahdanau, Dzmitry; Cho, Kyunghyun; Bengio, Yoshua. 2014. "Neural Machine Translation by Jointly Learning to Align and Translate". arXiv:1409.0473 [cs.CL].

Ballesteros, L., Croft, W.B.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, CA, 1–8, 1997.

Baker, M., Gill, F. and Tognini-Bonelli, E. (Eds.) 1993. Text and Technology: In honour of John Sinclair. Philadelphia: John Benjamins.

Baker, Paul and Jesse Egbert (eds). 2016. Triangulating Methodological Approaches in Corpus-Linguistic Research. Routledge.

Baker, J.; Li Deng; Glass, J.; Khudanpur, S.; Chin-Hui Lee; Morgan, N.; O'Shaughnessy, D. 2009. "Developments and Directions in Speech Recognition and Understanding, Part 1". IEEE Signal Processing Magazine. 26 (3): 75–80. Bibcode:2009ISPM...26...75B. doi:10.1109/MSP.2009.932166.

Baker, Paul; Egbert, Jesse, (eds.) 2016. Triangulating Methodological Approaches in Corpus-Linguistic Research. New York: Routledge.

Bar-Hillel, Yehoshua.1960. "Automatic Translation of Languages". Available online at http://www.mt-archive.info/Bar-Hillel-1960.pdf

Bar-Hillel, Yehoshua.1964. Language and Information: Selected Essays on Their Theory and Application. Reading, MA: Addison-Wesley. pp. 174–179.

Barnbrook, G. 1998. Language and Computers. Edinburgh: Edinburgh University Press.

Barrault, Loïc; Bojar, Ondřej; Costa-jussà, Marta R.; Federmann, Christian; Fishel, Mark; Graham, Yvette; Haddow, Barry; Huck, Matthias; Koehn, Philipp; Malmasi, Shervin; Monz, Christof. 2019. "Findings of the 2019 Conference on Machine Translation (WMT19)". Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). Florence, Italy: Association for Computational Linguistics: 1–61. doi:10.18653/v1/W19-5301. Retrieved on May 20 2020.

Bahdanau, Dzmitry .2016. "End-to-End Attention-based Large Vocabulary Speech Recognition". arXiv:1508.04395 [cs.CL]. Retrieved 22 June 2019.

Bird, Jordan J.; Wanner, Elizabeth; Ekárt, Anikó; Faria, Diego R. 2020. "Optimisation of phonetic aware speech recognition through multi-objective evolutionary algorithms". Expert Systems with Applications. Elsevier BV. 153: 113402. doi:10.1016/j.eswa.2020.113402. ISSN 0957-4174.

Barron, Brenda; Owner, Business. 2019. "Babel Fish: What Happened To The Original Translation Application?: We Investigate". Digital.com. Retrieved 22 June 2019.

"Basic OCR in OpenCV | Damiles". Blog.damiles.com. November 20, 2008. Retrieved June 16, 2020.

Baskaran, S. 2002. Semantic analyser for word sense disambiguation. MS thesis. Madras Institute of Technology, Anna University, Chennai 2002.

Baskaran Sankaran, Vaidehi V, 2002, "Role of Collocations and Case-Markers inWord Sense Disambiguation: A Clustering-Based Approach". In Proceedings of IEEE International Symposium on Natural Language Processing and Knowledge Engineering 2002. Vol. I. pp. 625-630. Hammamet, Tunisia.

Baskaran S, Vaidehi V, 2003, "Collocation based Word Sense Disambiguation using Clustering for Tamil". Communicated to International Journal of Dravidian Linguistics, Thiruvananthapuram, India.

Bates A.W. 1995 “Technology, open learning and distance education.” London: Routledge.

Bengio, Y. (1991). Artificial Neural Networks and their Application to Speech/Sequence Recognition (Ph.D.). McGill University.

Béjoint, H. 2000. Modern Lexicography: An Introduction. Oxford: Oxford University Press.

Bendeck, Fawsy .2008. WSM-P workflow semantic matching platform. München: Verl. Dr. Hut. ISBN 9783899638547. OCLC 501314022.

Benesty, Jacob; Sondhi, M. M.; Huang, Yiteng. 2008. Springer Handbook of Speech Processing. Springer Science & Business Media. ISBN 978-3540491255.

Bergmanis, Toms; Goldwater, Sharon. "Context Sensitive Neural Lemmatization with Lematus" (PDF). Retrived on 22 July 2020.

Bernadini, S. 2000. Competence, Capacity, Corpora. Bologna: CLUEB.

Bernini G. (Ed.) 2001. Pragmatic Organisation of Discourse in the languages of Europe. Vol. 1. Berlin: Mouton de Gruyter.

Bharati, A. , Chaitanya, V. and Sangal, R. 1995. Natural Language Processing : A Paninian Perspective; Prentice Hall India, 1995.

Bharati, A. and Sangal, R. 1993. "Parsing free word order languages in the Paninian Framework", Proceedings of the Annual Meeting of Association for Computational Linguistics, New York, 1993.

Bharati, A., Bhatia, M., Chaitanya, V. and Sangal, R. 1997. "Paninian Grammar Framework Applied to English", South Asian Language Review, Creative Books, New Delhi, 1997.

Bharati, A., Chaitanya, V. And Sangal, R.1995. Natural Language Processing: A Paninian Perspective. New Delhi: Prentice-Hall.

Bhattacharya, Indrajit, Lise Getoor, and Yoshua Bengio. 2004. Unsupervised sense disambiguation using bilingual probabilistic models. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

Bhattacharyya, Pushpak. 2010. IndoWordNet. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Malta, Irec, May, 2010.

"BitCrawl by Hobson Lane". Archived from the original on October 27, 2012. Retrieved 2012-05-29.

Bibcode:2004JEI....13..146S. doi:10.1117/1.1631315. Retrieved May 2, 2020.

Biber, D. 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D., CONRAD, S. and REPPEN, R. 1998. Corpus linguistics, Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Black, Alan W. 2002. Perfect synthesis for all of the people all of the time. IEEE TTS Workshop 2002.

Black, A.H. and Black, C.A. 1997. A Conceptual Introduction to Morphological Parsing Using AMPLE System. Dallas, Texas: Summer Institute of Linguistics.

Black, E., Garside, R. and Leech, G. (Eds.) 1993. Statistically-driven Computer Grammars of English: the IBM/Lancaster Approach. Amsterdam: Rodopi.

Blechman, R. O.; Blechman, Nicholas (23 June 2008). "Hello, Hal". The New Yorker. Archived from the original on 20 January 2015. Retrieved 17 January 2015.

Bloom, L. 1970. Language Development: Form and Function in Emerging Grammars. Cambridge, Mass.: MIT Press.

Bordes, Antoine; Usunier, Nicolas; Garcia-Duran, Alberto; Weston, Jason; Yakhnenko, Oksana 2013.

Braschler, M., Krause, J., Peters, P., Schäuble, P.: Cross-Language Information Retrieval (CLIR) Track Overview, In Proceedings of the Seventh Text Retrieval Conference (TREC-7). NIST, Gaithersburg, MD, 1999.

Boguraev, B. and Pustejvsky, J. (Eds.) 1996. Corpus Processing for Lexical Acquisition. Cambridge, Mass.: MIT Press.

Booker, Ellis (14 March 1994). "Voice recognition enters the mainstream". Computerworld. p. 45.

Boretz, Adam. 2009. "AppTek Launches Hybrid Machine Translation Software" SpeechTechMag.com (posted 2 MAR 2009)". Speechtechmag.com. Retrieved 12 June 2020.

Botley, S.P., A.M. Mcenery, and A. Wilson (Eds.) 2000. Multilingual Corpora in Teaching and Research. Amsterdam-Atlanta, GA: Rodopi.

Bouillon, P. and BUSA, F. (Eds.) 2001. The Language of Word Meaning. Cambridge: Cambridge University Press.

"British English definition of voice recognition". Macmillan Publishers Limited. Retrieved on 21 February 2020.

Brian (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups". IEEE Signal Processing Magazine. 29 (6): 82–97. Bibcode:2012ISPM...29...82H. doi:10.1109/MSP.2012.2205597.

Brown, R. 1973. A First Language: The Early Stages. Cambridge, Mass.: Harvard University Press.

Brownlee, Jason. 2017. A Gentle Introduction to Text Summarization. In Deep Learning for Natural Language Processing. Last Updated on August 7, 2019.

Buitelaar, P.; B. Magnini, C. Strapparava and P. Vossen. 2006. Domain-specific WSD. In Word Sense Disambiguation: Algorithms and Applications, E. Agirre and P. Edmonds, Eds. Springer, New York, NY.

Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z. (eds.), 2018. "Translating Embeddings for Modeling Multi-relational Data" (PDF), Advances in Neural Information Processing Systems 26, Curran Associates, Inc., pp. 2787–2795. Retrieved on 29 June 2020.

Burton, D.M. 1973. Shakespeare's grammatical style: a computer-assisted analysis of Richard II and Anthony and Cleopatra. Austin: University of Texas Press.

Butler, C.S. (Ed.) 1992. Computers and Written Texts. Oxford: Blackwell Publishers.

Cao, M.; X.Sun and H. Zhuge, 2018. The contribution of cause-effect link to representing the core of scientific paper—The role of Semantic Link Network, PLOS ONE, 2018, doi:10.1371/journal.pone.0199303.

Caridakis, George; Castellano, Ginevra; Kessous, Loic; Raouzaiou, Amaryllis; Malatesta, Lori; Asteriadis, Stelios; Karpouzis, Kostas (19 September 2007). Multimodal emotion recognition from expressive faces, body gestures and speech. IFIP the International Federation for Information Processing. 247. Springer US. pp. 375–388. doi:10.1007/978-0-387-74161-1_41. ISBN 978-0-387-74160-4.

Carter, R. and Mccarthy, M. (Eds.) 1988. Vocabulary and Language Teaching. London: Longman.

Carol, E. and C.F. Meyer (Eds.) 1997. Synchronic Corpus Linguistics. Bergen, Norway: ICAME.

Madsen, Mathias Winther. 2009. The Limits of Machine Translation. M.A. thesis, University of Copenhagen. Retrieved on 12 May 2020.

Cassidy, F.G. Ed. 1985. Dictionary of American Regional English. Vol.1: Introduction and A-C. Cambridge, MA.: Harvard University Press.

Cernak, M.; A. Asaei, and A. Hyafil, 2018. “Cognitive speech coding: examining the impact of cognitive speech processing on speech compression,” IEEE Signal Processing Magazine, vol. 35, no. 3, pp. 97–109, 2018.

Chan, Y. S.; H. T. Ng. 2005. Scaling up word sense disambiguation via parallel texts. In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI, Pittsburgh, PA).

Chan, William; Jaitly, Navdeep; Le, Quoc; Vinyals, Oriol (2016). "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition" (PDF). ICASSP.

Cerf, Vinton; Wrubel, Rob; Sherwood, Susan. "Can speech-recognition software break down educational language barriers?". Curiosity.com. Discovery Communications. Archived from the original on 7 April 2014. Retrieved 26 March 2014.

Chan, William; Zhang, Yu; Le, Quoc; Jaitly, Navdeep (10 October 2016). "Latent Sequence Decompositions". arXiv:1610.03035 [stat.ML].

Charniak, E. 1997. “Statistical parsing with a context-free grammar and word statistics”, Proceedings of the 14th AAAI, Menlo Park, 1997.

Chen, Danqi; Fisch, Adam; Weston, Jason; Bordes, Antoine (2017). "Reading Wikipedia to Answer Open-Domain Questions". arXiv:1704.00051 [cs.CL].

Chitu, Alex. 2007. "Google Switches to Its Own Translation System". Googlesystem.blogspot.com. Retrieved on 13 August 2020.

Chomsky, A.N. 1968. Language and Mind. New York: Harcourt Brace.

Chomsky, A.N. 1972. Studies in Generative Grammar. The Hague: Mouton.

Chorowski, Jan; Jaitly, Navdeep (8 December 2016). "Towards better decoding and language model integration in sequence to sequence models". arXiv:1612.02695 [cs.NE].

Chung, Joon Son; Senior, Andrew; Vinyals, Oriol; Zisserman, Andrew (16 November 2016). "Lip Reading Sentences in the Wild". arXiv:1611.05358 [cs.CV].

Chung, Yu-An .2018. "Semi-Supervised Training for Improving Data Efficiency in End-to-End Speech Synthesis". arXiv:1808.10128 [cs.CL]. Retrieved on May 22, 2020.

Chu, W. C. Speech Coding Algorithms: Foundation and Evolution of Standardized Coders, John CLL POS-tagger.

Ciaramella, Alberto. "A prototype performance evaluation report." Sundial workpackage 8000 (1993).

Cimiano, Philipp; Christina Unger; John McCrae (1 March 2014). Ontology-Based Interpretation of Natural Language. Morgan & Claypool Publishers. ISBN 978-1-60845-990-2.

Claburn, Thomas (25 August 2017). "Is it possible to control Amazon Alexa, Google Now using inaudible commands? Absolutely". The Register. Archived from the original on 2 September 2017.

Coleman, J. and Kay, C.J. (Eds.) 2000. Lexicology, Semantics and Lexicography: Selected Papers from the 4th G.L. Brook Symposium. Amsterdam/ Philadelphia: John Benjamins.

Coldewey, Devin.2017. "DeepL schools other online translators with clever machine learning". TechCrunch. Retrieved on March 20, 2020..

Collins, Allan M.; M. R. Quillian.1969. "Retrieval time from semantic memory". Journal of Verbal Learning and Verbal Behavior. 8 (2): 240–247. doi:10.1016/S0022-5371(69)80069-1.

Collins, Allan M.; M. Ross Quillian.1970. "Does category size affect categorization time?". Journal of Verbal Learning and Verbal Behavior. 9 (4): 432–438. doi:10.1016/S0022-5371(70)80084-6.

Collins, Allan M.; Elizabeth F. Loftus. 1975. "A spreading-activation theory of semantic processing". Psychological Review. 82 (6): 407–428. doi:10.1037/0033-295x.82.6.407.

Collins, P. and BLAIR, D. (Eds.) 1989. Australian English. St. Lucia: University of Queensland Press.

"Comparison of MT systems by human evaluation, May 2008". Morphologic.hu. Archived from the original on 19 April 2012. Retrieved on 12 June 2020.

”Computational linguistics”. Retrieved Wikipedia on 12 June 2020.

“Corpus linguistics”. Retrieved Wikipedia on 12 June 2020.

Crystal, D. 1995. The Cambridge Encyclopaedia of the English Language. Cambridge: Cambridge University Press.

Cuyckens, H. and Zawada, B. (Eds.) 2001. Polysemy in Cognitive Linguistics. Amsterdam/Philadelphia: John Benjamins.

Dahl, George E.; Yu, Dong; Deng, Li; Acero, Alex.2012. "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition". IEEE Transactions on Audio, Speech, and Language Processing. 20 (1): 30–42. doi:10.1109/TASL.2011.2134090.

Dash, Niladri Sekhar and Someone. 2000. “The process of designing a multidisciplinary monolingual sample corpus”. International Journal of Corpus Linguistics. 5(2): 179-197.

Dash, Niladri Sekhar. 2005. Corpus Linguistics, and Language Technology: With Reference to Indian Languages. New Delhi: Mittal Publications.

Dash, Niladri Sekhar. 2007. “Frequency-based analysis of words and morphemes in Bengali text corpus”. Indian Journal of Linguistics. Vol. 25. No. 26. Pp. 223-253, 2007.

Deng, L.; Hassanein, K.; Elmasry, M. 1994. "Analysis of the correlation structure for a neural predictive model with application to speech recognition". Neural Networks. 7 (2): 331–339. doi:10.1016/0893-6080(94)90027-2. Retrieved on March 22 2020.

Deng, L.; M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton.2010. Binary Coding of Speech Spectrograms Using a Deep Auto-encoder. Interspeech.

Deng L., Li, J., Huang, J., Yao, K., Yu, D., Seide, F. et al. 2013. Recent Advances in Deep Learning for Speech Research at Microsoft. ICASSP, 2013.

Deng, L.; Li, Xiao. 2013. "Machine Learning Paradigms for Speech Recognition: An Overview" (PDF). IEEE Transactions on Audio, Speech, and Language Processing. 21 (5): 1060–1089. doi:10.1109/TASL.2013.2244083.

Deng, Li; Yu, Dong .2014. "Deep Learning: Methods and Applications" (PDF). Foundations and Trends in Signal Processing. 7 (3–4): 197–387. CiteSeerX 10.1.1.691.3679. doi:10.1561/2000000039. Archived (PDF) from the original on 22 October 2014.

Derose, S.J. 1988. "Grammatical category disambiguation by statistical optimization." Computational Linguistics 14(1): 31–39.

Derose, S.J. 1990. "Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages." Ph.D. Dissertation. Providence, RI: Brown University Department of Cognitive and Linguistic Sciences. Electronic Edition available at

Desai, S.K. 1974. Experimentation with language in Indian Writing in English (Fiction). Monograph of the Dept. of English, Shivaji University, Kohlapur, India.

Dewey, Godfrey.1923. Relativ Frequency of English Speech Sounds. Harvard: Harvard University Press.

Dhanalakshmi V., Padmavathy P., Anand Kumar M., Soman K.P., and Rajendran S. 2009. “Chunker For Tamil Using Machine Learning” 7th International Conference on Natural Language Processing 2009 (ICON2009), IIIT Hyderabad, India, December 2009.

Dhanalakshmi V, Padmavathy P, Anand Kumar M, Soman K P, and Rajendran S 2009, “Chunker for Tamil”, Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, IEEE Press, doi: 10.1109/ARTCom. 2009.191.

Dhanalakshmi V, Anand kumar M, Shivapratap G, Soman K.P, and Rajendran S. 2009. “Tamil POS Tagging using Linear Programming”. International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May-2009. ISSN 1797-9617.

Dhanalakshmi V, Anand kumar M, Rekha R U, Arun kumar C, Soman K P, and Rajendran S. 2009. “Morphological Analyzer For Agglutinative Languages Using Machine Learning Approaches”. Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, ARTCom 2009, Oct 2009 Kottayam, India

Dhanalakshmi V , Anand kumar M, Rekha R U, Soman K P and Rajendran S. 2010. “Grammar Teaching Tools for Tamil”, Proceedings of Technology for Education Conference (T4E), IIT Bombay, India, 2010.

Diab, Mona, and Philip Resnik. An unsupervised method for word sense tagging using parallel corpora. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002.

Diamantini, C.; Mircoli, A.; Potena, D.; Storti, E. (2015). "Semantic disambiguation in a social information discovery system". 2015 International Conference on Collaboration Technologies and Systems (CTS): 326–333. doi:10.1109/CTS.2015.7210442. ISBN 978-1-4673-7647-1. S2CID 13260353.

Drahota, A. 2008. "The vocal communication of different kinds of smile" (PDF). Speech Communication. 50 (4): 278–287. doi:10.1016/j.specom.2007.10.001. Archived from the original (PDF) on 2013-07-03.

Drew, Harwell (2019-09-04). "An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft". washingtonpost.com. Washington Post. Retrieved 2019-09-08.

Dukes, K., Atwell, E. and Habash, N. 2011. 'Supervised Collaboration for Syntactic Annotation of Quranic Arabic'. Language Resources and Evaluation Journal.

DuPont, Quinn (January 2018). "The Cryptological Origins of Machine Translation: From al-Kindi to Weaver". Amodern (8).

Dutoit, T.; V. Pagel, N. Pierret, F. Bataille, O. van der Vrecken. The MBROLA Project: Towards a set of high quality speech synthesizers of use for non commercial purposes. ICSLP Proceedings, 1996.

Eaton, H. 1940. Semantic Frequency List for English, French, German and Spanish. Chicago: Chicago University Press.

Edwards, A.W. and R.L Chambers.1964. “Occurrence of various language properties in English”. Journal of the Association for Computing Machinery. 2: 465-482.

Edwards, J.A. and Lampert, M.D. (Eds.) 1993. Talking Data: Transcription and Coding in Discourse Research. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Eggins, S. 1994. An Introduction to Systemic Functional Linguistics. London: Pinter Publishers.

Electronic Dictionary Research (Edr) Project. Down loaded from internet on 15.06.2020.

Elderton, W.P. 1949. “A few statistics on the length of English words”. Journal of Royal Statistics. Series A. CXII: 436-445.

Englund, Christine (2004). Speech recognition in the JAS 39 Gripen aircraft: Adaptation to speech at different G-loads (PDF) (Masters thesis). Stockholm Royal Institute of Technology. Archived (PDF) from the original on 2 October 2008.

Eniko Hja. 2010. Dictionary building based on parallel corpora and word alignment. In Proceedings of the XIV Euralex International Congress, Leeuwarden, pages 6-10, Leeuwarden/Ljouwert, Netherlands, July, 2010.

Eyland, E. Ann (1987), 'Revelations from Word Counts', in Newing, Edward G.; Conard, Edgar W. (eds.) Perspectives on Language and Text: Essays and Poems in Honor of Francis I. Andersen's Sixtieth Birthday, July 28, 1985, Winona Lake, IN: Eisenbrauns, p. 51, ISBN 0-931464-26-9.

Facchinetti, R. 2007, Theoretical Description and Practical Applications of Linguistic Corpora. Verona: QuiEdit, ISBN 978-88-89480-37-3

Facchinetti, R. (ed.) Corpus Linguistics 25 Years on. New York/Amsterdam: Rodopi, 2007 ISBN 978-90-420-2195-2

Facchinetti, R. AND Rissanen M. (eds.) Corpus-based Studies of Diachronic English. Bern: Peter Lang, 2006 ISBN 3-03910-851-4.

Farias, R. C. and J. M. Brossier, 2013. "Adaptive quantizers for estimation," Signal Processing, vol. 93, no. 11, pp. 3076–3087, 2013.

Farwell, David; Gerber, Laurie; Hovy, Eduard. 2003. Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas, AMTA'98, Langhorne, PA, USA, October 28–31, 1998 Proceedings. Berlin: Springer. p. 276. ISBN 3540652590.

Fehr, Tiff, How We Sped Through 900 Pages of Cohen Documents in Under 10 Minutes, Times Insider, The New York Times, March 26, 2020.

Fellbaum, Christiane. 1999. WordNet. Blackwell Publishing Ltd, United Kingdom.

Fellbaum, Christiane; Vossen, Piek. 2012. "Challenges for a multilingual wordnet". Language Resources and Evaluation. 46 (2): 313–326. doi:10.1007/s10579-012-9186-z.

Fernandez, Santiago; Graves, Alex; Schmidhuber, Jürgen 2007). "Sequence labelling in structured domains with hierarchical recurrent neural networks" (PDF). Proceedings of IJCAI. Archived (PDF) from the original on 15 August 2017.

"First-Hand:The Hidden Markov Model – Engineering and Technology History Wiki". ethw.org. Retrieved 1 May 2018.

Flesch, R. 1946. The Art of Plain Talk. New York: Harper & Brothers.

Fiori, Alessandro. 2014. Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding. IGI Global; 1 edition.

Flowerdew, J. and Peacock, M. (Eds.) 2001. Research Perspectives on English for Academic Purposes. Cambridge: Cambridge University Press.

Follensbee, Bob; McCloskey-Dale, Susan .2000. "Speech recognition in schools: An update from the field". Technology And Persons With Disabilities Conference 2000. Retrieved 26 March 2020.

Forgrave, Karen E. 2002. "Assistive Technology: Empowering Students with Disabilities." Clearing House 75.3 (2002): 122–6. Web.

Francis, W.N. and Kucera, H. 1964. Manual of information to accompany A standard Corpus of present-day edited American English. Dept. of Linguistics, Brown University, USA.

Frakes, W.B., Baeza-Yates, R. 1992. Information Retrieval: Data Structures and Algorithms, Prentice-Hall, 1992.

Francis, W.N. and Kucera, H. 1982. Frequency Analysis of English Usage: Lexicon and Grammar. Boston: Houghton Mifflin.

Fries, U., MÜ, V. and Schneider, P. (Eds.) 1997. From Aelfric to the New York Times. Amsterdam: Rodopi.

Fuchs, C. and Vitorri, B. (Eds.) 1994. Continuity in Linguistic Semantics. Amsterdam and Philadelphia: John Benjamins.

Fung, Pascale. 2000. A statistical view on bilingual lexicon extraction. In Jean Veronis, ´ editor, Parallel Text Processing, pages 219–236. Kluwe.

Furbach, Ulrich, Ingo Glöckner, and Björn Pelzer. "An application of automated reasoning in natural language question answering." Ai Communications 23.2-3 (2010): 241-265.

Ganesh Ambedkar, 2019. “Corpus analysis of poola and maatiri”. 18th Tamil Internet Confeernce Papers, Compiled by Vasu Ranganathan & Sobha L. International Forum for Information Technology in Tamil (INFIT, California, USA , pages 111-116.

Ganesh Ambedkar. Landscape Conflicts – an Analysis. (personal copy from the author).

Ganesh Ambedkar. Corpus Analysis of 'caavu' and 'maraNam’. (personal copy from the author).

Ganesh Ambedkar. Tokens of Appreciation. (personal copy from the author).

Ganesh Ambedkar. A measurement on tanittamizh iyakkam. (personal copy from the author).

Galitsky, Boris (2005). "Disambiguation Via Default Rules Under Answering Complex Questions". International Journal on Artificial Intelligence Tools. 14: 157–175. doi:10.1142/S0218213005002041.

Galitsky, Boris.2003. Natural Language Question Answering System: Technique of Semantic Headers. International Series on Advanced Intelligence. Volume 2. Australia: Advanced Knowledge International. ISBN 978-0-86803-979-4.

Galitsky, B & Pampapathi R. 2005. Can many agents answer questions better than one. First Monday. 2005;10. doi:10.5210/fm.v10i1.1204.

Gangemi, A.; Navigli, R.; Velardi, P. (2003). The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet (PDF). Proc. of International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE 2003). Catania, Sicily (Italy). pp. 820–838.

Garrett, Jennifer Tumlin; et al. 2011. "Using Speech Recognition Software to Increase Writing Fluency for Individuals with Physical Disabilities". Journal of Special Education Technology. 26 (1): 25–41. doi:10.1177/016264341102600104.

Garside, R., Leech, G. and Mcenery, A. (Eds.) 1997. Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman.

Garside, R., Leech, G. and Sampson, G. (Eds.) 1987. The Computational Analysis of English: A Corpus Based Approach. London: Longman.

Gaussier, E.. 1998. Flow network models for word alignment and terminology extraction from bilingual corpora. In Proceedings of the joint 17th International Conference on Computational Linguistics and 26th Annual Meeting of the Association for Computational Linguistics, pages 444–45.

Geer, David, "Statistical Translation Gains Respect", pp. 18 – 21, IEEE Computer, October 2005". Ieeexplore.ieee.org. 27 September 2011. doi:10.1109/MC.2005.353.

Gerbig, A. 1997. Lexical and Grammatical Variation in a Corpus: A Computer-Assisted Study of Discourse on the Environment. London: Peter Lang Publishing.

Ghadessy, M., Henry, A. and Roseberry, R.L. Eds. 2001. Small Corpus Studies and ELT: Theory and Practice. Amsterdam/ Philadelphia: John Benjamins.

Gibson, H.N. 1962. The Shakespeare Claimants: A Critical Survey of the Four Principle Theories Concerning the Authorship of the Shakespearean Play. London: Methuen and Co.

Gliozzo, A.; B. Magnini and C. Strapparava. 2004. Unsupervised domain relevance estimation for word sense disambiguation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP, Barcelona, Spain).

"Inside Google Translate – Google Translate".

http://www.mt-archive.info/10/HyTra-2013-Tambouratzis.pdf

Goel, Vaibhava; Byrne, William J. 2000. "Minimum Bayes-risk automatic speech recognition". Computer Speech & Language. 14 (2): 115–135. doi:10.1006/csla.2000.0138. Archived from the original on 25 July 2011. Retrieved 28 March 2011.

Good, I.J. 1957. “Distribution of word frequencies”. Nature. 179: 595.

"Google Translator: The Universal Language". Blog.outer-court.com. 25 January 2007. Retrieved on 12 June 2020.

Google Blog: The machines do the translating (by Franz Och)

"Google's neural network learns to translate languages it hasn't been trained on".

https://blogs.microsoft.com/ai/chinese-to-english-translator-milestone/. Missing or empty |title= (help).

Gonzalo, Julio, Braschler, Martin, Kluck, Michael (Eds.) 2003. Comparative Evaluation of Multilingual Information Access Systems. 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers Springer-Verlag Berlin and Heidelberg GmbH & Co. KG.

Gordin, Michael D. 2015. Scientific Babel: How Science Was Done Before and After Global English. Chicago, Illinois: University of Chicago Press. ISBN 9780226000299.

Gove, P.B. (Ed.) 1961. Webster's Third New International Dictionary of the English Language Sentences. New York: Hartcourt, Brace and World.

Graf, D. 1996. Relative Clauses in Their Discourse Context: A Corpus-Based Study. Unpublished M.A. Thesis: Freiburg.

Granger, S. and Tyson, S.P. (Eds.) 2003. Extending the Scope of Corpus-Based Research: New Applications, New Challenges. Amsterdam: Rodopi.

Granger, S., Hung, J. and Tyson, S.P. Eds. 2002. Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Amsterdam: John Benjamins.

Graves, Alex; Mohamed, Abdel-rahman; Hinton, Geoffrey. 2013. "Speech recognition with deep recurrent neural networks". arXiv:1303.5778 [cs.NE]. ICASSP 2013.

Graves, Alex. 2014. "Towards End-to-End Speech Recognition with Recurrent Neural Networks" (PDF). ICML.

Gray, Robert M. 2010. "A History of Realtime Digital Speech on Packet Networks: Part II of Linear Predictive Coding and the Internet Protocol" (PDF). Found. Trends Signal Process. 3 (4): 203–303. doi:10.1561/2000000036. ISSN 1932-8346.

Green JR, Bert F; et al. 1961. "Baseball: an automatic question-answerer" (PDF). Western Joint IRE-AIEE-ACM Computer Conference: 219–224.

Greenbaum, S. (Ed.) 1996. Comparing English Worldwide: The International Corpus of English. Oxford: Clarendon.

Greene, B. and Rubin, G. 1971. Automatic Grammatical Tagging of English. Technical Report. Department of Linguistics. Brown University, RI, USA.

Grefenstette, G. (ed.). 1998. Cross-Language Information Retrieval, The Kluwer International Series on Information Retrieval, Kluwer Academic Publishers, Boston, 1998.

Gupta, Maya R.; Jacobson, Nathaniel P.; Garcia, Eric K. (2007). "OCR binarisation and image pre-processing for searching historical documents" (PDF). Pattern Recognition. 40 (2): 389. doi:10.1016/j.patcog.2006.04.043. Archived from the original (PDF) on October 16, 2015. Retrieved May 2, 2020.

Gallafent, Alex. 2011. "Machine Translation for the Military". PRI's the World. Retrieved 17 September 2013.

Habib, Raza .2019. "Semi-Supervised Generative Modeling for Controllable Speech Synthesis". arXiv:1910.01709 [cs.CL].

Han et al. 2012. "LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors," in Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Posters, pages 441–450, Mumbai, India.

Halliday, M.A.K. 1987. Spoken and Written Modes of Meaning, Comprehending Oral and Written Language. San Diego, CA: Academic Press.

Halliday, M.A.K. 1989. Spoken and Written Language. Oxford: Oxford University Press.

Halliday, M.A.K. and Hasan, R. 1976. Cohesion in English. London: Longman.

Halteren, H.V. (Ed.) 1999. Syntactic Word Class Tagging. Dordrecht: Kluwer Academic Press.

Hanzo L., F. C. A. Somerville, and J. P. Woodard, 2007. Voice and Audio Compression for Wireless Communications, John Wiley & Sons Ltd., Chichester, England, 2nd edition, 2007.

Harabagiu, Sanda; Hickl, Andrew .2006. "Methods for using textual entailment in open-domain question answering". Association for Computational Linguistics. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: 905–912. doi:10.3115/1220175.1220289.

Hausmann, F.J., Reichmann, O., Wiegand, H.E. and Zgusta, L. (Eds.) 1990. International Encyclopedia of Lexicography. Vol.2. Berlin: Walter de Gruyter.

Hercules, Dalianis et al . 2003. Porting and evaluation of automatic summarization. Down loaded form internet on 26.06.2020.

Herden, G. 1956. Language as Choice and Chance. Groningen, Holland: P. Noordhoff Ltd.

Herden, G. 1962. Calculus of Linguistic Observation. Hague: Mouton & Co.

Héja, E. 2010. The Role of Parallel Corpora in Bilingual Lexicography. In N. Calzolari, K. Choukry, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valetta, Malta: European Language Resources Association (ELRA).

History and Development of Speech Synthesis, Helsinki University of Technology, Retrieved on July 4, 2020.

"History of Speech Recognition". Dragon Medical Transcription. Retrieved on 17 January 2020.

Holley, Rose (April 2009). "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs". D-Lib Magazine. Retrieved on July 5, 2020.

Holmes John and Wendy Holmes .2001. Speech Synthesis and Recognition (2nd ed.). CRC. ISBN 978-0-7484-0856-6.

Hovy, E., Ide, N., Frederking, R. (eds.): Multilingual Information Management: Current Levels and Future Abilities, NSF/EC/DARPA, April 1999.

"How to optimize results from the OCR API when extracting text from an image? - Haven OnDemand Developer Community". Retrieved on July 5, 2020.

"How OCR Software Works". OCRWizard. Retrieved June 16, 2020.

Hsu, Wei-Ning (2018). "Hierarchical Generative Modeling for Controllable Speech Synthesis". arXiv:1810.07217 [cs.CL].

Hu, Hongbing; Zahorian, Stephen A. 2010. "Dimensionality Reduction Methods for HMM Phonetic Recognition" (PDF). ICASSP 2010. Archived (PDF) from the original on 6 July 2012.

Hull, D.A., Grefenstette, G.: Querying Across Languages. A Dictionary-based Approach to Multilingual Information Retrieval. In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, 49–57, 1996.

Helbig, H. 2006. Knowledge Representation and the Semantics of Natural Language (PDF). ISBN 978-3540244615.

Hermajakob, U., Knight, K., & Hal, D. 2008. Name Translation in Statistical Machine Translation Learning When to Transliterate. Association for Computational Linguistics. 389–397.

Hickey, R. and Stanislaw, P. (Eds.) 1996. Language History and Linguistic Modelling: A Festschrift for Jacek Fisiak. Vol. 2. Berlin: Mouton de Gruyter.

Hirschman, L. & Gaizauskas, R. 2001. Natural Language Question Answering. The View from Here. Natural Language Engineering (2001), 7:4:275-300 Cambridge University Press.

Hinton, Geoffrey; Deng, Li; Yu, Dong; Dahl, George; Mohamed, Abdel-Rahman; Jaitly, Navdeep; Senior, Andrew; Vanhoucke, Vincent; Nguyen, Patrick; Sainath, Tara; Kingsbury,

Hoffman, C. 1955. The Man Who was Shakespeare. New York: Julias Messner Inc.

Hofland, K. and Johansson, S. 1982. Word Frequencies in British and American English. Bergen: Norway Computing Centre for the Humanities.

Huang, Xuedong; James Baker; Raj Reddy. "A Historical Perspective of Speech Recognition". Communications of the ACM. Archived from the original on 20 January 2015. Retrieved on January 20, 2020.

Hulpus, Ioana; Prangnawarat, Narumol. 2015. "Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation". The Semantic Web – ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11–15, 2015, Proceedings, Part 1. International Semantic Web Conference 2015. Springer International Publishing. p. 444.

Hunston, S. 2002. Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Hutchins, W. J. 1986. Machine Translation: Past, Present, and Future. Chichester: Ellis Harwood.

Hybrid approaches to machine translation. Costa-jussà, Marta R.,, Rapp, Reinhard,, Lambert, Patrik,, Eberle, Kurt,, Banchs, Rafael E.,, Babych, Bogdan. Switzerland. ISBN 9783319213101. OCLC 953581497.

"IBM-Shoebox-front.jpg". androidauthority.net. Retrieved on April 4, 2020.

Ide, Nancy and Jean Véronis (Eds.) 1998. Introduction to the special issue on word sense disambiguation: the state of the art. Computational Linguistics. 24(1): 1-40.

Ide, N.; T. Erjavec, D. Tufis. 2002. Sense discrimination with parallel corpora. In Proceedings of ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions (Philadelphia, PA).

Ilson, R.F. (Ed.) 1986. Lexicography: An Emerging International Profession. Manchester: Manchester University Press.

"Improvements in voice recognition software increase". TechRepublic.com. 27 August 2002.

Ingram, D. 1989. First Language Acquisition. Cambridge: Cambridge University Press.

"In-Q-Tel". In-Q-Tel. Archived from the original on 20 May 2016. Retrieved 12 June 2020.

Jackson, William .2003. "GCN – Air force wants to build a universal translator". Gcn.com. Retrieved 12 June 2020.

Jayant N. S. and Noll, P. 1985. "Digital coding of waveforms. Principles and applications to speech and video," Signal Processing, vol. 9, no. 2, pp. 139-140, 1985.

"James Baker interview". Retrieved on February 9, 2020.

Jensen, J.T. 1990. Morphology: Word Structure in Generative Grammar. Amsterdam: John Benjamins.

Jespe rsen, O. 1909-1949. Modern English Grammar on Historical Principles. 7 Vols. London: Allen and Unwin.

Jia, Ye; Zhang, Yu; Weiss, Ron J. (2018-06-12), "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis", Advances in Neural Information Processing Systems, 31: 4485–4495, arXiv:1806.04558

Johansson, S. and Hofland, K. (Eds.) 1982. Computer Corpora in English Language Research. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S. and Stenström, A-B. (Eds.) 1991. English Computer Corpora: Selected Papers and Research Guide. Berlin: Mouton de Gruyter.

Juang, B. H.& Rabiner, Lawrence R. 2014. "Automatic speech recognition–a brief history of the technology development" (PDF): 6. Retrieved 17 July 2020.

Juan-Manuel Torres-Moreno. (Ed.). 2014. Automatic Text Summarization (Cognitive Science and Knowledge Management) 1st Edition. Wiley-ISTE.

Joshi, A.K. and Schabes, Y. 1997. "Tree-Adjoining Grammars", Handbook of Formal Languages, G. Rozenberg and A. Salomaa (eds.), Vol. 3 pp. 69 - 124; Springer, Berlin, New York, 1997.

Jurafsky, Daniel; Martin, James H. (2009). Speech and Language Processing. Pearson. pp. 906–908.

Kalchbrenner, Nal; Blunsom, Philip. 2013. "Recurrent Continuous Translation Models". Proceedings of the Association for Computational Linguistics: 1700–1709.

Kamakshi, S. and Rajendran, S. 2004. Preliminaries to the Preparation of a Machine Aid to Translate Linguistics Texts in English into Tamil.. Dravidian Linguistics Association, Thiruvananthapuram, Publication 86, 2004.

Kapidakis, Sarantos; Cezary Mazurek, Marcin Werla.2015. Research and Advanced Technology for Digital Libraries. Springer. p. 257. ISBN 9783319245928. Retrieved April 3, 2020.

Katamba, F. 1993. Morphology. London: Macmillan Press.

Kennedy, G. 1998. An Introduction to Corpus Linguistics. New York: Addison-Wesley Longman Inc.

Kenny, A.J.P. 1982. The Computation of Style. Oxford: Pergamon Press.

Keynote talk: Recent Developments in Deep Neural Networks. ICASSP, 2013 (by Geoff Hinton).

Keynote talk: "Achievements and Challenges of Deep Learning: From Speech Analysis and Recognition To Language and Multimodal Processing," Interspeech, September 2014 (by Li Deng).

Kettemann, C.B. and Marko, G. (Eds.) 2002. Teaching and Learning by Doing Corpus Analysis. Language and Computers: Studies in Practical Linguistics 42. Amsterdam-Atlanta, GA.: Rodopi.

Kilgarriff, Adam.1996. “Corpus similarity and homogeneity via word frequency”. Proceedings of the EURALEX Conference. Gothenburg, Sweden, August.

Kilgarriff, Adam. 2003. Thesauruses for natural language processing. In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pages 5 13, Beijing, China, October, 2003.

Kilgariff, Adam and J. Palmer (Eds.) 2000. Computer and the Humanities: Special Issue on Word Sense Disambiguation. Vol. 34. No.1. 2000.

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., & Rychlý, P. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In Proceedings of EURALEX 2008. Barcelona: Universitat Pompeu Fabra: 425-433

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. 2004. The Sketch Engine. In Proceedings of EURALEX 2004 (pp. 105–116). Lorient,France.

Kilgarriff, A., & Tugwell, D. 2002. Sketching words. In M.-H. Corréard (Ed.), Lexicography and natural language processing: a festschrift in honour of BTS Atkins (pp. 125–137). Euralex.

Kirk, J.M. (Ed.) 2000. Corpora Galore: Analyses and Techniques in Describing English. Amsterdam; Atlanta, GA: Rodopi.

Klatt, Dennis H. (1977). "Review of the ARPA speech understanding project". The Journal of the Acoustical Society of America. 62 (6): 1345–1366. Bibcode:1977ASAJ...62.1345K. doi:10.1121/1.381666.

Knight, Kevin. 1994. "Building a large ontology for machine translation (1993)". arXiv:cmp-lg/9407029.

Knight, K. and Hatzivassiloglou, V. 1995. “Two Level, Many Paths Generation”; Proceedings of the ACL-95. Cambridge, MA, 199.

Knight, Kevin and Vasileios Hatzivassiloglou. 1995. Two-level, many-paths generation. In Proceedings of the 33rd annual meeting on Association for Computational Linguistic, pages 252-260, Massachusetts, USA, June, 1995.

Knowlson, James. 1975. Universal Language Schemes in England And France 1600-1800.

Koehn, Philipp. 2010. Statistical Machine Translation. Cambridge: Cambridge University Press. p. 15. ISBN 9780521874151.

Koehn, Philipp and Kevin Knight. 2002.Learning a translation lexicon from monolingual corpora. In Proceedings of the Workshop on Unsupervised Lexical Acquisition, volume 9, pages 916, Philadelphia, USA, July, 2002. Association for Computational Linguistics (ACL).

Kominek, John and Alan W. Black. 2003. CMU ARCTIC databases for speech synthesis. CMU-LTI-03-177. Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Kondoz, A. M. 2004. Digital Speech: Coding for Low Bit Rate Communication Systems, John Wiley & Sons Ltd., Chichester, UK, 2nd edition, 2004.

Kowlaski, Gerald. 1997. Information Retrieval System: Theory and Implementation. Kluwar Academic Publishers.

Krishnamurthy, R. 2002. The Corpus Revolution in EFL Dictionaries. Kernerman Dictionary News, 10.

Krishnamurthy, R. 2008. Corpus-driven Lexicography. International Journal of Lexicography, 21(3), 231–242.

Kruisinga, E. 1931-32. A Handbook of Present-Day English. Goringen: Noordhoff.

Kucera, H. and Francis, W.N. 1967. Computational Analysis of Present Day American English. Providence, RI: Brown University Press.

Kuhn, T. S. 2012. The Structure of Scientific Revolutions-50th Anniversary Edition, vol. 3, 4th edition, 2012.

Kurath, H. Ed. 1954. Handbook of the linguistic geography of New England. Washington DC: American Council of Learned Societies.

Kytö, M., Ihalainen, O. and Rissanen, M. (Eds.) 1988. Corpus Linguistics, hard and soft: Proceedings of the 8th International Conference on English Language Research on Computerised Corpora. Amsterdam: Rodopi.

Lam, Khang Nhut and Jugal Kalita. 2013.“Creating reverse bilingual dictionaries.” In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 524-528, Atlanta, USA, June, 2013.

Lamel, L.F.; J.L. Gauvain, B. Prouts, C. Bouhier, R. Boesch. Generation and Synthesis of Broadcast Messages, Proceedings ESCA-NATO Workshop and Applications of Speech Technology, September 1993.

Landau, S.I. 2001. Dictionaries: The Art and Craft of Lexicography. Cambridge: Cambridge University Press.

Laufer, B. 1992. Corpus-based versus lexicographer examples in comprehension and production of new words. In Proceedings of the Fifth Euralex International Congress (pp. 4–9). Tampere: University of Tampere.

Leech, G., G. Myers, and J. Thomas (Eds.) 1995. Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman.

Leech, Geoffrey, B. Francis, and X. Xu .1994. “The use of computer corpora in the textual demonstrability of gradience in linguistic categories”. In, Fuchs, C. and B. Vitorri (Eds.) Continuity in Linguistic Semantics. Amsterdam and Philadelphia: John Benjamins. Pp. 57-76.

Lehmann, Fritz; Rodin, Ervin Y. eds. 1992. Semantic networks in artificial intelligence. International series in modern applied mathematics and computer science. 24. Oxford; New York: Pergamon Press. p. 6. ISBN 978-0080420127. OCLC 26391254.

Lehrberger, John. 1988. Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation. John Benjamins Publishing. ISBN 90-272-3124-9.

Lenders, W. Computational lexicography and corpus linguistics until ca. 1970/1980, in: Gouws, R. H., Heid, U., Schweickard, W., Wiegand, H. E. (eds.) Dictionaries - An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin: De Gruyter Mouton, 2013 ISBN 978-3112146651.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the 5th SIGDOC (New York, NY). 24–26.

Levy, M. 1997. Computer Assisted Language Learning. Oxford: Oxford University Press.

Lin, J. 2002. The Web as a Resource for Question Answering: Perspectives and Challenges. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002).

"Li Deng". Li Deng Site.

Linden, Krister and Lauri Carlson. 2010. FinnWordNet-WordNet pa finska via overs attning. LexicoNordica, 17:119140.

"LipNet: How easy do you think lipreading is?". YouTube. Archived from the original on 27 April 2017. Retrieved on May 5, 2017.

Liu, H.; Christiansen, T.; Baumgartner, W. A.; Verspoor, K. 2012. "BioLemmatizer: A lemmatization tool for morphological processing of biomedical text". Journal of Biomedical Semantics. 3: 3.

Ljung, M. (Ed.) 1997. Corpus-Based Studies in English. Papers from the 17th International Conference on English-Language Research Based on Computerized Corpora. Amsterdam: Rodopi.

Lock, A. (Ed.) 1978. Action, Gesture and Symbol: The Emergence of Language. London: Academic Press.

Lorge, I. 1949. Semantic Content of the 570 Commonest English Words. New York: Columbia University Press.

"Machine Translation Service". 5 August 2011.

"Machine Translation: No Copyright On The Result?". SEO Translator, citing Zimbabwe Independent. Retrieved on 24 July 2020.

Macwhinney, B. 1991. The CHILDES Project: Tools for Analyzing Talk. Hillsdale, N.J.: Lawrence Erlbaum.

Mair, C. and Hundt, M. (Eds.) 2000. Corpus Linguistics and Linguistics Theory. Amsterdam-Atlanta, GA: Rodopi.

Makhoul, John "ISCA Medalist: For leadership and extensive contributions to speech and language processing". Archived from the original on 24 January 2018. Retrieved 23 January 2018.

Mani, Inderjeet & Mark T. Maybury (Eds.). 1999. Advances in Automatic Text Summarization The MIT Press. abridged edition.

Mann, Bruce. 2009. Computer-Aided Instruction.

https://www.researchgate.net/publication/228176284. Down loaded from intenet on 04./08/2020.

Manning, C. D.; Raghavan, P.; Schütze, H.. Introduction to Information Retrieval. Cambridge University Press.

Manish Sinha, Mahesh Kumar, Prabhakar Pande, Laxmi Kashyap, and Pushpak Bhattacharyya. Hindi word sense disambiguation. In International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India, 2004.

Manning and Schutze, 1999. Manning, C. and Schutze, H.; Foundations of Statistical Natural Language Processing, pp. 407 – 409; MIT Press, 1999.

Markoff, John (2011-02-16). "On 'Jeopardy!' Watson Win is All but Trivial". The New York Times.

Markoff, John (23 November 2012). "Scientists See Promise in Deep-Learning Programs". New York Times. Archived from the original on 30 November 2012. Retrieved 20 January 2015.

Mary S. Neff Michael C. McCord.1990. "Acquiring Lexical Data from Machine-Readable Dictionary Resources for Machine Translation". IBM T. J. Watson Research Center, P. O. Box 704, Yorktown Heights, New York 10598: 85–90. CiteSeerX 10.1.1.132.8355.

Maas, Andrew L.; Le, Quoc V.; O'Neil, Tyler M.; Vinyals, Oriol; Nguyen, Patrick; Ng, Andrew Y. 2012. "Recurrent Neural Networks for Noise Reduction in Robust ASR". Proceedings of Interspeech 2012.

Manone, V.K. Rajendran S., Anandkumar, M & Soman, K.P. 2014. “A new TAG formalism for Tamil parser analysis”. iDravidian 2014: Symposium on Natural Language Processing for Dravidian Languages, 17th December 2014, University of Goa, Goa as a part of ICON 2014 from 18th to 19th December 2014.

Manone, V.K. Rajendran S., & Soman, K.P. 2015. A Synchronised Tree Adjoining Grammar for English to Tamil Machine Translation.

Manone, V.K. Rajendran S., & Soman, K.P. 2015. “A Synchronous Syntax for English-Tamil language pair for Machine Translation”. Fourth International Symposium on Natural Language Processing (NLP’15). Kochi, Kerala: Co-affiliated with Fourth International Conference in Computing, Communications and Informatics (ICACCI-2015).

Menon, Vijay Krishna. S Rajendran and K.P. Soman. 2015. "Training Tee Adjoining Grammars with Huge Text Corpus using Spark MapReduce", Soft Computing Models for Big Data, volume 05, issue 04, July 2015, pages 1021-1026. (ISSN 0976-6561).

Melby, Alan. 1995. The Possibility of Language. Amsterdam:Benjamins, 1995, 27–41. Benjamins.com. 1995. ISBN 9789027216144. Retrieved on 12 June 2020.

Mcarthur, D., and Sampson G. Corpus Linguistics: Readings in a Widenting Discipline, Continuum, ISBN 0-8264-8803-X.

Mcarthur, T. 1981. Longman Lexicon of Contemporary English. London: Longman.

McCarthy, J. 1982. Formal Problems in Semitic Phonology and Morphology. New York: Garland.

McCarthy, M. 1998. Spoken Language and Applied Linguistics. Cambridge: cambridge University press.

McCarthy, D.; R. Koeling, J. Weeds, J. Carroll. 2007. Unsupervised acquisition of predominant word senses. Computational Linguistics 33(4): 553–590.

McCarthy, D.; R. Navigli. 2009. The English Lexical Substitution Task, Language Resources and Evaluation, 43(2), Springer.

McCormack C.& D Jones, 1998. “Building a Web-Based Education System”, John Wiley & Sons. 1998; ISBN: 0471191620.

McCusker, James P. & Chastain, Katherine. 2016. "What is a Knowledge Graph?". authorea.com. Retrieved 15 June 2020.

McDermott, A. .2002. “Early dictionaries of English and historical corpora: in search of hard words”. In, Vera, J.E. D. (Ed.) A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics. Amsterdam-New York, NY: Rodopi. Pp. 197-226.

Mcenery, T. and Wilson, A. 1996. Corpus Linguistics. Edinburgh: Edinburgh University Press.

Mcenery, T., Rayson, P. and Wilson, A. (eds.) 2002. A Rainbow of Corpora: Corpus Linguistics and the Langauges of the World. München: Lincom Europa.

McKean, Kevin (8 April 1980). "When Cole talks, computers listen". Sarasota Journal. AP. Retrieved 23 June 2020.

McLoughlin, I. V. 2016. Speech and Audio Processing: a MATLABBased Approach, Cambridge University Press, Cambridge, UK, 2016.

Melamed, Dan. 1996. Automatic construction of clean broad-coverage translation lexicons. In 2nd Conference of the Association for Machine Translation in the Americas, Montreal, Canada.

Merriam-Webster. (1984). Merriam-Webster's dictionary of synonyms. Springfield, US: Springfield, Mass.: Merriam-Webster,

Meyer, C.F. 2002. English Corpus Linguistics. Cambridge: Cambridge University Press.

"Microsoft researchers achieve new conversational speech recognition milestone". 21 August 2017.

Milestones in machine translation – No.6: Bar-Hillel and the nonfeasibility of FAHQT Archived 12 March 2007 at the Wayback Machine by John Hutchins.

Milyaev, Sergey; Barinova, Olga; Novikova, Tatiana; Kohli, Pushmeet; Lempitsky, Victor (2013). "Image binarisation for end-to-end text understanding in natural images" (PDF). Document Analysis and Recognition (ICDAR) 2013. 12th International Conference on: 128–132. doi:10.1109/ICDAR.2013.33. ISBN 978-0-7695-4999-6. Retrieved May 2, 2020.

Mittal, Sparsh & Ankush Mittal. 2011. "Versatile question answering systems: seeing in synthesis", Mittal et al., IJIIDS, 5(2), 119-142, 2011.

Miller, G.A. 1951. Language and Communication. New York: McGraw-Hills.

Miller, G.A.. 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11):39-41, 1995.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. 1990. Five Papers on WordNet. CSL Report 43, Cognitive Science Laboratory, Princeton University, Princeton.

Miller, G.A., E.B. Newman, and E.A. Friedman. 1958. “Length-frequency statistics for written English”. Information and Control. 1(2): 370-389.

Milyaev, Sergey; Barinova, Olga; Novikova, Tatiana; Kohli, Pushmeet; Lempitsky, Victor.2013. "Image binarisation for end-to-end text understanding in natural images" (PDF). Document Analysis and Recognition (ICDAR) 2013. 12th International Conference on: 128–132. doi:10.1109/ICDAR.2013.33. ISBN 978-0-7695-4999-6. Retrieved on May 2, 2020.

Mindt, D. 1995. An Empirical Grammar of the English Verb: Modal Verbs. Berlin: Cornelsen Verlag.

Mitkov, R. (Ed.) 2003. The Oxford Handbook of Computational Linguistiucs. Oxford: Oxford University Press.

Mittal et al. 2011. "Versatile question answering systems: seeing in synthesis", International Journal of Intelligent Information Database Systems, 5(2), 119-142.

"molto-project.eu". molto-project.eu. Retrieved 12 June 2020.

Mohammad, S; G. Hirst. 2006. Determining word sense dominance using a thesaurus. In Proceedings of the 11th Conference on European chapter of the Association for Computational Linguistics (EACL, Trento, Italy).

Mohan Raj, S.N. and Rajendran S. 2016. “Tamil Oriented Machine Translation under Indian Language to Indian Language Machine Translation (ILILMT) consortium.” In: Proceedings of 15th World Tamil Internet conference 2016, held at Gandhi Gram Rural University, Tamil Nadu, September 8-11, 2016, pages 393-402.

Mohri, M. (2002). "Edit-Distance of Weighted Automata: General Definitions and Algorithms" (PDF). International Journal of Foundations of Computer Science. 14 (6): 957–982. doi:10.1142/S0129054103002114. Archived (PDF) from the original on 18 March 2012. Retrieved 28 March 2011.

Moldovan, Dan, et al. "Cogex: A logic prover for question answering." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

Moore, Robert C. 2001. Towards a simple and accurate statistical approach to learning translational relationships among words. In Proceedings of the workshop on datadriven machine translation, Toulouse, July. 39th annual meeting of the Associate for Computational Linguistics.

Moore, R. K. 2005. “Cognitive informatics: the future of spoken language processing?,” in Proceedings of the 10th International Conference on Speech and Computer (SPECOM), Patras, Greece, October 2005.

Morante, Roser; Martin Krallinger, Alfonso Valencia and Walter Daelemans. 2012. Machine Reading of Biomedical Texts about Alzheimer's Disease. CLEF 2012 Evaluation Labs and Workshop. September 17, 2012.

Morgan, Bourlard, Renals, Cohen, Franco.1993. "Hybrid neural network/hidden Markov model systems for continuous speech recognition. ICASSP/IJPRAI"

Muegge .2006., "Fully Automatic High Quality Machine Translation of Restricted Text: A Case Study," in Translating and the computer 28. Proceedings of the twenty-eighth international conference on translating and the computer, 16–17 November 2006, London, London: Aslib. ISBN 978-0-85142-483-5.

Muegge, Uwe. 2006. "An Excellent Application for Crummy Machine Translation: Automatic Translation of a Large Database", in Elisabeth Gräfe (2006; ed.), Proceedings of the Annual Conference of the German Society of Technical Communicators, Stuttgart: tekom, 18–21.

Muller, Thomas; Cotterell, Ryan; Fraser, Alexander; Schütze, Hinrich. "Joint Lemmatization and Morphological Tagging with LEMMING" (PDF).

Muralishankar, R; Ramakrishnan, A.G.; Prathibha, P .2004. "Modification of Pitch using DCT in the Source Domain". Speech Communication. 42 (2): 143–154. doi:10.1016/j.specom.2003.05.001.

Nadia, Lachtar. 2014. Design and implementation of information retrieval system based ontology. Downloaded from internet on 29.06.2020. 978-1-4799-3824-7/14/$31.00 ©2014 IEEE

Nation, I.S.P. 1997. "Vocabulary size, text coverage, and word lists", in Schmitt; McCarthy (eds.), Vocabulary: Description, Acquisition and Pedagogy, Cambridge: Cambridge University Press, pp. 6–19, ISBN 978-0-521-58551-4.

Nagao, M. 1981. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in Artificial and Human Intelligence, A. Elithorn and R. Banerji (eds.) North-Holland, pp. 173–180, 1984.

Navigli, R. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. Proc. of the 44th Annual Meeting of the Association for Computational Linguistics joint with the 21st International Conference on Computational Linguistics (COLING-ACL 2006), Sydney, Australia.

Navigli, R.; A. Di Marco. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. Computational Linguistics, 39(3), MIT Press, 2013, pp. 709–754.

Navigli, R.; G. Crisafulli. Inducing Word Senses to Improve Web Search Result Clustering. Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), MIT Stata Center, Massachusetts, USA.

Navigli, R.; M. Lapata. An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 32(4), IEEE Press, 2010.

Navigli, R.; K. Litkowski, O. Hargraves. 2007. SemEval-2007 Task 07: Coarse-Grained English All-Words Task. Proc. of Semeval-2007 Workshop (SemEval), in the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic.

Navigli, R.;P. Velardi. 2005. Structural Semantic Interconnections: a Knowledge-Based Approach to Word Sense Disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 27(7).

Nelson, G., Wallis, S. and Aarts, B. 2002. Exploring Natural Language: Working with the British Component of the International Corpus of English. Amsterdam/ Philadelphia: John Benjamins.

Nesi, H. 1999. A User's Guide to Electronic Dictioaries fo Language Leaners. Intenatioal Jornal of Lexicography, 12(1) 55-66.

Nesi, H. 2000a Electronic dictionaries in second language vocabulary comprehension and acquisition: The state of the art. In U. Heid, S. Evert, E. Lehmann, & C. Rohrer (eds) , IX EURALEX International Conference (pp.839-847). Stuttgart.

Nesi, H. 2000. On Screen or in Print? Students' use of a learner's dictionary on CD-ROM in book form. In P. Howarth & R. Herington (eds.), Issues in EAP Leaning Technologies (pp. 106-114). Leeds: Leeds University Press.

Nguyen P. 2010. "Automatic classification of speaker characteristics". International Conference on Communications and Electronics 2010. pp. 147–152. doi:10.1109/ICCE.2010.5670700. ISBN 978-1-4244-7055-6.

Nikolic J. and Z. H. Peric, 2008. “Lloyd-Max’s algorithm implementation in speech coding algorithm based on forward adaptive technique,” Informatica (Lithuanian Academy of Sciences), vol. 19, no. 2, pp. 255–270, 2008.

Nikolic, J.; Z. H. Peric, and A. Z. Jovanovic, 2016. “Two forward adaptive dual-mode companding scalar quantizers for Gaussian source,” Signal Processing, vol. 120, pp. 129–140, 2016.

Nida, E. A., & Taber, C. R. 2003. The Theory and Practice of Translation (4th ed.). Brill.

Nihalani, P., Tongue, R.K. and Hosali, P. 1979. Indian and British English: A handbook of Usage and pronunciation. New Delhi: Oxford University Press.

Nino, Ana. 2009. "Machine Translation in Foreign Language Learning: Language Learners' and Tutors' Perceptions of Its Advantages and Disadvantages" ReCALL: the Journal of EUROCALL 21.2 (May 2009) 241–258. Retrieved on 20 March 2020.

NIPS Workshop: Deep Learning for Speech Recognition and Related Applications, Whistler, BC, Canada, Dec. 2009 (Organizers: Li Deng, Geoff Hinton, D. Yu).

Nirenburg, Sergei. 1989. "Knowledge-Based Machine Translation". Machine Trandation 4 (1989), 5 - 24. Kluwer Academic Publishers. 4 (1): 5–24. JSTOR 40008396.

"Nuance Exec on iPhone 4S, Siri, and the Future of Speech". Tech.pinions. 10 October 2011. Archived from the original on 19 November 2011. Retrieved 23 July 2020.

Nye, Mary Jo. 2016. "Speaking in Tongues: Science's centuries-long hunt for a common language". Distillations. 2 (1): 40–43. Retrieved 20 March 2020.

Oakes, M.P. 1998. Statistics for Corpus Linguistics. Edinburgh: Edinburgh University Press.

Oard, Douglas. "Multilingual Information Access." Understanding Information Retrieval Systems(2011): 373-80. Web.

Oard, D.W.: Web Language Distribution. Web site for Research Resources on Cross-Language Text Retrieval. See: http://www.clis.umd.edu/dlrg/filter/papers/

"Obituaries: Stephen Balashek". The Star-Ledger. 22 July 2012.

Och, F. J., & Ney, H. 2003. A systematic comparison of various statistical alignment models. Computational linguistics, 29(1), 19–51.

Ofoghi, Bahadorreza; John Yearwood & Liping Ma. 2008. The impact of semantic class identification and semantic role labeling on natural language answer extraction. The 30th European Conference on Information Retrieval (ECIR'08). Springer Berlin Heidelberg. pp. 430–437. doi:10.1007/978-3-540-78646-7_40.

Ofoghi Bahadorreza; John Yearwood & Liping Ma. 2009. "The impact of frame semantic annotation levels, frame-alignment techniques, and fusion methods on factoid answer processing". Journal of the American Society for Information Science and Technology. 60 (2): 247–263. doi:10.1002/asi.20989.

"Optical Character Recognition (OCR) – How it works". Nicomsoft.com. Retrieved June 16, 2020.

OCR Introduction". Dataid.com. Retrieved June 16, 2020.

O'keeffe, Anne AND Mccarthy, Michael (ed.) 2010. The Routledge Handbook of Corpus Linguistics. London: Routledge.

Oltramari, A.; Gangemi, A.; Guarino, N.; Masolo, C. (2002). Restructuring WordNet's Top-Level: The OntoClean approach. OntoLex'2 Workshop, Ontologies and Lexical Knowledge Bases (LREC 2002). Las Palmas, Spain. pp. 17–26. CiteSeerX 10.1.1.19.6574.

Olvera-Lobo, María-Dolores. "Cross-Language Information Retrieval on the Web." Handbook of Research on Social Dimensions of Semantic Technologies and Web Services(n.d.): 704-19. Web.

Ooi, V.B.Y. 1997. Computer Corpus Lexicography. Edinburgh: Edinburgh University Press

Oostdijk, N. and Dehann, P. (Eds.) 1994. Corpus Based Research into Language. Amsterdam-Atlanta, GA: Rodopi.

Orton, H. 1962. Introduction to Survey of English Dialects. Leads: E.J. Arnold and Son Ltd.

Ordentlich O. and U. Erez, 2019. “Performance analysis and optimal filter design for sigma-delta modulation via dualitywith DPCM,” IEEE Transactions on Information 9eory, vol. 65, no. 2, pp. 1153–1164, 2019.

Pajzs, J. 2009. On the Possibility of Creating Multifunctional Lexicographical Databases. In H. Bergenholtz, S. Nielsen, & S. Tarp (eds.), Lexicography at a crossroads. Dictionaries and encyclopedias today, lexicographical tools tomorrow (pp. 327–354). Bern: Lang.

Palmer, H. 1933. Second Interim Report on English Collocations. Tokyo: Institute for Research in English Teaching.

Partington, A. 1998. Patterns and Meanings - Using Corpora for English Language Research and Teaching. Amsterdam/Philadelphia: John Benjamins.

Palmer, M.; O. Babko-Malaya and H. T. Dang. 2004. Different sense granularities for different applications. In Proceedings of the 2nd Workshop on Scalable Natural Language Understanding Systems in HLT/NAACL (Boston, MA).

Partha Lal June 13, 2002. Text Summarization. Downloaded form interenet on 26.06.2020.

Pașca, Marius .2005. "Book Review New Directions in Question Answering Mark T. Maybury (editor) (MITRE Corporation) Menlo Park, CA: AAAI Press and Cambridge, MA: The MIT Press.

Pati, P.B.; Ramakrishnan, A.G. (May 29, 1987). "Word Level Multi-script Identification". Pattern Recognition Letters. 29 (9): 1218–1229. doi:10.1016/j.patrec.2008.01.027.

Paul, J. D. 2019. “Re-creating the sigsaly quantizer: this 1943 analog-to-digital converter gave the allies an unbreakable scrambler-(resources),” IEEE Spectrum, vol. 56, no. 2, pp. 16-17, 2019.

Perera, R., Nand, P. and Naeem, A. 2017. Utilizing typed dependency subtree patterns for answer sentence generation in question answering systems.

Percy, C., Meyer, C.F. and Lancashire, I. (Eds.) 1996. Synchronic Corpus Linguistics. Amsterdam-Atlanta, GA: Rodopi.

Perera, R. and Perera, U. 2012. Towards a thematic role based target identification model for question answering.

Peric Z. H. and J. Nikolic, 2012. “An adaptive waveform coding algorithm and its application in speech coding,” Digital Signal Processing, vol. 22, no. 1, pp. 199–209, 2012.

Peric Z. and J. Nikolic, 2012. “High-quality Laplacian source quantisation using a combination of restricted and unrestricted logarithmic quantisers,” IET Signal Processing, vol. 6, no. 7, pp. 633–640, 2012.

Peters, B.P., Collins, P. and SMITH, A. (Eds.) 2002. New Frontiers of Corpus Research. Language and Computers. Amsterdam-Atlanta, GA: Rodopi.

Petrov, S.; Das, D.; Mcdonald, R. 2011.. "A Universal Part-of-Speech Tagset". 11 Apr 2011, arXiv:1104.2086.

Pierce, John R. 1969. "Whither speech recognition?". Journal of the Acoustical Society of America. 46 (48): 1049–1051. Bibcode:1969ASAJ...46.1049P. doi:10.1121/1.1911801.

Pilehvar, M. T. Jurgens D. and Navigli R. 2013. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity.. Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria, August 4–9, 2013, pp. 1341-1351.

Pinker, S. 1995. The Language Instinct: The New Science of Language and Mind. Middlesex, England: Penguin Books Ltd.

Pinola, Melanie (2 November 2011). "Speech Recognition Through the Decades: How We Ended Up With Siri". PC World. Retrieved 22 October 2020.

"Pioneering Speech Recognition". 7 March 2012. Retrieved on July 18, 2020.

Poddar, Arnab; Sahidullah, Md; Saha, Goutam (March 2018). "Speaker Verification with Short Utterances: A Review of Challenges, Trends and Opportunities". IET Biometrics. 7 (2): 91–101. doi:10.1049/iet-bmt.2017.0065.

Pollack, Pickett, Sumby .1974. Experimental phonetics. MSS Information Corporation. pp. 251–258. ISBN 978-0-8422-5149-5.

Ponzetto, S. P.; R. Navigli. Knowledge-rich Word Sense Disambiguation rivaling supervised systems. In Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), 2010.

Poon, Hoifung, and Pedro Domingos. 2009. "Unsupervised semantic parsing." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009.

Poplack, S. 'The Care and Handling of Mega-corpus, in Fasold, R and Schiffrin, D. (eds.) Languge Change and Variation, Amsterdam: Benjamins, 1989, 411-451.

"POS tags". Sketch Engine. Lexical Computing. 2018-03-27. Retrieved on July 2020.

Poutsma, H. 1926-29. A Grammar of Late Modern English. Groningen: Noordhoff.

Prathosh, A. P.; Ramakrishnan, A. G.; Ananthapadmanabha, T. V. (December 2013). "Epoch extraction based on integrated linear prediction residual using plosion index". IEEE Trans. Audio Speech Language Processing. 21 (12): 2471–2480. doi:10.1109/TASL.2013.2273717.

Prenger, Ryan .2018. "WaveGlow: A Flow-based Generative Network for Speech Synthesis". arXiv:1811.00002 [cs.SD].

Preyer, W. 1989. The Mind of a Child. New York: Appleton.

Pustejovsky, J. 1995. The Generative Lexicon. Cambridge, MA: MIT Press.

Quillian, M. R. (1967). 2009. Word concepts: A theory and simulation of some basic semantic capabilities. Behavioral Science, 12(5), 410-430.

Quarteroni, Silvia, and Suresh Manandhar. "Designing an interactive open-domain question answering system." Natural Language Engineering 15.1 (2009): 73-95.

Quillian, M. R. (1968). Semantic memory. Semantic information processing, 227–270.

Quillian, M. R. (1969). "The teachable language comprehender: a simulation program and theory of language". Communications of the ACM. 12 (8): 459–476. doi:10.1145/363196.363214.

Quillian, R. 1966. Semantic Memory. Unpublished doctoral dissertation, Carnegie Institute of Technology, 1966.

Qing-An Zeng (October 28, 2015). Wireless Communications, Networking and Applications: Proceedings of WCNA 2014. Springer. ISBN 978-81-322-2580-5.

Quirk, R. 1960. 'Towards a Description of English Usage', Transactions of the Philological Society 1960, 40-61.

Quirk, R., Gimson, A.C. and Warburg, J. 1968. The Use of English. 2nd Edition. London: Longman Group Ltd.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. A Comprehensive Grammar of the English Language. London: Longman.

Rabiner, Lawrence.1984.. "The Acoustics, Speech, and Signal Processing Society. A Historical Perspective" (PDF). Retrieved 23 January 2020.

Ralf D. Brown. "Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation" (PDF). Language Technologies Institute (Center for Machine Translation) Carnegie Mellon University Pittsburgh, PA 15213-3890 USA. Retrieved 2 Junly 2020.

Rajendran S. & N. Gejeswari. 2019. Word Order Typology and Its Implication in Translation. Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:4 April 2019

Rajendran S. & G. Vasuki. 2019. English to Tamil Machine Translation System Using Parallel Corpus. Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:5 May 2019.

Rajendran and Krishanakumar K. A Comprehensive Study Of Shallow Parsing and Machine Translatin In Malayalam. Uploaded in Academia.edu.

Ray, Pradipta Ranjan; Harish V. Sudeshna Sarkar Anupam Basu Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi. Downloaded from internet on 08.07.2020.

"Ray Kurzweil biography". KurzweilAINetwork. Archived from the original on 5 February 2014. Retrieved 25 September 2014.

Ravin, Y. and Leacock, C. (Eds.) 2000. Ploysemy: Theoretical and Computational Approaches. New York: Oxford University Press Inc.

Rekha R U, Anand Kumar M, Dhanalaksmi V, Soman K P and Rajendran S. A Novel Approach to Morphological Generator for Tamil. Data Engineering and Management, Second International Conference, ICDME, Springer-Verlag, Berlin, Heidelberg, 2012, 249-251.

Remez, R.; Rubin, P.; Pisoni, D.; Carrell, T. (22 May 1981). "Speech perception without traditional speech cues" (PDF). Science. 212 (4497): 947–949. Bibcode:1981Sci...212..947R.

doi:10.1126/science.7233191. PMID 7233191. Archived from the original (PDF) on 2011-12-16. Retrieved 2011-12-14.

Ren, Yi. 2019. "Almost Unsupervised Text to Speech and Automatic Speech Recognition". arXiv:1905.06791 [cs.CL].

Renouf, A. J. 1986. "The Exploitation of a Computerized Corpus of English Text", in: M. Rivas (ed.), Actes du VÎIIème Colloque G.E.R.A.S. Paris: Université de Paris-Dauphine.

Reynolds, Douglas; Rose, Richard (January 1995). "Robust text-independent speaker identification using Gaussian mixture speaker models" (PDF). IEEE Transactions on Speech and Audio Processing. 3 (1): 72–83. doi:10.1109/89.365379. ISSN 1063-6676. OCLC 26108901. Archived (PDF) from the original on 8 March 2014. Retrieved 21 February 2020.

Rubin, P.; Baer, T.; Mermelstein, P. 1981. "An articulatory synthesizer for perceptual research".

Roget Peter Mark. 1911. Roget's Thesaurus of English Words and Phrases. TY Crowell Company, USA, 1911.

Roget, Peter Mark. 2008. Roget's International Thesaurus. 3/E**. Oxford and IBH Publishing.

Robinson T. 1992. "A real-time recurrent error propagation network word recognition system". [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 617–620 vol.1. doi:10.1109/ICASSP.1992.225833. ISBN 0-7803-0532-9.

Journal of the Acoustical Society of America. 70 (2): 321–328. Bibcode:1981ASAJ...70..321R. doi:10.1121/1.386780.

Rudnicka, Ewa; Bond, Francis; Grabowski, Łukasz; Piasecki, Maciej; Piotrowski, Tadeusz 2018. "Lexical Perspective on Wordnet to Wordnet Mapping". Proceedings of the 9th Global WordNet Conference (GWC 2018): 210.

Rundell, M. 2002. Good Old-fashioned Lexicography: Human Judgment and the Limits of Automation. In M.-H. Corréard (Ed.), Lexicography and Natural Language Processing. A Festschrift in Honour of BTS Atkins (pp. 138–155). Euralex.

Rundell, M., & Stock, P. 1992. The corpus revolution. English Today, 8(04), 45–51.

Sarangi, Susanta; Sahidullah, Md; Saha, Goutam (September 2020). "Optimization of data-driven filterbank for automatic speaker verification". Digital Signal Processing. 104. doi:10.1016/j.dsp.2020.102795.

Saveski, Martin and Igor Trajkovski. 2010. Automatic construction of wordnets by using machine translation and language modeling. In Proceedings of the 13th International MultiConference Information Society, volume C, Ljubljana, Slovenia,

Schantz, Herbert F. 1982. The history of OCR, optical character recognition. Retrieved on 12th July 2020.

Schank, Roger C. 2014. Conceptual Information Processing. New York: Elsevier. p. 5. ISBN 9781483258799.

Schmidhuber, Jürgen (2015). "Deep Learning". Scholarpedia. 10 (11): 32832. Bibcode:2015SchpJ..1032832S. doi:10.4249/scholarpedia.32832.

Schmidt, Thomas and Kai Worner. 2012. Multilingual corpora and multilingual corpus analysis. John Benjamins Publishing, Amsterdam, Netherlands.

Schubotz, Moritz; Philipp Scharpf; et al. (12 September 2018). "Introducing MathQA: a Math-Aware question answering system". Information Discovery and Delivery. Emerald Publishing Limited. 46 (4): 214–224. doi:10.1108/IDD-06-2018-0022.

Schulze, B.M. et al. 1994. “Comparitive State-of-the-art Survey and Assessment of General Interest Tools”, Technical Report D1B – I, DECIDE Project, Institute for Natural Language Processing, Stuttgart, 1994.

Schutte, John (15 October 2007). "Researchers fine-tune F-35 pilot-aircraft speech system". United States Air Force. Archived from the original on 20 October 2007.

Schütze, H. 1997. Ambiguity Resolution in Language Learning: Computational and Cognitive Models. Cambridge: Cambridge University Press.

Selkirk, E.O. 1983. The Syntax of Words. Cambridge, Mass.: MIT Press.

Selting M. and Couper-Kuhlen, E. (Eds.) 2001. Studies in Interactional Linguistics. Amsterdam/ Philadelphia: John Benjamins.

Sezgin, Mehmet; Sankur, Bulent (2004). "Survey over image thresholding techniques and quantitative performance evaluation" (PDF). Journal of Electronic Imaging. 13 (1): 146. Bibcode:2004JEI....13..146S. doi:10.1117/1.1631315. Archived from the original (PDF) on October 16, 2015. Retrieved May 2, 2020.

Sharman, R. 1990. Hidden Markov model methods for word tagging. Report 214. Winchester: IBM UK Scientific Centre.

Shillingford, Brendan; Assael, Yannis; Hoffman, Matthew W.; Paine, Thomas; Hughes, Cían; Prabhu, Utsav; Liao, Hank; Sak, Hasim; Rao, Kanishka (13 July 2018). "Large-Scale Visual Speech Recognition". arXiv:1807.05162 [cs.CV].

Simmons, Robert F. 1963. "Synthetic language behavior". Data Processing Management. 5 (12): 11–18.

Simmons, Robert F. 1982."Themes From 1972." In: 20th Annual Meeting of the Association for Computational Linguistics. Toronto, Ontario, Canada: Association for Computational Linguistics. Pages 100–101, URL: https://www.aclweb.org/anthology/P82-1022.

Sinclair, J.M. 1985. "Lexicographic Evidence", in: R. Ilson (ed.), Dictionaries, Lexicography and Language Learning. Oxford: Pergamon.

Sinclair, J.M. ed. 1987. Looking Up: An Account ofthe COBUILD Project in Lexical Computing. London/Glasgow: Collins.

Sinclair, J. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sinclair, J. 1992. The Automatic Analysis of Corpora', in Svartvik, J (ed.) Directions in Corpus Linguistics (Proceedings of Nobel Symposium 82). Berlin: Mouton de Gruyter.

Smith, L. (Ed.) 1981. English for Cross-cultural Communication. London: Macmillan.

Smith, Ray. 2007. "An Overview of the Tesseract OCR Engine" (PDF). Archived from the original (PDF) on September 28, 2010. Retrieved on May 23, 2020.

Soergel, Dagobert. 1974. Indexing languages and thesauri: construction and maintenance. Melville Pub. Co., New York, USA, 1974.

Souter, C. and Atwell, E. (Eds.) 1993. Corpus Based Computational Linguistics. Amsterdam: Rodopi.

Sowa, John F. 1987. "Semantic Networks". In Stuart C Shapiro (ed.). Encyclopedia of Artificial Intelligence. Retrieved 29 April 2008.

Sowa, John F. (ed.). 1991. Principles of Semantic Networks: Explorations in the Representation of Knowledge. Morgan Kaufmann Publishers, INC, San Mateo, California.

"Speaker Identification (WhisperID)". Microsoft Research. Microsoft. Archived from the original on 25 February 2014. Retrieved 21 February 2014.

"Speaker Independent Connected Speech Recognition - Fifth Generation Computer Corporation". Fifthgen.com. Retrieved on 15 June 2020.

"Speech Recognition for Learning". National Center for Technology Innovation. 2010. Retrieved 26 March 2020.

"Speech synthesis". World Wide Web Organization. Retrieved on March 26, 2020.

Spencer, A. 1991. Morphological Theory. Oxford: Basil Blackwell.

SPIEGEL ONLINE, Hamburg, Germany (13 September 2013). "Google Translate Has Ambitious Goals for Machine Translation". SPIEGEL ONLINE.

Sperberg-Mcqueen, C.M. and Burnard, L. 1994. Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: ACH-ACL-ALLC Text Encoding Initiative.

Stenström, A-B, Andersen, G. and Has und, I.K. 2002. Trends in Teenage Talk: Corpus Compilation, Analysis and Findings. Amsterdam: John Benjamins.

Stern, W. 1924. Psychology of Early Childhood up to Six Years of Age. New York; Holt.

Stubbs, M. 1996. Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture. Oxford: Blackwell. Publishers.

Summers, D. 1991. Longman/Lancaster English Language Corpus: Criteria and Design. Harlow: Longman.

Sun X. and H. Zhuge, Summarization of Scientific Paper through Reinforcement Ranking on Semantic Link Network, IEEE ACCESS, 2018, doi:10.1109/ACCESS.2018.2856530.

Sun, Haitian; Dhingra, Bhuwan; Zaheer, Manzil; Mazaitis, Kathryn; Salakhutdinov, Ruslan; Cohen, William .2018. "Open Domain Question Answering Using Early Fusion of Knowledge

Bases and Text". Association for Computational Linguistics. Brussels, Belgium: 4231–4242. arXiv:1809.00782.

Sussna, Michael. "Word sense disambiguation for free-text indexing using a massive semantic network." Proceedings of the second international conference on Information and knowledge management. ACM, 1993.

Steyvers, M.; Tenenbaum, J.B. 2005. "The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth". Cognitive Science. 29 (1): 41–78. arXiv:cond-mat/0110012. doi:10.1207/s15516709cog2901_3. PMID 21702767.

Storey, M. A.; B. Phillips, M. Maczewski and M. Wang “Evaluating the usability of Web-based learning tools .“ Retrieved on 22 July 2020.

Stuart M. Shieber .1992. Constraint-based Grammar Formalisms: Parsing and Type Inference for Natural and Computer Languages. MIT Press. ISBN 978-0-262-19324-5.

Suen, C.Y.; Plamondon, R.; Tappert, A.; Thomassen, A.; Ward, J.R.; Yamamoto, K. (May 29, 1987). Future Challenges in Handwriting and Computer Applications. 3rd International Symposium on Handwriting and Computer Applications, Montreal, May 29, 1987. Retrieved on July 3, 2020.

Sutskever, Ilya; Vinyals, Oriol; Le, Quoc Viet .2014. "Sequence to sequence learning with neural networks". arXiv:1409.3215 [cs.CL].

Svartvik, J. (Ed.) 1990. The London Corpus of Spoken English: Description and Research. Lund Studies in English 82. Lund: Lund University Press.

Svartvik, J. (Ed.) 1992. Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 - Stockholm, 4-8 August 1991. Berlin, New York: Mouton De Gruyter.

Svartvik, J. and Qurick, R. 1980. A Copurs of English Conversation. Lund: C.W.K. Gleerup.

Swigger, Kathleen. "Semantic.ppt". Retrieved on August 23, 2020.

Steyvers, M.; Tenenbaum, J.B. (2005). "The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth". Cognitive Science. 29 (1): 41–78. arXiv:cond-mat/0110012. doi:10.1207/s15516709cog2901_3. PMID 21702767.

Tang, K. W.; Kamoua, Ridha; Sutan, Victor. 2004. "Speech Recognition Technology for Disabilities Education". Journal of Educational Technology Systems. 33 (2): 173–84. CiteSeerX 10.1.1.631.3736. doi:10.2190/K6K8-78K2-59Y7-R9R2.

Tannen, D. (Ed.) 1982. Spoken and Written Language: Exploring Orality and Literacy. Norwood, New Jersey: Ablex Publishing Corporation.

Tapanainen, P. and Jarvinen, T. 1997. “A non-projective dependency parser”, Proceedings of the 5th Conference on Applied Natural Language Processing, Washington D.C.,1997.

Tapanainen, P. and Voutilainen, A. 1994. “Tagging Accurately – Don’t Guess If You Don’t Know”, Proceedings of 4th ACL Conference on Applied Natural LanguageProcessing, ACM, Stuttgart, 1994.

Tappert, C. C.; Suen, C. Y.; Wakahara, T. 1990. "The state of the art in online handwriting recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence. 12 (8): 787. doi:10.1109/34.57669.

Taylor, Paul .2009. Text-to-speech synthesis. Cambridge, UK: Cambridge University Press. p. 3. ISBN 9780521899277.

"The History of OCR". Data Processing Magazine. 12: 46. 1970.

"The HMM-based Speech Synthesis System". Hts.sp.nitech.ac.j. Retrieved on June 22, 2020

"The Mailbag LG #114". Linuxgazette.net. Retrieved 15 June 2020.

Tu, Zhaopeng; Lu, Zhengdong; Liu, Yang; Liu, Xiaohua; Li, Hang. 2016. "Modeling Coverage for Neural Machine Translation". arXiv:1601.04811 [cs.CL].

"the Association for Computational Linguistics – 2003 ACL Lifetime Achievement Award". Association for Computational Linguistics. Archived from the original on 12 June 2010. Retrieved on 10 March 2020. "Kitt.cl.uzh.ch [CL Wiki]" (PDF).

Thai, Perishan."An Introduction to Cross-Language Information Retrieval Approaches". Web. Web.simmons.edu

Thomas, J. and Short, M. (Eds.) 1996. Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech. London and New York: Addison Wesley Longman.

Tognini-Bonelli, E. 2001. Corpus Linguistics at Work. Amsterdam: John Benjamins.

Universal POS tags.

Trier, Oeivind Due; Jain, Anil K. 1995. "Goal-directed evaluation of binarisation methods" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 17 (12): 1191–1201. doi:10.1109/34.476511. Retrieved May 2, 2020.

Trivi, Jean-Michel (2009-09-23). "An introduction to Text-To-Speech in Android". Android-developers.blogspot.com. Retrieved on July 20, 2020.

Tüske, Zoltán; Golik, Pavel; Schlüter, Ralf; Ney, Hermann. 2014. "Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR" (PDF). Interspeech 2014. Retrieved on May 2, 2020.

"Using machine translation in clinical practice".

Uta, F. (Ed.) 1979. Cognitive Processes in Spelling. London: Academic Press.

Vaishnavi Ramaswamy. 2003. A Morphological Analyzer for Tamil. Ph.D Dissertaion Subbmitted to University of Hyderabad.

Vaishnavi, Ramaswamy. 2003. “Parsing in AMPLE, KIMMO & PERL: Nouns in Tamil”.

Vaishnavi, T. and Roxanna Samuel. 2016. "Individual Document Keyword Extraction for Tamil". International Journal of Computer Technology & Applications,Vol 7(3), 448-452 IJCTA May-June 2016 Available online@www.ijcta.com

Vaishnavi, T. and Roxanna Samuel. 2016. "Individual Document Keyword Extraction for Tamil". International Journal of Computer Technology & Applications,Vol 7(3), 448-452 IJCTA May-June 2016 Available online@www.ijcta.com.

Van Santen, J. (April 1994). "Assignment of segmental duration in text-to-speech synthesis". Computer Speech & Language. 8 (2): 95–128. doi:10.1006/csla.1994.1005.

Van Santen, Jan P. H.; Sproat, Richard W.; Olive, Joseph P.; Hirschberg, Julia (1997). Progress in Speech Synthesis. Springer. ISBN 978-0-387-94701-3.

Van de Riet, R. P. 1992. Linguistic Instruments in Knowledge Engineering (PDF). Elsevier Science Publishers. p. 98. ISBN 978-0444883940.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. 2005. Parallel corpora for medium density languages. In Proceedings of the RANLP 2005 (pp. 590–596). Borovets.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia (2017-12-05). "Attention Is All You Need". arXiv:1706.03762 [cs.CL]. Retrived on July 12, 2020.

"Voice Recognition To Ease Travel Bookings: Business Travel News". Retrived on July 12, 2020.

Vossen, Piek. 1998. A multilingual database with lexical semantic networks. Kluwer Academic Publishers, Dordrecht, Netherlands.

Vossen, P. Ed. 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer, Dordrecht, The Netherlands.

Vossen, Piek. Ontologies. 2003. In: Mitkov, Ruslan (ed.) (2003): Handbook of Computational Linguistics, Chapter 25. Oxford: Oxford University Press.

Vossen, Piek. 2005. Building Wordnets. http://www.globalwordnet.org/gwa/ BuildingWordnets.ppt,

Vera, D.E.J. (Ed.) 2002. A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics. Amsterdam: Rodopi.

Véronis, J. (Ed.) 2000. Parallel Text Processing: Alignment and Use of Translation Corpora. Dordrecht: Kluwer Academic Publishers.

Veronis, Jean 2000. From the rosetta stone to the information society: A survey of parallel text processing. In Jean Veronis, ´ editor, Parallel Text Processing, pages 1– 25. Kluwer.

V´eronis, Jean. 2001. Parallel Text Processing: Alignment and Use of Translation Corpora. Computational Linguistics. 27. Dordrecht: Kluwer Academic Publishers (Text, speech and language technology series, edited by Nancy Ide and Jean V´eronis, volume 13), 2000, xxiii+402 pp; hardbound. pp. 592–595. doi:10.1162/coli.2000.27.4.592. ISBN 978-0-7923-6546-4. S2CID 14796449.

"voice recognition, definition of". WebFinance, Inc. Retrieved on February 21, 2020.

Waibel, Alex. 1989. "Modular Construction of Time-Delay Neural Networks for Speech Recognition" (PDF). Neural Computation. 1 (1): 39–46. doi:10.1162/neco.1989.1.1.39. Retrieved form 29 June 2020.

Waibel, Hanazawa, Hinton, Shikano, Lang. 1989. "Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing."

Waibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K.; Lang, K. J. 1989. "Phoneme recognition using time-delay neural networks". IEEE Transactions on Acoustics, Speech, and Signal Processing. 37 (3): 328–339. doi:10.1109/29.21701. hdl:10338.dmlcz/135496.

Wallis, S. And Nelson G. 2001. Knowledge discovery in grammatically analysed corpora. Data Mining and Knowledge Discovery, 5: 307–340.

Wang, William Yang and Kallirroi Georgila. 2011. Automatic Detection of Unnatural Word-Level Segments in Unit-Selection Speech Synthesis, IEEE ASRU 2011.

Wang, Jianqiang, and Douglas W. Oard. 2012. "Matching meaning for cross-language information retrieval." Information Processing & Management48.4 (2012): 631-53.

"WaveNet: A Generative Model for Raw Audio". Deepmind.com. 2016-09-08. Retrieved on July 2020.

Way, Andy; Nano Gough (20 September 2005). "Comparing Example-Based and Statistical Machine Translation". Natural Language Engineering. 11 (3): 295–309. doi:10.1017/S1351324905003888.

Weigand, E. and Dascal, M. (Eds.) 2001. Negotiation and Power in Dialogic Interaction. Amsterdam/Philadelphia: John Benjamins.

"WebBANC: Building Semantically-Rich Annotated Corpora from Web User Annotations of Minority Languages" (PDF).

Wichmann, A., Fligelstone, S., Mcenery, T. and Knowles, G. (Eds.) 1997. Teaching and Language Corpora. London: Longman.

Wills, J.D. 1990. The Lexical Syllabus. London: Collins.

Winograd, T. 1983. Language as a Cognitive Process. Vol. I. Mass.: Addison-Wesley.

Wiseby, R.A. (Ed.) 1971. Computer in Literary and Linguistic Research. Papers from the Cambridge Symposium. Cambridge: Cambridge University Press.

Wołk, Krzysztof; Marasek, Krzysztof. 2015. "Neural-based Machine Translation for Medical Text Domain. Based on European Medicines Agency Leaflet Texts". Procedia Computer Science. 64 (64): 2–9.

Woods, William A; Kaplan, R. 1977. "Lunar rocks in natural English: Explorations in natural language question answering". Linguistic Structures Processing 5. 5: 521–569.

Wooten, Adam. 2006. "A Simple Model Outlining Translation Technology" T&I Business (February 14, 2006)". Tandibusiness.blogspot.com. Retrieved 12 June 2020.

“WordNet” from Wikipedia retrieved on 22 July 2020.

Wu, J.; Chan, C. 1993. "Isolated Word Recognition by Neural Network Models with Cross-Correlation Coefficients for Speech Dynamics". IEEE Transactions on Pattern Analysis and Machine Intelligence. 15 (11): 1174–1185. doi:10.1109/34.244678.

Xu, X. 1996, Building and Aligning an English-Chinese Parallel Corpus. Unpublished MA Dissertation. The University of Lancaster. UK.

Yamamoto, Ryuichi .2019. "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram". arXiv:1910.11480 [eess.AS].

Yarowsky, D. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proc. of the 14th conference on Computational linguistics (COLING), 1992.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics.

Yih, Wen-tau, Xiaodong He, and Christopher Meek. "Semantic parsing for single-relation question answering." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014.

Young, S. and G. Bloothooft (Eds.) 1997. Corpus-Based Methods in Language and Speech Processing. Vol-II. Dordrecht: Kluwer Academic Press.

Yu, D.; Deng, L.; Dahl, G. 2010. "Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition" (PDF). NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

Yu, D.; Deng, L. 2014. "Automatic Speech Recognition: A Deep Learning Approach (Publisher: Springer)".

Yule, G.U. 1964. The Statistical Study of Literary Vocabulary. Cambridge: Cambridge University Press.

Zahorian, S. A.; A. M. Zimmer, and F. Meng, 2002. "Vowel Classification for Computer based Visual Feedback for Speech Training for the Hearing Impaired," in ICSLP 2002.

Zanettin, Federico. 1998. Bilingual comparable corpora and the training of translators. Meta 43, (4):616-630.

Zernik, U. (Ed.) 1991. Lexical Acquisition. Englewood Cliff, NJ: Erlbaum.

Zhang, Julia. 2002. Language Generation and Speech Synthesis in Dialogues for Language Learning, masters thesis, Section 5.6 on page 54.

Zhao, L., Kipper, K., Schuler, W., Vogler, C., & Palmer, M. (2000). A Machine Translation System from English to American Sign Language. Lecture Notes in Computer Science, 1934: 54–67.

Zhu, Linchao, et al. "Uncovering the temporal context for video question answering." International Journal of Computer Vision 124.3 (2017): 409-421.

Zhuge, H. Knowledge Grid, World Scientific Publishing Co. 2004.

Zhuge, H. Inheritance rules for flexible model retrieval. Decision Support Systems 22(4)(1998)379–390

Zhuge, H. Active e-document framework ADF: model and tool. Information & Management 41(1): 87–97 (2003)

Zhuge H.and Zheng, L.Ranking Semantic-linked Network, WWW 2003

H.Zhuge, The Semantic Link Network, in The Knowledge Grid: Toward Cyber-Physical Society, World Scientific Publishing Co. 2012.

Zhuge, H.; L. Zheng, N. Zhang and X. Li, 2004. An automatic semantic relationships discovery approach. WWW 2004: 278–279.

Zhuge, H. 2009. Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning, IEEE Transactions on Knowledge and Data Engineering, 21(6)(2009)785–799.

Zhuge, H. 2011. Semantic linking through spaces for cyber-physical-socio intelligence: A methodology, Artificial Intelligence, 175(2011)988–1019.

Zhuge, H. 2016. Multi-Dimensional Summarization in Cyber-Physical Society, Morgan Kaufmann, 2016.

Zhuge, H. 2008. The Web Resource Space Model, Springer, 2008.

Zhuge, H. 2020. Cyber-Physical-Social Intelligence on Human-Machine-Nature Symbiosis, Springer, 2020.

Zhuge H.and Y.Xing, 2012. Probabilistic Resource Space Model for Managing Resources in Cyber-Physical Society, IEEE Transactions on Service Computing, 5(3)(2012)404–421.

Zipf, G.K. 1936. The Psycho-biology of Language: An Introduction to Dynamic Philology. London: G. Routledge.

Zipf. G.K. 1949. Human Behaviour and the Principle of Least Effort: An Introduction of Human Ecology. Cambridge, Mass.: Addison-Wesley.

Zurini, Madalina. Word sense disambiguation using aggregated similarity based on WordNet graph representation. Informatica Economica.

==========